International Society for Photogrammetry and Remote Sensing
Société Internationale de Photogrammétrie et de Télédétection
Internationale Gesellschaft für Photogrammetrie und Fernerkundung

INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES
ARCHIVES INTERNATIONALES DE PHOTOGRAMMÉTRIE, DE TÉLÉDÉTECTION ET DE SCIENCES DE L'INFORMATION SPATIALE
INTERNATIONALES ARCHIV FÜR PHOTOGRAMMETRIE, FERNERKUNDUNG UND RAUMBEZOGENE

Volume XXXVIII – Part 2

1910 2010

isprs

a century of information from imagery

Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science

ISPRS Technical Commission II Symposium
IGU International Symposium on Spatial Data Handling
IGU International Conference on Modelling Geographical Systems

Hong Kong
26 – 28 May 2010

Editors

Eric Guilbert, ISPRS Technical Commission II
Brian Lees, IGU Commission on Geographical Information Science
Yee Leung, IGU Commission on Modelling Geographical Systems

Organisers

ISPRS Technical Commission II Theory and Concepts of Spatial Information Science
International Geographical Union Commission on Geographic Information Science
International Geographical Union Commission on Modelling Geographical System

ISSN 1682-1750 (book)
ISSN 1682-1777 (CD)

Published by
ISPRS Technical Commission II

ISPRS Headquarters 2008-2012
c/o CHEN JUN, ISPRS Secretary GeneralNational Geomatics Centre of China
No. 1 Baishengcun, Zizhuyuan
Beijing 100048, PR CHINA
Tel: +86 10 6842 4072
Fax: +86 10 6842 4101
Email: chenjun@nsdi.gov.cn
ISPRS WEB Homepage: http://www.isprs.org

Available from
GITC bv
P.O.Box 112
8530 AC Lemmer
The Netherlands
Tel: +31 (0) 514 56 18 54
Fax: +31 (0) 514 56 38 98
E-mail: mailbox@gitc.nl
Website: www.gitc.nl

TABLE OF CONTENTS

INTRODUCTION

KEYNOTE ADDRESSES

SESSION 1 – SPATIAL MODELLING

SESSION 2 - SPACE AND TIME

SESSION 3 - SPATIAL ANALYSIS

SESSION 4 - SPATIAL DATA MINING

SESSION 5 - UNCERTAINTY MODELLING

SESSION 6 - SPATIAL DATABASE

SESSION 7 - ADVANCES IN CARTOGRAPHY

SESSION 8 - LOCATION-BASED SERVICES

SESSION 9 - MOBILE DATA MODELS

SESSION 10 - SPATIAL DATA PROCESSING ALGORITHMS

SESSION 11 - WEB GIS

SESSION 12 - GEO-VISUALIZATION

SESSION 13 - SPATIAL INFORMATION FOR ENVIRONMENTAL STUDIES

SESSION 14 - SPATIAL INFORMATION FOR LAND USE STUDY

SESSION 15 - APPLICATION OF GIS AND REMOTE SENSING

Organising Committee

Wenzhong Shi (Chair), The Hong Kong Polytechnic University
Qiming Zhou (Co-Chair), Hong Kong Baptist University
Yee Leung (Co-Chair), The Chinese University of Hong Kong
Chenghu Zhou (Co-chair), Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
Eric Guilbert (Secretary), The Hong Kong Polytechnic University
Mang Lung Cheuk, Hong Kong Baptist University
Bo Huang, The Chinese University of Hong Kong
Brian Lees, University of New South Wales, Australia
Bo Wu, The Hong Kong Polytechnic University
Anthony Yeh, The University of Hong Kong

Programme Committee

Yee Leung (Chair), Hong Kong
Qiming Zhou (Co-Chair), Hong Kong
Brian Lees (Co-Chair), Australia,
Wenzhong Shi (Co-Chair), Hong Kong

Ozgun Akcay, Turkey
Masatoshi Arikawa, Japan
Ali Bennasr, Tunisia
Jean-Paul Bord, France
Tao Cheng, UK
Christophe Claramunt, France
Graham Clarke, UK
Hande Demirel, Turkey
Manfred M. Fischer, Austria
Stewart Fotheringham, Ireland
Andrew Frank, Austria
Tung Fung, Hong Kong
Christopher Gold, UK
Michael Goodchild, USA
Daniel A. Griffith, USA
Lars Harrie,Sweden
Francis Harvey, USA
Andy Hudson-Smith, UK
Gerhard Joos, Denmark
Pavlos Kanaroglou, Canada
Marinos Kavouras, Greece
Zhilin Li, Hong Kong
Hui Lin, Hong Kong
Yaolin Liu, China
Shattri Mansor, Malaysia
Lingkui Meng, China
Martien Molenaar, Netherlands
Mir Abolfazl Mostafavi, Canada
Christopher J. Pettit, Australia
Alias Abdul Rahman, Malaysia
Docent Juri Roosaare, Estonia
Anne Ruas, France
Yukio Sadahiro, Japan
Mauro Salvemini, Italy
Nadine Schuurman, Canada
Monika Sester, Germany
A. Rashid B. M. Shariff, Malaysia
Ali Sharifi, Netherlands
Ryosuke Shibasaki, Japan
Ake Sivertun, Sweden
Therese Steenberghen, Belgium
Alfred Stein, Netherlands
Vladimir Tikunov, Russia
Abdülvahit Torun, Turkey
Vit Vozenilek, Czech Republic
Jinfeng Wang, China
Shuliang Wang, China
Robert Weibel, Switzerland
Stephan Winter, Australia
Michael Worboys, USA
Anthony Yeh, Hong Kong
Chenghu Zhou, China

Invited Reviewers

Jean-Philippe Aurambout, Australia
Itzhak Benenson, Israel
Pawel Boguslawski, UK
Lei Chen, Hong Kong
Kun Cheng, Hong Kong
Mang Lung Cheuk, Hong Kong
M.R Delavar, Iran
Trudie Dockerty, UK
Matt Duckham, Australia,
Paul Gessler, USA
Eric Guilbert, Hong Kong
Qingsheng Guo, China
Jan-Henrik Haunert, Germany
Bo Huang, Hong Kong
Margarita Kokla, Greece
Thangavelu Kumaran, India
Rainer Laudien, Germany
I-Chieh Lee, USA
Rong Li, USA
Jie Lian, Canada
Steve Liang, Canada
Chun Liu, China
Desheng Liu, USA
Yaolin Liu, China
Janet Nichol, Hong Kong
Xutong Niu, USA
Henk Ottens, Netherland
Lilian Pun, Hong Kong
Inge Sandholt, Denmark
Anthony Stefanidis, USA
Georg Treu, Germany
Caixia Wang, USA
Shuliang Wang, China
King Wong, Hong Kong
Bo Wu, Hong Kong
Bisheng Yang, China
Wenze Yang, USA
Qing Zhu, China

Host organisations

The Hong Kong Polytechnic University
The Chinese University of Hong Kong
Hong Kong Baptist University
Chinese Academy of Science
The University of Hong Kong

Funding organisations

Faculty of Construction and Land Use, The Hong Kong Polytechnic University
Chung Chi College and the Faculty of Social Science, CUHK
Department of Geography, Hong Kong Baptist University
State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences

Supporting organisations

The Hong Kong Institute of Surveyors
Hong Kong Geographical Association
Hong Kong GIS Association

Sponsors

ESRI
Institute of Space and Earth Information Science, CUHK
Department of Land Surveying and Geo-Informatics, PolyU
Leica Geosystems
South Survey and Mapping Instrument
Pitney Bowes
Taylor & Francis

INTRODUCTION

The **Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science** was held on May 26th to 28th 2010 in The Hong Kong.Polytechnic University This conference was organised by: Commission II of the International Society of Photogrammetry and Remote Sensing (ISPRS) and Commission of Geographic Information Science and Commission of Modelling Geographical Systems of the International Geographical Union (IGU). The conference joined together the **Symposium of Technical Commission II of ISPRS**, the **Symposium on Spatial Data Handling** and the conference on **Modelling Geographical Systems** from IGU.

ISPRS is a society regrouping scientific societies from more than 100 countries working in domains related to photogrammetry, remote sensing and geographical information science. 2010 will be a special year for the society as it will be celebrating its centenary anniversary. The **ISPRS Technical Commission II Symposium** is organised every four years alternately with the ISPRS Congress and is among the major events in ISPRS calendar regrouping leading scholars from the GISc and related communities. Last editions of the symposium were held in Ottawa, Canada (2002) and Vienna, Austria (2006) jointly with SDH.

The **14th International Symposium on Spatial Data Handling** (SDH) is the premier biennial international research forum for Geospatial Information Science (GISc). It commenced in 1984, in Zurich, Switzerland and has been held in Seattle, USA; Sydney, Australia; Zurich, Switzerland; Charleston, USA; Edinburgh, UK; Delft, The Netherlands; Vancouver, Canada; Beijing, China; Ottawa, Canada; Leicester, UK; Vienna, Austria; and Montpellier, France.

The IGU Commission on **Modelling Geographical Systems** has been organising its conferences over the years on the modeling and analysis of geographical data and systems in different parts of the world. It is the first time that it joins hand with ISPRS and IGU Commission on Geographical Information Science to co-organise the conference to encourage interaction and collaboration among researchers and professionals in the three organisations.

These proceedings gather the papers presented during the conference including keynote addresses by renowned experts in the field of GeoSpatial Information Science and research papers organised in theme sessions.

TWENTY YEARS OF PROGRESS: GISCIENCE IN 2010

Michael F. Goodchild

University of California, Santa Barbara

ABSTRACT

The concept of a science of geographic information has its roots in the early 1990s, and discussions over whether GIS is more than a tool. Three major lines of thought developed at that time: those centered on the individual, on society, and on technology. Many substantial research results have been obtained in all three areas.

Today geospatial technology is more important than ever, and new research directions are emerging, again based in the same conceptual framework. The presentation ends with some speculations on the future of GIScience.

THE NEW ERA FOR GEO-INFORMATION

Deren Li

Wuhan University

ABSTRACT

Along with the forthcoming of Google Earth, Virtual Earth, the next generation of Internet, Web 2.0, Grid Computing and smart sensor web, comes the new era for Geo-Information. In this paper, main features of new Geo-Information era are discussed. This new era is characterized by these features: serviced users are extended from professionals to all public users, the users are data and information providers as well, provided geospatial data are no longer measurement-by-specification but measurement-on-demand through smart sensor web, and services are transferred from data-driving to application-driving. Such problems as out-of-order issues in geographic data collection and information proliferation, quality issues in geographic information updating, security issues in geographic information services, privacy issues in sharing geographic information and property issues on sharing geographic information, which are brought about by new geo-information era, especially problems and challenges confronted in geo-information science and geo-spatial information industry, are analyzed. Then strategies concerning standards, planning, laws, technology and applications are proposed.

PRINCIPLES OF NEURAL SPATIAL INTERACTION MODELLING

Manfred M. Fischer
Vienna University of Economics and Business
manfred.fischer@wu.ac.at

Keywords: Spatial interaction, neural networks, non-linear function approximation, model performance, network learning
JEL Classification: C31, C45, R19

ABSTRACT:

The focus of this paper is on the neural network approach to modelling origin-destination flows across geographic space. The novelty about neural spatial interaction models lies in their ability to model non-linear processes between spatial flows and their determinants, with few – if any – a priori assumptions of the data generating process. The paper draws attention to models based on the theory of feedforward networks with a single hidden layer, and discusses some important issues that are central for successful application development. The scope is limited to feedforward neural spatial interaction models that have gained increasing attention in recent years. It is argued that failures in applications can usually be attributed to inadequate learning and/or inadequate complexity of the network model. Parameter estimation and a suitably chosen number of hidden units are, thus, of crucial importance for the success of real world applications. The paper views network learning as an optimization problem, describes various learning procedures, provides insights into current best practice to optimize complexity and suggests the use of the bootstrap pairs approach to evaluate the model's generalization performance.

1. INTRODUCTION

The development of spatial interaction models is one of the major intellectual achievements and, at the same time, perhaps the most useful contribution of spatial analysis to social science literature. Since the pioneering work of Wilson (1970) on entropy maximization, there have been surprisingly few innovations in the design of spatial interaction models. Fotheringham's (1983) competing destinations version, Griffith's eigenvector spatial filter versions[1] (see Griffith 2003; Fischer and Griffith 2008), the spatial econometric interaction models[2] (see LeSage and Pace 2009; LeSage and Fischer 2010), and neural network based (briefly neural) spatial interaction models (see Fischer and Gopal 1994; Fischer 2002) are the principal exceptions.

The focus in this paper is on neural networks as efficient non-linear tools for modelling interactions across geographic space. The term "neural network" has its origins in attempts to find mathematical representations of information processing in the study of natural neural systems (McCulloch and Pitts 1943; Rosenblatt 1962). Indeed, the term has been used very broadly to include a wide range of different model structures, many of which have been the subject of exaggerated claims to mimic neurobiological reality[3]. As rich as neural networks are, they still ignore a host of biologically relevant features. From the perspective of applications in spatial interaction modelling, however, neurobiological realism is not necessary. In contrast, it would impose entirely unnecessary constraints.

From the statistician's point of view neural network models are analogous to non-parametric, non-linear regression models. The novelty about neural spatial interaction models lies in their ability to model non-linear processes with few – if any – *a priori* assumptions about the nature of the data generating process. We limit ourselves to models known as feedforward neural models[4]. Spatial interaction models of this kind can be viewed as a general framework for non-linear function approximation where the form of the mapping is governed by a number of adjustable parameters. The network inputs are origin, destination and separation variables, and the network weights the model parameters.

[1] Eigenvector spatial filtering (see Griffith 2003) enables spatial autocorrelation effects to be captured, and shifts attention to spatial autocorrelation arising from missing origin and destination factors reflected in flows between pairs of locations.

[2] Note that spatial econometric interaction models are – in general – formally equivalent to regression models with spatially autocorrelated errors, but differ in terms of the data analysed and the manner in which the spatial weights matrix is defined.

[3] Neural networks can model cortical local learning and signal processing, but they are not the brain, neither are many special purpose systems to which they contribute (Weng and Hwang 2006).

[4] Feedforward neural networks are sometimes also called multilayer perceptrons even though the term perceptron is usually used to refer to a network with linear threshold gates rather than with continuous non-linearities. Radial basis function networks, recurrent networks rooted in statistical physics, self-organizing systems and ART [Adaptive Resonance Theory] models are other important classes of neural networks. For a fuzzy ARTMAP multispectral classifier see, for example, Gopal and Fischer (1997).

The paper is organized as follows. The next section continues to provide the context in which neural spatial interaction modelling is considered. Neural spatial interaction models that have a single hidden layer architecture with K input nodes (typically, K=3) and a single output node are described in some detail in Section 3. They represent a rich and flexible class of universal approximators. Section 4 proceeds to view the problem of determining the network parameters within a framework that involves the solution of a non-linear optimization problem with an objective function that recognizes the integer nature of the origin-destination flows. The section that follows reviews some of the most important training (learning) procedures and modes that utilize gradient information for solving the problem. This requires the evaluation of derivatives of the objective function – known as error or loss function in the machine-learning literature[5] – with respect to the network parameters.

Section 6 addresses the issue of network complexity and briefly discusses some techniques to determine the number of hidden units. This problem is shown to essentially consist of optimizing the complexity of the neural spatial interaction model (complexity in terms of free parameters) in order to achieve the best generalization performance. Section 7 then moves attention to the issue of how to appropriately test the generalization performance of the estimated neural spatial interaction model. Some conclusions and an outlook for the future are given in the final section.

2. CONTEXT

Spatial interaction models of the gravity type represent a class of models used to explain origin-destination flows across geographic space. Examples include migration, journey-to-work and shopping flows, trade and commodity flows, information and knowledge flows. Origin and destination locations of interaction represent points or areas (regions) in geographic space. Such models typically recognize three types of factors to explain mean interaction frequencies between origin and destination locations: (i) origin-destination variables that characterize the way spatial separation of origins from destinations constrains or impedes the interaction, (ii) origin-specific variables that characterize the ability of the origins to produce or generate flows, and (iii) destination-specific variables that represent the attractiveness of destinations.

Suppose we have a spatial system consisting of n regions, where i denotes the origin region $(i=1, ..., n)$ and j the destination region $(j=1, ..., n)$. Let $m(i, j)$ $(i, j=1, ..., n)$ denote observations on random variables, say $M(i, j)$, each of which corresponds to flows of people, commodities, capital, information or knowledge from region i to region j. The $M(i, j)$ are assumed to be independent random variables. They are sampled from a specified probability distribution that is dependent upon some mean, say $\mu(i, j)$. Let us assume that no a priori information is given about the row and column totals of the observed flow matrix $[m(i, j)]$. Then the mean interaction frequencies between origin i and destination j may be modelled by

$$\mu(i, j)=C\ A(i)^{\alpha}\ B(j)^{\beta}\ F(i, j) \qquad i, j=1, ..., n \quad (1)$$

where $\mu(i, j)=E[M(i, j)]$ is the expected flow, C denotes a constant term, the quantities $A(i)$ and $B(j)$ are called origin and destination variables, respectively. α and β indicate their relative importance, and $F(i, j)$ represents a distance deterrence function that constitutes the very core of spatial interaction models. Hence, a number of alternative specifications of $F(\cdot)$ have been proposed in the literature (see, for example, Sen and Smith 1995, pp. 92-99). But the negative exponential function is the most popular choice (with theoretical relevance from a behavioural viewpoint):

$$F(i, j)=\exp[-\theta\ d(i, j)] \qquad i, j=1, ..., n \qquad (2)$$

where θ denotes the so-called distance sensitivity parameter that has to be estimated.

Inserting Eq. (2) into Eq. (1) yields the well known class of exponential spatial interaction models that can be expressed equivalently as a log-additive model of the form

$$Y(i, j)=\kappa+\alpha\ a(i)+\beta\ b(j)+\theta\ d(i, j)+\varepsilon(i, j) \qquad (3)$$

where $Y(i, j)\equiv\log[\mu(i, j)]$, $\kappa\equiv\log C$, $a(i)\equiv\log[A(i)]$ and $b(j)\equiv\log[B(j)]$. Of note is that the back transformation of this log-linear specification results in an error structure of the exponential spatial interaction model being multiplicative. The parameters κ, α, β and θ have to be estimated if future flows are to be predicted.

There are n^2 equations of the form (3). Using matrix notation we may write these equations more compactly as

$$Y=X\ \theta+\varepsilon \qquad (4)$$

where Y denotes the N-by-1 vector of observations on the interaction variable, with $N=n^2$ (see Table 1 for the data organization convention). X is the N-by-4 matrix of observations on the explanatory variables including the origin, destination, separation variables, and the intercept. θ is the associated 4-by-1 parameter vector, and the N-by-1 vector $\varepsilon=[\varepsilon(1,1), ..., \varepsilon[n, n)]^T$ denotes the vectorized form of $[\varepsilon(i, j)]$.

If the spatial interaction model given by Eq. (4) is correctly specified, then provided that the regressor variables are not perfectly collinear, θ is estimable under the assumption that the error terms are *iid* with zero mean and constant variance, and the OLS estimator is the best linear unbiased estimator. A violation of these assumptions may lead to spatial autocorrelation.

[5] We will use the terms error function, loss function and cost function interchangeably in this paper.

It is noteworthy that the above spatial interaction model can not guarantee that the predicted flows when summed by rows or columns of the spatial interaction data matrix will necessarily have the property to match observed totals leaving the origins *i* $(i = 1, ..., n)$ or terminating at the destinations j $(j = 1, ..., n)$ in the given spatial interaction system. If the outflow totals for each origin zone and/or the inflow totals into each destination zone are a priori known, then the log-linear model given by Eq. (4) would need to be modified to incorporate the explicitly required constraints to match exact totals. Imposing origin and/or destination constraints leads to so-called production-constrained, attraction-constrained and production-attraction-constrained spatial interaction models that may be convincingly justified using entropy maximizing methods (see Wilson 1967).

Dyad Label	ID$_{\text{origin}}$	ID$_{\text{destination}}$	Flow	Origin Variable	Destination Variable	Separation (Origin, Destination)
1	1	1	$Y(1, 1)$	$a(1)$	$b(1)$	$d(1, 1)$
2	2	1	$Y(2, 1)$	$a(2)$	$b(1)$	$d(2, 1)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	n	1	$Y(n, 1)$	$a(n)$	$b(1)$	$d(n, 1)$
n+1	1	2	$Y(1, 2)$	$a(1)$	$b(2)$	$d(1, 2)$
n+2	2	2	$Y(2, 2)$	$a(2)$	$b(2)$	$d(2, 2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2n	n	2	$Y(n, 2)$	$a(n)$	$b(2)$	$d(n, 2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
(n-1)n	1	n	$Y(1, n)$	$a(1)$	$b(n)$	$d(1, n)$
(n-1)n+1	2	n	$Y(2, n)$	$a(2)$	$b(n)$	$d(2, n)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n^2	n	n	$Y(n, n)$	$a(n)$	$b(n)$	$d(n, n)$

Table 1. Data organization convention

Moreover, note that this widely used log-normal specification of the spatial interaction model has several shortcomings[6]. Most importantly, it suffers from least-squares and normality assumptions that ignore the true integer nature of the flows and approximate a discrete-valued process by an almost certainly misrepresentative continuous distribution.

3. FEEDFORWARD NEURAL SPATIAL INTERACTION MODELS

Neural spatial interaction models represent the most recent innovation in the design of spatial interaction models. For concreteness and simplicity, we consider neural spatial interaction models based on the theory of single hidden layer feedforward networks. Single hidden layer feedforward neural networks consist of nodes (also known as processing units or simply units) that are organized in layers. Figure 1 shows a schematic diagram of a typical feedforward neural spatial interaction model containing a single intermediate layer of processing units separating input from output units. Intermediate layers of this sort are called *hidden* layers to distinguish them from the input and output layers. In this network there are three input nodes representing the origin, destination and separation variables (denoted by x_1, x_2, x_3); *H* hidden units (say $z_1, ..., z_H$) representing hidden summation units (denoted by the symbol $\textcircled{\Sigma}$); and one (summation) output node representing origin-destination flows. Weight parameters are represented by links between the nodes. Observe the feedforward structure where the inputs are connected only to units in the hidden layer, and the outputs of this layer are connected only to the output layer that consists of only one unit.

Any network diagram can be converted into its corresponding mapping function, provided that the diagram is feedforward as in Fig. 1 so that it does not contain closed directed cycles[7]. This guarantees that the network output can be described by a series of functional transformations as follows. First, we form a linear combination[8] of the *K* input variables $x_1, ..., x_K$ (typically *K*=3) to get the input, say net_h, that hidden unit *h* receives

$$net_h = \sum_{k=1}^{K} w_{hk}^{(1)} x_k + w_{ho}^{(1)} \tag{5}$$

for $h = 1,...,H$. The superscript $^{(1)}$ indicates that the corresponding parameters are in the first parameter layer of the network. The parameters $w_{hk}^{(1)}$ represent connection weights going from input *k* $(k = 1,...,K)$ to hidden unit *h* $(h = 1,...,H)$, and $w_{ho}^{(1)}$ is a bias[9].

These quantities, net_h, are known as activations in the field of neural networks. Each of them is then transformed using a non-linear *transfer* or *activation* function[10] φ to give the output

[6] Flowerdew and Aitkin (1982), for example, question the appropriateness of this model specification, and suggest instead that the observed flows follow a Poisson distribution, leading to models termed Poisson spatial interaction models.

[7] Networks with closed directed cycles are called *recurrent* networks. There are three types of such networks: *first*, networks in which the input layer is fed back into the input layer itself; *second*, networks in which the hidden layer is fed back into the input layer, and *third*, networks in which the output layer is fed back into the input layer. These feedback networks are useful when input variables represent time series.

[8] Note, we could alternatively use product rather than summation hidden units to supplement the inputs to a neural network with higher-order combinations of the inputs to increase the capacity of the network in an information capacity sense. These networks are called *product unit* rather than summation unit networks (see Fischer and Reismann 2002b).

[9] This term should not be confused with the term bias in a statistical sense.

[10] The inverse of this function is called link function in the statistical literature. Note that radial basis function networks may be viewed as single hidden layer networks that use radial basis function nodes in the hidden layer. This class of

$$z_h = \varphi(net_h) \qquad (6)$$

for $h = 1, ..., H$. These quantities are again linearly combined to generate the input, called net, that the output unit receives

$$net = \sum_{h=1}^{H} w_h^{(2)} z_h + w_o^{(2)} . \qquad (7)$$

The superscript $^{(2)}$ indicates that the corresponding parameters are in the second parameter layer of the network. The parameters $w_h^{(2)}$ represent the connection weights from hidden units h $(h = 1, ..., H)$ to the output unit, and $w_o^{(2)}$ is a bias parameter. Finally, net is transformed to produce the output $\psi(net)$, where ψ denotes an activation function of the output unit.

Information processing in such networks is, thus, straightforward. The input units just provide a 'fan-out' and distribute the input to the hidden units. These units sum their inputs, add a constant (the bias) and take a fixed transfer function φ_h of the result. The output unit is of the same form, but with output activation function ψ. Network output can then be expressed in terms of an output function

$$\phi_H(x, w) = \psi\left[\sum_{h=1}^{H} w_h^{(2)} \varphi_h\left(\sum_{k=1}^{K} w_{hk}^{(1)} x_k + w_{k0}^{(1)}\right) + w_0^{(2)}\right] \qquad (8)$$

where the expression $\phi_H(x, w)$ is a convenient short-hand notation for the model output since this depends only on inputs and weights. Vector $x = (x_1, ..., x_K)$ is the input vector and w represents the vector of all the weights and bias terms. $\varphi(\cdot)$ is a non-linear [generally sigmoid] hidden layer activation function and $\psi(\cdot)$ an output unit [often quasi-linear] activation function, both continuously differentiable of order two on □ . The function ϕ is explicitly indexed by the number of hidden units, H, in order to indicate the dependence, but will be dropped for convenience.

Note that the bias terms $w_{k0}^{(1)}$ and $w_0^{(2)}$ in Eq. (8) can be absorbed[11] into the set of weight parameters by defining additional input and hidden unit variables, x_0 and z_0, whose values are clamped at one so that $x_0 = 1$ and $z_0 = 1$. Then the network function given by Eq. (8) becomes

$$\phi_H = \psi\left[\sum_{h=0}^{H} w_h^{(2)} \varphi_h\left(\sum_{k=0}^{K} w_{hk}^{(1)} x_k\right)\right] . \qquad (9)$$

The main power of neural spatial interaction models accrues from their capability for universal function approximation. Cybenko (1989); Funahashi (1989); Hornik, Stinchcombe and White (1989) and many others have shown that single hidden layer neural networks such as those given by Eq. (9) can approximate arbitrarily well any continuous function.

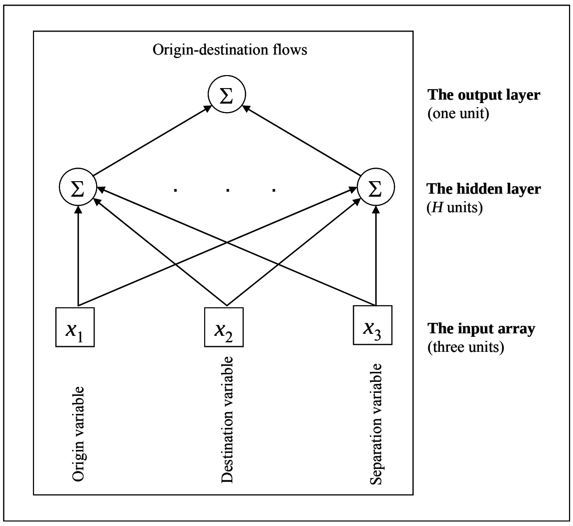


Figure 1. Network diagram of the neural spatial interaction model as defined by Eq. (8), for *K*=3 (bias units deleted)

Neural spatial interaction modelling involves three major stages (Fischer and Gopal 1994):

- The *first stage* consists of the identification of a model candidate from the general class of neural spatial interaction models of type (9). This involves both the specification of appropriate transfer functions ψ and φ, and the number, *H*, of hidden units.
- The *second stage* involves solving the network training [network learning, parameter estimation] problem, and hence determines the optimal set of model parameters where optimality is defined in terms of an error [loss, performance] function.
- The *third stage* is concerned with testing and evaluating the out-of-sample [generalization] performance of the chosen model.

Both the theoretical and practical side of the model selection problem have been intensively studied (see Fischer 2001, 2000 among others). The standard approach for finding a good neural spatial interaction model is to split the available set of samples into three subsets: training, validation and test sets. The training set is used for parameter estimation. In order to avoid overfitting, a common procedure is to use a network model with sufficiently large *H* for the task at hand, to monitor – during training – the out-of-sample performance on a separate validation set, and finally to choose the model that corresponds to the minimum on the validation set, and employ it for future purposes such as the evaluation on the test set.

neural networks asks for a two stage approach for training. In the first stage the parameters of the basis functions are determined, while in the second stage the basis functions are kept fixed and the second layer weights are found (see Bishop 1995, 170 pp.).

[11] This is the same idea as incorporating the constant term in the design matrix of a regression by inserting a column of ones.

4. A RATIONALE FOR THE ESTIMATION APPROACH

If we view a neural spatial interaction model as generating a family of approximations (as *w* ranges over *W*, say) to an unknown spatial interaction function, then we need a way to pick a best approximation from this family. This is the function of network learning or network training which might be viewed as an optimization problem[12].

We develop a rationale for an appropriate objective (loss or cost) function for this task. Following Rumelhart et al. (1995) we propose that the goal is to find that model which is the most likely explanation of the observed data set, say *D*. We can express this as attempting to maximize the term

$$P\big(\phi(w)\,|\,D\big)=\frac{P\big(D\,|\,\phi(w)\big)\;P\big(\phi(w)\big)}{P\big(D\big)} \qquad (10)$$

where ϕ represents the neural spatial interaction model (with w denoting the vector of weights) in question. $P\big(D\,|\,\phi(w)\big)$ is the probability that the model would have produced the observed data *D*. Since sums are easier to work with than products, we will maximize the log of this probability, and since the log is a strictly monotonic transformation, maximizing the log is equivalent to maximizing the probability itself. In this case we have

$$\begin{aligned}&\log P\big(\phi(w)\,|\,D\big)=\\&\log P\big(D\,|\,\phi(w)\big)+\log P\big(\phi(w)\big)-\log P\big(D\big).\end{aligned} \qquad (11)$$

The probability of the data, $P\big(D\big)$, is not dependent on the model. Thus, it is sufficient to maximize $\log P\big(D\,|\,\phi(w)\big)+\log P\big(\phi(w)\big)$. The first of these terms represents the probability of the data given the model, and hence measures how well the neural network model accounts for the data. The second term is a representation of the model itself; that is, it is *a priori* probability, that can be utilized to get information and constraints into the learning procedure.

We focus solely on the first term, the performance, and begin by noting that the data can be broken down into a set of observations, $D=\big\{q_u=\big(x_u,y_u\big):u=1,...,U\big\}$, each q_u we will assume to be chosen independently of the others. Hence we can write the probability of the data given the model as

$$\begin{aligned}&\log P\big(D\,|\,\phi(w)\big)=\\&\log\prod_u P\big(q_u\,|\,\phi(w)\big)=\sum_u\log P\big(q_u\,|\,\phi(w)\big).\end{aligned} \qquad (12)$$

Note that this assumption permits to express the probability of the data given the model as the sum of terms, each term representing the probability of a single observation given the model. We can still take another step and break the data into two parts: the observed input data x_u and the observed target data y_u. Therefore we can write

$$\begin{aligned}&\log P\big(D\big|\phi(w)\big)=\\&\sum_u\log P\big(y_u\big|x_u\text{ and }\phi(w)_u\big)+\sum_u\log P\big(x_u\big).\end{aligned} \qquad (13)$$

Since we assume that x_u does not depend on the model, the second term of the right hand side of the equation will not affect the determination of the optimal model. Thus, we need only to maximize the term $\sum_u\log P\big(y_u\,|\,x_u\text{ and }\phi(w)_u\big)$.

Up to now we have – in effect – made only the assumption of the independence of the observed data. In order to proceed, we need to make some more specific assumptions, especially about the relationship between the observed input data x_u and the observed target data y_u, a probabilistic assumption. We assume that the relationship between x_u and y_u is not deterministic, but that for any given x_u there is a distribution of possible values of y_u. But the model is deterministic, so rather than attempting to predict the actual outcome we only attempt to predict the expected value of y_u given x_u. Therefore, the model output is to be interpreted as the mean bilateral interaction frequencies (that is, those from the location of origin to the location of destination). This is, of course, the standard assumption.

To proceed further, we have to specify the form of the distribution of which the model output is the mean. Of particular interest to us is the assumption that the observed data is the realization of a sequence of independent Poisson random variables. Under this assumption we can write the probability of the data given the model as

$$P\big(y_u\big|x_u\text{ and }\phi(w)_u\big)=\prod_u\frac{\phi(w)_u^{y_u}\exp\big(-\phi(w)_u\big)}{y_u!} \qquad (14)$$

and, hence, define a maximum likelihood estimator as a parameter vector $\hat{w}$ which maximizes the log-likelihood *L*

$$\begin{aligned}&\max_{w\in W}\;L\big(x,y,w\big)=\\&\max_{w\in W}\sum\nolimits_u\big[y_u\log\phi(w)_u-\phi(w)_u-\log(y_u!)\big].\end{aligned} \qquad (15)$$

Instead of maximizing the log-likelihood it is more convenient to view learning as solving the minimization problem

[12] This directs attention to the literature on numerical optimization theory, with particular reference to optimization techniques that use higher-order information such as conjugate gradient procedures and Newton's method. The methods use the gradient vector (first-order partial derivatives) and/or the Hessian matrix (second-order partial derivatives) of the loss function to perform optimization, but in different ways. A survey of first-order and second-order optimization techniques applied to network training can be found in Cichocki and Unbehauen (1993).

$$\min_{w \in W} \lambda(x, y, w) = \min_{w \in W} \left[-L(x, y, w) \right] \quad (16)$$

where the loss (cost) function λ is the negative log-likelihood L. λ is continuously differentiable on the Q-dimensional real parameter space ($Q = HK + H + 1$) which is a finite dimensional closed bounded domain and, thus, compact.

5. LEARNING MODES AND PROCEDURES

It can be shown that $\lambda(w)$ assumes its weight minimum under certain conditions, but characteristically there exist many minima in real world applications all of which satisfy

$$\nabla\lambda(w) = 0 \quad (17)$$

where $\nabla\lambda$ denotes the gradient of λ. The minimum for which the value of λ is smallest is termed the global minimum and other minima are called local minima.

There are many ways to solve the minimization problem (16). Closed-form optimization via the calculus of scalar fields rarely admits a direct solution. A relatively new set of interesting techniques that use optimality conditions from calculus are based on evolutionary computation (Goldberg 1989; Fogel 1995). But gradient procedures which use the first partial derivatives $\nabla\lambda(w)$, so-called first order strategies, are most widely used. Gradient search for solutions gleans its information about derivatives from a sequence of function values. The recursion scheme is based on the formula[13]

$$w(\tau+1) = w(\tau) + \eta(\tau)\, d(\tau) \quad (18)$$

where τ denotes the iteration step. Different procedures differ from each other with regard to the choice of step length $\eta(\tau)$ and search direction $d(\tau)$, the former being a scalar called *learning* rate and the latter a vector of unit length.

The simplest approach to using gradient information is to assume $\eta(\tau)$ being constant and to choose the parameter update in Eq. (18) to comprise a small step in the direction of the negative gradient so that

$$d(\tau) = -\nabla\lambda(w(\tau))\,. \quad (19)$$

After each such update, the gradient is re-evaluated for the new parameter vector $w(\tau+1)$. Note that the loss function is defined with respect to a training set of size $U1$, say D_{U1}, to be processed to evaluate $\nabla\lambda$. One complete presentation of the entire training set during the training process is called an *epoch*. The training process is maintained on an epoch-by-epoch basis until the connection weights and bias terms of the network stabilize and the average error over the entire training set converges to some minimum.

Gradient descent optimization may proceed in one of two ways: pattern mode and batch mode. In the *pattern mode* weight updating is performed after the presentation of each training example. Note that the loss functions based on maximum likelihood for a set of independent observations comprise a sum of terms, one for each data point. Thus

$$\lambda(w) = \sum_{u1 \in U1} \lambda_{u1}(w) \quad (20)$$

where λ_{u1} is called the *local error* (loss) while λ the *global error* (loss), and pattern based gradient descent makes an update to the parameter vector based on one training example at a time so that

$$w(\tau+1) = w(\tau) - \eta\nabla\lambda_{u1}(w(\tau)). \quad (21)$$

Rumelhart et al. (1986) have shown that pattern based gradient descent minimizes Eq. (16), if the learning parameter η is sufficiently small. The smaller η, the smaller will be the changes to the weights in the network from one iteration to the next and the smoother will be the trajectory in the parameter space. This improvement, however, is attained at the cost of a slower rate of training. If we make the learning rate parameter η too large so as to speed up the rate of training, the resulting large changes in the parameter weights assume such a form that the network may become unstable.

In the *batch mode* of learning, parameter updating is performed after the presentation of all the training examples that constitute an epoch. From an online operational point of view, the pattern mode of training is preferred over the batch mode, because it requires less local storage for each weight connection. Moreover, given that the training patterns are presented to the network in a random manner, the use of pattern-by pattern updating of parameters makes the search in parameter space stochastic in nature which in turn makes it less likely to be trapped in a local minimum. On the other hand, the use of batch mode of training provides a more accurate estimation of the gradient vector $\nabla\lambda$. Finally, the relative effectiveness of the two training modes depends on the problem to be solved (Haykin 1994, 152 pp.)

For batch optimization there are more efficient procedures, such as conjugate gradient and quasi-Newton methods, that are much more robust and much faster than gradient descent (Nocedal and Wright 1999). Unlike steepest gradient, these algorithms have the characteristic that the error function always decreases at each iteration unless the parameter vector has arrived at a local or global minimum. Conjugate gradient methods achieve this by incorporating an intricate relationship between the direction and gradient vectors. The initial direction vector $d(0)$ is set equal to the negative gradient vector at the initial step $\tau = 0$. Each successive direction vector is then computed

[13] When using an iterative optimization algorithm, some choice has to be made of when to stop the training process. There are various criteria that might be used. For example, learning may be stopped when the loss function or the relative change in the loss function falls below a prespecified value.

as a linear combination of the current gradient vector and the previous direction vector. Thus,

$$d(\tau+1) = -\nabla\lambda(w(\tau+1)) + \gamma(\tau)\, d(\tau) \tag{22}$$

where $\gamma(\tau)$ is a time varying parameter. There are various rules for determining $\gamma(\tau)$ in terms of the gradient vectors at time τ and $\tau+1$, leading to the Fletcher-Reeves and Polak-Ribière variants of conjugate gradient algorithms (see Press et al. 1992). The computation of the learning rate parameter $\eta(\tau)$ in the update formula given by Eq. (18) involves a line search, the purpose of which is to find a particular value of η for which the loss function $\lambda(w(\tau)+\eta\, d(\tau))$ is minimized, given fixed values of $w(\tau)$ and $d(\tau)$.

The application of Newton's method to the training of neural networks is hindered by the requirement of having to calculate the Hessian matrix and its inverse, which can be computationally expensive for larger network models[14]. The problem is further complicated by the fact that the Hessian matrix has to be non-singular for its inverse to be computed. Quasi-Newton methods avoid this problem by building up an approximation to the inverse Hessian over a number of iteration steps. The most commonly variants are the Davidson-Fletcher-Powell and the Broyden-Fletcher-Goldfarb-Shanno procedures (see Press et al. 1992).

Quasi-Newton procedures are today the most efficient and sophisticated (batch) optimization algorithms. But they require the evaluation and storage in memory of a dense matrix at each iteration step τ. For larger problems (more than 1,000 weights) the storage of the approximate Hessian can be too demanding. In contrast, the conjugate gradient procedures require much less storage, but an exact determination of the learning rate $\eta(\tau)$ and the parameter $\gamma(\tau)$ in each iteration τ, and, thus, approximately twice as many gradient evaluations as the quasi-Newton methods.

When the surface modelled by the loss function in its parameter space is extremely rugged and has many local minima, then a local search from a random starting point tends to converge to a local minimum close to the initial point. In order to seek out good local minima, a good training procedure must thus include both a gradient based optimization algorithm and a technique like random start that enables sampling of the space of minima. Alternatively, stochastic global search procedures might be used. Examples of such procedures include Alopex (see Fischer, Reismann and Hlavácková-Schindler 2003), genetic algorithms (see Fischer and Leung 1998), and simulated annealing. These procedures guarantee convergence to a global solution with a higher probability, but at the expense of slower convergence.

Finally, it is worth noting that the technique of error backpropagation provides a computationally efficient technique to calculate the gradient vector of a loss function for a feedforward neural network with respect to the parameters. This technique – sometimes simply termed backprop – uses a local message passing scheme in which information is sent alternately forwards and backwards through the network. Its modern form stems from Rumelhart, Hinton and Williams (1986), illustrated for gradient descent optimization applied to the sum-of-squares error function. It is important to recognize, however, that error backpropagation can also be applied to our loss function and to a wide variety of optimization schemes for weight adjustment other than gradient descent, both in pattern or batch mode.

6. NETWORK COMPLEXITY

So far we have considered neural spatial interaction models of type (9) with *a priori* given numbers of input, hidden and output units. While the number of input and output units is basically problem dependent, the number *H* of hidden units is a free parameter that can be adjusted to provide the best testing performance on independent data, called testing set. But the testing error is not a simple function of *H* due to the presence of local minima in the loss function. The issue of finding a parsimonious model for a real world problem is critical for all models but particularly important for neural networks because the problem of overfitting is more likely to occur.

A neural spatial interaction model that is too simple (i.e. small *H*), or too inflexible, will have a large bias and smooth out some of the underlying structure in the data (corresponding to high bias), while one that has too much flexibility in relation to the particular data set will overfit the data and have a large variance. In either case, the performance of the network model on new data (i.e. generalization performance) will be poor. This highlights the need to optimize the complexity in the model selection process in order to achieve the best generalization (Bishop 1995, p. 332; Fischer 2000). There are some ways to control the complexity of a neural network model, complexity in terms of the number of hidden units or, more precisely, in terms of the independently adjusted parameters. Practice in neural spatial interaction modelling generally adopts a trial and error approach that trains a sequence of neural networks with an increasing number of hidden units and then selects that one which gives the best predictive performance on a testing set[15].

There are, however, other more principled ways to control the complexity of a neural network model in order to avoid overfitting[16]. One approach is that of *regularization*, which involves adding a regularization term $R(w)$ to the loss

[14] Note that computational time rises with the square of *Q*, the dimension of the parameter space.

[15] Note that limited data sets make the determination of *H* more difficult if there is not enough data available to hold out a sufficiently large independent test sample.

[16] A neural network is said to be overfitted to the data if it obtains an excellent fit to the training data, but gives a poor representation of the unknown function which the neural network is approximating.

function in order to control overfitting, so that the total error function to be minimized takes the form

$$\tilde{\lambda}_p(w) = \lambda_p(w) + \mu\, R(w) \tag{23}$$

where μ is a positive real number, the so-called regularization parameter, that controls the relative importance of the data dependent error $\lambda_p(w)$, and $R(w)$ the regularization term, sometimes also called complexity term. This term embodies the *a priori* knowledge about the solution, and therefore depends on the nature of the particular problem to be solved. Note that $\tilde{\lambda}_p(w)$ is called the *regularized error* or *loss function*.

One of the simplest forms of a regularizer is defined as the squared Euclidean norm of the parameter vector *w* in the network, as given by

$$R(w) = \|w\|^2 . \tag{24}$$

This regularizer[17] is known as weight decay function that penalizes larger weights. Hinton (1987) has found empirically that a regularizer of this form can lead to significant improvements in network generalization.

Sometimes, a more general regularizer is used, for which the regularized error or loss takes the form

$$\lambda(w) + \mu \|w\|^m \tag{25}$$

where *m*=2 corresponds to the quadratic regularizer given by Eq. (24). The case *m*=1 is known as the 'lasso' in the statistics literature (Tibshirani 1996). The regularizer given by Eq. (25) has the property that – if μ is sufficiently large – some of the parameter weights are driven to zero in sequential learning algorithms, leading to a sparse model. As μ is increased, so an increasing number of parameters are driven to zero.

One of the limitations of this regularizer is inconsistency with certain scaling characteristics of network mappings. If one trains a network using original data and one network using data for which the input and/or target variables are linearly transformed, then consistency requires that the regularizer should be invariant to re-scaling of the weights and to shifts of the biases (Bishop 2006, p. 258). A regularized loss function that satisfies this property is given by

$$\lambda(w) + \mu_1 \left\|w_{q1}\right\|^m + \mu_2 \left\|w_{q2}\right\|^m \tag{26}$$

where μ_1, μ_2 and *m* are regularization parameters. w_{q1} denotes the set of the weights in the first parameter layer, that is $w_{11}^{(1)},\ldots,w_{h1}^{(1)},\ldots,w_{HK}^{(1)}$, and w_{q2} those in the second layer, that is $w_1^{(2)},\ldots,w_h^{(2)},\ldots,w_H^{(2)}$.

The more sophisticated control of complexity that regularization offers over adjusting the number of hidden units by trial and error is evident. Regularization allows complex neural network models to be trained on data sets of limited size without severe overfitting, by limiting the effective network complexity. The problem of determining the appropriate number of hidden units is, thus, shifted to one of determining a suitable value for the regularization parameter(s) during the training process.

The principal alternative to regularization as a way to optimize the model complexity for a given training data set is the procedure of *early stopping*. As we have seen in the previous sections, training of a non-linear network model corresponds to an iterative reduction of the loss (error) function defined with respect to a given training data set. For many of the optimization procedures used for network training (such as conjugate gradient optimization) the error is a non-decreasing function of the iteration steps τ. But the error measured with respect to independent data, called *validation data set*, say D_{U2}, often shows a decrease first, followed by an increase as the network starts to overfit, as illustrated in Fischer and Gopal (1994). Thus, training can be stopped at the point of smallest error with respect to the validation data, in order to get a network that shows good generalization performance. But, if the validation set is mall, it will give a relatively noisy estimate of generalization performance, and it may be necessary to keep aside another data set, the test set D_{U3}, on which the performance of the network model is finally evaluated.

This approach of stopping training before a minimum of the training error has been reached is another way of eliminating the network complexity. It contrasts with regularization because the determination of the number of hidden units does not require convergence of the training process. The training process is used here to perform a directed search in the weight space for a neural network model that does not overfit the data and, thus, shows superior generalization performance. Various theoretical and empirical results have provided strong evidence for the efficiency of early stopping (see, for example, Weigend, Rumelhart and Huberman 1991; Baldi and Chauvin 1991; Finnoff 1991; Fischer and Gopal 1994). Although many questions remain, a picture is starting to emerge as to the mechanisms responsible for the effectiveness of this approach. In particular, it has been shown that stopped training has the same sort of regularization effect [i.e. reducing model variance at the cost of bias] that penalty terms provide.

7. GENERALIZATION PERFORMANCE

Model performance may be measured in terms of Kullback and Leibler's (1951) information criterion, *KLIC*, which is a natural performance criterion for the goodness-of-fit of ML estimated models

[17] In conventional curve fitting, the use of this regularizer is termed *ridge regression*.

$$KLIC\left(D_{U3}\right) = \sum_{u3\in U3} \frac{y_{u3}}{\sum\limits_{u3'\in U3}^{U3} y_{u3'}} \ln \left[\frac{y_{u3}\left[\sum\limits_{u3'\in U3} y_{u3'}\right]^{-1}}{\phi\left(x_{u3},\hat{w}\right)\left[\sum\limits_{u3'\in U3} \phi\left(x_{u3'},\hat{w}\right)\right]^{-1}} \right] \qquad (27)$$

where $\left(x_{u3}, y_{u3}\right)$ denotes the $u3$-th pattern of the data set D_{U3}, and ϕ is the estimated neural spatial interaction model under consideration.

The standard approach to evaluate the out-of-sample (generalization or prediction) performance of a neural spatial interaction model (see Fischer and Gopal 1994) is to split the data set *D* of size *U* into three subsets: the *training* [in-sample] *set* $D_{U1} = \{q_{u1} = (x_{u1}, y_{u1})\}$ of size $U1$, the internal validation set $D_{U2} = \{q_{u2} = (x_{u2}, y_{u2})\}$ of size $U2$ and the testing [out-of-sample, generalization, prediction] set $D_{U3} = \{q_{u3} = (x_{u3}, y_{u3})\}$ of size $U3$, with U1+U2+U3=U. The training set serves for parameter estimation. The validation set is used to determine the stopping point before overfitting occurs, and the test set to evaluate the generalization performance of the model, using some measure of error between a prediction and an observed value, such as $KLIC\left(D_{U3}\right)$.

Randomness enters into this standard approach to neural network modelling in two ways: in the splitting of the data samples, and in choices about the parameter initialization. This leaves one question wide open. What is the variation of test performance as one varies training, validation and test sets? This is an important question, since there is not just one ‘best’ split of the data or obvious choice for the initial weights. Thus, it is useful to vary both the data partitions and parameter initializations to find out more about the distribution of generalization errors. One way is to use the bootstrap pairs approach (Efron 1982) with replacement to evaluate the performance, reliability, and robustness of the neural spatial interaction model.

The bootstrap pairs approach[18] is an intuitive way to apply the bootstrap notion to combine the purity of splitting the data set into three data sets with the power of a resampling procedure. The basic idea of this approach is to generate *B* pseudo-replicates of the training sets D^{*b}_{U1}, *B* internal validation sets D^{*b}_{U2} and *B* testing sets D^{*b}_{U3}, then to re-estimate the model parameters $\hat{w}^{*b}$ on each training bootstrap sample q^{*b}_{u1}, to stop training on the basis of the associated validation bootstrap sample q^{*b}_{u2} and to test generalization performance, measured in terms of *KLIC*, on the test bootstrap sample q^{*b}_{u3}. In this bootstrap world, the empirical bootstrap distribution of the performance measure can be estimated, pseudo-errors can be computed, and used to approximate the distribution of the real errors. The approach is appealing, but characterized by very demanding computational intensity in real world contexts (see Fischer and Reismann 2002b for an application). Implementing the approach involves the following steps:

Step 1: Conduct totally independent re-sampling operations, where

(i) *B* independent training bootstrap samples are generated, by randomly sampling *U*1 times (*U*1<*U*), with replacement, from *D* for *b*=1, …, *B*

$$D^{*b}_{U1} = \left\{q^{*b}_{u1} = \left(x^{*b}_{u1},\, y^{*b}_{u1}\right)\right\}, \qquad (28)$$

(ii) *B* independent validation bootstrap samples are generated [in the case of early stopping only], by randomly sampling *U*2 times (*U*2<*U*), with replacement, from *D* so that for *b*=1, …, *B*

$$D^{*b}_{U2} = \left\{q^{*b}_{u2} = \left(x^{*b}_{u2},\, y^{*b}_{u2}\right)\right\}, \qquad (29)$$

(iii) *B* independent test bootstrap samples are generated, by randomly sampling *U*3 times (*U*3<*U*), with replacement, from *D* so that for *b*=1, …, *B*

$$D^{*b}_{U3} = \left\{q^{*b}_{u3} = \left(x^{*b}_{u3},\, y^{*b}_{u3}\right)\right\}. \qquad (30)$$

Step 2: Use each training bootstrap sample q^{*b}_{u1} to compute the bootstrap parameter estimates $\hat{w}^{*b}$ by solving Eq. (16) with q^{*b}_{u1} replacing q_u:

$$\hat{w}^{*b} = \arg\,\min\left\{\lambda\left(q^{*b}_{u1},\, w^{*b}\right) : w^{*b} \in W \subseteq \mathbb{R}^{Q}\right\} \qquad (31)$$

where *Q* is the number of parameters, and

$$\lambda\left(q^{*b}_{u1},\, w^{*b}\right) = \sum_{u1=1}^{U1}\left[y_{u1}\ln\phi(w)_{u1} - \phi(w)_{u1}\right]. \qquad (32)$$

Note: During the training process the generalization performance of the model (in terms of the *KLIC* criterion) is monitored on the corresponding bootstrap validation set, in the case of early stopping. The training process is stopped if the validation error starts to increase.

Step 3: Calculate the *KLIC*-statistic $KLIC\,(D^{*b}_{U3})$ for each test bootstrap sample.

Step 4: Replicate Steps 3-4 many times, say *B*=100 or *B*=1,000.

[18] This approach contrasts to residuals bootstrapping that treats the model residuals rather than $q_u = (x_u,\ y_u)$ as the sampling units and creates a bootstrap sample by adding residuals to the model fit. In this latter case bootstrapping distribution is conditional on the actual observations.

Step 5: The statistical accuracy of the generalization performance statistic can then be evaluated by looking at the variability of the statistic between the different bootstrap test sets. Estimate the standard deviation $\hat{\sigma}$ of the statistic as approximated by bootstrap

$$\hat{\sigma}^B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\widehat{KLIC}^{*b}\left(D_{U3}^{*b}\right) - \widehat{KLIC}^{*}(\cdot) \right]^2 \right\}^{\frac{1}{2}} \quad (33)$$

with

$$\widehat{KLIC}^{*}(\cdot) = \sum_{b=1}^{B} \widehat{KLIC}^{*b}\left(D_{U3}^{*b}\right). \quad (34)$$

The true standard error of $\widehat{KLIC}$ is a function of the unknown density function, say *F*, of *KLIC*, that is $\sigma(F)$. With the bootstrapping approach described above one obtains $\hat{F}_{U3}^{*}$, which is supposed to describe closely the empirical distribution $\hat{F}_{U3}$, in other words $\hat{\sigma}_{U3}^{B} \approx \sigma(\hat{F}_{U3})$. Asymptotically, this means that the sample size tends to infinity, i.e. $U3 \to \infty$, the estimate $\hat{\sigma}^B$ tends to $\sigma(F)$. For finite sample sizes, however, there will be deviations in general.

8. CLOSING REMARKS

In this paper a modest attempt has been made to provide a unified framework for neural spatial interaction modelling, based upon maximum likelihood estimation under distributional assumptions of Poisson processes. In this way we avoid the weaknesses of least squares and normality assumptions that ignore the true integer nature of the origin-destination flows and approximate a discrete-valued process by an almost certainly misrepresentative continuous representation.

Randomness enters in two ways in neural spatial interaction modelling: in the splitting of the data set into training, validation and test sets on the one side, and in choices about parameter initialization on the other. The paper suggests the bootstrapping pairs approach to evaluate the performance, reliability and robustness of neural spatial interaction models. The approach is attractive, but computationally intensive.

Despite significant improvements in our understanding of the fundamentals and principles of neural spatial interaction modelling, there are many open problems and directions for future research. The design of a neural network approach suited to deal with the doubly constrained case of spatial interaction, for example, is still missing. Finding good global optimization procedures for solving the non-convex learning problems is still an important issue for further research even though some relevant work can be found in Fischer, Hlavácková-Schindler and Reismann (1999). Also the model identification problem deserves further attention to come up with techniques that go beyond the current rules of thumb.

From a spatial analytic perspective an important avenue for further investigation is the explicit incorporation of spatial dependency in the network representation that received less attention in the past than it deserves. Another is the application of Bayesian inference techniques to neural networks. A Bayesian approach would provide an alternative framework for dealing with issues of network complexity and would avoid many of the problems discussed in this paper. In particular, confidence intervals could easily be assigned to the predictors generated by neural spatial interaction models, without the need of bootstrapping.

REFERENCES

Baldi, P., Chauvin, Y., 1991. Temporal evolution of generalization during learning in linear networks. *Neural Computation*, 3(4), pp. 589-603

Bishop, C. M., 2006. *Pattern recognition and machine learning*. Springer, New York

Bishop. C. M., 1995. *Neural networks for pattern recognition*. Clarendon Press, Oxford

Cichocki, A., Unbehauen, R., 1993. *Neural networks for optimization and signal processing*. John Wiley, Chichester

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals and Systems*, 2, pp. 303-314

Efron, B., 1982. *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia [PA]

Finnoff, W., 1991. Complexity measures for classes of neural networks with variable weight bounds. In: *Proceedings of the International Geoscience and Remote Sensing Symposium* [IGARSS'94, Volume 4]. IEEE Press, Piscataway [NJ], pp. 1880-1882

Fischer, M. M., 2002. Learning in neural spatial interaction models: A statistical perspective. *Journal of Geographical Systems*, 4 (3), pp. 287-299

Fischer, M. M., 2001. Neural spatial interaction models. In Fischer M M, Leung Y (eds) GeoComputational modelling. Techniques and applications. Springer, Berlin, Heidelberg and New York, pp. 195-219

Fischer, M. M., 2000. Methodological challenges in neural spatial interaction modelling: The issue of model selection. In Reggiani A (ed.) Spatial economic science: *New frontiers in theory and methodology*. Springer, Berlin, Heidelberg and New York, pp. 89-101

Fischer, M. M., Gopal, S., 1994. Artificial neural networks. A new approach to modelling interregional telecommunication flows. *Journal of Regional Science*, 34(4), pp. 503-527

Fischer, M. M., Griffith, D. A., 2008. Modeling spatial autocorrelation in spatial interaction data: An application to patent citation data in the European Union. *Journal of Regional Science*, 48(5), pp. 969-989

Fischer, M. M., Leung, Y., 1998. A genetic-algorithm based evolutionary computational neural network for modelling spatial interaction data. *The Annals of Regional Science*, 32(3), pp. 437-458

Fischer, M. M., Reismann, M., 2002a. Evaluating neural spatial interaction modelling by bootstrapping. *Networks and Spatial Economics*, 2(3), pp. 255-268

Fischer, M. M., Reismann, M., 2002b. A methodology for neural spatial interaction modeling. *Geographical Analysis*, 34(2), pp. 207-228

Fischer, M. M., Hlavácková-Schindler, K., Reismann, M., 1999. A global search procedure for parameter estimation in neural spatial interaction modelling. *Papers in Regional Science*, 78(2), pp. 119-134

Fischer, M. M., Reismann, M., Hlavácková-Schindler, K., 2003. Neural network modelling of constrained spatial interaction flows: Design, estimation and performance issues. *Journal of Regional Science*, 43(1), pp. 35-61

Flowerdew, R., Aitken, M., 1982. A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science*, 22(2), pp. 191-202

Fogel, D. B., 1995. *Evolutionary computation: Toward a new philosophy of machine intelligence*. IEEE Press, Piscataway [NJ]

Fotheringham, A. S., 1983. A new set of spatial interaction models: The theory of competing destinations. *Environment and Planning A*, 15(1), pp. 15-36

Funahashi, K., 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3), pp. 183-192

Griffith, D. A., 2003. *Spatial autocorrelation and spatial filtering*. Springer, Berlin, Heidelberg and New York

Goldberg, D. E., 1989. *Genetic algorithms*. Addison-Wesley, Reading [MA]

Gopal, S., Fischer, M. M., 1997. Fuzzy ARTMAP – a neural classifier for multispectral image classification. In: Fischer, M. M., Getis, A. (eds) *Recent developments in spatial analysis*. Springer, Berlin, Heidelberg and New York, pp. 306-335

Hassoun, M. H., 1995. *Fundamentals of artificial neural networks*. MIT Press, Cambridge [MA] and London, England

Haykin, S., 1994. *Neural networks. A comprehensive foundation*. Macmillan College Publishing Company, New York

Hinton, G. E., 1987. Learning translation invariant recognition in massively parallel networks. In: Bakker, J. W. de, Nijman, A. J., Treleaven, P. C. (eds) *Proceedings PARLE Conference on Parallel Architectures and Languages Europe*. Springer, Berlin, Heidelberg and New York, pp. 1-13

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), pp. 359-368

LeSage, J. P., Fischer, M. M., 2010. Spatial econometric modelling of origin-destination flows. In Fischer, M. M., Getis, A. (eds) *Handbook of Applied Spatial Analysis*. Springer, Berlin, Heidelberg and New York, pp. 409-433

LeSage, J. P., Pace, R. K., 2009. *Introduction to spatial econometrics*. CRC Press (Taylor and Francis Group), Boca Raton [FL], London and New York

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. *Annals of Mathematical Statistic*s, 22, pp. 78-86

McCulloch, W. S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, pp. 115-133

Moody, J. E., 1992. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In: Moody, J. E., Hanson, S. J., Lippman, R. P. (eds) *Advances in neural information processing systems* 4. Morgan Kaufmann, San Mateo [CA], pp. 683-690

Nocedal, J., Wright, S. J., 1999. *Numerical optimization*. Springer, Berlin, Heidelberg and New York

Press, W. H., Teukolky, S. A., Vetterling, W. T., Flannery, B. P., 1992. *Numerical recipes in C. The art of scientific computing*, 2nd edn. Cambridge University Press, Cambridge

Rosenblatt, F., 1962. *Principles of neurodynamics*. Spartan Books, Washington DC

Rumelhart, D. E., Durbin, R., Golden, R., Chauvin, Y., 1995. Backpropagation: The basic theory. In: Chauvin, Y., Rumelhart, D. E. (eds) *Backpropagation: Theory, architectures and applications*. Lawrence Erlbaum Associates, Hillsdale [NJ], pp. 1-34

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. Learning internal representations by error propagation. In: Rumelhart, D. E., McClelland, J. L., PDP Research Group (eds) *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press Cambridge [MA], pp. 318-362

Sen, A., Smith, T. E., 1995. *Gravity models of spatial interaction behaviour*. Springer, Berlin, Heidelberg and New York

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* B, 58, pp. 267-288

Weigend, A. S., Rumelhart, D. E., Huberman, B. A., 1991. Generalization by weight elimination with application to forecasting. In: Lippman, R., Moody, J., Touretzky, D. (eds) *Advances in neural information processing systems* 3. Morgan Kaufmann, San Mateo [CA], pp. 875-882

Weng, J., Hwang, W.-S., 2006. From neural networks to the brain: Autonomous mental development. *IEEE Computational Intelligence Magazine*, 1(3), pp. 15-31

Wilson, A. G., 1970. *Entropy in urban and regional planning*. Pion, London

Wilson, A. G., 1967. A statistical theory of spatial distribution models. *Transportation Research*, 1, pp. 253-269

Zapranis, A., Refenes, A.-P., 1999. *Principles of neural model identification, selection and adequacy. With applications to financial econometrics*. Springer, London

Deriving space-time variograms from space-time autoregressive (STAR) model specifications

Daniel A. Griffith[1], Gerard B.M. Heuvelink[2]

[1]Ashbel Smith Professor, University of Texas at Dallas
Email: dagriffith@utdallas.edu
[2]Wageningen University and Research Centre
Email: gerard.heuvelink@wur.nl

KEY WORDS: additive semi-variogram, autoregressive model, multiplicative semi-variogram, space, space-time, STAR, time

ABSTRACT:

Many geospatial science subdisciplines analyze variables that vary over both space and time. The space-time autoregressive (STAR) model is one specification formulated to describe such data. This paper summarizes STAR specifications that parallel geostatistical model specifications commonly used to describe space-time variation, with the goal of establishing synergies between these two modeling approaches. Resulting expressions for space-time correlograms derived from 1st-order STAR models are solved numerically, and then linked to appropriate space-time semivariogram models.

1. Introduction

Geostatistics furnishes techniques for modeling the covariance matrix, whereas spatial autoregression furnishes techniques for modeling the inverse covariance matrix, for a set of n geographically distributed values of a single random function. Both seek to capture spatial autocorrelation effects in georeferenced data.

Although, in practice, both geostatistics and spatial autocorrelation techniques mostly are applied to static spatial variables, a growing interest among researchers is to utilize these techniques to address change over both space and time. Incorporating time is more than just adding another dimension, because the behavior of a variable over time differs from its behavior over space, and characteristics of temporal processes often are known to some degree. Accordingly, a space-time geostatistical or autoregressive model must capture the fundamental differences between spatial and temporal variation, and must include these differences in its structure and parameterization.

The purpose of this paper is to establish the basis for a synergy between space-time geostatistics and autoregressive (STAR) approaches to the modeling of correlation structures latent in space-time data. The mutually advantageous conjunction of these two approaches follows that established for the static case by Griffith and Csillag (1993) and Griffith and Layne (1997), and seeks to create an enhanced combined approach to the modeling of space-time correlation structures. Simple 1st- and 2nd-order geographic neighbor direct dependency structures are addressed, with conceptualizations furnished by especially Gasim (1988) allowing them to be extended to larger neighborhoods. In doing so, we exploit the notion that a space-time semivariogram is valid (i.e., non-negative definite) when any linear combination of values of the associated random function at any finite number of space-time points has non-negative variance.

2. The Configurational Structure of Georeferenced Data

Consider a variable $Z = \{Z(\mathbf{s}, t) \mid \mathbf{s} \in S, t \in T\}$ that varies within a spatial domain S and a time horizon T. Let Z be observed at n space-time points $(\mathbf{s}_i, t_j)$, i = 1, 2, ..., m and j = 1, 2, …, k, where n = mk. These space-time observations constitute a time series of length k at each of the m spatial locations, and imply the use of a regular sampling scheme (i.e., the observations are uniformly spaced over time at each spatial location).

The set of n points can be converted to a surface partitioning by constructing its associated set of Thiessen (Voronoi) polygons; these become volumes in three dimensions. This conversion allow the generation of a Delauany triangulation (the dual graph) that furnishes a topology-based articulation of the configurational structure of the set of n points. Interpoint distances furnish another. Suppose variable Z is an areal unit aggregate observed for m regions in time, where k is the frequency of observations per region over time. Let these areal units form a mutually exclusive and collectively exhaustive partitioning of a surface. If these polygons are convex hulls (all internal angles < 180^o), then the geometric centroid of each polygon can be computed, and this set of points can be used both to convert the surface partitioning into a geographic distributions of points, and to construct a dual graph (similar to a Delaunay triangulation). This graph commonly is constructed using criteria based on chess piece movements: the rook's case (i.e., links connect points for polygons that share a common non-zero length boundary), and the queen's case (i.e., links connect points for polygons that share a common zero—i.e., point—or non-zero length boundary). For concave hulls (e.g., polygons with at least one internal angle > 180^o) or for nested areal units (e.g., one contained completely inside another), a judiciously selected arbitrary point may be the dual graph node. Meanwhile, time can be represented with a simple line graph comprising a linear sequence of links and points.

In all three geographic cases, the graphs in question can be converted to adjacency matrices, **C**, which are binary 0-1 matrices with all diagonal entries being 0. Because these graphs are planar or near-planar and connected, the number of ones in the m-by-m matrices representing geographic arrangement is at least 2(m-1), usually does not exceed 3(m-2), and never exceeds 8m. These matrices are symmetric here, in part because geographic dependencies are being cast as non-directional. The number of ones in the k-by-k time sequencing matrix is 2(k-1). This set of matrices furnishes the building blocks of n-by-n space-time data matrices. Eigenfunctions extracted from each of these binary matrices can be used to summarize their respective structure.

3. STAR Model Specifications

STAR model specifications (see Cliff et al. 1975) are explicit formulations describing how a variable Z varies in space **s** = (x, y) and time (t) in some joint fashion (x, y, t). The following two linear discrete cases are of interest here:

$$Z(x, y, t) = a\cdot Z(x, y, t-\Delta t) + b\cdot\{Z(x-\Delta x, y, t-\Delta t) + Z(x+\Delta x, y, t-\Delta t) + Z(x, y-\Delta y, t-\Delta t) + Z(x, y+\Delta y, t-\Delta t)\} + \varepsilon(x, y, t)\ , \qquad (1)$$

and

$$Z(x, y, t) = a\cdot Z(x, y, t-\Delta t) + b\cdot\{Z(x-\Delta x, y, t) + Z(x+\Delta x, y, t) + Z(x, y-\Delta y, t)+Z(x, y+\Delta y, t)\} + \varepsilon(x, y, t)\ . \qquad (2)$$

Equation (1) specifies a value at location (x, y, t) as a function of the preceding *in situ* value (time t-Δt) as well as the preceding neighboring values, a lagged specification. Equation (2) specifies a value at location (x, y, t) as a function of the preceding *in situ* value (time t-Δt) as well as the contemporaneous neighboring values, a spatially contemporaneous specification. The random process ε is white noise, which is uncorrelated in space and time. In the STAR model, correlation in space and time is captured by the autoregressive structure of the model (i.e., the response variable appears in both sides of the equations). Feedback loops or cycles make equation (2) fundamentally different from equation (1). An initial field for t = 0 and spatial boundary conditions are needed in these formulations. In this paper, interest is in the case where sufficient time has transpired and the spatial extent is sufficiently large to allow negligible effects from boundary conditions.

3.1 Theoretical Space-time Correlations

Theoretical correlations can be posited for equations (1) and (2). Consider an infinite regular square (i.e., equal-sized pixels) tessellation lattice for which spatial adjacency (i.e., geographic neighbors) is defined by whether or not two square cells share a non-zero length common boundary (i.e., the rook's definition). Let {Z(**s**)} be a Gaussian random variable distributed across the vector of locations **s** (i.e., cells), such that {Z(**s**)} and {Z(**s**+**h**)}, for locations shifted by **h** units, are stochastically equivalent (i.e., complete stationarity). Spectral theory (Bartlett 1975; Haining 1978) reveals that the appropriate correlation function for the additive specification [i.e., equation (1)] is given by

$$\rho_{h,g,k} = \frac{\displaystyle\int_0^\pi\!\!\int_0^\pi\!\!\int_0^\pi \frac{\cos(h\times t)\cos(g\times u)\cos(k\times v)}{\{1-\rho_T\cos(t)-\rho_s\cos(u)-\rho_s\cos(v)\}^\eta}\,dt\,du\,dv\ ,}{\displaystyle\int_0^\pi\!\!\int_0^\pi\!\!\int_0^\pi 1/\{1-\rho_T\cos(t)-\rho_s\cos(u)-\rho_s\cos(v)\}^\eta\,dt\,du\,dv} \qquad (3)$$

whereas that for the multiplicative specification [i.e., equation (2)] is given by

$$\rho_{h,g,k} = \frac{\displaystyle\int_0^\pi\!\!\int_0^\pi\!\!\int_0^\pi \frac{\cos(h\times t)\cos(g\times u)\cos(k\times v)}{[1-\cos(t)\{\rho_s[\cos(u)+\cos(v)]+\rho_T\}]^\eta}\,dt\,du\,dv\ ,}{\displaystyle\int_0^\pi\!\!\int_0^\pi\!\!\int_0^\pi 1/[1-\cos(t)\{\rho_s[\cos(u)+\cos(v)]+\rho_T\}]^\eta\,dt\,du\,dv} \qquad (4)$$

for temporal lag h (h = 0, 1, …), and spatial lags g (g = 0, 1, …) and k (k = 0, 1, …), where a positive integer value of η yields a η^{th}-order model, ρ_s is the spatial and ρ_T is the temporal autoregressive parameter, and (h, g, k) denotes the space-time lag involved.

3.2 Space-time Autoregressive Structures

The eigenvalues of the n-by-n connectivity matrix **C** for a linear surface partitioning containing P cells are $2\cos(\frac{p\,\pi}{P+1})$, p = 1, 2, …, P. The 2 can be absorbed into the autoregressive parameter values, ρ_j, doubling the size of each feasible. This solution can be extended to two- and three-dimensional regular square lattice structures. Ord (1975) first reported the eigenvalues of the PQ-by-PQ connectivity matrix for a square tessellation surface partitioning forming a P-by-Q (n = PQ) complete rectangular region as $2[\cos(\frac{p\,\pi}{P+1})+\cos(\frac{q\,\pi}{Q+1})]$, p = 1, 2, …, P, and q = 1, 2, …, Q. Gasim (1988) presents extensions to Ord's results. And, Basilevsky (1983) summarizes the conventional time-series results. Here the three-dimensional matrix representation is given by

$$\mathbf{C} = \mathbf{I}_T\otimes\mathbf{I}_s - \rho_s\mathbf{I}_T\otimes\mathbf{C}_s - \rho_T\mathbf{C}_T\otimes\mathbf{I}_s,$$

where ⊗ denotes Kronecker product, $\mathbf{I}_T$ denotes the T-by-T identify matrix, $\mathbf{I}_s$ denotes the PQ-by-PQ identity matrix, and $\mathbf{C}_s = \mathbf{C}_P\otimes\mathbf{I}_Q + \mathbf{C}_Q\otimes\mathbf{I}_P$, for a P-by-Q rectangular square lattice, where $\mathbf{C}_j$ is a matrix of 0s except for the upper- and lower-off diagonals, which contain 1s (j = P, Q, and T). $\mathbf{C}_P$ and $\mathbf{C}_Q$ have the same structure as $\mathbf{C}_T$.

Equation (3) describes the correlogram values for space-time data characterized by equation (1), whereas equation (4) describes space-time data characterized by equation (2). The three-dimensional connectivity matrix representation is given by

$$\mathbf{C} = \mathbf{I}_T\otimes\mathbf{I}_s - \rho_s\mathbf{C}_T\otimes\mathbf{C}_s - \rho_T\mathbf{C}_T\otimes\mathbf{I}_s\ ,$$

where 1 - $\cos(t)\{\rho_s[\cos(u)+\cos(v)]+\rho_T\}$ are the limiting eigenvalues of the space-time connectivity matrix **C**. Additional discussion of this topic appears in Griffith (1996).

Because the eigenvalues define the spectrum of a matrix, they appear in the denominator of equations (3) and (4); these denominators are based upon the limiting eigenvalues of the connectivity matrix representation of the space-time three-dimensional structure of data. In addition, Griffith and Csillag (1993) note, in contrast to the current thinking of that time, that a simultaneous autoregressive model can be portrayed by letting η = 2 in the denominator of equations (3) and (4)—it becomes a 2nd-order covariance specification; Bartlett (1975, pp. 19, 25) reports this result. Furthermore, Griffith and Layne (1997) summarize the close numerical connections between the spatial-only form of equations (3) and (4) and geostatistical semivariogram models.

3.3 Space-time Covariance Functions in Geostatistics

An important issue in the space-time geostatistical literature concerns whether or not the space and the time components of a formulated function are: separable such that they factor (Gneiting et al., 2006); or, nonseparable such that they form a linear combination (Ma 2008). Mitchell et al. (2005) propose a modified multivariate repeated measures likelihood ratio test coupled with bootstrapping for this purpose. Brown et al. (2001) note that separability requires that the expected value for some random variable at location (x, y) in time t+1, given its values in a neighborhood of location (x, y) in time t, must equal the conditional expectation just for location (x, y).

Stein (2005) furnishes an overview of space-time covariance and aspects of spatial-temporal interaction, and proposes a new class of space-time covariances. Ma (2003, 2008) presents methods for constructing spatio-temporal stationary covariance models, and supplements the set presented by Kolovos et al. (2004). Gneiting et al. (2006) posit theorems for symmetric and separable specifications, the Cressie-Huang and the Gneiting model, and stationarity. Fuentes et al. (2008) propose a nonstationary and nonseparable spectral density specification for which separability is a special case. Finally, Calder (2007) proposes a Bayesian specification that includes priors on initial points in time.

The space-time separability assumption (Bogaert 1996) states that the space-time covariance function C(h, u) can be written as a product of a spatial, $C_S(h)$, and a temporal, $C_T(u)$, covariance function, such that

$$C(h, u) = C_S(h)\cdot C_T(u)\ . \tag{5}$$

One non-separable specification expresses the space-time covariance function as a linear combination of these two components (De Cesare et al. 2001), such that

$$C(h, u) = C_S(h) + C_T(u) + p\ C_S(h)\cdot C_T(u)\ , \tag{6}$$

which is statistically valid if both $C_S(h)$ and $C_T(u)$ are valid covariance functions and parameter p satisfies certain conditions (De Cesare et al. 2001). This product-sum model appears to perform well in practice (De Iaco et al. 2003, Gething et al. 2007).

Another alternative is the metric model (Dimitrakopoulos and Luo 1994), which reduces the space-time covariance function to

$$C(h,u) = C_{ST}(\sqrt{h^2 + (\alpha \cdot u)^2})\ , \tag{7}$$

whose essential characteristic is that distance in space is made comparable to distance in time through the scaling parameter α. Equation (7) is rather restrictive because it assumes that the variances in time and space are equal. The following more flexible specification results from combining equation (6) with p = 0 and equation (7):

$$C(h,u) = C_S(h) + C_T(u) + C_{ST}(\sqrt{h^2 + (\alpha \cdot u)^2})\ . \tag{8}$$

The third term in the right-hand side represents a joint space-time interaction effect.

4. Numerical Experiments

Only numerical integration solutions to the definite integrals in equations (3) and (4) are available here. Because this integration is numerical intensive, and 1st- and 2nd-order results are similar, only 1st-order models are assessed. Because the exponent in the denominators of the integrands is 1, equations (1) and (2) refer to a space-time conditional autoregressive (CAR) specification. Numerical results for equations (3) and (4) were calculated for time lags h = 0, 1, …, 65 and space lags g and k = 0, 1, …, 50, using the autoregressive parameter pairs $\{(\rho_S, \rho_T)$: (0.49, 0.01), (0.40, 0.19), (0.30, 0.39), (0.20, 0.59), (0.10, 0.79), (0.01, 0.97)} (see Table 1). Theoretical nugget and sill values for equations (5)-(8) respectively are 0 and 1. Deviations from these values represent specification error; the numerical integration error is negligible.

The stable, the Bessel, and the exponential variogram models were evaluated in terms of their fits to these numerical data, with the exponential variogram model performing the best. Estimation results for this model appear in Tables 1 and 2, and suggest that equation (5) does not furnish a good description of the space-time structure generated by equations (1) and (2). Equation (8) fails to provide any improvement in the description furnished by equation (7), because equations (1) and (2) generate realizations from a random function that have the same sill (variance) in time and space; in cases where the variances differ, equation (8) will almost surely do better than equation (7). Equation (7) appears to yield a marginally better description than the one provided by equation (6). The principal difference between the relationship between equation (7) and equations (1) and (2) is the estimated α parameter, the anisotropic weight attached to the time distance in order to differentiate it from space distance, which is included in the specification of equation (7), but not equations (1) and (2).

5. Conclusions

In summary, numerical evaluation suggests that the STAR model equations (1) and (2) yield the metric model equation (7) with exponential-shaped variograms. However, real-world processes may, in addition to the space-time models characterized by equations (1) and (2), have purely spatial and purely temporal components. Whittle (1954) shows that purely spatial AR models have Bessel function-shaped covariance functions, whereas linear one-dimensional time series models have exponential variograms. Thus, processes that also have purely temporal and/or spatial components should be characterized by variogram models given by equation (8) rather than equation (7). Assuming that the generating

Table 1. Parameter estimates for the exponential variogram model and contemporaneous spatial dependence

ρ_s	ρ_T	space			time			a	RESS	space-time		
		C0	C1	r	C0	C1	r			C0	C1	r
Equation (5)												
0.01	0.97	1.2052	0.0992	0.5212	0.7661	0.0006	4.9569		0.8594			
0.10	0.79	1.2315	0.0514	0.7712	0.7788	0.0007	2.7328		0.9279			
0.20	0.59	1.3987	0.0459	0.8321	0.6917	0.0006	2.4328		0.9451			
0.30	0.39	1.1825	0.1004	0.0000	0.7787	0.0008	1.9168		0.9986			
0.40	0.19	1.2671	0.0309	0.8431	0.7697	0.0007	1.6549		0.9620			
0.49	0.01	1.3535	0.0283	0.8041	0.7231	0.0006	1.3169		0.9721			
Equation (6)												
0.01	0.97	0.0000	1.0000	0.4543	0.0096	0.9892	4.4235	-1.0000	0.0054			
0.10	0.79	0.0008	0.9992	0.5960	0.0281	0.9685	1.9877	-1.0000	0.0441			
0.20	0.59	0.0000	1.0000	0.5964	0.0270	0.9691	1.3773	-1.0000	0.0684			
0.30	0.39	0.0105	0.9894	0.5675	0.0106	0.9852	1.0190	-1.0000	0.0739			
0.40	0.19	0.0079	0.9920	0.5076	0.0049	0.9907	0.7222	-1.0000	0.0673			
0.49	0.01	0.0006	0.9992	0.2962	0.0000	0.9952	0.3038	-1.0000	0.1187			
Equation (7)												
0.01	0.97							0.0078	0.0029	0.0052	0.9948	0.3869
0.10	0.79							0.0714	0.0340	0.0208	0.9792	0.5214
0.20	0.59							0.1585	0.0560	0.0207	0.9793	0.5414
0.30	0.39							0.2784	0.0601	0.0174	0.9826	0.5381
0.40	0.19							0.4727	0.0526	0.0126	0.9874	0.5145
0.49	0.01							0.8494	0.1137	0.0044	0.9956	0.8494
Equation (8) [a]												
0.01	0.97	0	***	***	0	0.0014	***	0.0792	0.0029	0	0.9986	0.3869
0.10	0.79	0	0.0001	32.9987	0	0.0029	1.9357	0.0737	0.0337	0	0.9970	0.5135
0.20	0.59	0	0.0001	16.0388	0	0.0034	2.1593	0.1635	0.0552	0	0.9966	0.5317
0.30	0.39	0	0.0001	13.7028	0	0.0035	2.2861	0.2857	0.0588	0	0.9964	0.5280
0.40	0.19	0	0.0001	10.1898	0	0.0037	2.2348	0.4825	0.0509	0	0.9963	0.5046
0.49	0.01	0	0.0002	7.2160	0	0.0044	2.1385	0.8653	0.1112	0	0.9954	0.3806

[a] The three C0 terms were set equal to 0—the theoretical value— in order to achieve convergence.
NOTE: *** denotes an estimate at the limit of the numerically calculated space-time data cube.

processes satisfies the linear ARMA model, the temporal and spatio-temporal variograms may be described with exponential functions, whereas the spatial component may be described with a Bessel function.

6. References

Bartlett, M. (1975). *The Statistical Analysis of Spatial Pattern*. London, Chapman-Hall.

Basilevsky, A. (1983). *Applied Matrix Algebra in the Statistical Sciences*. NY, North-Holland.

Bogaert, P. (1996) "Comparison of kriging techniques in a space-time context." *Mathematical Geology* 28, 73-86.

Brown, P., P. Diggle, M. Lord, and P. Young. (2001). "Space-time calibration of radar rainfall data." *Applied Statistics* 50, 221-241.

Calder, C. (2007). "Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment." *Environmental and Ecological Statistics* 14, 229-247.

Cliff, A., P. Haggett, J. Ord, K. Bassett, and R. Davies. (1975). *Elements of Spatial Structure*. Cambridge, Cambridge U. Press.

De Cesare, L., D.E. Myers and D. Posa. (2001). "Product-sum covariance for space-time modeling: an environmental application." *Environmetrics* 12, 11–23.

De Iaco, S., D.E. Myers and D. Posa. (2003). "The linear coregionalization model and the product-sum space-time variogram." *Mathematical Geology* 35, 25–38.

Dimitrakopoulos, R. and X. Luo. (1994). Spatiotemporal Modeling: Covariances and Ordinary Kriging Systems - *Geostatistics for the Next Century*. R. Dimitrakopoulos. Dordrecht, Kluwer. 88–93.

Fuentes, M., L. Chen, and J. Davis. (2008). "A class of nonseparable and nonstationary spatial temporal covariance functions." *Environmetrics* 19, 487-507.

Gasim, A. (1988). "First-order autoregressive models: a method for obtaining eigenvalues for weighting matrices." *J. of Statistical Planning and Inference* 18, 391–398.

Gneiting, T., M. Genton, and P. Guttorp. (2006). Chapter 4 - *Statistical Methods of Spatio-Temporal Systems*. B. Finkenstaedt, L. Held, and V. Isham. Boca Raton, Chapman-Hall. 151-175.

Gething, P.W., P.M. Atkinson, A.M. Noor, P.W. Gikandi, S.I. Hay and M.S. Nixon. (2007). "A local space–time kriging approach applied to a national outpatient malaria data set." *Computers & Geosciences* 33, 1337–1350.

Griffith, D. (1996). Spatial Statistical Analysis and GIS: Exploring Computational Simplifications for Estimating the Neighborhood Spatial Forecasting Model - *Spatial Analysis: Modelling in a GIS Environment*. P. Longley and M. Batty. London, Longman GeoInformation. 255–268.

Griffith, D., and F. Csillag. (1993). "Exploring relationships between semi-variogram and spatial autoregressive models." *Papers in Regional Science* 72, 283-295.

Griffith, D., and L. Layne. (1997). Uncovering Relationships Between Geo-Statistical and Spatial Autoregressive Models - *1996 Proceedings on the Section on Statistics and the Environment*. Washington, D.C., American Statistical Association. 91–96.

Haining, R. (1978). *Specification and Estimation Problems in Models of Spatial Dependence*. Evanston, IL, Department of Geography, Northwestern University.

Kolovos, A., G. Christakos, D. Hristopulos, and M. Serre. (2004). Methods for generating non-separable spatiotemporal covariance models with potential environmental applications." *Advances in Water Resources* 27, 815-830.

Ma, C. (2003). "Families of spatio-temporal stationary covariance models." *Journal of Statistical Planning and Inference* 116, 489-501.

Ma, C. (2008). (2008). "Recent developments on the construction of spatio-temporal covariance models." *Stochastic Environmental Research and Risk Assessment* 22, S39-S47.

Mitchell, M., M. Genton, and M. Gumpertz. (2005). Testing for separability of space-time covariances." *Environmetrics* 16, 819-831.

Ord, J. (1975). "Estimation methods for models of spatial interaction." *Journal of the American Statistical Association* 70, 120–126.

Stein, M. (2005). "Space-time covariance functions." *Journal of the American Statistical Association* 100, 310-321.

Whittle, P. (1954). "On Stationary Processes in the Plane." *Biometrika* 41, 434–449.

Table 2. Parameter estimates for the exponential variogram model and time-lagged spatial dependence

ρ_s	ρ_T	space			time			a	RESS	space-time		
		C0	C1	r	C0	C1	r			C0	C1	r
Equation (5)												
0.01	0.97	1.2950	0.1065	0.5216	0.7130	0.0006	4.9405		0.8597			
0.10	0.79	1.3720	0.0563	0.7786	0.6995	0.0007	3.4037		0.9366			
0.20	0.59	1.3902	0.0440	0.8213	0.6966	0.0007	2.0341		0.9525			
0.30	0.39	1.3016	0.0337	0.9219	0.7481	0.0009	1.5630		0.9607			
0.40	0.19	1.3459	0.0292	0.9110	0.7263	0.0009	1.1788		0.9711			
0.49	0.01	1.2671	0.0210	1.0209	0.7750	0.0015	0.4744		0.9816			
Equation (6)												
0.01	0.97	0.0000	1.0000	0.4564	0.0092	0.9897	4.4207	-1.0000	0.0054			
0.10	0.79	0.0003	0.9996	0.6219	0.0240	0.9733	1.9716	-1.0000	0.1266			
0.20	0.59	0.0000	0.9999	0.6627	0.1955	0.9779	1.3631	-1.0000	0.1093			
0.30	0.39	0.0000	0.9999	0.6925	0.0122	0.9856	1.0026	-1.0000	0.0650			
0.40	0.19	0.0039	0.9961	0.7478	0.0000	0.9986	0.7322	-1.0000	0.0874			
0.49	0.01	0.0000	0.9663	0.7875	0.8039	0.1919	0.2914	-0.9999	0.7325			
Equation (7)												
0.01	0.97							0.0079	0.0030	0.0047	0.9953	0.3880
0.10	0.79							0.0788	0.1190	0.0147	0.9853	0.5386
0.20	0.59							0.2001	0.1008	0.0104	0.9896	0.5885
0.30	0.39							0.4346	0.0575	0.0044	0.9955	0.6321
0.40	0.19							1.0732	0.0826	0.0000	1.0000	0.7076
0.49	0.01							10.3725	0.1179	0.0000	1.0000	0.9785
Equation (8) [a]												
0.01	0.97	0	***	***	0	0.0014	***	0.0080	0.0030	0	0.9986	0.3880
0.10	0.79	0	0.0001	18.4139	0	0.0019	3.2382	0.0802	0.1187	0	0.9981	0.5319
0.20	0.59	0	0.0001	19.6642	0	0.0017	2.6567	0.2029	0.1005	0	0.9982	0.5830
0.30	0.39	0	0.0001	8.1486	0	0.0019	2.9423	0.4382	0.0570	0	0.9981	0.6274
0.40	0.19	0	0.0001	7.6018	0	0.0010	3.6334	1.0772	0.0824	0	0.9989	0.7054
0.49	0.01	0	0.0003	0.7745	0	0.0002	7.8400	10.3960	0.1178	0	0.9996	0.9770

[a] The three C0 terms were set equal to 0—the theoretical value— in order to achieve convergence.

NOTE: *** denotes an estimate at the limit of the numerically calculated space-time data cube.

GIS AS PLANNING SUPPORT SYSTEM

Anthony G. O. Yeh

The University of Hong Kong

ABSTRACT

The development of GIS has a very close relationship with urban planning. GRIDS in the 1960s and IMGRID in the 1970s were developed to carry out map overlap analysis which is fundamental to urban and regional planning. In the early days of the development of GIS in the 1960s and 1970s, there were very few planning departments that installed GIS because of their expensive hardware and limited software and data. The decrease in the price of hardware, computer storage and devices, and accompanying improvement in the performance of hardware and software (particularly the speed of computer processors) and advancement in the data structure and related algorithms of vector-based GIS, has made the once expensive and time consuming GIS to be more affordable and workable. GIS is now more accessible to planners and is an important tool and database for urban planning both in the developed and developing countries. Recent development in the integration of GIS with planning models, 3-D visualization, virtual reality and the internet will make GIS more useful as a planning support system for urban and regional planning. Today, the main constraints in the use of GIS in urban planning are no longer technical issues, but the availability of data, organizational change, and training. If these constraints can be removed, GIS will be a more effective planning support system for urban and regional planning.

A UNIFIED SPATIAL MODEL FOR GIS

Maciej Dakowicz and Chris Gold

Dept. Computing and Mathematics, University of Glamorgan, Wales, UK.

KEY WORDS: GIS, Data Structures, Voronoi Diagrams, Delaunay Triangulations

ABSTRACT:

GIS (Geographic Information Systems) are concerned with the manipulation and analysis of spatial data at a “Geographic” scale. Apart from issues of storage, database query and visualization, they must deal with several significantly different types of spatial information. These may be roughly classified as: discrete objects; networks; polygonal maps; and surfaces. Each of these has a specific set of assumptions associated with it, a specific data structure, and a specific set of algorithms. This produces a high level of complexity in the construction, manipulation, analysis and comparison of these datasets. This paper reports the results of an attempt to integrate these based on a slight modification of the fundamental spatial query: “What is here?” where “here” is usually an (x, y) location. It summarizes our previous work on this topic, and presents the “big picture” for GIS.

We demonstrate that when “here” is replaced by “closest to here” the resulting proximal query (a Voronoi diagram) may be used to manipulate the four categories described above, with a resulting simplification of the system. All discrete objects become “fields”, with a value at any location. The catch is that in order to represent non-point objects a line-segment Voronoi diagram is required, and this has been shown to be extremely complex to construct on digital computers. Nevertheless, with the availability of several potential solutions, including our own, we can show the implementation of a basic GIS with a common data structure and set of operations. In addition, several other “awkward” spatial queries can be shown to be simplified.

1. INTRODUCTION

Most GIS systems use separate thematic “layers” (pages) to store different types of data. Layers can be composed of object or field type data. Each layer contains specific features or characteristics of the area, so there is a separate layer for the road network, the distribution of buildings or the terrain relief. The layers can be stacked on top of each other and various operations can be performed on them. Some of the simplest are queries using a single layer, including finding what is present at a specified location, what is the elevation at a specified point or what is the area of a selected parcel. Using more than one layer the operations can be performed on different objects and characteristics, so we can ask what is the nearest mailbox to a selected house, find houses located within a specified postcode area, or find the areas where tobacco is cultivated or if the land value is higher than a specified value.

However, although it is possible to compare and perform operations on different layers, there is no consistent method applicable to all data types. GIS has traditionally separated field and object layers and used different data structures to manage them (Burrough and McDonnell (1998)).

The solution used in this work is to represent field and object models using proximal maps. This would standardize the GIS operations and make the topology information available in every model. Each object in a proximal map is associated with a region enclosing the part of the map closer to that object than to any other. The adjacency between objects is clearly defined, as neighbouring objects share boundaries. Converting discrete object layers into proximal maps transforms the query “What is here?” to “What is closest to here?”, so the answer is available at any location. We have previously reported on various aspects of this project (Gold and Dakowicz (2006), Dakowicz and Gold (2006)). In this article we integrate the various components and show the validity of the “big picture” for GIS. We should perhaps emphasize that this is an exercise in producing a good “model of space” to form the framework for a good and consistent set of operations and algorithms. Given the many years of traditional GIS development there is no claim that the current work will immediately produce faster execution times. However, a consistent spatial model leads to a consistent spatial data structure appropriate to all the examples given. We believe the current work to be a good demonstration of the model’s viability: it is based on the extremely difficult problem of implementing the Line Segment Voronoi Diagram (LSVD) in a finite-precision computing environment, especially in its kinetic and interactive form.

2. A UNIFIED SPATIAL MODEL

The main objective of this research is to develop a unified spatial model for objects and fields and demonstrate its usefulness. In this work, a vector Voronoi Diagram (VD) considered to be “the fundamental spatial data structure” (Aurenhammer (1991); Okabe et al. (2000)) is used. The VD is a proximal map of the input data. The creation of the Voronoi diagram of a set of discrete points converts the model to a continuous field. The resulting diagram covers the whole map, the attribute value at any location inside each cell is available from its node and the adjacency relationships between nodes are clearly defined. The VD is associated with the Delaunay Triangulation (DT), which is its dual graph with edges connecting neighbouring points. Both are well studied and there are many algorithms allowing their construction and modification (Guibas and Stolfi (1985); Devillers (1998)). Two other structures - the crust and skeleton (Amenta et al. (1997); Gold and Snoeyink (2001)) - can be easily extracted from the VD/DT. These have many interesting properties and can be used in various GIS operations, including digital terrain construction or watershed generation. Additionally, it is possible to convert VD/DT models to rasters using various interpolation techniques to perform traditional analysis on them, such as slope estimation.

However, modelling line and polygon objects with the ordinary point Voronoi diagram is problematic, as the connectivity of two nodes in the VD depends on the distance between them and the configuration of neighbouring nodes. The ordinary VD does not guarantee preserving connectivity between selected nodes and this is why the Constrained Delaunay Triangulation (CDT) (Lee and Lin (1986); Chew (1987)) and the Line Segment Voronoi Diagram (LSVD) (Gold et al. (1995); Imai (1996); Held (2001) and Karavelas (2004)) were developed. Both structures preserve the configuration of the input line segments, but in different ways. In the CDT the input segments (constraints) are "forced" into the triangulation as edges and the only thing distinguishing constrained edges from ordinary edges is a flag, stating whether the edge is constrained or not. The resulting triangulation is not fully Delaunay and the structure of its dual Voronoi diagram is different from the ordinary VD. On the other hand, in the LSVD, the input segments are separate objects and each of them has an associated Voronoi region, just like point objects.

This work presents methods that can be used to construct and update the ordinary VD/DT as well as the more complex CDT and LSVD diagrams. Depending on the applications, the input data and intended operations, the ordinary VD/DT, CDT or LSVD can be produced from the input set of points or segments and points. Dynamic methods for mesh management are proposed. These allow the insertion and deletion of points and line segments in the VD/DT/CDT/LSVD, making local updates possible at any time. The same idea and mechanism are used to construct different types of Voronoi diagrams. The method is based on the idea of moving points in the VD, and in the case of the CDT/LSVD the moving point leaves a trace behind, which becomes the line segment.

In most GIS there are four major categories of features and related data structures:

1. Discrete objects. These can be points, lines or polygons. Points are used to represent locations of features or objects having relatively small size. Examples include lamp posts locations or mailboxes. Lines consist of series of connected points and can be used to represent linear features, e.g. roads or geological faults. Polygons store relatively large single objects, usually with clearly defined boundaries, such as buildings or lakes. Some objects may be mobile, changing their position in the map, e.g. people or cars.
2. Networks. These are a special case of connected lines with defined topologies. Networks consist of segments, with nodes where segments join. They are used for example to model flow in rivers or roads.
3. Polygonal maps. These are space exhausting, so the whole area is covered by non-overlapping polygons. Examples include postcode boundaries or land ownership information.
4. Surfaces. They are space-filling, and store information about the relief of the terrain or other "field" information.

We provide examples of these four categories of spatial data types represented by Voronoi diagrams, as in Figure 1. Discrete objects may be points – which are converted to fields by calculating the Voronoi diagram. They may also be polygonal objects, such as houses – which may be converted to fields by calculating the LSVD or CDT from boundary segments. Networks, the second category, can be converted to fields by the LSVD or CDT constructed for their segments. These preserve connectivity of the links, so flow can easily be simulated, and adds the possibility of additional analyses, such as the proximity to the nearest network segment. The third category is polygonal maps, which are fields covering the whole map area. Representing them with the LSVD adds topology (connectivity) to the map automatically as edges are added. The fourth category, surfaces, can be constructed from point, contour or raster data, giving a field model based on the DT. Mobile objects can be handled in any of these four modes by using the kinetic Voronoi diagram: applications include collision detection and map updating and downdating.

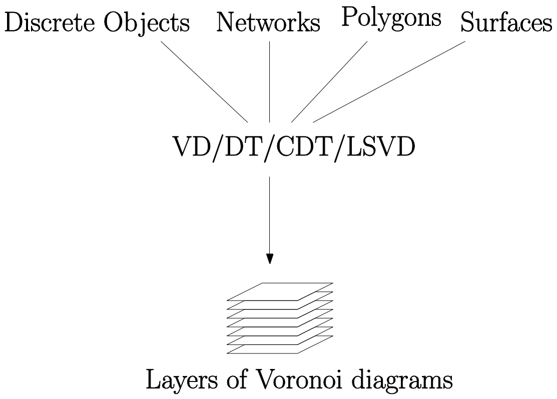


Figure 1. Integration of various models using Voronoi diagrams.

3. DISCRETE OBJECTS

We can distinguish three main types of discrete objects: points, polygons and mobile objects. Points can be converted to fields simply by constructing the Voronoi diagram. Figure 2 shows the VD of a set of rock outcrops of various types distinguished by the numbers assigned to different locations. The heavy boundary lines separate cells with different values of labels, and thus gives an approximate geological map.

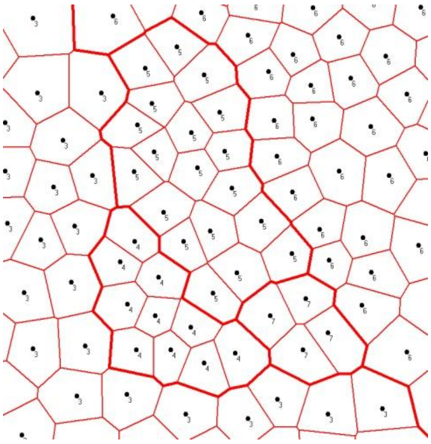


Figure 2. Boundaries between different types of rocks.

It should be noted that the VD could be applied to other types of point data: for example forestry. In this case the cell area may be an approximation of the tree crown size, and the tree density (trees per unit area) may be determined as the reciprocal of the cell size (area per unit tree), eliminating the traditional "counting circle" approach, emphasizing the generality of the proximal model.

Discrete polygons can be converted to fields by constructing the CDT or LSVD for the boundary nodes and lines, so each polygon can be represented using line segments or constrained edges. Figure 3 presents two different representations of the same polygonal map of buildings. In Figure 3a the map is converted to a Line Segment VD (Gold and Dakowicz (2006)), and buildings are represented using polygons formed by closed loops of line segment objects. Each complete polygon has a general Voronoi cell associated with it, and neighbourhood relationships are established. Figure 3b shows the same set of polygons converted to the CDT, where boundaries of polygons are represented by constrained edges. However, polygons here are not objects as in the LSVD.

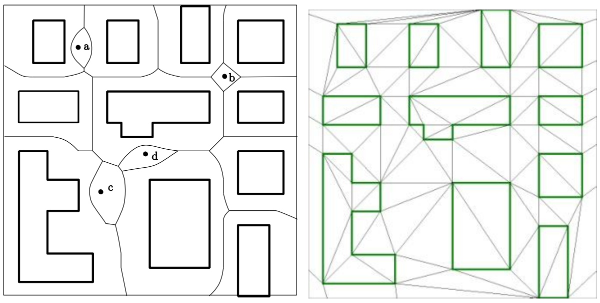


Figure 3. Polygonal objects. a) Represented by the LSVD, plus points. b) Represented by the CDT

Mobile objects can also be managed by incorporating them into the Voronoi diagram and using the moving point approach to change their locations. Figure 3a shows four point objects, which could be cars or people: objects c and d are adjacent and share a common Voronoi edge.

4. NETWORKS

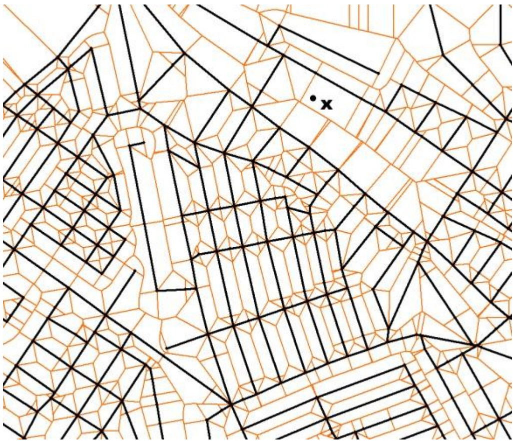


Figure 4. The LSVD of a road network.

Networks using various structures associated with the VD can be modelled in several ways. Figure 4 shows a drawing of an imaginary road network. Such an image is usually converted to a digital network by digitizing it. However, this often leads to errors at junctions, where several segments of the network meet, and the challenge is to assure that they are joined. Our kinetic algorithm (Gold and Dakowicz (2006)) eliminates problems at the junctions, as all objects have buffer zones associated with them and are merged when their buffer zones overlap. The result is a space-filling tessellation with all cells connected, including those for adjacent road segments. Figure 4 shows the LSVD created from the data. Each part of the network is represented by a line segment object and they are correctly joined at the junctions. The topology is defined and various analyses, such as shortest path queries, can be performed. Additionally various proximity relationships are readily available within the 2D space, and not just in the network, so it is simple to calculate the distance from point x in Figure 4 to the nearest road segment, for example.

4.1 River Networks: The Crust and Skeleton

If a set of points represents samples on the boundary of a polygon or along a network then constructing the VD/DT permits reconstruction of the boundary or "crust" if the samples are close enough (Amenta et al. (1997)). The medial axis, or "skeleton" (Blum (1967)) can be easily formed in a similar manner (Thibault and Gold (2000)).

The river network, even without any elevation data, provides enough information to construct a very reasonable watershed model (Gold and Dakowicz (2005)). The idea of Blum (1967), who assigned elevation values to the medial axis based on the distance from the crust, may be used here. The watershed is often roughly equidistant between river segments, as slopes on each side of the river are often similar, so it can be approximated with the medial axis.

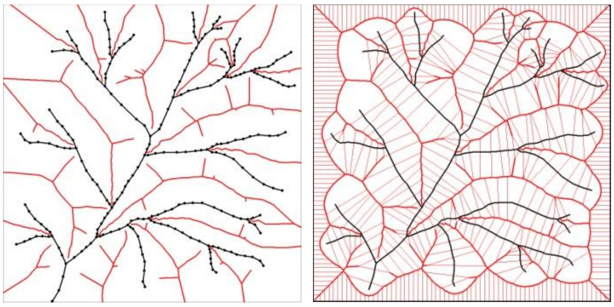


Figure 5. The river dataset. a) Crust and skeleton (watersheds). b) Flow cells.

Figure 5a shows the crust (river channel) and skeleton (watershed) constructed from the samples. Figure 5b shows the Voronoi cells associated with each sample. Traditionally this is estimated from grids, failing to preserve the Euclidean metric and making various dubious approximations of flow between adjacent cells. Here full connectivity of the original data is preserved by the data structure.

4.2 Cumulative Catchment Areas

Assuming a fixed rainfall throughout the map, the river network data can be used to compute cumulative catchment areas (Gold and Dakowicz (2005)). In the TIN model constructed from the samples each node of the river network has a Voronoi cell associated with it. The volume of water in each cell is proportional to the cell size. The cumulative sum of these volumes downstream gives an estimate of the total water flow at any river node. This can be seen in Figure 6a, where the height of the bars represents the volume of water. This cumulative catchment model provides a first-order approximation of total flow, using the proximal model to associate rainfall areas with river samples. Clearly our proximal model has transformed a network into a surface model, while preserving the network connectivity for subsequent flow analysis.

Note the buffer zones around the river system in Figure 6b. These are simple partitions of the Voronoi cells, and may be used to select regions within a fixed distance of the network (or of any other cartographic object). Itself a polygon object, it can be used for a variety of analyses, such as limiting forestry close to a river. This is often used in GIS analysis by overlaying it with choropleth (polygonal) maps of soils, political zones, etc. This is considerably more efficient than traditional buffer zone calculation if the VD has been previously calculated.

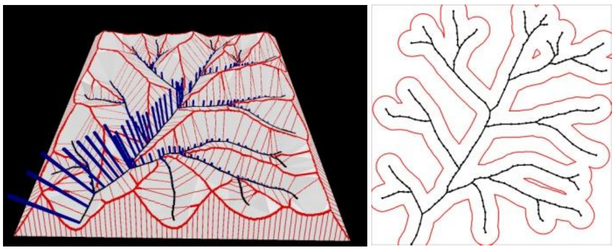


Figure 6. Modelling river networks. a) The cumulative catchment areas. b) River with buffer zones.

5. POLYGONAL MAPS

Polygonal maps can be converted to fields by making a CDT or LSVD from the edges defining the polygon boundaries. Figure 7 shows a sample polygon map and the resulting LSVD, with line segment objects approximating shapes of polygons. Each map polygon is the sum of the Voronoi cells of its boundary interiors. These may be overlaid for subsequent analysis.

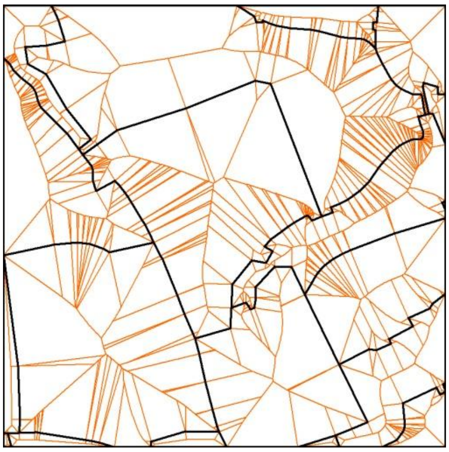


Figure 7. A polygon map and the resulting LSVD.

6. SURFACES

Surfaces form the fourth category of spatial models discussed. There are three main types of surface data: elevation data points, grids and contours. These can readily be converted to fields by generating the VD/DT. Contours can be converted to the TIN model by extracting their samples and triangulating them. Cases of "flat triangles" where all three vertices are on the same contour in peaks, pits, ridges and valleys can be handled by generating skeleton points from the contours and assigning elevations to them based on the principle of constant slope (Thibault and Gold (2000)). More generally, interpolation assumes that a value may be estimated at any location, whether from the VD/DT or by more complex methods that attempt to preserve slope continuity. While "counting circles" and the "gravity model" are sometimes still used, the VD approach is based on a consistent spatial model and produces good results. The VD based Sibson method is also called "area-stealing" or "natural neighbour" interpolation ((Gold (1989), Sibson (1982), Watson (1987)). It is based on the idea of measuring the areas that a dummy point inserted at the interpolated location would "steal" from its neighbours, and then using them as the weights for the weighted-average. These neighbours (called natural neighbours) are well defined, since the insertion of a point in the VD/DT produces a unique result, and the Sibson method is particularly appropriate for poor data distributions.

6.1 Runoff Modelling

The two most widely used formulations for water flow modelling are the Finite Element Method (FEM) and the Finite Difference Method (FDM). FEM methods use irregular meshes, FDM methods are often grid based. A form of the FDM, the Integrated Finite Difference Method (IFDM), is based on non-structured meshes. A manual method of irregular cell construction originally suggested by MacNeal (1953) was fully defined by Narasimhan and Witherspoon (1976), and was automated using the Voronoi diagram by Lardin (1999), showing that an iterative finite difference scheme could be developed for Voronoi cells ("buckets") to hold the water, and using the DT to define the slopes between cells. The volume of water moved between adjacent cells depended on the gradients of the triangle edges. Water is distributed irregularly, based on the distribution of cells adjacent to the processed cell.

The efficiency and stability of IFDM simulation is determined by the shape of the cells. The method is based on the idea of moving water volumes between neighbouring Voronoi cells with the amounts determined mostly by the gradient of the cells and the width of the common edge. A problem occurs when there is a large elevation difference between two cells sharing a very narrow common Voronoi edge. Then a large amount of water may accumulate in that cell but only leave it slowly through this narrow edge, giving a poor simulation. To avoid such cases a random pattern of points is generated and inserted into the existing TIN with a guaranteed minimum spacing, as in Figure 8a. A relatively large disk radius is assigned to points and used for collision detection to prevent insertion of points too close to each other. This leads to a fairly regular distribution of Voronoi generators. Additional points have elevation values assigned using Sibson interpolation (Sibson (1982); Gold (1989)), although other interpolation methods leading to a natural and smooth surface can be used. Figure 8a shows the 3D view of the resulting TIN model with the original contour lines draped over the terrain. Figure 8b shows the 3D view of the Voronoi cells of the same model part-way through a simulation.

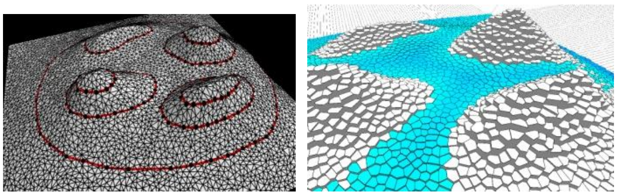


Figure 8. Runoff modelling. a) The 3D view of the Delaunay triangulation with superimposed contour lines. b) Voronoi view, part way through the simulation.

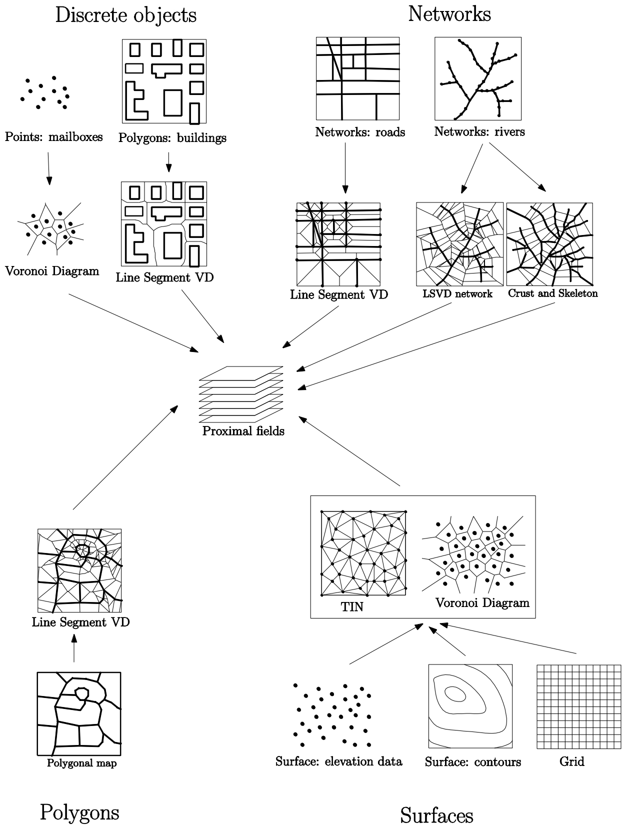


Figure 9. Transforming various data types into proximal maps.

7. INTEGRATION

Figure 9 summarizes the main types of data that may be transformed into proximal field models (Voronoi diagrams or Delaunay triangulations). The result is a set of layers or overlays: different data sets covering the same area and using the same spatial structure. Spatial analysis, which in GIS frequently consists of overlaying data sets to identify potential conflicts, may then be performed in a consistent and straightforward fashion. Merging two layers is performed by drawing the secondary layer onto the primary layer, snapping lines or points together whenever collisions, or close collisions, occur. For the classical polygon-polygon overlay care must be taken to preserve the attributes associated with each half-edge: the resulting overlaid map must have polygons with one attribute set from each original layer. Other combinations of overlays are equally straightforward: locating points within polygons, road segments within counties, houses within city wards, mailboxes adjacent to roads, roads on terrain models, trees on landscapes, cars on two-lane highways, surface-water runoff, and many more. Conversion of Voronoi (TIN) to raster is trivial; the inverse is frequently done in terrain or 3D model simplification. The resulting data structure (a graph) is in a form known to be well handled by computer.

Albrecht (1996) suggested a set of 20 universal analytic GIS operations, in six categories: Search (including interpolation); Location Analysis (including buffer, overlay, Voronoi); Terrain Analysis; Neighbourhood analysis; Spatial analysis (including pattern and shape) and Measurements (metric properties). Our spatial model directly addresses most of these.

Figure 10 shows the process of extracting buffer zones from a network map represented by the LSVD (Figure 10a). Figure 10b shows buffer zones drawn on top of the line segment Voronoi diagram.

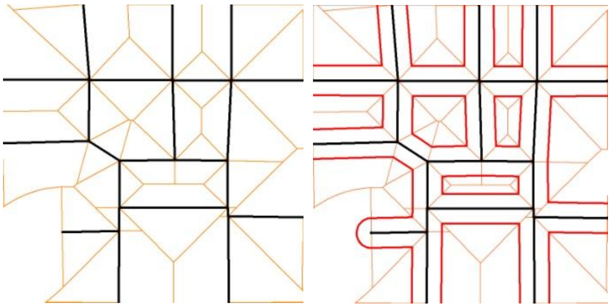


Figure 10. Road network and buffer zones. a) The LSVD. b) Buffer zones of network segments

Figure 11 shows the process of polygon overlay. Figure 11a shows a LSVD created from polygons defining building boundaries. Figure 11b shows the diagram after adding the road network of Figure 10a to the map of buildings. Figure 11c shows the result of overlying the map of buildings with the buffer zones of roads from Figure 10b. Note that the overlaps of buildings and buffer zones (perhaps where the proposed road widening causes a conflict) are themselves polygons.

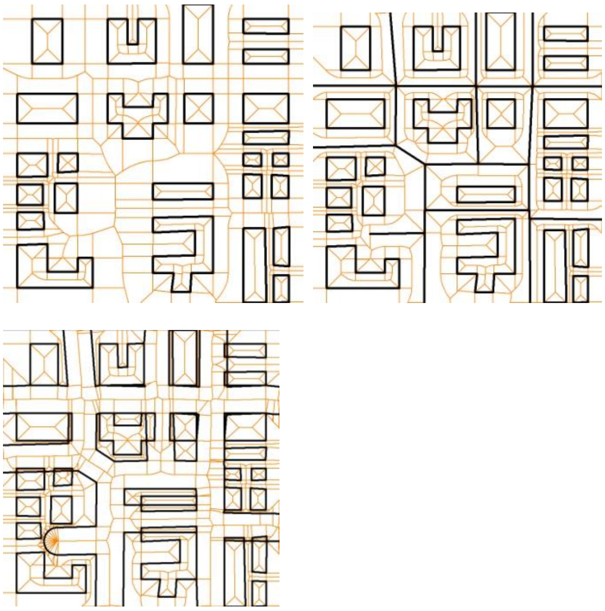


Figure 11. Polygon overlay. a) buildings, b) merged buildings and roads, c) merged buildings and road buffer zones.

8. CONCLUSION

In conclusion, a single spatial model for all (or many) types of spatial data provides a solid algorithmic framework, as well as a clarification of many types of spatial query and analysis that are currently performed with a wide range of frequently inconsistent heuristics. We believe that the proximal (Voronoi) model greatly simplifies the formulation and architecture of geographic spatial analysis. We suggest that the associated spatial model may replace many of the varied data structures used by traditional GIS: while conversion is straightforward this is not usually necessary, and our data structure permits the direct implementation of most basic forms of GIS analysis.

9. REFERENCES

Albrecht, J. H. 1996. Universal GIS operations for environmental modelling. In Third International Conference/Workshop on Integrating GIS and Environmental Modelling, Santa Fe, Mew Mexico.

Amenta, N. and Bern, M. and Eppstein, D., 1997. The Crust and the beta-Skeleton: Combinatorial Curve Reconstruction., Research Report, Xerox PARC.

Aurenhammer, F., 1991. Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure., ACM Computing Surveys., pp. 345-405.

Blum, H., 1967. A Transformation for Extracting New Descriptors of Shape., MIT Press, pp. 362-380.

Burrough, P. A. and McDonnell, R., 1998. Principles of Geographical Information Systems., Oxford University Press. New York, NY, USA.

Chew, L. P., 1987. Constrained Delaunay Triangulations. Proceedings of the Third Annual Symposium on Computational Geometry (SoCG), pp. 215-222.

Dakowicz, M. and Gold, C. M., 2006. Structuring Kinetic Maps. In Reidl, A. and Kainz, W. and Elmes, G., editors, Progress in Spatial Data Handling-12th International Symposium on Spatial Data Handling, Springer-Verlag Berlin, pp. 477-493.

Devillers, O., 1998. On Deletion in Delaunay Triangulation., Research Report 3451, INRIA.

Gold, C. M., 1989. Surface Interpolation, Spatial Adjacency and GIS., In Raper, J., editor, Three Dimensional Applications in Geographic Information Systems, Taylor and Francis, London, England, Taylor & Francis, pp. 21-35.

Gold, C. M. and Remmele, P. R. and Roos, T., 1995. Voronoi Diagrams of Line Segments Made Easy. In Gold, C. M. and Robert, J. M., editors, Proceedings of the 7th Canadian Conference on Computational Geometry, Quebec, QC, Canada, pp. 223-228.

Gold, C. M. and Snoeyink, J., 2001. A One-Step and Skeleton Extraction Algorithm., Algorithmica. pp. 144-163.

Gold, C. M. and Dakowicz, M., 2005. The crust and skeleton - applications in GIS. In Proceedings 2nd International Symposium on Voronoi Diagrams in Science and Engineering, Seoul, Korea, pp. 33-42.

Gold, C. M. and Dakowicz, M., 2006. Kinetic Voronoi/Delaunay Drawing Tools. In Proceedings of the 3rd International Symposium on Voronoi Diagrams in Science and Engineering. pp. 76-84.

Guibas, L. and Stolfi, J., 1985. Primitives for the Manipulation of General Subdivisions and the Computation of Voronoi Diagrams., ACM Transactions on Graphics, pp. 74-123.

Held, M., 2001. VRONI: An Engineering Approach to the Reliable and Efficient Computation of Voronoi Diagrams of Points and Line Segments., Computational Geometry: Theory and Applications, pp. 95-123.

Imai, T., 1996. A Topology Oriented Algorithm for the Voronoi Diagram of Polygons., In Proceedings of the 8th Canadian Conference on Computational Geometry, Carleton University Press, pp. 107-112.

Karavelas, M. I., 2004. A Robust and Efficient Implementation for the Segment Voronoi Diagram., In Proceedings of the International Symposium on Voronoi Diagrams in Science and Engineering, pp. 51-62.

Lardin, P., 1999. Le diagramme Voronoi generalise comme support a la simulation des ecoulements d`eau souterraine par differences finies integrees., M.Sc. Thesis., Laval University, Quebec City, Canada.

Lee, D. T. and Lin, A. K., 1986. Generalized Delaunay Triangulation for Planar Graphs., Discrete Computational Geometry, pp. 201-217.

MacNeal, R.H., 1953. An Asymmetrical Finite Difference Network., Quarterly of Applied Mathematics, pp. 295-310.

Narasimhan, T. N. and Witherspoon, P. A., 1976. An Integrated Finite Difference Method for Analyzing Fluid Flow in Porous Media., Water Resources Research, pp. 57-64.

Okabe, A. Boots, B., Sugihara, K. and Chiu, S. N., 2000. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams (2nd ed.), John Wiley & Sons.

Sibson, R., 1982. A Brief Description of Natural Neighbor Interpolation., In Barnett, V., editors, Interpreting Multivariate Data. John Wiley & Sons, New York, pp. 21-36.

Watson, D. F. and Philip, 1987. G. M., Neighborhood-based Interpolation, Geobyte, pp. 12-16.

SPATIAL MORPHOLOGICAL CONCEPTUAL MODEL OF BAY

D. D. Zhang [a,] *, X.M.Yang [a], F.Z. Su [a], Y.Y. Du [a]

[a] LREIS, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China - (zhangdd, yangxm, sufz, duyy)@ lreis.ac.cn

Commission VI, WG VI/4

KEY WORDS: Morphological, Conceptual model, Bay, Spatial Analysis

ABSTRACT:

Located in the conjunct part of sea area and land area, the morphology of bays is complicated and is changing seriously by the mutual action of natural and artificial factors. Therefore, the scientific and systematic description of bay is rather difficult. The research of the spatial morphological conceptual model of bays is meaningful for the representation of the bay and the construction of digital bays. Furthermore, it is also important for the change monitoring, exploitative intensity evaluation, as well as the further use of the bays. Nevertheless, the correlated research is rather less. A multi-scale expression model of bay is put forward in the paper. Firstly, the general morphological model of the bay which is in a bay-range scale is abstracted by Geometric Abstract according to the characteristics of bays, and then in the coast scale, five categories of coasts—— rocky coast, silty coast, sandy coast, artificial coast, and other type of coast were classified, a coast-range scale model of the bay based on these categories were proposed. Finally, through the construction of the spatial morphological conceptual model of the several typical bays—Daya Bay, Zhelin Bay and Jiaozhou Bay , the validity of the conceptual model is verified.

1. INTRODUCTION

The bay is a body of water partially enclosed by land, and with a baymouth opening to the sea (United Nations, 1983). It extends from the landward limit of tidal facies at its head to the seaward limit of coastal facies at its mouth, and covers the range of intertidal zone flat and offshore water area (Buatois,L.A,et al.,2003).With ascendant natural resources, dominant geographic location and environment condition, the bay area has become a natural region where is not only the most active of the earth surface, but also a part of coastal zone with highest intensity of human activities(Edition Committee of the Bay Chorography in China, 1999; Hugo,V.Z.et al.,2008). Meanwhile, by the mutual action of natural and artificial factors, the morphology of bays is changing all the time. Nevertheless, as a complex system with rich resource, multi-fields, multi-elements, and multi-level (Edition Committee of the Bay Chorography in China, 1999), the description of bays is rather difficult, therefore the systematic expressing and analyzing method of bays is rather less. The exiting research of the bay's spatial morphology can be categorized into two types: qualitative description and quantitative expression. For the qualitative way, the bays were described as bell-mouthed bay, door-shape bay, ζ -shape bay(Halligan,1906), half-heart-shape bay(Silvester,R.,1960), arc-shape bay(Rea, C. C., Komar, P. D,1975) and so on. For the quantitative way, indicators like open degree, shape coefficient were abstracted, and the bays were categorized into open bay, half-open bay etc. according to the values of open degree, shape coefficient and other indicators. Nevertheless, Neither of these shape description method of the bay can express the complex dynamic environment, as well as the land-sea interactions well. Therefore, an effective way of describing the morphology of bays is urgently needed.

This paper aimed at the morphological and structural characteristics of bays, a multi-scale conceptual model based on the differences of bays' type and characteristics of coasts were put forward. The following of this paper is organized as follows: part two is the abstraction method of bays' morphology and structure, and the construction method of conceptual model in the bay's scale as well as in the coast's scale; part five is case study of the conceptual model of three typical bays in china; part five is conclusions.

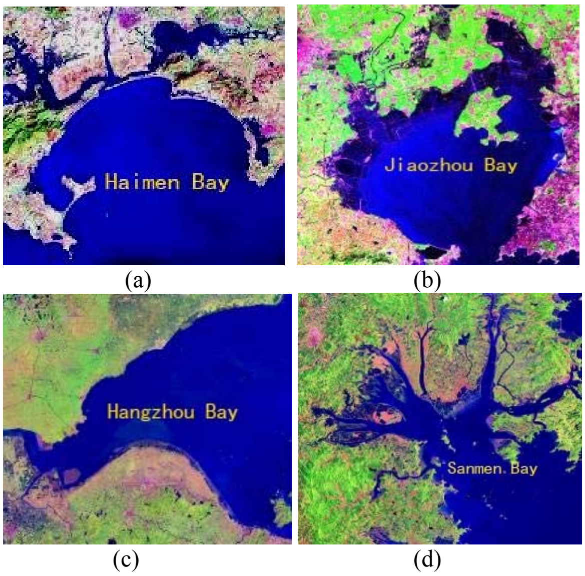
2. METHODOLOGY

2.1 Physical Characteristics of Bays in China

There are more than 200 bays with the area bigger than 5km^2 distributed along China's coastal zone. And most of these bays are distributed in Liaoning, Shandong, Zhejiang, Fujian, Guangdong, Guangxi, and Hainan province. Even located near each other, the morphological, dynamic characteristics are always different. The difference of the causing type, geographical environment, landform and hydrological characteristics lead to the morphology of bays are very different. They are not only appears in regular rectangle-shaped, half-circle-shaped, circle-shaped, door-shaped, but also in narrow-mouth-wide-bingy shaped, horn shaped, multi-vessel shaped and so on(Shown in Figure 1).From the quantitative point of view, the bays can be categorized into several types according to water area rate(WR), open degree(OD), and morphology coefficient(MC) (Wu S.,Y. ,2000; Zhang,D.D., 2008). Through these indicators, and the result of the coastal survey of the nation in 1980s, there werev31 Entire-water bays, 9 Much-water bays, 17 Middle- water bays,9Little-water bays,2Dry-water bays;15 Open bays,17 Half-open bays,21Half–close

* Corresponding author. This is useful to know for communication with the appropriate person in cases with more than one author.

bays,3Close bays; 17Long-narrow bays,13Wide-long bays,7 Square-round bays,5Long-wide bays,11Short-wide bays.



(a) (b)

(c) (d)

Figure 1 Different morphology of bays.(a) half-arc shaped;(b) heart-shaped;(c) horn-shaped;(d) multi-vessel shaped

2.2 General Model— bay-range scale model

According to the definition of China Bay Records, an integrated bay includes three parts: alongshore land area, tidal zone, and alongshore sea area. The baymouth is the borderline of bay and sea, the mean low-tide line is the borderline of sea area and tidal zone, while the mean high-tide line is the borderline of tidal zone and coastline. Therefore, a bay can be seen as a systematic structure formed by four axes: baymouth, mean low-tide line, coastline, land boundary, and three areas surrounded by these axes, namely alongshore land area, tidal zone, and alongshore sea area. Among these four axes, the up-boundary of land area can be establishment according to the research objective, coastal type, and the influencing range to land of different type of bank. However, due to the sea-land interaction, as well as the exploitative activities of bays by human beings, a large volume of silty soil fill up, coastline was eroded and is falling back gradually, tidal level was raising, the location and shape of the four axes were changing. As a result, the shape, structure, location and range were changing all the time, and the dynamic axis-area structure was forming. Among the axes, the coastline is most important for the shaping of the bay's structure, and fluxed obviously, because it endures intense land-sea interaction. Therefore, it was seen as the main axis of the structure. Figure 2 is the sketch map of the conceptual model of the spatial morphological of bay.

In this way, the complex of bay can be dissolved into separate geographical features, namely axis and area. And the bay can be expressed formally through formula (1), Formula (2) ~formula (4) is the expression of each areal geographical feature:

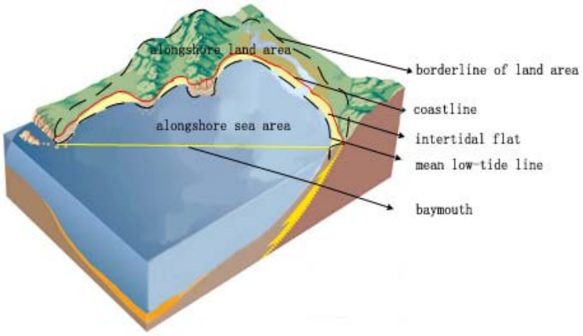


Figure 2 Sketch map of the conceptual model of spatial morphological of bay

$$M_{Bay} = f(L_C, L_{ML}, L_{UB}, L_{BM}, R_{AS}, R_{AL}, R_{TF}) \quad (1)$$

$$R_{AS} = f(L_{ML}, L_{BM}) \quad (2)$$

$$R_{AL} = f(L_C, L_{UB}) \quad (3)$$

$$R_{TF} = f(L_C, L_{ML}) \quad (4)$$

Where M_{Bay} = the morphology of the bay,

L_C = the location of coastline;

L_{ML} = the mean low- tide line

L_{UB} = the location of the upper borderline of the land area

L_{BM} = the location of the baymouth

R_{AS} = the range of alongshore sea area

R_{AL} = the range of alongshore land area

R_{TF} = the range of intertidal flats.

2.3 Conceptual Model in Coast Scale

According to the difference of the offshore coasts, conceptual models based on the characteristics of them are established. Considering of the differences of influencing intensity of offshore coasts to each axis and area, the establishing method of the location of axis and borderline of the area is different (Dandan Zhang, 2008).The location of coastline, baymouth, mean low-tide line is easy to established. Comparatively, the range of land area of different bays is complicated. A method of dynamic buffer zone which use dynamic degree of the most sensitive land use type at the alongshore coasts to establish the range of land area（SL）is brought out. The dynamic degree is established through Dynamic Degree Model of single land use change which can express the changing rate of some land use type, and it can be expressed in formula (5).

$$K = \frac{U_b - U_a}{U_a} \times \frac{1}{\mathrm{T}} \times 100\% \quad (5)$$

Where K = the dynamic degree of certain kind of land use

U_a = the area of land use in the beginning of the research period of time

U_b = the area of land use in the ending of the research period of time

T = represents the research time

Formula (6) ~ formula (9) is the formal expression of the range of alongshore land area with different coast types. Four types of coasts were categorized: rocky coast, silty coast, sandy coast, artificial coast, and other type of coast. The main sediment materials of rocky coast are rocks, and the intertidal flat is always sandy beach, or rocky shore platform, while the alongshore land area always developed dune, ridge plain, cliff , sloping bedrock terrain; Sandy coast develops sandy beach well in the intertidal flat, while the alongshore land area always develops bedrock terrain and dune; The intertidal flat of silty coasts are mainly loose sediments ,while the alongshore land area can develop cliff, bedrock terrain. Figure 3 is the profile sketch map of the sandy coast, rocky coast, and silty coast. In the point view of land use, different types of landform on coastal zones is suitable for different kind of land use type ,and the land use types suggests obviously regional characteristics, Figure 4 is the sketch map of the land use regional profile.

$$SL_{rock} = f(\mathrm{DD}_{\mathrm{woodland}}) \tag{6}$$

$$\mathrm{SL}_{\mathrm{silt}} = \mathrm{f}\ (\mathrm{DD}_{\mathrm{auquculture}}) \tag{7}$$

$$\mathrm{SL}_{sand} = \mathrm{f}(\mathrm{DD}_{\mathrm{grassland}}) \tag{8}$$

$$\mathrm{SL}_{artifical} = \mathrm{f}(DD_{\mathrm{farmland}},\ DD_{\mathrm{Resident}}) \tag{9}$$

Where

SL_{rock} = the range of the bays with most rocky coasts

$\mathrm{SL}_{\mathrm{silt}}$ = the range of the bays with most silty coasts

SL_{sand} = the range of the bays with most sandy coasts

$\mathrm{SL}_{artifical}$ = the range of the bays with most artificial coasts

DD_{woodland} = the Dynamic Degree of woodland

$\mathrm{DD}_{\mathrm{aquaculture}}$ = the Dynamic Degree of aquaculture

$\mathrm{DD}_{\mathrm{grassland}}$ = the Dynamic Degree of grassland

DD_{farmland} = the Dynamic Degree of farmland

DD_{Resident} = the Dynamic Degree of residential land area.

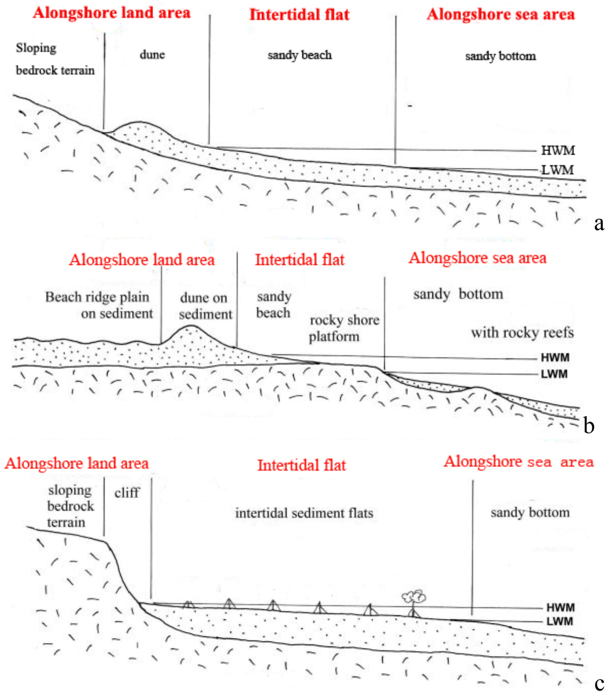


Figure3 Profile sketch map of three coastal type of bays.(a) sandy coast; (b) rocky coast; (c) silty coast.HWM stands for meanHigh-water Mark, LWM stands for mean Low-water Mark

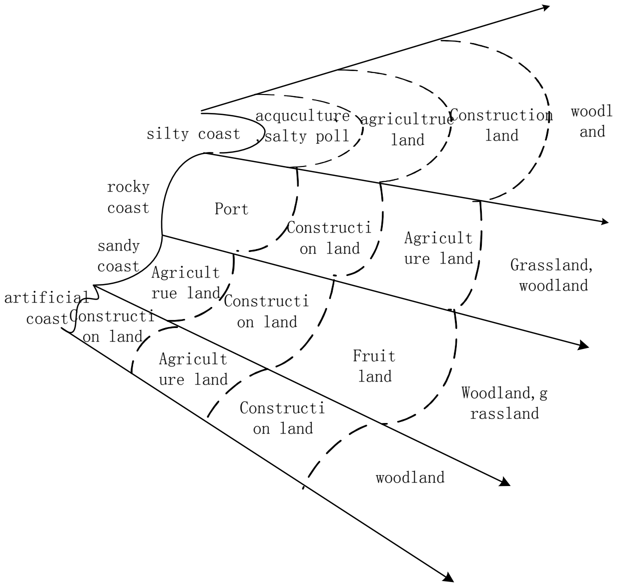


Figure 4 The sketch map of the land use regional profile

3. CASE STUDY

3.1 Study Area

For the further explanation of the construction process, Daya Bay, Zhelin Bay and Jiaozhou Bay with different coastal types and morphologies were taken as examples.

For these three bays, Daya Bay and Zhelin Bay is located in Guangdong province, while Jiaozhou Bay is in Shandong province; For the sea-area belonged, the formal two belongs to the South China Sea, while the later belongs to the Yellow Sea. The Main characteristics of the bays are shown as Table1.

Category	Values		
	Daya Bay	***Zhelin Bay***	***Jiaozhou Bay***
Open Degree	Half-open	Half-open	Half-close
Water Rate	Entire-Water	Entire-Water	Much-Water
Morphology Coefficient(MC)	Long-Wide	Wide-long	Square-round
Location	Huiyang,Guangdong	Raoping, Guangdong	Jiaozhou,Shandong
Main use type	Nuclear Power Station, port, aquaculture	Aquaculture,	aquaculture, port
Coast types	sand,rock, artifical	Sand,artifical	sand,rock,artifical

Table 1 Characteristic Comparisons of the three typical bays (Wu,S.Y.,2000;Zhang,D.D.,2008)

3.2 Data Source

Data used were for two objectives: one is for the establishment of the bay's coastal type in a quantitative manner; the other is for the Dynamic Degree analysis of different land use type in the establishment of the upper-border of the land area. Meanwhile, some basic data like the basic geographical features are also needed. Therefore, these data can be categorized into following types:

(1) Seamap. It is used for the definition of coastline as well as the location of the contour with 0m. The production date of these maps are 1984,1984 and 2005 for the Daya Bay, Zhelin Bay and Jiaozhou Bay respectively, while the scale is 1:60 000 ,1:25 000,1:35 000 respectively.
(2) Thematic data from the National Coastal Survey of china in 1980s.These data include the thematic of land use and landform, the scale is 1:20 000, while the projection is WGS_1984.
(3) Satellite imageries. TM, ETM, as well as SPOT imageries are all required. The usage of these images is for two purposes, one is for the background browse, and the other is for the acquirement of the thematic data. TM data in 2000 for Daya Bay, SPOT5 data in 2003 for Zhelin Bay, while ETM data in 2006 for Jiaozhou Bay is used.

3.3 Analysis Results

Through analysis of the coastal characteristics of Daya Bay,Zhelin Bay and Jiaozhou Bay by remote sensing images ,sea maps, and landform thematic maps of the coast in the 1980s, it is indicated that most part of the coast of Daya Bay is rocky coast, most part of the coast of Zhelin Bay is silty coast, while the Jiaozhou Bay is a sandy coast bay. The ratio of each type of coast of the three bays is shown in Table1.

Ratio of coastal types	**Values (%)**		
	Daya Bay	Zhelin Bay	Jiaozhou Bay
Sandy coast	34.1	14.8	38.8
Rocky coast	37.2	16.4	10.1
Silty coast	16.5	37.4	29.3
Artifical coast	11.1	30.6	16.2
Other type	1.1	0.8	4.6

Table 1 Ratio of the main coastal types of the typical bays

According to the coast-range model proposed, the ranges of the three typical bays were established. The land area of the Daya Bay,Zhelin Bay,Jiaozhou Bay is 3km buffer, 2km buffer and 1km buffer from the coastline respectively.Figure 5~Figure 7 is the range of them.

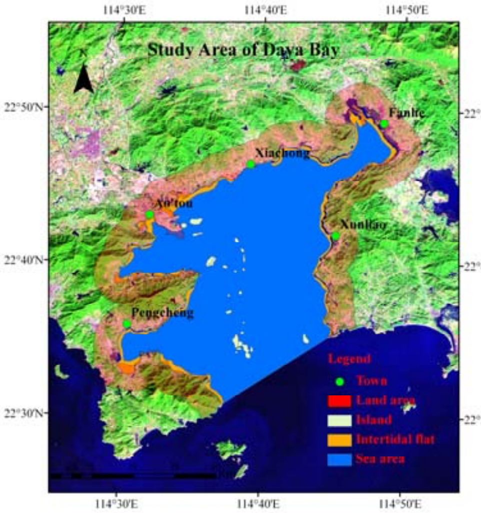


Figure 5 The range of Daya Bay with 3 km buffer zone from the coastline of land area

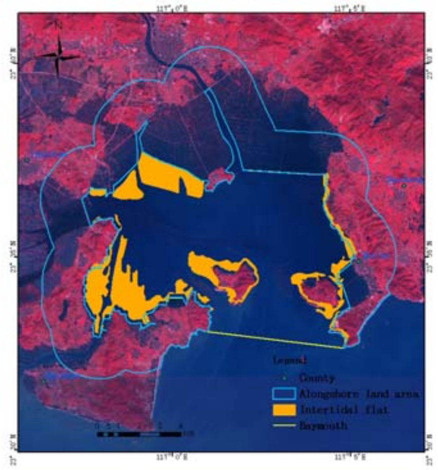


Figure 6 The range of Zhelin Bay with 2 km buffer zone from the coastline of land area

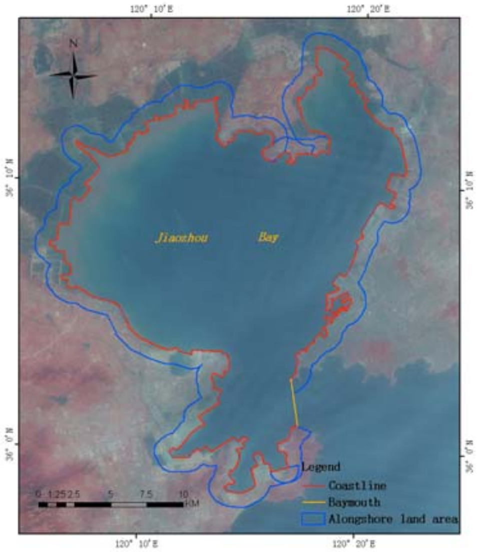


Figure 7 The range of Jiaozhou Bay with 1 km buffer zone from the coastline of land area

4. CONCLUSIONS

Aiming at the characteristics of multi-level and multi-field of bays, this paper proposed a multi-scale method for the expression of bays, and conceptual models in the bay-range scale, as well as the coast-range scale were constructed. Through the case study of three typical bays of Daya Bay, Zhelin Bay and Jiaozhou Bay, it is conclude that the conceptual model proposed in this paper is valid for the description of the morphology of bays, and therefore is meaningful for the construction of bay database and information management system.

ACKNOWLEDGEMENT

This study was funded by the National 973 Program of China (No.2006CB701305), National 863 Program of China(No. 2009AA12Z148).

REFERENCES

United Nations. UN Convention on the Law of the Sea, with Index and Final Act of the Third United Nations Conference on the Law of the Sea. New York,1983

Luis A. Buatois, M. Gabriela Mángano. Sedimentary facies, depositional evolution of the Upper Cambrian–Lower Ordovician Santa Rosita formation in northwest Argentina. Journal of South American Earth Sciences.16(5), 343-363(2003)

Hugo,V.Z., Barbour T., Hamann,R.,et al.Assessment of socio-economic impacts of sea harvest's operations on Saldanha Bay and the west coast district.EEU report on Sea Harvest socio-economic impacts,2008

Rea, C. C., Komar, P. D. Computer simulation models of hooked beach shoreline configuration [J].Sedimentary Petrology, 1975, 866-872

Silvester R. Stabilization of sedimentary coastlines [J]. Nature, 1960,467-469

Compilation Committee of Record of Bays in China.Record of bays in China: Vol. 14, Ocean Press, Beijing, 1998

Wu Sangyun,Wang Wenhai. Study on the classification system of bays. Acta Oceanologica Sinica,22(4),83-89(2000)

Dandan Zhang, Xiaomei Yang,Fenzhen Su,Xiaoyu Sun. A Dynamic Axis-area Analysis Method of Bay Use Change. IITA Conference on Geoscience and Remote Sensing (IITA-GRS 2008).Dec.21,2008

W-BASED VS LATENT VARIABLES SPATIAL AUTOREGRESSIVE MODELS: EVIDENCE FROM MONTE CARLO SIMULATIONS

An Liu[a], Henk Folmer[b], Han Oud[c]

[a]Department of Spatial Sciences, University of Groningen, P.O. Box 800, NL-9700AV Groningen, The Netherlands, an.liu@rug.nl, corresponding author

[b]Department of Spatial Sciences, University of Groningen, P.O. Box 800, NL-9700AV Groningen, The Netherlands, and Department of Social Sciences, Wageningen University, PO Box 8130, NL-6700 EW Wageningen, The Netherlands, henk.folmer@wur.nl

[c]Behavioural Science Institute, Radboud University Nijmegen, P.O. Box 9104, NL-6500 HE Nijmegen, The Netherlands, j.oud@pwo.ru.nl

KEY WORDS: spatial autoregressive model, structural equation model, latent variable, Monte Carlo simulation, bias, RMSE

ABSTRACT:

The paper evaluates by means of Monte Carlo simulations the estimator of the regression coefficient obtained by the classical W-based spatial autoregressive model and the structural equations model with latent variables (SEM) on the basis of data sets that contain two types of spatial dependence: spillover from (i) a hotspot and (iia) first order queen contiguity neighbors or (iib) inverse distance related neighbors. The classical models are either correctly specified or ignore (i), as is common in practice. SEM takes spatial dependence into account by means of a fixed number of nearest neighbors as well as the dependent variable in the hotspot weighted by inverse distance. The estimation results are analyzed in terms of bias and root mean squared error (RMSE) for different values of the spatial lag parameters, specifications of weights matrices and sample sizes. The simulation results show that compared to the misspecified models SEM frequently has smaller bias and RMSE and even outperforms the correctly specified models in many cases. These trends increase when the spatial lag parameter for spillover increases. The lead of SEM also increases by sample size. Finally, SEM is more stable in terms of both bias and RMSE over various dimensions.

1. INTRODUCTION

The conventional spatial regression model is based on a spatial weights matrix, usually denoted W, that accounts for spatial dependence and spill-over effects among the spatial units of observation. The latent variables approach (denoted SEM below), introduced by Folmer and Oud (2008), replaces the spatially lagged variables in the structural model by latent variables and models the relationship between latent spatially lagged variables and their observed indicators in the measurement model. SEM not only can produce virtually the same estimates as obtained by the classical approach but also is more general.

In order to gain insight into the characteristics of the estimators of the regression coefficients including the spatial autocorrelation coefficient produced by the classical approaches and SEM, Liu et al (2010a) carried out a series of Monte Carlo simulations on the basis of Anselin’s (1988) Columbus, Ohio, crime data set which was also used by Folmer and Oud (2008) for illustrative purposes. The latent spatial lag variable in the SEM model was measured by a number of nearest neighbors. Data was generated on the basis of a first order queen contiguity or an inverse distance weights matrix. The main result was that the classical approach (estimated with weights matrix consistent with the data generation matrix) had lowest bias and RMSE in the majority of cases. SEM outperformed the classical approach for some W matrices, however. Particularly, it had the smallest bias in several cases. Liu et al (2010b) examined the performances of the two approaches in the context of spatial dependence due to spillover from hotspots. In that case spatial lag variable was measured by the values of the dependent variable in hotspots weighted by inverse distance. The simulation results indicated that both approaches performed better for smaller values of the spatial lag parameter and larger sample sizes. SEM tended to outperform the classical approach in term of bias but the classical model based on first order contiguity matrix had lowest RMSE in most cases. Furthermore, SEM was most stable in terms of variations in both bias and RMSE. Globally speaking, the performances of both approaches do not differ much.

In this paper we further evaluate the performances of the classical W-based approach and SEM in a more general setting that combines the two different types of spatial dependence considered in the previous simulations. The remainder of the paper is organized as follows. Section 2 briefly specifies the model structures of the classical W-based approach and SEM. A description of the experimental design is given in section 3. In section 4 we report the simulation results and section 5 concludes the paper.

2. MODEL SPECIFICATIONS

The classical spatial autoregressive model reads:

$$y = \rho Wy + X\beta + \varepsilon \qquad (1)$$

$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad (2)$$

where y is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times k$ data matrix of explanatory variables with associated coefficient vector β, ε is an $n \times 1$ vector of error terms. W is the $n \times n$ spatial weight matrix, with spatial autoregressive or spatial lag parameter ρ. (For further details see amongst others LeSage and Pace, 2009)

A SEM in general form consists of two basic equations:

$$y = \Lambda\eta + \varepsilon \quad \text{with} \quad \operatorname{cov}(\varepsilon) = \Theta \tag{3}$$

$$\eta = \mathrm{B}\eta + \zeta \quad \text{with} \quad \operatorname{cov}(\zeta) = \Psi \tag{4}$$

Equation (3) is the measurement model with *y* the *p*-vector of observed variables or indicators, Λ the matrix of loadings of the observed variables (indicators) on the *k*-vector of latent variables η [1], and Θ is the $p\times p$ measurement error covariance matrix. In the structural model (4), B specifies the structural relationships among the latent variables and Ψ is the $k\times k$ covariance matrix of the errors in the structural model. The measurement errors ε are assumed to be uncorrelated with the latent variables η as well as with the structural errors ζ who are supposed to be uncorrelated with η. (For details on identification, estimation, testing and respecification of structural equation models see Jöreskog and Sörbom, 1996)

The SEM spatial autoregressive approach replaces the spatially lagged variable *Wy* in the W-based equation (1) by a latent variable η in the structural model. In the measurement model η is measured by a set of observed variables. This model structure implies that both spatially lagged variables related to neighboring regions and hotspots can be indicators of the same latent variable. For detailed model specifications see Folmer and Oud (2008).

3. EXPERIMENTAL DESIGN

The dependent variable (*y*) in each spatial unit is affected by the dependent variable in one or more neighboring units as well as by hotspots. For data generation this implies that besides the spatial structure, the hotspot also needs to be known. However, it is not until the samples are generated that we get to know which region is the hotspot (defined by highest value of *y*). To solve this problem we take a step backward and choose the 'potential' hotspot on the basis of the independent variable *x* instead. That is, we designate the hotspot according to the largest value of *x*.[2]

We consider regular lattice structures of dimensions 7×7 (N=49), 10×10 (N=100) and 15×15 (N=225). The spatial weight matrices are defined on these lattice maps. To generate samples we rewrite equation (1) as:

$$y = \rho_1 W_1 y + \rho_2 W_2 y + x\beta + \varepsilon \tag{5}$$

or

$$y = (I - \rho_1 W_1 - \rho_2 W_2)^{-1}(x\beta + \varepsilon) \tag{6}$$

where W_1 is the weights matrix representing the spatial hotspot structure. Particularly, W_1 is the inverse distance matrix with elements equal to $1/d_{ij}$ for cell *i* and hotspot *j* and 0 elsewhere), W_2 is the conventional first order contiguity or inverse distance matrix, and ρ_1 and ρ_2 are corresponding spatial lag parameters.

Next, *y* is generated as follows:

1. Generate the exogenous variable *x* by drawing from a uniform (0, 10) distribution.
2. Fix the regression coefficient for all simulation runs: $\beta = 1$.
3. The spatial lag parameters ρ_1 and ρ_2 take values 0, 0.1, 0.3, 0.5, 0.7 and 0.9 consecutively[3].
4. Generate values for the error term ε by randomly drawing from a normal distribution with mean zero and variance 2.0.
5. Choose the hotspot according to the values of *x* generated in step 1 and compute *y* according to equation (6).

Both a first order contiguity queen and an inverse distance W_2 is used to generate data.[4] We estimate two types of classical models. One, the TRUE model, estimated with the same W_1 and W_2 as in the model used for data generation. The second is in line with common estimation practice and considers only one overall type of weights matrix, viz. W_2 only. However, we consider both the first order contiguity and inverse distance matrix. Estimation of the SEM model is always based on the first three nearest neighbors and spillover from the hotspot. The estimators are compared in terms of bias and RMSE of β, the coefficient of the regressor *x*.[5] The number of replications is set to 500.

4. SIMULATION RESULTS

In this section we present the main simulation results for TRUE (estimated with W_1 and W_2 used to generate the samples), CONT (estimated with W_2 specified as first order contiguity matrix only), DINV (estimated with W_2 specified as inverse distance matrix only) and SEM (estimated with the first three nearest neighbors and spillover from the hotspot *j* as indicators for cell *i*). Observe that due to the restrictions on ρ_1 and ρ_2, not all parameter combinations are feasible, as explained in the previous section.

Table 1 reports the biases of the estimators of β for samples generated by spillover from hotspot and from first order queen contiguity neighbors for 49 observations. It shows that when $\rho_1 = 0$ and 0.1, CONT has lower biases than SEM in most cases and outperforms TRUE a few times, although the differences are quite small among all models. When $\rho_1 = 0$, $\rho_2 = 0$ or 0.7 and $\rho_1 = 0.1$, $\rho_2 = 0$ or 0.1,

[1] Observe that a SEM will not be identified if the latent variables have not been assigned measurement scales. It is convenient to fix the measurement scale of a latent variable by fixing one λ_i, usually at 1. That is, one often chooses $\lambda_1 = 1$.

[2] Here we only consider one hotspot. However, it is possible to consider several hotspots simultaneously (see Liu et al., 2010b)

[3] Note that in a spatial autoregressive model, the asymptotic properties of the ML estimator require $|I - \rho W| > 0$ (Anselin, 1988). In the present case this constraint is $|I - \rho_1 W_1 - \rho_2 W_2| > 0$.

[4] W_1 is always an inverse distance matrix.

[5] Since they are not directly comparable, we do not compare the spatial autoregressive coefficients (see Liu et al.,2010b).

SEM outperforms TRUE or performs equally well. When $\rho_1 \geq 0.3$, SEM outperforms CONT in almost every case and its dominance becomes more distinct as the value of ρ_1 increases. It is not surprising that CONT is better than SEM for small values of ρ_1, since it is the genuinely true model when $\rho_1 = 0$. Also observe that for each value of ρ_1, the biases of SEM tend to increase as ρ_2 increases. But the increase is not uniform over the interval of ρ_1 for a fixed ρ_2. This is probably due to the limited number of indicators, especially the fixed numbers of neighbors included in SEM estimation.

The RMSE for the model generated on the basis of a first order contiguity matrix is presented in Table 2. The table shows that CONT has smallest RMSE in more than half of the cases. Moreover, the RMSEs of SEM and CONT basically follow the pattern of bias. The lead of CONT diminishes when ρ_1 grows larger and for values of $\rho_1 \geq 0.5$, SEM beats CONT in most cases. Moreover, for the same range of values of ρ_1, SEM even outperforms TRUE in more than half of the cases.

For samples generated with spillover from hotspot and from other regions according to inverse distance, the biases of β are summarized in Table 3. Comparison of this table and Table 1 shows that in the present case both approaches perform worse. It also shows that SEM has lower bias than DINV most of the time. As the value of ρ_1 goes up to 0.9, SEM still remains stable in terms of bias while the estimation results of DINV get extremely biased. Another interesting finding is that SEM outperforms TRUE more often than in the previous case.

Table 4 shows that the RMSEs and bias follow similar patterns in the present case. Moreover, The RMSEs of SEM and DINV tend to diverge more rapidly when ρ_1 increases. Besides, SEM also outperforms TRUE more frequently than in the previous case.

Tables of results for sample sizes 100 and 225 are not presented here due to length limitations. They are available upon request from the first author. The main results are the following. When sample size goes up to 100 and 225, SEM outperforms the classical W-based approaches more in terms of bias, but it does not uniformly outperform them. The comparison in terms of RMSE as a function of the number of observations is very much in line with the bias pattern.

The above analyses of bias and RMSE of the estimator of β show that SEM outperforms the classical W-based models in most cases with more obvious dominance in terms of bias than RMSE. Specifically, it tends to increasingly outperform the classical approach when ρ_1 goes up. These conclusions hold for the correctly specified classical models but even more so for the misspecified models which ignore spillover from hotspots. As far as the type of weights matrix used in sample generation is concerned, both approaches have larger biases and RMSEs for samples generated with spillover from hotspot in combination with inverse distance matrix. Although SEM is also at a disadvantage in the type of samples as it only considers three neighbors in contrast to the correct and much larger number (total sample size) of units that inverse distance matrix model takes into account, DINV produces the most biased results in the majority of cases. SEM makes a winner in terms of stability of bias and RMSE over changing values of the spatial lag parameters, sample sizes and types of weights matrix used for sample generation.

	Hotspot (ρ_1) + Contiguity (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE		*TRUE*		*TRUE*		*TRUE*		*TRUE*		*TRUE*	
	CONT	SEM	CONT	SEM	CONT	SEM	CONT	SEM	CONT	SEM	CONT	SEM
$\rho_1 = 0$	*-0.004*		*-0.003*		*-0.001*		*0.001*		*0.004*		*-0.013*	
	-0.006	-0.004	-0.005	0.005	-0.003	0.006	-0.001	0.015	0.000	-0.004	0.003	0.014
0.1	*-0.004*		*-0.003*		*-0.001*		*0.001*		*0.002*		*-0.023*	
	-0.001	0.000	-0.001	-0.003	-0.001	0.004	-0.002	-0.009	-0.011	-0.025	-0.061	0.051
0.3	*-0.004*		*-0.003*		*-0.002*		*0.000*		*0.002*			
	-0.002	0.000	-0.007	-0.004	-0.020	-0.009	-0.028	-0.017	-0.013	-0.023		
0.5	*-0.004*		*-0.003*		*-0.003*		*0.007*					
	-0.010	0.001	-0.017	-0.003	-0.012	-0.010	-0.015	-0.023				
0.7	*-0.004*		*-0.003*		*0.049*							
	0.002	0.000	0.016	-0.001	0.075	-0.037						
0.9	*0.000*		*-0.004*									
	0.051	-0.001	0.060	-0.030								

Table 1. Bias of the estimator of β for spillover from hotspot and first order queen contiguity neighbors

	Hotspot (ρ_1) + Contiguity (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE		*TRUE*		*TRUE*		*TRUE*		*TRUE*		*TRUE*	
	CONT	SEM	CONT	SEM	CONT	SEM	CONT	SEM	CONT	SEM	CONT	SEM
$\rho_1 = 0$	*0.061*		*0.060*		*0.060*		*0.060*		*0.061*		*0.081*	
	0.060	0.071	0.060	0.075	0.060	0.075	0.060	0.072	0.061	0.072	0.063	0.138
0.1	*0.061*		*0.061*		*0.060*		*0.061*		*0.063*		*0.079*	
	0.061	0.072	0.061	0.071	0.061	0.072	0.061	0.073	0.063	0.081	0.100	0.078
0.3	*0.061*		*0.061*		*0.061*		*0.061*		*0.060*			
	0.061	0.063	0.061	0.067	0.062	0.062	0.062	0.063	0.100	0.068		
0.5	*0.061*		*0.061*		*0.061*		*0.075*					
	0.061	0.065	0.062	0.063	0.060	0.063	0.127	0.064				
0.7	*0.061*		*0.061*		*0.111*							
	0.062	0.061	0.064	0.061	0.166	0.071						
0.9	*0.064*		*0.060*									
	0.112	0.061	0.220	0.069								

Table 2. RMSE of the estimator of β for spillover from hotspot and first order queen contiguity neighbors

	Hotspot (ρ_1) + Inverse-distance (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM
$\rho_1 = 0$	*-0.012*		*-0.011*		*-0.011*		*-0.010*		*-0.009*		*-0.008*	
	-0.012	-0.004	-0.012	-0.014	-0.011	-0.027	-0.010	-0.060	-0.010	-0.116	-0.009	-0.022
0.1	*-0.012*		*-0.011*		*-0.010*		*-0.009*		*-0.009*		*-0.004*	
	0.000	0.000	0.001	-0.005	0.003	-0.023	0.007	-0.065	0.015	-0.143	0.070	-0.009
0.3	*-0.011*		*-0.010*		*-0.010*		*-0.009*		*-0.008*			
	0.046	0.002	0.049	0.001	0.058	-0.032	0.078	-0.095	0.141	-0.180		
0.5	*-0.010*		*-0.009*		*-0.008*		*-0.008*					
	0.124	0.004	0.134	-0.008	0.168	-0.054	0.256	-0.131				
0.7	*-0.008*		*-0.008*		*-0.007*							
	0.257	0.006	0.296	-0.015	0.506	-0.096						
0.9	*-0.006*		*-0.007*									
	2.518	-0.001	3.244	-0.087								

Table 3. Bias of the estimator of β for spillover from hotspot and inverse distance related neighbors

	Hotspot (ρ_1) + Inverse-distance (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM
$\rho_1 = 0$	*0.066*		*0.066*		*0.066*		*0.065*		*0.065*		*0.064*	
	0.066	0.071	0.065	0.077	0.065	0.084	0.065	0.097	0.064	0.143	0.064	0.127
0.1	*0.066*		*0.066*		*0.066*		*0.065*		*0.065*		*0.063*	
	0.065	0.068	0.065	0.070	0.065	0.079	0.065	0.093	0.065	0.166	0.088	0.144
0.3	*0.066*		*0.066*		*0.066*		*0.065*		*0.065*			
	0.077	0.067	0.078	0.067	0.083	0.067	0.094	0.107	0.144	0.184		
0.5	*0.066*		*0.066*		*0.065*		*0.065*					
	0.130	0.064	0.138	0.062	0.170	0.074	0.256	0.133				
0.7	*0.065*		*0.065*		*0.064*							
	0.258	0.061	0.296	0.061	0.506	0.102						
0.9	*0.064*		*0.063*									
	2.518	0.060	3.244	0.097								

Table 4. RMSE of the estimator of β for spillover from hotspot and inverse distance related neighbors

5. CONCLUSIONS

The paper evaluates by means of Monte Carlo simulations the estimator of the regression coefficient obtained by the classical W-based spatial autoregressive model and the structural equations model with latent variables (SEM) on the basis of data sets that contain two types of spatial dependence: spillover from (i) a hotspot and (iia) first order queen contiguity neighbors or (iib) inverse distance related neighbors. Two types of classical models were considered. SEM takes spatial dependence into account by means of a fixed number of nearest neighbors as well as the dependent variable in the hotspot weighted by inverse distance. The estimation results are analyzed in terms of bias and root mean squared error (RMSE) for different values of the spatial lag parameters, specifications of weights matrices and sample sizes.

The simulation results show that both approaches perform better for samples generated with spillover from the hotspot and from first order queen contiguity neighors. Moreover, compared to the misspecified W-based models, SEM frequently has smaller bias and RMSE and even outperforms the correctly specified models in many cases. These trends increase when the spatial lag parameter for spillover increases. The lead of SEM also increases by sample size. Finally, SEM was found to be more stable in terms of both bias and RMSE over various dimensions.

Finally, note that in the case of SEM not all model search options were exploited. Specifically, the number of observed spatially lagged variables was a priori fixed whereas it offers ample opportunities to search and test the optimal number of observed indicators (see Folmer and Oud, 2008). Another option of SEM that was not exploited was the use of several latent variables to take spatial dependence into account (Folmer and Oud, 2008). Exploitation of this option would have brought SEM closer to the correctly specified model. Full exploitation of all its model search options might improve the performance of SEM in comparison with the classical W-based approaches.

References

Anselin, L., 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Florax, R.J.G.M, Folmer, H. and Rey, S.J., 2003. Specification searches in spatial econometrics: the relevance of Hendry's methodology. *Regional Science and Urban Economics, 33, 5,* 557-579.

Florax, R. and Folmer, H., 1992. Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional Science and Urban Economics, 22,* 405–432.

Folmer, H. and Oud, J., 2008. How to get rid of W? A latent variables approach to modeling spatially lagged variables. *Environment and Planning A, 40,* 2526-2538.

Jöreskog K.G. and Sörbom, D., 1996. *Lisrel 8: User's Reference Guide.* Scientific Software International, Chicago, IL.

LeSage, J and Pace, R. K., 2009. *Introduction to Spatial Econometrics*. Chapman & Hall/CRC.

Neale M.C., Boker S.M., Xie G., Maes H.H., 2003. *Mx: Statistical Modeling.* VCU Box 900126, Richmond, VA 23298: Department of Psychiatry. 6th Edition.

Oud, J. and Folmer, H., 2008. A structural equation approach to models with spatial dependence. *Geographical Analysis, 40,* 152-166.

GENERALIZATION OF TILED MODELS WITH CURVED SURFACES USING TYPIFICATION

Richard Guercke [a], Junqiao Zhao [b], Claus Brenner [a] and Qing Zhu [b]

[a] Leibniz Universität Hannover, Institute of Cartography and Geoinformatics , Appelstr. 9a, 30167 Hannover, Germany – (richard.guercke, claus.brenner)@ikg.uni-hannover.de
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, P.R. China – johnzjq@gmail.com, zhuqing@lmars.whu.edu.cn

KEY WORDS: Multi-Scale Representation of Spatial Data, Generalization, Building Models, Tiles, Roof, Curved

ABSTRACT:

Especially for landmark buildings or in the context of cultural heritage documentation, highly detailed digital models are being created in many places. In some of these models, surfaces are represented by tiles which are individually modeled as solid shapes. In many applications, the high complexity of these models has to be reduced for more x efficient visualization and analysis. In our paper, we introduce an approach to derive versions at different scales from such a model through the generalization method of typification that works for curved underlying surfaces. Using the example of tiles placed on a curved roof – which occur, for example, very frequently in ancient Chinese architecture, the original set of tiles is replaced by fewer but bigger tiles while keeping a similar appearance. In the first step, the distribution of the central points of the tiles is approximated by a spline surface. This is necessary because curved roof surfaces cannot be approximated by planes at large scales. After that, the new set of tiles with less rows and/or columns is distributed along a spline surface generated from a morphing of the original surface towards a plane. The degree of morphing is dependent on the desired target scale. If the surface can be represented as a plane at the given resolution, the tiles may be converted to a bump map or a simple texture for visualization. In the final part, a perception-based method using CSF (contrast sensitivity function) is introduced to determine an appropriate LoD (level of detail) version of the model for a given viewing scenario (point of view and camera properties) at runtime.

1. INTRODUCTION

1.1 Generalization

The research behind the results in this paper was done in the context of the generalization of building models. Especially roof and wall surfaces may be modeled by individual tiles where "tiles" refers not only to roof tiles but also to bricks in walls or any other basic unit of which a surface is constructed.

In the context of the generalization of building models, our approach can be used to produce different LoD models from an original model. The typification procedure described in this paper can be used in large-scale visualization if the representation of the surface by a plane with the tiles as textures and / or bump maps is too coarse, and the number of tiles and their geometric complexity make the rendering of all tiles an expensive and unnecessary operation.

The basic idea of the approach is to interpret the distribution of the tiles in the original model as a sampling of a surface. This surface is approximated by another surface – in our experiments, we used a Bézier spline surface. According to the desired resolution or the viewing scenario, this surface may be simplified – we implemented, for example, a linear transition from the original surface to a plane. On this new surface, enlarged and possibly geometrically simplified instances of the tile model are distributed along paths derived from an interpolation of the traces of the rows and columns of the original model on the original surface.

In order to optimally select these simplified models during real-time visualization, the contrast sensitivity function (CSF) is employed as a criterion. We evaluate the spatial frequency and the contrast of tiles to obtain the perceptibility of the tile pattern. Then the automatic LoD management is achieved.

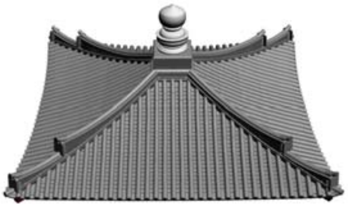


Figure 1. One tiled roof in Chi Lin Nunnery

In order to evaluate if our approach can provide suitable results for real-world data, we used a detailed model of a real roof surface from an ancient Chinese temple of the Chi Lin nunnery in Hong Kong (see Figure 1) for our experiments.

1.2 Related Work

Building generalization often models roof surfaces as planes (such as in [Kada, 2007]), and surfaces composed of individual tiles are not an issue because there are only very few models of such fine granularity. [Buchholz, 2006] describes an approach to procedurally derive textures for the visualization of 3D city models at different levels of detail. The issue of individually modeled tiles is, however, not addressed there either.

Human perception is introduced into polygon reduction in order to allocate more geometry to visually more important places [O'Sullivan et al., 2004]. The proposed methods can be divided into two kinds: one is implemented in model space and the other is in screen space. The former is usually by evaluating the perception of geometry using projected errors or curvature [Luebke et al., 2003; Winkler, 2000]. The latter evaluates the perception of an image on the screen by introducing more rigorous human vision system (HVS) models derived from research on image processing [Winkler, 2000]. However, most existing methods treat the perception information as weights to adjust the sequence of simplification operations such as edge collapse etc. [Qu and Meyer, 2008], which is not suitable for the aggregation of details. A new perceptually driven primitive location method was proposed to overcome this shortage by introducing the top-down constrain [Du et al., 2008]. However, following the idea of the previously proposed generalization framework [Guercke and Brenner, 2009], proper generalization methods for model parts with specific semantics are needed, such as the tiled roof model.

2. MODELING AND FEATURE EXTRACTION ISSUES

2.1 Introduction: A Pragmatic Approach

Generalization and feature extraction are closely related because generalization procedures can greatly benefit from or depend on semantic knowledge that may not be present in the data set to be processed.

As the general problem of feature extraction is in itself a wide field for research, we used heuristic approaches to extract the necessary semantic information for the generalization process. For many applications, these heuristics may work as well – perhaps with some small adaptations.

We tested our approach using the model of a roof surface from the Chi Lin Nunnery in Hong Kong [Li et al., 2006]. Especially in the feature extraction step, we were content with getting satisfying results for this data set. For other surfaces, more sophisticated methods may be necessary.

2.2 The Tile Model

In order to be suitable for the approach that is presented in this paper, the model should contain only a limited number of structurally different types of tiles with a clearly defined relationship.

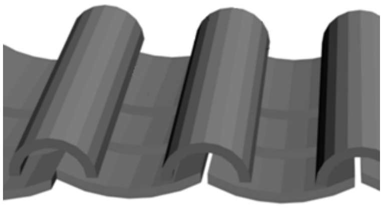


Figure 2. Placement of the upper and lower tiles

In the case of the roof, there are two different types of tiles: one for the upper (Tong Wa) and one for the lower layer (Ban Wa). Figure **2** illustrates the layout of the upper and lower tiles.

For each type of tile, a template tile is stored. This template tile stores the geometry of the tile in tile coordinates. The center of (the bounding box of) the tile is supposed to be the origin of the tile coordinate system. The x and y direction define the main directions of the tile, the z axis is the "normal" of the tile.

The distribution of the tiles is modeled by virtual tile objects that store a reference to the template tile, the position of the tile's center, the direction of its local coordinate axes in world coordinates, and scaling factors of the tile in its local x, y, and z directions. Using these parameters, the geometry of the tiles can be obtained by a simple linear transformation of the vertices in the template model.

2.3 Approximation of the Underlying Surface

Especially in the case of curved surfaces, the exact underlying shape is often the result of a combination of planning and aging. Therefore, there is usually no analytic model for the shape. For this reason, we approximate it by a Bezier spline surface because a representation based on control points is useful for the generalization process (for the simplification of the underlying shape described in section 3.2).

This surface was obtained by a least squares adjustment procedure minimizing the vertical distance between the centers of the tiles and the approximating spline surface. In order to reduce the complexity of the adjustment procedure, we used the fact that the surface could be described as a height function z=f(x,y). This way, only the z values for the control points had to be estimated from the set of the central points of the tiles.

In our approach, we used C^2 continuous Bezier surfaces for the approximation to ensure smooth transitions between the patches. The borders between the patches that form the surface are aligned with the coordinate axes for simplicity.

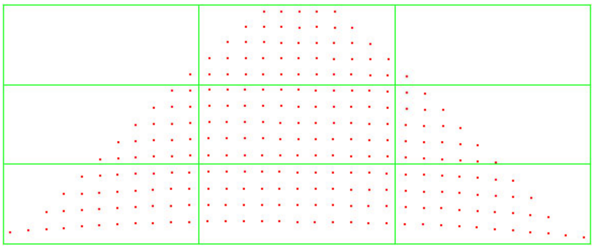


Figure 3. Boundaries for 3x3 Patches

The boundaries of the patches were chosen to be spaced equally in the range defined by the bounding box of the tiles' centers. In the adjustment process, the sum of the squared vertical distances of the tiles' centers to the surface was minimized.

Figure **3** shows the patch boundaries for three by three patches.

Because both versions yielded good results with maximum vertical distances of all tiles to the surface of less than one centimeter, we used only 2x1 patches for our experiments.

2.4 Identifying Rows and Columns

The generalization algorithm that is described in this paper relies on the tiles' being arranged in columns and rows. When the model is created, it is no problem to mark or organize the tiles in such a way that the rows and columns are marked explicitly.

For our sample data set, we had to detect the rows and columns from the distribution of the tiles. A general solution to this problem is a rather complex feature extraction task and beyond the scope of this paper.

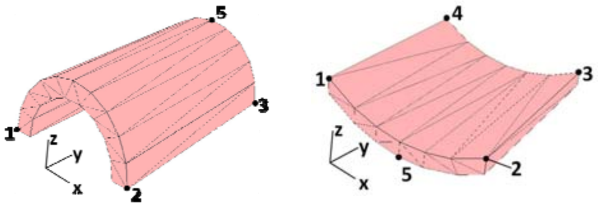


Figure 4: The upper (Tong Wa) and lower tiles (Ban Wa)

Figure 4 illustrates the distribution of characteristic vertices on the shapes of the different tiles. The characteristic vertices 1-5 in the given model could be determined by their distances to the center and their z-value. Using the characteristic points, the orientation of the tiles could be reconstructed. Since the rows and columns of the roof surface in our test data set were roughly aligned with the local x and y directions of the tiles, they could easily be identified once the orientation of the tiles had been established.

3. GENERATION OF LOD MODELS

3.1 Overview

The basic idea of the algorithm is to use a surface that approximates the original distribution of the tiles to place the tiles in the generalized model.

Depending on the desired resolution, fewer enlarged – and possibly geometrically simplified – instances of the tile models are distributed on an underlying surface which may be the original surface or – at lower resolutions – a simplified version of it. In our experiments, we used a surface resulting from a morphing of the original surface towards a plane.

In its current version, the algorithm uses the first (base) row to determine the number of columns, assuming that this is the widest row, containing the maximum number of tiles (columns).

We did not incorporate a simplification of the basic shape of the tiles into our approach. This would, however, offer considerable potential in a real application because the tile models may be quite complex – the lower tiles in our experiment had 28 vertices and the upper ones had 42 vertices which definitely means that there is potential for simplification at smaller scales.

The layers containing the different kinds of tiles were treated separately in our experiment – only the placing of the columns of the upper tiles was adapted to make the upper tiles cover the seams between the lower tiles properly.

3.2 Simplification of the Underlying Surface: Morph towards a Plane

The adjustment of the underlying surface serves several purposes. The first is a reduction of the geometric complexity of the underlying shape. Additionally, it offers a smooth transition to smaller scale models in which the tiles are modeled as bump maps or textures on a plane.

For our test case, we used a plane through the tiles on the ridge and the corner tiles in the base row. Although this plane is lying almost completely above the original shape, it has the advantage that the different roof surfaces are going to form a quite regular hipped roof when they are combined.

The control points of the spline are moved towards the plane in a linear way:

$$c_{i,new} = c_i + t \cdot (p(c_i) - c_i)$$

where c_i is the position of the i^th control point of the original spline, $p(c_i)$ the vertical projection of c_i on the target plane, and $c_{i,new}$ the position of the i^th control point in the spline representing the new underlying surface – for t=0, the surface is unchanged, for t=1, the surface is identical to the plane. According to the purpose of the generalization, appropriate values of t for have to chosen for the different LoD models.

In the following sections, the term “underlying surface” will refer to the morphed surface.

3.3 Building the Base Row

In a first step, a Bezier curve y=b(x) approximating the distribution of the centers of the original tiles in the first row is calculated.

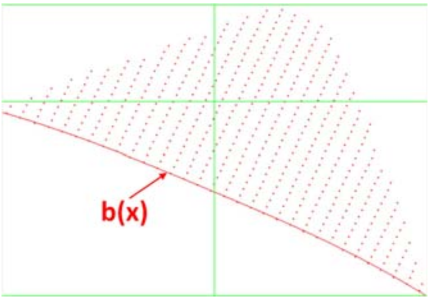


Figure 5. Trace b(x) of the base row in the (x,y)-plane

Figure 5 shows the distribution of the original tiles in the (x,y)-plane and the approximation spline curve b(x) through the centers of the tiles in the first row.

The restriction that b should be a function of x is introduced by the approximation process rather than the generalization algorithm. In fact, any function defining a surface curve B(t) on the underlying surface s=s(x,y) can be used. This function B may be defined directly or through function c: (x,y)=c(t) with B(t) = s(c(t)). In our example, the control parameter t is the x-value, so we can set

$$B(t) = B(x) = s \cdot \binom{x}{b(x)}.$$

The distribution of the tiles is done in parameter space (on the x axis). In the current version, the x values for the tiles t_i were evenly between the positions of the first and last tiles in the original first row: All tiles have the same distance of Δx in x direction with

$$\Delta x = \Delta x_{ori} \cdot \frac{k}{k_{ori}}$$

where k_{ori} is the original number of columns, k is the desired number of columns, and Δx_{ori} the average distance between the original tiles on the x axis. For the roof model in our testing scenario, this yielded acceptable results but in general, it is a problematic approach because it does not take the curvature of the surface into account.

A better solution is to use the actual distances on the surface for the distribution of the tiles. In the first step, an initial scaling factor f=k/ k_{ori} is defined. The first tile is placed in a starting position. The parameter t_0 of this first tile has to be known. The initial position of the next tile is determined as

$$\vec{x}_{i,step=0} = B(t_{i-1} + f \cdot \Delta t_{ori})$$

where t_{i-1} is the parameter of the previous tile and Δt_{ori} the average difference of the parameters for neighboring tiles in the base row of the original model. $t_{i,step=0} = f \cdot \Delta t_{ori}$ is the initial value of the curve parameter *t* for the current tile *i*.

After that, the error is calculated:

$$e = |\vec{x}_{i,step} - \vec{x}_{i-1}| - d_{tile}$$

with d_{tile} being the average distance between tile centers in the base row scaled by f, $\vec{x}_{i-1}$ the position of the preceding tile, and $\vec{x}_{i,step}$ the current position of the current tile. If |e| is smaller than a threshold value (depending on the scale), the tile position can be accepted. If e > 0, then the actual distance between the two tiles is too large. In this case, $t_{i,step}$ has to be reduced by some amount. If e < 0, then the distance between the tiles is too small and $t_{i,step}$ has to be increased.

The process terminates when either *k* tiles have been distributed or if a tile would have to be placed outside of a given boundary. A simple boundary condition might be that the center of the last tile should not be beyond the center of the last tile in the original base row – with some buffer to allow for the increased size of the tiles.

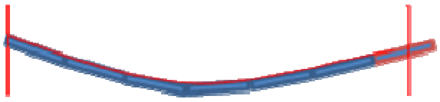


Figure 6. Tiles in the base row are too large

It may happen that due to the curvature of the underlying surface, it is not possible to place k tiles in the base row or that after placing k tiles, there is a gap to the intended end of the row. In the first case, the tiles are too big and the process for building the base can be repeated with a decreased scaling factor f for the tiles. If there is a gap, the tiles are too small, and the process is repeated with an increased scaling factor f. In Figure 6, the last tile could not be placed because it would fall outside the range specified for the tiles. The rectangles represent the tiles, and the bars on the left and right side are the boundaries beyond which no tile should extend.

3.4 Distribution of the Tiles on the Surface

The distribution of the tiles along the columns is done in a similar way to the building of the base row: the first tile of column i is the ith tile in the base row, and the tiles are distributed along the surface curve defined by the trace of the column. The scaling factor f is taken from the previous step.

In our sample data set, all columns are parallel in the (x,y) plane. In the more general case, one can approximate the traces of the columns by spline functions and perform an interpolation of the parameters to get the traces of the new columns. The end of the columns can be established by interpolating the ends of the original columns.
The main directions of the tiles were aligned with the direction of the columns: The local y axis of the tiles pointed in the direction of the traces of the columns. This property was preserved in the distribution of the new tiles in the LoD models:

$$\vec{y} = \vec{c}_{i+1} - \vec{c}_i + \vec{v}_{off}$$

where y is the direction of the tile's local y axis in world coordinates, c_{i+1}-c_i the vector from the center of the current tile to the center of the next one, and v_{off} an offset vector that is introduced to model the tilting of the tiles to avoid overlaps. The x vector can be obtained as the cross product of y and the global up vector $(0,0,1)^T$. The normal z of the tile is defined as the cross product of x and y. All local direction vectors of the tile are normalized to make them form an orthonormal basis.

In our experiments, we applied the scaling factor f only along the local x and y directions of the tiles to support a smooth transition towards a plane. If the height of the tiles is scaled as well, then v_{off} will also have to be scaled.

4. PERCEPTION-BASED SELECTION OF AN APPROPRIATE LOD MODEL

Using the established generalization method above, a number of LoD models for a tiled roof can be generated automatically. However, how to optimally select the right model for real time visualization is still an unsolved question. It's understandable to leave it to the user to set the switch distances for each LoD. But this trial and error process is obviously too time-consuming. Moreover, the improper selection of LoD would cause heavy popping effects which decrease the efficiency of transmitting both apparent and semantic information of such kind of models.

A number of perceptually driven methods had been proposed in recent decades [Luebke et al., 2003]. Reddy firstly introduced the principle perceptual model into the LoD selection issue [Reddy, 1997]. The spatial frequency of objects was analyzed by image segmentation using rendered images from multiple viewpoints. If the spatial frequency difference of two LoD models is above (or below) the visual acuity, a coarser (or finer) LoD is to be selected. Similar works have been done in [Luebke and Hallen, 2001] and [Cheng et al., 2006]. However, most of the existing methods evaluate the HVS factors using the curvature of vertices or faces, which obviously does not fit the component structured-model i.e. tiled roof or walls made of bricks. Therefore, a new perceptually based LoD selection method is needed for our test model.

As the most important component of HVS, CSF describes the quantified relationship between the visual perception and the factors of spatial frequency and contrast threshold, as illustrated in
Figure **7**. The expression of CSF is as follows:

$$A(\alpha) = 2.6(0.0192 + 0.144\alpha)\exp(-(0.144\alpha)^{1.1})$$

where $A(\alpha)$ is the contrast threshold of spatial frequency α (c/deg). If the current contrast is lower than $A(\alpha)$, the signal is invisible.

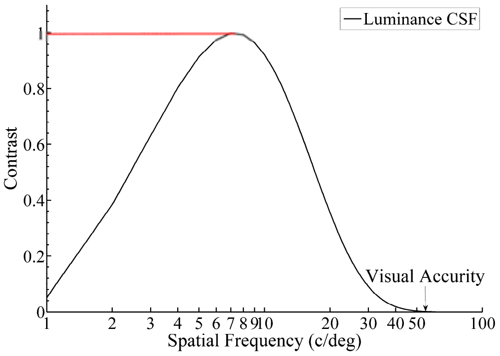


Figure 7. The CSF

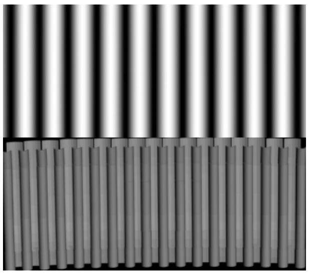


Figure 8. The grating signals (Upper: the theoretically sine wave. Lower: the top view of tiled roof)

Because the upper tiles hide the seams between lower tiles, we can treat the lower tile as the uniform background while the upper tile as the foreground, the represented pattern, as shown in Figure 8, matches the contrast grating signal which is used to evaluate the CSF quite well. Based on this fact, an approximated CSF model is proposed by evaluating the contrast and spatial frequency of roof pattern: For contrast, the luminance of tile depends on the normals of the tile, so it can be represented as:

$$L = k\cos\left(\frac{\pi}{2} - \theta\right) = k\sin\theta$$

Where L is the luminance of a tile, θ is the angle between the normal vector and the direction of light, k is the coefficient of simple illuminate model obeying Lambert's cosine law. Because the shape of the upper tile is a kind of semicircle which would likely to have both the highest luminance as well as the lowest luminace at most of the viewing angles, as illustrated in Figure 9. The Michaelson contrast of tiled roof, defined as (Lmax−Lmin)/(Lmax+ Lmin), can be calculated by:

$$C = \frac{\sin(min(\theta)) - \sin(max(\theta))}{\sin(min(\theta)) + \sin(max(\theta))}$$

where θ is the angle between the direction of light and the normal vectors of the triangles of the upper tile.

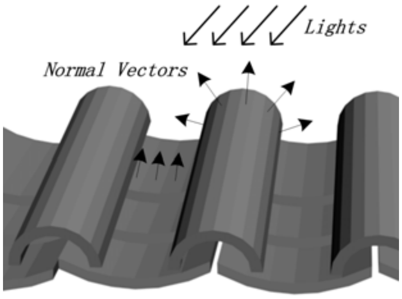


Figure 9. The computing of contrast

For spatial frequency, the projected distance d_{prj} between vertices 5 can be treated as the wave length of the contrast grating signals represented by the upper tiles. Therefore,

$$f = \frac{1}{d_{prj}}$$

where f is the spatial frequency evaluated at d_{prj}, the projected distance between upper tiles on screen. The CSF can then be rewritten as follows:

$$A(d_{prj}) = 2.6\left(0.0192 + 0.144\frac{1}{d_{prj}}\right)exp\left(-\left(0.144\frac{1}{d_{prj}}\right)^{1.1}\right)$$

In order to find the best LoD model for a given viewing scenario, the d_{prj} between two adjacent columns is an important criterion. At a given viewing position during real time visualization, we firstly find the nearest upper tiles in two nearby columns. d_{prj} can then be calculated by:

$$d_{prj} = \frac{Z \cdot {d_1}^2}{\tan(\alpha)} + \frac{Z \cdot {d_2}^2}{\tan(\beta)}$$

where z is the distance from the camera to the nearest clipping plane, d_1 and d_2 are the distances from vertices 5 to the viewing orientation, α , β are the angles between vertices 5 to the viewing orientation, as shown in Figure **10**.

If the realtime evaluated contrast threshold $A(d_{prj})$ is lower than C, a LoD model containing a larger d is to be selected. It is to be noted that if other factors like velocity as well as eccentricity are to be taken in consideration, a more sophisticated HVS model is needed.

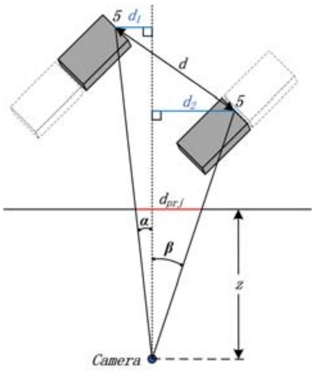


Figure 10. Computing of projected *d*

5. EXAMPLES

Figure 11 shows the original model and detailed views of parts of the roof in different LoDs. The first detail drawing shows a part of the roof from the model with 32 rows – the same number of rows as in the original model. One can see that the tiles fit neatly without noticeable gaps or overlaps.

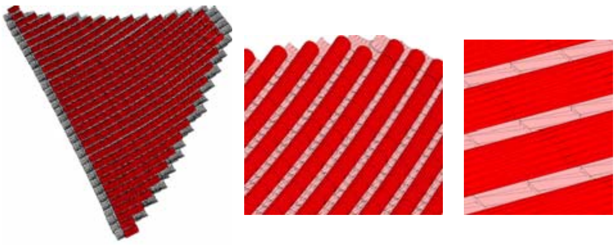


Figure 11. The original roof surface and details from different LoD models

There is some overlap between the upper tiles because the tiles were distributed along the columns without compensating for the loss of length in the columns resulting from the transition towards the plane. The effect could have been avoided had the iterative approach described in section 3.3 been used. In this case, the effect is only visible in the wireframe model from close range, so it can be accepted for visualization purposes.

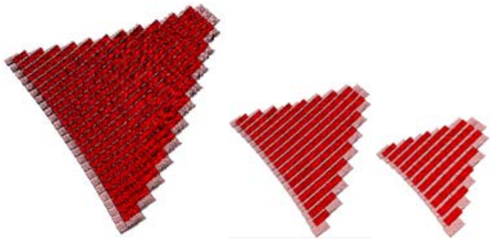


Figure 12. LoD models with 32, 25 and 12 tiles in the base row

Figure 12 shows LoD models for 32, 25, and 12 tiles in the base row. The corresponding values for d (see section 4) are 27, 34, and 76 centimeters.

6. CONCLUSION AND OUTLOOK

We have presented ideas for the generalization of tiled surface models and their application to the model of a roof surface of an ancient Chinese temple.

Though more research is definitely necessary to get a precise understanding of the potential and limitations of our approach, the results from the real-world data set are promising.

The main purpose of the experiments described in this paper was to test if our ideas are suitable for real-world situations. Especially in the context of generalization for visualization, we think that we could produce promising results.

Concerning the iterative algorithm for the distribution of tiles along a surface curve, a closer inspection of the potential and limitations is necessary.

At smaller scales, using simplified tile models can reduce the complexity of the model considerably without causing a noticeable loss of quality. Because the tiles are enlarged in the process, the selection of an appropriate simplified tile model has to be done with care. In our experiments, we did not introduce simplified tile models.

In this paper, we described how Contrast Sensitivity Function (CSF) could be used to select an appropriate LoD from a set of such models that were derived in advance. Recently, a novel perceptually based method for planning the discrete LoD for complex model façade was introduced [Zhu et al., 2009]. By using the idea, the visual model could also be employed in the process of building the LoD models to select values for the morphing parameter t depending on the desired number of columns k and the original surface.

For applications beyond visualization, techniques for the detection and avoidance of gaps or overlaps between the tiles may be necessary. Especially when different layers or neighboring surfaces have to be aligned, these problems can cause irritating effects.

ACKNOWLEDGEMENTS

The research for this paper was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the GDI-Grid project and by the National Basic Reseach Program of China (No. 2010CB731800) as well as the National High Technology Research and Development Program of China (No. 2008AA121600).

REFERENCES

Buchholz, H., 2006. Real-time Visualization of 3D City Models. PhD thesis, Universität Potsdam.

Cheng, I., Shen, R., Yang, Y. and Boulanger, P., 2006. Perceptual Analysis of Level-of-Detail: The JND Approach, Proceedings of the Eighth IEEE International Symposium on Multimedia: 533–40.

Du, Z.Q., Zhu, Q. and Zhao, J.Q., 2008. Percetion-driven simplification methodology of 3D complex building models, ISPRS2008, Beijing.

Guercke, R. and Brenner, C., 2009. A Framework for the Generalization of 3D City Models, Proceedings of 12th AGILE Conference on GIScience.

Kada, M., 2007. 3D Building Generalisation by Roof Simplification and Typification, Proceedings of the 23rd ICC.

Lee, C.H., Varshney, A. and Jacobs, D.W., 2005. Mesh saliency. ACM Transactions on Graphics, 24(3): 659-666.

Li, D.R., Zhu, Y.X., Du, Z.Q. and Hong, T., 2006. Virtual Tang-Style Timber-frame building complex. Lecture Notes In Computer Science, 4282:880-888.

Luebke, D. and Hallen, B., 2001. Perceptually Driven Simplification for Interactive Rendering, Proceedings of the Eurographics Workshop. Rendering Techniques 2001. Springer, London, United Kingdom, pp. 223-234.

Luebke, D. et al., 2003. Level of Detail for 3D Graphics. Morgan Kaufmann, San Francisco.

O'Sullivan, C., Howlett, S., McDonnell, R., Morvan, Y. and O'Conor, K., 2004. Perceptually adaptive graphics. Eurographics State of the Art Reports: 141-164.

Qu, L. and Meyer, G.W., 2008. Perceptually guided polygon reduction. IEEE Transactions on Visualization and Computer Graphics, 14(5): 1015-1029.

Reddy, M., 1997. Perceptually Modulated Level of Detail for Virtual Environments, University of Edinburgh, Edinburgh, United Kingdom.

Winkler, S., 2000. Vision models and quality metrics for image processing applications, Swiss Federal Inst. Technol, Lausanne, Switzerland

Zhu, Q., Zhao, J.Q., Du, Z.Q. and Zhang, Y.T., 2009. Quantitative analysis of discrete 3D geometrical detail levels based on perceptual metric. Computers and Graphics, 34(1):55-65.

AN OPTIMISED CELLULAR AUTOMATA MODEL BASED ON ADAPTIVE GENETIC ALGORITHM FOR URBAN GROWTH SIMULATION

Yan Liu [a], Yongjiu Feng [b],
[a] School of Geography, Planning and Environmental Management, The University of Queensland
yan.liu@uq.edu.au
[b] College of Marine Sciences, Shanghai Ocean University, 999 Huchenghuan Road, Shanghai, P.R. China,
yjfeng@shou.edu.cn

KEY WORDS: Cellular automata, Adaptive genetic algorithm, Model optimization, Urban growth

ABSTRACT:

This paper presents an improved cellular automata (CA) model optimized using an adaptive genetic algorithm (AGA) to simulate the spatio-temporal process of urban growth. The AGA technique can be used to optimize the transition rules of the CA model defined through conventional methods such as logistic regression approach, resulting in higher simulation efficiency and improved results. Application of the AGA-CA model in Shanghai's Jiading District, Eastern China demonstrates that the model was able to generate reasonable representation of urban growth even with limited input data in defining its transition rules. The research shows that AGA technique can be integrated within a conventional CA based urban simulation model to improve human understanding on urban dynamics.

1. INTRODUCTION

There is a long-standing interest in understanding urban dynamics though development and application of geographical models (Batty and Xie 1994; Batty, Xie, and Sun 1999; Couclelis 1997; He, *et al.* 2006; Li and Yeh 2002a, b; Liu and Phinn 2003; Muzy *et al.* 2008; Wu 1998, 2002). Compared to many modelling approaches that were developed based on exclusive use of certain mathematical formula, models based on cellular automata (CA) have strong power to capture the non-linear, spatial and stochastic processes of urban growth in more realistic ways (Liu 2008; Stevens, Dragicevic and Rothley 2007; White and Engelen 1993).

Conventionally, CA based urban models require strict definition of various spatial variables and parameters representing different spatial and non-spatial factors driving the development of urban growth (Li, He and Liu 2009). Many CA models have been developed using a diverse range of methods to define such variables and parameters; these methods include multi-criteria evaluation, logistic regression, principal component analysis, and partial least squares regression methods, to name a few. However, limitations of such methods in defining suitable transition rules, or the values of relevant parameters of the transition rules, or in constructing the architecture of the models have been identified and reported in the literature (Al-kheder, Wang and Shan 2008; Li and Yeh 2002a). As a result, there are significant differences between the simulation results and the actual patterns of urban growth, making such models less effective in simulating the actual process of urban growth (Li and Yeh 2002b; Liu and Phinn 2003).

The development of genetic algorithm (GA) and adaptive genetic algorithm (AGA) methods have provided researchers with new ways to identify and search for suitable transition rules and their defining parameters in urban modelling (Bies, *et al.* 2006; Srinivas and Patnaik 1994). This method has been used in satellite imagery classification (Huang *et al.* 2007), site selection (Li, He and Liu 2009), and problem clustering (Lorena and Furtado 2001).

This paper presents a method applying an adaptive genetic algorithm to define and search for transition rules and parameters of a cellular automata model to simulate the spatio-temporal processes of urban growth. The following section presents a generic CA model based on logistic regression method first, followed by the adaptive genetic algorithm method to optimize CA parameters based on minimizing differences between the simulated results and the actual urban development. Section 3 applies the AGA-CA model to simulate the urban growth of Shanghai's Jiading District, Eastern China. Results from the model are also presented and discussed in this section, followed by conclusions in the last section.

2. THE ADAPTIVE GENETIC ALGORITHM BASED CA MODEL (AGA-CA)

2.1 A generic CA model based on logistic regression

Generally, CA defines the state of a cell at one time as a function of the state of the cell and its neighbourhoods at a previous time in accordance with a set of transition rules, which can be generalized as follows (Wu 1998):

$$S_{ij}^{t+1} = f(S_{ij}^{t+1}, \Omega_{ij}, con()) \qquad (1)$$

where ${S_{ij}}^{t}$ and ${S_{ij}}^{t+1}$ represent the states of a cell at location *ij* at time *t* and *t*+1 respectively; Ω_{ij} is a neighbourhood function; $con()$ defines a set of constraints or factors affecting the transition of cell states; and *f* is the transition function.

Assume that a cell can take one of only two states, urban or non-urban. A square neighbourhood is defined with u×u cells and all cells within the neighbourhood have equal opportunity for development. Thus, the probability a cell changes its state from non-urban to urban can be defined as:

$$P_{\Omega} = \frac{\sum_{u \times u} S_{ij} = Urban}{u \times u - 1} \qquad (2)$$

where P_Ω is the probability a cell can change from one state to another; $\sum_{u\times u} S_{ij} = Urban$ represents the total number of urban cells within the u×u cells neighbourhood.

However, in practice, not all cells have equal opportunity for development. For instance, some non-urban areas such as large-scale water bodies or areas with critical physical constraints such as very steep slope may not be able to develop into urban areas. Other areas such as the primary farmland may be prevented from urban development through institutional control, i.e., land use planning regulation.

In order to represent the unequal opportunity of cells for urban development, a stochastic factor can be introduced into the CA based urban models (Wu 2002). With a stochastic control factor, the probability a cell converts from non-urban to urban state can be defined as:

$$\begin{aligned} P_G^t &= \frac{1}{1+\exp(-(a_1x_1+\ldots+a_mx_m))}\times P_\Omega \\ &\times con(cell_{ij}^t = suitable)\times(1+(-\ln\gamma)^\beta) \end{aligned} \tag{3}$$

where

P_G^t is the probability of the cell converting from one state to another at time *t;*

$con(cell_{ij}^t = suitable)$ is a constraint function, the value of which ranges from 0 to 1, with 0 meaning the cell is constrained from changing its current state, and 1 meaning the cell is able to change its state at the following time step;

$1+(-\ln\gamma)^\beta$ represents a stochastic factor, where γ is a random real number ranging from 0 to 1, and β is a parameter controlling the effect of the stochastic factor. The value of β ranges from 0 to 10;

x_i (i = 1, 2, … , m) are various spatial driving factors to urban growth, which can be represented by the distances from a cell to urban centres, town centres, main roads, and so on. These distance factors are also called spatial variables or independent variables; and

a_1 , a_2 ,…, a_m are used to assign different weight to each of the distance variables.

One of the common challenges in developing a logistic regression based CA model is how to choose the distance variables and configure the relevant parameters defining the impact of such distance factors on urban growth. Consequently, results generated by a logistic based CA model may show poor match for the actual patterns of urban growth. This indicates that there is a need to search for other techniques in identifying and defining the model's transition rules. Such a challenge in model development can be addressed by incorporating the adaptive genetic algorithm approach to randomly search for an optimized conversion probability for each cell, and subsequently minimize the differences between simulated results and actual urban growth patterns.

2.2 Adaptive Genetic Algorithm (AGA) based CA Modelling

A genetic algorithm (GA) is a search technique used in computing to find solutions to optimization and search problems. Inspired by evolutionary biology, genetic algorithms work in computer simulations to search for an exact or approximate solution from a population of solutions (Liao *et al.* 2008). This search and optimization process is achieved according to natural selection, including inheritance, selection, crossover and mutation.

There are two elements of a genetic algorithm, including a genetic representation of the solution domain, such as an array of cells in a cellular urban space, and a fitness function to evaluate and quantify the optimality of a solution.

The efficiency of a standard GA depends largely on the setting of its parameters such as the selection, crossover and mutation rates, which are difficult to adjust manually. Such difficulties can be overcome by the adaptive genetic algorithm (AGA) as the AGA could dynamically modify the parameters of the genetic algorithm (Espinoza, Minsker and Goldberg 2001; Kee Airey and Cye 2001). AGAs not only keep population diversity effectively but also improve the performance of local and premature convergences. Such genetic diversity is important to ensure the existence of all possible solutions in the solution domain and the identification of optimized solution. In addition, the adaptive genetic algorithm also enhances the search speed and precision of the genetic algorithm. Hence, the searching and optimization process for problem solutions can be accelerated.

2.2.1 Fitness function: A fitness function is an objective function to quantify the optimality of a solution. This function was created by selecting sample of cells within the cellular urban space to minimise the differences between the simulation results produced by a logistic regression based CA model and the actual urban growth patterns identified from remotely sensed images. The fitness function is defined and optimised through the modelling process as:

$$f(x) = \sum_{i=1}^{n}(f_i(x_1,x_2,\ldots,x_m) - f_i^0(x_1,x_2,\ldots,x_m))^2 \tag{4}$$

where

$f(x)$ is a fitness function;

n is the number of samples selected from the cell space which were used to retrieve the CA transition rules;

f_i is the conversion probability of the state of cell *i* based on the logistic regression model, i.e., $f_i = P_G$ as defined in Equation (3); and

f_i^0 is the actual conversion decision of cell *i*. f_i^0 can only take one of the two values, 0 or 1, with $f_i^0 = 0$ meaning the state of the cell *i* remains as non-urban and $f_i^0 = 1$ meaning the state of the cell has changed from non-urban to urban.

The process of urban growth can be affected by many factors, including socio-economic, physical and environmental, as well as institutional control factors. These factors can be built into the cellular automata model through a set of transition rules. With the fitness function, the simulation process of urban growth can be calibrated by dynamically updating the various parameters of the transition rules to minimise the value of the fitness function so the simulated urban patterns can better match with observed patterns of urban growth. The model calibration process is completed once the fitness function reaches a stable value over time and the model's transition rules and parameters can be considered suitable for operation to the whole cell space.

2.2.2 Coding of chromosomes: Chromosomes are the abstract representations of candidate solutions, which can also be called individuals. A chromosome is a set of parameters which define a proposed solution to the problem that the GA is trying to solve. In the CA based urban modeling practice, all possible CA transition rules factors affecting urban growth are considered as chromosomes. Each chromosome is coded as a simple string like:

$$C = [\alpha_1^k, \alpha_2^k, ..., \alpha_m^k] \tag{5}$$

where C represents a string of candidate solutions; m is the number of spatial driving factors (as in Equation (3)); k means the k^{th} individual (candidate solution). α_1^k to α_m^k represent the weight of each spatial driving factor in the k^{th} candidate solution. In fact, the values α_1^k to α_m^k in the optimized candidate solution are the parameter values required by the CA model as defined in Equation (3).

Initially a number of chromosomes were randomly generated to form the possible solutions for the adaptive genetic algorithm to begin its searching process. After many generations of selection, crossover and mutation operations, only those chromosomes which acquire lower fitness values will remain, resulting in the emergence of a good chromosome structure.

2.2.3 Selection operator: Selection is the key operation of the AGA method in which individual genomes are chosen from a population of candidate solutions for later breeding, including recombination and crossover. During each successive generation, individual solutions are selected through a fitness-based process, where solutions with lower fitness values are typically more likely to be selected. Using the Hamming distance that measures the minimum number of substitutions required to change one string into the other as a selection criterion, one chromosome is selected from every randomly selected pair of chromosomes on a competitive selection process. The selection process ensures that the diversity of chromosomes is reserved during the selection process.

2.2.4 Crossover operator: Crossover is an exchange of genetic material between homologous individuals for final genetic recombination. While many crossover operators available in genetic algorithm, this research employs the adaptive genetic operator proposed by Srinvias and Patnaik (1994). The crossover probability P_c is used to allow the crossover between chromosomes. This probability value changes continuously with the change of fitness value during the search process. This crossover probability is defined as:

$$P_c = \begin{cases} P_{c1} - \dfrac{(P_{c1} - P_{c2})(f' - f_{avg})}{f_{\max} - f_{avg}} & f' \ge f_{avg} \\ P_{c1} & f' < f_{avg} \end{cases} \tag{6}$$

where

P_{c1} is the maximum crossover probability, the value is 0.95;

P_{c2} is the minimum crossover probability, the value is 0.45;

f' is the fitness value;

$f_{\max}$ is the maximum fitness value; and

f_{avg} is the minimum fitness value.

Mutation operator: In genetic algorithms, mutation is used to maintain genetic diversity from one generation of a population of chromosomes to the next. Similar to the crossover operators, the mutation operator proposed by Srinivas and Patnaik (1994) was adopted in this research. The mutation probability P_m, is defined as:

$$P_m = \begin{cases} P_{m1} - \dfrac{(P_{m1} - P_{m2})(f_{\max} - f_{avg})}{f_{\max} - f_{avg}} & f \ge f_{avg} \\ P_{m1} & f < f_{avg} \end{cases} \tag{7}$$

where

P_{m1} = maximum mutation probability, 0.1

P_{m2} = minimum mutation probability, 0.001

$f_{\max}$ = maximum fitness value

f_{avg} = minimum fitness value

f = fitness value

2.2.5 Threshold of conversion probability: The AGA technique was used to minimize the fitness function $f(x)$ corresponding to the selected spatial samples. With the minimum value of $f(x)$, a set of optimized chromosome can be achieved together with its defining parameters. This leads to the generation of the conversion probability of each cell from non-urban to urban in the urban growth process. Hence, by comparing the conversion probability of a cell at time t with a pre-defined threshold value (Wu 1998; Wu 2002; Li and Yeh 2002b), if the conversion probability of the cell at time t is larger than the pre-defined threshold value the cell will be converted to an urban state at the following step. Otherwise, the state of the cell will remain unchanged.

$$S_{ij}^{t+1} = \begin{cases} Urban, & if\ P_G^t > P_{\text{threshold}} \\ Non\text{-}urban, & if\ P_G^t \le P_{\text{threshold}} \end{cases} \tag{8}$$

3. APPLICATION AND RESULTS

3.1 Study area and data

The proposed AGA-CA model was applied to simulate the urban growth in Shanghai's Jiading District, which is located in the Yangtze River Delta of Eastern China. The study region consists of eleven blocks (towns) with a total area of 463.6 km^2. Rapid urban expansion had occurred in the 1990s due to the fast economic development and population growth. Urban growth of this region from 1989 to 2006 was mapped out using data from various sources, including two Landsat-5 Thematic Mapper (TM) images acquired on August 6, 1989 and April 30, 2006 respectively to obtain spectrum information of the study area. In addition, essential ancillary data include a 1:50,000 topographic map, cadastre and transportation maps which were collected from the local government.

3.2 Model configuration and implementation

While many spatial factors can make an impact on urban development, in practice, not all factors can be quantified into a simulation model, especially when data reflecting such factors are either not available or not accessible. Considering the process of Jiading urban growth historically, urban development in this district is largely related to the distribution of existing urban and towns, accessibility to transport as well as preservation of primary agriculture land. Therefore, five spatial factors were selected, including distance to urban centre (x_1), distance to town centre (x_2), distance to main roads (x_3), distance to cropland (x_4), and distance to orchard field (x_5). The impact of each factor on urban development may be different, hence, different weights were assigned to each of this factors which were represented by a_1, a_2, a_3, a_4, a_5, respectively.

A total of 5000 sample cells were randomly selected from the Landsat TM images for the AGA model to commence the searching process. The distances of each of these samples to the urban centre, town centre, main road, cropland and orchard field were computed in GIS. These distance values were normalized to have a standard value ranging from 0 to 1.

A modelling framework was developed within ESRI's ArcGIS environment based on Microsoft Visual Basic .NET and ArcGIS Engine 9.2 technologies. This modelling framework incorporates the AGA-CA model as well as a number of other CA based modelling approaches. The user-friendly graphical user interface makes ease the sophisticated computation process of the model (Figure 1).

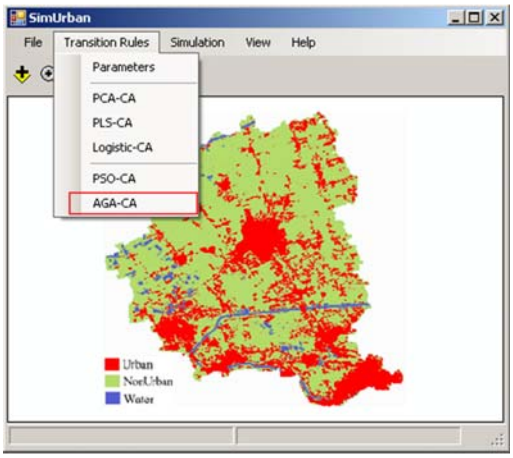


Figure 1. GUI of the modelling framework

Moreover, the strong coupling of the model within a GIS framework makes it possible to use the various display and analysis functions of GIS in raster based data integration and modelling. Hence, this modelling framework becomes an important component of the AGA-CA program for modelling urban growth.

3.3 Results and Discussion

Using the adaptive genetic algorithm proposed in this research, the model was executed to start the search and optimization process with the sample data selected from the 1989 Landsat TM imagery. Figure 2 shows the fitness track in the evolutionary computation of the AGA model, which demonstrates a rapid convergence rate after over 30,000 times of iteration, with a convergence fitness value of 391.9855 (Figure 2).

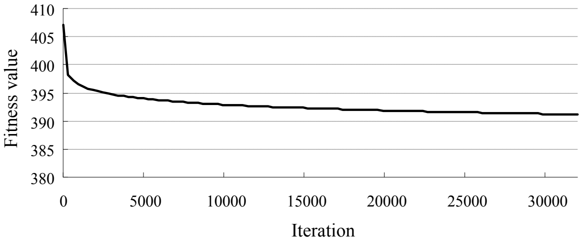


Figure 2. Fitness track of the AGA model

The convergence of the fitness track leads to the identification of a set of optimized chromosome or solution as:

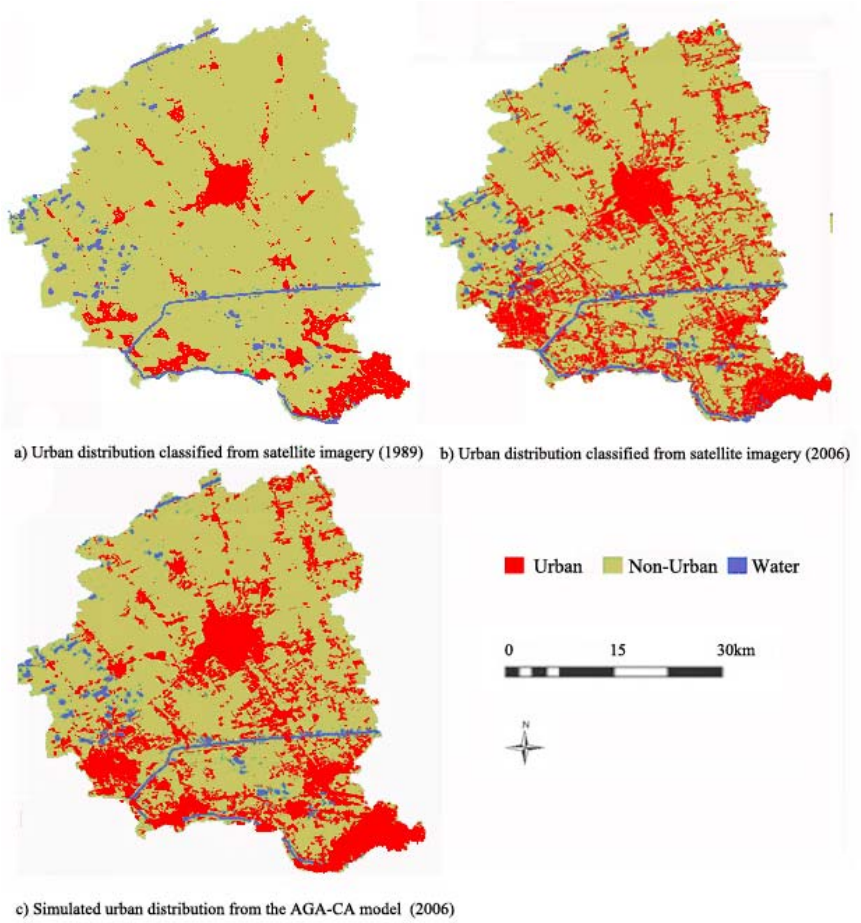
$$C = \begin{cases} a_1 = -0.5083 \\ a_2 = -0.6417 \\ a_3 = -0.4962 \\ a_4 = +0.3723 \\ a_5 = +0.2174 \end{cases} \tag{9}$$

In fact, this optimized solution becomes the initial input data for the CA model to compute the conversion probability of cells from non-urban to urban state. Hence, Equation (3) can be re-written as:

$$\begin{aligned} P_G &= [1 + \exp(-(-0.5083x_1 - 0.6417x_2 \\ &\quad - 0.4962x_3 + 0.3723x_4 + 0.2174x_5))]^{-1} \\ &\quad \times P_\Omega \times con(cell_{ij}^t = suitable) \times (1 + (-\ln\gamma)^\beta) \end{aligned} \tag{10}$$

The optimized chromosome displayed in Equation 9 shows different effect of the spatial factors on urban growth in Shanghai's Jiading District. According to Equation 9, a negative $a_i (i = 1,...,5)$ will lead to a larger P_G value, i.e., a higher possibility for a cell to convert from non-urban to urban state. Likewise, a positive a_i will result in a lower P_G value, hence, a lower possibility for the cell to be converted into an urban state in the next time step. The optimised result generated from the AGA approach shows that the distance to town centres has the most significant impact on the development of cells within its neighbourhood. This is reflected by the smallest value of a_2 ($a_2 = -0.6417$). Likewise, the spatial proximity of a cell to urban centre and main road also positive roles to its urban development (with $a_1 = -0.5083$ and $a_3 = -0.4962$ respectively). On the other hand, factors such as distances to cropland and orchard field have negative impact on urban development (with $a_4 = 0.3723$ and $a_5 = 0.2174$). Hence, the close a cell is to cropland and orchard field, the less opportunity the cell is to be developed into an urban state. This is largely in consistent with the conservation of primary agricultural land policies in practice.

Using the urban distribution pattern defined from the 1989 satellite image classification as the initial input data for the urban CA model (Figure 3a), and the transition rules generated from the AGA approach, the CA model was operated to generate a series of urban scenarios. Each iteration of the model represents one year. After 16 iterations the model generates a map representing urban growth patterns of Jiading District in 2006 (Figure 3c). This simulated urban scenario was compared with the actual urban distribution as defined by classifying the 2006 Landsat TM image (Figure 3b).


a) Urban distribution classified from satellite imagery (1989) b) Urban distribution classified from satellite imagery (2006)

c) Simulated urban distribution from the AGA-CA model (2006)

Figure 3. Actual and simulated urban scenarios of Jiading District using the AGA-CA model

Figure 3 b and c show high similarity even by visual inspection and comparison. By comparing the two maps on a cell-by-cell basis, an error matrix analysis was carried out (Table 1). The results show that the producer's accuracy for non-urban and urban areas were 85.7 and 76.3 per cent respectively, while the user's accuracy for the non-urban and urban categories were 81.1 and 86.7 per cent, respectively. Consequently, the model generated an overall accuracy of 82.7 per cent and a Kappa coefficient of 60.9 per cent. These simulation accuracies are considered good given that only five spatial distance variables were considered in the model. Should other factors such as the social demographic controls, institutional policy effects concerning sustainable urban development as well as other economic constraints included into the model, the AGA-CA model would also be able to can be used to generate and evaluate various urban growth scenarios.

		Simulation Results		
		Non-urban	Urban	Row Total
Satellite-based Land Use Classification	Non-urban	31556	5261	36817
	Urban	4115	13211	17326
	Column Total	38901	15242	54143
		Producer's Accuracy		Omission Error
Non-urban		85.7%		14.3%
Urban		76.3%		23.8%
		User's Accuracy		Commission Error
Non-urban		81.1%		18.9%
Urban		86.7%		13.3%
Overall Accuracy			82.7%	
Kappa Coefficient			60.9%	

Table.1 The confusion matrix between remote sensing-based land use classification and the simulated urban categories using the AGA-CA model of Jiading District in 2006

4. CONCLUSION

This paper presents an improved CA model optimized by adaptive genetic algorithm technique, which has been widely used as an evolutionary computation technique. By using the adaptive genetic algorithm technique, a set of transition rules and their defining parameters have been identified and optimised using the limited data available as input data sources. The AGA technique is particularly useful in optimizing the CA transition rules which can be used by conventional CA models based on logistic regression approach. The application of the APA-CA model in Shanghai's Jiading District demonstrates the effectiveness of the AGA technique in transition rule optimization for CA based urban models, which can contribute positively to human studies on urban dynamics.

ACKNOWLEDGEMENT

This study was supported by funding from the Selected and Cultivated Outstanding Young Scholars in Shanghai (Project No. SSC09018), Science and Technology Commission of Shanghai Municipality (Project No. 08230510700), and the Key Project on Science & Technology of China's Ministry of Education (Project No. 209047).

REFERENCES

Al-kheder, S., Wang, J. and Shan, J., 2008. Fuzzy inference guided cellular automata urban-growth modelling using multi-temporal satellite images. *International Journal of Geographical Information Science*, 22(11), pp. 1271-1293.

Batty, M. and Xie, Y., 1994. From cells to cities. *Environment and Planning B*, 21, pp. 531-548.

Batty, M., Xie, Y. and Sun, Z., 1999. Modelling urban dynamics through GIS-based cellular automata. *Computers, Environment and Urban System*, 23, pp. 205-233.

Bies, R. R., Muldoon, M. F., Pollock, B. G., Manuck, S., Smith, G. and Sale, M. E., 2006. A Genetic Algorithm-Based, Hybrid Machine Learning Approach to Model Selection. *Journal of Pharmacokinetics and Pharmacodynamics*, 33(2), pp. 196-221.

Couclelis, H., 1997. From cellular automata to urban models: New principles for model development and implementation. *Environment and Planning B: Planning and Design*, 24, pp. 165-174.

Espinoza, F., Minsker, B. S. and Goldberg, D., 2001. A Self-Adaptive Hybrid Genetic Algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference*, San Francisco, Morgan Kaufmann Publishers.

He, C. Y., Okada, N., Zhang, Q. F, Shi, P. J. and Zhang, J. S., 2006. Modelling urban expansion scenarios by coupling cellular automata model and system dynamic model in Beijing, China. *Applied Geography*, 26, pp. 323-345.

Huang, M. X., Gong, J. H., Zhou, S., Liu, C. B. and Zhang, L. H., 2007. Genetic algorithm-based decision tree classifier for remote sensing mapping with SPOT-5 data in the HongShiMao watershed of the loess plateau, China. *Neural Computing and Applications*, 6(6), pp. 513-517.

Kee, E., Airey, S. and Cye, W., 2001. An Adaptive Genetic Algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 391-397.

Li, X. and Yeh, A. G. O., 2002a. Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, 16(4), pp. 323-343.

Li, X. and Yeh, A. G. O., 2002b. Urban simulation using Principal components analysis and cellular automata for land-use planning. *Photogrammetric engineering and remote sensing*, 68(4), pp. 341-351.

Li, X., He, J. Q. and Liu, X. P., 2009. Intelligent GIS for solving high-dimensional site selection problems using ant colony optimization techniques. *International Journal of Geographical Information Science*, 23(4), pp. 399-416.

Liao, Y. L., Wang, J. F., Meng, B. and Li, X. H. 2008. Integration of GP and GA for mapping population distribution, *International Journal of Geographical Information Science,* DOI: 10.1080/13658810802186874.

Liu, Y., 2008. Modelling urban development with geographical information systems and cellular automata. New York: CRC Press.

Liu, Y. and Phinn, S. R., 2003. Modelling urban development with cellular automata incorporating fuzzy-set approaches. *Computers, Environment and Urban Systems*, 27, pp. 637-658.

Lorena, L. A. N. and Furtado, J. C., 2001. Constructive Genetic Algorithm for Clustering Problems. *Evolutionary Computation*, 9(3), pp. 309-327

Muzy, A., Nutaro, J. J., Zeigler, B. P. and Coquillard, P., 2008. Modelling landscape dynamics in an Atlantic Rainforest region: Implications for conservation. *Ecological modelling*, pp. 219: 212-225.

Schmitt, L. M., Nehaniv, C. L. and Fujii, R. H., 1998. Linear analysis of genetic algorithms. *Theoretical Computer Science*, 208, pp. 111-148

Stevens, D., Dragicevic, S. and Rothley, K., 2007. iCity: A GIS-CA modelling tool for urban planning and decision making. *Environmental Modelling & Software*, 22, pp. 761-773.

White, R. W. and Engelen, G., 1993. Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land use patterns. *Environment and Planning A*, 25, pp. 1175 -1193.

Wu, F., 2002. Calibration of stochastic cellular automata: the application to rural-urban land conversions. *International Journal of Geographical Information Science*, 16(8), pp. 795 - 818.

Wu, F., 1998. Simulating urban encroachment on rural land with fuzzy-logic-controlled cellular automata in a geographical information system. *Journal of Environmental Management*, 53(16), pp. 293-308.

VALIDATION OF PLANAR PARTITIONS USING CONSTRAINED TRIANGULATIONS

Hugo Ledoux and Martijn Meijers

Delft University of Technology
OTB—section GIS Technology
Delft, the Netherlands
{h.ledoux—b.m.meijers}@tudelft.nl

KEY WORDS: validation, planar partitions, triangulation, topology, Simple Features

ABSTRACT:

Planar partitions—full tessellations of the plane into non-overlapping polygons—are frequently used in GIS to model concepts such as land cover, cadastral parcels or administrative boundaries. Since in practice planar partitions are often stored as a set of individual objects (polygons) to which attributes are attached (e.g. stored with a *shapefile*), and since different errors/mistakes can be introduced during their construction, manipulation or exchange, several inconsistencies will often arise in practice. The inconsistencies are for instance overlapping polygons, gaps and unconnected polygons. We present in this paper a novel algorithm to validate such planar partitions. It uses a constrained triangulation as a support for the validation, and permits us to avoid different problems that arise with existing solutions based on the construction of a planar graph. We describe in the paper the details of our algorithm, our implementation, how inconsistencies can be detected, and the experiments we have made with real-world data (the CORINE2000 dataset).

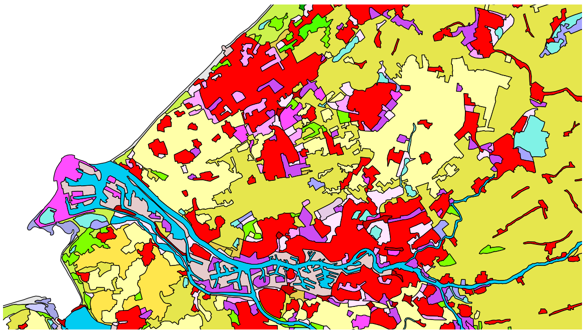


Figure 1: Part of the CORINE2000 dataset for a part of the Netherlands.

1 INTRODUCTION

Planar partitions are frequently used in GIS to model concepts such as land cover, the cadastral parcels or the administrative boundaries of a given area. As shown in Figure 1, a planar partition is a full tessellation of the plane into non-overlapping polygons. The spatial extent is partitioned into polygons, and every location must be covered by one and only one polygon (gaps are thus not allowed). In GIS-related disciplines planar partitions, like other geographical phenomena, are often represented and stored in a computer as a set of individual polygons to which one or more attributes are attached, and the topological relationships between polygons are not stored. The preferred method is with the *Simple Features* paradigm, which is an international standard (OGC, 2006); the *de facto* standard ESRI's shapefile and most databases (e.g. PostGIS) are based on this standard. We discuss in Section 2 details of the Simple Features paradigm that complicate the representation of planar partitions.

If a planar partition is stored as a set of individual polygons, then in practice errors, mistakes and inconsistencies will often be introduced when the planar partition is constructed, updated or exchanged. The inconsistencies most likely to occur are: (i) overlapping polygons (e.g. slivers); (ii) gaps between polygons; (iii) polygons not connected to the others.

In this paper we present a novel algorithm to *validate* such a planar partition, i.e. given a set of polygons stored with the Simple Features paradigm, our algorithm verifies if this set forms a planar partition, or not. As explained in Section 2 different solutions currently exist, these are based on the construction of the planar graph of the polygons and on the use of geometrical and topological validation rules. The solution we propose—using a constrained triangulation as a supporting structure for validation—is described in Section 3 and has in our opinion several advantages over existing methods. We report in Section 4 on our implementation of the algorithm (it uses the stable and fast triangulator of CGAL[1]) and on the experiments we have made with the CORINE Land Cover 2000 dataset. Finally, we discuss the advantages of our method in Section 5.

2 RELATED WORK

Validation of planarity of area partitions has its roots in the definition of what is a valid surface representation for real world features (i.e. a *polygon*). In this section we will first review the Simple Features specification (SFS) that describes what is a valid polygon and secondly how a set of polygons can be validated, so that it forms a planar partition.

2.1 Simple Features

The Simple Features specification is a recognised and used international standard for the storage and access of geographical objects in vector format such as points, lines and polygons. SFS defines a polygon by stating that: “A Polygon is a planar Surface defined by 1 exterior boundary and 0 or more interior boundaries. Each interior boundary defines a hole in the Polygon.” (OGC, 2006). In the specification 6 assertions are given that together define a valid polygon. Essential for a valid polygon is that the boundaries of the polygon must define one connected area (each point inside the polygon can be reached through the interior of the polygon from any other point inside the polygon). Additionally, a polygon can contain holes. We say that the exterior boundary of the polygon is the *outer ring*, and a hole is an *inner ring*. As shown in Figure 2, these holes can be filled by one or more

[1] The Computational Geometry Algorithms Library: `http://www.cgal.org`

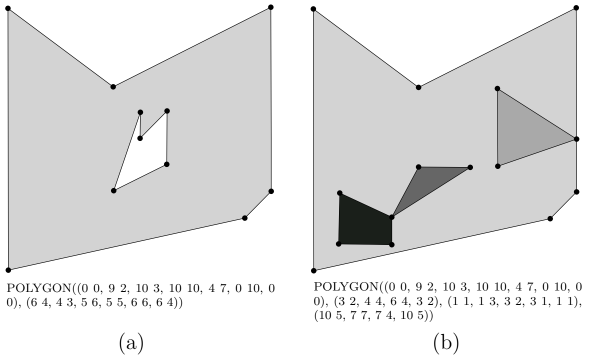

POLYGON((0 0, 9 2, 10 3, 10 10, 4 7, 0 10, 0 0), (6 4, 4 3, 5 6, 5 5, 6 6, 6 4))
(a)
POLYGON((0 0, 9 2, 10 3, 10 10, 4 7, 0 10, 0 0), (3 2, 4 4, 6 4, 3 2), (1 1, 1 3, 3 2, 3 1, 1 1), (10 5, 7 7, 7 4, 10 5))
(b)

Figure 2: Two examples of polygons and their WKT. **(a)** One polygon with one hole. **(b)** Another polygon with three holes, and each of them is filled with an island. Observe that two holes/islands touch each other at one point, and that one hole touch the outer boundary of the polygon at one location.

polygons, which we call islands. Island polygons can recursively contain holes which are filled by islands. Observe also that holes are allowed to interact with each others and the outer boundary under certain circumstances, e.g. they are allowed to touch at one point (as in Figure 2b), as long as the interior of the polygon stays one connected area.

The polygons can be represented either in text (well-known text–WKT) or binary (well-known binary–WKB) formats. Each polygon is stored independently from other polygons; even those adjacent (it is not possible to store topological relationships between the polygons).

Integrity checking of an individual polygon entails checking whether the polygon fulfils the above definition. A naive way of validity checking could be based on checking each segment of each linear ring with all other segments of the polygon for intersections, which is apparently quite costly with respect to computation.

In this work, we have adopted the Simple Feature definition for what we consider is a valid polygon. In the paper by Oosterom *et al.* it was shown in an experiment with different systems and a set of 37 'killer polygons' that in practice the use of this definition is not self-evident and that different products have different interpretations of what is a valid polygon. The authors concluded that "the consistent use of a polygon definition is not yet a reality" (van Oosterom et al., 2002). The validation of one polygon according to the SFS specification has since then found its way into software implementations and is easily possible with different libraries, e.g. GEOS [2] and JTS [3] being two of those (open source) libraries.

2.2 Planar partitions

Planar partitions, such as the CORINE2000 dataset, are freely available in shapefile format where each polygon has one value attached (its code for the land cover). Polygons in such datasets are usually fairly complex (see for instance Figure 3) and the number of polygons is generally very large. The specifications of the dataset states that all polygons form a planar partition, but in practice this is not the case (see Section 4).

Having a definition for what is a valid polygon alone is not enough for certain applications: for those applications it is necessary to define what a valid *set* of polygons is. Therefore, the *disjoint* spatial relation has to be introduced. Two polygons are said to be disjoint, when their interior does not overlap (i.e. does not have any spatial relation). A brute force approach to enforce planarity of a partition with independent, loose-lying polygons is rather cumbersome: it is necessary to check whether each polygon its interior is disjoint with all other polygons, which means a lot of computation (in the order of $O(n^2)$ where n is the number of features to be checked). Furthermore detection of holes between polygons in the partition is only possible by obtaining the union of all individual polygons, which again is computational intensive. As final remark: this also assumes that each individual polygon has already been checked.

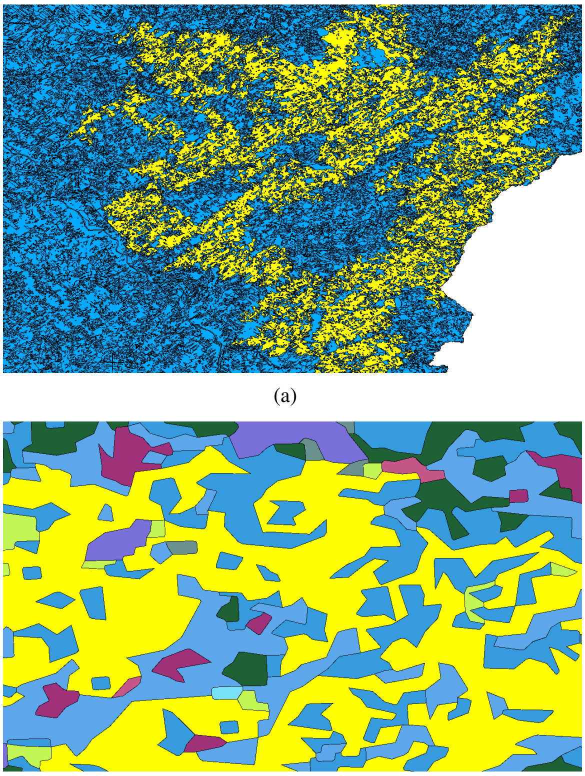


Figure 3: **(a)** One polygon (in yellow) from the Canadian Land Cover Map. Its outer boundary has around 35000 points and it has around 3200 holes. **(b)** Zoom in on this polygon, observe that holes are filled by islands, and that these touch other rings at several points.

In a series of papers on the topic of formal and correct spatial object modelling (Plümer and Gröger, 1996, Gröger and Plümer, 1997, Plümer and Gröger, 1997), a set of mathematical axioms is given for checking the 'validity' of a map, i.e. a collection of polygons. The axioms that Gröger and Plümer form are a correct and complete set of rules to enforce correctness for all polygons in a map together with their adjacency relationships. Checking of the axioms has static and dynamic aspects. Static integrity checking "concerns the question whether a given database as a whole is consistent" (Gröger and Plümer, 1997). An example of the dynamic aspects—how to keep a (in this case cadastral) dataset consistent under updates or transactions—is found in (Matijević et al., 2008). Our approach only examines the state of a geographic dataset as a whole (thus enforcing the integrity rules for a given set of polygons). For the remaining part of this paper, we will focus on static integrity checking.

Furthermore, it is important to note that Plümer and Gröger (1997)

[2] Geometry Engine Open Source: `trac.osgeo.org/geos`

[3] Java Topology Suite: `www.vividsolutions.com/jts/jtshome.htm`

base their axioms on concepts from graph theory, but they also highlight the fact that a graph-based approach alone is not enough: the graph has to be augmented with geometrical knowledge (each vertex has geometry attached, i.e. the coordinates of points have to be stored). Validation is thus underpinned by both geometrical and topological concepts and systems thus have to deal with those two concepts at the same time.

For validating all polygons in a dataset in a single operation, it is necessary to perform a conversion to a graph-based description, which is consecutively checked for consistency (following a set of rules similar to the axioms described by Gröger and Plümer). For this conversion different approaches are available (Shamos and Hoey, 1976, van Roessel, 1991). Implementation of this conversion to a graph-based representation is sometimes difficult, especially if the polygon contains holes. The graph of the boundary is then unconnected and extra machinery is necessary to still represent the knowledge on holes in the graph structure. The fact that holes are also allowed to touch complicates the task of validation even further: holes are supposed to form an unconnected planar graph, but if they touch the graph is connected.

3 VALIDATION WITH THE CONSTRAINED TRIANGULATION

Our approach to validation of planar partitions uses a constrained triangulation (CT) as a supporting structure because, as explained in Section 3.1, CTs are by definition planar partitions. The workflow of our approach is as follows:

1. the CT of the input segments forming the polygons is constructed;
2. each triangle in the CT is flagged with the ID of the polygon inside which it is located;
3. problems are detected by identifying triangles having no IDs, and by verifying the connectivity between triangles.

The flagging and the verification of the connectivity of the input polygons is performed by using graph-based algorithms on the dual graph of the CT.

We describe in this section the concepts needed and we give a detailed description of the different steps. It should be noticed that we assume that each input polygon to our approach is individually valid (as explained in Section 2 this is an easy task and tools are readily available).

3.1 Constrained triangulation

A triangulation decomposes an area into triangles that are non-overlapping. As shown in Figure 4a–b, given a set S of points in the plane, a triangulation of S will decompose its convex hull, denoted conv(S). It is also possible to decompose the convex hull of a set T where points and straight-line segments are present, with a constrained triangulation (CT). In CT(T) every input segment of T appears as an edge in the triangulation (see Figure 4c–d).

If T contains segments forming a loop (which defines a polygon), it permits us to triangulate the interior of this loop (i.e. a triangulation of the polygon). It is known that any polygon (also with holes) can be triangulated without adding extra vertices (de Berg et al., 2000, Shewchuk, 1997). Figure 5 shows an example.

In our approach, the triangulation is performed by constructing a CT of all the segments representing the boundaries (outer +

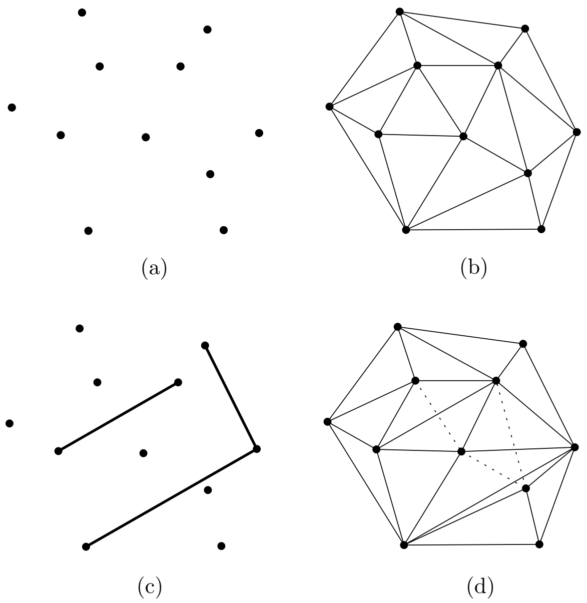


Figure 4: **(a)** A set S of points in the plane. **(b)** A triangulation of S; the union of all the triangles forms conv(S). **(c)** The set S with 3 constrained segments. **(d)** The constrained triangulation of the set of points and segments. The dashed lines are the edges of the triangulation of S that are removed since they are not conform to the input segments.

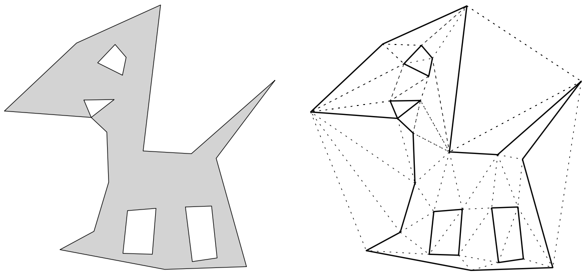


Figure 5: **(a)** A polygon with 4 holes. **(b)** The constrained triangulation of the segments of this polygon.

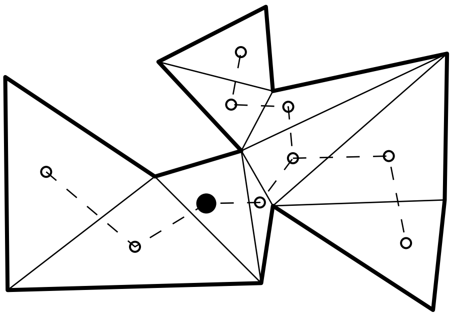


Figure 6: One polygon (thick lines) with its triangulation (normal black lines). The dual graph of the triangulation is drawn with dashed lines, and the filled black point is the centroid of the polygon from where the walk starts.

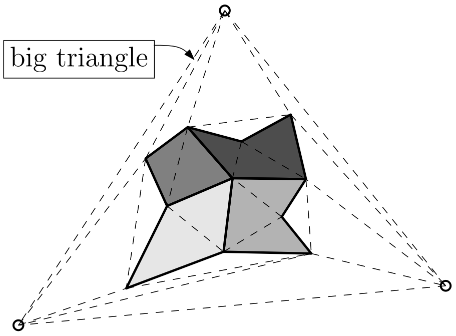


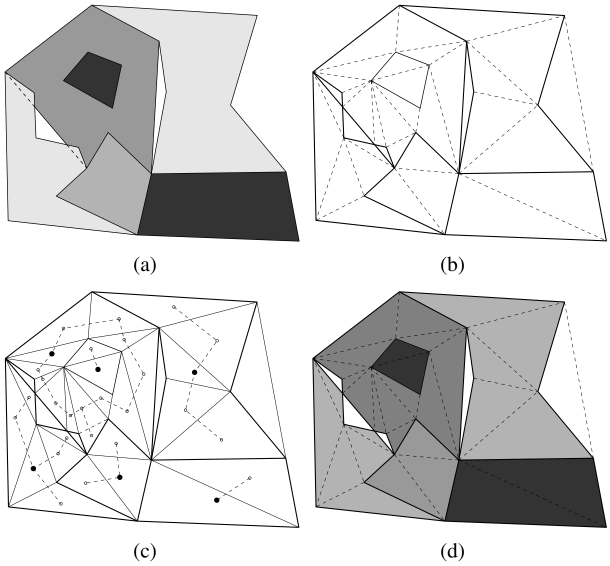
Figure 7: The 4 input polygons are triangulated and are inside the big triangle. A walk from one location outside the 4 polygons would appropriately flag as "universe" the 4 triangles inside the convex hull of the 4 polygons.

inner) of each polygon. If the set of input polygons forms a planar partition, then each segment will be inserted twice (except those forming the outer boundary of the set of input polygons). This is usually not a problem for triangulation libraries because they ignore points and segments at the same location (as is the case with the solution we use, see Section 4).

3.2 Flagging triangles

Flagging triangles means assigning the ID of each polygon to the triangles inside that polygon (the triangles that decompose the polygon). To assign this ID, we first compute one point inside each polygon. This point is what we subsequently call the "centroid" — observe here that this cannot be always the geometric centroid of the polygon as this could be outside the polygon. Our algorithm finds a location inside the polygon and makes sure that this location is not inside one of the holes of the polygon. Then for each centroid c we identify the triangle that contains c, and we start a "walk" on the dual graph of the triangulation, as shown in Figure 6. The walk is a depth-first search (DFS) on the dual graph, and observe that constrained edges in the triangulation will act as blockers for the walk. Observe also that islands are not a problem (see Figure 8c).

Big triangle. To appropriately flag all the triangles of the CT (those inside the convex hull of the input points/segments but not inside an input polygon) we exploit one particularity of libraries to compute triangulation: the so-called "big triangle", which is also being called the "far-away point" (Liu and Snoeyink, 2005). Many implementations indeed assume that the set S of points is entirely contained in a big triangle τ_{big} several times larger than the range of S. Figure 7 illustrates the idea. With this technique the construction of the CT is always initialised by first constructing τ_{big}, and then the points/segments are inserted. Doing this has many advantages, and is being used by several implementations (Facello, 1995, Mücke, 1998, Boissonnat et al., 2002). To assign an ID "universe" to the triangles, we simply start at one triangle incident to one vertex of τ_{big} and perform the same walk as for the other polygons.



(a) (b)

(c) (d)

Figure 8: **(a)** Six polygons form the input planar partition. **(b)** The constrained triangulation of the boundaries of the input polygons. **(c)** The dual graph of the triangles is drawn with dashed lines; the dark points are the points from which the walk in each polygon starts. **(d)** The result contains triangles that not flagged (white triangles). The white triangle on the right is not a problem since it is a "universe" triangle.

3.3 Identifying problems

If the set of input polygons forms a planar partition then all the triangles will be flagged with one and only one ID. Notice that because of the big triangle, triangles outside the spatial extent of the planar partitions will be flagged as "universe". Notice also that if a polygon contains a hole, then for the planar partition to be valid this hole must be filled completely by another polygon (an island).

If there are gaps and/or overlaps in the input planar partition then some triangles will not be flagged. We can detect these easily by verifying the IDs. Figure 8 illustrates one input planar partition that contains 6 polygons; notice that one has an island and that some polygons overlap and that there are also gaps. The walk starting from each centroid is shown in Figure 8c, and the resulting flagging of triangles is shown in 8d (the grey shadings represent the IDs). When 2 or more polygons overlap then depending on the location of the centroids some triangles will not be flagged (because the constrained edges block the walks).

Another problem that could arise is when the union of the input polygons forms more than one polygon. Figure 9 shows one example with 5 input polygons: 4 of them form a valid planar partition but one is not connected to the others (thus the 5 polygons do not form a planar partition). We solve that problem by starting a walk from any centroid, but that walk is not stopped by the constrained, only by the triangles flagged as "universe". The connectivity problem simply boils down to ensuring that all the triangles flagged with an ID other than "universe" can be reached.

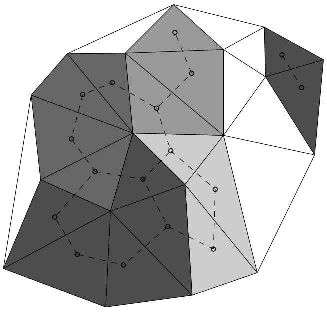


Figure 9: Five polygons, with one unconnected to the other ones. The dual graph for the flagged triangles is shown in with dashed lines.

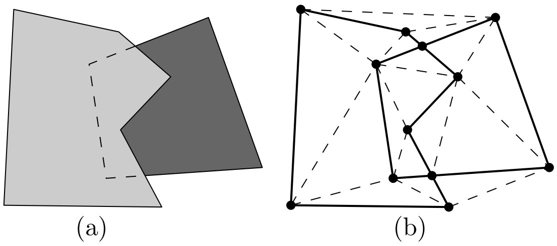


Figure 10: **(a)** Two overlapping polygons. **(b)** CGAL's constrained triangulation of the polygons.

4 IMPLEMENTATION AND EXPERIMENTS

We implemented the algorithm described in this paper with the Python language[4]. Our implementation reads as input either a *shapefile* or a set of WKTs, and tells the user what problems are present in the input polygons (if any).

For the constrained triangulation, we rely entirely on the implementation of CGAL (we use the Python bindings of CGAL[5]). Each segment of the input polygons is inserted incrementally in the CT. When 2 segments are identical, the second one is simply ignored. Since the input if formed of individual polygons, it is faster (and simpler) to rely on the spatial indexing scheme of CGAL to detect the duplicate edges than to pre-process them with an auxiliary data structure. It should be noticed that we use the default tolerance in CGAL to determine if 2 points are at the same location.

We also rely on CGAL for ensuring that a valid triangulation is formed when 2 or more polygons overlap. As shown in Figure 10, if 2 polygons overlap their segments will intersect (which would not be a valid planar graph). However, CGAL has built-in operations to calculate the intersection of 2 segments and to create new sub-segments.

We have tested our implementation with different parts of the CORINE2000 dataset. This is a dataset modelling the land cover for the whole of Europe, and it is freely available[6]. The dataset is divided into tiles and each tile can be downloaded as a shapefile. Although the specifications of CORINE2000 state that the polygons form a planar partition and that validation rules are used, we found several errors.

One example is when creating one planar partition from two adjacent tiles, as shown in Figure 11. The process of tiling the whole dataset has obviously introduced errors because several sliver polygons were detected during our experiments. We have also found one case where a polygon had been obviously "shifted" manually by a user (see Figure 12).

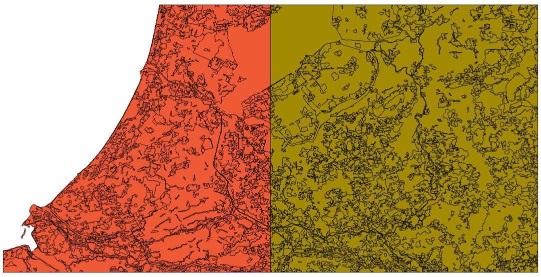


Figure 11: CORINE2000's tiles E39N32 and E40N32.

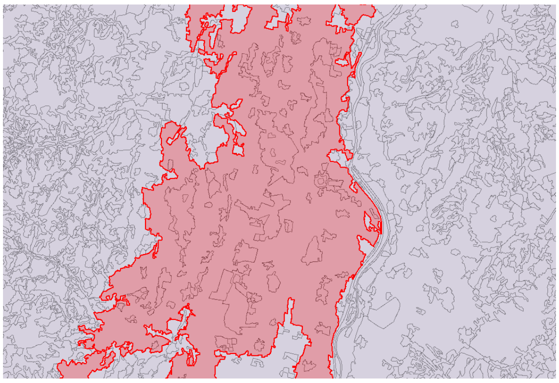


Figure 12: A polygon manually shifted (from CORINE2000 tile E41N27) – it is overlapping with neighbours on one side and gaps are present on the opposite side.

5 DISCUSSION AND CONCLUSIONS

The problem of validating a planar partition stored with Simple Features is theoretically a simple one: construct the planar graph of the input, and define a set of geometric and topological validation rules. Unfortunately, the implementation of a planar graph construction algorithm and of the validation rules is far from being trivial (especially when the input polygons contain holes) and can often not scale to big datasets containing millions of polygons.

We have presented in this paper a new algorithm and we have successfully implemented it. Our approach solves most of the current problems and has in our opinion several advantages:

1. The algorithm is simple and can be implemented easily over a CT library such as CGAL. The only things needed are: (i) to be able to add attributes to triangles (for the IDs); (ii) having access to the data structure. All the validation rules simply boil down to flagging triangles and graph-based searches.

2. The holes/islands inside polygons are easily handled by the CT. No additional data structure or special mechanisms are necessary, as is the case with planar graph approaches.

3. The implementation can be built over well-known and optimised CT implementations, which are fast and can handle millions of objects. It is known that triangulations of several millions points can be managed in main memory (Amenta et al., 2003, Blandford et al., 2005).

[4] `http://www.python.org`

[5] `http://cgal-python.gforge.inria.fr`

[6] More information can be found on `http://www.eea.europa.eu/themes/landuse/clc-download`

4. If problems are present in the input, we believe the CT could be used to *automatically repair* the planar partition. That would simply involve (re)flagging the IDs of problematic triangles (based on some user-defined rules) and then "following" the boundaries between IDs to reconstruct polygons and give them back to the user in Simple Features format. We see great potential for such an application.

5. Apart from static integrity checking, our approach could be used for keeping a dataset consistent under a set of edits (dynamic checking). The CT can then be used to locally check the validity of an update.

For future work, we plan on implementing the algorithm in C++ to be able to scale to massive datasets, and we also plan on working on the automatic repairing and incremental updates with the help of the CT. Finally, the ideas presented in this paper are all valid in higher dimensions and we plan on implementing them for constrained tetrahedralization (Si, 2004), which would permit us to validate 3D city models for instance.

6 ACKNOWLEDGEMENTS

We would like to thank Gustavo Adolfo Ken Arroyo Ohori for proof-reading this document and suggesting useful improvements.

REFERENCES

Amenta, N., Choi, S. and Rote, G., 2003. Incremental constructions con BRIO. In: Proceedings 19th Annual Symposium on Computational Geometry, ACM Press, San Diego, USA, pp. 211–219.

Blandford, D. K., Blelloch, G. E., Cardoze, D. E. and Kadow, C., 2005. Compact representations of simplicial meshes in two and three dimensions. International Journal of Computational Geometry and Applications 15(1), pp. 3–24.

Boissonnat, J.-D., Devillers, O., Pion, S., Teillaud, M. and Yvinec, M., 2002. Triangulations in CGAL. Computational Geometry—Theory and Applications 22, pp. 5–19.

de Berg, M., van Kreveld, M., Overmars, M. and Schwarzkopf, O., 2000. Computational geometry: Algorithms and applications. Second edn, Springer-Verlag, Berlin.

Facello, M. A., 1995. Implementation of a randomized algorithm for Delaunay and regular triangulations in three dimensions. Computer Aided Geometric Design 12, pp. 349–370.

Gröger, G. and Plümer, L., 1997. Provably correct and complete transaction rules for GIS. In: Proceedings 5th ACM international workshop on Advances in geographic information systems, New York, NY, USA, pp. 40–43.

Liu, Y. and Snoeyink, J., 2005. The "far away point" for Delaunay diagram computation in $\mathbb{E}^d$. In: Proceedings 2nd International Symposium on Voronoi Diagrams in Science and Engineering, Seoul, Korea, pp. 236–243.

Matijević, H., Biljecki, Z., Pavičić, S. and Roić, M., 2008. Transaction processing on planar partition for cadastral application. In: Proceedings FIG Working Week 2008—Integrating Generations, Stockholm, Sweden.

Mücke, E. P., 1998. A robust implementation for three-dimensional Delaunay triangulations. International Journal of Computational Geometry and Applications 8(2), pp. 255–276.

OGC, 2006. OpenGIS implementation specification for geographic information—simple feature access. Open Geospatial Consortium inc. Document 06-103r3.

Plümer, L. and Gröger, G., 1996. Nested maps—a formal, provably correct object model for spatial aggregates. In: Proceedings 4th ACM International Symposium on Advances in GIS, ACM, New York, NY, USA, pp. 76–83.

Plümer, L. and Gröger, G., 1997. Achieving integrity in geographic information systems—maps and nested maps. GeoInformatica 1(4), pp. 345–367.

Shamos, M. I. and Hoey, D., 1976. Geometric intersection problems. In: FOCS, IEEE, pp. 208–215.

Shewchuk, J. R., 1997. Delaunay Refinement Mesh Generation. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburg, USA.

Si, H., 2004. Tetgen: A quality tetrahedral mesh generator and three-dimensional Delaunay triangulator. User's manual v1.3 9, WIAS, Berlin, Germany.

van Oosterom, P., Stoter, J., Quak, W. and Zlatanova, S., 2002. The balance between geometry and topology. In: D. Richardson and P. van Oosterom (eds), Advances in Spatial Data Handling—10th International Symposium on Spatial Data Handling, Springer, pp. 209–224.

van Roessel, J. W., 1991. A new approach to plane-sweep overlay: Topological structuring and line-segment classification. Cartography and Geographic Information Science 18, pp. 49–67.

SPACE-TIME KERNELS

J.Q. Wang, T. Cheng*, J. Haworth

Department of Civil, Environmental and Geomatic Engineering, University College London,
Gower Street, WC1E 6BT London, United Kingdom {w.jiaqiu; tao.cheng; j.haworth}@ucl.ac.uk;

Commission II, WG II/3

KEY WORDS: Space-Time Kernels; Space-Time Analysis; Support Vector Regression;

ABSTRACT:

Kernel methods are a class of algorithms for pattern recognition. They play an important role in the current research area of spatial and temporal analysis since they are theoretically well-founded methods that show good performance in practice. Over the years, kernel methods have been applied to various fields including machine learning, statistical analysis, imaging processing, text categorization, handwriting recognition and many others. More recently, kernel-based methods have been introduced to spatial analysis and temporal analysis. However, how to define kernels for space-time analysis is still not clear. In the paper, we firstly review the relevant kernels for spatial and temporal analysis, then a space-time kernel function (STK) is presented based on the principle of convolution kernel for space-time analysis. Furthermore, the proposed space-time kernel function (STK) is applied to model space-time series using support vector regression algorithm. A case study is presented in which STK is used to predict China's annual average temperature. Experimental results reveal that the space-time kernel is an effective method for space-time analysis and modelling.

1. INTRODUCTION

Kernel methods are a class of algorithms for pattern recognition. The general task of pattern recognition is to find and study various patterns (such as clusters, correlations, classifications, regressions, etc) in different types of data (such as time series, spatial data, space-time series, vectors, images, etc) (Scholkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). To date, kernel-based methods have been applied to a range of areas including machine learning and statistical analysis amongst others and have subsequently become a very active research area (Kanevski et al, 2009). Some of the best known algorithms capable of operating with kernels are support vector machines (Vapnik, 1995), general regression and probabilistic neural networks (Specht, 1991), canonical correlation analysis (Melzer et al, 2003), spectral clustering (Dhillon et al, 2004) and principal components analysis (Hoffmann, 2007).

Recently, kernel functions have been introduced to spatial analysis (Fotheringham et al, 2002; Hallin et al, 2004; Pozdnoukhov and Kanevski, 2008) and temporal analysis (Rüping, 2001; Ralaivola and d'Alché-Buc, 2004; Sivaramakrishnan et al, 2007). In the field of spatial analysis; Fotheringham et al (2002) developed a method using a Gaussian kernel function for the analysis of spatially varying relationships called Geographically Weighted Regression (GWR). GWR has been widely used for spatial analysis including house price prediction, ecological distribution, etc. Pozdnoukhov and Kanevski (2008) present a methodology for data modelling with semi-supervised kernel methods, which is applied to the domain of spatial environmental data modelling. They demonstrate how semi-supervised kernel methods can be applied in this domain, starting from feature selection; to model selection and up to visualization of the results. A case study of topo-climatic mapping reveals that the described methodology of data-driven modelling of complex environmental processes using machine learning methods improves the modelling considerably. In the field of temporal analysis, Ralaivola and d'Alché-Buc (2004) proposed a new kernel-based method as an extension to linear dynamical models. The kernel trick is used twice; first, to learn the parameters of the model, and second, to compute preimages of the time series predicted in the feature space by means of Support Vector Regression (SVR). Their model shows strong connection with the classic Kalman Filter model. Kernel-based dynamical modelling is tested against two benchmark time series and achieves high quality predictions. Sivaramakrishnan et al

(2007) propose a novel family of kernels for multivariate time-series classification problems. Each time-series is approximated by a linear combination of piecewise polynomial functions in a reproducing kernel Hilbert space by a novel kernel interpolation technique. Through the use of a kernel function, a large margin classification formulation is proposed, which can discriminate between two classes. The formulation leads to kernels, between two multivariate time-series, which can be efficiently computed. Furthermore, the proposed kernels have been successfully applied to writer independent handwritten character recognition.

The use of kernel methods in spatial and temporal analysis has been widely covered in the literature; however, how to accommodate kernels in spatio-temporal analysis is still unclear and hence forms the focus of the current study. The structure of the paper is as follows; in section two, a review of the relevant kernels that can be applied to spatial and temporal analysis is carried out; in section three; a space-time kernel (STK) function is proposed based on the principle of a convolution kernel that combines spatial and temporal kernels; in section four, a support vector regression machine is developed that makes use of STK (SVR-STK) to model space-time series. The final section summaries the major findings and proposes the direction of further research.

2. REVIEW OF KERNELS IN SPACE-TIME ANALYSIS

2.1 Kernels in spatial analysis

In spatial analysis, kernels are used as weighting functions to model and explain local spatial autocorrelation and heterogeneity features. For example, in Geographically Weighted Regression (GWR) (Fotheringham et al, 2002), a Gaussian kernel is used to model geographical data whose weights decrease continuously as the distance between the two points increases (note, Fotheringham et al (2002) also recommend the bi-square kernel function as an alternative). A Gaussian kernel, as seen in Figure 1, is defined as a symmetric monotonic function that decreases in value as the distance increases between the target spatial unit z_i and the neighbouring spatial unit z_j.

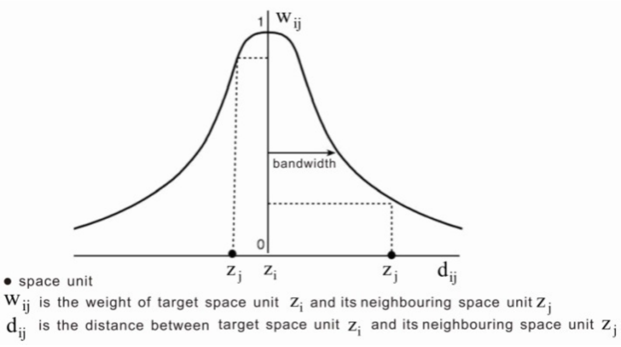

1
w_{ij}
bandwidth
0
z_j
z_i
z_j
d_{ij}
● space unit
w_{ij} is the weight of target space unit z_i and its neighbouring space unit z_j
d_{ij} is the distance between target space unit z_i and its neighbouring space unit z_j

Figure 1. Sketch map of spatial kernel (Fotheringham et al, 2002)

The Gaussian kernel function takes the following form:

$$w_{ij} = \alpha \cdot e^{\frac{d_{ij}^2}{2\sigma^2}} \qquad (1)$$

where d is the distance between target spatial unit z_i and its neighbouring spatial unit z_j and σ^2 is variance; also referred to as bandwidth (Fotheringham et al, 2002). The parameter σ^2 can change the smoothing degree of the Gaussian function curve; which alters the contribution of each neighbouring spatial unit z_j localized to a region nearby target spatial unit z_i. For a given regression point, the weight of a data point is at a maximum when it shares the same location as the regression point. This weight decreases continuously as the distance between the two points increases according to σ^2. In this way, a regression model is calibrated locally simply by moving the regression point across the region. For each location, the data will be weighted differently so that the results of any one calibration are unique to a particular location.

Kanevski et al (2009) apply a multi-scale kernel to deal with the problem of spatial interpolation of environmental data at different scales; the usual spatial interpolation methods are global and smoothing and can only deal with an average scale. This issue is addressed by considering a linear combination of Gaussian radial basis functions of different bandwidths. For a spatial modelling problem, multi-scale Radial Basis Functions (RBF) can be used:

$$f(x,\alpha) = \sum_{i=1}^{N}\sum_{p=1}^{K}\left(\alpha_i^{*} - \alpha_i^{(p)*}\right)e^{-\frac{(x-x_i)^2}{2\sigma_p^2}} + b \qquad (2)$$

where k is the number of kernels and $\alpha_i^{(p)}$ is the weight corresponding to i -th training point and p -th kernel. A potential issue with this technique is that the choice of parameter k increases the dimension of the optimization problem, which is

$2N(k+1)$. Moreover, k and bandwidths σ_p have to be tuned, which can reflect the change of spatial process in scale.

2.2 Kernels in temporal analysis

Rüping (2001) provides an overview of some of the kernel functions that can be applied to time series analysis, and discusses their relative merits. Typically, time series analysis requires a higher level of reasoning than simple numerical analysis can provide and therefore model assumptions must be carefully considered. Experiments are carried out to discover if these different model assumptions have effects in practice and if kernel functions exist that allow time series data to be processed with support vector machines without intensive pre-processing. Rüping (2001) tests various kernel functions that are capable of being applied to time series analysis, including linear kernels, RBF kernels, Fourier kernels, Subsequence Kernels, PHMM Kernels, Polynomial kernels, etc. To give an example, a linear kernel $k(x,y)=x \cdot y$ is the most simple kernel function. The decision function takes the form $f(x)=w \cdot x+b$. When one uses the linear kernel to predict time series, i.e. $x_T=f(x_{T-1},\cdots,x_{T-k})=\sum_{t=1}^{k} w_t x_{T-t}+b$, the resulting model is a statistical autoregressive model of the order k ($AR[k]$). With the kernel, time series are taken to be similar if they are generated by the same AR-model.

Of most interest to this study is the Fourier kernel; since it can handle Fourier transformations. This representation is useful if the information of the time series does not lie in the individual values at each time point but in the frequency of some events. It was noted by Vapnik (1995) that the inner product of the Fourier expansion of two time series can be directly calculated by the regularized kernel function:

$$K_1(x,y)=\left\{\frac{1-q^2}{2(1-2q\cos(x-y)+q^2)}\middle| 0<q<1\right\} \quad (3)$$

where q is regularization multiplier, which controls degree of attenuation of high frequency component in Fourier expanded equation. With the increase of q, SVR can express high frequency component more and enhance complexity of model. Conversely, with the reduction of q, high frequency component in data will attenuate quickly. Thus, the choice of q will influence the characterization ability of SVM for explaining the degree of data complexity. The schematic graph of Fourier kernel can be seen in Figure 2.

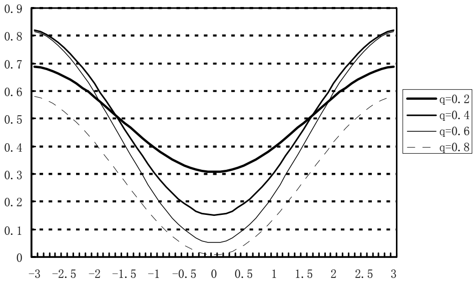


Figure 2. Schematic graph of Fourier kernel

3. SPACE-TIME KERNELS FUNCTION (STK)

The design of kernels for particular tasks is an open research problem. Kernel design methodology that incorporates prior knowledge into the kernel function is an important part of the successful application of the method (Kanevski et al, 2009). As discussed above, kernel functions can tackle spatial and temporal analysis using *kernel tricks* in machine learning and statistical models. The *kernel trick* is a method for using a linear classifier or regression algorithm to solve a nonlinear problem by mapping the original input space into a higher-dimensional feature space (Kanevski et al, 2009). According to kernel theory, a convolution kernel is a kind of construction kernel function, whose operation will be enclosed based on a standard kernel function (i.e. Polynomial kernel, Gaussian kernel, etc) (Haussler, 1999). A convolution kernel has following form:

$$K(x,y)=\sum_{x \in R^{-1}, y \in R^{-1}} \prod_{i=1}^{J} K_i(x_i,y_i) \quad (4)$$

where R^{-1} is finite set and K is convolution of basic kernel functions $K_1,K_2,\cdots,K_D(K_1 \times K_2 \times \cdots \times K_j)$. We assume space-time kernel as $K_{ST}(x,y)$ and its form is:

$$K_{ST}(x,y)=\sum_{t=1}^{□} \text{□EMBED Equation.3 □□□}(K_S(x,y) \cdot K_T(x,y$$

(5)

where $K_{ST}(x,y)$ is space-time kernel, which processes space-time convolution; $K_S(x,y)$ is a spatial kernel, which processes spatial convolution; $K_T(x,y)$ is a temporal kernel, which processes temporal convolution; λ is the order of the kernel function. Generally, a bigger λ can improve the learning ability of the kernel function. To avoid overfitting, λ should not be too large.

As discussed in Section 2.1, a Gaussian function is an important function that is able to tackle local spatial heterogeneous characteristics in geographical data. Additionally, Gaussian kernels have proven learning ability in machine learning regardless of the dimensionality of the sample data. Therefore, it can be used in the spatial kernel $K_S(x,y)$ discussed in Section 2.1 with following form:

$$K_S(x,y) = \left\{ \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2} \right) \middle| \sigma > 0 \right\} \quad (6)$$

where $\|\mathbf{x}-\mathbf{y}\|$ is the distance between target spatial unit x and its neighbouring spatial unit; and σ^2 is the kernel bandwidth, which is a parameter for spatial kernel $K_S(x,y)$. σ^2 changes the smoothing degree of Gaussian curve, which varies the contribution of each neighbouring spatial unit x localized to a region nearby target spatial unit y .

Convolution theorem states that Fourier transformations can convert complex convolution operations to simple product operations (Nussbaumer, 1982). This indicates that Fourier kernels can be used to tackle convolution in time. The Fourier kernel has been discussed in Section 2.2. Additionally, it should be noted that the Fourier kernel is well suited to modelling periodic series (including *sine* and *cosine* frequency components). As for sequences there is no periodicity so a polynomial kernel is more appropriate due to its stronger generalization ability. A polynomial kernel takes the following form:

$$K_2(x,y) = \left\{ \left((x \cdot y) + 1 \right)^d \middle| d > 0 \right\} \quad (7)$$

where d is the order of the polynomial kernel. With reduction of d , generalization ability of the polynomial kernel will become stronger. Larger d will improve the complexity of the machine learning algorithm, resulting in the decline of generalization ability.

As discussed above, Fourier kernels and Polynomial kernels strongly complement each other. Therefore, we can combine them to approximate any series as long as kernel parameters are exact to the right degree. Thus, the temporal kernel $K_T(x,y)$ can be expressed mathematically as equation (8) where α is a coefficient to give more impact to the Fourier kernel K_1 and Polynomial kernel K_2 ; d and q are kernel parameters of the two kinds of basic kernel functions

According to Equation 5, 6 and 8, the expression of the space-time kernel can be derived as equation (9).

The function of Equation 9 is called the space-time kernel function (STK).

4. APPLICATION OF STK

To test the performance of STK, it is applied to the modelling of space-time series, which are sets of location-related time series (Bennett, 1975; Martin and Oeppen, 1975). The Support vector algorithm, one of the basic and most advanced algorithms, is a natural field of application for kernels. Hence, here an SVR model with STK is constructed and used to analyze and model nonlinear space-time series. Figure 3 describes the structure and target function of the SVR machine with STK. The output expression in Figure 3 is the objective function of SVR with STK (called SVR-STK) which is a regression function rather than a classification function.

.

$$\left\{ \alpha \cdot \frac{1-q^2}{2(1-2q\cos(x-y+q^2))@ + (1-\alpha)\cdot((x\cdot y)+1)^d} \middle| 0 \le \alpha \le 1; d > 0; 0 < q < 1 \right\} \quad (8)$$

$$\left\{ \left(\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2} \right) \right) \cdot \left(\alpha \cdot \frac{1-q^2}{2(1-2q\cos(x-y+q^2))@ + (1-\alpha)\cdot((x\cdot y)+1)^d} @ \middle| \sigma > 0; 0 \le \alpha \le 1; d > 0; 0 < q < 1 \right) \right\} \quad (9)$$

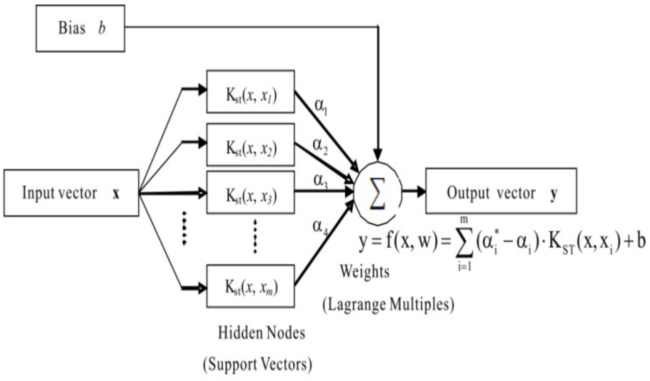
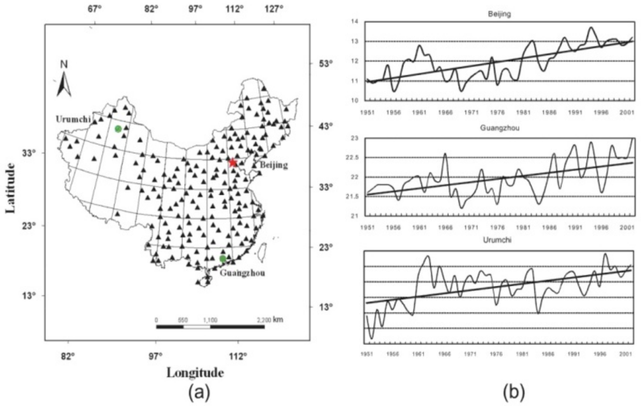


Figure 3. Architecture of support vector regression machine with space-time kernel (STK)

The model of Figure 3 is validated using data obtained from the national meteorological centre of P. R. China, including yearly temperature at 194 national meteorological stations (with geographical coordinates - longitude x and latitude y) from 1951-2002 as seen in Figure 4 (Cheng and Wang, 2009).



	Fitted (1951-1992)		
	RMSE		
	Plain SVR	Time series SVR	SVR-STK
Beijing	0.981	0.462	0.209
Guangzhou	0.910	0.314	0.084
Urumchi	1.173	0.853	0.306
	Forecasting (1993-2002)		
	RMSE		
	Plain SVR	Time series SVR	SVR-STK
Beijing	0.802	0.316	0.403
Guangzhou	0.813	0.418	0.387
Urumchi	0.837	0.551	0.541

Figure 4. Meteorological stations in study area: (a) spatial location distribution of the 194 stations; (b) graph of time series and trends of annual average temperature from 1951 to 2002 at the three stations of Beijing, Guangzhou, and Urumchi.

Of the 194 observation stations, there are huge data gaps in 57 stations. The data of these 57 stations are discarded, and data of 137 stations are used for the following test. To train and validate the models the data sets are split into two subsets: 80% as a sample set to train the model, and 20% as a validation set to test and validate the model. Thus, in this case, the meteorological data between 1951 and 1992 (42 years in total, nearly 80% of 52 years) is chosen as the training dataset for the forecasting between 1993 and 2002 (10 years in total, nearly 20% of 52 years).

Next, the SVR-STK model is constructed and trained after exploratory space-time analyses are undertaken. Each spatial unit is predicted in the experiment. Since the parameters of Equation 9 are numerous, selection of the arguments is tedious. The parameters of Equation 9 are adjusted and chosen according to the cross-validation method in order to obtain the best results. One-step-ahead forecasting, which is the most common testing standard, is considered in this case study. The SVR-STK results are compared firstly against a standard SVR model with inputs:

$$x_i, y_i, t_j \quad | \{i = 1, \dots n, j = 1, \dots m\} \tag{10}$$

Where x_i and y_i are the geographic coordinates of the i th station and t_j is the j th time period. Secondly, they are compared against pure time series SVR for the three individual test stations. The RBF kernel is used for both comparison tests; parameters were tuned separately for each station. Table 1 summarizes the accuracy measures using RMSE index for the fitted and forecasting results. SVR-STK significantly outperforms the plain SVR model for fitting and forecasting, achieving forecasting improvements of 49.75%, 52.4% and 35.36% for Beijing, Guangzhou and Urumchi respectively. SVR-STK also outperforms pure time series SVR for two of the three stations; Guangzhou and Urumchi, by 7.42% and 1.81% respectively. There is no improvement for Beijing, but given that SVR-STK requires only one set of parameters to be trained for all stations, the results are promising.

Table 1. Accuracy (RMSE) measures for three meteorological stations Beijing, Guangzhou and Urumchi in 52 years

5. CONCLUSIONS AND DISCUSSION

In the present paper, a space-time kernel function (STK) is presented, and the proposed STK is applied to the modelling of space-time series by support vector regression algorithm. An illustrative case study is presented in which China's annual average temperature at 137 international meteorological stations from 1993-2002 is predicted using a support vector regression model with STK (SVR-STK). Although good results are achieved, further validation is still needed. Moreover, the following problems are identified; firstly, more research is needed into whether the proposed space-time kernel can be used to model and explain local space-time autocorrelation and heterogeneity, and secondly; whether the space-time kernel can be introduced to GWR modelling using some *kernel tricks*. The above two problems should be considered in further research.

Acknowledgements

This research is support by Chinese 973 Programme (2006CB701305) and 863 programme (2009AA12Z206), and UK EPSRC (EP/G023212/1).

References

Aizerman, M., Braverman, E., and Rozonoer, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25, pp. 821-837.

Bennett. R. J., 1975. The Representation and identification of spatio-temporal systems: an example of population diffusion in north-west England. *Transaction of the Institute of British Geographers*, 66, pp. 73-94.

Cheng, T., Wang, J.Q., 2009. Accommodating spatial associations in DRNN for space–time analysis. *Computers, Environment and Urban Systems*, 33(6), 409-418.

Dhillon, I.,Guan, Y., and Kulis, B. 2004. Kernel k-means, spectral clustering and normalized cuts. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* Seattle, WA, USA, pp, 551-556.

Fotheringham, S., Chris Brundson, A., and Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Hallin, M., Lu, Z., Tran, L.T., 2004. Kernel density estimation for spatial processes: the L1 theory. *Journal of Multivariate Analysis,* 88, pp. 61-75.

Haussler, D., 1999. *Convolution kernels on discrete structures*. Technical report, University of Santa Cruz.

Hoffmann, H., 2007. Kernel PCA for Novelty Detection. *Pattern Recognition*. 40. 863-874.

Kanevski, M., Pozdnoukhov, A., and Timonin, V., 2009. *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*. EPFL Press.

Martin, R.J., and Oeppen, J.E., 1975. The identification of regional forecasting models using space-time correlation functions. *Transactions of the Institute of British Geographers*, 66, pp. 95-118.

Melzer, T., Reitera, M., and Bischof, H., 2003. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition.* 36,pp, 1961-1971.

Nussbaumer, H. J., 1982. *Fast fourier transform and convolution algorithms*. Springer, Berlin.

Pozdnoukhov, A., and Kanevski, M., 2008. GeoKernels: modeling of spatial data on GeoManifolds. *In M. Verleysen, editor, ESANN 2008: European Symposium on Artificial Neural Networks – Advances in Computational Intelligence and Learning, Bruges*, Belgium, 23-25, April.

Ralaivola, L., and d'Alché-Buc F., 2004. Dynamical modeling with kernels for nonlinear time series prediction. *Advances in neural information processing systems*, 16, pp. 129 - 136.

Rüping, S., 2001. SVM kernels for time series analysis. In: R. Klinkenberg, S. Rüping, A. Fick, N. Henze, C. Herzog, R. Molitor, and O. Schröder (ed.), LLWA 01-Tagungsband der GI-Workshop-Woche Lernen-Lehren-Wissen-Adaptivitet, pp. 43-50.

Scholkopf, B., Smola, A., 2002. *Learning with kernel: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Shawe-Taylor, J., Cristianini, N., 2004. *Kernel methods for Pattern Analysis*. Cambridge University Press.

Specht, D., 1991. A general regression neural network. *IEEE Transaction on Neural Network*. 2, pp. 568-576.

Sivaramakrishnan, K.R, Karthik, K., and Bhattacharyya, C., 2007. Kernels for large margin time-series classification. *IEEE Int Joint Conference on Neural Networks*. pp. 2746-2751.

Vapnik, V., 1995. *The nature of statistical learning theory*. New York, Springer-Verlag.

MODELLING LAND ALLOCATION PROCESS IN TIME AND SPACE

M.A. Sharifi[a], M. Karimi[b], and M.S. Mesgari[b]

a: Faculty of Geo-Information Science and Earth Observation (ITC), Tewente University, The Netherlands - alisharifi@itc.nl
b: GIS Department, Geodesy and Geomatics Faculty, K.N. Toosi University of Technology, Tehran, Iran - mesgari@kntu.ac.ir , karimi_paveh@yahoo.com

Keywords: land allocation; land use change; regional level; SMCDA; cellular automata; expert knowledge

ABSTRACT

Modeling land allocation for different land uses at regional level is a multi-dimensional problem, as it is influenced by spatial, temporal and dynamics of environmental and socio-economic factors in the complex process of land use change. It is therefore a challenge to develop a proper model that can handle these complexities. In this context, a new model has been conceptualized and developed. The model is considering the expert knowledge, the heuristics and human decision-making in modeling the process of land use change. This paper briefly presents the conceptual model, which is developed based on cellular automata concept and tested with the Borkhar and Meymeh township datasets in Esfahan, Iran.

1. INTRODUCTION

Land use allocation process is the result of interaction between land suitability and land demand in an environment affected by socio-economic, political and administrative rules and regulations. Therefore, it is a complex, dynamic and non linear process. The assessment of land suitability for land use types is normally carried out by comparisons of the land use requirement with the land use characteristics. This process is mainly implemented at polygon or pixel “micro level”. On the other hand, the demands for different land use types are estimated at administrative “macro level”, considering various scenarios. The interaction of the above two processes “in micro and macro levels” derived by various socio-economic variables of the environment causes the land use change, and can be used for land allocation.

This paper briefly presents a new conceptual model for land allocation process at regional level. The model is based on application of the basic tools and techniques such as geographic information systems, remote sensing, fuzzy logic, cellular automata, analytical hierarchical process, and spatial multi-criteria decision analysis. The developed model is implemented in Borkhar and Meymeh Township, in Esfahan province, Iran, for the periods of 1986-1998 and 1998-2005. The paper first presents the conceptual model, followed by introduction to the study area, some implementation results followed by discussion and conclusions.

2. CONCEPTUAL LAND ALLOCATION MODEL

Land allocation for different land use types at regional level is a multi-dimensional problem, as it is influenced by spatial, temporal and dynamics of environmental and socio-economic factors in the complex process of land use change. In this context, a model has been conceptualized and presented in Figure 1.

The model includes three major processes i) land suitability assessment, ii) land demand assessment, and iii) land allocation. In the land suitability assessment model, suitability of each location for different land use types is assessed by determining and integrating the effects and impacts of environmental and socio-economic factors. In land demand assessment model, using different concepts such as analysis of trends and scenarios, demand for each land uses are estimated. Finally in the allocation model, considering land suitability, land demand, the current land use/cover and land use conversion rules, land use types are assigned to each locations. In the following each of these models are briefly described.

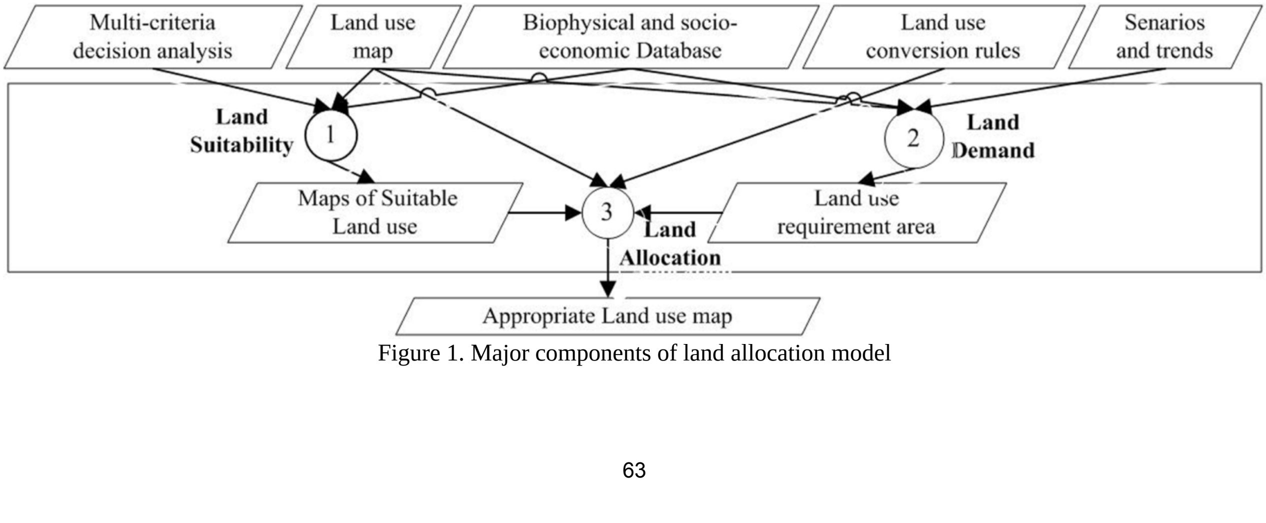

Multi-criteria decision analysis
Land use map
Biophysical and socio-economic Database
Land use conversion rules
Senarios and trends
Land Suitability
1
2
Land Demand
Maps of Suitable Land use
3
Land Allocation
Land use requirement area
Appropriate Land use map

Figure 1. Major components of land allocation model

3. LAND SUITABILITY ASSESSMENT MODEL

Study of land suitability concept in literatures shows consideration of series of environmental and socio-economic criterion. In this research major characteristics of land are considered to determine the land suitability of each piece of land "pixel" for different land use types. These are physical (intrinsic) suitability; accessibility to infrastructure, and major population and industrial centers; neighborhood effects and impacts and land use restrictions (Figure 2).

3.1 Physical suitability

Physical suitability is assessed by comparisons of the ecological land characteristics with the requirements of land use types "LUT" (F.A.O, 1976; Makhdoum, 1999; Ahamed et al., 2000; Kalogirou, 2002; Sante-Riveira et al., 2008, Zucca et.al., 2008). In this model this has been carried out using fuzzy spatial multiple criteria evaluation and applying Makhdoum's ecological capability model (Karimi et al., 2009).

3.2 Assessment of accessibility to infrastructures

In this research, accessibility to relevant infrastructures (road network, electricity transmission, gas pipelines and water canals) and major activity centers (major population and industrial center) are considered. As described by karimi et al., (2010a), accessibility is assessed in the following 3 steps. In the first step, accessibility to relevant infrastructures is estimated based on equation 1 (Engelen et al., 1997):

$$A_{\mathrm{ijk}} = \frac{1}{1 + D_{\mathrm{ij}} \big/ a_{\mathrm{jk}}}, \; j = 1,2,3,4 \quad (1)$$

Where A_{ijk} = the accessibility of pixel i, with land use type k, to the infrastructure j
Dij = the Euclidean distance of pixel i to the nearest pixel of the infrastructure j
a_{jk} = the relative importance of access of land use type k to the infrastructure j.

In the second step, accessibility of population centers to major activity centers is determined using gravity model (Geurs and van Wee, 2004), as presented in equation 2.

$$A_{\mathrm{g}} = (\sum_{k=1}^{K} P_k * e^{-\beta T_{gk}}) \quad (2)$$

Where A_g = the accessibility of population center g to all major activity centers
P_k = the importance of activity center k
T_{gk} = the distance of population center g to activity center k
β = adjustment parameter

In this equation, β plays a similar role as to a_{jk} in equation 1. The importance of population and industrial centers are assumed to be represented by their population and number of employee respectively. In most models, usually accessibility is defined using the concept of Euclidean distance. In this research, time is taken as the basis of calculating accessibility. Using the road networks and corresponding allowed speed, time is estimated and used for the assessment.

The resulted values of accessibility to activity centers are assigned to the population centers "points", in contrast to accessibility to linear infrastructure which is calculated for each pixel. To integrate these two, the point-based values of accessibility to activity centers are needed to be propagated to the regions. Therefore, thiessen Polygon analysis is used to assign the accessibility values to the regions surrounding population centers.

In the third step, overall accessibility is calculated using the Weighted Linear Combination (WLC) of accessibility to infrastructures and activity centers (equation 3).

$$A_{ia} = \sum_{j} w_{aj} * A_{iaj} + \sum_{k} w_{ak} * A_{iak} \quad (3)$$

where A_{ia} = the overall accessibility of pixel i with land use type a
A_{iaj} and $A_{iak\,=}$ the accessibilities of pixel i with land use type a, to infrastructure j and activity center k
w_{aj} and $w_{ak\,=}$ the relative importance of infrastructures and major activities in relation to land use type a

3.3 Land use interactions and neighborhood impacts

Land use interactions are usually modeled using cellular automata concept. Neighborhood effect for each cell is estimated through agglomeration of land use interactions effects of all adjacent cells, located in the influence region (White and Engelen, 2000; van Delden et al., 2007).

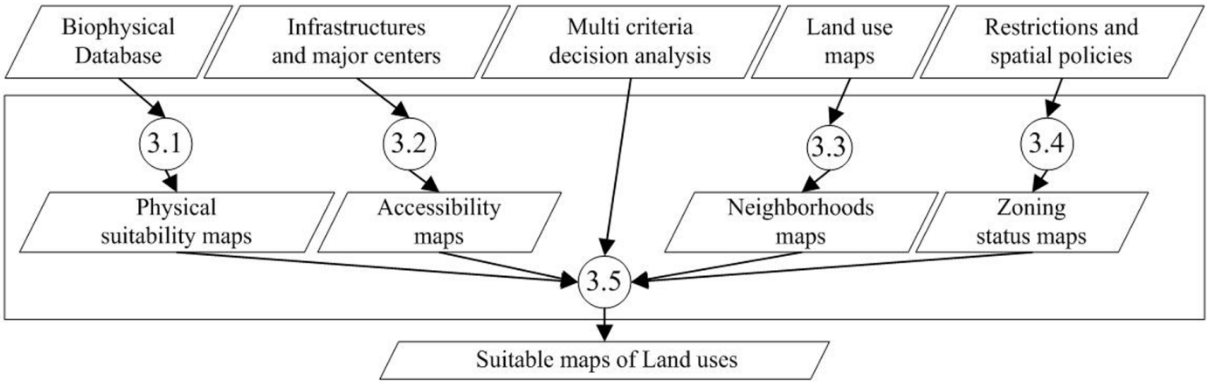

Biophysical Database
Infrastructures and major centers
Multi criteria decision analysis
Land use maps
Restrictions and spatial policies
3.1
3.2
3.3
3.4
Physical suitability maps
Accessibility maps
Neighborhoods maps
Zoning status maps
3.5
Suitable maps of Land uses

Figure 2. Processes of land suitability assessment

Land use interactions effects are usually represented as transition rules, showing the changes in interactions among land-use categories over distance. These interactions are usually called spatial externality. In generating transition rules, three main elements of influence radius, intensity, and distance decay should be defined (Hagoort et al., 2008). Land use interactions are conceptualized and modeled using linguistic variables, spatial metrics and expert knowledge. Various steps of this model are reported in (karimi et al, 2010b) and described briefly in the followings:

Modeling influence region: Land use interaction is mostly modeled considering eight neighboring cells, as consideration of more cells strongly increase the computation time. To overcome this limitation, hierarchical concept is developed by (van Vliet et al., 2009). As an improvement to that concept, in this research, a structure based on circular radius is proposed as a replacement for the square based structure of the hierarchical concept. This structure is more realistic, regarding the distance-based declining effect of neighborhood. In this representation, space is divided into circular levels, with distances flexible to the requirements of the experts.

Assessment of spatial externalities: The priced and un-priced radiated effects are assumed as a representation of the spatial externality intensity (Hagoort et al., 2008). These effects can be seen in land values. Moreover, often, land value is the main motivation of land use changes. Therefore, here, relative land values are determined using Analytical Hierarchical Process (AHP) (Saaty, 1980) and used as indicator of “intensity”.

The intensity of interaction between land use types usually declines over space. Yet, no mathematical or numerical methods have been developed for representation of these changes. This promotes the usage of expert knowledge. On the other hand, expression of expert knowledge using linguistic variables is easier and more straightforward. Linguistic variables of ‘very high’, ‘high’, ‘medium’, ‘low’ and ‘very low’ are considered for distance-decay. The spatial metrics used in this study is enrichment factor (Verburg et al. 2004), which is derived from land use types, extracted from the land use maps of the past two decades.

Finally, based on interpretation of spatial metrics by expert knowledge, distance decay of spatial externalities is assigned using linguistic variables (Table 1).

Table 1: Assessing the distance decay of spatial externalities

Land use type		Distance				
1	2	D1	D2	D3	D4	D5
U	U	+VH	+H	+M	+L	+L
I	U	-M	-L	-VL	0	+VL
R	U	+M	+M	+L	+L	+0
A	U	-VL	-VL	0	+VL	+VL
U	I	-L	-L	-VL	0	+VL
I	I	+VH	+H	+M	+L	+VL
R	I	-M	-L	-VL	0	0
A	I	-L	-L	-VL	0	0
U	R	-L	-L	+VL	+VL	+VL
I	R	-M	-L	-VL	+VL	+VL
R	R	+VH	+M	+VL	0	0
A	R	+M	+L	+L	+VL	+VL
U	A	-L	-VL	0	0	0
I	A	-M	-L	-VL	+VL	+VL
R	A	+H	+M	+VL	+VL	+VL
A	A	+VH	+H	+M	+VL	0

Spatial externalities of one land use on another in each distance or level of influence regions, is estimated by equation 4:

$$W_{k1,k2,r} = V_k * DD_{k1,k2,r} \quad (4)$$

Where $W_{k1\,,\,k2\,,\,r}$ = the spatial externalities of a cell with land use *k1*
$V_{k\,=}$ the relative land value of neighbor cells with land use *k2*
$DD_{k1,\ k2,\ r\ =}$ the distance decay of neighbor cells, located in the rth level

The quantification of linguistic variables is carried out using structured pair-wise comparison (Sharifi et al., 2006).

Classification of neighborhood effects: Intensity and distance decay for different land use types are different. LUI can be grouped into three classes. Spatial externalities of a land use type on itself, is defined as compactness. Positive and negative spatial externality of a land use type on a different land use type is defined as dependency and incompatibility respectively. Regarding the comparison of these classes, the following points should be mentioned. Distance decay of compactness is steeper than of dependency and incompatibility. Yet, the distance decay of compactness effect is different for various land use types. Some of the negative spatial externalities, decline over distance and finally converts to positive externality.

The classification of the surrounding cells, into the classes of cells with compactness, dependency and incompatibility effects is on the bases of the two land use types involved. The spatial externality of the surrounding cells falling in each class is aggregated separately. Finally, neighborhood effect for each cell is the weighted average of the aggregated effect of three classes, as presented in equation 5.

$$N_{\mathrm{il}} = w_{\mathrm{C}} * C_{\mathrm{il}} + w_{\mathrm{D}} * D_{\mathrm{il}} - w_{\mathrm{I}} * I_{\mathrm{il}} \quad (5)$$

Where N_{il} = Neighborhood effect for pixel *i* with land use *l*
C_{il} = Compactness for pixel *i* with land use *l*
D_{il} = Dependency for pixel *i* with land use *l*
I_{il} = Incompatibility for pixel *i* with land use *l*
w_C, w_D and w_I = the corresponding weights of compactness, dependency and incompatibility

Value of neighborhood effect which is estimated for various land use types does not have absolute meaning and can only be used for comparison. Therefore, these values are normalized.

3.4 Zoning and committed land

Land use restrictions have a deterministic effect on the pattern of land use change. When determining the areas with unchangeable land use, spatial policy regulations, committed land and environmental and socio-economic hazards are needed to be considered. Some spatial policies, such as those regarding environmentally protected areas, restrict all possible land use changes. Others might restrict only a limited set of land use conversions. In addition, status of restriction may be altered during time and even be enforced, more moderately, as a weighted indicator.

3.5 Integrated assessment of land suitability

In order to determine the overall neighborhood effects, various effects should be integrated. In this research, an additive and multiplicative weighted average is used to integrate the above mentioned factors, as presented in equation 6.

$$P_{k,j} = \left(A_{k,j}\right)^{w_a} * \left(S_{k,j}\right)^{w_s} * \left(N'_{k,j}\right)^{w_n} * \left(Z_{k,j}\right) \qquad (6)$$

Where P = the overall suitability for cell j with land use k
S = the physical suitability for cell j with land use k
A = the accessibility for cell j with land use k
N = the neighborhood effect for cell j with land use k
w_S, w_A, w_N, and w_Z = the relative importance of those parameters is represented by weights

4. LAND DEMAND ASSESSMENT

Demand for each land use type at each time step is estimated for each Land Demand Unit (LDU). Usually, two policies are considered in deciding on LDU. In the first policy, entire region is assumed as one large LDU, for which, the required land use types areas are estimated. As a result of this policy, the required land use types are mostly allocated around the major activity centers; this can be called agglomerated allocation. According to the second policy, the region is divided into a number of LDUs (small administration units) and the demand for each separate LDU is estimated. In such a distributed approach, the demand and allocation will be distributed over the LDU's. The non-linear historical model based on past trends, as presented in equation 7 is used for calculation of annual demands, in each LDU.

$$A_n = A_0 * (1 + r)^n \qquad (7)$$

Where A_n = the required area for the nth year
A_0 = the area in the base year
r = the growth rate of land demand

Using the available land use maps, the A_0 and A_n, which are the land use areas in years 1986 and 1998, for example, and having the time step as 12, the r can be calculated. Later, using r value, the land demands for each year can be calculated.

5. LAND USE ALLOCATION

In land use allocation, beside land suitability and land demand, both the current land use/cover, and land use conversion rules and policies are playing a very important role. Changes of various land use types have diverse environmental and socio-economic effects. Conversion of land use types with high investment into other land uses is often very costly or even impossible. Usually possibility and degree of difficulty of land use conversion are quantified through structure pair-wise comparisons by experts. This concept of land use change difficulty can be integrated with the concept of land suitability. The result of such integration shows the total potential of a pixel with a present land use to change to another specific land use type. Therefore, the potential of each pixel to change to another land use is estimated based on equation 8.

$$\mathrm{TP}_{k,j} = P_{k,j} * CM_{e,k} \qquad (8)$$

Where $TP_{k,j =}$ total potential of cell j with land use k
$P_{k,j}$ = overall suitability of cell j with land use k
$CM_{e,k}$ = the degree of difficulty regarding change of land use type from e to k.

In present land allocation models, first, initial land uses are allocated using maximum suitability values only. Then, some of the initially allocated land uses are changed to balance land suitability and satisfied land demands. This process is called land demand adjustment (White and Engelen, 2000) or iteration variable (Verburg et al. 2002). In this research, a new pixel-based stepwise procedure for land use allocation is proposed (karimi et al, 2010a), which is shown in Figure 3.

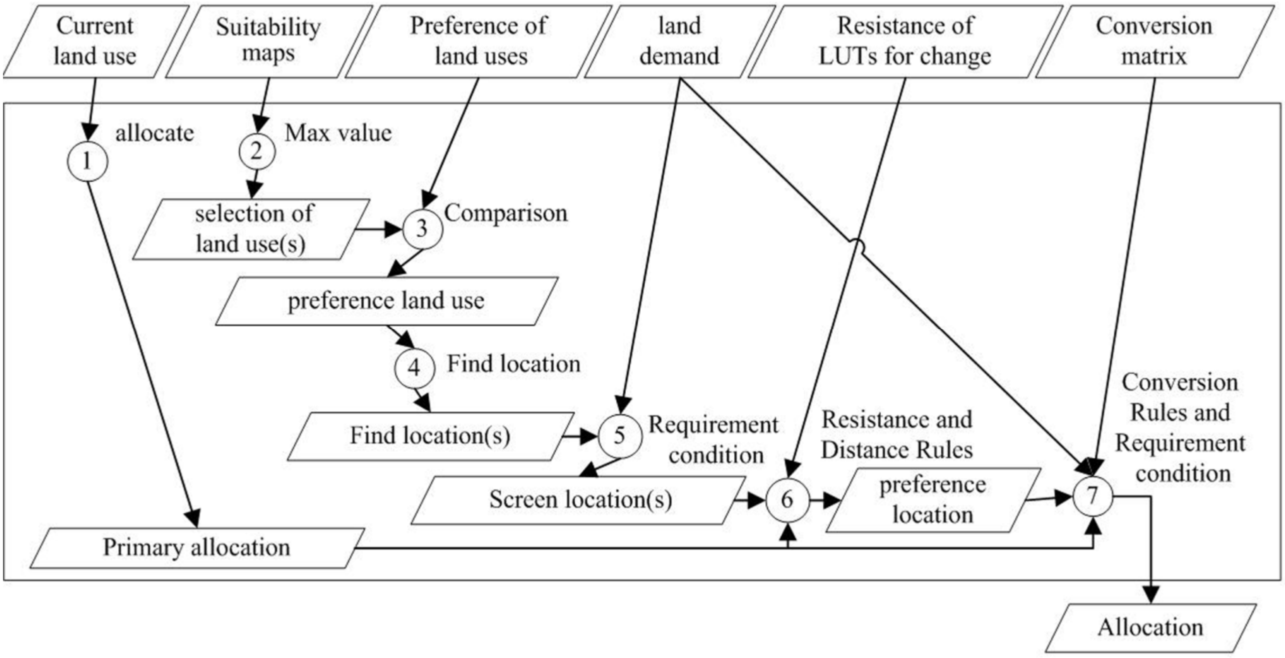


Figure 3. Land allocation process

6. IMPLEMENTATION AND RESULTS

6.1 Case study area

The study area is Borkhar and Meymeh Township located in the center and northwest of Esfahan province, Iran (Figure 4). It consists of 6 districts (Dehestans), 9 cities and 28 residential villages. Approximately 86% of the population lives in urban areas. In this region, mean annual population growth was 2.58% from 1986 until 1996 and 2.39% from 1996 until 2006, which shows high concentration and population growth.

The available date of the study area was studied and the periods of 1986-1998 and 1998-2005 were selected for the implementation and test of the model. There was no suitable land use map for the study area related to those time steps. Therefore, such maps were created by supervised classification of the Landsat images for years 1986 and 1998, and the ASTER images for the year 2005. The resulted land use maps consists of six classes of urban residential, rural residential, industry, agriculture, pasture and others.

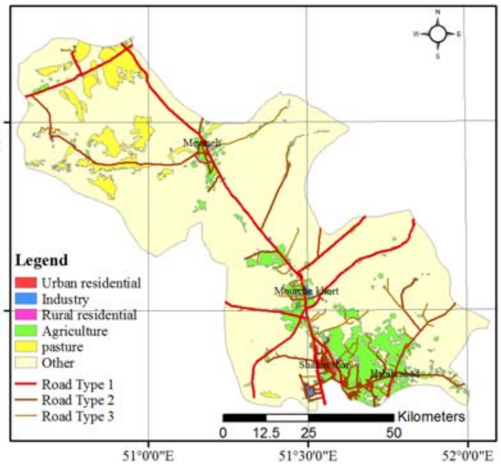


Figure 4: General map of the case study area

6.2 Implementation and results

The conceptual model was implemented in the case study area. In this process, the corresponding maps of physical suitability, accessibility, neighborhood effects, committed lands, demand and current land use map of the area were prepared in 1/250,000 scale. The demands for different land uses, was derived for the periods of 1986-1998 and 1998-2005. Next the simulated land use maps of the years 1998 and 2005 were generated, using the developed allocation algorithm (Figure 5). For evaluation of the results, the real land use maps of these years were also extracted from the relevant satellite images.

In LUC modeling, usually, three processes of calibration, evaluation and prediction are performed. In fact, the processes of calibration and evaluation are to ensure the quality of the prediction. These two processes were carried out using real data and expert knowledge. The calibration parameters extracted from expert knowledge were used in the model to derive the simulated map of future land use (Figure 5). Such map was then evaluated by its comparison with the actual land use map using Kappa coefficient. In addition, the Kappa coefficient for a randomly created land use map, called RCM, was also calculated. Those coefficients are compared in Table 2. This comparison shows the amount of similarity between the simulated map and the real land use map.

Table 2. Comparison of kappa coefficient for 1986-1998 (U: Urban residential, I: Industry, R: Rural residential, A: Agriculture)

	Total	U	I	R	A
RCM map	0.810	0.832	0.641	0.825	0.846
Simulated map	0.902	0.864	0.734	0.863	0.858
Difference	0.092	0.032	0.093	0.038	0.012

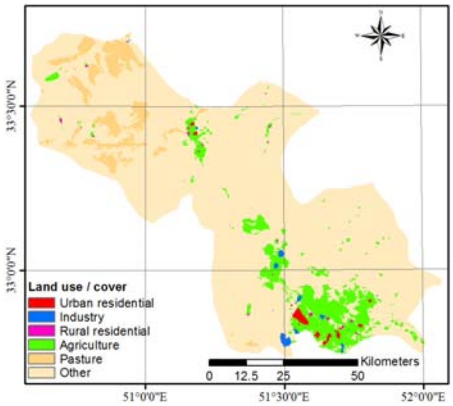


Figure 5: model-extracted land use map in 1998

A general idea about of the potential values of the areas with LUC can be conceived much better using a visual representation. In other words, the occurred LUC can be analyzed spatially, from the point of view of affecting factors. As an example, the total suitability of the areas changed during 1986-1998 to the land use types of agriculture are shown in Figure 6.

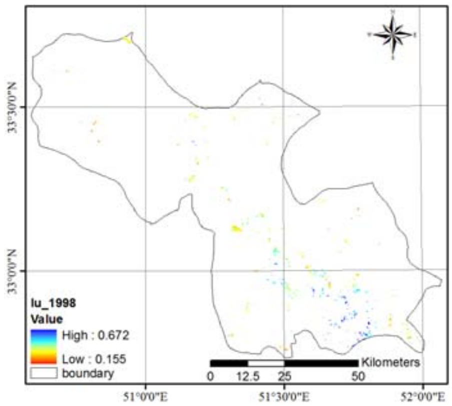


Figure 6: potential of occurred LUCs in 1988 for agriculture

7. CONCLUSIONS AND DISCUSSIONS

As a result of this research, a new method for LUC modeling in regional level is developed, tested and evaluated using the data of Borkhar and Meymeh Township. The presented conceptual model, which is developed based on cellular automata concept, contains the following novelties:

- One of the challenging issues in modeling land allocation is to calculate the land use interaction (neighborhood effect). In this research, neighborhood effect is broken down in to three components namely incompatibility, dependency and compactness. These components are modeled separately and combined in a multi-criteria decision analysis model, which leads to a new method for assessing the land use interaction.

- In the existing land allocation models calibration of parameters related to neighborhood effect are set through a trial & error process, which is considered disadvantage. In this research, attempt is made to estimate simultaneously the spatial interaction of land uses, accessibility to infrastructure, accessibility to major centers and physical suitability parameters based on heuristics of land use change and expert knowledge formulated through linguistic variables.
- In the physical suitability assessment model, fuzzy inference rules is implemented in the process of integrating various environmental layers (spatial multi criteria evaluation)
- In literature, usually accessibility to transportation network alone is considered. In this research, accessibility to all the relevant infrastructures is considered in the process of land suitability assessment. Infrastructure systems include transportation, electric power lines, gas pipelines and water canals.
- In this research, accessibility to major populated and industrial locations are considered and added to infrastructural accessibility.
- Integrated assessment of Land suitability is carried out using multi-criteria decision analysis.
- In this research a new allocation algorithm is developed based on land suitability, land demand, the current land use/cover, land use conversion rules, hubristic rules of land use interactions and land use change.

ACKNOWLEDGEMENTS

Authors would also like to acknowledge the research opportunity provided by the faculty of Geo-Information Science and Earth Observation (ITC), University of Tewente, The Netherlands. In addition, the authors would like to thank Esfahan province governance office for the providence of the research data.

REFERENCES

Ahamed, T.R.N., Rao, K.G., Murthy, J.S.R., 2000. GIS-based fuzzy membership model for crop-land suitability analysis. Agricultural Systems. 63, 75-95

Engelen, G., White, R., Uljee, I., 1997. Integrating constrained cellular automata models, GIS and decision support tools for urban planning and policy making. In: Timmermans, H. (Ed.), Decision Support Systems in Urban Planning. Chapman and Hall.

Engelen, G., Lavalle, C., Barredo, J.I., van der Meulen, M., White, R., 2007. The Moland modelling framework for urban and regional land-use dynamics, E. Koomen et al. (eds.), Modelling Land-Use Change, 297–319

F. A. O., 1976. A framework for land evaluation. F.A.O soils bulletin. pb No 32. Rome.

Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport Geography. 12, 127–140.

Hagoort, M., Geertman, S., Otters, H., 2008. Spatial externalities, neighborhood rules and CA land-use modeling. Ann Reg Sci. 42,39–56

Kalogirou, S., 2002. Expert systems and GIS: an application of land suitability evaluation. Computers, Environment and Urban Systems. 26, 89–112

Karimi, M., Mesgari, M.S., Sharifi., 2009. Modelling ecological capability using fuzzy logic (case study area: Borkhar and Meymeh township). Journal of Iranian GIS and RS Society, 1, 1-24 (in Persian)

Karimi, M., Mesgari, M.S., Scarify, M.A., 2010a. Modeling land use change in space and time in Borkhar and Meymeh Township, Iran. Landscape and urban planning (forthcoming)

Karimi, M., Sharifi, M.A., Mesgari, M.S., 2010b. Modeling land use interaction using linguistic variables. International Journal of Applied Earth Observation and Geoinformation (forthcoming)

Makhdoum, M.F., 1999. Fundamental of Land use Planning. publication of Tehran University. Third edition. (in Persian).

Saaty, T.L., 1980. The Analytical Hierarchy Process. McGraw-Hill, NewYork.

Sante-Riveira, I., Crecente-Maseda, R., Miranda-Barros, D., 2008. GIS-based planning support system for rural land-use Allocation, computers and electronics in agriculture. doi:10.1016/j.compag.2008.03.007

Sharifi, M.A, Boerboom, L., Shamsudin, K.B., Veeramuthu, L. 2006. Spatial multiple criteria decision analysis in integrated planning for public transport and land use development study in Klang valley, Malaysia, Proc. of the ISPRS Technical Commission II Symposium, Vienna.

van Delden, H., Luja, P., Engelen, G., 2007. Integration of Multi-Scale Dynamic Spatial Models of Socio-Economic and Physical Processes for River Basin Management. Environmental Modelling & Software. 22, 223-238

van Vliet, J., White, R., Dragicevic, S., 2009. Modelling urban growth using a variable grid cellular automaton. Computers, Environment and Urban Systems 33, 35–43

Verburg, P.H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., Mastura, S.S.A., 2002. Modelling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. Environmental Management. 30 (3), 391–405

Verburg, P.H., de Nijs, T.C.M., van Eck, J.R., Visser, H., de Jong, K., 2004, A method to analyze neighborhood characteristics of land use patterns, Computers, Environment and Urban Systems. 28, 667–690.

White, R., Engelen, G., 2000. High-resolution integrated modelling of the spatial dynamics of urban and regional systems. Computers, Environment and Urban Systems. 24, 235-246

Zucca, A., Sharifi, M.A., and Fabbri, A, (2008). “Application of spatial multi-criteria analysis in site selection for Local Park: a case study in the Bergamo Province, Italy”. Journal of Environmental Management “YJEMA 1645” pp. 1-18.

A SPATIO-TEMPORAL POPULATION MODEL FOR ALARMING, SITUATIONAL PICTURE AND WARNING SYSTEM

Z. Zhang *, R. Sunila, K. Virrantaus

Department of Surveying, Faculty of Engineering and Architecture, School of Science and Technology, Aalto University
zhangzhe@cc.hut.fi

KEY WORDS: Population model, spatio-temporal model, population estimation methods, spatio-temporal object, object-oriented model.

ABSTRACT:

Natural and man-made hazards, disasters, and concerns such as the increasing amount of terrorism, cause insecurity for people and society. Therefore, crisis management with respect to urban structure becomes one of the core tasks of governments. Population information is essential during the rescue planning process. Population density can be estimated by using various statistical methods. Often it is impossible to estimate the size of the population of a certain area. Most population distribution models do not take into consideration the temporal variation in population location; this causes poor estimation results. In this paper, we present a spatio-temporal population model, which is used for calculating the number of people in a certain area at a certain time. The model combine object-oriented spatio-temporal model with users´ knowledge. Population size is estimated on the basis of built-up environments in the real world. Built-up environments such as buildings and roads are divided into different categories; each category is one ***spatio-temporal object***. The model produces reasonably accurate results since each spatio-temporal object is modelled separately according to its own characteristics. The model is also flexible, as most of the estimation methods are based on user knowledge. The model is simulated using Java programming language. As a result, the model produces a text file, which can easily be adapted to most information systems.

1. INTRODUCTION

This research work is part of the "Alarm, Situational Picture and Warning System for Chemical, Biological, Radiological, and Nuclear (CBRN) and the Natural Disaster Incidents "(UHHA) project (Molarius et al., 2009). Several other organisations, such as Valtion Teknillinen Tutkimuskeskus (VTT), the Finnish Meteorological Institute (FMI), Insta Ltd, Helsinki University and Finnish Chemicals Ltd., are involved in this UHHA project, the goal of which is to set up an alarm, warning, and information exchange system that rescue personnel can use in the event of an emergency. A chemical accident in Kuusankoski city is selected as a study case for this research project. Finnish Chemicals Ltd. is a company that produces hazardous chemicals. It is located in Kuusankoski in the south-eastern part of Finland. To cope with an event such as a chemical leakage accident, a dispersion model has been created by FMI. This dispersion model can define the risk area based on an evaluation of the releases, source terms, and atmospheric dispersion of hazardous chemicals (Finnish Meteorological Institute). This risk area is the location input to the spatio-temporal population model. The spatio-temporal population model is used to estimate the exact number of people inside the risk area at a specific time. The UHHA server is the main UHHA user interface and was implemented by VTT (Molarius et al., 2009). It is a web user interface that requests population information from the population model and sends the population model results to the interorganisational Crisis Manager System (iCM) (Insta Oy, 1997). The Insta iCM System is designed to support decision making and the coordination of operations between crisis management organisations, both in international and domestic crisis situations (Insta Oy, 1997). Figure 1 shows the whole system architecture of the UHHA project.

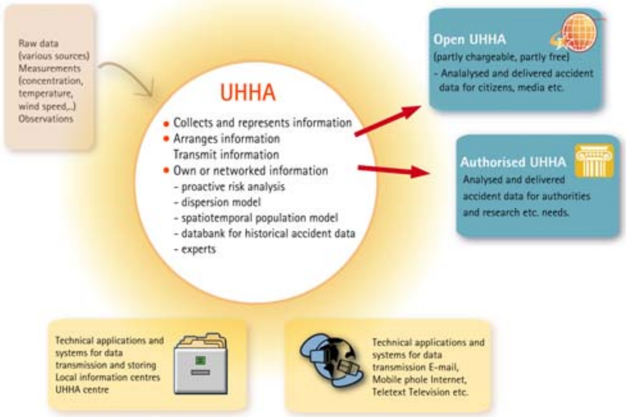

Raw data
(various sources)
Measurements
(concentration,
temperature,
wind speed,..)
Observations
UHHA
Collects and represents information
Arranges information
Transmit information
Own or networked information
- proactive risk analysis
- dispersion model
- spatiotemporal population model
- databank for historical accident data
- experts
Open UHHA
(partly chargeable, partly free)
- Analalysed and delivered accident data for citizens, media etc.
Authorised UHHA
Analysed and delivered accident data for authorities and research etc. needs.
Technical applications and systems for data transmission and storing Local information centres UHHA centre
Technical applications and systems for data transmission E-mail, Mobile phole Internet, Teletext Television etc.

Figure 1. UHHA project architecture (Molarius et al., 2009).

Ahola combined Yuan´s three-domain model (Yuan, 1996) and Langran and Chrisman´s (Langran, 1992) space-time composite model to create a spatio-temporal population model to support risk assessment and damage analysis (Ahola, 2007). The accuracy of the model is low since it estimates people’s location based on their occupation information. Time is stored as an attribute in the database, which causes a lack of flexibility. The data from SeutuCD are used as the main location and attributes input dataset, which only exists in Helsinki Metropolitan Area. All these drawbacks indicate that we need a spatio-temporal population model that can be applied to any municipality of Finland and that will produce results that are more accurate.

In addition to Ahola´s idea, The population density maps, which represent the average population density for a statistical unit, for instance city or country, can be considered the simplest models of population density distribution (Sweitzer and Langaas,1994). Longley has introduced a population density estimation method by using kernel functions (Longley, 2005). Liu introduced a regression and area to point residual kriging method to improve the population density interpolation accuracy (Liu et al., 2008).

The main emphasis of this paper is on developing a spatio-temporal population model, prototyping it by using programming language and examining how the model can be used in the application in question. The population model has been designed to be generic, so it could be used not only in the city in question but also in any Finnish municipality for crisis management.

2. BACKGROUND AND THEORY

2.1 Spatio-temporal models

A model is an approximation and simplification of reality. Several spatio-temporal modelling methods have been developed during the past 10 to 15 years. El-Geresy identified conceptual modelling of spatio-temporal domains and classified them into several categories: location-based models, time-based models, event-based models, object- or feature-based models, process-oriented models and causal models (El-Geresy, 2002). In a location-based model, the space is divided into locations by means of a grid; for each location, the changes are recorded in a list representing successive changes in the features of that specific location, together with their corresponding time (El-Geresy, 2002). One example of a location-based model is Langran’s temporal raster data model (Langran, 1992). Time-based models include Armstrong´s ***snapshot models*** (Armstrong, 1988; Hunter and Williamson, 1990), and ***space-time cube models*** (Hägerstrand, 1970). Peuquet (Peuquet and Duan, 1995) introduced an event-based model that deals with abstracted relations based on event. An object- or feature-based model is an extension of a vector model; these include Langran’s ***base state with amendment model*** (Langran, 1992), Langran and Chrisman’s ***space-time composite model*** (Langran, 1992), and Worboys’ ***spatio-temporal object model*** (Worboys, 1992). Process-oriented models classify the spatial relations into specific processes (El-Geresy, 2002). ***Causal models*** specify the temporal relation in terms of cause and effect (El-Geresy, 2002).

In addition to the above-mentioned spatio-temporal models, Yuan presented a three-domain model (Yuan, 1996). In his three-domain model, semantic, temporal, and spatial objects are defined as three separate domains and time is modelled as an independent concept. Geographic concepts and entities are represented by linking these three types of objects from the object perspective or layer dynamically (Yuan, 1996).

2.2 Object-oriented spatio-temporal data model

2.2.1 Definition of spatial object

Before talking about object-oriented data modelling, the definition of ***object*** has to be clear. From the static point of view, an object is a collection of named ***attributes***, each of which takes a value from a specified domain (Worboys, 1994). From the dynamic or behavioural point of view, an object can perform a set of operations under appropriate conditions

(Worboys, 1994). Therefore, Worboys has defined an object as state (static) plus functionality (dynamic) (Worboys, 1994). ***Object types*** that have similar behaviours form object groups. The objects associated with an object type are called ***occurrences*** (Worboys et al., 1994). Mattos et al. state that an object must be identifiable, relevant and describable (Mattos et al., 1993).

Spatial dimensions have always played an important role in GIS research. ***Spatial object*** refers to the object that contains a spatial domain. Spatial object in general can be divided into discrete such like point, lines or field model that represents the spatial phenomena as continuous surface or layer (Zhang an d Goodchild, 2002)There are several dimensions in which spatial object attributes may be measured. These include spatial, graphical, temporal and textual/numeric (Worboys, 1994). For instance, a road network is a spatial object as it has its coordinates in the real world (spatial), a line representing its cartographic form at a different level of generalization (graphical), times when it was created in the real world and system (temporal), and attributes describing the length of street address (textual/numeric).

2.2.2 Temporal objects

Traditional Geographic Information System (GIS) applications describe reality in a static manner; the time dimension is not taken into account. In real life, most phenomena change with time and because of this temporal information has played an important role in the GIS modelling process. Time can be divided into two dimensions in information systems (Worboys, 1994). ***Database time*** refers to the time when transactions take place in information systems, such as the time when data is stored in the database. It is also called ***system*** or ***transaction time***. ***Event time*** refers to the real-world or valid time when events actually occur in the application domain.

During the spatio-temporal modelling process, spatio-temporal data is the first core element that needs to be considered. Based on the above defined time dimensions and duration of the time, dynamic data can be divided into ***real-time data***, ***near-real-time data***, and ***time-stamped data*** (Nadi and Delavar, 2003). ***Real-time data*** refers to those geospatial data, such as traffic volume, that are collected and imported to GIS as soon as an event occurs. ***Near-real-time data*** are related to the data that need updating at any given moment and also need visualization and analysis before they can be used in GIS. ***Time-stamped data*** refers to data that have a time attribute attached to them (Nadi and Delavar, 2003).

2.2.3 Object-oriented spatio-temporal model

In the object-oriented model, the information space is decomposed into objects. Geographical information can be modelled by two classes of model, i.e. the ***fielding-based*** model and ***entity-based*** model. In the class of ***entity-based*** models, the information space is treated as populated by discrete, identifiable entities (or objects) with a geo-reference (Worboys, 1994). The association between entities is called ***relationship***. Worboys has extended the original object-oriented model in the temporal dimension and so developed the object-oriented spatio-temporal model (Worboys, 1992). This model represents the world as a set of discrete objects consisting of spatio-temporal atoms by incorporating a temporal dimension orthogonal into the 2D space (Worboys, 1992). Each spatio-temporal atom contains changes in both space and time, but no changes are recorded between them (Worboys, 1992). The object-oriented spatio-temporal model is suitable for modelling the complex phenomena since it breaks through the limitation of relational form and directly captures the in-depth semantics of the application domain.

2.3 Combine object-oriented spatio-temporal modelling with users´ knowledge.

Based on Worboys' idea, a spatio-temporal population model can be created by combining object-oriented spatio-temporal model and users´ knowledge. Figure 2 illustrates the modelling structure. This model is not concerned with how the spatial object's static attributes change according to the temporal dimension, which means spatial objects that presented in this model are stable (not time dependent), but the knowledge that gain from the entity relationship with the spatial object is time dependent. The temporal attribute is not necessarily stored in the database. For instance, this spatio-temporal population modelling started with an entity-based model. People stay inside spatial objects (residential buildings) at certain times. A relationship is an association between entities, for example "stay inside". In this case, the temporal dimension of spatial object refers to the time when there are people inside. Population size can be gained by using the attribute values of the residential buildings together with this entity relationship and users´ knowledge. This knowledge is updated according to time. For instance, the estimated average population size for a residential building during the morning hours was 10 since most people leave to go work or school. At night, most people stay at home, so estimated population size changed to 50. Further detailed population size estimation methods will be introduced in Section 3.2. Smith and Smith have proposed aggregation and generalization abstraction constructs that were used in this model as well (Smith and Smith, 1997). Worboys has extended these abstraction constructs to generalization, specialization, aggregation, association, ordered association and polymorphism (Worboys et al., 1990). Generalization constructs enable groups of entities of similar types to be considered as a single higher-order type (Worboys et al., 1990). For instance, all the major roads, minor roads or walking paths are grouped into one spatio-temporal object called ***roads***. For each spatio-temporal object, the population size changes individually according to time.

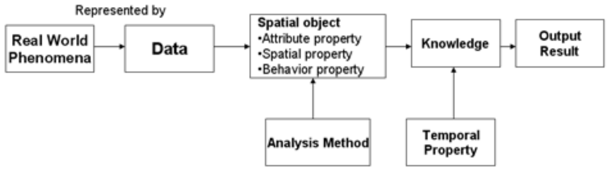


Figure 2. Combine object-oriented spatio-temporal model with users´knowledge .

3. MATERIAL AND METHODS

3.1 Data processing

The dataset used in spatio-temporal population modelling comprised Finnish Base Register data, National road and street database (Digiroad), the Topographic Database, and a dispersion model.

In Finland, every municipality is responsible for collecting its regional land information data. The law relating to the Finnish Base Register data is derived from the Population Information Act and the Population Information Decree. One dataset used in this research project is part of the Kuusankoski Base Register data, called the Building Information System. It was in Excel format. This dataset was used to estimate size of population inside buildings in the risk area.

Another dataset comprised Digiroad data, which can be used for modelling road networks. It was in vector format. Digiroad data is a national database that describes the geometry and physical features of roads and streets in Finland. The Finnish Road Administration is responsible for maintaining and updating the data.

The Topographic Database maintained by the National Land Survey was chosen as the background map. In addition, a dispersion model was created by experts from the Finnish Meteorological Institute. It was delivered in a Shape file format.

The Kuusankoski Building Information System and Digiroad data were first converted into Shape files by using ArcMap. A dispersion model was used to represent the risk area and added to the ArcMap as a separate layer. Within the risk area, the Shape file dataset (Kuusankoski Building Information System data and Digiroad data) was first selected and then imported to separate Geodatabase files. Figure 3 illustrate the dataset attributes and data processing.

3.2 Methods

Object-oriented spatio-temporal model combine with users´knowledge is chosen for implementing this spatio-temporal population model. For the case of Kuusankoski, eleven spatio-temporal objects, namely roads, office buildings, old people's homes, day care centres, residential buildings, shops, hospitals, hotels, schools, industrial buildings, and restaurants and bars, are defined. For each spatio-temporal object, the population size changes individually according to time.

3.2.1 Modelling roads and residential buildings

The average number of people on the road network inside the risk area was calculated by summing together the "Traffic density" values of different road segments in Road Geodatabase file (see figure 3). Traffic density attribute refers to how many people pass by certain road segment in one minute. A similar method can also be used to estimate population size in residential buildings. The maximum number of people inside the residential buildings was calculated by summing together the "Total amount of inhabitants" values in Buildings Geodatabase (see figure 3). Total amount of inhabitants attribute refers to how many people registered this building as their home address.

3.2.2 Modelling office buildings, old people´s homes, day care centres, restaurants and bars

For office buildings, old people's homes, day care centres, and restaurants, it was assumed that the average floor area used by one person is a constant. In Buildings Geodatabase file, there is an attribute value representing the floor area of the building. Therefore, the maximum number of people inside the building can be estimated on the basis of the total floor area of the building divided by the average floor area that one person uses. The floor area used by one person is estimated on the basis of the user's knowledge; therefore, this spatio-temporal population is also a knowledge-based model.

Population size = (Total floor area of building) / (average floor area that one person uses) (1)

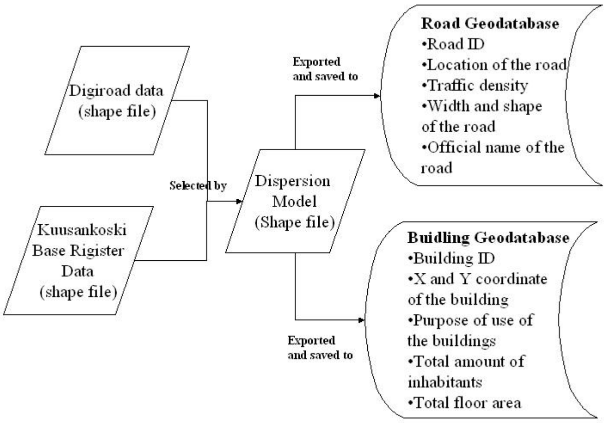

Digiroad data (shape file)
Kuusankoski Base Rigister Data (shape file)
Selected by
Dispersion Model (Shape file)
Exported and saved to
Road Geodatabase
•Road ID
•Location of the road
•Traffic density
•Width and shape of the road
•Official name of the road
Exported and saved to
Buidling Geodatabase
•Building ID
•X and Y coordinate of the building
•Purpose of use of the buildings
•Total amount of inhabitants
•Total floor area

Figure 3. Data processing.

3.2.3 Modelling hospitals, schools, hotels, and industrial buildings

For these buildings, the relationship between the population size and the floor area is not linear. In Finland, the number of such buildings is limited due to their purpose or use. The X and Y coordinates of the buildings can be found in Buildings Geodatabase file; ArcGIS address locator can be used to find a building´s address on the background map. After that, an organisation's name can be defined according to its address. The population size for the organisation can be found through, for example, the organisation's homepage. This method might be time-consuming, but it works for most Finnish municipalities because there are not an excessive number of hospitals, schools, hotels and industrial buildings inside any particular limited area. Although it is time-consuming, this method can nevertheless give quite accurate estimation results.

3.2.4 Modelling shops

In the Building Geodatabase file, the floor area for shops can be found; this varies from 200 square metres to 4000 square metres. Shops can be grouped into four categories according to their floor area. Average number of customers for different shop categories can be estimated based on statistical material or user´s knowledge and saved in Buildings Geodatabase file with attribute name "***Customers***". The number of people inside shops in the risk area was calculated by summing the number of "***Customers***" values together in the Buildings Geodatabase file. Table 1 shows the shop types and their corresponding floor area and number of customers.

Shop type	Floor area (m^2)	Estimated number of customers
Small food shop	200-500	40
Small supermarket	700-2000	130
Small shopping centre	2000-4000	300
Big shopping centre	4000-	500

Table 1. Estimated number of customers for different types of shops.

3.2.5 Modelling the temporal aspect

The temporal aspect is taken into account by assigning weights to the estimated maximum or average number of people inside each spatio-temporal object. The weights vary according to time. Table 2 shows an example of assigning weight to estimate the average number of people using road networks and the maximum number of people inside shops according to time. For instance, on weekdays between 7 and 10 o'clock, the average number of people using selected roads was 500, whereas in reality the number was greater because at that time most people were travelling to work. The weight assigned to this population size was 3, which means the estimate of population size in the road network at this time was approximately 3 times 500, which equals 1500. In Finland, most shops open at 10 o'clock, so the estimated weight of population size in shops on weekdays during the period 7 to 10 o'clock was only 1/10(some shop workers come to work before 10 o´clock).

	Weekday 07.00-10.00	Weekday 10.00-16.00	……	Weekday 19.00-22.00
Roads	3	1/2		1/3
Shops	1/10	2/3		2/3
……				

Table 2 Estimate of the weight of population size using roads and in shops according to time.

Office buildings, day care centres, schools, and industrial buildings have regular opening and closing hours. Estimate of the weight of population size inside these buildings is zero outside the opening hours and one inside the opening hours. Interviews and some other statistical materials are needed to compute people´s daily travelling habits, such as the time that people usually travel to work, back home and have a meal. This information can be used to estimate the weight of the population size inside residential buildings, hotels, old people's homes, restaurant and bars. For instance, estimate of the weight of population size inside a restaurant is one or at least ½ during the lunch time. It is also possible to get statistical information about the number of visitors during different time periods in big shopping centres in Finland. This information can be used to estimate the weight of population size inside shops. Hospitals have regular opening and closing hours. Therefore, population size decreases outside the opening hours, but not to zero because some patients and staff remain in the hospitals during the closing hours. Estimating the weight of population size for the roads is problematic. The population size on the roads increases during the time when people travel to work and back to home. However, it is difficult to give the exact weight of population size during different time periods on the road network and this should be included in the future development plan.

4. RESULTS

This spatio-temporal population model was implemented by using Java programming language on ESRI GIS platform. Figure 4 shows the UML representation of the software architecture. Figure 5 shows the user interface. In the upper part of window, there are lists of all the tools that help the user to interact with the background map. After the user has uploaded the map, different map layers can be turned on or off by clicking the on/off check box. Figure 5 shows the roads, buildings and buffer-layer check boxes were turned on. The buffer layer was produced by the Finnish Meteorological Institute. It includes four zones, the smallest of which has the highest level of pollution. The legend button is used to show the legend of the selected layer. Time was grouped into 17 categories and listed in a Set Time combo box. The time was specified as hours, so the user can easily choose the time category according to accident time.

Figure 6 shows that a new window, 'Buffer Analysis Results', pops up after the user has clicked the 'Analysis' button. It shows all the spatio-temporal objects defined in Section 3. For each spatio-temporal object, a different estimation method is used for estimating the population size. The user can click the 'Show Result' button to estimate the population size for the corresponding spatio-temporal object. Figure 6 illustrates the simulation of population size estimation methods for office buildings. After the user has clicked on the office building's 'Show Result' button, a dialog box pops up that shows the total floor area of the selected office building was 348 square metres. In the dialog box, the user can set a parameter that is used to define how many square metres one worker uses. In this case it was 23. The result, which was called 'Count People', was calculated by using the total floor area of the selected building divided by the parameter that the user set, which gave the result as 15. After that, a weight 2/3 was given according to time. The final estimation of the population size was 15 times 2/3, i.e. 10. A similar user interface and method was also used for old people's homes, day care centres, restaurants, and bars.

For roads, hospitals, hotels, schools, industrial buildings, restaurants and shops, the estimated population size for the corresponding buildings will appear in the corresponding text field after the user clicks on each 'Show Result' button. The methods introduced in Section 3 were used, so the results came by querying Geodatabase files and summing the corresponding attribute value together. The weight for different spatio-temporal objects was added after the query process.

The 'Total' button was used to sum all the spatio-temporal objects analysis results together. In this case, estimated population size inside the risk area is 3164. The user can use the 'Report' button to create a text file of analysis results. The 'Save' button is used to save the report.

accuracy. As a result, the model will produce a text file, which can easily be adapted to other systems.

5. CONCLUSION

The model is flexible, because most of the estimation methods are based on the user's knowledge. The user can update his/her knowledge frequently in order to produce more accurate results. The model also gives reasonably accurate results. Each spatio-temporal object is modelled individually according to the character of the object. Especially in the case of industrial buildings, hospitals, schools, and hotels, the population size information comes from the web pages of the organisation concerned, which have information with a very high level of

One expert is needed to process the data and set up the software environment before the model can be used. He/she should be familiar with the Java programming language and basic GIS tools, such as ArcGIS and ArcEngine API. Some real-world objects, such as places for hobbies and leisure, fire stations, police stations, sports centres etc., are ignored in this model. The reasons for this are a lack of data and, in some cases, the rather small population size inside these buildings. There are no data available for modelling forests, water regions, and airplanes in this project. This will be added to the future development list and implemented in the next version.

Because of the time limitation, only a text file format result is achieved. With a little more effort, a Shape file could also be produced as an output of the model. Buildings and roads were selected manually by using ArcMap selection with the 'location' tool. In the future, it will also be possible to add a selection function to the model so that the selection can be made automatically after the buffer layer has been turned on.

The uncertainty of the spatio-temporal population model is not discussed in this paper. Uncertainty may be caused by the lack of data of locations, incorrect attribute data and incorrect weight of population size. Model can be calibrated by doing interviews and collecting statistical data about people´s daily habits in order to get more correct weighting of population size for each spatio-temporal object. Data quality should also be concerned in the future development work.

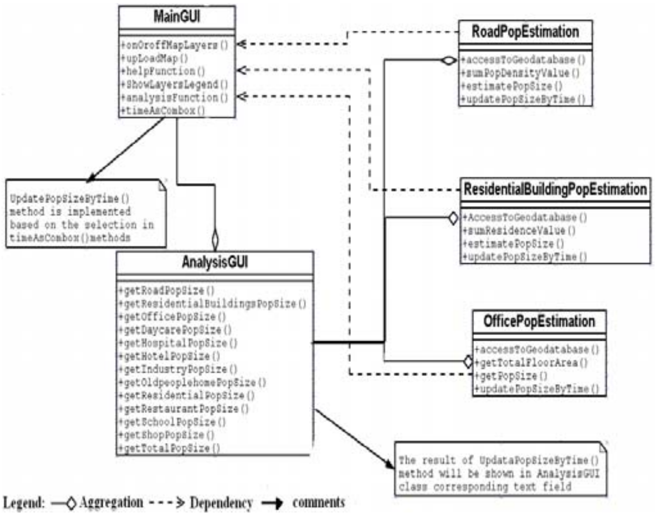


Figure 4. Spatio-temporal population model software UML diagram.

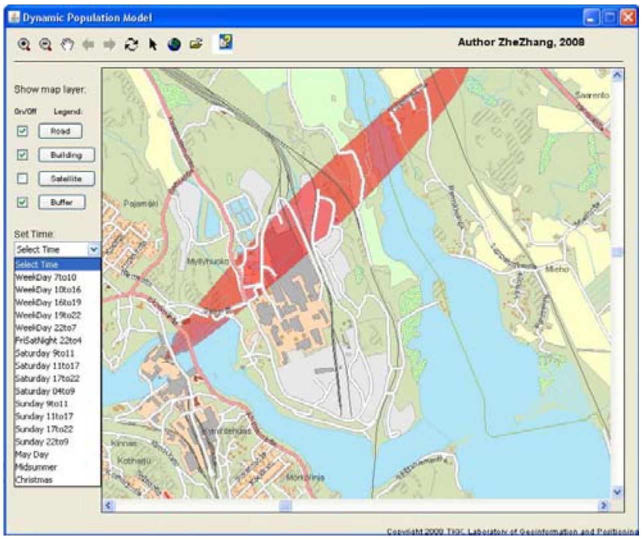
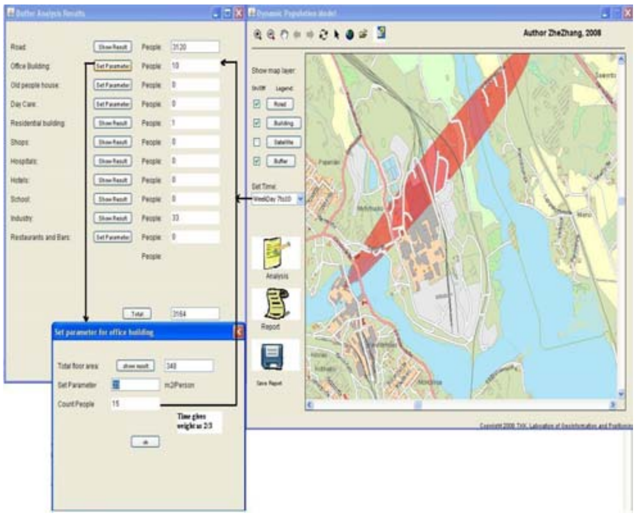


Figure 5. Spatio-temporal population model user interface.



Buffer Analysis Results

Road: People 3120

Office Building: People 10

Old people house: People 0

Day Care: People 0

Residential building: People 1

Shops: People 0

Hospitals: People 0

Hotels: People 0

School: People 0

Industry: People 33

Restaurants and Bars: People 0

People

Total 3164

Set parameter for office building

Total floor area: 340

Set Parameter: m2/Person

Count People: 15

Time gives weight as 2/3

Dynamic Population Model

Author ZheZhang, 2008

Show map layer

On/Off Legend

☑ Road

☑ Building

☐ Satellite

☑ Buffer

Set Time: WeekDay 7to10

Copyright 2008 TKK, Laboratory of Geoinformation and Positioning

Figure 6. An example of estimating population size in an office building.

6. REFERENCES

Armstrong, M., 1988. Temporality in spatial databases. *Proceedings of the GIS/LIS´ 88 Conference*, San Antonio, USA, pp.880-889.

Ahola T., Virrantaus K., Krisp J. M. and Hunter G.J., 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision-making , *International Jounal of Geogrpahical Informaiton Science,21*,pp.935-953.

El-Geresy B. A., Abdelmoty A. I., and Jones C. B., 2002. Springer-Verlag publications. "Spatio-temporal Geographic Information System: A Causal Perspective", Berlin Heidelberg,Germany. http://www.springerlink.com/content/8p6emwnag2qkb65r/fulltext.pdf (accessed 25 October 2009)

Finnish Meteorologival institute. Escape: a validated assessment tool for consequence analysis of accidents involving hazardous materials. http://www.fmi.fi/research_air/air_55.html(accessed 20 October 2009)

Hunter,G.J. and Williamson, I.P., 1990. The development of a historical digital cadastral database. *International Journal of Geographical Information System*, 4, pp. 169-179.

Hägerstrand, T., 1970. What about people in regional science?, *Papers of the Regional Science Association*, 24, pp. 1-12

Insta Oy, 1997. Insta inter-organisational crisis manager user manual.http://www.insta.fi/@Bin/1559978/insta_icm_05-2007_www.pdf (accessed 20 October 2009)

Longley P.A., 2005. Geographic information systems and science, pp.226-337.

Liu X. H, Kyriakidis P.C., Goodchild M. F., 2008, Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science*, 22, pp.431-447.

Mattos N.M., Eyer-Wegerner K., and Mitschang B., 1993. Grand tour of concepts for object-orientation from a database point of view. *Data and knowledge engineering*, 9, 321-352

Molarius R., Rantanen H., Huovila H., Korpi J., Yllaho J., Wessberg N., Virrantaus K., Rouhlainen V.,2009. Assuring the information flow from accident sites to decision makers - a finnish case study", In proceedings of First International Conference on Disaster Management and Human Health Risk: Reducing Risk, Improving Outcomes, New Forest, UK.

Nadi S. and Delavar M. R.,2003. ScanGIS conference paper."Spatio-temporal modeling of dynamic phenomena in GIS", Espoo, Finland. http://www.scangis.org/scangis2003/papers/11.pdf (accessed 20 October 2009)

Peuquet D.J. and Duan, N., 1995. An event-based spatio-temporal data model for temporal analysis of geographic data. *International Journal of Geographical Systems*, 9, pp.2-24

Smith, J.M., and Smith, D.C.P., 1997. Database abstractions:aggregation and generalization. *Association for Computing Machinery Transactions on Database Systems*, 2 , pp.105.

Sweitzer J., Langaas S.,1994. Modelling Population Density in The Baltic States Using the digital Chart of The World and other small data sets, Coastal Conservation and Management in the Baltic Region, In:*Proceedings of the EUCC-WWF conference, Riga, pp.257-267.*

Worboys M. F.,1992. A model for spatio-temporal information. In:*The 5th International Symposium on Spatial Data Handling*, 2, P. Bresnahan, E. Corwin and D.

Worboys M., 1994. Object-oriented approaches to georeferenced information. *International Jounal of Geographical Information Science*.8, pp. 385-399.

Worboys M.F., Hearnshaw H.M., Maguire D.J., 1990. Object-oriented data modelling for spatial databases. *International Journal of Geographical Information Systems*, 4, 369-383.

Yuan M., 1996. Temporal GIS and spatio-temporal modelling. In: *proceedings of the international conference/workshop integrating GIS and Environmental modelling ,USA* .

Zhang J. and Goodchild M.,2002. Uncertainty in Geographical Informaiton, Taylor&Francis pp. 9.

DEFINING DYNAMIC SPATIO-TEMPORAL NEIGHBOURHOOD OF NETWORK DATA

Tao Cheng [1] Berk Anbaroglu [1, 2]
[1] Dept. of Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT, UK
[2] Dept. of Geodesy and Photogrammetry Engineering, Hacettepe University, Ankara, Turkey
{tao.cheng, b.anbaroglu}@ ucl.ac.uk

Commission II, WG II/3

KEY WORDS: Spatio-temporal neighbourhood, spatio-temporal clustering, network complexity

ABSTRACT:

To improve the accuracy and efficiency of space-time analysis, spatio-temporal neighbourhoods (STNs) should be investigated and analysed in the classification, prediction and outlier detection of space-time data. So far most researches in space-time analysis use either spatial or temporal neighbourhoods, without considering both time and space at the same time. Moreover, the neighbourhoods are mostly defined intuitively without quantitative measurement. Furthermore, STNs of network data are less investigated compared with other types of data due to the complexity of network structure. This paper investigates the existing approaches of defining STNs and proposes a quantitative method to define STNs of network data in which the topology of the network does not change but the characteristics of the edges (i.e. thematic attribute values) change with time which requires dynamic STNs adapted to the properties of the network. The proposed method is tested by using London traffic network data.

1. INTRODUCTION

The amount of data which has spatial and temporal dimensions increases dramatically via the wide usage of fixed or mobile sensors. Extracting useful information from these data gains importance day by day which is referred as spatio-temporal data mining. Classification, prediction, outlier detection and clustering are among the major tasks of spatio-temporal data mining and analysis. These tasks involve analyzing spatio-temporal neighbourhoods (STNs) because STNs of an instance give important clues on the evolution of the instance itself. Better defining and identifying the neighbouring instances will lead to better modelling of the phenomenon under investigation.

So far most STNs root from spatial neighbourhoods (SNs) which then elongated in time. Different data forms as point, grid, polygon or network will exhibit different forms of STNs. Thus, different methods are used to define STNs on these data forms. For point data, a distance threshold is usually used to define spatial neighbourhood and temporal neighbourhood is usually defined intuitively (Celik et al., 2006). Lu et al. (2003) and Chen et al. (2008) used k-Nearest Neighbour (k-NN) based on Mahalanobis distance to find the SNs of an instance. Then STN is treated as a static entity where spatial neighbourhood is simply elongated in time dimension (GeoPKDD, 2006; Zhang et al., 2008). Grid data possesses a regular structure where several intuitive definitions for spatial neighbourhood exist. These neighbourhood strategies can be considered as metaphor of chess pieces and named as rook, queen and bishop neighbourhood. Yin and Collins (2007) used rook neighbourhood as the spatial neighbourhood and considered one time step backward as temporal neighbourhood to detect moving objects on videos. For polygon data, if two polygons share a common edge then they are thought to be spatial neighbours. Billard et al. (2007) predicted an epidemic in time across 12 states of US based on spatial neighbourhoods considering one former time stamp. Network data is the least investigated among all the data types and networks are treated as graph structures where the spatial neighbourhoods are defined by graph connectivity. Shekhar et al. (2001) defined STNs as spatial neighbourhoods that are adjacent at the graph with temporal neighbourhood consisting of previous time stamps of the spatial neighbourhoods. Although space and time are integrated in STNs, the neighbourhoods are fixed for an individual like in space and time. However, due to the complexity of networks, the neighbourhoods are actually dynamic.

This paper is motivated from aforementioned facts: 1) space and time are not treated in an integrated manner; 2) there is not a quantitative method for defining dynamic STNs for network data. To achieve this quest, literature related with spatial and spatio-temporal clustering is given under the second section. Third section discusses the proposed algorithm to cluster on spatio-temporal network data in which the topology of the network does not change but the thematic attribute associated with the edges of the network changes with time. The case study is discussed and results are shown in the fourth section. Conclusion and future work is given in the fifth section.

2. RELATED WORK ON SPATIAL AND SPATIO-TEMPORAL CLUSTERING

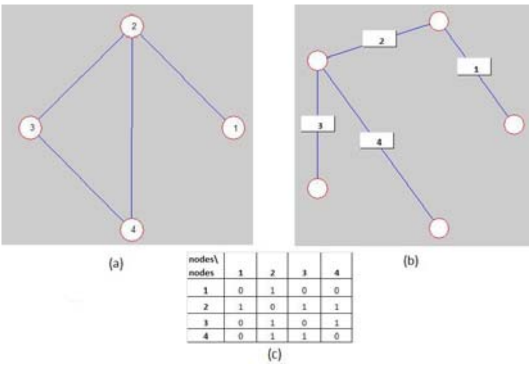
We believe that the originating point to attack the problem to define STNs should be spatio-temporal clustering, because the underlying ideas of both “neighbourhood” and “clustering” is same: to group the observations so that similar observations will fall into the same grouping (i.e. neighbourhood or cluster respectively). Considering these, the research problem can be restated as: dynamic spatio-temporal clustering on spatially embedded network data. There are three domains under spatio-temporal clustering: thematic, spatial and temporal attributes. Thematic attribute gives the information about the phenomenon observed, spatial and temporal attributes give the location and timing of the observation respectively. This section describes the literature conducted on spatial and spatio-temporal clustering.

Spatial clustering problem is also referred as ‘Dual Clustering’ by Lin (2005). They proposed two distance functions: one in spatial and another in thematic domain. These functions are combined with a pre-defined weight value to get one distance function. Choosing the pre-defined weight value is not trivial and it is chosen intuitively in their research. Wang (2007) used only spatial neighbourhood relations to cluster network data without considering the temporal domain. As a result, the dynamics in the network are not captured.

Spatial clustering is not sufficient to understand ‘events’ since to describe an event, one needs to answer the questions of *what*, *when* and *where*. In other words, thematic, temporal and spatial domains should be combined in a consistent way to have a better understanding of spatial phenomenon. Wei (2009) divided the time line into fixed size intervals and calculated the similarity based on the thematic domain. Spatial domain is used by means of defining a spatial distance threshold. However, how to choose the spatial distance threshold was not discussed. In addition, clustering results depend on the size of the chosen temporal interval. Neill (2005) emphasized on the significance of temporal domain. They used a probabilistic approach to detect emerging spatio-temporal clusters. However, the spatio-temporal process is assumed to follow a Poisson distribution which may not be the real case or time-consuming tests should be done to verify this assumption. Chan et al. (2008) captured the temporal dynamics of a graph by inspecting on the presence or absence of an edge. Their main task is to detect the regions where the change (absence/presence of an edge) is spatio-temporally correlated.

3. SPATIO-TEMPORAL CLUSTERING ON SPATIALLY EMBEDDED NETWORKS

Theoretically, one can represent the spatio-temporal objects as either vertices or edges in an undirected graph. An example of this is shown at figure 1. Figure 1 (a) represents the objects at vertices and Figure 1 (b) represents the same objects at edges. Figure 1(c) is the adjacency matrix for both of the graphs shown at Figure 1(a-b). To be consistent with the case study, from now on the representation shown at Figure 1 (b) will be used. Thus, spatio-temporal objects are the edges of the graph and vertices connect the objects coincident to them. In either case, the idea behind the representation is to obtain the adjacency matrix.



nodes\ nodes	1	2	3	4
1	0	1	0	0
2	1	0	1	1
3	0	1	0	1
4	0	1	1	0

Figure 1: Different graph representations of same data

Although the algorithm is designed for network data, this algorithm could be used for spatio-temporal clustering whenever the spatio-temporal phenomenon (which can exhibit in point, line or polygon) could be represented as a graph structure (G = (V, E) where V represents the set of vertices and E represents the set of edges).

Once the graph structure of the spatio-temporal phenomenon is acquired, then a matrix showing the connectivity between vertices (or edges); adjacency matrix; is created for the graph structure. While creating the adjacency matrix (if exists), the direction of the edges could be incorporated.

Up to now, the spatial domain is used to acquire the adjacency matrix of the spatio-temporal phenomenon. Temporal and thematic domains are exploited at this stage. Temporal domain is divided into equal parts where each part will have only two consecutive observations in the thematic domain. This is called as the *basic temporal interval*. For example basic temporal interval *k* of the object *p* consists of the two thematic attribute observations of p^{th} object at consecutive times of *k-1* and *k*. At each comparison step, basic temporal interval is shifted one time step. Thus, if the time-series has a length of *t*, there will be *t - 1* similarity results for the two adjacent objects’ similarity comparison. Since it consists of two consecutive (in temporal domain) observations, it is possible to derive several different similarity metrics (slope of change, difference/mean of the two observations,..) to compare between an object and the objects which are adjacent to it. Also, all of the possible similarities/dissimilarities between the two compared time series will be captured by this way (since it is not sound to have a basic temporal interval of size one). This is the first novelty of this research, since there is no need to specify a window size at temporal domain and it is designed to be the simplest possible, having two consecutive observations. In addition, this will allow capturing all of the possible similarities between two time series.

The similarity function is defined at basic temporal interval of *k* for two adjacent objects *p* and *q* with *at least* four inputs (i.e. p_{k-1}, p_k, q_{k-1}, q_k) where the thematic attribute value of *p* and *q* at times *k-1* and *k* are denoted as p_{k-1} and p_k and q_{k-1} and q_k respectively. Similarity function takes at least these four inputs, because some other parameters (which should be defined using background knowledge) may be needed to define the flexibility of similarity comparison.

For the objects to be labelled as positively similar at *basic temporal interval k* two requirements should be fulfilled: Firstly, the direction of change in thematic attribute values (i.e. slope) should be same and secondly, the thematic values of both objects should be similar which is quantified by the parameter *δ*. This requirement needs to be symmetric (e.g. if spatial object *p* is found to be positively similar at basic temporal interval *k* with the spatial object *q*, then *q* should also be positively similar with *p* at k^{th} basic temporal interval) , thus has two parts separated by a logical *or* operator. These two requirements for a positive similarity are illustrated at equations 1 and 2 respectively. If either of these conditions hasn’t met, then the similarity function will return a negative similarity result.

$$\frac{p_k - p_{k-1}}{q_k - q_{k-1}} > 0 \tag{1}$$

$$\begin{gathered}(1-\delta)(q_k + q_{k-1}) < p_k + p_{k-1} < (1+\delta)(q_k + q_{k-1}) \\ \vee \\ (1-\delta)(p_k + p_{k-1}) < q_k + q_{k-1} < (1+\delta)(p_k + p_{k-1})\end{gathered} \tag{2}$$

These similarity criteria are one of the many possibilities, however we tried to make it as generic as possible.

This time-series similarity comparison is done for all the adjacent objects which will constitute the dynamic spatial neighbourhood (i.e. spatial neighbourhood changes with time) of an object rather than STN. This is because; the comparison between adjacent objects is the done at the same basic temporal intervals. To capture the STN, temporal comparison should also be done. This will be achieved by comparing the object *p* at basic temporal interval *k* with itself at basic temporal interval *k+1*. The requirements for this comparison are same as spatial comparison stated above with the only difference that q_{k-1} and q_k is replaced with p_k and p_{k+1} respectively.

After the similarity comparisons are done for all adjacent objects, spatio-temporal clustering search is extended through the adjacencies of the adjacent objects. This time, adjacent objects are compared with their adjacencies at the intervals where a positive similarity is determined in the previous stage, so that spatio-temporal clusters will be spatially linked within themselves. Similar with the spatial search, temporal search will also extend to temporal adjacencies as long as there is positive similarity. This combination of spatial and temporal searches based on the similarity metric will constitute the spatio-temporal clusters and indeed the STNs. This search continues, until there is no more similarity is found. Therefore, the second novelty of this proposed algorithm is that there is no need to define spatial or distance threshold. By this way, the spatio-temporal clustering is conducted on all three domains: thematic, spatial and temporal domains.

4. CASE STUDY

4.1 Data Description

Algorithm presented in the third section is applied on the spatio-temporal traffic data of London road network. The data consists of average journey time of 11 objects obtained at 5 minutes intervals on 28 December 2009. These 11 links (i.e. 'objects' of this case study) are near the Blackwall Tunnel (figure 2) which is known as its unexpected congestions due to traffic accidents. Each link has an id (number) and a direction indicator (N or S tells traffic flow is towards north or south). Thematic attribute is converted from average journey time into average excess time per kilometre.

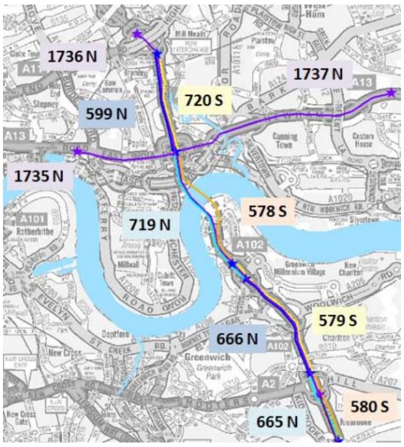
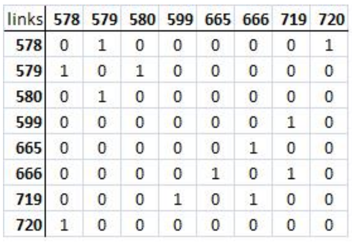


Figure 2: Map of the links in case study

To calculate the excess time per kilometre one need to define the average free flow journey time which we defined as the average of journey times occurred between 02:00 – 06:00 which is a common time interval used to observe the free flow characteristics of the link. Equation 3 shows the calculation of excess time per kilometre, where $jt_{d_{i,t}}$ denotes the average journey time of the link *i* at time *t* on day *d*. Similarly $jt_{d_{i,avg(02:00-06:00)}}$ denotes the free flow journey time at the link *i* and day *d*. Excess journey time is calculated as the difference between the observed journey time and free flow journey time. Then, this difference is divided by the length of the link to get the thematic attribute to be used to do spatio-temporal clustering. This metric is used to for spatio-temporal clustering, since it is a good metric to measure congestion.

$$excess_{per_{km}(d,i,t)} = \frac{jt_{d_{i,t}} - jt_{d_{i,avg(02:00-06:00)}}}{length(i)}$$
(3)

Another thing to define is the rules to define the adjacency of the links. An intuitive rule; flow of traffic should be in same direction and links should coincide at a vertex, is used to create the adjacency matrix. For instance, links 665N and 580S coincides at the same vertices; however they are not considered as adjacent since the flow of traffic is in opposite directions. With this definition of adjacency, adjacency matrix for the objects of figure 2 is created as shown at figure 3 where the links (1735-1737)N are not included because they overlap with other objects.



links	578	579	580	599	665	666	719	720
578	0	1	0	0	0	0	0	1
579	1	0	1	0	0	0	0	0
580	0	1	0	0	0	0	0	0
599	0	0	0	0	0	0	1	0
665	0	0	0	0	0	1	0	0
666	0	0	0	0	1	0	1	0
719	0	0	0	1	0	1	0	0
720	1	0	0	0	0	0	0	0

Figure 3: Adjacency Matrix of the objects of Figure 2

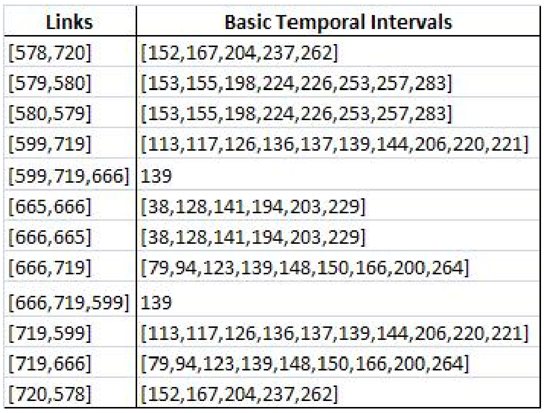
Rules to define the adjacency matrix can even be more specified by considering a traffic rule which is; an event happening at downstream will effect the upstream but not the vice versa. With this definition, 599N will be adjacent with 719N, but 719N will not be adjacent with 599N because what ever happened at 719N will not affect 599N because the traffic is flowing from 719N to 599N. However, this idea is not incorporated when defining the adjacency matrix in this research.

Final thing to define is the δ parameter in the similarity function. It is decided intuitively as 0.1 leading to the interpretation that two adjacent objects will be similar as long as their sum of the thematic attribute (i.e. excess time per km) values at a basic temporal interval lies in 10 percent zone of other adjacent object's sum of the thematic attribute (i.e. excess time per km) values at that basic temporal interval.

4.2 Results

Implementations were conducted on Matlab 2008a, however due to time constraints only spatial search part was implemented without doing the temporal search part. Once the similarities of adjacent objects are determined and the similarity search is extended based on the adjacency matrix and positive similarities. Thus, the results stated at figure 4 shows the dynamic spatial neighbourhood discovery rather than the STN.

On the left column of figure 4 the clustered links and on the right column the basic temporal intervals in which they are clustered are shown. Basic temporal intervals of a day start with 1 which indicates the time between 00:00-00:05 and ends with 288 which indicates the time between 23:55-00:00.



Links	Basic Temporal Intervals
[578,720]	[152,167,204,237,262]
[579,580]	[153,155,198,224,226,253,257,283]
[580,579]	[153,155,198,224,226,253,257,283]
[599,719]	[113,117,126,136,137,139,144,206,220,221]
[599,719,666]	139
[665,666]	[38,128,141,194,203,229]
[666,665]	[38,128,141,194,203,229]
[666,719]	[79,94,123,139,148,150,166,200,264]
[666,719,599]	139
[719,599]	[113,117,126,136,137,139,144,206,220,221]
[719,666]	[79,94,123,139,148,150,166,200,264]
[720,578]	[152,167,204,237,262]

Figure 4: Clustering Results of 28 December 2009

Since the main program searches for all of the individual objects (this can be seen from the first object under the different clusters) and creates the clusters based on that search, the adjacent links are clustered at the same basic temporal intervals which implies the symmetric nature of the clusters. For example when searching for the object 579, similarities were found at the basic temporal intervals of 113, 117, 126, 136, 139, 144, 206, 220 and 221 with object 719. When the main program runs for the object 719, it clusters with object 599 at the same basic temporal intervals. This search is redundant and similarity of the matrix can be exploited to eliminate this redundant search. However, this depends on the adjacency matrix and if the adjacency matrix is not symmetric then, all the objects should be searched.

Results indicate that, as the similarity search extends towards the adjacencies of adjacent links, detected similarities decreases. This is an expected result, since spatio-temporal correlation will decrease with the increase of the spatial distance between the objects. There is only one basic temporal interval (i.e. 139) and where a spatio-temporal cluster is formed among the links 599, 719 and 666.

These results show that the dynamic nature of the spatio-temporal clusters can be captured by using the proposed algorithm: spatio-temporal clusters grow and shrink with time (since the property of the network change with time).

It is clear that, these results utterly depend on the chosen similarity function and its parameters as well as the rules to define the adjacency among the objects. As aforementioned, the similarity function needs to be defined by considering background knowledge about the phenomenon.

5. CONCLUSION

This paper addressed the issue of importance of defining STNs. It is seen that most of the spatio-temporal data mining tasks use the STN concept. Commencing from the analogy between neighbourhood and cluster, this research proposed an algorithm on spatio-temporal network clustering for the task to determine the STN. It is shown that, the proposed algorithm captures the dynamics of the network. Both spatial and temporal information are used effectively to reduce the computational cost of detecting spatio-temporal clusters. Other than choosing the similarity function and related parameters of the similarity function, the user is not involved in choosing any spatial or temporal parameter.

There are several drawbacks of the proposed algorithm as well. Firstly, it cannot handle the cases where there is a loop at an edge (i.e. edge from a vertex to itself). Secondly, similarity is treated as a binary characteristic. And finally, the algorithm can only handle spatially embedded (network topology does not change with time) network objects.

The case study did not capture the STN due to time limitations but only captured the dynamically changing spatial neighbourhood. Implementing and testing the temporal search part is left as a future work. In most the researches, as well as the case study mentioned in this paper, temporal dimension exhibits in different scales (time-of-day, day-of-week, etc.). Future research will also focus on the extending the spatio-temporal clustering to different time scales. In addition understanding the dynamic nature of the spatio-temporal clusters (e.g. finding the patterns of growth and shrink) is another challenging task to be sought.

ACKNOWLEDGEMENTS

This research is jointly supported by UK EPSRC (EP/G023212/1), Chinese NSF (40830530) and 863 Programme (2009AA12Z206). Second author thanks to Hacettepe University and Higher Education Council of Turkey for the PhD scholarship they provide. Data is kindly provided from TfL. We also would like to thank to the anonymous reviewers for their comments.

REFERENCES

Billard, L., Lee, S.D., Kim, D.K., Lee, K.M., Lee, C.H., Kim, S.S., 2007. Modeling Spatial-Temporal Epidemics Using STBL Model. In: *International Conference on Machine Learning and Applications,* 629-633.

Celik, M., Shekhar, S., Rogers, J.P., Shine J.A., 2006. Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining: A Summary of Results. In *18th IEEE International Conference on Tools with Artificial Intelligence,* 106-115.

Chan, J., Bailey, J., Leckie, C., 2008. Discovering correlated spatio-temporal changes in evolving graphs. *Knowledge and Information Systems*, 16, pp. 53-96.

Chen, D., Lu, C.T., Kou, Y., Chen Y., 2008. On Detecting Spatial Outliers. *GeoInformatica*, 12(4): 455-475.

GeoPKDD. 2006. Last visited: 11 December 2009 http://www.geopkdd.eu/files/dev_public/D2.2M.pdf

Lin, C.R., Liu, K.H., Chen, M.S., 2005. Dual clustering: integrating data clustering over optimization and constraint domains. *IEEE Transactions on Knowledge and Data Engineering, ,* 17, pp. 628-637.

Lu, C. T., Chen, D., Kou, Y., 2003. Detecting spatial outliers with multiple attributes. *15th IEEE International Conference on Tools with Artificial Intelligence,* 122-128.

Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K., 2005. Detection of emerging space-time clusters. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, Chicago, Illinois, USA, pp. 218-227.

Shekhar, S., Lu, C., Zhang, P., 2001. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining,* San Francisco, California, USA

Wang, Y., Chen, Y., Qin, M., Zhu, Y., 2007. SPANBRE: An Efficient Hierarchical Clustering Algorithm for Spatial Data with Neighborhood Relations. In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery,*pp. 665-669.

Wei, L., Peng, W., 2009. Clustering Data Streams in Optimization and Geography Domains. In: *Advances in Knowledge Discovery and Data Mining*, pp. 997-1005.

Yin, Z., Collins. R., 2007. Belief Propagation in a 3D Spatio-temporal MRF for Moving Object Detection. In: *IEEE Conference on Computer Vision and Pattern Recognition.*

Zhang., S., Yao, H., Liu, S., 2008. Dynamic background modeling and subtraction using spatio-temporal local binary patterns. In: *ICIP 2008,* 1556-1559.

AN APPROACH OF DISCOVERING SPATIAL-TEMPORAL PATTERNS IN GEOGRAPHICAL PROCESS

Siyue Chai[a,b] ,Fenzhen Su[a],Weiling Ma[a,b,]

[a] Institute of Geographic Sciences and Natural Resources Research, Beijing 100101, China - chaisy@lreis.ac.cn
[b] Chinese Academy of Science, Beijing 100101, China – sufz@lreis.ac.cn

KEY WORDS: spatial data mining, spatial process, spatial-temporal pattern

ABSTRACT:

Spatial data mining focuses on searching rules of the geographical statement, the structures of distribution and the spatial patterns of phenomena. However, many methods ignore the temporal information, thus, limited results describing the statement of spatial phenomena. This paper focuses on developing a mining method which directly detects spatial-temporal association rules hidden in the geographical process. Through such approach, geographical process can be extracted as a particle which exists in spatial-temporal- attribute dimensions. By setting customized fixed-window, geographical process in one time interval is organized as a record with attribute value and spatial orientation change. Spatial-temporal association rules can be found in geographic process mining table.

$$[\text{TimeInterval}_i, \text{MovingDirection}_m, P] \Rightarrow [\text{TimeInterval}_i, \text{MovingDirection}_n, Q]$$

To verify this mining approach, it is applied on AVHRR MCSST thermal data for extracting Indo-Pacific warm pool's frequent movement patterns. The raw data provided by PO.DAAC, whose time spans of 20years from 1981 to 2000 with 7days' time particle, has been used to mining spatial temporal association rules. In the experiment, we extract warm pool within 30°N-30°S, 100°E-140°W and use 28°C as temperature threshold. After which Warm Pool's geographical process table is established so as to describe the variation of warm pool in spatial-temporal-attribute dimension. In the mining process, 18 spatial-temporal process frequent models can be found by setting minimal support threshold at 10% and confidence threshold at 60%. The result shows such a methodology can mine complicated spatial-temporal rules in realistic data. At the same time, the mining result of warm pool's frequent movement patterns may provide reference for oceanographers.

1. INTRODUCTION

In data mining, Association Rule Mining (ARM) is a popular and prevailing researched method for discovering interesting relations between variables in large databases. Previously Spatial Association Rules Mining (SARM) and Sequence Mining (SM) are two different subareas on spatial rules and temporal rules respectively. SARM interests on how to find connection rules which is represent by spatial topology, such as "is_Close" and "is_Next_To", among spatial objects(Koperski and Han 1995). So a spatial association rule (SAR) is an association rule where at least one of the predicates is spatial(Shekhar and Huang 2001). On the other hand, SM concerns how to dig association rules under temporal constraints(R. Agrawal 1995; Zaki 2001). However, with the development of information capture devices, data containing both time stamp and location information is recorded. Especially, when process Geographical Information System (PGIS) (Fen-zhen SU 2006) is under discussing, mining spatial- temporal rules are expected. Therefore a burning question is how to find a solution that could mine spatial-temporal association rules.

From research objects prospects, spatial-temporal association rules can be divided into two parts. One is to find rules from group objects, the other focuses on mining movements of one object which has flexible shape and moving patterns, such as Warm Pool. In this paper, we propose an approach that takes into account spatial-temporal attributes to mine the frequent trajectory patterns of flexible shaped objects. Such approach contains three steps. Fundamentally, we scan entire database and organise geographic process. Then the process record pre-processing is required. Finally, we apply Apriori-based algorithm to mine spatial-temporal association rules. To prove validation of such approach, we set an experiment using AVHRR remote sensing image provided by PO.DAAC, whose time spans is 20 years from 1981 to 2000 with 7days' time particle, to mine Indo-Pacific warm pool's frequent trajectory. Some interesting rules are found which comply to the general rules of Warm Pool's seasonal fluctuation, while leaving some rules still unresolved.

2. RELATED WORK

From analytical prospective, the target of STARs mining is to search the correlations hidden behind objects' spatial and temporal data. Automatic statistic approaches are used for extracting such correlations from 2-dimensional table structure which covers realistic objects' statement or abstracted objects process. Such statistic approach is well developed by related database techniques and traditional data mining algorithms. In traditional data mining, R.Agrawal etc(Rakesh Agrawal 1993) proposed an Apriori algorithm to mine frequent item set from transactional database. The form of the rule is abstracted as X=>Y. The algorithm employs prior knowledge of frequent item set properties, which are all non-empty subsets of a frequent item set that must also be frequent. Under such prior knowledge, the number of candidate sets' combination is reduced prominently. Apriori algorithm sets a mile stone in ARM. Two years later, J.Han(Jiawei Han 1995) developed Agrawal's research and proposed an algorithm which utilised Multi –Level logic structure to abstract target rules and mined cross level schema so that complicated logic rules could be described.

After such fundamental work, many data mining researchers (R. Agrawal 1995; J.Pei 2001; Zaki 2001;

Hwang, Wei et al. 2004; Winarko and Roddick 2007; Kong, Wei et al. 2009)demonstrated how to combine both temporal information and other attributes into sequence rules and temporal rules. In rich temporal description forms, from basic time stamp representation to process representation, temporal topology was introduced. From the view of data structure, Zaki et al's SPADE algorithm decomposed the original mining problem into several smaller sub-problems so that each sub-problem could be independently solved in memory, while J.Pei et al's PrefixSpan algorithm reduced search space and accelerated the mining process by using the projected databases. In temporal topology, Kong et al proposed an algorithm which used four kinds of temporal topology predicates, namely before, during, equal and overlap, to mine temporal association rules. Despite temporal topology, Winarko et al proposed a method called ARMADA, whose accommodating temporal intervals offered rules that were richer still.

Geographic researchers discussed ways of taking advantages of multi level logic structures and cross mining schema in Geographic Information System so as to automatically find dependent patterns in spatial data. Han etc al (Koperski and Han 1995) set multilevel concept descriptions for spatial rules, such as "g_clse_to", "not_disjoint, close_to", "Intersects, Inside, Contain, Equal" to show the spatial topology, the final rules can be represented as "is_a(x, house) A close_to(x, beach) → is_expensive(x) (90%)". Their work made spatial information and none spatial-temporal information consistant. AGM algorithm (Inokuchi, Washio et al. 2000) and FSG algorithm (Kuramochi and Karypis 2001) used an Apriori-based approach to combine frequent sub-graphs mined at the previous level to generate all candidates at the next level. Appice etc al (Appice and Buono 2005) summarizes advantages from this taxonomic knowledge on spatial data to mine multi-level spatial association rules. Bembenik etc al (Bembenik and Rybiński 2009) defined the neighborhood in terms of the Delaunay diagrams, instead of predefining distance thresholds with extra runs.

However, how to handle both spatial and temporal data is still under research. But rear works have been published yet. Lee etc al (Lee, Chen et al. 2009) developed an approach using TI-lists structure and GBM algorithm to find trajectory patterns. Huang etc al (Huang, Kao et al. 2008) summarized the correlation between sea salt and temperature in spatial-temporal distribution. The final rules can be described as "if the salinity rose from 0.15 psu to 0.25 psu in the area that is in the east–northeast direction and is near Taiwan, then the temperature will rise from 0° C to 1.2° C in the area that is in the east–northeast direction and is far away from Taiwan next month", which set temporal constraints that limit temporal information in two adjacent time intervals. Verhein etc al (Verhein and Chawla 2008) proposed source-thoroughfare-sink model describing trajectory pattern, while left none connection charts unexplained. In conclusion, traditional algorithm may lose some important information during searching frequent trajectory patterns of flexible shaped objects.

3. GEOGRAPHIC PROCESS MINING APPROACH

This section describes the concepts associated with applying the association rule mining to discover spatial-temporal patterns. Some definitions will also be given before describing the proposed algorithm. Figure 1 below shows the steps of processing.

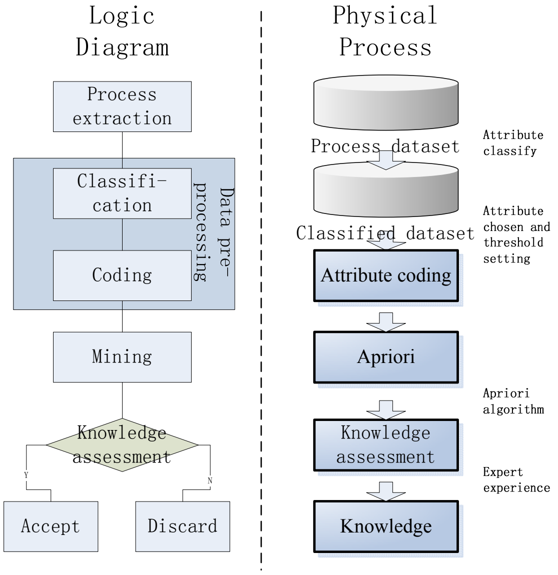


Figure 1. Mining approach flow chart

3.1 Description of geographic process

Modern physics shows that time and space form an indivisible entirety called four-dimension space. In this space, continuity dominates every attribute of existing objects (low speed), which means objects' snap shot always exists. In geography, modern science provides powerful devices to capture such snapshot of geographic phenomena rapidly and continuously. Therefore, to mine geographic patterns from these primary data needs strict organization and description of spatial-temporal statement.

Definition 1: In 4-dimensional space, given exactly T_i, objects *I* can be represented as I=[$Time_i$, $Location_i$, $Attribute_i$] , i∈N , $T_i \in (-\infty, -\infty)$. Location is the object's spatial vector, usually Location = [Longitude, Latitude, Altitude]^T, Attribute is the object's attribute vector.

Every objects can be described as I=[T_i , L_i, A_i]. Attribute A is a high dimensional vector, whose item depicts object from different aspects. To represent flexible appearance object's geometry, kinds of index can be used, such as degree of fragmentation, and fractal feature (M Coster 1985). Therefore, flexible shaped object can be recorded as kinds of index. And the description of spatial–temporal statement is compatible with 2-dimensional table structure, which means the characterization of object can be stored into formalized database table (Table 1).

ID	Time	Longitude	Latitude	Temperature	Area (million km^2)
1	1981-11-3	156.62	2.61	28.96	42.55
2	1981-11-10	158.57	1.43	29.44	42.1
3	1981-11-17	160.9	0.17	29.47	42.42
4	1981-11-24	161.46	-1.17	29.23	40.77

5	1981-12-1	163.04	-1.50	29.33	38.84

Table 1. An Example of Spatial-temporal statement table

Definition 2: Geographical Process (GP) is the change from starting statement to ending statement. Given fixed time window TI, TI = $[T_i,T_j]$, i < j and i, j∈N, GP = $^{\Delta}I$=[TI, $^{\Delta}L$, $^{\Delta}A$].

Definition 3: Time interval in GP is the process particle marked as PP.

Geographic process, which reflects object's spatial movement in time evolution, is composed by sets of quantification or qualification statement. Object's spatial change $^{\Delta}L$ contains the change of reshaping and moving. To simplify such variety, we view target object as a point.

Definition 4: Abstracted object as a point, called spatial-temporal point.

As a point, by introducing the orientation model, spatial topology is limited to just moving direction (Table2) so that the complexity of association rule is reduced.

TI ID	Direction	Temperature	Area **(million km^2)**	Start Time	End Time
1	N	-0.10	-6.98	1981 Winter	1982 Spring
2	NE	-0.24	4.50	1982 Spring	1982 Summer
3	SE	0.09	1.25	1982 Summer	1982 Autumn
4	SW	-0.06	-7.72	1982 Autumn	1982 Winter
5	NW	-0.26	-2.04	1982 Winter	1983 Spring

Table 2. An Example of Spatial-temporal process table (PP = 3 months)

From logical view, the description of spatial-temporal association rule is the instantiation of the form of target knowledge. Spatial-temporal association rule format prepares the foundation data set for pre-processing. As from mining view, it is the goal of the final result.

Definition 5: Geographic Process Association Rule is one kind of spatial-temporal association rule, defined as GPAR: GP_m=> GP_n, m,n∈N

Precisely, spatial temporal process association rule is written as [TI_m, $^{\Delta}L_m$, $^{\Delta}A_m$]=> [TI_n, $^{\Delta}L_n$, $^{\Delta}A_n$], m,n∈N. Such rule is distinguished from traditional association rule in three characters:

a) Spatial-Temporal scalability. The spatial or temporal information can be divided into different scale. In question-driven data mining process, the detail of final patterns is influenced by spatial-temporal particle.

b) Evolution entirety. As process association rule contain spatial temporal coupled information. So the process association rule describes the integrated 4-dimensional change process. In this experiment, setting time particle of Indo-Pacific warm pool's movement as 3 month，every statement change will be recorded while each attribute will be combined together for providing entirety of spatial- temporal information.

c) Rule flexibility. The formation of rule is just a description of process; it separates mining algorithm. So different mining target GP_m=> GP_n can be interpreted into diverse meaning. To take advantage of such character, different mining algorithm shares the same process of data sets.

3.2 Geographic Process data pre-processing

This section contains three parts. In the above, the pre-processing approach is discussed. Then, Geographic Process Mining Table (GPMT) is organized through pre-processing. As target GPAR rule described above, a mining method has been developed based on Apriori algorithm.

3.2.1. Process data set pre-processing

Data pre-processing is an approach which extracts data from primary concept level to upper ones. It is a question-driven procedure. All attribute's spatial-temporal particle, which directly influences the statistics in final rules is classified. After extracting the dataset of geographic process, one record in the table is called meta-process. Different attribute has different kind of classification threshold in the way of raising its concept level. Such as Table 2, using variance based binning method, to classify "Temperature". By setting 3 bins which represents "fall", "remain practically unchanged", "rise" in turn. The result is shown below.

class	Lower threshold	upper threshold
0		< -0.1140654
1	>= -0.1140654	<= 0.1126115
2	> 0.1126115	< 0.067004

Table 3. Classification after Binning

Reorganization and conversion makes every subset of process consistent (Figure 2). In this experiment's process [TI, $^{\Delta}L$, $^{\Delta}A$], as using orientation model, the number of changing domain "ΔL" sets 8, and the number of changing domain "A" are both 3. So taking the first row in Table2 as example, the result of coding is saved as "421" (Table 4).

Meta process: XXXXXXXXXX
└─X is classification number

Figure 2. Coding Format

TI_ID	process	Start Time	End Time
1	111	1981 Winter	1982 Spring
2	201	1982 Spring	1982 Summer
3	411	1982 Summer	1982 Autumn
4	610	1982 Autumn	1982 Winter

Table 4. Process Dataset

3.2.2. Building GPMT

Definition 6: GPAR's Time Span (TS) is the time span between TI_m and TI_n. So the TS∈[1,K], K= (Starting Time – Ending Time) / PP,K∈N.

The temporal connection hidden in the GPAR: GP_m=> GP_n is limited. For example, in this experiment, the maximum TS in 20 years is 75. To mine such GPAR, GPMT should be constructed by connecting Process Dataset joint with Cloned Process Dataset with TS time intervals' lag. (Figure 3)

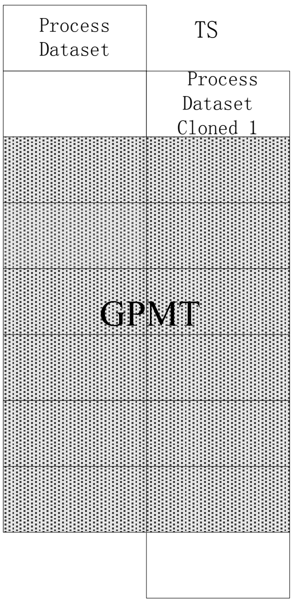


Figure 3 Geographic Process Mining Table (TS = 1 process)

3.3 Process association rules data mining

In this section, we introduce Apriori algorithm (Rakesh Agrawal 1993)to mine GPARs based on GPMT. By setting Maximum length of association rules and Support threshold = min_sup (support(A=>B) = P(A ∩ B)), confidence threshold = min_conf (confidence(A=>B) = P(B|A)), the Apriori algorithm traverses process dataset is not larger than such maximum length times. After first Scan, Apriori captures frequent set rule Length equals 1 and whose support and confidence are not less than min_sup and min_conffidence. Then build Length+1 candidate, and then Scan DB the second times, using Apriori traits pruning false candidate, generate Length+1 candidate. Such procedures continue until no Length+1 frequent sets can be found or Length equals Maximum length.

Pseudocode:

```
L1= {large 1-item sets} //Scan DB once, capture frequent set, L1
for (k=2; $L_{k-1}\neq \Phi$; k++)
{ $C_k$=apriori-gen($L_{k-1}$)//according to $L_{k-1}$ frequent set, digging new candidate sets, Ck
  for each t ∈D
  { Cspt =subset ($C_k$, t);
    for each c ∈$C_t$ c.count ++;
  }
  Lk={c∈$C_k$|c.count≥minsupp}
}
L=$\cup_k L_k$ ;
```

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section we present realistic remote sensing images and the mine spatial temporal association rules generated from them. To analysis results, we quoted some oceanographer's work to assess rules mined.

4.1 SST Data introduction

These experiments using AVHRR sensor's SST to analyze the data provided by PO.DAAC whose size is 2048*1024pixel^2. Under Mercator's projection, spatial resolution is 5.689 pixels per Longitude and 2.845 pixels per Latitude. Temporal resolution is 7 days and the data covers 20 years (1981-2001). According to Wang etc al (Wanying 1998), the Indo-Pacific warm pool forms an entire Warm Pool (WP) and its envelope should be 100°E-140°W,30°N-140°S within which the warm pool influences the weather of Eastern Asian mostly. Meanwhile, we use 28°C as Warm Pool's edge accepted by many oceanographers (Zhang Qilong and Weng Xuechuan 1997; Enfield, Lee et al. 2005) to extract geometries. (Figure 4)

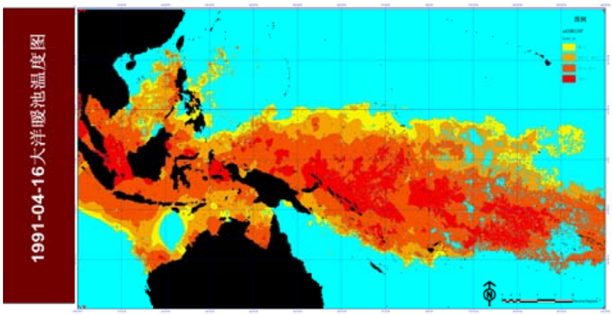


Figure 4. Warm Pool extraction image

4.2 Experiment

Using data processing approach mentioned above, the Warm Pool is described as the quantitatively and spatial-temporal statement table of Warm Pool which is captured by sensor in 7days. Set the maximum sequential rules is 3, min_support is 10%, confidence is 60%. We consider the temporal scales are 3months so as to slice data into big scale part.

4.2.1 Experiment

4 solar terms are chosen like spring, summer, autumn and winter. So the split points are Spring Equinox, Summer Solstice, Autumnal equinox and Winter Solstice. Seasonal process recorded (Table5) can be attained From Table2.

T I -I D	**Direct ion**	**Temper ature**	**Area change (million km^2)**	**Start time**	**End Time**
1	1	-0.10	-0.70	Spring	Summer
2	2	-.24	4.50	Summer	Autumn
3	4	0.95	1.25	Autumn	Winter
4	6	-0.06	-7.72	Winter	Spring
5	7	-0.27	2.04	Spring	Summer

Table 5 Seasonal process table

The Top 4 GPARs Results are listed:

Conseq-uent	**antece-dent**	**GPAR**	**support %**	**Confidence %**
Summer to autumn = 111	Autumn to winter = 511	[Summer to autumn, N,1,1] is_before [Autumn to winter, S,1,1]	27.78	80
Winter to Spring = 511	Spring to Summer = 810	[Winter to Spring , S,1,1] is_after[Spring to Summer, NW,1,0]	16.7	100
Autumn to winter = 512	Spring to Summer = 810	[Autumn to winter , S,1,2] is_after[Spring to Summer , NW,1,0]	16.7	66.7
Summer to autumn = 112	Spring to Summer = 111	[Summer to autumn ,N,1,2]is_after [Spring to Summer ,N,1,1]	16.7	66.7

Table 6 Mining Results

4.2.2 Rule analysis

El Niño-Southern Oscillation (ENSO), is a climate pattern that occurs across the tropical Pacific Ocean on average every five years, but over a period which varies from three to seven years, and is therefore, widely and significantly, known as "quasi-periodic." The two components are coupled: when the warm oceanic phase (known as *El Niño*) is in effect, surface pressures in the western Pacific are high, and when the cold phase is in effect (*La Niña*), surface pressures in the western Pacific are low. (K.E. Trenberth, P.D. Jones et al. 2007) Southern Oscillation is an oscillation in air pressure between the tropical eastern and the western Pacific Ocean waters. Low atmospheric pressure tends to occur over warm water and high pressure occurs over cold water, in part because deep convection over the warm water acts to transport air.(Figure 5) Normal equatorial winds warm as they flow westward across the Pacific. Cold water is pulled up along west coast of South America. Warming water is pushed toward west side of Pacific.(http://en.wikipedia.org/)

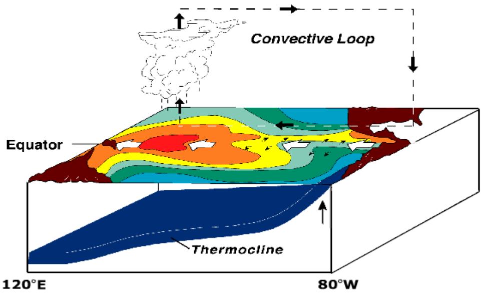


Figure 5 Normal Pacific pattern (http://en.wikipedia.org/)

The movement of warm water forms a periodical geographical process. After mining, GPARs may have association with Southern Oscillation. Here is the comparison.

GPAR1: In the same year, if during autumn to winter, Warm Pool moves South without Temperature and Area change prominently, then in the previous summer to autumn, Warm Pool moved North without Temperature and Area change prominently. This rule reflects warm pool's annual latitude infatuation. Two none spatial-temporal attributes remain unchanged. The rules appeared in 1984, 1992, 1993, 1995, 1997. Almost Earlier than all this rules El Niño phenomena happened. El Niño happens between October to next years February. It occured in 1982-1983,1986-1987, 1991-1992, 1993, 1994-1995, 1997-1998.And this rules are just after El Niño, excluding 1984. But during 1982-1983 ,one of the most powerful El Ninos happened , so the connection of GPAR1 and El Ninos have shown some kind of relation which needs further research or explanation.

GPAR2: In the same year, if during Spring to Summer, Warm Pool moves Northwest with Area shrinking prominently, then during winter to spring, Warm Pool will move South without Temperature and Area change prominently. Unlike GPAR1, rules applied in 1993, 1998, 1999, with 100% confidence and warm pool dominant area shrink prominently. In all dataset, during spring to summer, the area of Warm Pool shrink just appeared in 4 years in 1993, 1997, 1998, and 1999. This phenomena may be explained by too many rains above sea or powerful cold stream. It still needs oceangrapher to explain.

GPAR3: In the same year, if during Spring to Summer, Warm Pool moves Northwest with Area shrinking prominently, then during Autumn to winter, Warm Pool will move South with Area expanding prominently. Such rules happened in 1998, 1999, with pool dominant area shrink in spring to summer, expanding right into autumn to winter. In all dataset, during spring to summer, the area of Warm Pool shrinking just appeared in 4 years , which are 1993, 1997, 1998, 1999, while during autumn to winter in 1988, 1989, 1996, 1998, and 1999, more rules would be generated if data classification scale expaned. However,this rule may also be just a coincidence because its support and confidence are not high enough.

GPAR4: In the same year, if during spring to summer, Warm Pool moves North without Temperature and Area change prominently, then during summer and autumn, Warm Pool will move North with Area expanding prominently , just the same as GPAR3.This rule needs more data to prove.

So each of GPARs listed here has high support and confidence. The index of Lift which defines as Confidence / Support, are all greater than 1, which means all these rules are strong rules showing GP1=>GP2 are positively correlated. Despite these statistics, the appearance years of these GPARs may really correlate with El Niño and Southern Oscillation. Many oceanographers have worked on such areas and found patterns between Warm Pool and Southern Oscillation. (M. J. McPhaden 1990; Brijker, Jung et al. 2007; Cheng, Qi et al. 2008)

5. CONCLUSION

In this paper, an approach is proposed for mining spatial-temporal association rules of a flexible-shaped object. Such approach has three steps: geographic process is firstly generated from original image by designing process table. Then the attained table is processed so as to make attributes of geographic process consensus to form the GPMT table. Finally, spatial temporal association rules are mined using Apriori algorithm with rich spatial temporal information.

Although we have shown the rules mined by such approach some issues still need to be addressed in the future research. Above all, how to enrich spatial- temporal topologies so that rules can be represented in a more complicated way and can be dealt with is a sophisticated question. Anyway, the GPARs described in this paper is just an example to spatial–temporal association rules, it could also be used for other applications such as mobile advertisements, shoppers' trajectory analysis, and animal trajectory analysis to find other interesting rules.

References

Appice, A. and P. Buono ,2005. Analyzing Multi-level Spatial Association Rules Through a Graph-Based Visualization. Innovations in Applied Artificial Intelligence: 448-458.

Bembenik, R. and H. Rybiński ,2009. "FARICS: a method of mining spatial association rules and collocations using clustering and Delaunay diagrams." Journal of Intelligent Information Systems 33(1): 41-64.

Brijker, J. M., S. J. A. Jung, et al. ,2007. "ENSO related decadal scale climate variability from the Indo-Pacific Warm Pool." Earth and Planetary Science Letters 253(1-2): 67-82.

Cheng, X., Y. Qi, et al. ,2008. "Trends of sea level variations in the Indo-Pacific warm pool." Global and Planetary Change 63(1): 57-66.

Enfield, D. B., S.-K. Lee, et al. ,2005. "How are large western hemisphere warm pools formed?" Progress In Oceanography 70(2-4): 346-365.

Fen-zhen SU, C.-h. Z. ,2006. "A framework for Process Geographical Information System." Geographical Research 25(3): 477-483.

Huang, Y.-P., L.-J. Kao, et al. ,2008. "Efficient mining of salinity and temperature association rules from ARGO data." Expert Systems with Applications 35(1-2): 59-68.

Hwang, S.-Y., C.-P. Wei, et al. ,2004. "Discovery of temporal patterns from process instances." Computers in Industry 53(3): 345-364.

Inokuchi, A., T. Washio, et al. ,2000. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Springer-Verlag.

J.Pei, J. H., B.Mortazavi-Asl,Q.Chen,U.Dayal,M.C.Hsu ,2001. "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth " International Conference on Extending Database Technology:Advances in Database Technology: 215-224.

Jiawei Han, Y. F. (1995). "Discovery of Multiple-Level Association Rules from Large Databases." Very Large Data Bases 420-431.

K.E. Trenberth, P.D. Jones, et al. ,2007. Observations: Surface and Atmospheric Climate Change. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. S., D. Qin, M. Manninget al. Cambridge,UK: 235-336.

Kong, X., Q. Wei, et al. ,2009. "An approach to discovering multi-temporal patterns and its application to financial databases." Information Sciences In Press, Uncorrected Proof.

Koperski, K. and J. Han ,1995. Discovery of spatial association rules in geographic information databases. Advances in Spatial Databases: 47-66.

Kuramochi, M. and G. Karypis ,2001. "Frequent Subgraph Discovery." Data Mining, IEEE International Conference on 0: 313-320.

Lee, A. J. T., Y.-A. Chen, et al. ,2009. "Mining frequent trajectory patterns in spatial-temporal databases." Information Sciences 179(13): 2218-2231.

M Coster, J. C. ,1985. "Précis d'analyse d'images."

M. J. McPhaden, J. P. ,1990. "El Niñ-Southern Oscillation Displacements of the Western Equatorial Pacific Warm Pool." Science 250: 1385-1388.

R. Agrawal, R. S. ,1995. "Mining sequential patterns." 11th International Conference on Data Engineering (ICDE'95).

Rakesh Agrawal, T. I., Arun Swami ,1993. "Mining association rules between sets of items in large databases." ACM SIGMOD Record 22(2): 207 - 216

Shekhar, S. and Y. Huang ,2001. Discovering Spatial Co-location Patterns: A Summary of Results. Advances in Spatial and Temporal Databases: 236-256.

Verhein, F. and S. Chawla ,2008. "Mining spatio-temporal patterns in object mobility databases." Data Mining and Knowledge Discovery 16(1): 5-38.

Wanying, L. K. Z. C. S. ,1998. "Basic features of the warm pool in the western pacific and its impact on climate." Acta Geographica Sinica 53(6): 511-519.

Winarko, E. and J. F. Roddick ,2007. "ARMADA - An algorithm for discovering richer relative temporal association rules from interval-based data." Data & Knowledge Engineering 63(1): 76-90.

Zaki, M. J. ,2001. "SPADE: An Efficient Algorithm for Mining Frequent Sequences." Machine Learning 42(1): 31-60.

Zhang Qilong and Weng Xuechuan ,1997. "Analysis of some oceanographic characteristics of the tropical western pacific warm pool." Studia Marina Sinica 38(01): 31-38.

http://en.wikipedia.org/. "El Niño-Southern Oscillation."

SPATIO-TEMPORAL TRAJECTORY ANALYSIS OF MOBILE OBJECTS FOLLOWING THE SAME ITINERARY

Laurent ETIENNE[a], **Thomas DEVOGELE**[a] **and Alain BOUJU**[b]

[a]French naval academy research institute (IRENAV), CC 600, 29240 BREST, (laurent.etienne,thomas.devogele)@ecole-navale.fr
[b]Data-processing and Industrial Imagery Laboratory (L3I), Avenue Crépeau, 17042 La Rochelle Cedex 1, abouju@univ-lr.fr

KEY WORDS: spatio-temporal patterns, data mining, mobile object, outlier detection, maritime GIS.

RÉSUMÉ:

More and more mobile objects are now equipped with sensors allowing real time monitoring of their movements. Nowadays, the data produced by these sensors can be stored in spatio-temporal databases. The main goal of this article is to perform a data mining on a huge quantity of mobile object's positions moving in an open space in order to deduce its behaviour. New tools must be defined to ease the detection of outliers. First of all, a zone graph is set up in order to define itineraries. Then, trajectories of mobile objects following the same itinerary are extracted from the spatio-temporal database and clustered. A statistical analysis on this set of trajectories lead to spatio-temporal patterns such as the main route and spatio-temporal channel followed by most of trajectories of the set. Using these patterns, unusual situations can be detected. Furthermore, a mobile object's behaviour can be defined by comparing its positions with these spatio-temporal patterns. In this article, this technique is applied to ships' movements in an open maritime area. Unusual behaviours such as being ahead of schedule or delayed or veering to the left or to the right of the main route are detected. A case study illustrates these processes based on ships' positions recorded during two years around the Brest area. This method can be extended to almost all kinds of mobile objects (pedestrians, aircrafts, hurricanes, ...) moving in an open area.

1 INTRODUCTION

More and more mobile devices are equipped with tracking systems broadcasting accurate information about their movements. Those sensors generate a large amount of data which can be stored in spatio-temporal databases ($STDB$) in order to be analysed. Mobile objects monitoring is commonly used in various fields such as the study of meteorological phenomena, animal migration (Lee et al., 2008), crowd or pedestrian displacement (Knorr et al., 2000), vehicle trips (cars, planes, ships) (Baud et al., 2007, Giannotti et al., 2007). This mobile object monitoring can be linked with intelligent system analysis to improve the system's performance (to ease freight transport planning, for example). Using spatio-temporal databases led to new capabilities. Indeed, the displacement of these mobile objects can be analyzed over a long period of time in order to deduce the general behaviour of mobile objects following the same route. Detecting outliers that behave in an unusual way in such large amounts of data is a very active research field linked to data mining and statistical analysis.

Assuming that mobile objects following a same itinerary behaves in a similar optimized way, these behaviours can be deduced by data mining on $STDB$. It pave the way to analysis of mobile objects' trajectories and detection of unusual behaviour. Different ways to detect outliers in a large dataset could be applied to our issue. These outlier detections are classified according to the method used which can be based on distribution (Barnett and Lewis, 1994), distance (Knorr et al., 2000, Ramaswamy et al., 2000, Lee et al., 2008) or density (Aggarwal and Yu, 2001, Papadimitriou et al., 2003, D'Auria et al., 2006, Kharrat et al., 2008, Lee et al., 2008). The distance and density methods are merged in Lee's works (Lee et al., 2008) based on a "partition and detect framework" that identifies subsets of trajectories which have fewer neighbours. These parts of trajectories are considered as locally unusual regarding density and distance criteria. Unfortunately, time criteria is not taken into account in these methods. In this paper, we propose a process to qualify the position of a mobile object both on spatial and temporal criteria.

The main goal of this study is to define spatio-temporal analysis tools to describe mobile objects' behaviour. Assuming that similar mobile objects following the same itinerary behave in a similar way and move along an optimized main route, it could be useful to analyse the trajectories of these objects in order to deduce spatio-temporal patterns and then, to qualify their behaviour by comparing their trajectories to these patterns. Such trajectory analysis tools coupled with a visualization process could be useful for traffic monitoring operators to focus on outliers (mobile objects behaving in an unusual way) for safety purpose. In some areas, mobile objects' traffic is very dense and the amount of data to be processed in real time can be important. In order to create these tools, notion of trajectories of mobile objects following a same itinerary have to be defined. Then homogeneous subset of trajectories of mobile objects following a same itinerary have to be extracted from the $STDB$. The main goal of this study is to analyze this subset of trajectories in order to infer the behaviour of mobile objects following a similar path.

The paper is organized into 6 main sections. The first section of this article introduces the main goal of this paper and related research in data mining mobile objects movement. The second section proposes a method to extract and filter trajectories of mobile objects following a similar itinerary. The third section deals with statistical computation of spatial and spatio-temporal route and channel. This section also describes how to qualify the position of a mobile object following an itinerary using these tools. The fourth section present the results of our process applied to a case study focused on passenger ships in the Brest area followed by some discussions. Finaly, the last section concludes pointing to further future work.

2 SPATIO-TEMPORAL TRAJECTORIES EXTRACTION AND FILTERING

The general process proposed to identify unusual mobile object behaviour is presented in figure 1. First of all, information about mobile objects' positions are stored into a spatio-temporal database (figure 1 step 2). The zone graph of the area of interest is set up in the knowledge database (figure 1 step 5). A cluster of trajectories of similar mobile objects following the same itinerary is extracted from the $STDB$ (figure 1 step 3). Then, a statistical analysis is performed to compute spatio-temporal patterns (figure 1 step 4) which are then stored in a knowledge database (figure 1 step 5). Each new position of a mobile object can be compared with these spatio-temporal patterns in order to qualify the mobile object behaviour (figure 1 step 6). The next sections of this article describe this general process.

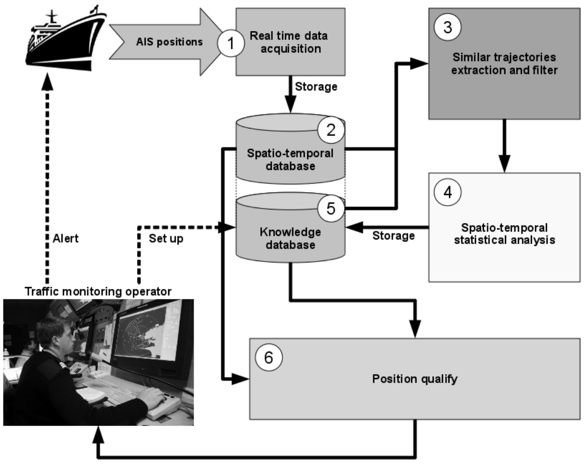

AIS positions
1
Real time data acquisition
Storage
2
Spatio-temporal database
3
Similar trajectories extraction and filter
4
Spatio-temporal statistical analysis
5
Knowledge database
Storage
Alert
Set up
Traffic monitoring operator
6
Position qualify

FIG. 1: General process of spatio-temporal trajectory analysis

To analyze a large amount of moving objects' trajectories, both spatial and temporal information about their positions must be stored. $STDB$ are employed to store sets of discrete data having spatial and temporal properties (Güting, 1994). These $STDB$ offer tools to perform queries on these sets of data on spatial and temporal criteria. Formally, the position of a moving object (O) is composed of spatial coordinates with a timestamp corresponding to the date on which the moving object was at that position (absolute time). So, the trajectory T_o of a mobile object O can be defined as a sequence of temporally ordered positions P_{oj} so that $T_o = (P_{od}, ..., P_{oj}, ..., P_{oa})$ where P_{od} stands for the departure position of the trajectory and P_{oa} for the arrival one.

2.1 Definition of a zone graph

In order to deduce main routes, the trajectories of same-type mobile objects following the same itinerary are extracted from the $STDB$ and then grouped together. The concept of itinerary can be defined as an ordered sequence of spatial zones. In our study, the space is a wide open area which allows mobile objects to navigate from one place to another using the most effective path.

The concept of zone graph used in this section will now be formalized. Zones of this graph represent important areas. These important areas can be manually defined by an operator according to various criteria such as regulations (waiting areas, traffic channels, restricted areas), geography (obstacles, isthmuses, straits, inlets), economy (shops, loading sites, ports, fishing areas). This directed zone graph can be used to describe an itinerary. Using the previously defined vertices of this zone graph (G_Z), an itinerary (I) is defined as a sequence of ordered zones linked by arcs (a path of the zone graph). An itinerary is made up of at least one arc, therefore it has a departure zone (Z_D) and an arrival zone (Z_A). A trajectory T_o follows an itinerary I through the vertices of the zone graph G_Z if it satisfies the following conditions :
Let an itinerary be defined as $I = \{Z_D, ..., Z_i, ..., Z_A\}$
Let a trajectory be defined as $T_o = (P_{od}, ..., P_{oj}, ..., P_{oa})$
Trajectory T_o follows the itinerary I if :

$$\forall Z_i \in I, \exists P_{oj} \in T_o, P_{oj} \subset Z_i \tag{1}$$

$$\begin{aligned}\forall P_{oj} \in T_o \wedge P_{oj} \subset Z_l, \forall P_{ok} \in T_o \wedge P_{ok} \subset Z_m,\\ Z_l <_I Z_m \rightarrow P_{oj} < P_{ok}\end{aligned} \tag{2}$$

$$\forall P_o j \in T_o \wedge P_{oj} \subset Z_i \rightarrow Z_i \in I \tag{3}$$

$$P_{oj} \subset Z_D \rightarrow P_{oj} = P_{od} \tag{4}$$

$$P_{oj} \subset Z_A \rightarrow P_{oj} = P_{oa} \tag{5}$$

In other words, for each zone of the itinerary I, there is at least one position P_o of the trajectory T_o inside this zone (Eq. 1) which respects the time order relation previously defined (Eq. 2). Taking into account the frequency of trajectory samples and the speed of the mobile object, trajectories that cross a zone of the graph should have at least one position within this zone. No other position P_o of the trajectory T_o is within a zone that does not belong to the itinerary (Eq. 3). Only the first position P_{od} of the trajectory belongs to the departure area of the itinerary Z_D (Eq. 4). In the same way, only the last position P_{oa} of the route belongs to the last area of the route Z_A (Eq. 5).

2.2 Extraction of an homogeneous group of trajectories

Now that the concepts of trajectory and itinerary have been formalized, the criteria used to extract trajectories following the same arc A of an itinerary I can be detailed. The goal of this part is to extract the $STDB$ trajectories of same type T objects moving along the same arc A of an itinerary I. This set is called homogeneous group of trajectories of same type mobile objects following the same arc of an itinerary (HGT_{AIT}). Thus, the first selection criterion is the type of the mobile object. The second selection criterion is a geographical one. The first position of the trajectory must be the only one within the departure zone (Z_D) of the arc (Eq. 4), and the last position of the trajectory must be the only one within the arrival zone (Z_A) of the arc (Eq. 5). Finally, the last selection criterion used is time. Some moving objects can follow this itinerary periodically, these different trajectories can be distinguished using a time interval. These selection criteria applied to the $STDB$ are used to extract all the spatio-temporal positions of a same mobile object in the meantime between positions P_{od} and P_{oa} forming the trajectory of the mobile object (Tr_o) ordered by timestamp. Finally, the trajectory should not intersect zones of the graph G_Z that do not belong to the itinerary I (Eq. 3). All valid trajectories previously extracted from the $STDB$ compose the HGT_{AIT} to be analyzed.

2.3 Erroneous trajectory filtering

Once the HGT_{AIT} has been extracted from the database, trajectories with an important gap between two consecutive positions or erroneous positions are filtered from the HGT_{AIT} in order to improve statistical analysis. First of all, trajectories containing important communication loss compared to normal transmission rate of the studied group of trajectories are discarded. Then, some tracks may contain erroneous positions due to a malfunction of the geolocation system or transmission errors. These erroneous positions can be detected using the calculated speed of the position compared to the maximum speed of a moving object of this type. Trajectories having either erroneous positions or transmission gaps are removed from the HGT_{AIT}.

2.4 Spatial shifting

In order to compute trajectories for which departure and arrival positions are independent from time of transmission, starting and ending positions of the trajectory within the departure and arrival zones must also be filtered. Without this filtering, a bias can be measured in the spatio-temporal patterns defined in the next section of this article. The cloud of initial starting positions of the HGT_{AIT} is represented in figure 2.a. The new starting positions are computed by interpolation between a virtual starting line (border of Z_D) and each trajectory of the HGT_{AIT}. The same process is applied to the arrival zone Z_A. The result of space shifting applied to our example is illustrated in figure 2.b.

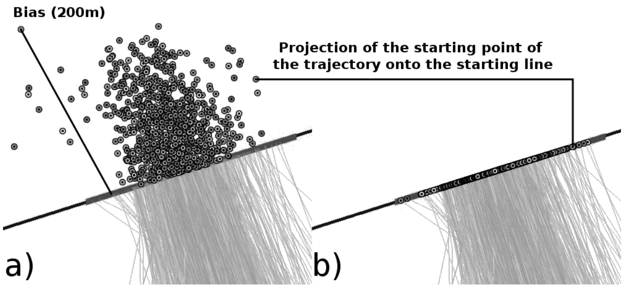

Bias (200m)
Projection of the starting point of the trajectory onto the starting line
a)
b)

FIG. 2: Spatial shifting of trajectories

2.5 Spatio-temporal Douglas & Peucker filter

Once the spatial shifting is done, in order to optimize the computation time, trajectories can be both indexed according to a spatio-temporal method (Rasetic et al., 2005) and simplified using a filter initially proposed by Douglas & Peucker (Douglas and Peucker, 1973). Several different algorithms are based on this work. Some of them have been compared by Wu (Wu and Pelot, 2007). In this study, a spatio-temporal Douglas & Peucker filter (Bertrand et al., 2007, Cao et al., 2006, Meratnia and de By, 2004) is used. The goal of this filter is to retain only significant positions of a trajectory while keeping information about speed or heading changes. To do this, the greatest distance d_{max} between each positions P_i of the trajectory and their spatio-temporal projections P'_i on the line between the starting positions P_d and arrival P_a is calculated. If this distance d between P_i and P'_i exceeds a threshold, the farthest position P_{max} is retained. The trajectory is then split at that position (P_{max}) and the algorithm is recursively applied to both trajectory subparts. If the distance d is smaller than the threshold, only positions P_d and P_a are kept. This algorithm also filters inaccuracies of measuring devices (Bertrand et al., 2007).

2.6 Position normalized relative timestamps computation

In order to ease distance and time comparison between trajectories, a relative timestamp is computed for each position of a trajectory. Timestamps of positions are very useful to compute speed and order each position within a trajectory. Initial positions of trajectories are all set up with an absolute timestamp (t_A). In order to compare these trajectories, a new relative timestamp (t_R) is computed for each position. This relative timestamp stands for the interval of time since the starting position of the trajectory. Thereby, every starting position of the trajectories of the HGT_{AIT} have a null relative timestamp ($t_0 = 0$). Finally, to avoid spatial distortions introduced by slightly different speeds of mobile objects of the HGT_{AIT}, timestamps of all the trajectories of the HGT_{AIT} must be normalized. This relative normalized timestamp t_{NR} stands for the normalized time elapsed since the starting position of a trajectory. To compute this relative normalized timestamp, first of all, the median duration D_{med} of the HGT_{AIT} is calculated. The choice of the median duration is less disturbed by outliers. Using this duration, a normalization process is applied to all trajectory positions so that each trajectory begins at a time $t_0 = 0$ and ends at the same relative normalized time $t_m = t_0 + D_{med}$.

3 STATISTICAL ANALYSIS OF TRAJECTORIES

Once the HGT_{AIT} has been extracted and filtered, it is worthwhile to perform a statistical analysis of this group of trajectories. This statistical analysis aims at qualifying positions and trajectories of moving objects following an itinerary using spatial and temporal criteria. To do this, spatio-temporal patterns are defined to compare positions and trajectories of a moving object with patterns which stand for normal behaviour of mobile objects of the same type following the same itinerary.

3.1 Main route computation

First of all, a main route followed by most of the trajectories of the HGT_{AIT} is computed by statistical analysis. The first stage of this process consists in setting up a new relative normalized timestamp for each position of each trajectory of the HGT_{AIT} as explained in section 2.6. Then, for each position of each trajectory of the HGT_{AIT}, positions of other trajectories of the HGT_{AIT} are interpolated using their normalized time. This second step of the main route computation generates a subset of positions at each normalized time (note that only meaningful positions kept by the spatio-temporal Douglas & Peucker algorithm are used, so that the computation process is only applied on subparts of trajectories where mobile object behaviour changes). Median positions are computed at each normalized time using median values of coordinates (latitudes and longitudes) of each position subset. Then, these computed median positions are ordered according to their normalized time to create the main route of the itinerary. Finally, this main route is also filtered using the Spatio-temporal Douglas & Peucker algorithm (section 2.5). Algorithm 1 summarizes the main route computation steps.

Algorithm 1 Main route computation

```
Require:
for each trajectory $Tr$ of the $HGT_{AIT}$ do
    Delete erroneous trajectories
    Spatial shifting of starting and ending positions
    Douglas_Peucker_ST(Trajectory $Tr$)
    Temporal normalization using median duration $t_m$
end for
Algorithm Main_Route_Computation($HGT_{AIT}$)
for each trajectory $Tr_i$ of the $HGT_{AIT}$ do
    for each position $P_i$ of $Tr_i$ do
        Let $tn_i$ be the normalized time of $P_i$
        for each other trajectoiries $Tr_j$ of the $HGT_{AIT}$ do
            Interpolate the positions $P_j$ at normalized time $tn_i$
            Add $P_j$ to the subset of positions $EP_i$
        end for
        Compute median position $P_{med}$ of $EP_i$
        Add $P_{med}$ to the main route $R_{IT}$ at normalized time $t_{ni}$
    end for
end for
return Douglas_Peucker_ST(Trajectory $R_{IT}$)
```

3.2 Spatial channel computation

As the studied mobile objects move in an open area, some of them can move away from the main route. These slight deviations must be distinguished from outliers. The goal of the spatial channel computation is to detect outlier positions of trajectories that spread out of this spatial channel. These unusual deviations affect a small subset of positions within some trajectories of the HGT_{AIT}. In order to distinguish normal and unusual trajectories, a spatial channel is calculated using a statistical analysis of all the trajectories of the HGT_{AIT} compared to the main route.

Positions of all trajectories of the HGT_{AIT} are ordered by distance and side to the main route using crossing positions between trajectories of the HGT_{AIT} and the line perpendicular to the heading (LPH) of each previously calculated position of the main route. On each side of the main route, the positions of trajectories which intersect with the LPH are ordered by distance from the main route's position. The sorted position corresponding to the ninth decile of each side of the main route is used to create the border of the channel. Positions outside of this spatial limit are considered as outliers. The choice of this statistical decile provides a channel within which most of the mobile objects following this itinerary move along. Algorithm 2 summarizes the different steps used to calculate the spatial channel (SC).

Algorithm 2 Spatial channel computation

```
Algorithm Spatial_Channel_Computation($HGT_{AIT}$)
for each position $P_i$ of the main route $R_{IT}$ do
    Compute line $[LPH]$ perpendicular to the heading of $P_i$
    for each trajectory $Tr_j$ of the $HGT_{AIT}$ do
        Compute crossing position $P_i'$ between $Tr_j$ and $[LPH]$
        if $P_i'$ is right $P_i$ then
            Store $P_i'$ in array $A_{right}$
        else
            Store $P_i'$ in array $A_{left}$
        end if
    end for
    Sort $A_{right}$ and $A_{left}$ by distance to $P_i$
    $Pfi_{right}$ = ninth decile of $A_{right}$
    $Pfi_{left}$ = ninth decile of $A_{left}$
    Set $Pfi_{right}$ and $Pfi_{left}$ timestamp to $P_i$ one
    Add $Pfi_{right}$ to the right border $Tr_{right}$ of spatial channel
    Add $Pfi_{left}$ to the left border $Tr_{left}$ of spatial channel
end for
```

3.3 Spatio-temporal zones calculation

Given that a moving object is travelling in the spatial channel of a main route, one other interesting element is to know wheter this object is on time compared to the main route. As for positions, temporal zones can be computed in order to temporally qualify the mobile object's position (ahead of schedule, on time, late).

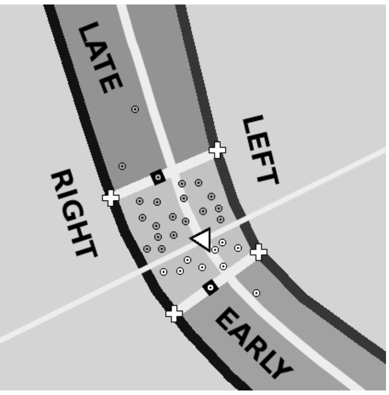


FIG. 3: Spatio-temporal zone at a relative time

To generate these temporal zones, once the spatial channel is computed, the trajectories outside the spatial channel are first removed from the HGT_{AIT}. Then, for each position of the main route P_{Ri} (represented by a white triangle on figure 3) using its relative time tP_{Ri}, all other positions of the HGT_{AIT} are interpolated. These positions are converted into a new polar system using P_{Ri} as pole and P_{Ri}'s heading as polar axis. This conversion defines a total order for each position subset according to distance. Distances r_{ij} and azimuth θ_{ij} of each interpolated position from the HGT_{AIT} are then divided into two subsets according to the azimuth $((\theta_{ij} > 90^\circ \wedge \theta_{ij} <= 270^\circ) \rightarrow P_j delayed)$ and then sorted by distance (white dots for early positions and grey dots for late ones as shown in figure 3). Finally the positions whose distances r_{ij} match the ninth decile of each subset (P_{Li} for late positions and P_{Ei} for early positions) are selected (shown as black squares in figure 3).Then, the projected positions of P_{Ei} and P_{Li} on the main route are computed (P'_{Ei} and P'_{Li}). The crossing positions (white crosses in figure 3) between the spatial channel and the lines perpendicular to P'_{Ei} and P'_{Li} are used to create the temporal normality zone Z_N for each tP_{Ri}. Spatial channel and temporal zones at each relative time can be combined to create the spatio-temporal channel which is then stored in the knowldege database. As new positions are frequently acquired by the system, this spatio-temporal channel can be improved by updating it periodically.

Algorithm 3 Temporal zones computation

```
Require:
Let $R_{IT}$ be the main route of the $HGT_{AIT}$
Let $SC$ be the spatial channel of the $HGT_{AIT}$
Algorithm Temporal_Zone_Computation($HGT_{AIT}$)
Remove every trajectory of the $HGT_{AIT}$ which lies out of $SC$
for each position $P_{Ri}$ of the main route $R_{IT}$ do
    Let $tP_{Ri}$ be the relative time of $P_i$
    Let $H_{Ri}$ be the heading of $P_i$
    Change the polar system using $P_{Ri}$ as pole and $H_{Ri}$ as polar axis
    for each trajectory $Tr_j$ of the $HGT_{AIT}$ do
        Interpolate position $P_j$ of $Tr_j$ at relative time $tP_{Ri}$
        Compute $r_{ij}$, the distance between the pole and $P_j$
        Compute $\theta_{ij}$, the angle between the polar axis and $P_j$
        if $(\theta_{ij} > 90^\circ \wedge \theta_{ij} <= 270^\circ)$ then
            Store $r_{ij}$ in array $A_{late}$
        else
            Store $r_{ij}$ in array $A_{early}$
        end if
        Sort $A_{late}$ and $A_{early}$ by distance $r_{ij}$ to $P_i$
        $r_{late}$ = ninth decile of $A_{late}$
        $r_{early}$ = ninth decile of $A_{early}$
        Store $r_{late}$ and $r_{early}$ for relative time $tP_{Ri}$
        Using $P_{Ri}$ speed, $r_{late}$ and $r_{early}$, interpolate positions on $SC$ and $R_{IT}$
        Create normality zone $Z_N$ using interpolated positions
    end for
end for
```

4 EXPERIMENT

This section presents the results of the process exposed in previous sections applied to a maritime context. The shipping freight traffic is constantly increasing and traffic surveillance operators can have to visualy monitor up to 250 ships displayed simultaneously on theirs displays. For safety purposes, ships are fited out with *Automatic Identification System* (AIS) to track ships' positions in real time using GPS receivers and VHF transmission systems (IMO, 2007). The spatio-temporal database studied in this example contains 1005 ships and 4 821 447 positions stored since May 2007 in the Brest area (Iroise sea). This spatio-temporal database works using a PostgreSQL/PostGIS server. Each position is associated to a ship whose features are also stored in this database.

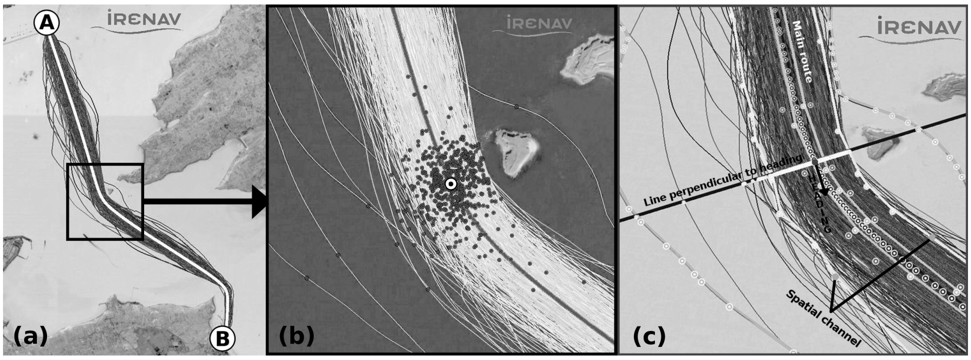


FIG. 4: Main route and spatial channel computation, position cloud at same normalized time

Using the $STDB$ spatial extraction tools, ship trajectories can be distinguished and extracted from this database. As explained in section 2.1, a spatial zone (Z) can be defined and represented by geometric areas ($Z.g$) of points of interest. In a place where mobile objects usually stop or interact, where traffic is limited by the geography or by regulations, a zone is defined. As the mobile object move in an open space, there is no forced network between these zones (except for limited traffic due to regulation or geography), the space is a wide open area which allows ships to navigate from one place to another using the most effective path. A position P_o is included into an area Z_i if its coordinates are included into the geometrical surface $Z_i.g$ of the zone. The geometry of the zone must also be large enough to include at least one position of each trajectory that cross this zone (otherwise interpolated positions may have to be calculated). The zone graph of our example is depicted in figure 5 where labeled white circles stands for zones of interest.

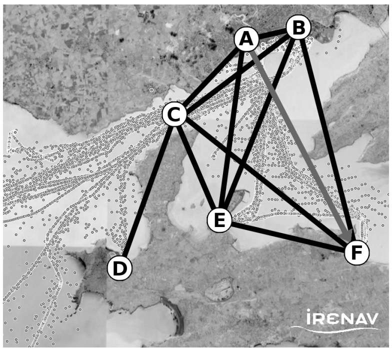


FIG. 5: Zone graph of the $STDB$

Thus, the itinerary shown in Figure 5 by the arc (A, F) (Brest Arsenal → Lanvéoc Naval Academy) of the zone graph G_Z is different from the one represented by the string (A, E, F) (Brest Arsenal → Ile Longue → Lanvéoc Naval Academy). The zone graph is incomplete and directed, all its vertices are not connected directly with each other by an edge and the way back of the itinerary may be different as navigation rules can set distinct channels in order to avoid collisions. Once set up, this zone graph is saved in the knowledge database (figure 1 step 5). The numerous dots shown in figure 5 represent positions of ships. The main routes used by most of the ships are visually noticeable as dense areas.

Once the graph zone established, an homogeneous group of trajectories is extracted from the $STDB$ as explained in section 2.2. The first selection criterion used to extract this set of trajectories is the type of the mobile object. Applied to our maritime example, only "passenger ships" are selected (30 vessels out of 1005) then the data mining extraction method identified 554 trajectories of passenger ships following the itinerary "Brest Arsenal ⇒ Lanvéoc Naval Academy" represented by the arc (A, F) on figure 5. Next, trajectories containing important communication loss compared to normal transmission rate (No position for 1 minute for the AIS system), erroneous positions or transmission gaps are discarded from the HGT_{AIT} as explained in section 2.3. Among 554 trajectories, 506 trajectories were kept after filtering out erroneous trajectories, which is enough to apply statistical analysis to this set of trajectories. The starting and ending position of the remaining trajectories are spatialy shifted as exposed in section 2.4. This spatial shifting avoid a maximum 200-meter distance between farthest starting positions and the projected one on the starting line as shown on figure 2.a. The spatio-temporal Douglas & Peucker filter exposed in section 2.5 applied to the HGT_{AIT} reduced the number of positions from 104 201 to 16 110 (compression rate of 84.54 %) for a threshold of 10 m (precision of a GPS device).

The extracted and filtered HGT_{AIT} composed of 506 trajectories plotted in black in figure 4.a is then used to compute spatio-temporal patterns presented in sections 3.1 and 3.2. Looking at figure 4.a visually shows that same-type mobile objects with the same itinerary globally follow a main route. The cloud of dark dots shown in figure 4.b represents the subset of positions at a same normalized time, the large white dot indicates the median position of the whole subset. All these median positions ordered by theirs normalized time compose the main route plotted in white in figure 4.

Once the main route calculated, the spatio temporal channel can be statisticaly computed using algorithms 2 and 3 presented in sections 3.2 and 3.3. Figure 4.c shows the calculated borders of the spatial channel applied to our example. Thus, unusual positions outside the spatial channel can be highlighted for each HGT_{AIT}. The distances between the main route and the spatial channel borders (right and left) are different. Indeed, it is easier for a moving object to deviate outward than to get closer to an obstacle in an open space area. Similarly, the width of the channel provides information about the trajectories spreading from the main route. In our example, this spreading is narrower at the start, the end and in the curves of the itinerary. However, in straight parts of trajectories, spatial channel width increases. The choice of the statistical decile used to compute the spatio-temporal channel gives a more or less wide spread of this channel within which most of the mobile objects following this itinerary move along.

Finaly, as shown in Figure 3, positions of the trajectory of a passenger ship going from Brest to Lanvéoc can be qualified using the five spatio-temporal zones previously-defined in section 3.3. Only 30 positions are displayed in order to keep Figure 3 readable. Positions of the ship are spatially and temporally qualified in order to alert the traffic operator about the unusual behaviour of a ship. The operator can then focus on a few ships within a huge set of vessels cruising in a wide area. Note that the distances between the main route's position and the early and late zone borders are quite different as it is more frequent for a moving object to be delayed than to be ahead of schedule. Moreover, a position outside the spatial channel cannot be temporally qualified as early or late, indeed the moving object moving away from the route can either take a shortcut or make a detour.

5 DISCUSSIONS

The novelty of the method is the use of meta-knowledge (HGT_{AIT}, main route, spatio-temporal channels) to describe the behaviour of mobile objects following an itinerary on both spatial and temporal criteria. Moreover, these meta-data could be used to qualify new mobile object's positions in real time. The graph zone used to define arc of itineraries can bridge this study to the network based approach of trajectory analysis. However, matching a position to an itinerary in real time remains a complex problem to solve as some arcs of an itinerary can be shared. Previous position of the mobile object coupled with its destination can facilitate the matching to an itinerary but every time a new position is obtained, this matching may change. Tracks for future research include extending our analysis to sections of trajectories. By analogy to the zone graph, it would be interesting to split trajectories into subsections to enhance analysis of the behaviour of a ship on a subpart of the trajectory sharing common properties (speed, heading, rate of turn...). Sections of trajectories could be compared to the main routes. Furthermore, computation time could be decreased by filtering the whole $BDST$ using the Douglas and Peucker spatio-temporal algorithm and adding a trajectory index. Indeed, 50,04% of CPU was used to extract and filter the HGT_{AIT}. Finally, the main selection criteria used in this analysis is the type of the ship which does not take into account the environment of the ship (such as the tide, wind or season).

6 CONCLUSION

This article focused on the specific problem of outlier detection in mobile object displacements in an open area. It was applied to a maritime context as shown in our case study based on an important dataset. Once the notion of itinerary and trajectory following an arc of an itinerary formaly defined, a general process to qualify mobile object behaviour based on spatio-temporal data mining was defined as previously exposed in figure 1. First of all, position data are acquired and a knowledge database is set up with the zone graph. Then, trajectories of same-kind mobile objects are clustered according to arcs of itineraries. A statistical analysis of each cluster allows to define the main route and spatio-temporal channel of this cluster. These meta-data are stored in the knowledge database. Each new position can be spatially and temporally qualifyed. These processes have been tested on an important dataset applied to the maritime context in different area. Thus, statistical analysis of a GHT_{AIT} gives us information about mobile object's behaviour. Thanks to the spatio-temporal channels, positions of a trajectory can be qualified on both spatial and temporal criteria. It could be worthwhile to validate this study by providing these tools to traffic surveillance operators who can monitor up to 250 ships displayed simultaneously in order to decrease the operator's cognitive load. However, real time analysis tools have not yet been implemented to this prototype.

RÉFÉRENCES

Aggarwal, C. C. and Yu, P. S., 2001. Outlier detection for high dimensional data. SIGMOD 30(2), pp. 37–46.

Barnett, V. and Lewis, T., 1994. Outliers in Statistical Data. John Wiley & Sons New York.

Baud, O., El-Bied, Y., Honore, N. and Taupin, O., 2007. Trajectory comparison for civil aircraft. In : Aerospace Conference, 2007 IEEE, pp. 1–9.

Bertrand, F., Bouju, A., Claramunt, C., Devogele, T. and Ray, C., 2007. Web and Wireless Geographical Information Systems. Lecture Notes in Computer Science, Vol. 4857, Springer Berlin / Heidelberg, chapter Web Architecture for Monitoring and Visualizing Mobile Objects in Maritime Contexts, pp. 94–105.

Cao, H., Wolfson, O. and Trajcevski, G., 2006. Spatio temporal data reduction with deterministic error bounds. VLDB Journal 15, pp. 221–228.

D'Auria, M., Nanni, M. and Pedreschi, D., 2006. Time-focused density-based clustering of trajectories of moving objects. Journal of Intelligent Information Systems 27(3), pp. 267–289.

Douglas, D. H. and Peucker, T. K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica : The International Journal for Geographic Information and Geovisualization 10, pp. 112–122.

Giannotti, F., Nanni, M., Pinelli, F. and Pedreschi, D., 2007. Trajectory pattern mining. In : KDD '07 : Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp. 330–339.

Güting, R. H., 1994. An introduction to spatial database systems. VLDB Journal 3, pp. 357–399.

IMO, 2007. Development of an e-navigation strategy. Technical report, International Maritime Organization.

Kharrat, A., Popa, I. S., Zeitouni, K. and Faiz, S., 2008. Clustering Algorithm for Network Constraint Trajectories. Springer Berlin Heidelberg, chapter Clustering Algorithm for Network Constraint Trajectories, pp. 631–647.

Knorr, E. M., Ng, R. T. and Tucakov, V., 2000. Distance-based outliers : algorithms and applications. The VLDB Journal 8(3-4), pp. 237–253.

Lee, J., Han, J. and Li, X., 2008. Trajectory outlier detection : A partition-and-detect framework. In : Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on Data Engineering, pp. 140–149.

Meratnia, N. and de By, R. A., 2004. Spatiotemporal compression techniques for moving point objects. In : Advances in Database Technology - EDBT 2004, Lecture Notes in Computer Science, Vol. 2992, Springer Berlin / Heidelberg, pp. 561–562.

Papadimitriou, S., Kitagawa, H., Gibbons, P. and Faloutsos, C., 2003. Loci : fast outlier detection using the local correlation integral. In : Data Engineering, 2003. Proceedings. 19th International Conference on Data Engineering, pp. 315–326.

Ramaswamy, S., Rastogi, R. and Shim, K., 2000. Efficient algorithms for mining outliers from large data sets. In : SIGMOD '00 : Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ACM, New York, NY, USA, pp. 427–438.

Rasetic, S., Sander, J., Elding, J. and Nascimento, M. A., 2005. A trajectory splitting model for efficient spatio-temporal indexing. In : VLDB '05 : Proceedings of the 31st international conference on Very large data bases, VLDB Endowment, pp. 934–945.

Wu, Y. and Pelot, R., 2007. Geomatics Solutions for Disaster Management. Jonathan Li and Sisi Zlatanova and Andrea G. Fabbri, chapter Comparison of Simplifying Line Algorithms for Recreational Boating Trajectory Dedensification, pp. 321–334.

SPATIO-TEMPORAL ANALYSIS OF PRECIPITATION AND TEMPERATURE DISTRIBUTION OVER TURKEY

P. A. Bostan [a], Z. Akyürek [b]

[a] METU, Geodetic and Geographic Inf. Tech. Natural and App. Sciences, 06531 Ankara, Turkey - aslantas@metu.edu.tr
[b] METU, Civil Eng. Dept., 06531 Ankara, Turkey - zakyurek@metu.edu.tr

"14th International Symposium on Spatial Data Handling"

KEY WORDS: Precipitation, Temperature, Co-kriging, Ordinary Least Square, Geographically Weighted Regression

ABSTRACT:

In this study**,** mean annual precipitation and temperature values observed at 225 meteorological observations over Turkey are used to disclose spatial distribution of mean annual precipitation and temperature values. Data components were obtained from the Turkish State Meteorological Service for 34 years period (1970-2003). The basic objectives of the study are: to infer the nature of spatial variation of precipitation and temperature over Turkey based on meteorological observations and to model the pattern of variability of these data components by using secondary variables extracted from SRTM and river network. Modeling the spatial distribution of data sets is implemented with Co-kriging (COK), Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR) techniques with using secondary variables such as elevation, aspect, distance to river, roughness, drop (elevation differences between station and grid), sd-grid (standard deviation of 5*5 km grid), and plan-profile curvature. Correlations among the listed variables were analyzed and highly correlated ones were removed from the analysis. The study found a presence of high spatial non-stationary in the strength of relationships and regression parameters. The co-kriging interpolation method gave strong relationship for temperature (r^2= 0.823) but comparatively weak relationship for precipitation (r^2= 0.542). OLS method resulted with lower relationships for temperature (r^2= 0.68) and for precipitation (r^2= 0.3). The highest adjusted r^2 values were obtained with GWR method; 0.96 for temperature and 0.66 for precipitation.

1. INTRODUCTION

Geographical variables can not be measured at all part of space, therefore researchers who work with those variables generally use interpolation techniques in some part of their studies. Thus observations are taken at points and spatial interpolation is used to obtain a full spatial coverage. There are many examples such as; soil physical properties, air quality, groundwater pressure, plant species abundance (Heuvelink, 2006). At the areas that haven't observed values discrete precipitation values are needed. For this purpose thiesen polygons, regression analysis, inverse distance weighted method, trend analysis, isohyets curves etc. can be used for prediction. However these methods are generally inadequate and tendentious and they do not give variation information about predictions (Çetin and Tülücü, 1998). A common theme in many similar studies is that techniques that make use of the relation between precipitation and secondary data, such as elevation data, often provide more accurate estimates than approaches that are based only one parameter like precipitation measurements (Lloyd, 2005). This study is concerned with mapping annual average precipitation and temperature for Turkey from sparse point data using Co-kriging (CO) technique, global ordinary least square regression technique known as OLS and a local regression technique known as Geographically Weighted Regression (GWR) methods. By using the spatial relationships between meteorological observations and variables derived from elevation, optimum spatial distributions of mean annual precipitation and temperature are aimed to be defined.

2. STUDY AREA AND DATA

The study area covers all Turkey. The data used in this study is obtained from the Turkish State of Meteorological Service. Data consist of mean monthly precipitation and temperature values measured at climate stations between 1970-2003 years. Analyses are performed on annual average values. Data from 225 meteorological stations as illustrated in Figure (1) were selected to use in the analyses because of their consistent number of data years and length in the observation period.

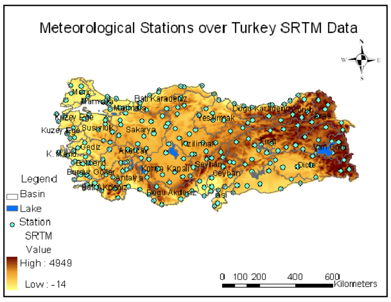


Figure 1: Study area and meteorological stations within the basins

2.1. Variables Obtained From Digital Elevation Model

The use of digital elevation data to guide the interpolation of monthly temperatures is becoming accepted practice, but the spatial variability of air temperature is significantly affected by topographic relief together with several other geographical factors such as latitude and distance to coast line (Rigol et al., 2000). Therefore in this study as well as topographical parameters, some geographical variables were selected as the ancillary variables in finding the distribution of precipitation and temperature. Totally 9 topographical and geographical variables as listed in Table 1, are used as additional input data for spatial interpolation analyses. All the variables except station elevation are obtained from elevation (SRTM 3 arc second spatial resolution) and river network digitized from 1/250000 scaled map.

Variable	Description
Elevation	Height of meteorological stations
Aspect	Function of aspect derived from elevation
Curvature	Degree of curvature derived from elevation
Roughness	Elevation cell height minus 5 km grid mean height
Drop	Elevation cell height minus 5 km grid minimum height
Sd_Grid	Stan. deviation of elevation in each 5 km*5 km grid
River	Distance to nearest river, calculated by using the Euclid. distance computation method

Table 1. Ancillary data derived from elevation and river network.

3. METHODOLOGY

i. Co-kriging (CO) method is an extension of ordinary kriging that takes into account the spatial cross-correlation from two or more variables. The usual situation is one where the primary or target variable, $Z_u(x)$, has been measured at many fewer places, x, than the secondary one, $Z_v(x)$, with which it is co-regionalized.

The influence of the secondary information on estimating *Z* depends on (i) the correlation between primary and ancillary variables, (ii) the spatial continuity of the attributes, and (iii) the sampling density and spatial configuration of primary and ancillary variables (Simbahan et al., 2005).

ii. Ordinary Least Squares (OLS) provides a global model of the variable to predict by creating a single regression equation to express that process. It can be regarded as the starting point for all spatial regression techniques (ESRI website).

Theoretical background is summarized below:

Suppose that X is an independent and Y is a dependent variable. We make n number of observations on two variables. The relationship between Y and X can be regressed using OLS as follows:

$$Y = X\beta + \varepsilon \qquad (1)$$

where Y is a vector of the observed dependent variable, X is a known model matrix including a column of 1 (for intercept) and n independent variables, β is a vector of unknown fixed-effects parameters, and ε is a random error term. The OLS estimate of β is obtained by the least-squares method as shown in Eq. (2).

$$\beta = (X^T X)^{-1} X^T Y \qquad (2)$$

where superscript T denotes the transpose of a matrix. The relationship represented by Eq. (1) is assumed to be universal or constant across the geographic area (Zhang et al, 2005).

iii. Geographically weighted regression (GWR) is a local statistical technique to analyze spatial variations in relationships which is based on Tobler's (1970) "First law of geography".

The simple linear model usually fitted by ordinary least squares (OLS) methods is given in Eq. (3).

$$P= Co+C1(H)+C2(A)+e \qquad (3)$$

- *P*=rainfall (mm)
- *Co*=rainfall at sea level (mm) and flat area
- *C1*= dimensionless rate of increase in rainfall with altitude, or height coefficient (mm/m)
- *H*=station altitude (m)
- *C2*= change of rainfall with aspect
- *A*= aspect of that station
- *e*= error term

In GWR by retaining the same linear model, we can allow parameters, the intercept constant, the height and aspect coefficient to change, or 'drift', over space. That is, if (*x*, *y*) is a coordinate pair, the simple linear model of Equation (1) can be expanded to Eq. (4).

$$P= Co(x,y)+C1\ (x,y)(H)+C2\ (x,y)(A)+e \qquad (4)$$

This revised model as seen in Eqn (2), allows the coefficients to vary as continuous functions over space, so that each may be thought of as a three-dimensional surface over the geographical study area rather than as a single, fixed, real number (Brunsdon et al., 2001).

The assessment of methods is performed with statistical measures of RMSE and r^2.

4. ANALYSIS AND RESULTS

i. Co-Kriging Application

In this method only three secondary variables can be used due to restrictions of employed software. Because of this limitation the secondary variables are grouped into nine different combinations where each of them consists of three variables. SRTM and aspect are used in all combinations since they are considered to be the most identifier variables to predict precipitation and/or temperature.

From the cross-validation table RMSE values were obtained for each variable set (Table 2). For precipitation (h) variable combination (aspect, drop and standard deviation of grid) had the minimum RMSE values (197 mm). For temperature (f) variable combination (SRTM, drop and standard deviation of grid) had the minimum RMSE values (1,532 C^o).

	Prec.	**Temp.**
	RMSE (mm)	RMSE (C^o)
a)SRTM-Asp-Curv	222,2	1,561
b)SRTM-Asp-Drop	215,4	1,555
c)SRTM-Asp-Rough	222,04	1,560
d)SRTM-Asp-River	222,9	1,593
e)SRTM-Asp-Sd-Grid	213,8	1,545
f)SRTM- Drop -Sd-Grid	207,6	1,532
g)SRTM-River-Rough	223,04	1,594
h)Asp –Drop- Sd-Grid	197,06	1,535

Table 2. RMSE values of variable combinations.

The predictions of the combinations giving minimum errors are presented in Figure 2a and 2b. For precipitation, estimated values of "h" variable combination are used to obtain prediction map (Figure 2a). According to the Figure north-west, north-east, south and south- east regions of Turkey have more precipitation than other regions (≈1600 mm).

In order to obtain temperature prediction map "f" variable combination estimated values are used (Figure 2b). According to the figure at the west, south and south-east regions of Turkey average annual temperature is higher than the other regions (highest value is 19 C^o).

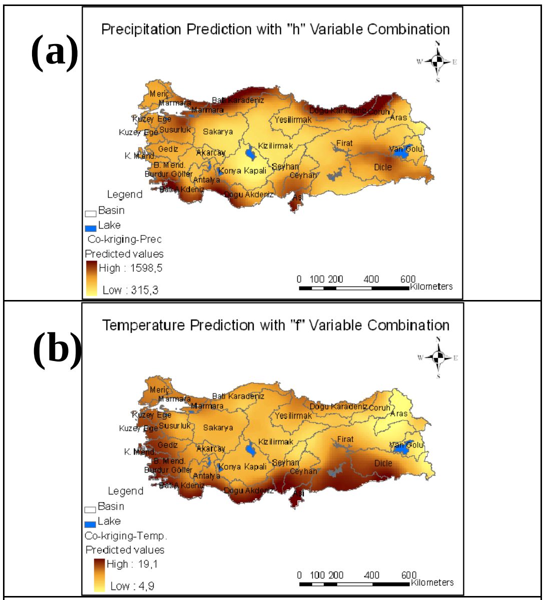


Figure 2. Co-Kriging prediction maps. In (a), precipitation; in (b) temperature prediction maps respectively obtained from "h" and "f"variable combinations are represented.

The RMSE values should not be used alone in order to decide whether an interpolation method yields the best interpolation. Other issues such as the density and location of measurement points (bias) need to be considered (Carrera-Hernandez and Gaskin, 2006). For this purpose comparison between measured and predicted values obtained from co-kriging methods are made. As it is presented in Table (3), "h" and "f" variable combinations resulted with the highest r^2 values for precipitation estimation (0,542 and 0,499). According to r^2 values of temperature "f" variable combination had the highest r^2 value (0,823). It can be said that parameters that show differences of topography have made better approximations to derive the nature of spatial variation of precipitation and temperature.

	Precipitation	**Temperature**
a)	0,424	0,818
b)	0,461	0,819
c)	0,425	0,818
d)	0,42	0,812
e)	0,469	0,821
f)	0,499	0,823
g)	0,419	0,812
h)	0,542	0,820

Table 3. R^2 values between measurements and predictions.

Interpolation residuals of precipitation and temperature are illustrated in Figure 3a and 3b. Generally precipitation prediction error is high at north and north-east regions of Turkey (Figure 3a). Underestimation and also overestimation is very high at these regions (-1067 and 557 mm). According to the temperature interpolation residuals, underestimation takes place at the south and north coasts of Turkey (Figure 3b).

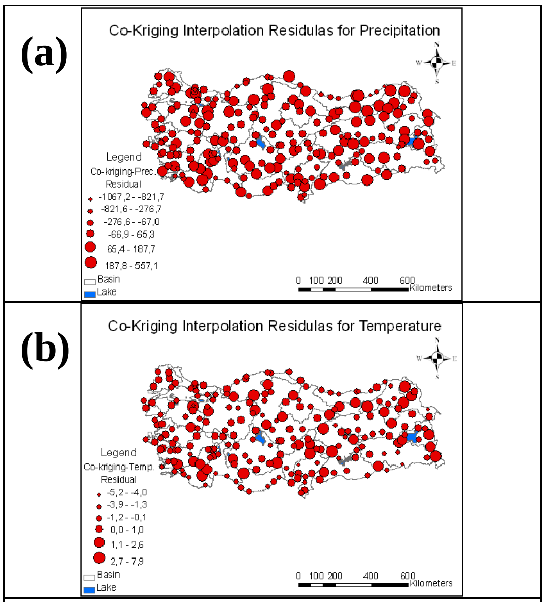


Figure 3. Co-kriging interpolation residuals. In (a), residuals of precipitation prediction; in (b), residuals of temperature prediction respectively obtained with "h" and "f" variable combinations are represented.

ii. OLS Application

OLS method is applied to meteorological variables to make predictions. By using seven secondary variables precipitation prediction is made. According to the t statistic only elevation and Sd_grid variables are statistically significant. Therefore OLS method is applied again by using these two variables. Adjusted r^2 is 0,3.

For temperature data set firstly all independent variables are used in OLS application. According to the t statistic only elevation variable is statistically significant parameter. So in second application only elevation is used as independent variable. Adjusted r^2 is 0,68.

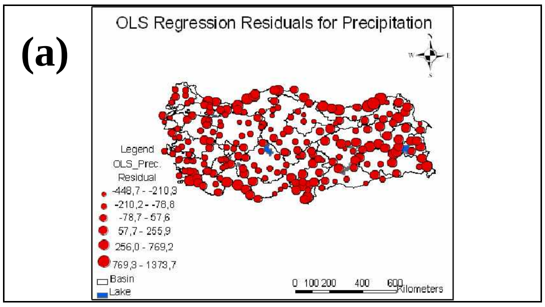
Moran's I technique is applied to precipitation and temperature residuals to control if the residuals are spatially autocorrelated or not. According to the precipitation residuals, index value is 0,23 and and Z score is 5,8. For temperature, index value is 0,73 and Z score is 17,2 (Table 4). For two datasets Moran's I index value indicates a strong clustering. Therefore null hypothesis of randomness is rejected. It can be easily concluded that according to the OLS results, there is statistically significant spatial autocorrelation for both datasets in the study area.

Global Moran's I Summary		
	Precipitation	**Temperature**
Moran's Index	0,238148	0,734636
Expected Index	-0,004464	-0,004464
Variance	0,001748	0,001830
Z Score	5,802919	17,278317

Table 4. Global Moran's I results for precipitation and temperature OLS regression.

In Figure (4a), regression residuals of precipitation prediction are illustrated with graduated symbols. In general, high residuals are located at the north-east Anatolia.

In Figure (4b), regression residuals of temperature prediction are represented. Overestimation is commonly concentrated at west and south regions; however underestimation is concentrated at north regions of Turkey. Figure also supports Moran's I index value with visually.



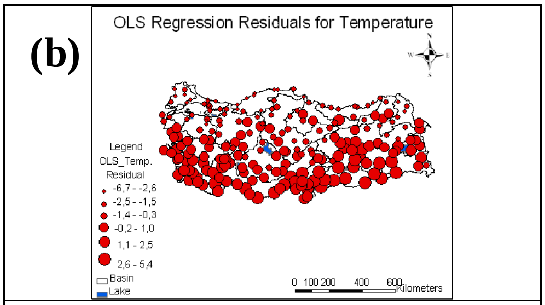


Figure 4. OLS regression residuals. In (a), residuals of precipitation estimation; in (b), residuals of temperature estimation are illustrated.

When there is statistically significant spatial autocorrelation of the regression residuals, the OLS regression is considered to be unreliable. In such circumstances using local regression techniques to model non-stationary variables instead of global techniques can be used to improve predictions. GWR is a frequently used local regression technique.

iii. GWR Application

The output obtained from GWR can be voluminous. Predicted precipitation, temperature values, local r-square values, root-mean-square errors (RMSE) and *t*-values of parameters for each meteorological station are used for discussion. In table (5), RMS errors are shown for two data sets.

Precipitation	Temperature
RMSE	RMSE
141,4 mm	0,6 C^o

Table 5. RMSE values of GWR predictions

Measured and predicted values obtained from GWR are compared for two meteorological variables. Having high r^2 values, 0,67 for precipitation and 0,96 for temperature, indicates truthful predictions are obtained with GWR.

Moran's I index values are calculated for GWR residuals (Table 6). For precipitation 0,004 Moran's I index and 0,15 Z score shows there is no spatially correlation. Residuals are randomly distributed across the study area. For temperature -0,11 Moran's I index and -2,0 Z score express tendency toward dispersion.

Global Moran's I Summary		
	Precipitation	**Temperature**
Moran's Index	0,004167	-0,116688
Expected Index	-0,004464	-0,004464
Variance	0,003077	0,003139
Z Score	0,155598	-2,002920

Table 6. Global Moran's I results for precipitation and temperature GWR applications.

The predicted precipitation and temperature values obtained from GWR are mapped and shown in Figure 5a and 5b.

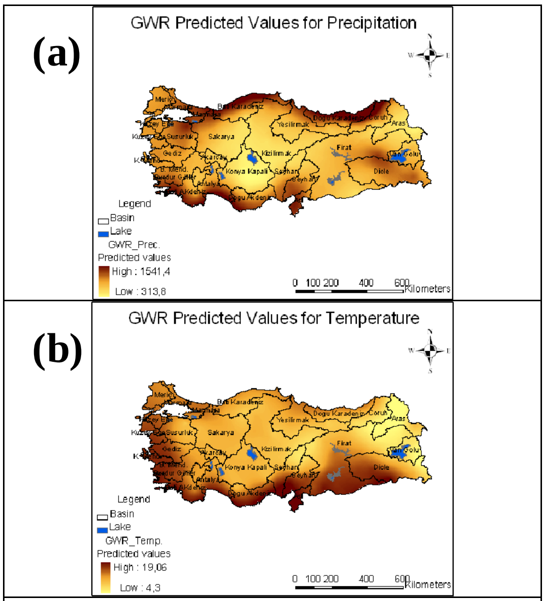


Figure 5. Prediction maps. In (a), precipitation; in (b), temperature prediction maps obtained with GWR application are shown.

North, south, and west coasts and south-eastern of Turkey have more precipitation. Average annual precipitation is comparatively lower than other regions in Central Anatolia (Figure 5a). In respect of Figure (5b), south, south-eastern and west coasts of Turkey have higher mean annual temperature values than other regions.

Local r-square values calculated for each meteorological station are interpolated with kriging operation (Figure 6a and 6b).

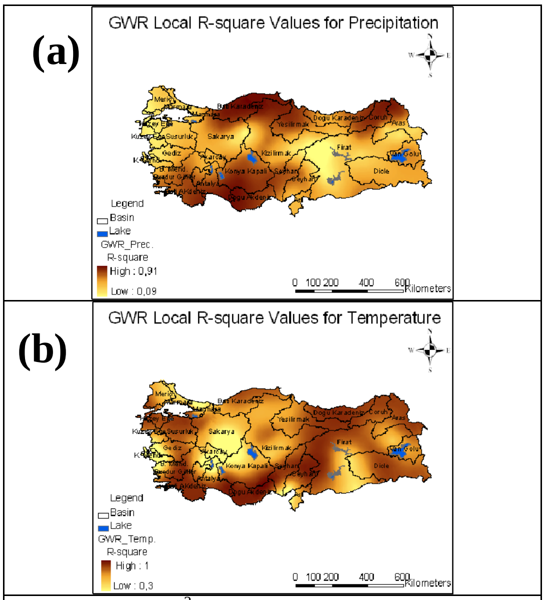

(a)
GWR Local R-square Values for Precipitation
Legend
Basin
Lake
GWR_Prec.
R-square
High : 0,91
Low : 0,09
0 100 200 400 600 Kilometers
(b)
GWR Local R-square Values for Temperature
Legend
Basin
Lake
GWR_Temp.
R-square
High : 1
Low : 0,3
0 100 200 400 600 Kilometers

Figure 6. Local r^2 values. In (a), precipitation; in (b), temperature local r^2 values obtained with GWR application are shown.

For precipitation high r-square values were observed at the north-east, north-west and south regions of Turkey (Figure 6a). This indicates that the model best fits these regions when predicting the precipitation values. Also it can be reported that, the effects of secondary variables on spatial distribution of precipitation are not so significant for south-east and central Anatolia.

In contrast to precipitation r-square map, r-square values of temperature were considerably high for all Turkey except some regions in central Anatolia (Figure 6b). At south parts, r-square values exceed to 0,99. Elevation parameter is very suitable when extracting temperature spatial distribution for whole Turkey according to local r^2 values of meteorological stations'.

Distribution of prediction errors of precipitation and temperature are mapped to highlight inaccurately estimated regions (Figure 7a and 7b). Both error values are lower than Co-kriging and OLS methods. In Figure 7a precipitation prediction errors are illustrated. Generally error is high at north-east regions of Turkey. In Figure 7b, temperature prediction errors are shown. At west, south-east and central Anatolia error is higher than the other regions.

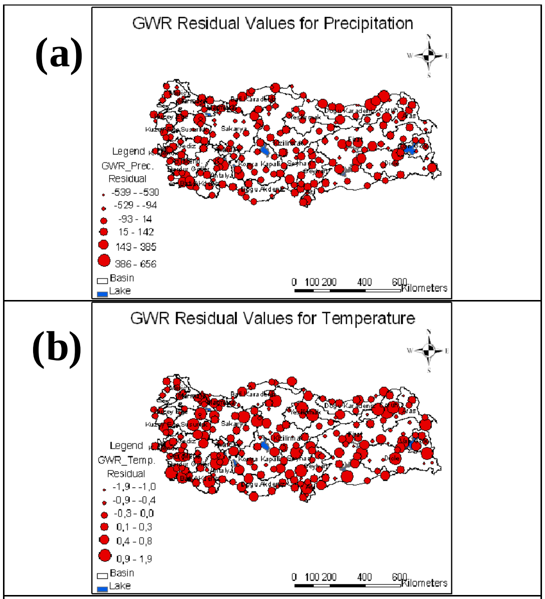


Figure 7. GWR residuals. In (a), precipitation; in (b), temperature residuals obtained with GWR application are shown.

5. CONCLUSION

In this study, mean annual precipitation and temperature values measured at 225 meteorological observations over Turkey are used for revealing spatial distribution of mean annual precipitation and temperature by using secondary variables derived from elevation and river network.

Co-kriging method takes into account the spatial cross-correlation from two or more variables. Different variable combinations are analyzed and cross validation results are evaluated. Temperature prediction mean and RMS errors are

lower than precipitation mean and RMS errors. Also the coefficient of determination, r^2, between measured and predicted values for temperature is very high than precipitation.
OLS provides a global model of the variable to predict by creating a single regression equation to express that process. According to the results r^2 values are lower for two datasets than Co-kriging and GWR (0,3 for precipitation, 0,68 for temperature). Residuals' Moran's I index value show significant positive autocorrelation. This means that either key secondary variable is missing or the global model is not suitable for this dataset.

GWR has provided a means of investigating spatial non-stationary in linear regression models (Brundson et al., 2000). From the outputs of GWR, predicted values, local r-square values and RMSE of each meteorological stations, and *t*-values of parameters are used to evaluate GWR results. These outputs are interpolated with kriging operation and results are analyzed. The lowest RMS errors are obtained by GWR analysis for both data sets (Table 5). Also it is understood that input data set is very appropriate when extracting temperature spatial distribution for whole Turkey due to high local r-square values according to GWR results (Figure 6b). The maps of the local r^2 indicate that, as the relation varies locally, the benefits in using secondary data to provide accurate estimation will vary locally. According to the Moran's I index values there is no spatial autocorrelation for precipitation, for temperature residual pattern is dispersed (Table 6).

According to the three methods, generally GWR gives better predictions for two variable sets in respects of r^2 values and RMS error values between predictions and measurements and Moran's I index values. Also this method can be applied to areas that have no observation with respect of leave-one-out cross validation method.

Spatial distribution of meteorological variables (precipitation and temperature) over Turkey has been defined. In future studies, the temporal variation and distribution of these variables by estimating temporally varying coefficients α_{st}, β_{st} for each meteorological station will be studied in future studies.

4. REFERENCES

Brundson, C., Fotheringham, S., Charlton, M., Geographically Weighted Regression as a Statistical Model, 2000.

Brundson, C., McClatchey, J., Unwin, D.J., Spatial Variations In The Average Rainfall–Altitude Relationship In Great Britain: An Approach Using Geographically Weighted Regression, Int. J. Climatol. 21: p.455–466 (2001).

Carrera-Hernandez, J.J, Gaskin, S.J., Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico, Journal of Hydrology 336: p.231-249 (2007).
Çetin, M., Tülücü, K., Doğu Akdeniz Bölgesinde Aylık Yağışların Yersel Değişimlerinin Jeoistatistik Yöntemle İncelenmesi, Tr. Journal of Engineering and Environmental Science, 22 (1998), p. 279-288.

Heuvelink, G.B.M., Incorporating process knowledge in spatial interpolation of environmental variables. 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (2006).

Jerosch, K., Schlüter, M., Pesch, R., Spatial analysis of marine categorical information using indicator kriging applied to georeferenced video mosaics of the deep-sea Håkon Mosby Mud Volcano, Ecological Informatics, vol 1, issue 4, p. 391-406, 2006.

Lloyd, C.D, 2005. Assessing the Effect of Integrating Elevation Data Into the Estimation of Monthly Precipitation in Great Britain. Journal of Hydrology 308: p.128–150 (2005).

Love, P., Future trends: water resources: meeting future demand. The journal of futures studies, strategic thinking and policy, vol.01, no.03, p.275-278 (1999).

Mas, J. F., Puig, H., Palacio, J. L., Sosa-López, A., Modelling deforestation using GIS and artificial neural networks, Environmental Modelling & Software, vol: 19, issue: 5, p. 461-471, 2004.

Pfeiffer, D.U., Issues related to handling of spatial data, Australian Veterinary Association Second Pan Pacific Veterinary Conference, p.83-105 (1996).

Rigol, J.P, Jarvis, C.H., Stuart, N., Artificial neural networks as a tool for spatial interpolation, International Journal of Geographical Information Science, , vol.15, no. 4, p.323-343 (2001).

Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., Haddix, M.L, Fine-resolution mapping of soil organic carbon based on multivariate secondary data, Geoderma, v132, 3-4, p.471-489 (2005).

Tobler, W.R., A computer movie simulating urban growth in the Detroit region, Economic Geography 46: p.234-240 (1970).

Weir-Smith, G., Schwabe, C.A., Spatial interpolation versus neural network propagation as a method of extrapolating from field surveys, GIS Centre, HSRC (Human Sciences Research Council), Pretoria, 2000.

Zhang, L., Gove, J. H., Heath L. S., Spatial residual analysis of six modeling techniques, Ecological Modelling 186 p. 154–177, 2005.

ESRI website:
http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=How%20OLS%20regression%20works (accessed 25 August 2009).

AUTOMATICALLY AND ACCURATELY MATCHING OBJECTS IN GEOSPATIAL DATASETS

L. Li [a, *], M. F. Goodchild [a]

[a] Dept. of Geography, University of California, Santa Barbara, CA, 93106 US - (linna, good)@geog.ucsb.edu

KEY WORDS: Object Matching, Linear Programming, Assignment Problem, Optimization, Greedy

ABSTRACT:

Identification of the same object represented in diverse geospatial datasets is a fundamental problem in spatial data handling and a variety of its applications. This need is becoming increasingly important as extraordinary amounts of geospatial data are collected and shared every day. Numerous difficulties exist in gathering information about objects of interest from diverse datasets, including different reference systems, distinct generalizations, and different levels of detail. Many research efforts have been made to select proper measures for matching objects according to the characteristics of involved datasets, though there appear to have been few if any previous attempts to improve the matching strategy given a certain criterion. This paper presents a new strategy to automatically and simultaneously match geographical objects in diverse datasets using linear programming, rather than identifying corresponding objects one after another. Based on a modified assignment problem model, we formulate an objective function that can be solved by an optimization model that takes into account all potentially matched pairs simultaneously by minimizing the total distance of all pairs in a similarity space. This strategy and widely used sequential approaches using the same matching criteria are applied to a series of hypothetical point datasets and real street network datasets. As a result, our strategy consistently improves global matching accuracy in all experiments.

1. INTRODUCTION

1.1 Motivation

High-quality data are always the prerequisite for meaningful analyses. Since no single geographical dataset is a complete and accurate representation of the real world, we usually require data from diverse sources in scientific research and problem solving. In a particular geographical application, we need to obtain data from multiple sources that represent different properties of objects of interest. Unlike old days when research was impeded due to lack of data, rapid development of technologies for data collection and dissemination creates abundant opportunities for manipulating and analyzing geographical information. However, it is not always straightforward to take advantage of large volumes of geospatial data because data created by different agencies are usually based on different generalization schemes, using different scales, and for different purposes.

As it is impossible to directly collect all data by ourselves, we often need to utilize secondary data sources. Thus it is usually inevitable to combine multi-source data in science, decision-making, and everyday life. For example, in an emergency such as Jesusita Fire in Santa Barbara, effective evacuation requires integrative geospatial information about the affected area, probably including DEM, land use, residence and facility locations. Another typical application is the creation of an integrated database from two input datasets. It is possible that one dataset has all necessary features and attributes, but the other one bears a higher accuracy of positions. For instance, we have an old street network stored as vector data, and a recent remote sensing image that covers the same area. After extracting streets in the image, we want to identify the same streets in the outdated vector database in order to improve its positional accuracy. In all these cases, accurate identification of objects that represent the same entity in reality is an essential prerequisite to further analyses.

1.2 Objective

Object matching can be divided into two steps: the first step is to define a proper similarity measurement between objects, and the second step is to search for matched pairs based on this measurement. This paper focuses on the second step of this procedure by providing a new strategy for matching objects in multiple sources given a certain criterion. Rather than adopting a sequential matching procedure that is widely used in existing literature, we propose a matching algorithm according to an optimization model by regarding object matching as an assignment problem. In the remainder of this paper, Section 2 discusses two types of methods for object matching: widely used greedy method and proposed optimization method. In section 3, we describe two sets of data used in our experiments for a comparison between two methods. In section 4, we present the percentage of correctly matched pairs using different methods, followed by some conclusions in section 5.

1.3 Related Work

Object matching in geospatial datasets has been a fundamental research problem for decades. Most efforts have focused on the definition of similarity between objects. If two objects in different datasets are similar in terms of positions, shapes, structures, and topologies and so on, it is probable that they represent the same entity in the real world. The similarity metric varies from one application to another due to the inherent characteristics of input data and the availability of data properties.

The most popular similarity measurement is the proximity between objects. One typical criterion is the absolute proximity

[*] Corresponding author.

measured as a distance, such as the Euclidean distance between points or Hausdorff distance between polylines (Yuan and Tao, 1999), and some other distances in particular applications, like the discrete Frechet distance (Devogele, 2002) and radial distance (Bel Hadj Ali, 1997). The Hausdorff distance has been proved proper in calculating the proximity between linear features (Abbas, 1994). It is defined as the maximum distance of the shortest distances between each point on one linear object and a set of points constituting another polyline. When the distance between two objects is smaller than a threshold, they may be regarded as a corresponding pair. In addition to a distance threshold, another measurement based on a relative proximity is usually called the nearest neighbour pairing. This criterion intends to find the nearest neighbour of a particular object in the other dataset regardless of its absolute distance. If an object A in the first dataset is the closest object for object A’ in the second dataset, and meanwhile object A’ is the closest one for object A in the second dataset, objects A and A’ are defined as a matched pair (Saalfeld, 1988; Beeri *et al*., 2004).

Besides proximity, other geometric information is also used in object matching. For example, matching between street segments may be reduced to node matching since nodes, especially intersections, are usually taken as control points (*e.g.* Cobb *et al*., 1998; Filin and Doytsher, 2000). The number and directions of connecting segments for a node are usually used to refine the matched candidates as a result of proximity criterion (Saalfeld, 1988). The angles between two street centrelines or between GPS tracks and street networks are also widely used in polyline and map matching (Walter and Fritsch, 1999; Quddus *et al*., 2003).

Another category of information for object matching is semantic similarity, including two important considerations: similarity between geographic types and similarity between individual geographic objects. In any dataset that involves geographical classes, it is critical to establish a mapping between different classification systems because any classification entails loss of information and usually subjective judgment. On the other hand, similarity between geographic objects may be defined according to attribute values, either numeric or string-similarity (Cohen *et al.*, 2003). Hastings (2008) used both types of semantic similarity - geotaxonomic and geonomial metrics - in conflation of digital gazetteers.

Furthermore, contextual information is also helpful for refining matching results based on the relationship between investigated objects and its surrounding environment. For instance, Filin and Doytsher (2000) developed an approach called “round-trip walk” to take into account contextual information. The counterpart nodes at two ends of the arc are called connected nodes. Two nodes are identified as matched only under the condition that they are similar enough and at the same time their connected nodes are also similar enough. When no explicit contextual information is available, Samal *et al.* (2004) proposed proximity graphs as an aid to incorporate context when landmarks are not connected with other features by constructing topology among them.

While these different methods all focus on the definition of similarity measurement in various datasets, few efforts, if there’s any, have been made to improve the search process and consequent matching results given a selected similarity criterion. Rather than comparing different similarity metrics, we propose a new search strategy that minimizes the global mismatch errors after a certain similarity measure is selected.

2. METHODS

Automatic object matching requires an objective function or a series of functions, the solutions to which lead to matched pairs. This function provides a rule to determine whether two objects should be matched and a search path to find all matched pairs. The variables in this function could be any similarity metrics, such as Euclidean distance or Hausdorff distance, or a combination of a set of measurements. In this section, we will discuss two search strategies in object matching after a similarity metric is selected: the first one is the popular greedy method that aims to always find the possible minimum dissimilarity between paired objects in each step, and the second one is our proposed optimization strategy that intends to minimize the total dissimilarity between all matched objects.

2.1 Matching Objects Using Greedy

Greedy is a simple way to achieve local optimum at each stage. Its essence is to make the optimal choice at each step even in a problem that requires multiple steps to solve. It has been studied in many fields such as operations research and computer science (Wu *et al*., 1990) and widely implemented in many applications. One obvious problem with the greedy algorithm is that an addition of a new item to the solution set may render the solution not optimal and it does not provide a mechanism to remove items already in the solution. For example, if we match two objects incorrectly in a previous step, there is no way to correct that mistake in later stages. Therefore, in a greedy-based algorithm, a mismatch error in any step will result in at least two mistakes because it will make it impossible for the omitted object to be matched to the correct one in a later stage.

Two greedy methods were implemented in MATLAB in our study. Greedy1 adopts a sequential identification and removal procedure: it identifies the closest pair of objects as corresponding counterparts and removes both from the candidate set; then it identifies the closest pair in the remaining objects and removes them, until all objects are matched. Greedy2 is a modified version of greedy1 by adding a random component to the procedure in order to jump out of local optima. It starts with a random object in one dataset and identifies the closest object in the other, followed by the elimination of matched pairs; then it selects another random object and identifies its matched correspondence until the process is finished. This procedure could be repeated as many times as necessary (*e.g.*, 100) and the best result would be the final result.

2.2 Matching Objects Using Optimization

In order to rectify mistakes introduced in previous stages in a greedy algorithm, we propose another strategy to rely on a global measurement of similarity by regarding object matching as an assignment problem that takes into account all corresponding pairs of objects simultaneously. The search for corresponding objects is based on minimization of dissimilarity between matched objects and can be formulated as the following objective function:

$$Minimize \ \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_{ij} \tag{1}$$

where i = index for the objects in the first dataset
j = index for the objects in the second dataset
n = the number of objects in each dataset
c_{ij} = the dissimilarity between object i in one dataset and object j in the other. c_{ij} could be any form of similarity measures or any combination of multiple metrics that jointly decide the similarity between two objects
x_{ij} = a Boolean indicator: when object i in the first dataset and object j in the second dataset are matched, it is assigned to 1, and assigned to 0 otherwise

The constraints for this objective function are as follows:

$$\sum_{j=1}^{n} x_{ij} = 1, \quad \forall i \tag{2}$$

$$\sum_{i=1}^{n} x_{ij} = 1, \quad \forall j \tag{3}$$

These two constraints ensure that every object in each dataset is matched to exactly one object in the other dataset.

This form of objective function is well known as the assignment problem in the operations research. It is generalized from the problem of assigning a set of tasks to a group of agents with the objective to minimize the total cost of performing all tasks, under the constraints that each task can only be assigned to one agent, and each agent can only accept one task (Hillier and Lieberman, 2001). Our task in object matching is to assign each object in one dataset to its corresponding counterpart in the other one, satisfying the objective function that minimizes the total dissimilarity between matched pairs.

In real applications, two datasets that represent the same area rarely have the same number of objects, so we relaxed the constraints:

$$\sum_{i}^{m} x_{i,j} <= 1, \forall j \tag{4}$$

$$\sum_{j}^{n} x_{i,j} = 1, \forall i \tag{5}$$

where m = the number of objects in dataset 1
n = the number of objects in dataset 2
$m<=n$

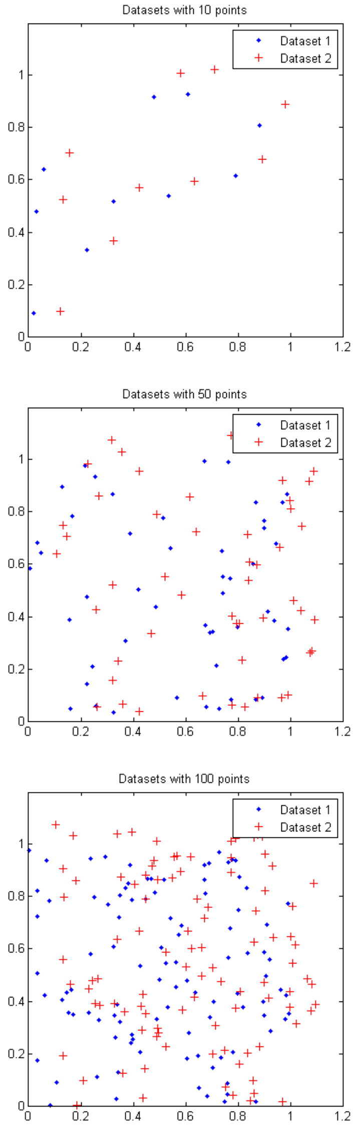
Therefore, each object in the smaller dataset is matched to one object in the other, and some objects in the larger dataset will be identified as having no corresponding pair. This assignment problem was implemented using the GNU MathProg modeling language in the GLPK (GNU Linear Programming Kit) package that provides a platform for solving linear programming problems. The similarity criterion is Euclidean distance in the point datasets, and Hausdorff distance in the polyline datasets.

3. DATA

Two sets of data were used to test the differences between greedy and optimization methods in object matching: hypothetical point datasets and real street network datasets.

3.1 Hypothetical Data

Hypothetical data were generated by a random process. The first set of point data were created by a bivariate point process and the second set of point data were created by the following formula: $x_2 = 0.1 + x_1$, $y_2 = 1.1 * y_1$, where x_1, y_1 are the coordinates of points in the first set of datasets, and x_2, y_2 are the coordinates of points in the second set of datasets. Within a square area, the number of points varies from 10 to 100 with an interval of 5. Some examples of these datasets are demonstrated in Figure 1.


Datasets with 10 points

Datasets with 50 points

Datasets with 100 points

Figure 1. Hypothetical datasets with different numbers of point objects

3.2 Real street data

Real street data are more complex than the hypothetical point data, since they are composed of multiple points and the offsets between objects are not uniform. In our experiment, street network data in Goleta CA were created under different standards by two agencies. These data represent approximately the same streets in a neighbourhood of Goleta (Figure 2). These two datasets have 236 and 223 objects, respectively. As shown in the figure, there are some discrepancies between these two datasets, and some streets are missing in one version of the data. These data were prepared in a way that they are under the same coordinate system and internally consistent.

Pre-processing was performed in the datasets to maximize 1:1 correspondences, since our optimized object matching strategy is designed for 1:1 matching. Due to the difference in generalization of real streets, the same street may be represented as different numbers of segments. For example, the street Hollister could be described as 5 segments (objects) in one dataset, and as 7 segments (objects) in the other. Therefore, it is helpful to make as many pairs of 1:1 correspondences as possible. In our experiment, we merged street segments based on the name attribute and the topology of polylines. In each dataset, if multiple street segments have the same name and they are connected, they are merged to form one object after pre-processing.

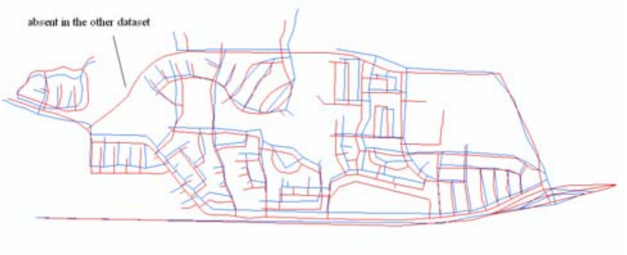


Figure 2. Street networks in a neighborhood of Goleta, CA.

4. RESULTS AND DISCUSSION

Both greedy and optimization methods were tested in these datasets. The sum of distances between matched objects using each of the three methods is displayed in Figure 3. When the number of points is small, the total distances calculated from different methods are similar. As the density of points becomes larger, the difference of total distance becomes more obvious between greedy and optimization methods, but the results are relatively close between the two greedy methods. In any dataset, the total distance of matched pairs is consistently smaller using the optimization method. In Figure 4, the relationship between the percentage of correctly matched pairs and the number of points is displayed. The trend shows that there is a drastic drop in the percentage of correct matches using the two greedy methods as the number of points becomes larger. However, the percentage of correct matches using the optimization method is stable and robust in all tested datasets. While the percentage of correct matches decreases from 100% or 80% to less than 20% using the greedy methods, the percentage of the optimization method maintains at a level close to 100% even in dense datasets. Therefore, when the density of points gets larger, the probability of mismatch becomes larger, and consequently the superiority of the optimization method becomes more obvious.

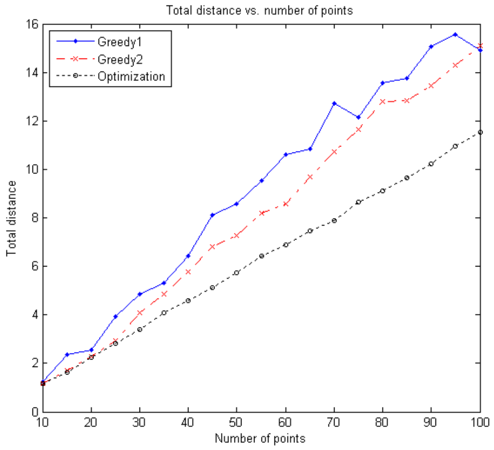


Figure 3. Total distance of matched pairs.

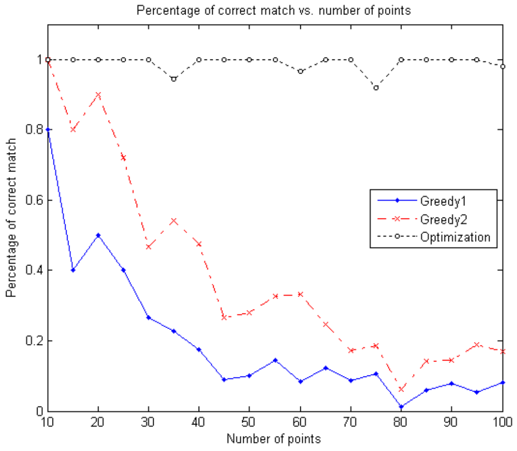


Figure 4. Percentage of correct matches.

The results of object matching in real street data using the three methods are demonstrated in Table 1. The total distance between matched pairs is smaller using the optimization method than using greedy methods. As a result, the percentage of correct matches using the optimization method is about 10% higher than that using the optimization method.

Table 1. Results of object matching for street datasets

	Total distance	Percentage of correct match
Greedy1	13104	88.14%
Greedy2	13078	88.56%
Optimization	12369	97.03%

In all experiments, either with hypothetical point data or real polyline data, object matching using the optimization method consistently achieves better results. When the density of a dataset increases, the probability of mismatch becomes larger, and consequently, the advantage of the optimization method becomes more obvious. While a denser dataset makes object matching more susceptible to mismatches, the spatial arrangement of objects within the study area is also another important factor that affects the matching result. These experiments indicate that the optimization method for object matching is more robust than greedy methods. In some datasets, object matching using a greedy method may also result in a good percentage of correct matches, but in other cases, the percentage could be not acceptable. Since it requires a lot of time and labour to identify and correct even a small number of mismatches, it is important to maximize the percentage of correct matches in the automatic stage of object matching.

In terms of the choice of similarity measurement in our experiments, when the total distance of matched pairs is small, the percentage of correctly matched objects is high. Therefore, Euclidean distance and Hausdorff distance are proper indicators of point and linear object similarity in these datasets, respectively. However, when more attributes are available, not only relying on the geometric distance in a geographical space, we can also construct a similarity space according to a weighted combination of these properties, and use that metric as a similarity measurement in our objective functions. Furthermore, additional attributes may also be used to reduce search space in particular applications. Although the emphasis of this paper is not the selection of similarity measurement, a proper similarity metric is a necessity for effective and efficient object matching. A measurement that is an adequate indicator of the likeness between two objects should be included in the objective function.

5. CONCLUSIONS

Object matching is a fundamental problem in spatial data handling and many related applications. How to identify objects in different data sources that represent the same entity in reality is a prerequisite for data manipulation and analyses in later stages, such as accuracy improvement, change detection, and geospatial analysis using multi-source data. There are two major components in the object matching process: selection of an appropriate similarity measurement and identification of matched objects according to this measurement. Most existing literature has focused on the definition of a proper similarity metric in particular applications. They usually adopt a sequential procedure to find object pairs one after another based on the chosen metric. In our paper, we focus on the other aspect of the problem: how to effectively search for corresponding objects once a similarity measurement is chosen. Rather than using a greedy strategy that consecutively adds more matched pairs into the solution set, and never removes any mismatched pairs from the solution, our optimized object matching takes into account all possible matched objects simultaneously with the aim to minimize the total dissimilarity between all corresponding objects.

Therefore, object matching is formulated as an assignment problem that intends to assign each object in one dataset to an object in the other dataset, with the objective to minimize the sum of dissimilarity between object pairs. Unlike the widely used greedy procedure for finding matched pairs, this strategy makes it possible to rectify mismatch errors made in early steps. Although only point and polyline data were tested in this paper, this method can also be applied to other types of data as long as the selected metric is adequately representative of the resemblance between objects. Our experiments demonstrate that optimized object matching method is robust and always achieves a higher percentage of correctly matched pairs in both hypothetical and real datasets.

Although our research points out a new research direction in object matching, there are some limitations. First, formulation of object matching as an assignment problem entails the constraints that one object can only be assigned to one or none object in the other dataset. Therefore, this strategy is appropriate for 1:1 correspondence. In real applications, there are cases when an object in one dataset is represented as several parts in the other dataset (1:*n* correspondence), or several objects are corresponding to a different number of objects (*m*:*n* correspondence). Therefore, one of our future research questions is to find a way to maximize the 1:1 correspondence in different datasets before the execution of the optimized object matching strategy. Another problem we are going to investigate is to directly tackle the 1:*n* and *m*:*n* relationships by examining partial similarity between objects. Finally, as the input datasets become larger, the matching procedure may degrade rapidly, and makes it difficult to finish matching within a reasonable time frame. Therefore, we will study the improvement of the algorithm using heuristics to reduce the search space, such as divide-and-conquer technique (Preparata and Shamos, 1985).

References

Abbas, I., 1994. Base de données vectorielles et erreur cartographique: problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorf. *Computer Science*. Paris, Université de Paris VII.

Bel Hadj Ali, A., 1997. Appariement geometrique des objets géographiques et étude des indicateurs de qualité. Saint-Mandé (Paris), Laboratoire COGIT.

Cobb, M. A., Chung, M. J., Foley III, H., Petry, F.E. and Shaw, K.B., 1998. A rule-based approach for the conflation of attributed vector data. *Geoinformatica*, **2**(1), pp. 7-35.

Cohen, W., Ravikumar, P. and Fienberg, S. E., 2003. A comparison of string distance metrics for name-matching tasks. *IJCAI-2003*.

Devogele, T., 2002. A new merging process for data integration based on the discrete Frechet distance. In: *Advances in Spatial Data Handling*. D. Richardson and P. van Oosterom. New York, Springer Verlag: pp. 167-181.

Filin, S. and Doytsher, Y., 2000. The detection of corresponding objects in a linear-based map conflation. *Surveying and Land Information Systems*, **60**(2), pp. 117-128.

Hastrings, J. T., 2008. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, **22**(10), pp. 1109-1127.

Hillier, F. S. and Lieberman, G. J., 2004. *Introduction to Operations Research* (McGraw-Hill).

Preparata, F. P. and Shamos, M. I., 1985. *Computational Geometry: An Introduction* (New York, NY: Springer-Verlag New York, Inc.).

Quddus, M., Ochieng, W., Zhao, L. and Noland, R., 2003. A general map matching algorithm for transport telematics applications. *GPS Solutions*, **7**(3), pp. 157-167.

Saalfeld, A., 1988. Conflation automated map compilation. *International Journal of Geographical Information Systems*, **2**(3), pp. 217-228.

Samal, A., Seth, S. and Cueto, K., 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, **18**(5), pp. 459-489.

Walter, V. and Fritsch, D, 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, **13,** pp. 445-473.

Wu, S., Manber, U., Myers, G. and Miller, W., 1990. An *O*(*NP*) Sequence comparison algorithm. *Information Processing Letters*, **35**, pp. 317-323.

Yuan, S. and Tao, C., 1999. Development of conflation components. *The Proceedings of Geoinformatics'99 Conference* (Ann Arbor).

The Estimation of Mesoscale Ocean Eddies Change Based on CBR

Yunyan Du [a], Chenghu Zhou [a] Lijing Wang [a, b], Guangya Qi [a], Xinzhong Yang [a, b]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101

[b] Geomatics College, Shandong University of Science and Technology, Qingdao 266510

KEY WORDS: Artificial Intelligence, Case-based Reasoning (CBR), Mesoscale Ocean Eddies, Spatial Relationships, Rough set

ABSTRACT:

Case-based reasoning (CBR) method has been widely used to study geographical problems during the past two decades. However it is still not perfect particularly when employed to solve complicated geographical problems. Urgently needed improvement includes development of geographic data representation modeling and design of algorithm for spatial similarity computation and reasoning. This paper reports an improved CBR-based method for studying the spatially and temporally complex Mesoscale Ocean Eddies (MOEs). After summarizes the basic advantages and challenges of current existing quantitative methods, the paper first proposes that CBR approach, with support of GIS, can be employed to study variation of MOEs. Representation model was constructed to describe the case, i.e., MOEs. This paper then provides an algorithm to retrieve the inherent spatial relationships among cases, as well as a CBR similarity reasoning algorithm to predict change of MOEs. The method was finally tested by examining the MOE in the South China Sea and yields an average estimation accuracy of 80%. In summary, the CBR-based approach proposed in this study provides an effective and explicit solution to quantitatively analyze and predict the change of MOEs.

1. INTRODUCTION

A variety of methods have been widely used to study change of MOE. These approaches can be categorized into either static or dynamic analysis. The first method indirectly studies ocean eddies by examining variations in marine water masses. The latter one studies characteristics of ocean eddies by analyzing the flow direction and velocity of ocean currents, either by using conventional analysis methods to study data derived from field survey or remote sensing, or by numerically simulating the ocean current and eddy(Du et al., 2004). Specifically, methods used to study MOE include numerical simulation, mathematical statistics, remote sensing information extraction (Hough change, multi-fractal), and the P vector (P-Vector) method. Numerical simulation method mainly uses the Princeton model (POM) to examine the characteristics of ocean eddy（Yang et al.,2000; Li et al.,2003）. Statistical methods study the spatial and temporal distribution pattern of ocean eddy by quantitatively examining long sequence of marine data derived from remote sensing techniques (TOPEX / Poseidon, altimeter, MODIS, NOAA and SeaWiFS)(Qian e tal.,2000; Lin, 2005). Remote sensing methods mainly focus on extracting quantitative information of MOE from different platforms (Wang et al., 2001, 2004; He et al., 2001; Ge et al., 2007).

Although these studies achieved valuable results in regarding to the developing mechanism and distribution pattern of MOE, as well as their own self parameters. However, some limitations still exist in these methods. The numerical simulation approach can provide continuous time series data to describe change of MOE. Unfortunately, time consuming is a big issue for this method as the parameters must be retuned and simulation must be rerun due to change of the boundary conditions when different regions are studied. Statistical methods, to somewhat extent, can present the spatial and temporal distribution patterns and movement trend of MOE. However, there is no way for these methods to quantitatively predict the occurrence of ocean eddy. While remote sensing is able to quantitatively extract the local spatial information and parameters of MOE, it is unable to analyze and predict its evolution trend. Therefore, a new method is in great of need. If the new method can absorb the advantages of those above-mentioned methods, it will serve better to study the MOE.

CBR is a method used to solve current geographical problems based on reasoning from historical similar cases. This method is capable to quantitatively analyze and predict geographic phenomena by examining enough existing historical data, even without any knowledge about their developing mechanism. CBR is a comprehensive problem-oriented analysis approach. CBR has already been widely used to study geographical problems since 1990s and yielded quite a few promising results (Jones et al., 1994; Du et al., 2002; Li, 2004). However, majority of these studies either focus on the direct application of traditional CBR approaches to solve geographical problems, or only consider the spatial attributes of geographical problems. Current CBR methods are not perfect enough to solve complicated geographical problems, particularly those with significant zonal and territorial differentiation. Therefore, research is necessary to develop improved case representation modelling and enhanced algorithms for similarity computation and reasoning]. This paper, from a new methodological perspective, uses CBR methods to quantitatively analyze and estimate the MOE change.

2. CBR FOR ESTIMATION OF MOE CHANGE

2.1 Case Representation Model

To better describe spatial distribution characteristics and relationships of the geographical problems, this paper proposed a three-component representation model by adding a new component, “geographical environment” into the traditional dual-mode model. As a result, case in this paper consists of three components, including “problem”, “geographical environment”, and “outcome”. By adding the “geographical environment” component, case representation model is able to consider the influence of marine environmental variables on the development of eddy. Spatial information of ocean eddy is also integrated into the "outcome" component.

2.1.1 Conceptual definition of three-mode representation model: The "problem" component of an ocean eddy case refers to "the situation of an eddy after a certain time interval". The "geographical environment" refers to those ocean physical environments that influence the development and variation of MOE, usually including eddy's spatial position, seabed terrain characteristics, water features (such as the ocean temperature, salinity, ocean current, density), and some other spatial-temporal information. The geographical environment can be described by 1- or n-dimensional GIS spatial feature layers or simply by some spatial indicators. The "outcome" component is defined as the situation of a MOE after a certain time interval, for instance, the variations in its travel speed, direction, and intensity.

2.1.2 Case representation and organization: In this research, the "problem" is described by some quantitative attributes of an ocean eddy, including, but not limited to, its travel direction, speed and intensity. The "geographic environment" component is represented by a series of quantitative indicators of MOE attributes and its spatial relationships (e.g., the relationships of direction, topology, and distance) to the adjacent ones. The last component, the "outcome", is the attributes of the successive eddy after a certain time interval. As a result, case can be described using equation (1):

$$Case_i = \left\{ \begin{array}{l} S_i, SA_{1i}, SA_{2i},..., SA_{ji}, SR_{1i}, SR_{2i},..., \\ SR_{li}, Vortex_{t1i} \rightarrow Vortex_{t2j} \end{array} \right\}$$

$$i=1,2,..K; j=1,2,..M; l=1,2,..N; \quad (1)$$

$$S_i = \{(x_i^1, y_i^1), (x_i^2, y_i^2), ..., (x_i^m, y_i^m)\}$$

Where i is the case number; Si is a set of spatial shape attributes of case i, i.e., the coordinates collection of polygonal boundary of an eddy; SA_{1i}， SA_{2i}， …,SA_{ji} are the attributes (totally M) of case i; SR_{1i}， SR2i， …,SR_{li} refer to the spatial relationships (totally N) between case i and the geographical environment factors; Vortex t1i → Vortext2i is the case "outcome", i.e., the situation of an ocean eddy after a certain time interval.

2.2 Extracting Spatial Characteristics of MOE

Rough set is an approach used to study the data representation, learning, and induction from incomplete knowledge and data with certain uncertainty. No prior information other than data set itself is required. This method is able to extract the decision-making or classification rules by knowledge simplifying while maintains enough classification accuracy (Ding, 2004). This paper uses rough set theory to extract the decision-making rules from historical cases of MOE. Specific algorithm used in this study includes three basic steps:

(1) Description of the prior spatial relationships among eddies based on rough set theory

As shown in Figure 1, several specific spatial relationships impacted MOE change, such as topological relationships, the distance to the Kuroshio axis, and the distance to shoreline and so on, were selected based on previous research results or experiences. These spatial relationships were then converted into quantitative indicators by using GIS spatial analysis methods. Spatial decision-making table is then constructed, with row representing historical cases and column showing attributes. The first part of column records conditional attributes, including indicators of spatial relationships while the other part of column documents decision-making attributes, i.e., the "result" of case.

(2) Discretizing continuous variables in the spatial decision-making table using different methods based on different conditional attributes.

(3) Simplifying spatial relationships in the decision-making table using attribute simplifying algorithm. Decision-making spatial relationships which determine the case outcome are extracted and decision-making rules are finally retrieved.

2.3 Case Similarity Calculation and Reasoning

Nearest neighbourhood method was usually used in traditional CBR approach to compute the similarity among cases based on the assumption that two cases have similar but completely independent attributes. As spatial relationships among cases, as well as between cases and environment, were all considered in the representation model proposed in this study, nearest neighbourhood method cannot be used to calculate the similarity among cases. In this study, general similarity among ocean eddy cases was calculated by equation (2).

$$Similarity_{Case(i,j)} = w_1 \times S_{r(Case(i,j))} + w_2 \times S_{a(Case(i,j))} + w_3 S_{s(Case(i,j))} \quad (2)$$

Where w_1 , w_2 , and w_3 are weights assigned to different similarity coefficients and w_1 + w_2 + w_3 =1. $S_{a(Case(i,j))}$, $S_{r(Case(i,j))}$, and $S_{s(Case(i,j))}$ are the similarity coefficients between case *i* and *j*'s attributes, spatial relationships, and shapes respectively. In equation (2), $S_{a(Case(i,j))}$ was calculated same as the traditional CBR using Euclidean distance algorithm.

In equation (2) $S_{r(Case(i,j))}$ was calculated using equation (3):

$$S_{r(Case(i,j))} = w_{dir} \times S_{dir(Case(i,j))} + w_{top} \times S_{top(Case(i,j))} + w_{dis} S_{dis(Case(i,j))} \quad (3)$$

Where w_{dir} 、 w_{top} and w_{dis} are weights assigned to different similarity coefficients and w_{dir} + w_{top} + w_{dis} =1. $S_{dir(Case(i,j))}$, $S_{top(Case(i,j))}$, and $S_{dis(Case(i,j))}$ are the similarity coefficients between case *i* and *j* in relationships of spatial direction, topology, and distance respectively. They can be calculated by the traditional GIS spatial relationship algorithm (Goyal, 2000).

Usually different similarity calculation methods are used to compute $S_{s(Case(i,j))}$ in equation (2) based on the geometric shape. For example, if the geographical cases are linear features, a "similarity calculation algorithm of Radius Vector Serial Analysis Model Based on Barycentre (RVSAMB)" can be used to calculate the $S_{s(Case(i,j))}$,(Du et al., 2002), while "an approach to similarity measures for polygonal shapes based on mechanics" is used for polygonal cases(Fan et al., 2003).

Case reasoning was then performed once similarity coefficients were calculated. Historical cases with similarity coefficients greater than an arbitrarily-set threshold were first selected. "Outcomes" of these cases were then screened and different weights were assigned to the "outcomes" based on different values of similarity. Weighted average was then calculated and accepted as the "outcome" of current case.

3. CASE STUDY

3.1 Study Area

Method proposed in this paper was tested by studying the MOE developed in the SCS (0°-23°N, 99°-12l°E) from November 2003 to February 2009. The study region has an area of 3.5 million square kilometres. The SCS is a semi-enclosed basin with complex seabed topography, usually showing unique mesoscale variations in marine environmental conditions due to the influence of East Asia Monsoon and the Kuroshio. Many researches have provided valuable historical experiments and solid basis for employing CBR approach to study MOE in the SCS (Lin et al.,2007).

Raw data used in this study include stratified numerically simulated global sea surface height abnormity (SSHA), sea surface temperature (SST), and marine current. These data have a spatial resolution of 1/32x1/32 degree and provided by Navy Research Laboratory (NRL). The data are substantiated by multiple satellite images. For instance, SSH is substantiated by ENVISAT, GFO and JASON-1 while SST by IR satellite-derived data. Cases (MOE) studied in this paper are identified by expert based on three data groups. As shown in Figure 1, Typical ocean eddy is identified as those with a diameter no less than 100 km, height difference between eddy centre and the outmost closed contour no less than 8 cm, life span no less than 20 days, visible annular flow on the current map of MOE, and a current speed more than 0.5m/s in the eddy centre.

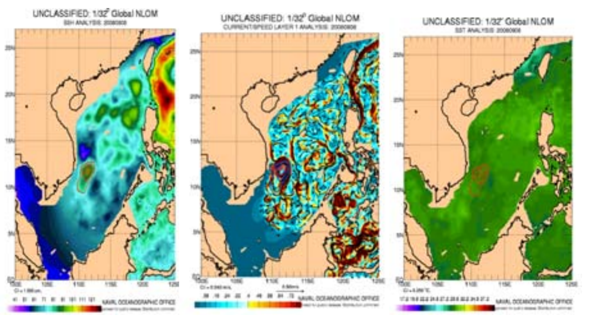

UNCLASSIFIED: 1/32° Global NLOM
UNCLASSIFIED: 1/32° Global NLOM
UNCLASSIFIED: 1/32° Global NLOM
NAVAL OCEANOGRAPHIC OFFICE

Figure 1 Example of a MOE in the SCS case. Data used to identify the eddy are also shown in this figure.

3.2 Estimation of Eddy Variation in the SCS

3.2.1 Representation and organization of cases: For this study, cases are represented by three components: "problem", "geographical environment", and "outcome". These three components are then quantitatively described. The "problem" is described by the shape and related spatial-temporal attributes of a specific eddy at a certain time, including the vortex number (*ID*), perimeter(*P*)，area(*A*), type(A_1), intensity(A_2), condition of the vortex at this time(A_3), horizontal scale(A_4), major axis length(A_5), minor axis length(A_6), time(A_8), and duration(A_9). Current research results indicate that development of ocean eddy in the SCS is significantly affected by the physical marine environment and the occurrence of other mesoscale phenomena in the same region. As a result, this paper uses four ocean environmental indicators and two spatial direction indicators to quantitatively describe the "geographic environment" component, including sea surface temperature in the vortex centre (F_1), temperature difference (F_2) between eddy centre and periphery, geographical longitude (*Lo*) and latitude (L_a) of the eddy centre, geographic azimuth of eddy opening (Dir_1), movement direction of eddy's main axis (Dir_2), and the eddy's movement velocity (S_p). The "outcome" refers to the eddy's intensity, direction, and movement speed at next moment. In summary, case of MOE can be represented by equation (4):

$$Case_i = \begin{cases} ID_i, P_i, A_i, A_{1i}, A_{2i}, \ldots, A_{8i}, F_{1i}, F_{2i}, \\ L_{oi}, L_{ai}, Dir_{1i}, Dir_{2i}, S_{pi}, \\ Vortex_{t1}\left(A_2, Dir_2, S_p\right) \rightarrow \\ Vortex_{t2}\left(A_2, Dir_2, S_p\right) \end{cases} \qquad (4)$$

3.2.2 Extracting spatial relationships and case library construction: Indicators used in equation (4) must be calculated before establishing case library. The four ocean environmental indicators are determined by GIS grid analysis while the two spatial direction indicators are calculated by examining direction relationships among polygonal objects [18]. All indicators are calculated by executing a VBA- algorithm in ArcMap. Fifty typical MOE were selected and imported into the case library to test the CBR approach. As the MOE usually lasts a long time, its formation process is divided into 5 stages (birth, development, stabilization, weakening and extinction) for the purpose to reduce the number of case in the library while maintaining related information about eddy's evolution. One historical case was identified to match each of these 5 stages respectively. Table 1 illustrates an example of the case library, with each row showing one case, i.e., one of the 5 development stages of an ocean eddy. As a result, information of one eddy is described in 5 rows. Columns in Table 1 show the indicators used to quantitatively describe case in the representation model. Ten eddies were randomly selected as test cases (not showing in this paper) to test the estimation accuracy of CBR method proposed in this research.

Tab.1 The case library of MOE in the SCS from 2003 to 2009

OID	*ID*	*A_1*	*A_2*	…	*F_1*	…	*Dir_1*	…	*P*
1	1	Warm	11.22	…	29	…	none	…	984352
2	1	Warm	13.26	…	29.45	…	W	…	1835143
3	1	Warm	28.48	…	30.66	…	E	…	2559812
4	1	Warm	20.77	…	29.08	…	W	…	4395433
5	1	Warm	12.37	…	30.16	…	none	…	1887941
6	2	Cold	12.31	…	23.49	…	none	…	1066029

7	2	Cold	15.50	…	23.07	…	SE	…	903146
8	2	Cold	17.86	…	22.77	…	none	…	980633
9	2	Cold	20.61	…	21.74	…	S	…	1415178
10	2	Cold	20.96	…	22.04	…	S	…	1347750
…	…	…	…	…	…		…	…	…
246	50	Warm	15.46	…	29.27	…	E	…	1049085
247	50	Warm	14.98	…	28.82	…	W	…	1566574
248	50	Warm	15.35	…	28.48	…	SW	…	2713391
249	50	Warm	15.97	…	28.65	…	SW	…	2814450
250	50	Warm	15.69	…	29.41	…	SW	…	3034540

Note: Unit of each field in this table varies. Intensity (A_2), surface temperature in eddy centre(F_1), eddy polygon perimeter (P), eddy polygon area(A) are measured in centimetre, degree, meter, and square meter respectively. "None" in the field of eddy opening direction (Dir_1) suggests a closed eddy while the other value showing its opening azimuth.

3.2.3 Similarity Calculation and Reasoning: Once the case library is constructed, equations in section 2.3 are used to calculate similarity among historical cases and predict the "outcome". Equation (5) plays a more important role in this study as it describes the direction relationship. Different weights are determined and then assigned to different attributes and spatial relationships before the general similarity among cases was calculated. Based on previous research results, following weights are directly assigned to different indicators: P: 0.05, A: 0.05, A1: 0.1; A2: 0.1; A3: 0; A4: 0.15; A5: 0.05; A6: 0.05; A7: 0.05; A8: 0.15, F: 0.05; F2: 0.05; Lo: 0.05: La: 0.05: and Sp: 0.05.The threshold value for similarity extraction is set as 70% in this test. After obtaining similar historical cases, algorithm in section 2.3 is used to perform the reasoning with different weights assigned to different extracted historical cases based on the similarity value. Weights of 0.2, 0.3, and 0.5 were respectively assigned to similarity values falling within the range of [0.7, 0.8), [0.8, 0.9), and [0.9, 1].Calculation results were shown in Tables 2, 3. Each row in the table represents one ocean eddy, while the "predicted value" columns correspond to the forecast outcome of the eddy direction and movement velocity. The predication accuracy in the table shows that how well the forecast result matches the actual value.

Table 2 Predication result and accuracy of movement direction

Case	Value(development)		AC.	Value (stabilization)		AC.(	value (weakening)		AC.	value (extinction）		AC.	Average
No.	Estimate	Actual	(%)	Estimate	Actual	%)	Estimate	Actual	(%)	Estimate	Actual	(%)	AC(%)
11	North 76.51	N.E 55.35	94.12	North 77.76	North 97.35	94.56	North 90.58	North 78.95	96.77	North 107.95	N.W 125.73	95.06	95.13
16	North 74.67	North 111.44	89.79	North 78.76	N.E 39.37	89.06	North 98.12	N.W 143.24	87.47	North 95.47	N.E 45.57	86.14	88.12
20	N.W 112.53	N.W 120.48	97.79	North 70.62	North 91.51	94.20	North 89.29	N.W 152.48	82.45	N.E 57.77	N.E 55.79	99.45	93.47
27	North 69.97	North 111.33	88.51	North 80.07	North 80.61	99.85	North 92	East 15.1	78.64	North 99.07	West 169.37	80.47	86.86
33	North 87.18	N.E 44.99	88.28	North 83.74	N.W 146.34	82.61	North 83.45	N.W 114.09	91.49	North 112.32	N.W 128.41	95.53	89.48
59	N.W 121.05	N.E 56.32	82.02	North 84.24	N.W 139.63	84.61	N.E 65.51	N.E 33.91	91.22	North 89.99	N.E 62.89	92.47	87.58
62	*	N.W 146.92	*	*	N.W 113.26	*	North 83.08	East 18.64	82.10	*	N.W 128.31	*	82.10
73	N.E 49.69	North 84.13	90.43	N.E 45.59	N.W 131.58	76.11	North 110.92	East 22.1	75.33	North 81.89	North 72.71	97.45	84.83
77	East 14.27	North 103.03	75.34	N.E 155.02	North 108.9	87.19	N.W 136.31	N.E 35.79	72.08	West 174.92	N.E 46.64	64.37	74.75
80	North 92.94	N.W 129.54	89.83	North 73.71	East 21.87	85.60	North 74.12	N.W 119.12	87.50	N.W 112.5	North 83.4	91.92	88.71
Average Accuracy			88.46			88.20			84.51			89.21	88

Note: ⋆indicates that no similar historical case can be identified in the case library under the given condition. As a result, no further attempts were made to predicate the movement velocity of this specific case. Direction is measured in angular degree. AC.means accuracy.

Table 3 Predication result and accuracy of movement speed

Case	Value (development)		AC	Value (stability)		AC	value (weakening)		AC	value (extinction）		AC.	Average
No.	P	Actual	(%)	Predicted	Actual	(%)	Predicted	Actual	(%)	Predicted	Actual	(%)	AC.(%)
11	10246.9	11961.7	92.9	8004.93	5154.61	88.16	6185.82	4145.33	91.52	10425.08	2358.33	66.49	84.76
16	11298.9	10027.2	94.7	9395.16	10587.7	95.05	6794.27	11462.74	80.61	10120.6	7577.43	89.44	89.96
20	6267.8	7226.6	96.0	5339.2	10978.3	76.58	5536.51	7941.14	90.01	5190.02	1685.5	85.44	87.01
27	10755.	10984	99.0	9323.92	6938.11	90.09	7230.63	4970.61	90.61	9765.13	7838.52	92	92.94
33	7111.8	5908.8	95	5541.57	8427.23	88.01	5969.03	5642.08	98.64	9512.17	6747.93	88.52	92.54
59	10520.1	942.4	60.2	8100.46	4002.56	82.98	5855.83	7465.94	93.31	7926.51	12293.4	81.86	79.59
62	*	2661	*	*	3263.79	*	5659.01	4748.39	96.22	*	15448.9	*	96.22
73	4525.4	5041.4	97.8	3191.43	2947.18	98.99	5589.74	1269.45	82.05	11959.4	1509.93	56.60	83.88
77	4726.7	618.4	82.9	1398.33	8695.48	69.69	1932.07	14675.5	47.07	2598.57	5322.16	88.69	72.10
80	5491.7	17164.2	51.5	5872.82	2751.12	87.03	5685.1	3204.71	89.70	9770.96	4529.37	78.23	76.62
Average accuracy			85.6			86.29			85.97			80.81	84.66

Note: ⋆indicates that no similar historical case can be identified in the case library under the given condition. As a result, no further attempts were made to predicate the movement velocity of this specific case. Velocity is measured in m/day.

3.2.4 Results: Test results suggest that estimation accuracy of majority of these 10 cases is over 80% with an average of 86.6%. Estimation accuracy of eddy's movement direction and velocity is slightly low. No similar historical cases were found for some cases when the threshold was set at 70%. However, for those cases with corresponding similar historical cases, estimation accuracy is above 80%. Average estimation accuracies of movement direction and velocity are 88% and 84.6% respectively. Due to limited number of ocean eddy case in the case library, not all cases have similar historical cases and thus the predication accuracy is very low. Once more cases are added into the library increases, this problem can be easily solved and estimation accuracy can be improved significantly.
Type subheadings flush with the left margin in bold upper case and lower case letters. Subheadings are on a separate line between two single blank lines. The blank line after is added automatically when using the provided Word template file.

4. CONCLUSIONS

A new "geographic environment" component was introduced into case representation model of the traditional CBR approach, which was then used to study the change of MOE in the SCS. Experiment result indicates that the method proposed in this paper is simple, flexible, and practical in studying MOE change with satisfactory estimation accuracy. As the case study indicates, the CBR method is able to provide solutions to some application-oriented problems and hence is capable to perform quantitative simulation and complex analysis to study change of MOE. In addition, case library built using CBR method can be dynamically updated. Self-training is also possible as more cases were accumulated. Continuously absorption of the high-quality data and improved research results will further boost estimation accuracy. In summary, this method shows great potentials in predicting rapid change of MOE.

5. REFERENCES

Ding H. 2004. the study of space Similarity theory and the calculation model. Doctoral Dissertation of Wuhan University.

Du YY, Su F Zh,, Zhang T Y,Yang X M,Zhou Ch H. 2005. Research of case-based reasoning on Marine information space vortex characteristics similar. *Journal of Tropical Ocean*, pp. 24(3):1-9.

Du YY, Zhou CH, Sh QQ. 2002. Theoretic and Application Research of Geo-Case Based Reasoning. *Geography Journal*, pp. 2(57): 151-158.

Fan LT, Wu SY, Chen J. 2003. Tthe similarity measure based on the mechanics of polygon. *Journal of Shanghai Jiao tong University*, pp. 37(6):874-877.

GeY, Du YY, Cheng Q M, Li C. 2007. the Application of Multifractal methods in remote sensing information processing and Extraction of eddies. *ACTC Oceanologica Sinica*, pp. 9(29) : 40-47.

Goyal P K. 2000. Similarity Assessment for Cardinal Directions Between Extended Spatial Objects: [Ph. D Dissertation] . Orono: The University of Maine.

He ZG, Wang DX. et al. 2001. the vortex structure of the South China Sea in Satellite tracking buoys and satellite remote sensing of sea surface height. *Journal of Tropical Ocean*, pp. 20(1), 27-35.

Ji GR, Chen X, Jia YZ. et al. 2002. An Automatic Detecting method of the marine mesoscale eddy in remote sensing image, *Oceanologia et limnologia Sinca*, pp. 33(2), 139-144.

Jones, E.K, Roydhouse, A. 1994. Intelligent retrieval of historical meteorological data. *Artificial Intelligence Appl*, pp. 8(3): 43-54.

Lan J, Hong JL, Li PX. 2006. the seasonal variation characteristics of cold vortex of the western South China Sea in summer. *Progress in Earth Science*, pp. 11(21):1145-1152.

Li L. 2002. The Research Overview of Mesoscale ocean phenomena in the South China Sea.

Lin PF. 2005. Statistical analyses on mesoscale eddies in the South China Sea and the Northwest Pacific .Chinese Academy of Sciences Institute of Marine Research.

Lin PF ,Wang F, Chen YL, Tang X H. 2007. Temporal and spatial variation characteristics on eddies in the South China Sea. *ACTC Oceanologica Sinica*, pp. 5(29):14-21.

Li X, Xie JA, Liao QF. 2004. The use of case-based reasoning (CBR) method of radar images for land use classification. *Journal of Remote Sensing*, pp. 8(3):246-252.

Li Y Ch, Cai W L, Li L, Xu D. 2003. The seasonal and inter-annual change of MOE in the northeastern South China Sea. *Journal of Tropical Ocean*, pp. 5(22): 61-70.

Qian YF, Wang QQ, Zhu BC. 2000. the Numerical simulation of the cold and warm eddies formed by the South China Sea wind. *Atmospheric Sciences*, pp. 9(24): 625-633.

Wang DX, Chen J. et al. 2004. the ocean temperature ,salinity and circulation features of the central and southern South China Sea in August 2000. *Ocean logia et limnology Sinca*, pp. 35(2), 97-109.

Wang DX, Shi P. et al. 2001. The mixed assimilation experiment on TOPEX sea surface height data in the South China Sea. *Ocean logia et limnology Sinca*, pp. 32(1): 101-108.

Wang GZ. 2001. Rough Set Theory and knowledge acquisition. Xi'an: Xi'an Jiao tong University, pp.1-25.

THE USE OF STATISTICAL POINT PROCESSES IN GEOINFORMATION ANALYSIS

Alfred Stein^a, Valentyn Tolpekin^a and Olga Spatenkova^b

^{a}ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands
^{b}Helsinki University of Technology, PO Box 1200, FIN-02015

ABSTRACT:

Many objects in space can best be modeled statistically by using point processes. Examples are fires in an urban environment, herds of animals in large areas, earthquakes and forest fires and large speckles on a radar image. Modern developments in point process theory now much better than before allow us to make statistical models to explain the observed patterns. In this paper, we will address the way that point processes can be modeled in space and time. The first application draws from domestic fires at the city level, where we apply a statistical point pattern analysis to derive major causes from related layers of information. The second application considers earthquakes as a marked point process. For earthquakes, large and complex data sets exist including many possibly relevant covariates that may influence their occurrence. The Strauss point process model is explored to analyze earthquake data in Pakistan recorded since 1973, in particular the major earthquake event occurring in 2005. The model, despite some limitations, is rigorous for applying it to such a marked point pattern, representing well the clustering behaviour as determined by a number of environmental factors. Finally, the Strauss point process model is suggested for the use in identifying and explaining the occurrences of speckles in a radar image.

1 INTRODUCTION

Spatial point pattern play an increasingly important role in modern image analysis and geographical information processing. On the one hand, we observe patterns of objects that show a point-like pattern, or at least can be modelled as such, whereas on the other hand several images inherently show point-like patterns. Typical examples of the first category are the locations of settlements in an area, the presence of wildlife herds observable from high resolution remote sensing, whereas in geographical information processing examples include the position of indoor fires in a large city, and the position of earthquakes in space and time. In particular several aspects of object-related noise may exhibit a point-like pattern, and the most common example of such is the presence of speckle on a radar image.

Spatial point pattern analysis is a powerful technique to detect relationships in spatial data distribution. The theory has rapidly grown in recent years and the background is described in an accessible way in (9) and (12), whereas a solid summary is given in (22). Classical examples exist in forestry (2, 13, 14), where either the positions of trees or the positions of gaps in fotrests are mdoeled as a point process. Other examples include studies in epidemiology (12, 17), or wildlife (23). (26) identifies practical difficulties when applying point pattern analysis methods in ecology and provides several relevant gudielines. The analysis methods usually first distinguish between clustering, regularity and randomness, and succeed by providing answers to questions about the scale of clustering and reasons behind the patterns. On the basis of an observed pattern we usually identify a process that generates these. This allows as well an analysis of spatial distributions in time (12). Typical examples discussed below are a Poisson process and a Strauss process, whereas also terms like a clustered or a regular process are used in the literature. It is usually the parameters of such process hat we are interested in, and that we may derive from a collection of observed points, e.g. within a limited window in space and time. Various software tools are now easily available for standard use. In this sense, an increasingly better match may arise between the patterns observable on images and understanding of processes occurring at the earth surface.

The aim of this paper is to briefly introduce the subject and then present some examples of data analysis and recognition. This will include some aspects of the Strauss process model as a specific model for application in spatial analysis.

2 METHODS

2.1 Point patterns

Basic concepts and analysis methods are in e.g. (9, 6, 19). Our interest concerns detection of systematics in the distribution, i.e. regularity or aggregation (clustersing) as deviation from randomness. Complete spatial randomness (CSR) is defined by the following criteria: (i) the number of events in a planar region A of size $|A|$ follows a homogeneous Poisson distribution with mean $\lambda|A|$, where λ is the constant density; (ii) given n events x_i in a region A, the x_i are an independent random sample from the uniform distribution on A (9). In other words, the density of the point pattern does not vary over the bounded region, and there are no interactions among the events.

Density estimation can be based on kernel functions (7) - a bivariate probability density function, which is symmetric around the origin located at a point of estimation. Incidents contribute to density estimation according to their distance from the kernel centre - the closer to the kernel centre, the larger the influence. The range of influence is limited by the kernel bandwidth controlling the smoothness of the result. Density plots with well-chosen bandwidth provide a good summary of the data, whereas a bandwidth that is too large leads to too much smoothing, and a bandwith that is too small over-emphasizes local events, like small variations in the incident pattern. Dependency relationships for local interactions can be described by the nearest neighbour distances defined as the distance from the ith event to the nearest other event in the bounded region of interest. Empirical cumulative probability distribution function $\hat{G}$ for the nearest neighbour distances summarises the incident pattern in an effective way:

$$\hat{G}(w) = \frac{\sum_{w_i \leq w} 1}{n},$$

where w_i is a nearest neighbour distance for the ith event and n is the number of events in the study region. Yet, the observed pattern is usually part of a larger region, where the distribution of events is unknown. Interaction between events lying inside and outside the study region cannot be properly accounted and cause edge effects. A simple but effective adjustment consists in reducing the sample by the buffer defined around the boundary. Events falling inside the buffer are not used for the analysis directly, but unveil the distribution behind the reduced study region.

To ease the interpretation, it is suitable to plot the $\hat{G}-$function against the theoretical curve for CSR, which is (ignoring the edge effects):

$$G(w) = 1 - exp(-\lambda \pi w^2).$$

Importance of the difference between $\hat{G}-$ and $G(w)$ is assessed by using Monte Carlo simulations. For this purpose, empirical cumulative probability distribution functions are generated for nearest neighbour distances for each of 99 realizations of a simulated CSR process with the same density as the original pattern. Its average provides a reference line, maximum and minimum values provide simulation envelopes.

A stochastic mechanism that generates a set of events in the study region is called a spatial point process. To model the dependence of domestic fires on exploratory variables we fit a process, termed a DF process. Its density function reflects the spatial distribution of the different influences. Assuming a stochastic dependence between the points, we use a class of Markov point processes (21), which allows flexible modelling of interpoint interactions. The Strauss process (24, 16) represents an example of Markov point process for pairwise interaction and can be used to simulate a wide range of patterns from simple inhibition to clustering (9, 15). The conditional density of Strauss process is

$$\lambda(u, x) = \beta(u) \cdot \gamma^{t(u,x)},$$

where $\beta(u)$ is the density at location u, $t(u, x)$ is the number of events x that lie within a distance r of u and the inhibition parameter γ controls the strength of interaction between points. For the special case that $\gamma = 1$ the Strauss model reduces to the homogeneous Poisson process with constant density β, the case that $\gamma = 0$ corresponds to a simple inhibition process, whereas for $\gamma > 1$ the model produces a clustered process. The effect of dependence on exploratory variables is expressed with a density being a loglinear function of covariates:

$$log\beta(u) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_n c_n,$$

where the c_i are the explanatory variables and β_i are parameters to be fitted. A linear form is chosen as a first approach in this exploratory study.

2.2 Goodness of fit

Modelling is an iterative procedure, aiming at finding a suitable representation of the data corresponding to observed relationships. The suitability of a model is checked according to several criteria. The Akaike Information Criterion AIC (1) is a versatile measure for model selection. In addition to goodness-of-fit it also considers the number of estimated parameters and the number of observations. A model with the lowest AIC value reflects the best trade-off between bias and variance.

The overall goodness-of-fit for the Strauss models can be assessed based on simulation envelopes of summary functions (9, 18, 3). The $K-$function provides a summary of the spatial pattern over a wide range of scales and is therefore more effective than measures based on the nearest neighbour distances. It is defined as the expected number of other points of the process lying within a distance d of a typical point of the process, divided by the density λ. A suitable estimate of this function given by (20):

$$\hat{K}(d) = \frac{R}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} I_d(d_{ij}),$$

where n is a number of points in the study region with area R, d_{ij} is the distance between ith and jth points, and $I_d(d_{ij})$ is an indicator function, which is 1 if $d_{ij} \leq d$ and 0 otherwise. After adjustment for inhomogeneity this becomes $\hat{K}_I(d, \lambda)$, defined as

$$\hat{K}_I(d, \lambda) = R^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{I_d(d_{ij})}{\lambda(x_i)\lambda(x_j)}.$$

We apply the reduced sample method to adjust the estimate for edge corrections. The $\hat{K}_I-$function is calculated for each of the realizations of a simulated models. In order to test the goodness-of-fit of the model, we consider global envelopes, which represent the largest absolute difference between the simulated and estimated theoretical curves over the entire distance interval. A significance level of 0.05 is achieved after 19 simulations.

Spatstat, an R package designed for analysing spatial point patterns was used for the analysis (4, 5, 3).

3 EXAMPLES AND ILLUSTRATIONS

3.1 Domestic fires

The first example considers domestic fires in Helsinki occurring within a single year. At the city scale, such domestic fires form a spatial pattern. We can derive various summary statistics describing pattern properties. These can represent first order effects describing the number of fires per unit area varying in a study region, or second order effects describing the dependency relationships between fires. A visual inspection of the $\hat{G}$ plot brings the spatial distribution of the pattern to light. An excess of nearest neighbors at short distances indicates clustering in the data, while an excess of long distances neighbors refers to regularity. Buildings and census records form an additional pattern of events. They carry information on types and age of buildings and socio-economical classes (density of population and workplaces, age of households, education, income, unemployment). Second order effects, in particular the $\hat{G}-$function between domestic fires and these patterns of selected influences, provide insight into the relationships in the data. If there are considerably more nearest neighbors at short distances than what would be expected for random distribution, we can assume a correlation between the events. In this way, processes underlying domestic fires are unveiled that indicate the importance of particular exploratory variables.

Modelling the distribution of domestic fires has been used to assess a probability of fire occurrence and analyse the contribution of explanatory variables. A point pattern analysis allows to preserve the level of detail offered by the data itself, in contrast to lattice methods that handle aggregated data. It avoids an ambiguous definition of a lattice scale and therefore enables to draw more accurate conclusions. The methods applied on buildings could be well applied to other cities and may include other phenomena that can be represented as a point pattern or a grid layer, such as crime distribution or house prices.

Conceptualization of studied phenomena as point pattern layers allows to apply well-established statistical methods for spatial point patterns analysis. In addition, spatial statistics offers more than a basis for accepting or rejecting null hypotheses about spatial randomness. The difference from randomness observed from the $\hat{G}-$function indicates a dependence between domestic fires and particular explanatory variables. The $\hat{G}-$function also provides an insight into the aggregation scales for separate variables. Comparison of plots for different variables helps to identify the most important influences. The Strauss model considers all the variables simultaneously and enables to quantify their influence to the distribution of domestic fires through the estimated parameters. The analysis of the distribution in time by splitting the set of events according to different time scales could be enhanced by using periodic splines to directly specify the time domain in the model. Yet, this is beyond the scope of the current manuscript and will be explored in the future.

The point pattern analysis can serve as a basis for generating new hypotheses and complement other data mining methods in the process of knowledge discovery. Still, before drawing reliable conclusions, the obtained results need to be discussed and confirmed with domain experts. Here we can benefit from the visual form of the density and $\hat{G}-$function plots.

The point pattern analysis is hampered by a large number of islands within the study area, which, as being built-up, need to be considered in the study. Rigorous distinction of land and water areas through the observation window and applying an edge correction would give a solution, however, it prolongs the processing time exceedingly. Having in mind that the frequency of incidents on the islands generally decreases with more tedious accessibility, we assume no significant effect of the observation window shape on the final results. To confirm this hypothesis, we performed the $\hat{G}-$function with the reduced sample edge correction method on a restricted area covering only a mainland of Helsinki. As expected, no major differences were observed in the results between the general and restricted observation windows. We therefore proceed the analysis using the simplified observation window.

Fitted models of domestic fires distribution reflect the empirical data. As there always exists a gap between the data and reality, the best model from a mathematical point of view does not need to be the best one in reality. Thus, it is desirable to consider also other criteria and keep the preferred model consistent with a priori knowledge.

Precise interpretation of the fitted parameter values indicates the relations between estimated density and variables involved in the model. However, the parameters can provide an accurate account of the process only in connection with the concrete variables values. The actual degree of influence of particular model variables can be assessed using AIC, for more details see (8). According to Akaike's rule of thumb, two models are significantly different, if the difference of their AIC is more than 2. Thus, model selection based on step-wise variable reduction comparing the AIC values leads to the model representing the most significant variables.

The method is data driven and the reliability of the results depends on the quality of the input datasets. We should therefore consider the data quality carefully and be aware of data quality problems that may occur. In this study we battled with the positional accuracy of the incident dataset. The coordinates of the incident location are inserted via an electronic report filled by the mission commander by clicking a mouse on the corresponding place in the map. This process should ensure the highest possible accuracy. As the commander's main responsibility is in extinguishing the fire, we may put some doubt on the precision of the coordinates of incidents, which may not correspond to incident addresses. Also, temporal variations of explanatory variables are unknown and may influence the results as, for example, population density data are based on permanent addresses. Additional uncertainty emerges with the data processing. We carried out the analysis by splitting set of events into various categories, that may have vague boundaries between them. Although we do not expect major changes in the results, this issue is postponed for a further analysis.

3.2 Earthquakes

In a recent study we investigated earthquake data in the Northern part of Pakistan, an active seismic zone. Data include 1403 earthquakes that occurred in the region between January 1973 and August 2008. The year 2005 is marked by a large seismic activity in the region as compared to the previous years. This is due to a major shock, the Kashmir earthquake of magnitude 7.6, which struck the region on Oct 8, 2005 followed by a range of aftershocks, causing great devastation and misery by killing more than 80000 people and damaging the whole infrastructure of the region. There were 22 earthquakes of magnitude 5.5, out of which 12 occurred the same day as the major earthquake and 15 earthquakes of magnitude 5.5 occurred within 15 days after the Kashmir earthquake. Only 7 other earthquakes of magnitude 5.5 occurred during the past 35 years. The seismicity of the area decreases after the first month after the Kashmir earthquake and the number of events in the preceding months is almost negligible as compared to the first month after the main shock. Locations of all earthquakes within one month after the Kashmir earthquake are in its close vicinity. Only four earthquake locations lie more than 50 km from the aftershocks region. The analysis of the data considering it as a point pattern will be based on this study region and the earthquakes located within it.

The epicentre region lies on the western edge of the Himalayan Arc, which denotes the area of continental convergence between the Indian and Eurasian tectonic plates. The Indian plate moves northwards at a rate of about 40mm/year and subducts below the Eurasian plate. The Kashmir earthquake is associated with fault rupture near the western end of the MBT in Kashmir region of Northern Pakistan. Location of tectonic plates boundaries plays a significant role in determining seismicity of the study area.

To assess the influence of geological faults located in the study region on the earthquakes distribution pattern, the distance of earthquake locations to faults could serve as additional information (covariates) in modelling the point pattern. For that purpose a a geo-referenced Tectonic map of Pakistan. From this map the study area of the earthquakes data was extracted using its bounding coordinates and the faults within the study area were digitized. Aftershocks earthquakes occurred along the plate boundaries with a dense cluster of aftershocks near the point where the two boundaries converge. Thus the location of plate boundaries can possibly serve as an important factor contributing to the distribution pattern of the earthquakes. To evaluate the contribution of plate boundaries location, a pixel image of the shortest distance of each pixel from the pate boundary was obtained. Similarly, to test the effect of active faults in the study area on the earthquake point pattern, distance of each earthquake location was calculated from the nearest fault.

An earthquake hypocenter is the three dimensional point in the earth where the rupture of an earthquake begins. For large earthquakes, the ruptures may extend up to several kilometres, and the hypocenter may be anywhere along the rupture. The epicentre of an earthquake event is the point location on the surface of the globe that represents the projection of the hypocenter onto the surface of the globe.

The explanatory variables, apart from the Cartesian coordinates, consisted of the information about the spatial location of the plate boundaries and geological faults in the study area given as pixel images showing shortest distance to the nearest plate boundary and nearest fault location for each pixel. The application of Strauss point process model proved satisfactory in explaining the spatial trends and capturing the sources of variability introduced by the explanatory variables. The application showed that the locations of plate boundaries and geological faults are significant determinants for the earthquake epicentre locations. When the effects of both these variables were combined along with the magnitudes and geographic locations of the earthquake epicentres, the modelling was significantly improved. The effects of the explanatory variables were quantified by improvement in AIC values. The improvement in the modelling of earthquake location can also be assessed visually by the plots of fitted trends for different types of earthquakes.

3.3 Speckle on remote sensing images

Radar images provide important information about the earth surface that is complementary to optical remote sensing images. In some cases it has advantage over optical images, e.g. in dense cloud cover conditions and observation at night time. Synthetic Aperture Radar (SAR) is a special case of radar system where relatively high spatial resolution is achieved due to coherent processing of many recorded responses.

Speckle is an inherent property of SAR images. It originates from interference of coherent responses coming from many scattering elements within a resolution cell. It results into a large variance of radar image compared to its mean value. Therefore SAR images are difficult to use for automatic classification purposes. In several instances, it is required to reduce the effect of speckle, being this goal of SAR image despeckling.

As an example we may consider the ERS-2 Single Look Complex (SLC) image covering Serowe region in Botswana (see Figure 1. The patterns of the spikes show a point structure that may identify important issues related to land processes. The image presented attached subset of HH image of Botswana. The bright spots in the left part of the image are results from strong reflectors in a city. Hence, also after despeckling, the remaining pattern of extremes shows a pattern that can be readily analyzed and interpreted using a statistical point pattern analysis.

4 DISCUSSION

At this stage, good results are obtained with the combination of spatial point pattern theory and remote sensing, as well as in its combination with geographical information processing. The combination of readily available software tools and the request for an increasingly better data quality may lead to a more regular use of the methodology, thus leading to answering relevant questions. In particular, modern methods on space-time point processes may become beneficial to better understand the development of patterns in space and time.

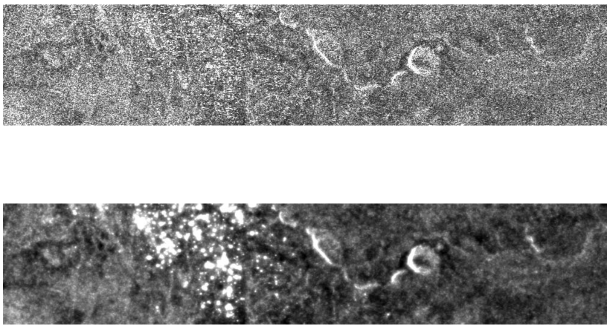


Figure 1: (an ERS-2 Single Look Complex (SLC) image covering Serowe region in Botswana and the same image after despeckling. .

Standard GIS packages do not yet contain easy to use and interpretable spatial statistical software. In the context of geoinformation processing, this is a clear deficiency, as such procedures are useful for a wide set of applications. A better integration of the spatial statistical software and GIS packages is a necessary step forward. An important reason in this respect is that a spatial statistical summary of collected or registered point data may be helpful to further communicate quantitative findings to the user.

A step to further explore concerns the issue of spatial data quality. Spatial data quality is firstly relevant in terms of accuracy of the observations,. In the study described above on earthquakes this plays an important role, as an earthquake occurs at some depth below the Earth crust, whereas its effects are mainly visible at the surface. Moreover, there is never a precise location of such an event, and only an approximate value. The second issue related to point patterns is their attribute, which may be difficult to define in full. The domestic fire example may at several instances relate the question whether any fire that is registered in a house is in fact domestic fire. Buildings may be used for different purposes, and there is usually an issue of not reporting such a fire to fire brigades, or reporting it in a deviate way. Such issues apply to a range of other spatial point patterns in a similar way. An as yet somewhat unexplored domain concerns the use of marked point processes in remote sensing images. When additional information comes available from images, it is not difficult to imagine that such methods can be useful for a range of applications. In particular, we see good opportunities in deforestation studies and in development of urban regions.

Finally, recent progress has been made on spatial processes in modeling of spatial extremes and of modeling point patterns in the space-time domain. A good example of the first type of study is in soil contamination. The second type of analysis is well presented in two recent papers (11), (10).

REFERENCES

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Atkinson, P. M., Foody, G. M., Gething, P. W., Mathur, A. and Kelly, C. K., 2007. Investigating spatial structure in specific tree species in ancient semi-natural woodland using remote sensing and marked point pattern analysis. *Ecography* 30, 88–104.

Baddeley, A., 2008. Analysing spatial point patterns in R. *CSIRO workshop notes [online]*. Available from: http://www.csiro.au/files/files/pn0y.pdf [Accessed 1 April 2008].

Baddeley, A. and Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* 12 (6), 1–42.

Baddeley, A. and Turner, R., 2007. *Spatial Point Pattern analysis* [online]. Available from: http://www.spatstat.org

Bailey, T. C. and Gatrell, A. C., 1995. *Interactive Spatial Data Analysis*. Longman, Harlow.

Bowman, A. W. and Azzalini, A., 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, New York.

Burnham, K. P. and Anderson, D. R., 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

Diggle, P. J., 2003. *Statistical Analysis of Spatial Point Patterns*. 2nd ed. Arnold, London.

Diggle, P. 2007. Spatio-Temporal Point Processes: Methods and Applications. In Finkenstadt, B., Held, L. and Isham, V. (eds) *Statistical methods for spatio-temporal systems* Chapman and Hall, Boca Raton, pp. 1–45.

Diggle, P. and Gabriel, E. 2009. Spatio-Temporal Point Processes. In Gelfand, A., Diggle, Guttorp, P. amd Fuentes, M. (eds) *Handbook of Spatial Statistics* Chapman and Hall, pp. 449 – 461.

Gatrell, A. C., Bailey, T. C., Diggle, P. J. and Rowlingson, B. S., 1996. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, NS 21(1), 256–274.

Getis, A. and Franklin, J., 1987. Second-order neighborhood analysis of mapped point patterns. *Ecology*, 68(3), 473–477.

Getzin, S., Dean, Ch., He, F., Troyfomow, J. A., Wiegand, K. and Wiegand, T., 2006. Spatial patterns and competition of tree species in a Douglas-fir chronosequence on Vancouver Island. *Ecography*, 29(5), 671–682.

Gregori, P. and Mateu, J., 2002. Spatial point processes: an overview. *In*: J. Mateu, Montes and F. eds. *Spatial Statistics Through Applications*. WIT Press, Southhampton, Boston.

Kelly, F. P. and Ripley, B. D., 1976. A note on Strauss's model for clustering. *Biometrika*, 63, 357–360.

Martínez-Beneito, M. A., Abellán, J. J., López-Quílez, A., Vanaclocha, H., Zurriaga, Ó., Jorques, G. and Fenollar, J., 2006. Source detection in an Outbreak of legionnaire's disease. *In*: A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan eds. *Case Studies in Spatial Point Process Modeling*. Lecture Notes in Statistics 185, Springer, New York.

Mattfeldt, T., Eckel, S., Fleischer, F. and Schmidt, V., 2007. Statistical modelling of the geometry of planar sections of prostatic capillaries on the basis of stationary Strauss hard-core processes. *Journal of Microscopy* 228 (3), 272–281.

O'Sullivan, D. and Unvin, D. J., 2003. *Geographic Information Analysis*. Wiley, Hoboken.

Ripley, B. D., 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13, 255–266.

Ripley, B. D. and Kelly, F. P., 1977. Markov point processes. *Journal of the London Mathematical Society*, 15, 188–192.

Schabenberger, O,. and Pierce, F.J., 2002. Contemporary statistical models for the plant and soil sciences. CRC Press, Boca Raton.

Stein, A. and Georgiadis, N., 2006. Spatial Marked Point Patterns for Herd Dispersion in a Savanna Wildlife Herbivore Community in Kenya. *In*: A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan eds. *Case Studies in Spatial Point Process Modeling*. Lecture Notes in Statistics 185, Springer, New York.

Strauss, D. J., 1975. A model for clustering. *Biometrika*, 62, 467–475.

Walter, C., McBratney, A.B., Viscarra Rossel, R.A. and Markus, J.A. 2005. Spatial point-process statistics: concepts and application to the analysis of lead contamination in urban soil. *Environmetrics* 16 339355

Wiegand, T. and Moloney, K. A., 2004. Rings, circles, and null-models for point pattern analysis in ecology. *OIKOS*, 104(2), 209–229.

9

URBAN ROAD NETWORK ACCESSIBILITY EVALUATION METHOD BASED ON GIS SPATIAL ANALYSIS TECHNIQUES

Hu Weiping, Wu Chi

School of Geography, South China Normal University,
Shipai, Guangzhou, P. R. China, huweipingok@sina.com

Commission II, WG II/3

KEY WORDS：Accessibility; Road Network Evaluation; GIS; Spatial Analysis

ABSTRACT:

The urban road network plays a key role in the urban spatial structure. It is the main city social-economy activities and transportation carrier. Today, more and more researchers pay attention on road network. One of the most important problems is how to evaluate the accessibility of road network. This paper tries to discuss it. Firstly, road accessibility concept and some appraisal methods are discussed. Then, the spatial analysis method on road network assessment has established based on the GIS spatial analysis technology, some urban road network accessibility evaluation models are built up. The models use ESRI Corporation's ArcGIS Engine components and Microsoft Corporation. Net Framework, and focus on the road network connectivity, the shortest travel time and the weighted average travel time. The paper presented three main road network accessibility evaluating indicators, introduced theory basis of the model construction in detail, and the model construction process. Taking Foshan city as an example, the models were tested using the urban road network data. Finally, further urban road network accessibility evaluation models are discussed.

1. INTRODUCTION

The urban road network plays a key role in the urban spatial structure. It is the main city social-economy activities and transportation carrier. Today, more and more researchers pay attention on road network. One of the most important problems is how to evaluate the accessibility of road network.

Hansen proposed the accessibility concept for the first time, defines it as the transport network in various nodes interaction opportunity (Hansen, 1959). Hereafter, the accessibility widely applied in research and road network plan, construction, and evaluation. In the transportation geography, the road network accessibility evaluation has taken as an important problem.

But accessibility does not have a unification concept till now. Generally speaking, the accessibility is the weight of a place to another place's simple degree and efficiency. Yang and Zhou thought that the accessibility is a place's convenience degree which arrives from other place. It can be a spatial distance, topological distance, trip distance, travel time or transportation costs(YANG Jiawen, ZHOU Yixing,1999).

The accessibility has both spatial and time features. It displays the convenience degree of a place as a spatial entity. And time is the main impedance factor of accessibility.

2. EVALUATION MODEL OF ROAD NETWORK ACCESSIBILITY

There are many evaluation indicators are proposed by researchers. We use three evaluation indicators to establish our model:

1-Shortest Time Distance(STD). It refers to the total time that one node need to all other nodes in the road network by the shortest time spending route. The lower STD value that a node has indicates that the node's accessibility is higher. The model expression is:

$$A_i = \sum_{j=1}^{n} T_{ij} \quad \left(i, j \subset (1, n)\right) \cdots (1)$$

In the formula, A_i for the node accessibility value, its value may change from 0 (the self node) to +∞ (not connect node); T_{ij} is the least travel time from node i to node j; n is the total number of road network nodes.

2-Weighted Average Travel Time(WATT). It is the weighted summation of the total time that a node needed to all other nodes in the road network by the shortest time spending route. The weight represents the importance of a node in the road network, it can be calculated by population density or economical indexes. The WATT value is mainly related to the node's position in the road network . For example, the node in the central region usually has a smaller value. The model expression is:

$$A_i = \sum_{j=1}^{n} T_{ij} \times M_j \bigg/ \sum_{j=1}^{n} M_j \cdots (2)$$

In the formula, A_i for the node accessibility value, its value may change from 0 (the self node)to +∞ (not connect node); T_{ij} is the least travel time from node i to node j; M_j is node j's weight; n is the total number of road network nodes.

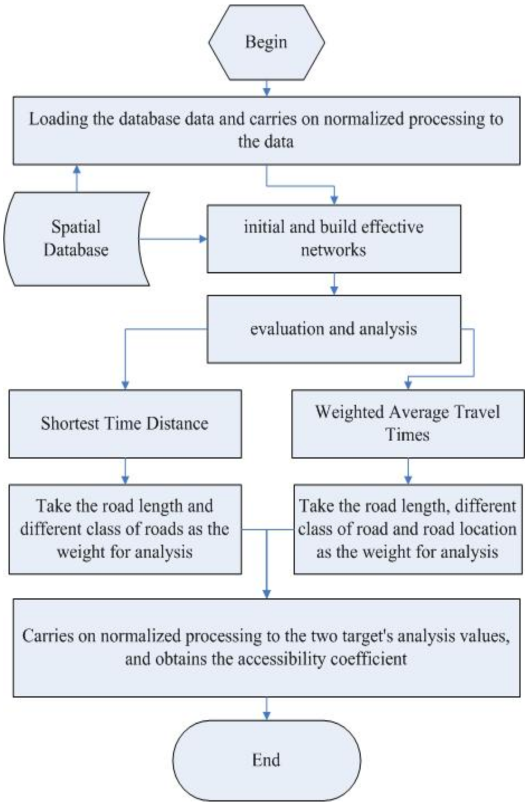
3-Accessibility index. It is a normalized index for the shortest travel time and the weighted average travel time. The formula is:

$$A_i^{'} = A_i \bigg/ \left(\sum_{i=1}^{n} A_i \Big/ n \right) \cdots (3)$$

In the formula, $A_i^{'}$ is the accessibility index; A_i is one node's accessibility value; $\sum_{i=1}^{n} A_i \Big/ n$ is the mean value of accessibility.

3. METHOD OF ROAD NETWORK ACCESSIBILITY EVALUATION

The work flow is shown in figure 1.



Begin
Loading the database data and carries on normalized processing to the data
Spatial Database
initial and build effective networks
evaluation and analysis
Shortest Time Distance
Weighted Average Travel Times
Take the road length and different class of roads as the weight for analysis
Take the road length, different class of road and road location as the weight for analysis
Carries on normalized processing to the two target's analysis values, and obtains the accessibility coefficient
End

Fig.1 Road network accessibility evaluation flowchart

3.1 The method of road network accessibility evaluation

- ◆ We select ESRI PersonalGeodatabase to deal with data, and build up the FeatureDataset, construct effectiveness network (it calls the Utility Network), use the INetworkCollection class to carry on the geometry network to built CreateGeometricNetwork. Its foundation method is:

```
        //network class
        INetworkCollection      m_NetworkCollection      =
(INetworkCollection)m_FeatureDataset;
        //create utility network
        IGeometricNetwork      m_GeometricNetwork      =
m_NetworkCollection.CreateGeometricNetwork(
            "MyNet", esriNetworkType.esriNTUtilityNetwork,
true));
```

- ◆ Carries on normalized processing to the data precision, the line feature connection precision is 0.001 meter (i.e. two line feature connected node tacitly approves in 0.001 meter for connection).
- ◆ Geometry network analysis
  - ■ Calculates each node network connectivity, the following is the key code:

```
        //An initialization network flows object
        ITraceFlowSolverGEN   traceFlowSolver   =   new
    TraceFlowSolverClass();
        //Use the transfer network flow method to carry on
    connection essential factor tracing
          traceFlowSolver.FindFlowElements(esriFlowMethod
            .esriFMConnected,
            esriFlowElements.esriFEJunctionsAndEdges,   out
            m_JunEnumNetEID, out m_EdgeEnumNetEID);
```

- Calculates each node's most short-path, the following is the key code:

```
        //set the Weight
        INetSchema NetSchema = Network as INetSchema;
        INetWeight                    NetWeight              =
     NetSchema.get_WeightByName(WeightName);
        NetSolverWeights.FromToEdgeWeight =
   NetWeight;
        NetSolverWeights.ToFromEdgeWeight                    =
NetWeight;
            //find the path
        traceFlowSolver.FindPath(esriFlowMethod.esriFM
Connected,
              esriShortestPathObjFn.esriSPObjFnMinSum,out
              m_JunEnumNetEID,  out  m_EdgeEnumNetEID,
              intCount - 1, ref vaRes);
```

◆ According to formula 1 and formula 2, we calculate the values of STD and WATT, then carry on normalized processing, get the accessibility coefficient value.

```
        //STD values
        for (int i = 0; i < n; i++)
            TotalSTDValue     =     TotalSTDValue     +
STDValue[i];
          STDAcc[i]= STDValue[i] / TotalSTDValue / n;
        //WATT values
        for (int i = 0; i < n; i++)
            TotalWATTValue   =   TotalWATTValue   +
WATTValue[i];
        WATTAcc[i]= WATTValue[i] / TotalWATTValue
     / n;
```

◆ The analysis data processing

- Uses the report form to express the values of STD, the values of WATT and the accessibility coefficient values;
- Carries on the sector classification to each node by the accessibility coefficient's difference, and carries on the rank exaggeration on the graph;

3.2 The road network evaluation system

We take the central region of Foshan city as an example, the road data of 2008 were used in the test. The nodes (the green dots) distribution is shown in figure 2.

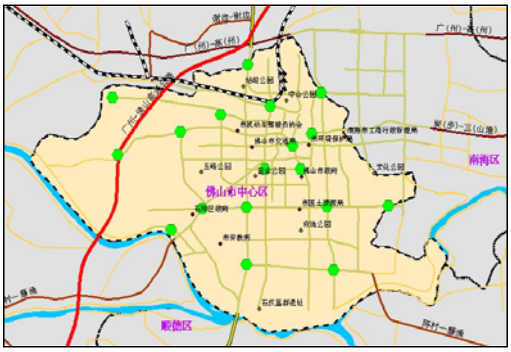
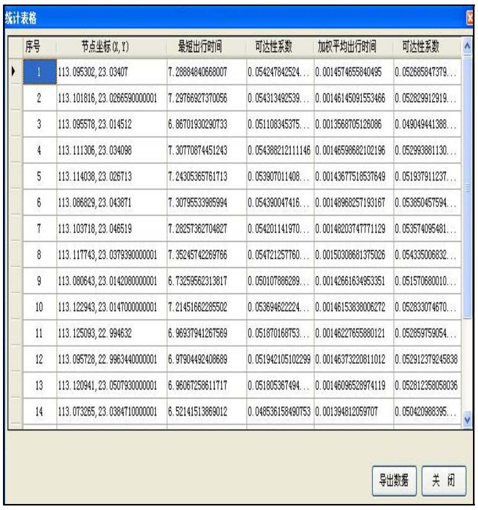


Fig.2 Sample nodes' location

The shortest travel time. The weighted average time as well as normalized process to the data, the processed result is as following (Fig.3).



统计表格

序号	节点坐标(X,Y)	最短出行时间	可达性系数	加权平均出行时间	可达性系数
1	113.095302,23.03407	7.28884840668007	0.054247842524...	0.0014574655840495	0.052685847379...
2	113.101816,23.0266590000001	7.29766927370056	0.054313492539...	0.00146145091553466	0.052829912919...
3	113.095578,23.014512	6.86701930290733	0.051108345375...	0.0013568705126086	0.049049441388...
4	113.111306,23.034098	7.30770874451243	0.054388212111146	0.00146598682102196	0.052993881130...
5	113.114038,23.026713	7.24305365761713	0.053907011408...	0.00143677518537649	0.051937911237...
6	113.086829,23.043871	7.30795533985994	0.054390047416...	0.00148968257193167	0.053850457594...
7	113.103718,23.046519	7.28257362704827	0.054201141970...	0.00148203747771129	0.053574095481...
8	113.117743,23.0379390000001	7.35245742269766	0.054721257760...	0.00150308681375026	0.054335006832...
9	113.080643,23.0142080000001	6.73259562313817	0.050107886289...	0.00142661634953351	0.051570680010...
10	113.122943,23.0147000000001	7.21451662285502	0.053694622224...	0.00146153838006272	0.052833074670...
11	113.125093,22.994632	6.96937941267569	0.051870168753...	0.00146227655880121	0.052859759054...
12	113.095728,22.9963440000001	6.97904492408689	0.051942105102299	0.00146373220811012	0.052912379245838
13	113.120941,23.0507930000001	6.96067258611717	0.051805367494...	0.00146096528974119	0.052812358058036
14	113.073265,23.0384710000001	6.52141513869012	0.048536158490753	0.001394812059707	0.050420988395...

Fig.3 Computed result

The Kriging space interpolation calculation. After the shortest travel time normalized into the accessibility coefficient, the interpolation result is as follows (Fig.4).

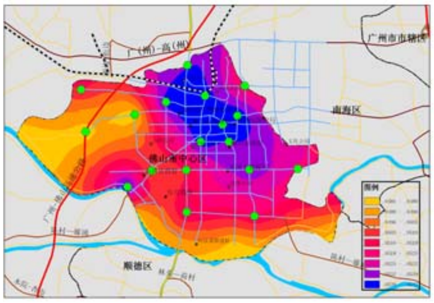


Fig. 4 Shortest travel time value interpolation chart

After the weighted average travel time normalized into the accessibility coefficient, the interpolation result is as follows(Fig.5).

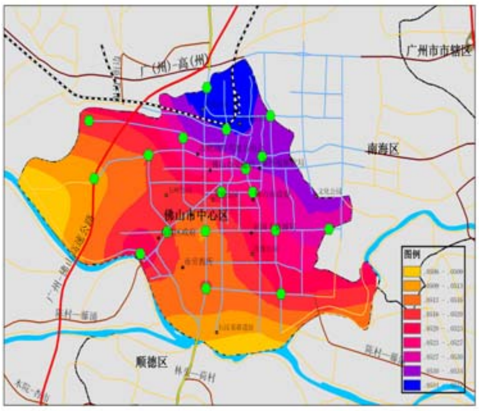


Fig.5 Weighted average travel time value interpolation chart

3.3 Discussion of the results

In Figure 4 and Figure 5, two assessment methods obtain the similar results. From North-east to the South-west, the Foshan central region's accessibility value decreases gradually.

In Figure 4, it mainly indicates the node’s shortest general time. The central region of Foshan has relatively higher value of accessibility.

In Figure 5, on the one hand, the accessibility value displays the node shortest travel time characteristic. On the other hand, as the node has joined the weights on centricity and transportation rank, the northern region with a railroad, a national highway and provincial highway and so on, then it has a higher accessibility value.

In Figure 4 and Figure 5, the western region has relatively lower value of accessibility.

4. SUMMARY

Based on GIS spatial analysis methods, using ESRI Corporation's ArcGIS Engine components and Microsoft Corporation .Net Framework, we built a weighted and normalized index to value the accessibility of road nodes. In the sample test of Foshan city, the results show that the index can explain the true situation of road network’s accessibility.

REFERENCES

Hansen W.G. How accessibility shapes land-use. Journal of the American Institute of Planners，1959, 25, pp. 73-76.

YANG Jiawen, ZHOU Yixing. Accessibility: Concept, Measure And Application. Geography and Territorial Research, 1999, 15(2), pp.61-66.

RELEVANCE-DRIVEN ACQUISITION AND RAPID ON-SITE ANALYSIS OF 3D GEOSPATIAL DATA

D. Eggert, V. Paelke

IKG, Institute for Cartography and Geoinformatics, Leibniz Universität Hannover, Appelstr. 9a, 30167 Hannover, Germany, {eggert, paelke}@ikg.uni-hannover.de

KEY WORDS: Information Visualization, 3D Geovisualization, Data Analysis, GPGPU, CUDA, OPENCL, Density Calculation, k-nearest-neighbors

ABSTRACT:

One central problem in geospatial applications using 3D models is the tradeoff between detail and acquisition cost during acquisition, as well as processing speed during use. Commonly used laser-scanning technology can be used to record spatial data in various levels of detail. Much detail, even on a small scale, requires the complete scan to be conducted at high resolution and leads to long acquisition time, as well as a great amount of data and complex processing.

Therefore, we propose a new scheme for the generation of geospatial 3D models that is driven by relevance rather than data. As part of that scheme we present a novel acquisition and analysis workflow, as well as supporting data-models. The workflow includes on-site data evaluation (e.g. quality of the scan) and presentation (e.g. visualization of the quality), which demands fast data processing.

Thus, we employ high performance graphics cards (GPGPU) to effectively process and analyze large volumes of LIDAR data. In particular we present a density calculation based on k-nearest-neighbor determination using OpenCL.

The presented GPGPU-accelerated workflow enables a fast data acquisition with highly detailed relevant objects and minimal storage requirements.

1. MOTIVATION

A wide variety of 3D geospatial based applications have been proposed in recent years, mostly in relation to city modeling but also for other domains. Application paradigms like 3D location based services and augmented reality rely on appropriate 3D models of the environment as basic constituents. Despite technical advances, the cost effectiveness of creating and maintaining the required 3D models, as well as their appropriate presentation to users, remain a key issue. This situation is further complicated by the fact that 3D geospatial information is subject to frequent changes and that the current developments in 3D computer games and film animation lead users to expect a high fidelity of the models used in an application.

One central problem in applications using 3D model is the tradeoff between detail and acquisition cost, during conception, as well as processing speed, during use. Much detail, even on a small scale, requires the complete scan to be conducted at high resolution and leads to long acquisition time, great amount of data, and complex processing. Fast scanning in contrast will be shorter in duration but will provide lower resolution and an overall coarse model. Adding more detail and using more realistic graphics may seem the obvious solutions. However, often they are neither cost effective nor viable using existing techniques.

We suggest looking for alternative ways to provide 3D information on a large scale. Recent research has found that in a variety of visual applications that use 3D city models, a high amount of detail is only required for objects that are of high relevance to the user (Cartwright 2005)(Elias, Paelke and Kuhnt 2005).

We propose the generation of large-scale geospatial models that are driven by relevance rather than data. Particularly, developing new progressive acquisition and modeling techniques that provide a more coherent view into the available sources of information. To achieve this, our plan is to use laser-scanning technology and novel user interface techniques, providing instant visual feedback which demands fast data analysis.

The established workflow consists of a data acquisition stage in the field and a following processing and analysis stage in a standard in-door office environment. This makes fast multi-core computers or even entire computer clusters available for the analysis. An on-site environment lacks this computation power; therefore we employ high performance graphic cards to process and analyze the LIDAR data at the recording site. Compared to common CPUs, even mobile GPUs (build into Laptops) have a significant increased computation power. Since these computation capabilities entail higher power consumption, most high performance mobile GPUs come along with an integrated low power GPU. In case the high performance GPU is not needed, it will be deactivated and the separate low power GPU takes over for longer battery life. Since the high performance GPU is only needed during the analysis, the concept of having a separate low power GPU fits the demands of our on-site analysis.

2. RELATED WORK

Almost all approaches to recognize salient objects and reconstruct their shape are data driven, targeting the extraction of every detail from the data, needed or not, see e.g., (Volsseman and Dijkman 2001), (Rottensteiner and Briese 2002), (Filin 2004), (Filin, Abo-Akel and Doytsher 2007), (Becker and Haala 2007). Recent advances in terrestrial laser scanning has shown that processing the point-clouds can be performed, under adequate representation, both efficiently in terms of processing time and with relatively limited computational resources (Zeibak and Filin 2007), (Gorte 2007), (Barnea and Filin 2007), (Barnea and Filin 2008). These results refer to the registration of the point clouds (Barnea and Filin 2007), (Barnea and Filin 2008), the extraction of primitives and objects (Gorte 2007), (Barnea, Filin and Alchanaties 2007), and to the association of scan pairs (Zeibak and Filin 2007).

The software currently used in the 3D reconstruction process and for data acquisition is designed for operation in standard in-door office environments, e.g., (InnovMetric 2008), (Cyclone 2008). Regarding on-site interaction, the user interface concepts of mixed and augmented reality (Milgram, et al. 1994), (Azuma 1997), (Azuma, Baillot, et al. 2001) that integrate the real environment into the user interface have shown high potential to support complex spatial interaction tasks. As an example, the Studierstube system (Schmalstieg, et al. 2002) demonstrates a number of promising spatial interaction concepts. Hedley (Hedley, et al. 2002) and others have demonstrated collaborative 3D geovisualization applications based on augmented reality techniques. While the technical challenges of mobile outdoor are great, there have been a number of demonstrators, e.g., the outdoor modeling application by (Piekarski and Thomas 2001). Another AR input/output device is the GeoScope (Paelke and Brenner 2007) that aims to avoid some of the central problems by providing high-precision video overlay in outdoor use-cases where high mobility is not required and seems well suited for acquisition applications.

A point cloud's density is an important indicator of its quality. In order to determine this density the k-nearest-neighbors (kNN) can be used. Most approaches for determining the kNN of a point in a point set rely on reducing the complexity of the required neighbor searches. They generally try to reduce the number of distances to calculate by arranging the data in spatial data structures, e.g. a kd-tree structure (Arya, et al. 1998) or by using Morton order or Z-order of points as in (Connor and Kumar 2008). Another recent proposal with promising results uses a brute-force search implemented using the C for CUDA API (Garcia, Debreuve and Barlaud 2008).

3. CONCEPT

The objective of our work is the effective creation of 3D geospatial models based on integrating global data (airborne laser or alternative sources) if available with local detail from terrestrial laser scans through progressive acquisition and modeling. Such modeling will be according to need, relevance, and controlled on-site with an augmented reality user interface.

The central idea is to control and limit the amount of detail in all processing stages to that actually required while providing feedback and on-site interaction capabilities. This will allow reducing acquisition time, modeling time, as well as storage and computation requirements in the actual use of the resulting models. Additionally, it will allow to focus on the relevant features needing further detailing. Such focus is almost impossible to achieve using uniform scans. For this we propose a demand-driven workflow, as shown in Figure 1, into which the acquisition, analysis, integration and presentation activities are embedded.

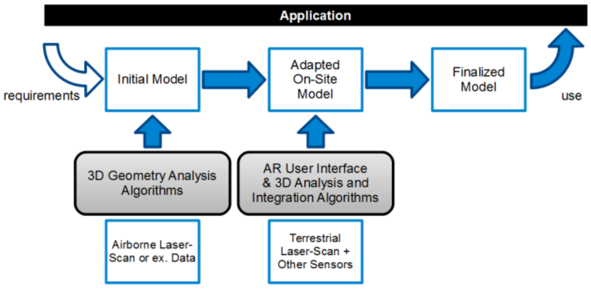

Application
requirements
Initial Model
Adapted On-Site Model
Finalized Model
use
3D Geometry Analysis Algorithms
AR User Interface & 3D Analysis and Integration Algorithms
Airborne Laser-Scan or ex. Data
Terrestrial Laser-Scan + Other Sensors

Figure 1: acquisition and analysis workflow

Based on the application and requirements, an initial model is acquired via airborne laser scanning where and if possible to provide a cost effective base model. Alternatively, existing 2D or 3D models can be used if available.

The central activity is on-site modeling, in which terrestrial laser scanning is employed. A user interface based on the paradigm of augmented reality (AR) in which the view of the real environment is augmented with information on the current model, its resolution and quality allows to control the acquisition and modeling process through intuitive decisions and selection of relevant features worth or need detailing. A mobile version of the previous mentioned GeoScope constitutes a suitable AR setup.

The AR user interface is closely coupled to 3D geometry analysis and integration algorithms that match and integrate data from different scans and data sources and provide measures of object distinctiveness, complexity and scan quality. In contrast to common off-line processing techniques, this approach requires 3D geometry analysis and integration schemes that operate at interactive speeds. Thus, we employ high performance graphics cards to speed up the analysis algorithms.

The resulting model can be further extended and refined either on-site or back in the office using established modeling tools or dynamically generated structure elements (e.g. pre-packed facade features from a library) to match the application requirements. It can then be applied in the intended application (e.g. visualization or precise positioning).

4. GEOMETRIC ANALYSIS ALGORITHMS

In order to create a 3D model from the LIDAR data, various geometry analysis algorithms must be applied. Beside several other algorithms the determination of the k nearest neighbors (kNN) of each point in the recorded point cloud is commonly needed. While algorithms like triangle-mesh reconstruction use kNN for surface reconstruction, our approach employs kNN to determine a density value for each point. The density in each point gives information about the quality of the recorded point set. The density d_P of point P is the inverted sum of the distances between P and the k nearest neighbors P_i of P.

$$d_P = \frac{k}{\sum_{i=0}^{k} |PP_i|}$$

The association of colors to the minimum and maximum density value (e.g. max density = green, min density = red) allows to visualize the calculated density values, as shown in Figure 2.

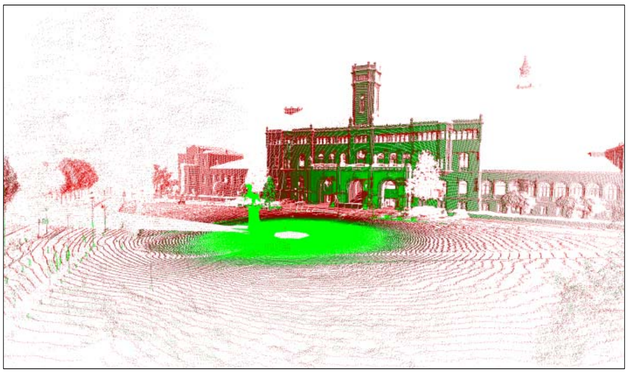


Figure 2: density visualization

Doing the kNN calculations on graphics processing units (GPU) raises two questions:

- Which kNN algorithm to implement
- Which programming language to use

The technique of using a GPU to perform calculations is often referred to as general purpose computation on GPUs (GPGPU). GPGPU is basically a kind of stream processing that exploits the parallel data processing capabilities of modern GPUs. Therefore the used algorithm needs to be highly parallelizable. Furthermore the processed data, or more specific the corresponding results, must be independent. Regarding the programming language to use leaves basically two options. Manufactures of graphics hardware like Nvidia and ATI provide their own proprietary GPU computing languages, like Nvidias “C for CUDA” or ATIs “Stream”. Their major drawback lies in the particular hardware support, since only the manufacturers own platforms are supported. On the other hand there are hardware independent languages like OpenCL and Microsoft’s DirectCompute. While both work with most graphics hardware, DirectCompute is still limited to operating systems supporting DirectX 11. In contrast the OpenCL language framework, which is managed by the Khronos Group, can be used in a cross-platform manner, regarding the used hardware (not only different GPUs, but even on CPUs and other processing units) as well as the used operating system, analogue to OpenGL. Therefore, we decided to use OpenCL to implement our algorithms in order to keep a maximum of flexibility.

4.1 Methods

Since all programming languages utilizing GPUs are lacking of an object-oriented modeling paradigm, the implemented kNN algorithms needed to be simple to keep the realization costs low. Considering the given GPGPU constraints, parallelizable processing and independent data respectively results, we evaluated two kNN algorithms.

First, we implemented the most simple brute-force kNN algorithm (BF kNN) following (Garcia, Debreuve and Barlaud 2008). BF kNN is by nature highly-parallelizable which makes it suitable for GPU computations. The second evaluated kNN algorithm is based on partitioning the point cloud in a pre-processing step. During the actual search only points of neighbored partitions are considered. This algorithm will be referred to as partitioned kNN (P kNN) search.

4.1.1 Brute-Force kNN Search

The brute-force kNN search calculates the distances between all points to determine the k nearest neighbors. This results in a quadratic runtime $O(n^2)$ with n being the number of points in the point cloud.

Since every distance between arbitrary points can be calculated independently, all distances could be calculated within a single step in parallel, assuming the corresponding number of computation units is present. This characteristic makes the BF kNN search well suitable for GPU computations.

4.1.2 Partitioned kNN Search

In contrast to BF kNN the partitioned kNN search needs a pre-processing step, which divides the space into partitions of equal space (and/or other constrains like equal number of contained points). In our case we just divided the space into equal sized partitions. This enables a linear runtime complexity $O(n)$ for the pre-processing step. The partition indices i_x for each point are calculated as follows:

$$i_x = \lfloor p_x * s \rfloor$$

where x is the corresponding dimension (in case of 3D: $x \in \{1,2,3\}$) and s is the number of partitions the particular dimension is divided in.

After creating the partitions the k nearest neighbors are determined by calculating the distances between all points in the same, as well as in the 26 neighbored partitions (partitions at the border of course have less). Assuming each of the three dimensions is divided into eight partitions ($s = 8$) the whole space is divided into $S = 8^3 = 512$ partitions. The kNN search situation for a single partition (solid red cube) for the described case is shown in Figure 3.

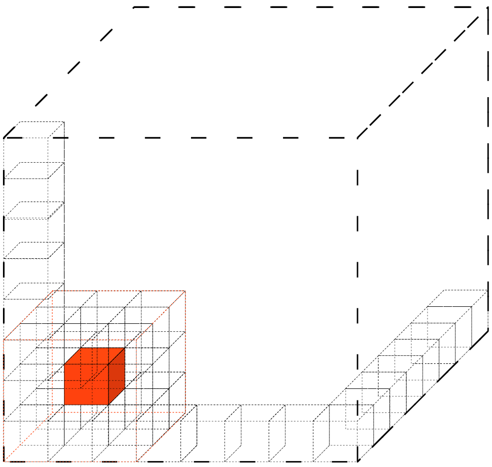


Figure 3: partitioned kNN search space

This also results in a quadratic runtime, however, with a considerably reduced number of distance calculations. Assuming the points are uniformly distributed and the space is divided into S partitions, there are $\frac{27}{S}n$ distances to calculate for each point of a partition, rather than n in case of the brute-force method.

However, the points in point clouds acquired using laser scanning technology are not uniformly distributed. The points accumulate at walls and nearby the scanning device, as seen in Figure 2. Since all partitions are processed in parallel, the resulting duration is the processing duration of the densest partition. Given that the number of points within partitions varies significantly, a more sophisticated processing scheme has high potential to improve the performance of the P kNN for such data sets.

5. TEST CASES AND ENVIRONMENT

We implemented the two kNN methods, BF kNN and P kNN, in OpenCL to run them on a GPU. In order to get comparable CPU-based results we implemented them in C++ as well.

The used test environment consists of an Intel Core 2 Duo E8400 with 3.0 GHz, 8 GB of dual channel DDR2 RAM, as well as an Nvidia Geforce GTX 285 with 240 stream processors.

While the C++ implementation is single-threaded the OpenCL implementation creates multiple calculation threads. In case of the BF kNN there are as much threads started as number of points in the point cloud. In case of the P kNN algorithm the thread count corresponds to the number of partitions.

The test scenarios included the calculation of $k = 10$ nearest neighbors for point clouds consisting of up to 60’000 points. Point clouds containing more points caused the OpenCL implementation to crash, so we assume the OpenCL/CUDA scheduler is unable to handle more threads. In case of P kNN the space was divided into 512 equal sized partitions. Overall, this test case was processed using BF kNN and P kNN running on the mentioned CPU, as well as GPU. In each case we ran three repetitions with different point sets.

6. RESULTS

As stated before we tested both algorithms with GPU-based, as well as with CPU-based implementations. While the x-axis shows the number of points in the point cloud, the y-axis shows the needed calculation time in milliseconds. The BF kNN method results are indicated with a red plus (+), in contrast the P kNN method results are indicated using a green x (x).

Figure 4 shows the CPU-based results. The computation time needed by the BF kNN method shows the expected quadratic complexity (i.e. processing 20’000 points needed approx. 10 seconds, while processing 40’000 points needed approx. 40 seconds). By contrast the P kNN is considerable faster, needing less than five seconds calculating the 10 nearest neighbors of 60’000 points, while the brute-force search needed about 90 seconds.

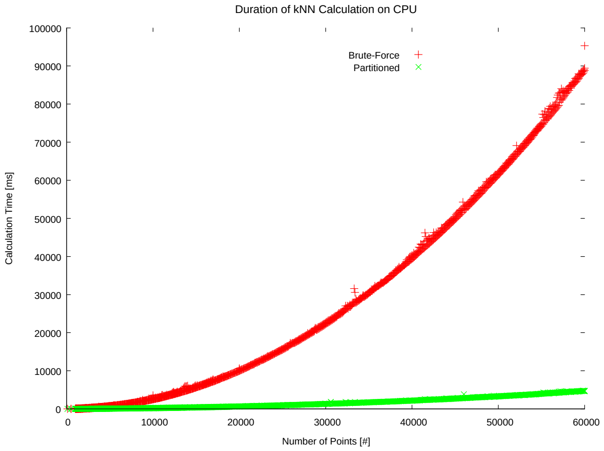


Figure 4: duration of kNN calculation using CPU-based implementation

The results of the GPU-based implementation are shown in Figure 5. Just as the CPU-based methods GPU-based alternatives have a quadric time complexity as well. Nonetheless both are significantly faster due to parallel execution (factor 45 in case of BF kNN and factor 3 in case of P kNN). In contrast with the CPU versions the different methods doesn’t show a real difference in time consumption, while the P kNN method is slightly faster.

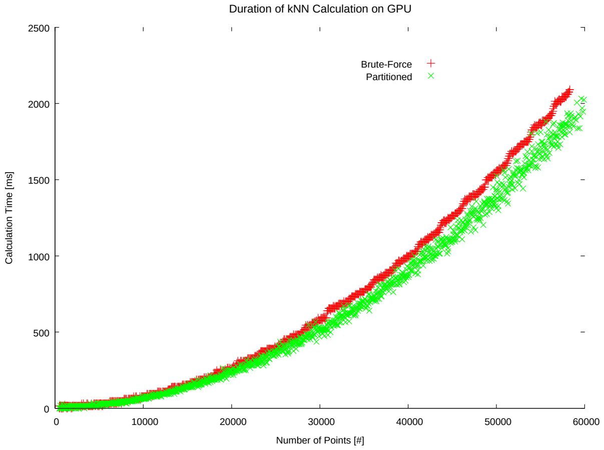


Figure 5: duration of kNN calculation using GPU-based implementation

The minimal difference is somehow unexpected, because the method using partitions calculates just a fraction of the distances the brute-force one does. The reason for this, might caused by the way the points, respectively the memory is accessed. While each GPU thread running the brute-force method always fetches the same point, memory address respectively, the partitioned search doesn’t. The former is called coalesced memory access, which appears to be more efficient than the latter un-coalesced one. This will be topic of further research. Furthermore a growing variance of the P kNN method result is evident. This might be caused by the different distribution of the points in the point cloud.

Once the kNN for each point has been calculated, the density in the corresponding point can be determined. After applying adequate colors to the resulting density space, as mentioned in section 4, a live visualization of the density quality of the recorded point set is possible as shown in Figure 2. This enables an assessment by the operator. Based on the results the operator can take appropriate subsequent actions.

7. CONCLUSION AND OUTLOOK

In this paper, we proposed a new scheme for generation of geospatial 3D models that is driven by relevance. The presented workflow, which includes on-site data evaluation and presentation, demands fast data processing. We are facing these demands by employing GPGPU to effectively process and analyze large volumes of LIDAR data.

In particular we presented a density calculation based on k-nearest-neighbor determination using OpenCL. The evaluated implementations using OpenCL accelerated the kNN search by up to a factor of 45 compared to the brute-force algorithm CPU-implementation. The P kNN algorithm acceleration reached a factor of up to 3. In summary the GPGPU analysis suites well the demands of the on-site environment and enables a much faster data analysis. Furthermore the GPGPU analysis concept isn’t just limited to field environment as it can be used on standard PC hardware as well.

As discussed the P kNN OpenCL-based implementation leaves room for improvements, for instance in accessing the same points/memory in each GPU computation thread (coalesced memory access). Furthermore the algorithm needs to be adapted to efficiently process non-uniformly distributed point sets produced by using laser scanner. Both goals could be reached by changing the way the partitions are processed, e.g. processing all partitions in a serial manner and processing the points within a single partition in parallel. As mentioned this will be discussed in further research.

In addition to the improvement of the current implementation our further research focuses on an appropriate AR setup, as well as on suitable integration algorithms.

8. ACKNOWLEDGEMENT

This joint research project was financially supported by the State of Lower-Saxony and the Volkswagen Foundation, Hannover, Germany.

9. REFERENCES

Arya, S., D. M. Mount, N. S. Netanyahu, R. Silverman, und A. Y. Wu. „An Optimal Algorithm for Approximate Nearest Neighbor Searching.“ *Journal of the ACM, 45*, 1998: 891-923.

Azuma, R. "A Survey of Augmented Reality." *Teleoperators and Virtual Environments, Vol. 6, No. 4.* 1997.

Azuma, R., Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. "Recent Advances in Augmented Reality." *IEEE Computer Graphics and Applications, Vol. 21, No. 6.* 2001.

Barnea, S., and S. Filin. "Keypoint Based Autonomous Registration of Terrestrial Laser Point Clouds." *ISPRS journal of Photogrammetry and Remote Sensing*, 2008.

—. "Registration of terrestrial laser scans via image based features." *International Archives of Photogrammetry and Remote Sensing. 36(3/W52).* 2007. 26-31.

Barnea, S., S. Filin, and V. Alchanaties. "A supervised approach for object extraction from terrestrial laser point clouds demonstrated on trees." *International Archives of Photogrammetry and Remote Sensing. 36(3/W49A).* 2007. 135-140.

Becker, S., and N. Haala. "Combined Feature Extraction for Façade Reconstruction." *International Archives of Photogrammetry and Remote Sensing. 36(3/W52).* 2007. 44-49.

Cartwright, W. "Towards an understanding of the importance of landmarks to assist perceptions of a space in web-delivered 3D worlds." *3rd Symposium on LBS & TeleCartography.* Vienna, Austria, 2005.

Connor, M., and P. Kumar. "Parallel Construction of k-Nearest Neighbor Graphs for Point Clouds." *Eurographics Symposium on Point-Based Graphics.* Los Angeles, CA, USA, 2008.

Cyclone. *Leica Cyclone.* 2008. http://www.leica-geosystems.com/corporate/de/ndef/lgs_6515.htm (accessed 8 1, 2008).

Elias, B., V. Paelke, and S. Kuhnt. "Concepts for the Cartographic Visualization of Landmarks." *Proc. 3rd Symposium on LBS & TeleCartography.* Vienna, Austria, 2005.

Filin, S. "Surface classification from airborne laser scanning data." *Computers & Geoscience 30(9-10)*, 2004: 1033-1041.

Filin, S., Avni, and A. Y. Baruch. "Quantification of Environmental Change in Receding Lake Environments." *Proceedings of FIG working week 2007 and GSDI-8.* Hong-Kong, 2007. 1-6.

Filin, S., N. Abo-Akel, and Y. Doytsher. "Detection and reconstruction of free form surfaces from airborne laser scanning data." *International Archives of Photogrammetry and Remote Sensing. 36(3/W52).* 2007. 119-124.

Garcia, V., E. Debreuve, and M. Barlaud. "Fast k nearest neighbor search using GPU." *CVPR Workshop on Computer Vision on GPU.* Anchorage, Alaska, USA, 2008.

Gorte, B. "Planar Feature Extraction in Terrestrial Laser Scans Using Gradient Based Range Image Segmentation." *International Archives of Photogrammetry and Remote Sensing. 36(3/W52).* 2007. 173-182.

Hedley, N., M. Billinghurst, L. Postner, R. May, and H. Kato. "Explorations in the use of Augmented Reality for Geographic Visualization." *Teleoperators and Virtual Environments, Vol. 11, No. 2.* 2002. 119-133.

InnovMetric. *InnovMetric Polyworks.* 2008. http://www.innovmetric.com/Manufacturing/home.aspx (accessed 1 8, 2008).

Milgram, P., H. Takemura, Utsumi A., and F. Kishino. "Augmented Reality: A class of displays on the reality-virtuality continuum." *SPIE Vol. 2351-34, Telemanipulator and Telepresence Technologies.* 1994.

Paelke, V., and C. Brenner. "Development of a Mixed Reality Device for Interactive On-Site Geo-visualization." *Proc. Simulation und Visualisierung.* Magdeburg, 2007.

Piekarski, W., and B. H. Thomas. "Tinmith-Metro: New Outdoor Techniques for Creating City Models with an Augmented Reality Wearable Computer." *5th Int'l Symposium on Wearable Computers.* Zurich, Switzerland, 2001. 31-38.

Rottensteiner, F., and C. Briese. "A new method for building extraction in urban areas from high-resolution LiDAR data." *International Archives of Photogrammetry and Remote Sensing 34(3A).* 2002. 295-301.

Schmalstieg, D., et al. "The Studierstube Augmented Reality Project." *Teleoperators and Virtual Environments, Vol. 11, No. 1.* 2002.

Volsseman, G., and S. Dijkman. "3D building model reconstruction from point clouds and ground plans." In *International Archives of Photogrammetry and Remote Sensing. 34(3W4)*, 37-43. 2001.

Zeibak, R., and S. Filin. "Change detection via terrestrial laser scanning." *International Archives of Photogrammetry and Remote Sensing. 36(3/W52).* 2007. 430-435.

FACIAL EXPRESSION RECOGNITION BASED ON CLOUD MODEL

Hehua Chi [a], Lianhua Chi [b] *, Meng Fang [a], Juebo Wu [c]

[a] International School of Software, Wuhan University, Wuhan 430079, China - hehua556@163.com
[b] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China - lianhua_chi@163.com
[c] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China - wujuebo@gmail.com

KEY WORDS: Cloud Model, Facial Expression Recognition, Backward Cloud Generator

ABSTRACT:

Facial expression is one of the major features of facial recognition in recent years, and it has become a hotspot. In this paper, we present a novel method of facial recognition based on cloud model, in combination with the traditional Facial expression system. Firstly, we carry out the transformation from images into grids with M by N, where M and N denote the actual image positioning of the grid. Each grid is a gray value (0-255) and the grids stand for the data from data points to data sets based on cloud model. Secondly, we do data pre-processing for the original facial expressions of input images. Cloud droplets image can be obtained as the input of backward cloud generator in order to extract the three numerical characteristics, that is, Ex, En and He. With these three characteristics, facial expression can be realized. Finally, in order to demonstrate the feasibility of the presented method, we conduct a case study of facial expression recognition based on cloud model. The results show that the method is feasible and effective in facial expression recognition.

1. INTRODUCTION

Expression is a basic way to express mankind's feelings and is one kind of effective communication. The facial expressions have corresponding change before people expressing their emotions. The facial expression can not only express their thoughts and feelings accurately and subtly, but also describe the others' cognitive attitudes and inner world. Facial expression contains rich human behaviours and is a kind of information resources in human-computer interaction with more effective, natural and direct way. If computers and robots have the ability to understand and express feelings as men's adapting to the environment, it will change the relationship between the computer and human fundamentally. If that, the computer can better service to mankind. It is the research meaning of this topic on facial recognition with emotion understanding and emotion expression [1-3]. Therefore, the facial expression recognition can be achieved through the observation and analysis of face images. The facial recognition is a identification task with a non-contact way, which is so vital to realize the interaction between nature and man-machine [4]. Facial expression recognition is a hot research topic in computer vision, emotion and image processing, which can be widely applied in human-computer interaction, multi-media, security, medical assistance, and behavioural science, etc.

Many scholars have launched a lot of studies on facial expression recognition and the main research results are as follows: M. Pantic et al. presented a method of emotional expression classification based on an expert system [5]. Y. L. Tian et al. introduced recognizing action units for facial expression analysis based on the behaviour identifying [6]. X. X. Yuan et al. gave a way for face recognition based on the wavelet analysis and support vector machine [7].

In this paper, we proposed a novel approach of facial expression recognition based on cloud model, aiming to mine the hidden knowledge of facial expression and the facial features with cloud model.

2. BASIC PRINCIPLE

2.1 Cloud model

Definition: Suppose U is a quantitative universe of discourse with precise numerical value: X $\subseteq$U and T is qualitative concept of space U. If the certainty of x (x $\in$X) belonging with T is a random number with stable tendency, that is, $C_T(x)$ $\in$[0, 1], then the distribution of concept T from U mapping to [0, 1] in data space is called cloud [8], where meets:

$$C_T(x):U \rightarrow [0,1] \quad \forall x \in X\ (X\subseteq U)\ x \rightarrow C_T(x)$$

Cloud model has three characteristics:

Expectation (Ex) is the prototype value (centre or standard value) of concept, and is the most representative value of the qualitative concept. Entropy (En) is the measurement of concept uncertainty while Hyper-entropy (He) is the measurement of entropy uncertainty, that is, the entropy of entropy.

Cloud model has the characteristics with macro accurate, micro fuzzy, macro controllable and micro uncontrollable. Its essential unit is concept cloud composed of cloud droplets, including randomness and fuzziness. It is the organic synthesis of fuzziness and randomness in nature language, and contains the mapping between quantitative and qualitative data. The theory is a breakthrough for limitations of hard computation in

* Corresponding author. lianhua_chi@163.com.

probability and statistics, but also solves the inherent defect of membership functions. It is a new method and new technology for solving problems in data mining, and breaks the limitation of boundary sets. As a general mathematics theory, the cloud model cleverly realizes the analysis between qualitative and quantitative data. With the mathematical conversion method and technology development, it has been widely and successfully applied in the knowledge discovery, the spatial data mining system, intelligent control, efficiency evaluation, solution or explain natural, social problems or phenomenon, and have achieved remarkable results.

2.2 Backward cloud generator

Backward cloud generator is the model of uncertainty transformation between numerical value and language value mapping from the qualitative to quantitative data [9-10].

It turns a certain amount of accurate data to corresponding qualitative value {Ex, En, He} effectively, and reflects the whole cloud droplets according to these accurate data. The more the amount of cloud droplets are, the more accuracy the concepts will be. Backward cloud generator is a process of cloud generator indirectly and reversely, which regards a group of cloud droplets Drop(x_i, $C_T(x_i)$) with a certain distribution as samples, and generates the three numerical characteristics {Ex, En, He} corresponding to the concepts (shown as figure 1). Through the forward and the backward generator, the cloud model makes the establishment between the qualitative and quantitative relationship.

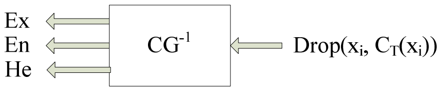


Figure 1. The input and output of backward cloud generator

Input: The sample points and their certainty degrees $C_T(x_i)$ (i=1, 2..., N)

Output: The numerical characteristics of qualitative concept, Ex, En and He.

The details are as follows:

(1) According the sample x_i, calculate the sample mean: $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$, the first-order absolute centre distance: $M_1=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\overline{X}\right|$, and the variance $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2$.

(2) Compute Ex, $Ex=\overline{X}$.

(3)Compute En, $En=\sqrt{\frac{\pi}{2}}\times M_1$.

(4)Compute He, $He=\sqrt{S^2-En^2}$

3. FACIAL EXPRESSION RECOGNITION

As a challenging cross-subject between the biological feature recognition and the affective computing field, the facial expression recognition technology driven by a variety of applications has the rapid development [11]. The facial expression recognition system mainly includes the following steps: the acquisition of the facial expression images, the face detection, the facial expression feature extraction and the facial expression recognition. Its structure is shown in Figure 2. For an automated facial expression recognition system, first, we should obtain static images or facial image sequences; The second step is the facial image pre-processing, including the face detection and the image normalization ; The third step is the facial image feature extraction, including the original feature acquisition, the feature dimensionality and extraction, the feature separation; The fourth step is the facial expression recognition, that is, according to the extracted features and some criteria, we realize the classification.

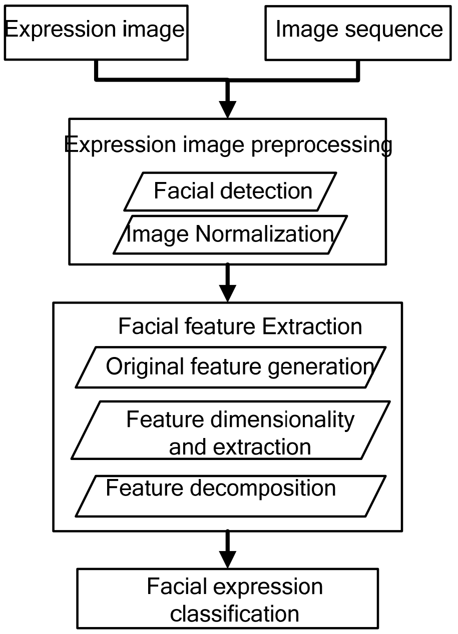

Expression image
Image sequence
Expression image preprocessing
Facial detection
Image Normalization
Facial feature Extraction
Original feature generation
Feature dimensionality and extraction
Feature decomposition
Facial expression classification

Figure 2: The model of facial expression recognition

3.1 Facial image pre-processing

In the process of pre-treatment, the facial detection and localization are applied firstly, namely, to find the position and existing face to face segmentation from the background from the input. Then the facial images are done data normalization, such as gray normalization, etc.

3.2 Face detection and localization

Face detection and localization is the primary problem to solve in the automatic identification system of facial expression, including the face detection in simple background and complex background. At the beginning of the study, the face images in database are all simple background, which means the difference between face and background is big and most of them are positive images. The research value of the latter is more practical and theoretical. In detecting and locating face image, the image must be normalized in order to facilitate subsequent processing.

3.3 Image normalization

Image normalization includes geometry normalization and gray normalization. The former means converting the face results to the same position and size. The latter is to stretch the image gray and improve the image contrast. It also includes light compensation and to overcome of the light changes. After that, we can get more suitable images in the process by image pre-processing and edge extraction.

3.4 Facial Feature Extraction

After the pre-processing, we can make the feature extraction and selection on the facial images, that is, extracting the facial expression feature information. The purpose is to obtain a set of category features, that is to say, we obtain the feature vectors that the number of features and the classification error rate are fewer. It is a very important part. The effect will directly affect the correct recognition rate of the facial expression.

3.4.1 The original feature acquisition: We use some information to obtain the original features of the expressions, such as features of shape, geometric relations, local texture, optical flow and so on. This step is called as the original features acquisition. However, these primitive characteristics generally exists the redundancy and other issues. In order to more effectively characterize the nature of the facial expression, we need to make the process on the original feature data.

3.4.2 Feature dimensionality and extraction: Because the dimensions of the original features are usually very large, we should convert them into the low-dimensional subspace. That not only make the dimension of the original features significantly reduced, but also the validity of these low-dimensional space will be increased. In recent years, we make some new research on the methods of the feature dimensionality and extraction.

3.4.3 Feature separation: Facial images contain a wealth of information. For different recognition tasks, the information also varies. The facial detection is to find the consistency of the facial images. The facial recognition need to present the individual differences among the facial expressions. Recently, a new solution is to separate the different factors of the human facial expression, such as the expression factors and individual factors, avoiding the interference of other factors.

3.5 Facial expression classification

Expression classification refers to the definition of a group of categories, and to design appropriate mechanisms for the expression recognition. If the expressions are classified according to the facial movements (FACS), facial actions are classified into 44 AUs (action units). In accordance with the emotion classification, the expressions are classified into seven kinds of basic emotions (crying, surprise, happiness, ecstasy, unwilling, frustration, fear).

4. FACIAL RECOGNITION BASED ON CLOUD MODEL

4.1 The process of facial recognition based on cloud model

Cloud theory portrays the distance relationship between each element in the domain and its core concept using the membership. The greater degree of the membership, the elements are much closer to the core concept. This feature is the same as the facial expression recognition's feature that we obtain the facial expression feature and make the classification. So we can use the cloud theory to obtain the facial expression feature, using the numerical characteristics of the cloud to express the facial expression features. Making use of the cloud model algorithm to extract facial expression features, we propose a new facial expression recognition method - facial expression recognition based on cloud model.

First, input a group of primitive facial images; Second, pre-process the input images to get a set of standard cloud droplet images; Third, make use of backward cloud generator to realize the image feature extraction and output the numerical characteristics (Ex, En, He) of this cloud droplet images; Fourth, make the numerical characteristics (Ex, En, He) as the facial expression features; At last, use the numerical characteristics (Ex, En, He) to realize the facial expression classification. The chart of facial expression recognition based on Cloud Model is shown in Figure 3.

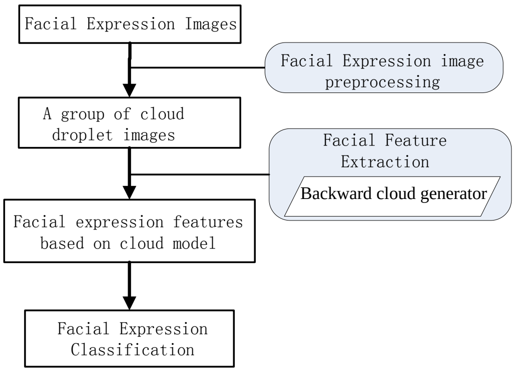

Facial Expression Images
Facial Expression image preprocessing
A group of cloud droplet images
Facial Feature Extraction
Backward cloud generator
Facial expression features based on cloud model
Facial Expression Classification

Figure 3: The structure of facial recognition based on cloud model

4.2 A group of images of cloud droplets

Macro accuracy and micro fuzziness are the features of cloud model, with macro controllable and micro uncontrollable. Its essential unit is cloud droplets, which can form cloud with such cloud droplets and realize the transformation between qualitative and quantitative data. It reflects the uncertainty of knowledge representation.

After image pre-processing，a group of cloud droplets images can be obtained for the original face expressions. Such images are considered as the standard input images for the following processing.

4.3 The characteristics of facial expression based on cloud model

The numerical characteristics of cloud reflects quantitative feature of qualitative concept with Ex, En and He. It is the numerical basis for describing cloud model and mining knowledge from uncertainty data. The facial expression is a kind of uncertainty data.

This paper uses cloud generator to find knowledge from facial expressions, that is {Ex, En, He}, and to achieve facial expression recognition upon such characteristics.

5. THE EXPERIMENT OF FACIAL EXPRESSION RECOGNITION BASED ON CLOUD MODEL

5.1 The experiment of facial recognition

The data source comes from Japanese Female Facial Expression (JAFFE) database, which is an open face image database (http://www.kasrl.org/jaffe_download.html). It contains 10 women's expressions, including the people KA, KL, KM, KR, MK, NA, NM, TM, UY and YM. Each person has 7 different expression as AN, DI, FE, HA, NE, SA and SU. Each expression has 3 or 4 samples and the total number is 216. In this experiment, we conduct knowledge mining by cloud model for facial expression images, aiming to find the numerical characteristics and realize the facial expression recognition.

5.2 Sample sets training

5.2.1 The experiment of different expressions for one person: In JAFFE database, the attribute KA is selected at the beginning, while the original images are chosen from the ten Japanese women's KAs, including AN, DI, FE, HA, NE, SA and SU. By backward cloud generator, the KAs of the same expression from different Japanese women can transform to the three numerical characteristics {Ex, En, He} of cloud, as shown in line 1, table 1.

By using the same method as KA processing, the corresponding numerical characteristics can be obtained for KL, KM, KR, MK, NA, NM, TM, UY, YM and the results are shown in table 1 from column 2 to column 10.

5.2.2 The same expressions of different people: In JAFFE database, the attribute AN is selected at the beginning, while the original images are chosen from the ten Japanese women's ANs, including KA, KL, KM, KR, MK, NA, NM, TM, UY and YM. By backward cloud generator, the ANs of the same expression from different Japanese women can transform to the three numerical characteristics {Ex, En, He} of cloud, as shown in column 1, table 1.

By using the same method as AN processing, the corresponding numerical characteristics can be obtained for DI, FE, HA, NE, SA, SU and the results are shown in table 1 from column 2 to column 7.

5.3 Sample sets training

(1) Every line in table 1 means the input is the different expressions of the same person, and the output is the numerical characteristics of cloud of such input images {Ex, En, He}.

(2) Each column in table 1 means the input is the same facial expression of ten persons, and the output is the numerical characteristics of cloud of such input images {Ex, En, He}.
(3) The problem is that how to identify the facial image belong to whose expression if existing a face image for recognition?

The method is as follows: Firstly, generate the numerical characteristics {Ex, En, He} of cloud for the original face expressions. Secondly, compute the numerical characteristics {Ex, En, He} of cloud by backward cloud generator by adding the image to be identified to such original face expressions. Finally, compare two groups of numerical characteristics {Ex, En, He} of cloud to find the differences. According to such differences, we can get the image category and achieve facial expression recognition.

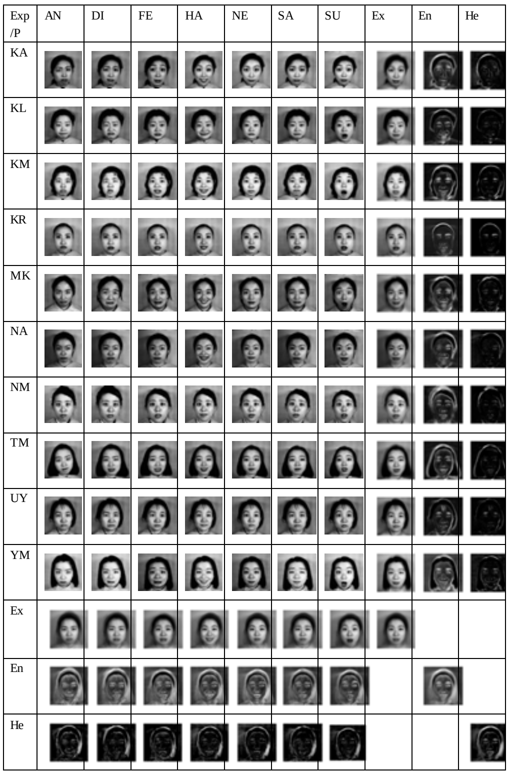


Table 1: The training samples of facial expression

5.4 Facial recognition based on the samples

After the training of samples, we choose two groups as the original facial image for person, as shown in table 2.

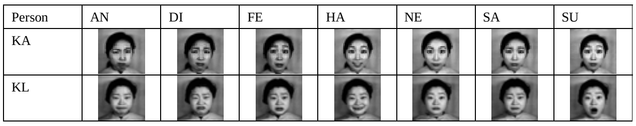
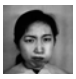


Table 2 two groups of original facial expression images

The face image for identification 

The experimental steps are:
Step 1: Choose the first line in table 2 as the original image with backward cloud generator and calculate the {Ex, En, He} of the image. The results are shown in line 1 in table 1.
Step 2: Add the facial expression images into line 1 in table 2 for identification. By backward cloud generator, compute the {Ex, En, He} of the image as shown in line 1 in table 3.

Step 3: Select line 2 in table 2 as the original facial expression image. Based on backward cloud generator, generate the {Ex, En, He} of image as shown in line 2 in table 1.

Step 4: Add the facial expression images into line 2 in table 2 for identification. By backward cloud generator, compute the {Ex, En, He} of the image as shown in line 2 in table 3.

The line 3 in table 3 shows the difference value of {Ex, En, He} for images in the second to first steps while the line 4 in table 3 gives the fourth to third steps.

Characteristic	The second	The fourth	Second-First	Fourth-Third
Ex				
En				
He				

Table 3 The results and comparative results

Observing from line 3 and 4 in table 3, it can be obviously seen that the {Ex, En, He} of the former different image is not clearer than the latter's. We can know that the facial expression of the image for identification is more close to A, which is correct. Therefore, it is feasible of that cloud model can achieve face facial image recognition. The research develops the cognition of cloud model theory and further expands the application fields of cloud model.

6. CONCLUSIONS

As a mathematical transformation model with knowledge uncertainty, the cloud model integrates the fuzziness with the randomness and forms the qualitative and the quantitative mapping between them. This paper put forward a new method of facial expression recognition based on cloud model. By using cloud model, the facial expression recognition can be carried out effectively, and it expressed the uncertainty of facial expression. The quantitative numerical characteristics {Ex, En, He} of facial expressions were mined by the backward generator of cloud model. In this paper, the hidden knowledge in facial expression images were obtained with the numerical characteristics {Ex, En, He} of cloud model. Ex is the characteristics of the facial image in common, En is the personality deviation of general common knowledge, and He is the discrete level of knowledge. In analyses of facial image knowledge, by the numerical characteristics {Ex, En, He}, the facial expression can be realized. The experimental results showed that this method can effectively achieve facial recognition. Furthermore, the facial expression recognition and its application based on cloud model should be further study in next step.

7. ACKNOWLEDGEMENTS

This paper is supported by National 973 (2006CB701305, 2007CB310804), National Natural Science Fund of China (60743001), Best National Thesis Fund (2005047), and National New Century Excellent Talent Fund (NCET-06-0618).

8. REFERENCES

[1] N. Zhang. Summarization for the facial expression recognition. Journal of Shandong Institute of Light Industry(Natural Science Edition), 21(4), 2007.

[2] X. M. Liu, H. C. Tan, Y. J. Zhang. New Research Advances in Facial Expression Recognition. Journal of Image and Graphics, 11(10), 2006.

[3] K. MaSe, A. Pentland. Recognition of Facial Expression From Optical flow [J]. IEICE Trans, 74(10), 1991, pp.3474-3484.

[4] M. Pantic, L. J. M. Rothkrantz, Automatic Analysis of Facial Expressions:the State of the Art. IEEE trans. Pattern Analysis and Machine Intelligence, 22(12), 2000, pp.1424-1445.

[5] M. Pantic, L. J. M. Rothkrantz. An Expert System for Multiple Emotional Classification of Facial Expressions. Pro.11th IEEE Int. Conf.on Tools with Artificial Intelligence, 1999, pp.113-120.

[6]Y. L. Tian, T. Kanade, J. f. Cohn. Recognizing Action Units for Facial Expression Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, 23(2), 2001, pp.97-115.

[7] X. X. Yuan, W. Jiang, L. Zhang. Facial expression recognition method based on wavelet energy feature and Support Vector Machines. Optical Technology, 34(2), 2008.

[8] K. C. Di, D. Y. Li, D. R. Li. Cloud Theory and Its Applications in Spatial Data Mining and Knowledge Discovery. Journal of Image and Graphics, 4(11), 1999.

[9] H. J. Lv, Y. Wang, D. Y. Li. The Application of Backward Cloud in Qualitative Evaluation. Chinese Journal of Computers, 26(8), 2003.

[10] D. R. Li, S. L. Wang, D. Y. Li. Spatial Data Mining Theories and Applications. Science Press, 2006.

[11] M. Qiao, Y. J. Chen. Feature Extraction Methods on Facial Expression Recognition. Journal of Chongqing Institute of Technology, 22(6), 2008.

GLACIER INFORMATION EXTRACTION BASED ON MULTI-FEATURE COMBINATION MODEL

J.M. Gong [a, *], X.M．Yang [a], T. Zhang [a], X. Xu [a], Y.W. He [a]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, CAS, 100101 Beijing, China - (gongjm, yangxm, zhangt, xux, heyw)@lreis.ac.cn

Commission VI, WG VI/4

KEY WORDS: Multi-feature Combination Model; Image Segmentation; Remote Sensing; Glacier Landform; Information Extraction; Feature Description

ABSTRACT:

As a typical landform class of Qinghai-Tibetan Plateau, glacier is widely distributed in alpine terrain. However, field measurement is impossible in those areas because of complex terrain and adverse weather. At first, on the basis of analyzing the features of glacier image spectrum, object shape, spatial relations and environment distribution including terrain and climate, this paper combines and develops the existing feature description algorithm of object-oriented method. Secondly, we build a series of combined extraction models for glacier landform by using high resolution remote sensing images and DEM data. At last, based on object-oriented method and combined extraction models, this paper tests glacier landform extraction in Qinghai-Tibetan Plateau study area of Western Mapping Project. Results demonstrate that the multi-feature combination model is feasible. The researches introduce a new approach to remote sensing auto-extraction of glacier information which is difficult to measure in the field. Moreover, the paper explores some new ideas in the researches of monitoring glacier ablation and climatic change.

1. INTRODUCTION

In recent years, the fact that the world's glacier is accelerating ablation has caused great concerns of native and international scholars (Aizen, 2007; Noferini, 2009; Scherler, 2008; Wolken, 2006; Yao, 2006). With an area of more than 2.5 million km^2, an average elevation of 4500m above, Qinghai-Tibet Plateau is located in Eurasia, which has unique alpine climatic characteristics of Qinghai-Tibet Plateau. As a typical landform landscape of the plateau, glacier is widely distributed in the alpine region. According to statistics of glacier catalogue in 2004, the glaciers area of Qinghai-Tibet Plateau covers about 47000 km^2, accounting for more than 80% total glacier area of China. Permafrost area is about 1.5 million km^2, accounting for more than 60% total area of Qinghai-Tibet Plateau (Pu, 2004). Glacier ablation of Qinghai-Tibet Plateau has great research value to the worldwide climate change. Therefore, the monitoring of glacier change is an important topic of current global change research. In the process of glacier ablation, retreat and thinning of glacier result in various types of glacial landforms. Based on remote sensing techniques, this paper explores to find a quick automatic extraction method of glacier information, which has great research significance to monitoring glacier ablation and global climate change.

Solely based on a single gray level or spectral information, traditional remote sensing image analysis methods often focus on gray-level statistical characteristics of image and calculate its variance, mean and other statistical parameters to achieve the purpose of image analysis. However, because the interrelated information of spatial characteristics contained in image is ignored, those image analysis methods often limit the accuracy of information extraction, even make wrong judgments since the rich information of shape, texture and context is treated as noise, resulting in the phenomenon of misjudgment and misclassification in image interpretation process (Chen, 2006; Tan, 2007). Compared to the traditional methods of image analysis, object-oriented approach, which primarily produces a certain criterion of polygon objects composed of homogeneous pixel cluster by image segmentation, is used in glacier information extraction of multi-feature combination. Further, we can extract varieties of landform classes based on the analysis of object features, including spectrum, shape, texture and spatial relations. This paper attempts to find a suitable combination model to describe glacier features and achieve the purpose of glacier information automatic extraction, which mainly includes the following steps: image segmentation, feature description and glacier information extraction of multi-feature combination model.

2. GLACIER IMAGE SEGMENTATION ALGORITHM

Image segmentation is a critical step of information extraction based on object-oriented method, in which its segmentation quality has a direct impact on image analysis accuracy. Image segmentation is a process that image is expressed as a number of region set , which fulfils the homogeneity standard including spectrum, shape and other features description while meeting the heterogeneity standards among the adjacent regions (Definiens, 2007). Starting from the pixel, the smaller homogeneous objects gradually merged into a large homogeneous image object by using region merging approach of bottom-up. Usually, in accordance with the different research purpose, image segmentation approaches are generally classified as three categories based on edge, region, and the mixture of the previous two.

* Corresponding author: Jianming Gong, E-mail: gongjm@lreis.ac.cn.

Since differences of target size and spatial structure still exist in high resolution remote sensing images, it is difficult to reflect the rich object and spatial semantic information only in a single spatial scale. Therefore, we need different scales to express and describe different size target. The structure of image object obtained in different segmentation scales represents different scales image object information, in which a smaller object is a child object of a larger object (Tan, 2007). Since each object has interrelated spatial feature information with adjacent objects, child object and parent object, the purpose of our experiment is to explore the spatial feature of interrelated objects and add it to the multi-feature combination model.

3. FEATURE DESCRIPTION

The purpose of information extraction is to distinguish interested regions in remote sensing image. There are different methods corresponding to different targets. Especially, some specific thematic objectives need to fully consider the characteristics of data (Datcu, 2002; Anthony, 1997; Zhou, 1999). Features description is object-oriented expression for latent knowledge in the primitive obtained from image segmentation. In addition to visual features such as spectrum, feature description also includes the object shape, spatial relationship and terrain features (Yang, 2009).

3.1 Spectrum features

In addition to traditional statistical value of image gray such as histogram, variance, mean, the normalized difference snow index is suitable for extracting glacier information since it is very sensitive to the changes of water content in snow and ice. In the following formula, $\rho_{\mathrm{Re}\,d}$ is the red band reflectance and ρ_{Green} is the green band reflectance (Guo, 2003).

$$NDSI = \frac{\rho_{\mathrm{Re}\,d} - \rho_{Green}}{\rho_{\mathrm{Re}\,d} + \rho_{Green}} \qquad (6)$$

3.2 Shape features

Generally, the boundary of ice cover is clear, around which there is often a large number of moraine. Meanwhile, some clear curved contours are left behind the ice tongue in the process of glacier retreat. In addition to the compactness and smoothness mentioned above, there are some shape indexes to describe glacier feature.

3.3 Spatial relation features

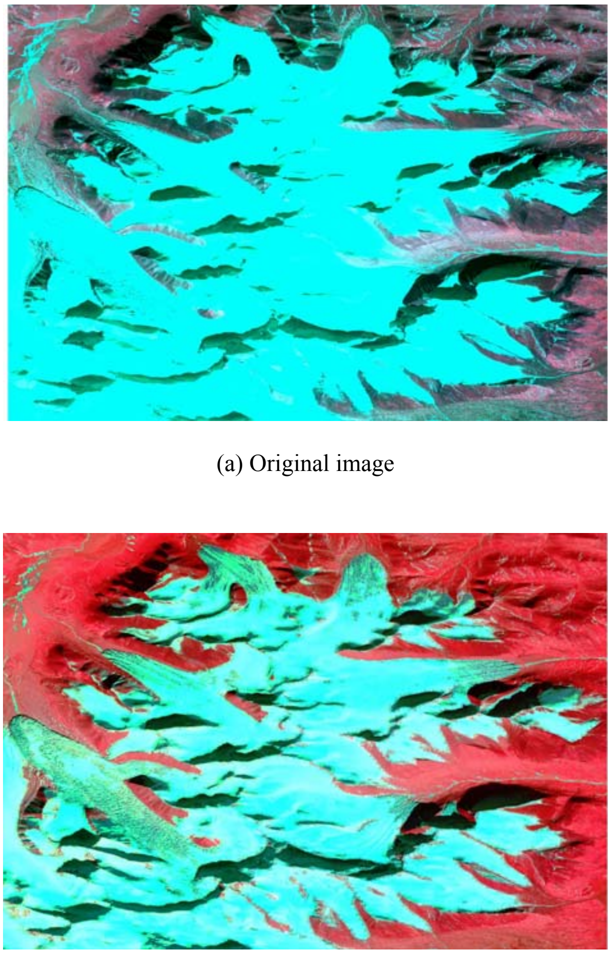
As for each object in the spatial relationship, we can calculate the mean difference among the objects, and give a weight according to border length or area size to achieve classification and clustering of object (Definiens, 2007). Some statistics can reflect the spatial distribution features of pixels enclosed by different object.

3.4 Terrain features

Glacier is distributed above the perennial snowline of the alpine region，in which valley glaciers are located in canyons of high mountain. Since the shapes of ice tongue, ice pillar and ice cliff are closely related to the terrain factor, the terrain features have more important role to the glacier information extraction. Digital elevation model (DEM) mainly describes spatial distribution of region landform, and can determine slope, aspect and relief degree of earth's surface (Song, 2007), so we can identify and extract glacier information by the DEM.

4. GLACIER INFORMATION EXTRACTION BASED ON MULTI-FEATURES COMBINATION MODEL

In our researches, the data we used are both of 2.5 m panchromatic and 10 m multi-spectral SPOT-5 images. After pretreatment, we select a sub scene fused image of eastern Qinghai-Tibet Plateau as trial data, with a rectangle area of 436 km^2. The original image is shown as Figure 1 (a), and the enhanced image by histogram is shown as Figure 1 (b). The glacier information extraction of multi-feature combination model mainly includes the following steps: (1) multi-scale image segmentation; (2) the bound identification of ice and snow; (3) glacier information extraction.



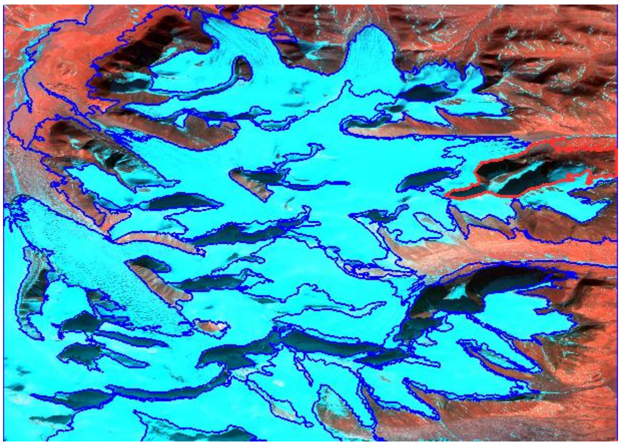
(a) Original image

(b) Enhanced image

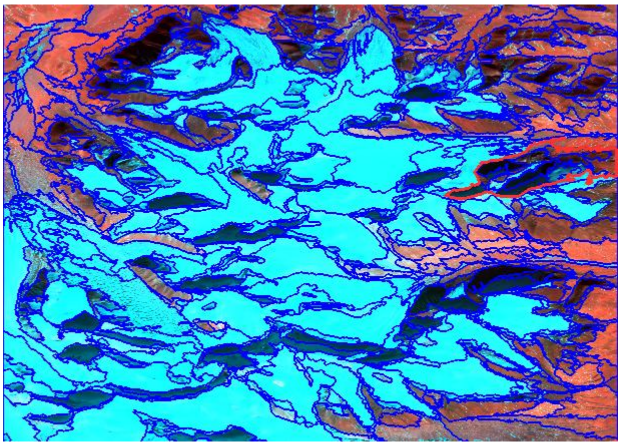
Figure.1 Image of Qinghai-Tibetan Plateau trial area

4.1 Multi-scale image segmentation

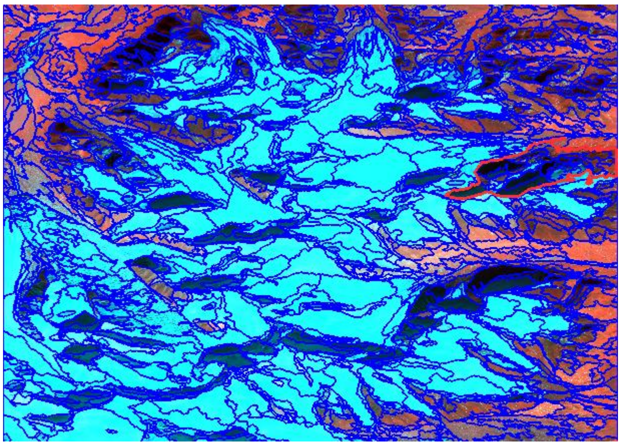
In the process of multi-scale image segmentation, we select 0.1 and 0.7 as the weight of shape and smoothness. Meanwhile, we set the segmentation scale as 500, 200,100 and 50 and get different results as Figure 2 (a) to (d), with 72, 354, 1036 and 3290 image objects, respectively. Obviously, over-segmentation phenomenon exists in the Figure 2 (c) and (d), and under-segmentation phenomenon exists in the Figure 2 (a) and (b).



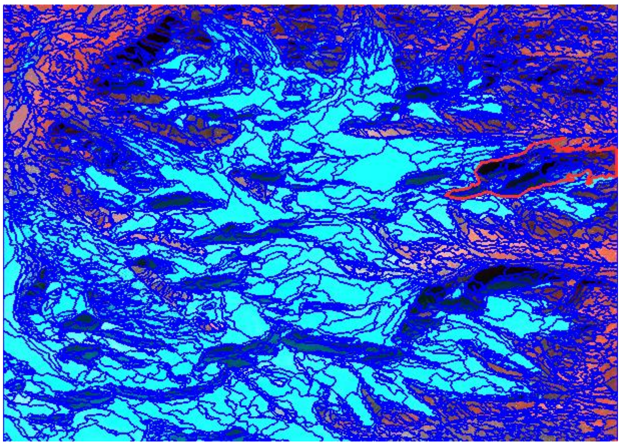
(a) segmentation scale is 500



(b) segmentation scale is 200



(c) segmentation scale is 100



(d) segmentation scale is 50

Figure.2 Image segmentation results in different scales

After repeated experiments, the image has the best segmentation results when the scale parameter is 150, the shape weight is 0.18 and the smoothness is 0.56. Under these conditions, the test image is segmented out 555 valid polygons, with medium size and even distribution.

For multiple object layer segmented in different scales, we can decide spatial membership by calculating the difference between adjacent objects and the difference between object and its supperobject or subobject. Usually, one superobject obtained in a larger segmentation scale can segment several subobjects. Because of the changes in spatial resolution, these subobjects may belong to or not belong to the class under superobject according to the subobject decision criterion. The ultimate purpose of segmentation is not segmentation, but re-clustering according to certain classification rules.

4.2 Glacier and snow boundary detection

Because glaciers and snow show a relatively strong reflectivity, it is easy to distinguish from the surrounding grassland and soil. However, it is difficult to distinguish the glaciers from snow. In our study, based on spectral analysis, we use an appropriate threshold of snow cover index to extract the common scope of glaciers and snow. In addition to snow and glaciers, the rest area is permafrost and tundra regions. Tundra is mainly distributed in the low zone of full sunlight, while most of the permafrost is distributed in the shady slope and valley areas. So we can further distinguish permafrost from tundra by using slope, aspect and spectral index analysis.

Since snows are different from glaciers and new snow gradually becomes coarse snow, the spectral reflectance of snow decreases over time and environment changes. Studies have shown that the best reflectance of new snow occurs in the wavelength range of 0.80 to 1.10μm, with reflectivity more than 80%, up to 95%, corresponding to MSS-7 band of Landsat. From the process of snow into glacier, the spectral reflectance gradually drops to around 60%. However, the best reflectance of glacier occurs in the wavelength of 3μm which is around the transition border of the mid-infrared to the thermal infrared. Basically, there is not reflection below 2.8μm. Due to the moraine and its uneven ablation, the reflectance of ice tongue gradually declines from 60% to 30% in the visible region (Song, 2007; Zeng, 1990). These features provide a new idea for extracting the boundaries of snow and glacier. Therefore, we can extract the boundary of new snow based on spectral calculation. Some regions covered by coarse snow and difficult to distinguish, can be classified as the scope of glaciers, taken into the next step to extract.

4.3 Glacier information extraction

Studies have shown that serac clusters appear in the process of glaciers retreat and thinning. Ice cliff is formed while the fracture is occurred during the glaciation. Usually, the leading edge and trailing edge of ice cliff have a certain high difference and slope. Serac clusters are mainly located in the relatively low altitude and flat areas, hence its relief degree is relatively small. On the contrary, the iceberg is mainly located in the relatively high altitude areas. Both of them have a clear change of arc-shaped contour in the image. Glacial lake is located in flat areas, while the moraine is mostly distributed in the leading edge of the ice tongue with a certain slope. In our combination extraction model, the glacial lakes and ice moraine are limited in the scope of 150 m below the snow line, mainly based on the recognized law of the temperature falls 6℃ by the altitude increase 1000 m. In theory, the temperature will rises 1℃ when the height decrease 150 m, so the possibility of emerging glacial lake and ice moraine is very small as the temperature difference. This information can be used in the combination extraction model as auxiliary decision data. Some glacier information extraction models developed in our study are described as follows:

Ice pillar: slope> 45°, 4π < roundness<15, and relief degree >0.8; Ice tongue: 0<ellipseness<100, 60° <main direction angle <120°, and relief degree<0.2; Serac clusters: 4π < roundness<15, slope> 30°, relief degree <0.2, and elevation is within 150m below the snow line; Glacier lake: slope<5°, spectral reflectance<10%, and elevation is within 150m below the snow line; Moraine: slope>15°, spectral reflectance<20%, and elevation is within 150m below the snow line.

In our researches, we design multi-feature combination model based on image spectrum, object shape, spatial relation, and terrain and climatic features of trial area. In theory, the model has a possibility that some glaciers do not belong to any of the rules. However, after the steps of border extraction of snow and glacier, and exclusion of snow scope through the mask, the remainder of the extraction area should be glacier types and very limited scope. Therefore, we can classify those objects which are attributed to glacier but not meet the combination extract model as other glacier type. Furthermore, we can add more description features to the model for further extraction research.

Based on above analysis, we extract 262 non-adjacent polygons of glacier class by using multi-feature combination model. Table 1 shows the statistical results of glacier information extraction. Figure 3 shows the glacier classification map. From the analysis of extraction results, the glacier information extraction based on multi-feature combination model is feasible.

Class	Amount	Area(km^2)	Proportion
Ice cliffs	28	26.2	6.01%
Ice pillar	37	25.6	5.87%
Glacier tongue	35	68.6	15.73%
Serac clusters	39	37.2	8.53%
Other glacier	29	56.7	13.00%
Glacial lake	3	0.8	0.18%
Moraine	32	21.6	4.95%
Permafrost	18	62.3	14.29%
Tundra	22	67.2	15.41%
Perennial snow	19	69.8	16.01%
Total	262	436.0	100.0%

Table.1 Statistical results of glacier information extraction

5. CONCLUSION

Multi-feature combination model can take full advantages of rich image spectrum, object shape, spatial relation, and terrain and climatic characteristics. But because the process from snow to glacier under the action of gravity is a long-term one, thickness and the particle size of snow bring difficulties to distinguish between snow and snow-covered glacier. Studies have shown that the model can achieve a good results as much as possible large segmentation scale in the condition of ensuring object accuracy. When we use the multi-feature combination model, the most typical features of object should first consider to be used by setting appropriate weight of shape, spectrum and spatial relation. When extracting ice tongue information, the main consideration feature is object shape; the main consideration feature is image spectral feature while extracting moraine. We need to comprehensively consider the features of the image spectrum, object shape, object spatial relation and terrain, climate auxiliary data of the region while extracting serac clusters, ice cliffs, ice pillar, and glacial lakes,. It is obvious that the accuracy of model is closely related the DEM resolution and image timeliness. Furthermore, we can add more description features and set flexibility threshold to further improve the extraction accuracy of the multi-features combination model.

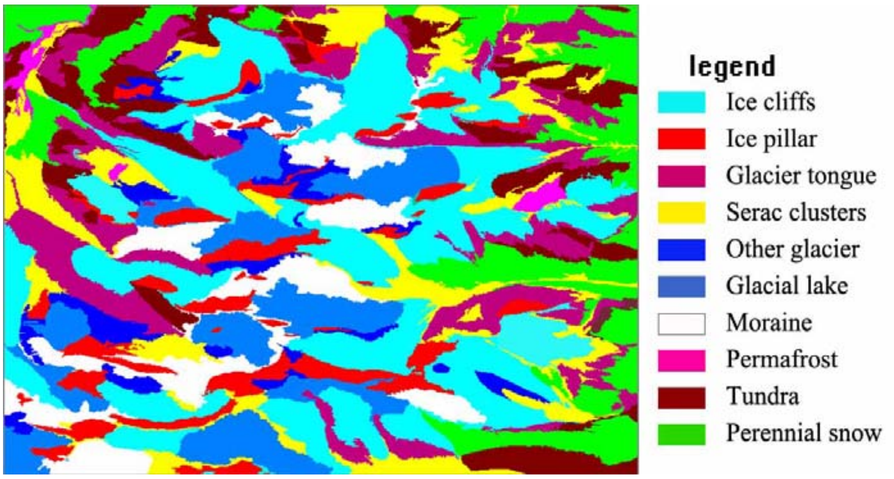


Figure.3 Results map of glacier landform class extraction

ACKNOWLEDGMENTS

This study was supported by the Western Region Mapping Project of China and the National High Technology Research and Development Program of China (No.2008AA121706 and No. 2009AA12Z148).

REFERENCE

Aizen V.B, Kuzmichenok V.A, 2007. Glacier changes in the Tien Shan as determined from topographic and remotely sensed data. *Global and Planetary Change*, 56(3), pp.328-340.

Anthony Gar On Yeh, 1997. An Integrated Remote Sensing and GIS Approach in the Monitoring and Evaluation of Rapid Urban Growth for Sustainable Development in the Pearl River Delta, China. *International Planning Studies*, 2(2), pp.193-210.

Chen Yunhao, 2006. Classification of Remot Object Oriented Sensing Image Based on and Class Rules, *Geomatics and Information Science of Wuhan University*, 31(4), PP.316-319.

Datcu M, Seidel K, D'Elia S, Marchetti PG, 2002. Knowledge-driven information mining in remote-sensing image archives. *Esa Bulletin-European Space Agency*, 110, pp.26-33.

Definiens AG, 2007. *Definiens Developer 7 User Guide*, Document Version 7.0.0.828. Definiens AG, Trappentreustr.1, D-80339 München, Germany, pp.156-183.

Feng Zhiming, Tang Yan, 2007. The Relief Degree of Land Surface in China and Its Correlation with Population Distribution. *Acta Geographica Sinica*, 62(10), pp.1073-1082.

Gong Jianya, 2006. *Basic Theory of Geographic Information System.* Science Press, Beijing,pp.254-271.

Guo Ni, 2003. Vegetation Index and Its Advances. *Arid Meteorology*, 21(4), pp.71-75.

Li Zhen, 1999. Deriving Glacier Change Information on the Xizang (Tibetan) Plateau by Integrating RS and GIS Techniques. *Acta Geographica Sinica*, 54(3), pp.263-268.

Noferini L, Mecatti D, Macaluso G, 2009. Monitoring of Belvedere Glacier using a wide angle GB-SAR interferometer. *Journal of Applied Geophysics*, 68(2), pp.289-293.

Pu Jianchen, YaoTandong, 2004. Fluctuations of the Glaciers on the Qinghai-Tibetan Plateau during the Past Century. *Journal of Glaciology and Geocryology*, 26(5), pp.517-522.

Scherler D, Leprince S and Strecker MR, 2008. Glacier-surface velocities in alpine terrain from optical satellite imagery - Accuracy improvement and quality assessment. *Remote Sensing of Environment*, 112(10), pp.3806-3819.

Song Bo, 2007. Identifying Automatically the Debris-covered Glaciers in China’s Monsoonal Temperate-Glacier Regions Based on Remote Sensing and GIS. *Journal of Glaciology and Geocryology*, 29(3), pp.56-462.

Tan Qulin, 2007. An Algorithm For Object-Oriented Multi-Scale Remote Sensing Image Segmentation. *Journal of Beijing Jiaotong University*, 31(4), pp.111-114.

Wolken G.J, 2006. High-resolution multispectral techniques for mapping former Little Ice Age terrestrial ice cover in the Canadian High Arctic. *Remote Sensing of Environment*, 101(1), pp.104-114.

Yang Xiaomei, Gong Jianming, Gao Zhenyu, 2009. The research on extracting method of microscale remote sensing information combination and application in coastal zone. *Acta Oceanologica Sinica*, 31(2), pp.40-48.

Yao Tandong, 2006. The Response of Environmental Changes on Tibetan Plateau to Global Changes and Adaptation Strategy. *Advance in Earth Sciences*, 21(5), pp.459-464.

Zeng Qunzhu, 1990. Glacier and Snow Dynamic Monitoring by Remote Sensing. *Remote Sensing Information*, 2, pp.28-29.

Zhou Chenghu, Luo Jiancheng, Yang Xiaomei,1999. *Understand and Analysis of Remote Sensing Image by Using Geoscience Knowledge*. Science Press, Beijing, pp.91-122.

EXPLORING SPATIOTEMPORALLY VARYING REGRESSED RELATIONSHIPS: THE GEOGRAPHICALLY WEIGHTED PANEL REGRESSION ANALYSIS

Danlin Yu

Department of Earth and Environmental Studies, Montclair State University, Montclair, NJ, 07043
yud@mail.montclair.edu

KEY WORDS: Geographic information; spatiotemporal variation of relationships; geographically weighted panel regression; Greater Beijing Area, China

ABSTRACT:

Regression analysis with geographic information needs to take into consideration the inherent spatial autocorrelation and heterogeneity of the data. Due to such spatial effects, it is found that local regression such as the geographically weighted regression (GWR) tends to capture the relationships better. In addition, in panel data analysis, the variable coefficient panel regression can borrow such ideas of spatial autocorrelation and heterogeneity to develop models that would fit the data better and produce more accurate results than the pooled models. Despite the fact that both methods are well developed and utilized, models that take advantage of both methods simultaneously have eluded the research community. Combination of GWR and panel data analysis techniques has an obvious benefit: the added temporal dimension enlarges the sample size hence contains more degrees of freedom, adds more variability, renders less collinearity among the variables, and gives more efficiency for estimation. This research for the first time attempts such combination using a short regional development panel data from 1995 – 2001 of the Greater Beijing Area (GBA), China. A geographically weighted panel regression (GWPR) model is developed and compared with both cross-sectional GWR and panel regression. The study reveals very promising results that the GWPR indeed produced better and clearer results than both cross-sectional GWR and the panel data model. This indicates the new method would potentially produce substantial new patterns and new findings that cannot be revealed via pure cross-sectional or time-series analysis.

1. INTRODUCTION

Geographically weighted regression (GWR) and panel data analysis are well developed data analytical methodologies in geography and econometrics. Recognizing the fundamental question in social science that social processes are not likely governed by any universal "laws", but might vary depending on **where** the processes are investigated, Fotheringham and colleagues (2002) proposed the geographically weighted regression to address this "spatial non-stationarity" issue (Fotheringham et al. 2002, p 9). Panel data analysis, on the other hand, has received increasing interests in econometrics due to its obvious advantages over conventional cross-sectional or time-series data analysis techniques and increasingly available panel datasets (Hsiao 2003; Baltagi 2005). The enlarged sample size gives the researcher more degrees of freedom, reduces the collinearity among explanatory variables hence improves the efficiency of econometric estimates. Studies on both fields have yielded fantastic progresses, yet analysis that takes advantages of both methodologies eludes the research community. Two particular reasons would attribute to the lack of such combination.

First, geographically weighted regression, as its name suggests, focuses almost entirely on the ***spatial*** non-stationarity. The method recognizes that a set of universal coefficients in regression analysis might not be adequate to address the underlying data generating process of the observed geographic dataset. Instead, due either to intrinsic varying mechanisms or potential model misspecification, the regressed relationships are different from location to location. Relationships in regression analysis using geographic information, as evidenced in many a study (Fotheringham et al. 1998; Huang and Leung 2002; Yu and Wu 2004; Yu 2006; Yu et al. 2007), do vary in geographic space. It is only very recently, however, that scholars start to explore the possibility that relationships are potentially varying in not only geographic space, but also **temporal** space (Crespo et al. 2007; Demsar et al. 2008; Yu 2009).

Second, panel data analysis has long been regarded as an important analytical technique for econometric analysis. Although panel data analysis that utilizes geographic information is receiving increased attention in the mainstream econometric analysis (Anselin 1988, 2001; Elhorst 2001, 2003; Baltagi 2005; Anselin et al. 2008; Yu 2009; among others), such development focuses primarily on treating geography as an agent for dependence among cross-section observations. It is well known that the effects of geography are twofold – spatial autocorrelation and heterogeneity (Anselin 2001). Anselin et al. (2008) point out that the case of spatial heterogeneity can be handled by means of standard panel analysis methods. As detailed in Hsiao (2003), there is a full set of methods dealing with the so-called "**variable-coefficient models**" (Hsiao 2003, Ch. 6). While reviewing these well-developed methods, I found they indeed acknowledge the heterogeneous properties of the cross-sectional units. Such treatment, however, doesn't necessarily reflect the important characteristics of **spatial heterogeneity**.

As argued in Fotheringham et al. (2002), spatial heterogeneity is not like statistical heterogeneity that might follow certain distribution (Fotheringham et al. 2002). Instead, spatial heterogeneity is very much **determined** by **distances**. In GWR analysis, the spatial structure that follows the "First Law of Geography" (Tobler 1970) and generates spatial heterogeneity can be well simulated via the distance decaying Gaussian or Gauss-like kernel functions in which distance is the parameter. While in the "variable-coefficient" panel data analysis, **such important characteristics of geographic information are barely utilized**.

It is with this recognition that this proposed research attempts for the first time to combine research merits of both GWR and panel data analysis to produce new geo-panel data analysis methodology. In this particular study, I will utilize a set of regional development panel data from 1995 – 2001 of the Greater Beijing Area (GBA), China to develop such methodology. The results from this geo-panel analysis will be compared to the ones acquired from conventional methods. It is hoped with the new methods, we'll be able to discover new insights that was previously hidden in the dataset. Such new findings would potentially bring significant new understandings of regional studies in China.

The following section will give detailed reviews of the methodological development in spatiotemporal analysis from both geographic and econometric perspectives. This is followed by an introduction to the study region, GBA, China and the data. The fourth section extends the discussion of GWR and panel analysis and elaborates the development of the geographically weighted panel regression (GWPR) and its implementation. Results from applying the methods to the dataset will be reported in the fifth section. The study concludes with summary and future research foci.

2. BACKGROUND

2.1 Studies on spatiotemporal models and processes

Spatial data analysis techniques have borrowed many ideas from time series analysis. One of the most important aspect of spatial data, spatial autocorrelation, for instance, resembles the series autocorrelation, though differs in the way **lags** are defined (Anselin 1988; Anselin et al. 2008). The fundamental similarity between spatial data and time series data is that both follows a "*neighbors are similar*" Law. In spatial data, this is Tobler's (1970) "First Law of Geography", which resembles the common wisdom in time series analysis that observations close together in time will be more closely related than observations further apart. Another aspect of spatial data is the spatial heterogeneity, which constitutes the other aspect of Tobler's Law that "*non-neighbors are dissimilar*". It is the investigation of this spatial heterogeneity that leads to the development and implementation of the geographically weighted regression (Fotheringham et al. 2002). However, the current GWR analysis utilizes largely cross-sectional data instead of panel data. Though recent studies start to consider temporal information in GWR analysis (see Desmar et al. 2008; Yu 2009; Yu and Lv 2009), integrating time series data in GWR analysis is still under-developed.

Integrating time series into geographic analysis is termed **spatiotemporal analysis**. This spatiotemporal modeling technique has been applied to a wide range of scientific and engineering fields. Studies in the genre, however, focus mainly on the spatiotemporal clustering of observations and interpolation. For instance, Knox (1964) investigates the space-time interaction of epidemics and develops the Knox test to determine whether or not there are apparent spatiotemporal clusters. Bilonick (1985) and Kyriakidis and Journel (2001) apply the spatiotemporal models to determine space–time trends in the deposition of atmospheric pollutants. Bras and Rodrígues-Iturbe (1984), Armstrong et al. (1993) apply spatiotemporal kriging procedure to estimate rainfall in various regions. Hohn et al. (1993) develop spatiotemporal model to characterize population dynamics in ecology, to name but a few.

As pointed out by Kyriakidis and Journel (1999), joint analysis of space and time in a spatiotemporal framework mainly builds on the extension of established spatial analytical techniques that are widely applied in the fields of geology (Journel and Huijbregts 1978), forestry (Matérn 1980), and meteorology (Gandin 1963). Such extension usually treats time as an added spatial dimension, hence enlarges the two-dimensional geographic space to a three-dimensional **geographic-time** space. However, simple extension as such might not be all that plausible due to the fundamental differences between geographic space and time (or ***geographic*** space and ***temporal*** space). Geographic space represents a state of coexistence, in which there can be multiple directions. While temporal space represents a state of successive existence, a nonreversible ordering in only one direction is present (Snepvangers et al. 2003). Isotropy is well defined in **geographic** space, but has no meaning in a space-time context due to the ordering and nonreversibility of time.

The majority of the above mentioned studies are largely confined in the field of geostatistics (Kyriakidis and Journel 1999). The primary goals of these studies are fairly similar (Snepvangers et al. 2003): to predict an attribute $z = \{z(s,t) \quad s \in S, t \in T\}$ defined on a geographical domain $S \subset R^2$ and a time interval $T \subset R^1$, at a space–time point (s_0, t_0), where z was not measured. The prediction is to be based on n geographic measurements at t time intervals which constitute the nt points (s_i, t_i), with i=1, … , n. Seldom do the studies focus on relationships between regressed variables in the spatiotemporal framework. Just as in a pure cross-sectional scenario, regressed relationships tend to vary from geographic location to geographic location (the essence of the GWR method); it is very tenable that regressed relationships might vary from spatiotemporal location to spatiotemporal location.

2.2 The variable coefficient panel data analysis model

Panel data analysis has been well developed in econometrics for decades (Baltagi, 2005). It differs from pure cross-sectional or time-series analysis by incorporating both dimensions. Apparently, the added dimension enlarges the sample size hence contains more degrees of freedom, adds more variability, renders less collinearity among the variables, and gives more efficiency for estimation (Hsiao, 2003). Panel data analysis with geographic data has only recently attracted scholarly attention (Anselin 1988, 2001; Elhorst 2001, 2003; Anselin et al. 2008; Lv and Yu 2009). The focus of this trend of *spatial panel data analysis*, as termed in both Elhorst (2001, 2003) and Anselin et al. (2008), is primarily an extension of the spatial data analysis techniques with cross-sectional data. Estimations of the parameters focus on the pooled model that either incorporates a spatial lag term in the RHS of the equation or a spatial error term. The potential of heterogeneous parameters are usually overshadowed due to the less accurate prediction performance than the pooled model (Baltagi 2005; Baltagi et al. 2008) or a willingness to trade bias over a reduction in variance (Toro-Vizcarrondo and Wallace 1968).

Of course, this is not to say that panel data analysis can't deal with heterogeneous parameters. As a matter of fact, Hsiao (2003) indicates that "when data do not support the hypothesis of coefficients being the same, yet the specification of the relationships among variables appears proper or it is not feasible to include additional conditional variables, then it would seem reasonable to allow variations in parameters across cross-sectional units and/or over time as a means to take account of the interindividual and/or interperiod heterogeneity" (p.141). Many a study also indicates that pooling parameters over cross-sectional units might not be very tenable (Robertson and Symons 1992; Pesaran and Smith 1995; Pesaran et al. 1999). This is especially true when the cross-sectional units are samples from geographic space, as dictated by the "First Law of Geography" (Anselin 1988). However, if all the coefficients are treated as fixed and different for different cross-sectional units in different time periods, there will be more unknown parameters than available observations (N by K by T unknown parameters with only N by T observations). Apparently, we won't be able to estimate the unknowns from the data. To solve this dilemma, we need to search for approaches that allow the coefficients to differ, yet reduce the unknown parameters to be less than the available data. Hsiao (2003) introduced two potential approaches to solve the dilemma. First the coefficient is separated to three components including a trend, an individual variation and a temporal variation. Then either by treating the individual and temporal variations as fixed or random, we can impose restrictions (when fixed) or assume/estimate a distribution (when random) to drastically reduce the unknown parameters. It is found, however, such treatments are usually rather computationally prohibitive. Applications of those methods are rather limited (Hsiao, 2003).

Other than the computational consideration, the variable coefficient panel analysis is largely an **aspatial** approach in dealing with geographic information. No matter the fixed or the random approach, if the cross-section is on geographic space, it is apparent that the important characteristics of geographic information (governed by the "First Law of Geography) are not utilized. Apart from the above fixed with restriction, and random with distribution approaches, a third approach, in which the varying coefficients can be obtained via functions of the spatiotemporal locations, might seem to be rather tenable an alternative, yet studies are seldom extended in this direction.

3. STUDY AREA: THE GREATER BEIJING AREA, CHINA

The GBA is located in the Northern China Plain, includes Hebei province and Beijing, Tianjin provincial municipalities. The region is also often called the Capital Economic Circle, or Jing-Jin-Ji region. The area has in total 170 county level spatial units (Fig. 1). During the pre-reform era, due to the central location of Beijing as the national capital, GBA was one of the most developed heavy industry centers in China. As pointed out by Lu (1997), during the 1950s, 95% of the national and local investment went to heavy industries. Such massive investment brought tremendous economic gains for GBA under Mao's China (Yu and Wei 2008), and also formed the heavy industry-centered and government-sponsored economic structure.

During the reform era, however, as China gradually integrates its own economy to the global economic system, the changed global and regional geopolitical environment enables the southern provinces to achieve a rapid economic recovery. While in the mean time, the central government takes a very cautious attitude towards the reform in its heart regions, the GBA. Reform policies are experimented in the southern provinces and gradually extended to other parts of the nation as they are proven successful. Under such scenarios, many a scholar discovers an interesting trend in China's regional development dynamics during the first decade of reform that regional inequality converges (Yu and Wei 2003). Such convergence, however, reflects only a residual effect of China's economic distribution before the reform era. As a matter of fact, regional inequality in China resumes and deepens after the 1990s (Yu and Wei 2003). Yet this cautious attitude of the government again creates a fairly different regional development pattern in GBA than those often observed and studied in the southern provinces.

Recent research focus on the southern provinces for the reform China is well-justified as these regions spearhead China's economic dynamics during the reform era. Yet it is quite unrealistic to assume that development status and dynamics in these regions would be representative of China's regional development. As argued above and presented in Yu and Wei (2008), the patterns and status of the GBA's development might differ drastically from its southern peers. Hence an exploration to this particular region might shed light towards a more complete understanding of China's regional development.

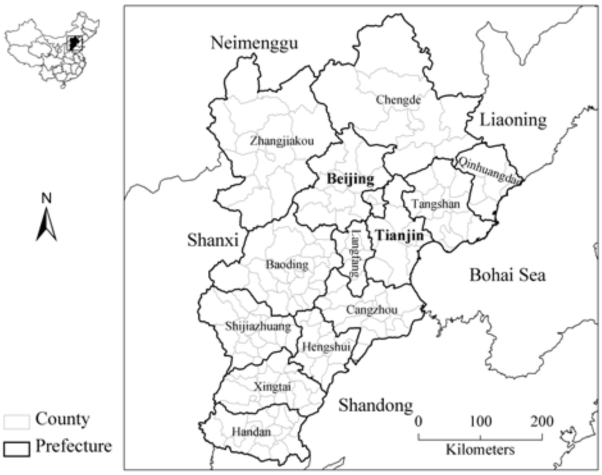


Figure 1: Location of GBA, China

Yu (2006) and Yu and Wei (2008) have pioneered the work in this direction. Their analyses of GBA indeed brought some fairly interesting results as different from the often studied southern provinces. For instance, they found that in contrast to the usually negative effect of investment in China's state-owned-enterprises (SOEs) in economic development, SOEs do not have significant impact in GBA (Yu and Wei 2008). Not surprisingly, they also identified that the governmental supports and investment dominate the performance of local economies. Agreeing with the results found in the southern provinces, attracting foreign direct investment seems to be an important factor to boost local economies as well. The increased urbanization, however, doesn't seem to be well associated with local economic performance.

Recent works in China's regional studies employ some rather recent development in GIS and spatial data analysis such as spatial regression and geographically weighted regression (Leung and Huang 2002; Yu 2006). These studies, however, resemble many others in that analyses are done from a cross-sectional aspect. Though data with time dimension are used, panel data analysis is left as an unexplored area. As argued by Baltagi (2005), cross-sectional analysis with relative stable distribution might hide a multitude of changes. Even with repeated cross-sectional analyses at different time periods, the dynamics of adjustment that are often of more interests will not be present. The current study hence intends provide better understanding of GBA's regional development via the application of advanced spatial and temporal analytical methodologies with a short panel from 1995 to 2001. In particular, to model the relationship between GBA's economic performance and a set of identified mechanisms, a geographically weighed panel regression analysis is developed and applied. The practice intends to capture the spatiotemporal dynamics of GBA's regional development from 1995 – 2001.

4. METHODOLOGY: GEOGRAPHICALLY WEIGHTED PANEL REGRESSION ANALYSIS

The central idea of geographically weighted panel regression (GWPR) analysis is fairly similar to the cross-sectional GWR analysis. In GWPR, however, it is assumed that the time series of observations at a particular geographic location is a realization of a smooth spatiotemporal process. Such spatiotemporal process follows a distribution that closer observations (either in geography or in time) are more related than distant observations. Depending on the panel analysis intends to pool over geographic (cross-sectional) or temporal observations, we can apply different models to simulate such process. In this particular study, since our panel data is a relatively short panel (7 years), but covers more cross-sectional units (170 counties), I will focus the discussion on developing models that simulate the spatiotemporal process over geographic space. The other scenario with more temporal observations can follow similar route of arguments.

If we only concern the regression coefficients vary over cross-sectional units (geographic space), the spatiotemporal process is effectively reduced to a spatial process just as in GWR analysis. Unlike GWR analysis, however, the spatial process is applicable to all the temporal observations simultaneously and is assumed to be temporally invariant (due to the short period). Based on such postulation, the GWPR on short panel can be seen as an expanded version of the cross-sectional GWR analysis to panel data. Following similar arguments as in GWR, a bandwidth (or bandwidths in adaptive kernel) can be obtained for each location to determine a set of local sampling locations. Observations within the local sampling locations will be weighted based on a kernel function just as in GWR (Fotheringham et al. 2002). Such weighting will be applied to all temporal periods. Within these local sampling locations, it is assumed that the panel is poolable over geographic space. A fixed or random effects model as detailed in Baltagi (2005) can be applied to obtain the coefficients of the explanatory variables at that specific location.

From the experiences of applying GWR with cross-sectional data, we found that Gaussian or Gaussian-like kernel density functions work rather well in simulating the spatial distance-decaying process (Fotheringham et al. 2002). Similar principles apply to the GWPR scenario. Specifically, a spatial kernel function will be established very much the same as the kernel functions in cross-sectional GWR analysis. The kernel function and its bandwidth will be used to determine the size of the subsample around any particular geographic location and assign weights to existing data points. Unlike the cross-sectional GWR model, this subsample will be a subsample of panel data that include both spatial and temporal observations. Weights generated from the spatial kernel function, however, will remain temporally invariant to keep the model simple. Temporally variant weights can certainly be generated by introducing a temporal scalar for each time period. The essence of the method would not change. After the sub-setting and weight-assigning, we can then apply a panel regression procedure for each location. Either a fixed effects or random effects panel analysis model will be applied to this subsample and obtain a unique coefficient for that particular location. The procedure can then be repeated for all the geographic locations to obtain the set of variable coefficients over geography.

One of the key components in applying locally weighted panel regression is the size of the local samples, per GWR terminology, the bandwidth of the (fixed) kernel function or the nearest neighbor of the (adaptive) kernel function. Two criteria are applied in cross-sectional GWR analysis. One is based on the cross-validation score (CV) and the other the Akaike Information Criterion (AIC) (see Fotheringham et al. 2002 for detail). At the current stage of development of GWPR, I focus only on utilizing the cross-validation score to determine the local sample size and the kernel weighting. Similar to how CV score is determined in cross-sectional GWR analysis, CV score is calculated based on the average of the dependent and independent variables over time:

$$CV = \sum_{i=1}^{n} [\bar{y}_i - \hat{\bar{y}}_{\neq i}(b)]^2 \tag{1}$$

where $\bar{y}_i$ is the average over time of the dependent variable at location *i*, $\hat{\bar{y}}_{\neq i}(b)$ is the estimated dependent variable with bandwidth *b* and excluding observation in location *i*.

Implementation of GWPR is done with R scripts (R Development Core Team, 2009). I have extended the cross-sectional GWR codes (SPGWR, Bivand and Yu, 2009) via incorporating panel analysis codes (PLM, Croissant, 2009). The codes are available upon request. At the current stage, estimation of the *geographically* variable coefficients, pseudo-significance t test for each coefficient are done.

5. RESULTS AND DISCUSSION

Based on previous studies in GBA, China (Yu, 2006; Yu and Wei 2008), five particular variables are identified for the exploration of regional development. Specifically, for each county, the per capita GDP (GDPPC) value is used as a proxy for regional development. Per capita fixed asset investment (FIXINVPC), per capita financial income (FININCPC), per capita foreign direct investment (FDIPC), and urbanization level (URB) are chosen as the development mechanisms. Among them, FIXINVPC represents the central government's support to local economic development. FININCPC indicates the local governments' financial capability. The financial capability of local governments would represent their potential possibility to support regional development. FDIPC is usually argued as the agent of globalization in China's regional development studies (Wei 2000, Fujita and Hu 2001). URB attempts to capture the co-movement between economic development and urbanization in China. The econometric relationship between development and mechanisms takes the form:

$$GDPPC = A \times FIXINVPC^{\beta_1} \times FININCPC^{\beta_2} \times FDIPC^{\beta_3} \times URB^{\beta_4} \tag{2}$$

A logarithm transformation of the above production-function alike equation yields a linear relationship between the logarithms of the above variables, and takes the usual form:

$$Y = X\boldsymbol{\beta} + \varepsilon \tag{3}$$

where Y is the logarithm transformed GDPPC; X is the matrix containing the four independent variables in their logarithm transformed forms and a constant term; $\boldsymbol{\beta}$ is the vector of model coefficients; and $\boldsymbol{\varepsilon}$ is the vector of unobservable noise.

For short panel data such as the one we are using, it is rather hard to justify the application of a random effect model (Baltagi, 2005). A Hausman's test suggests just that. In addition, *F* test indicates that the dataset used has strong individual effects than time effects, which justify our pooling over cross-sectional units instead of time. The analysis hence discusses results generated from fixed effect panel analysis that has individual (cross-sectional) effect.

By using an adaptive kernel function, cross-validation for GWPR points out an optimal local sample (which minimizes the CV score) contains 26 geographic observations. Table 1 presents the results of an individual fixed effect panel regression analysis. Figure 2 gives the results generated by GWPR. Only coefficients that are pseudo-significant at 95% confidence level via the pseudo-t test are greyed. For comparison purposes, a cross-sectional GWR analysis using only data from the year 2001 is presented in Figure 3 as well.

From reading the tables and figures, a few observations emerge. First, resonating with previous findings (Yu and Wei 2008; Yu 2009), it seems no matter in aspatial panel analysis or cross-sectional GWR or GWPR, per capita foreign direct investment, which was usually deemed the agent of globalization, doesn't really play much of a role in the Greater Beijing Area. Such an observation would trigger a very interesting question: as GBA *is* one of China's economic centers, and GBA *is* progressively globalizing, why isn't globalization contributing to local regional development. As a matter of fact, according the GWPR analysis, FDIPC actually significantly (at 95% confidence level) works against regional development in Beijing and the inland Hebei counties that are adjacent to Beijing (Figure 2c). Possible answers would include the fact that FDIPC might not be a very good agent of globalization in this specific geography as it was originally identified in studying the southern China. In this regard, it might be more appropriate to identify a different agent of globalization in GBA, such as number of international visits. It might also attribute to the fact that, however, GBA's globalization process is also heavily involved with localization process, as Beijing is not only an economic center, but a cultural and political center as well. In addition, it is understandable that comparing with their southern peers such as Zhejiang and Jiangsu, regions in GBA, especially counties in inland Hebei province were not quite attractive during the period from 1995 – 2001 to FDI.

	Estimate	Std. Error	t-value	Pr(>|t|)
FININCPC	0.528	0.016	32.209	0.000
FDIPC	0.002	0.002	0.798	0.425
FIXINVPC	0.073	0.012	6.280	0.000
URB	0.287	0.044	6.527	0.000

Total Sum of Squares: 54.651
Residual Sum of Squares: 10.921
F-statistic: 1017.07 on 4 and 1016 DF, p-value: < 2.22e-16

Table 1. Panel regression analysis of GBA, China

Second, all the analyses point to the most important regional development mechanism in GBA is the local financial capability (figures 2b and 3b). This further supports the fact that decentralization in China, even at a location that is so centralized is working in favor to regional development. Although the two geographically weighted analysis captured the fact that Beijing, as the centralization center, benefits rather less from the local financial capability than its peers in Hebei and Tianjin. The difference between GWPR and cross-sectional GWR in 2001, however, remains quite interesting. With more information available for estimation, GWPR clearly picks out an urban area oriented trend that more urbanized regions benefit more than the less urbanized ones. This shall not come as a surprise, however, considering the administrative characteristics and fiscal distribution in China. Counties usually don't have their own fiscal revenue per se. The decentralization of fiscal power stops at the prefecture level. Within a specific prefecture, it is like a small regime of a centralized entity, in which the ones that are at the top tier enjoy more of the benefits than the ones that are below. This feature, however, is rather obscured in the cross-sectional GWR analysis in 2001. Similar conclusions can be drawn for per capita fixed asset investment, which is used to represent the central government's support for regional development. It seems that the central government's support is rather important mainly in the peripheral counties than in the more urbanized ones. From Figure 2a, support from the central government is not even significantly related with local development in Beijing and Tianjin.

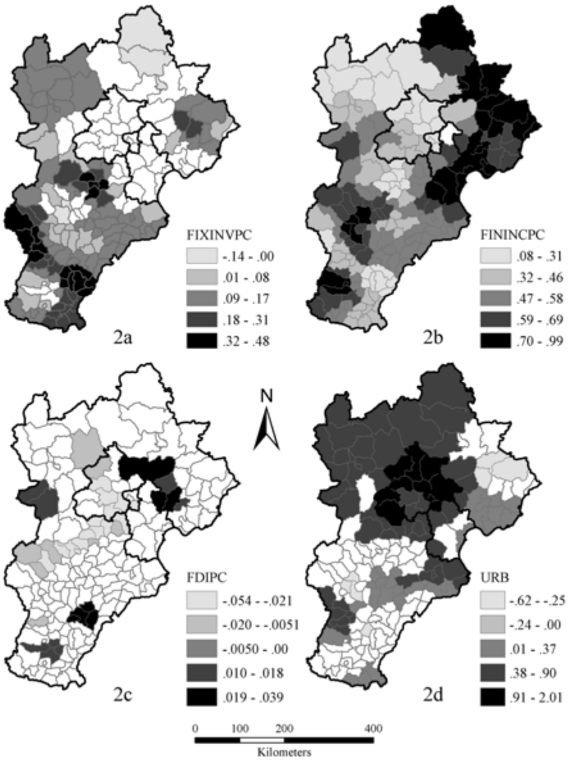


Figure 2. Coefficients surfaces generated from the GWPR, only locally pseudo-significant counties are greyed: 2a. coefficient surface for per capita fixed asset investment; 2b. coefficient surface for per capita financial income; 2c. coefficient surface for per capita foreign direct investment; 2d. coefficient surface for urbanization.

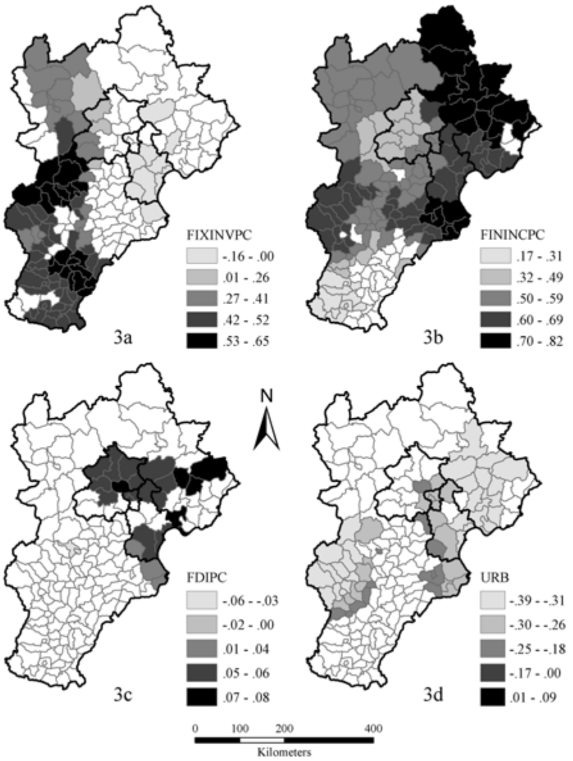


Figure 3. Coefficients surfaces generated from the cross-sectional GWR in 2001, only locally pseudo-significant counties are greyed: 3a. coefficient surface for per capita fixed asset investment; 3b. coefficient surface for per capita financial income; 3c. coefficient surface for per capita foreign direct investment; 3d. coefficient surface for urbanization.

Third, quite interestingly, when we are comparing Figures 2a and 2b, especially the shadings of the significant values, it is almost immediately clear that the two types of governments' supports, i.e., the central and local governments (represented by fixed asset investment and local financial income), are not only the strongest supportive mechanisms for regional development in GBA, but also complementary to each other across the region. Such mutual-complementing pattern is barely discernible in the cross-sectional GWR analysis with 2001 data (Figures 3a and 3b). It is, however, quite evident in the GWPR maps in which more information participated in the analysis. This mutual-complementing regional development mechanism is a significant discovery in the regional development studies in GBA, China. This result suggests a balanced investment strategy was on-going from 1995 – 2001 in GBA, in which the central government purposefully invested more on regions that had less financial self-dependence. Such an investment strategy reflects the developing history of GBA that it used to be one of the heavy industrial centers in China, and traditionally dependent heavily on government's supports for its economic development. Economic reform that started in 1978 changed the developing modes all across China drastically, yet the investment structure remains quite resistant. Such a pattern would not be immediately observable from cross-sectional analysis. With added dimension of temporal information, and the integration of geographic weighting techniques, the GWPR is able to make rather thorough discoveries.

Fourth, yet the most interesting conclusion drawn via applying GWPR is the relationship between urbanization and regional development in GBA. Our previous studies (Yu and Wei, 2008; Yu 2009) with cross-sectional analysis indicates urbanization is at best marginally contributing to regional economies. This is also reported via the cross-sectional GWR analysis (Figure 3d). The relationships between urbanization and per capita GDP are not only mostly negative, but also not significant at all in many counties. GWPR, however, suggests otherwise. As a matter of fact, via modeling with the added temporal information, it stands out immediately that more urbanized an area, higher the level of per capita GDP. This is especially true in Beijing, Tianjin and the capital city of Hebei, Shijiazhuang (figure 2d, place reference see figure 1). This finding supports the common wisdom in GBA, China that large cities tend to be more developed than less urbanized areas. More importantly, this finding solves a seemingly anti-intuitive dilemma that was usually obtained from cross-sectional analysis that urbanization is not significantly related with regional development. The advantage of modeling with more information speaks for itself again here.

6. REFERENCE

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. 2001. Spatial econometrics. In *A Companion to Theoretical Econometrics,* edited by B.H. Baltagi, Blackwell, Oxford, p. 310–330.

Anselin, L., Le Gallo, J. and Jayet. H. 2008. Spatial panel econometrics. In: *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice (3rd ed.)*, edited by L. Matyas and P. Sevestre, Springer, New York, p. 625-662.

Armstrong, M., Chetboun, G. and Hubert, P. 1993. Kriging the rainfall in Lesotho. In *Geostatistics Tróia '92, Vol. 2,* edited by A. Soares, Kluwer Academic Publisher, Dordrecht, p. 661–672.

Baltagi, B.H. 2005. *Econometric Analysis of Panel Data (3rd ed.)*, Wiley, New York.

Bilonick, R.A. 1985. The space-time distribution of sulfate deposition in the northeastern United States. *Atmospheric Environment*, 19(11): 1829–1845.

Bivand, R. and Yu, D.L. 2009. *Statistical package for geographically weighted regression analysis, SPGWR*, URL http://cran.r-project.org/web/packages/spgwr/index.html.

Crespo, R., Fotheringham, A.S. and Charlton, M.E. 2007. Application of geographically weighted regression to a 19-year set of house price data in London to calibrate local hedonic price models. In: *Proceedings of the 9th International Conference on Geocomputation 2007 (Maynooth, Ireland)*, National University of Ireland Maynooth: Maynooth, Ireland, 2007.

Croissant, Y. 2009. *Linear model for panel data*, URL: http://cran.r-project.org/web/packages/plm/plm.pdf

Demsar, U., Fotheringham, A.S. and Charlton, M.E. 2008. Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. *Information Visualization*, 1-17.

Elhorst, J.P. 2001. Dynamic models in space and time. *Geographical Analysis*, 33:119–140.

Elhorst, J.P. 2003. Specification and estimation of spatial panel data models. *International Regional Science Review*, 26(3): 244–268.

Fotheringham, A.S., Brunsdon, C.F. and Charlton, M.E. 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, 30: 1095-1927.

Fotheringham, A.S., Brunsdon, C.F. and Charlton, M.E. 2002. *Geographically Weighted Regression: the Analysis of Spatially Varying Relationships*. Wiley, West Sussex.

Fujita, M., & Hu, D. (2001). Regional disparity in China 1985-1994: the effects of globalization and economic liberalization. *The Annals of Regional Science*, *35*, 3-37

Gandin, L. 1963. *Objective Analysis of Meteorological Fields*, Gidrometeorologicheskoe Izdatel’stvo (GIMEZ), Leningrad.

Hohn, M.E. Liebhold, A.M. and Gribko, L.S. 1993. Geostatistical model for forecasting spatial dynamics of defoliation caused by the gypsy moth (Lepidoptera: Lymantriidae). *Environmental Entomology*, 22(5): 1066–1075.

Hsiao, C. 2003. *Analysis of Panel Data (2nd ed.)*, Cambridge University Press, New York.

Huang, Y. and Leung, Y. 2002. Analysing regional industrialisation in Jiangsu province using geographically weighted regression. *Journal of Geographical Systems*, 4: 233-249.

Journel, A. G. and Huijbregts, Ch. J. 1978, *Mining Geostatistics*, Academic Press, New York.

Knox, E.G. 1964. Epidemiology of childhood leukaemia in Northumberland and Durham. *British Journal of Preventive and Social Medicine*, 18: 17-24.

Kyriakidis, P.C. and Journel, A.G. 1999. Geostatistical space– time models: a review. *Mathematical Geology*, 31: 651–684.

Kyriakidis, P.C. and Journel, A.G. 2001. Stochastic modeling of atmospheric pollution: a spatial time-series framework: Part I. Methodology. *Atmospheric Environment* 35: 2331– 2337.

Lu, D. 1997. *Zhongguo Yanhai Diqu 21 Shiji Chixu Fazhan (Sustainable development of China’s coastal regions in the 21st century)*. Wuhan: Hubei Science and Technology Press

Lv, B.Y. and Yu, D.L. 2009. Improvement of China’s regional economic efficiency under the gradient development strategy: A spatial econometric perspective. *Social Science in China,* 20 (6): 60-72.

Matérn, B. 1980. *Spatial Variation, Lecture Notes in Statistics, (2nd ed.),* Springer, New York.

Pesaran, M.H. and Smith R. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68: 79–113.

Pesaran, M.H., Shin, Y. and Smith R. 1999. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, 94: 621–634.

R Development Core Team, 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robertson, D. and Symons, J. 1992. Some strange properties of panel data estimators. *Journal of Applied Econometrics*, 7: 175–189.

Rodríguez-Iturbe, I. and Mejía, J.M. 1974. The design of rainfall networks in time and space. *Water Resources Research*, 10(4): 713–728.

Snepvangers, J.J.J.C., Heuvelink, G.B.M. and Huisman J.A. 2003. Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma*, 112: 253– 271

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46: 234-240.

Toro-Vizcarrondo, C. and Wallace, T.D. 1968. A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63: 558–572.

Wei, Y. H. D. 2000. *Regional development in China: states, globalization, and inequality*. London: Routledge

Yu, D.L. 2006. Spatially varying development mechanisms in the Greater Beijing Area: a geographically weighted regression investigation. *Annals of Regional Science*, 40: 173-190.

Yu, D.L. 2007. Modeling housing market dynamics in the city of Milwaukee: a geographically weighted regression approach. *GIScience and Remote Sensing*, 44: 267-282.

Yu, D.L. 2009. Understanding regional development mechanisms in Greater Beijing Area, China, 1995 – 2001, from a spatial-temporal perspective. *GeoJournal*, in press

Yu, D.L. and Lv, B.Y. 2009. Measurement of Provincial Total Factor Production: Application of Geographically Weighted Regression from a Spatial Temporal Perspective. *Journal of Chinese Soft Science*, 16 (11): 160-170

Yu, D.L. and Wei, Y.H.D. 2003. Analyzing regional inequality in post-Mao China in a GIS Environment. *Eurasian Geography and Economics*, 44: 514-534.

Yu, D.L. and Wu, C. 2004. Understanding population segregation from Landsat ETM+ imagery: a geographically weighted regression approach. *GIScience and Remote Sensing*, 41: 145-164.

LAND SUITABILITY EVALUATION FOR WHEAT CULTIVATION BY FUZZY THEORY APPROACHE AS COMPARED WITH PARAMETRIC METHOD

M. Mokarram [a,*], K. Rangzan [a], A. Moezzi [b] , J. Baninemeh[c]

[a] Dept. of Remote Sensing and GIS, Shahid Chamran University,Ahwaz, Iran (m.mokarram.313, kazemrangzan @gmail.com)
[b] Dept. of soil science , Shahid Chamran University,Ahwaz, Iran (moezzi251@gmail.com)
[c] Scientific board member, Khuzestan Agriculture and Natural Resources Research Centre. Email: jamal_nn@ yahoo.com
*Member of young researcher club of Islamic Azad university of Safashahr, Iran

KEY WORDS: suitability, Fuzzy, wheat, AHP, GIS, IDW, Kappa, Con

ABSTRACT:

Nowadays infinity examples of irreparable damaging to natural resources have been occurred due to lack of attention and improper uses of soil and water. Land evaluation is a process of assessment of land performance when used for specified purposes. In other words, Land Evaluation is the estimation of the possible behaviour of the land when used for a particular purpose. The main purpose of this study is to prepare land suitability evaluation maps for Wheat using Fuzzy classification in Shavur area, Khuzestan province. In the model non-physical factors is included. The results are compared to a Crisp classification using the standard FAO framework (parametric) for land evaluation which, include non-physical parameters as well. In the present study, eight soil parameters, such as soil Texture, Wetness (ground water depth and hydromorphy), Cation Exchange Capacity (CEC) and Exchangeable Sodium Percentage (ESP), Gypsum (%), $CaCO_3$ (%),Topography, Soil depth and pH values, are chosen for crop-land suitability analysis and thematic maps are developed for each of the parameters with IDW model. Different Fuzzy membership functions obtained from the literature (Con function) were employed and the weights for each parameter were calculated according to an Analytic Hierarchy Process (AHP) that relies on pair wise comparisons. Climatic requirements, landscape and soil requirements for selected crop was determined based on parametric method. Finally classes of land suitability provided for each Land unit. The coefficient of Kappa is used for comparing these two methods and choosing the better one. The results with the parametric method showed 26% of the area as moderately suitable, 25% as marginally suitable and 49% as unsuitable. The results with the Fuzzy theory showed 31% of the study area as highly suitable for wheat 29 % as moderately suitable, 19% as marginally suitable and 21% as unsuitable. Based on the results it has been concluded that Fuzzy method allows obtaining results that seems to be corresponded with the current conditions in the area.

1. INTRODUCTION

Agriculture is important as a source of food and income, but How, Where and When to cultivate are the main issues that farmers and land managers have to face day to day. Land evaluation is carried out to estimate the suitability of land for a specific use such as arable farming or irrigated agriculture. Land evaluation can be carried out on the basis of biophysical parameters and/or socio-economic conditions of an area (FAO 1976). Planning and management of the land use suitability mapping and analysis is done by application of GIS (Geographic Information System) (McHarg, 1969; Brail and Klosterman, 2001; Collins et al., 2001). The GIS-based land use suitability analysis has been applied in a wide variety of situations including ecological approaches for defining land suitability/habitant for animal and plant species (Store and Kangas, 2001), geological favourability (Bonham-Carter, 1994), suitability of land for agricultural activities (Cambell et al., 1992; Kalogirou, 2002), landscape evaluation and planning (Miller et al., 1998), environmental impact assessment (Moreno and Seigel, 1988), selecting the best site for the public and private sector facilities (Church, 2002) are also other examples. The GIS-based approaches to this problem have their roots in the applications of hand-drawn overlay techniques used by American landscape architects in the late nineteenth and early 20th century (Collins et al., 2001). Several studies have been focused on this subject, including evaluation of many factors and aggregation of these factors in many different ways (Lukasheh et al. 2001; Kontos et al. 2003; Sener et al. 2006). The overlay procedures play a central role in many GIS applications (O'Sullivan and Unwin, 2003) including techniques that are in the forefront of the advances in the land use suitability analysis such as: multi-criteria decision analysis (MCDA) (Malczewski, 1999), artificial intelligence (AI) ingeo-computation methods (Ligtenberg et al., 2001; Xiao et al., 2002) and visualization methods (Jankowski et al., 2001). Over the last forty years or so GIS-based land use suitability techniques have increasingly become integral components of urban, regional and environmental planning activities (Collins et al., 2001). GIS are used for geographic data acquisition and processing. The analytical hierarchy process (AHP) developed by Saaty (1977) is the multi-criteria evaluation technique used, enhanced with Fuzzy factor standardization. Besides assigning weights to factors through the AHP, control over the level of risk and trade off in the siting process is achieved through a second set of weights, i.e., order weights, applied to factors in each factor group, on a pixel-by-pixel basis, thus taking into account the local site characteristics. The AHP has been incorporated in the GIS technology producing a flexible way of combining various criteria.

The main purpose of this study is to prepare land suitability evaluation maps for Wheat using Fuzzy classification and compare it with FAO method for Shavur area in Khuzestan in GIS.

2. METHODS

2.1 Fuzzy method

Fuzzy logic was initially developed by Lotfi Zadeh in 1965 as a generalization of classic logic. Zadeh (1965) defined a Fuzzy set as "a class of objects with a continuum of grades of memberships"; being the membership a function that assigns to each object a grade ranging between zero and one, the higher the grade of membership the closest the class value to one. Traditionally thematic maps are represented with discrete attributes based on Boolean memberships, such as polygons, lines and points. These types of entities have a value or do not have it; an intermediate option is not possible. With Fuzzy theory, the spatial entities are associated with membership grades that indicate to which extent the entities belong to a class (Hall et al, 1992). Mathematically, a fuzzy set can be defined as (Mc Bratney A. B. and Odeh I. O. A. 1997):

$$A = \{x, \mu_A(x)\} \qquad \text{For each } x \varepsilon X \qquad \text{Eq.1}$$

Where μ_A is the function (membership function MF) that defines the grade of membership of x in **A**. The MF $\mu_A(x)$ takes values between and including 1 and 0 for all **A**. If $X = \{x_1, x_2, ..., x_n\}$ the previous equation can be written as:

$$A = \{[x_1, \mu_A(x_1)] + [x_2, \mu_A(x_2)] + ... + [x_n, \mu_A(x_n)]\} \qquad \text{Eq.2}$$

In plain words equations 1 and 2 mean that for every x that belongs to the set **X**, there is a membership μ_A function that describes how the degree of ownership of x in **A** is.

Mc Bratney and Odeh (1997) expressed the fuzzy membership function μ_A as (X)→[0,1] with each element X belonging to **X** with a grade of membership μ_A (X) ε [0,1] this way μ_A=0 represents that the value of X does not belong to **A** and μ_A=1 means that the value belongs completely to **A**. Alternatively 0< μ_A (X)<1 implies that X belongs in a certain degree to **A**.

The membership function can take any shape and can be symmetrical or asymmetrical. The simplest function is of triangular form but Trapezoidal, Gaussian, Parabolic among others are also possible. Given the non-discrete characteristics of soils and land use, fuzzy theory suits well to the analysis of land suitability. With fuzzy representation the boundaries between suitability classes are not so strict and map units that are more or less suitable that is in an intermediate condition can be described properly. The development of GIS has contributed to facilitate the mapping of land evaluation results, both Boolean and fuzzy, but the topological rules imbibed in GIS software are based on Crisp theory.

Interpolation using of 64 sampling point are developed for each of the parameters with IDW (Inverse Distance Weighted) for production map for each one of parameters model. The calculation of the fuzzy memberships for the Soil depth and Wetness (water depth and hydromorphy) was evaluated using a linear function as given in Eq.7 (Moreno, 2007).

$$\mu_A(X) = f(x) = \begin{Bmatrix} 0 & x \le a \\ x - a / b - a & a \prec x \prec b \\ 1 & x \ge b \end{Bmatrix} \qquad \text{Eq.7}$$

Where x is the input data and a and b are the limit values according to Sys tables.

For Texture soil, Cation Exchange Capacity (CEC), Exchangeable Sodium Percentage (ESP), Gypsium (%), CaCO3 (%) ,Topography, and pH values, using a linear function as given in Eq.8 (Moreno, 2007).

$$\mu_A(X) = f(x) = \begin{Bmatrix} 1 & x \le a \\ b - x / b - a & a \prec x \prec b \\ 0 & x \ge b \end{Bmatrix} \qquad \text{Eq. 8}$$

For land suitability it is required to calculate the convex combination of the raster values containing the different fuzzy parameters. The convex combination means that "if $A_1, ... A_k$ are fuzzy subclasses of the defined universe of objects **X** and $w_1, ... w_k$ are non-negative weights summing up to unity, then the convex combination of $A_1, ... A_k$ is a fuzzy class **A** whose membership function is the weighted sum" (Burrough, 1989), where the weights $w_1, ... w_k$ were calculated using APH as described in the previous section and the fuzzy parameters μ_A have been calculated with the membership functions described in the previous sections and using conditional statements in ArcGIS. Equations 3 to 5 present the convex combination.

$$\mu_A = w_1 \cdot \mu_{A1} + ... \; w_k \cdot \mu_{Ak} \qquad \text{Eq.3}$$

$$\mu_A = \sum_{j=1}^{k} w_j \cdot \mu_{Aj(x)} \qquad x \in X \qquad \text{Eq.4}$$

$$\sum_{j=1}^{k} w_j = 1 \qquad w_j > 0 \qquad \text{Eq.5}$$

AHP relies on Pairwise Comparison Matrices which are matrices relating different components and assigning values according to their relative importance. These values are given by a scale from 1 to 9, where 1 means that the two elements being compared have the same importance and 9 indicates that from the two elements one is extremely more important than the other

2.2. FAO Framework method

In this study the FAO Framework for Land Evaluation (1976) has been employed to classify the potential land use. According to this framework, the structure for suitability classification is composed of four categories:

I. Land Suitability Orders: reflecting kinds of suitability. S: suitable, N: non suitable.

II. Land Suitability Classes: reflecting degrees of suitability within orders such as S1(highly suitable), S2(moderately suitable), S3 (marginally suitable) and N (not suitable).

III. Land Suitability Subclasses: reflecting kinds of limitation or main kinds of improvement measures required, within classes (e.g. S2m, S2e, etc.).

IV. Land Suitability Units: reflecting minor differences in required management within Subclasses such as S2e-1, S2e-2.

In evaluating of the qualitative land suitability, land properties were compared with the corresponding plant requirements. In this stage, in order to classify the lands, the Sys *et al.* (1991) parametric method was used. In parametric method land and climate characteristics are defined using different ratings. In this method impressive features in land suitability is ranked between a minimum and maximum value (usually between 0 and 100) according to Sys table. If a feature is so effective 100 and if it isn't effective zero will be assigned to that feature. These rankings are shown with A, B, C .....

To determine different characteristics and land indexes the following equation is used.

$$I = R\min \times \sqrt{\frac{A}{100}} \times \frac{B}{100} \times \frac{C}{100} \times ... \qquad \text{Eq. 6}$$

Where, $R_{\min}$ is a parameter with a minimum rank

And A, B, C …are parameters rank influencing the land suitability.

3. STUDY AREA

The study area, Shavur plain, lies in the Northern of Khuzestan province, Iran. It is located within coordinate of latitude 31˚37’30’’ and 32˚30’00’’ North and longitude 48˚15’00’’ and 48˚40’40’’ East with the area of 774 km^2 (Fig.1). Data used for the case study were consisting of: Topography, Wetness, Soil fertility, salinity and alkalinity and soil physical characteristics (Texture, Soil depth, CaCO3 and Gypsum in percent) which are extracted from the report of the land classification study (Ministry of Energy, 2006).

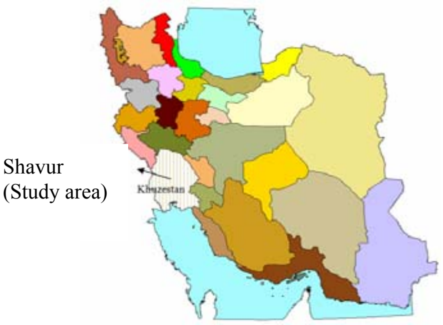


Fig.1. Location of the study area in Iran.

Climate data and those related to the stages of the plant growth were taken from Khuzestan Soil and Water Research Institute (2009) collected data and physiological requirements of the wheat plant were extracted from tables prepared specifically for Iran (Givi, 1997).

4. Results and Discussion

Fuzzy maps were prepared for each of the parameters are shown in Fig.5. AHP relies on Pair wise Comparison Matrices which are matrices relating different components and assigning values according to their relative importance. These values are given by a scale from 1 to 9, where 1 means that the two elements being compared have the same importance and 9 indicates that from the two elements one is extremely more important than the other. The table with the scale for Pair wise Comparison is shown in Table 1 (Saaty and Vargas 2001). As an example, pH has been considered more important than Slope and received a value of 5 when compared to it, while Slope when compared to pH received its reciprocal, 1/5. The final weight is the result of dividing each record value by the sum of the respective column and then calculating the average for the corresponding row. The results of Pair wise Comparison Matrix in the AHP method for preparation of the weights used for the overly of the Fuzzy maps are given in Table 2.

The classified land suitability evaluation based on the Fuzzy logic is shown in Fig. 2.

Table .1 Pair wise Comparison Matrix for Wheat Suitability according to Saaty

Parameters	CEC and ESP	Soil wetness	CaCO3	Gypsum	pH	Texture	Soil depth	Topography	Weight
CEC and ESP	1	2	3	4	5	6	7	8	0.3290
Soil wetness	1/2	1	2	3	4	5	6	7	0.2243
CaCO3	1/3	1/2	1	2	3	4	5	6	0.1526
Gypsum	1/4	1/3	1/2	1	2	3	4	5	0.1053
pH	1/5	1/4	1/3	1/2	1	2	3	4	0.0750
Texture	1/6	1/5	1/4	1/3	1/2	1	2	3	0.0525
Soil depth	1/7	1/6	1/5	1/4	1/3	1/2	1	2	0.0359
Topography	1/8	1/7	1/6	1/5	1/4	1/3	1/2	1	0.0254

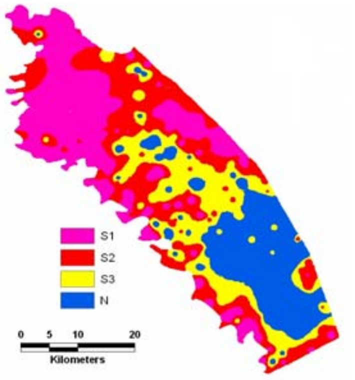


Fig.2. Classified land suitability map for wheat (Fuzzy method)

The results of the qualitative land suitability classes by using the guidelines given by Sys et al. (1993) in Eq.6 for wheat plant were determined and is given in Table 2 (the first 10 units are presented and the rest of the units are omitted from the Table) and the land suitability maps base on the parametric (FAO) method is shown in Fig.4.

Table 2. Samples results of the qualitative suitability evaluation of different land series for wheat using parametric method

Land units	Land index	Suitability classes
1	57.3	S2
2	38.6	S2
3	41.2	S2
4	56.8	S3
5	35.3	S3
6	14.1	S3
7	8.2	S2
8	52.2	S2
9	37	S3
10	21	N

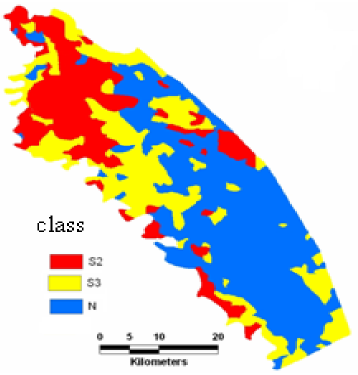


Fig.3. Land suitability map for wheat (FAO method)

As it is shown in fig3. The region is classified in to 3 classes: N,S2 and S3. There is no any instance of Class S1 because the features are discrete and higher weights which assigned to the limiting features in land suitability evaluation.To assess the agreement between the Fuzzy and the FAO methods, the Kappa statistic developed by Cohen (1960) was calculated. The Kappa coefficient is a measurement of the degree of agreement between two observations(maps) and its calculation is based on the difference between the tow maps. A Kappa value of 0 indicates that there is a poor agreement between the maps and a value of 1 indicates an almost perfect agreement. The value of Kappa coefficient for this study is calculated to be 0.28 between two maps (Fuzzy map and FAO) which shows a poor agreement between the two methods (maps) for the land suitability evaluation of Shavur plain. Fig. 4 shows the results of this comparison as a map.

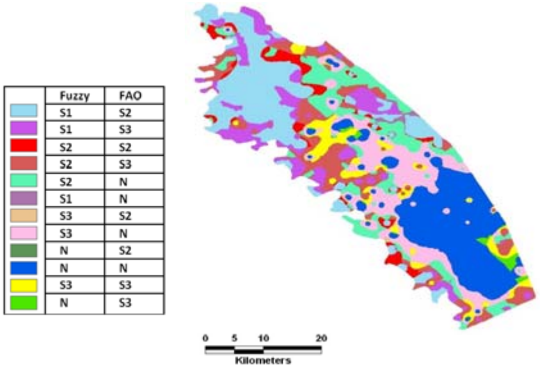


Fig.4. Comparison map showing correspondence between Fuzzy and FAO results.

The results of the FAO method show 26% of the land to be moderately suitable (S2 class), 25% as marginally suitable (S3 class) and 49% as not suitable (class N). In comparison, the results of the Fuzzy method show 31% of the land as highly suitable (S1 class) which the FAO method does not evaluate. Furthermore, the moderately suitable class for Fuzzy is 29% which is almost equivalent to the result of the FAO method. The class S3 (marginally suitable) is 19% and for class N (not suitable) is 21% for Fuzzy which they are quite different in compare with the FAO method results.

In order to evaluate and present the better method between these methods, five different cultivation fields were randomly chosen and the yields per hectare of the irrigated wheat were measured. The points are plotted on the prepared comparison map and are shown in Fig.6 and their information is given in Table 3. This Table shows the corresponding classes of the locations for different methods together with the production yield measured in the field. According to the Jihade Keshavarzi organization of the Khuzestan Province (The organization responsible for the agricultural affairs), the maximum, average and minimum yield for the wheat production in the Shavur plain are about less than 2, 3.5 and more than 5 tons/ha respectively.

The differences of the yield are due to the suitability of the soil and the categories of the land (Jihade Keshavarzi organization, 2009). Considering these figures we can consider the fields having yield of more than 4 tons/ha having soil class of suitability S1, between 3 and 4 tons/ha, S2, between 2 and 3 tons/ha, S3 and less than 2 tons/ha having class of N. Base on this consideration and the result of the measured yield of the field locations in the Fig.6 Fuzzy method is considered to be better than the FAO method.

Table.3 Information of the sampling points for comparison of the results of Fuzzy and FAO methods.

Location	X	Y	Class of Fuzzy	Class of FAO	yield tons/ha
1	249476	3539574	S1	S2	4.82
2	255364	3527325	S2	S3	3.93
3	256693	3538530	S2	N	3.47
4	260017	3535966	S1	S3	4.86
5	263245	3524476	S3	N	2.12

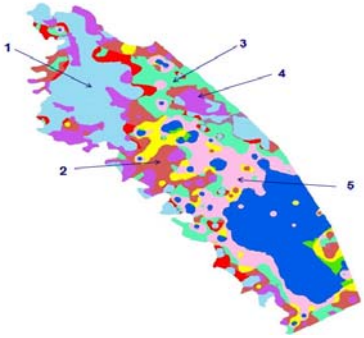


Fig.6. Sampeling locations for comparison of the Fuzzy and FAO methods (Table 3, shows the information of the points).

CONCLUSION

Since the soil properties have contineouse spatial change, Fuzzy method which is based on the continouse ahanges of the parameters used in the evaluation of the soil sutability can classify the soil better than FAO methd. This is proved by the field observation and the agreement with the work of Sanchez Moreno (2007).

REFERENCES

Brail, R.K., Klosterman, R.E., 2001. Planning Support Systems, ESRI Press, Redlands, CA.

Burrough, P. A. 1989. "Fuzzy Mathematical Methods for Soil Survey and Land Evaluation." Journal of Soil Science 40: 477-492.

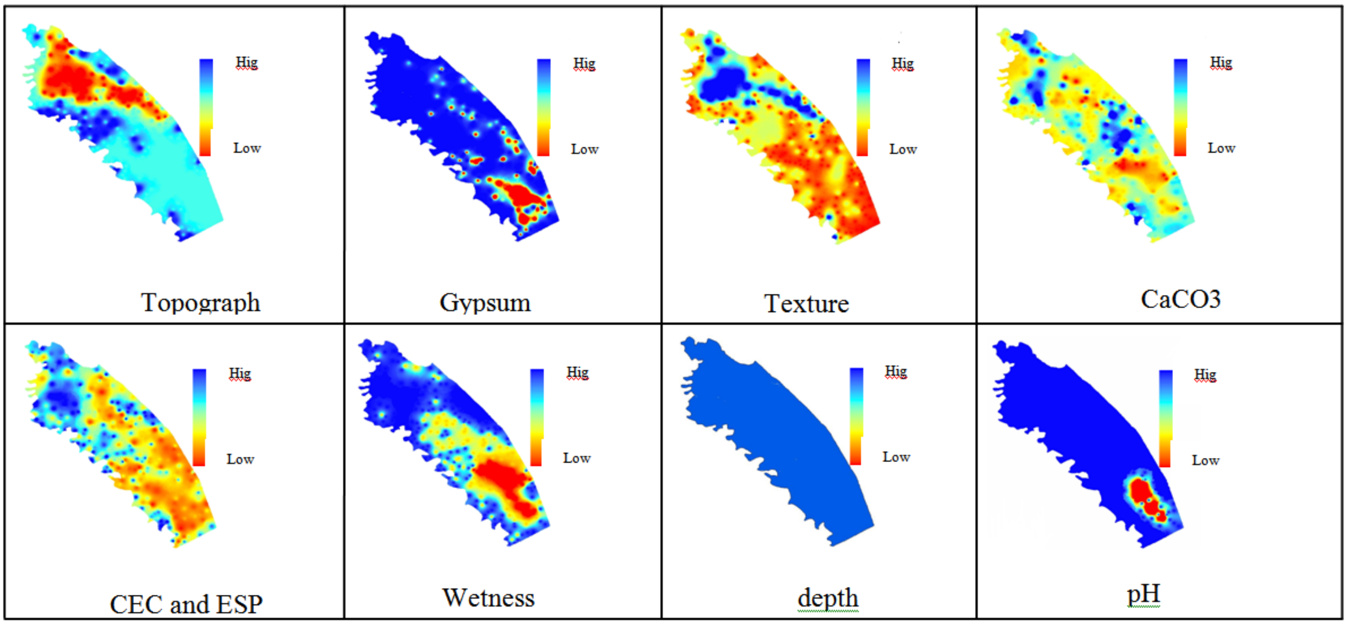


Fig.5. Fuzzy maps for each of parametrs

Cambell, J.C., Radke, J., Gless, J.T., Whirtshafter, R.M., 1992. An application of linear programming and geographic information systems: cropland allocation in antigue. Environment and Planning A 24, 535–549.

Church, R.L., 2002. Geographical information systems and location science. Computers and Operations Research 29 (6), 541–562.

Cohen, J. 1960. "coefficient of agreement for nominal scales." Educational and Psychological Measurement 20 (1): 37-46.

Collins, M.G., Steiner, F.R., Rushman, M.J., 2001. Land-use suitability analysis in the United States: historical development and promising technological achievements. Environmental Management 28 (5), 611–621.

Collins, M.G., Steiner, F.R., Rushman, M.J., 2001. Land-use suitability analysis in the United States: historical development and promising technological achievements. Environmental Management 28 (5), 611–621.

Food and Agriculture Organization of the United Nations FAO, 1976. A framework for land evaluation. Rome, FAO.

Hall, G.B., Wang, F., Subaryono, 1992. Comparison of Boolean and Fuzzy classification methods in land suitability analysis by using geographical information systems. Environment and Planning A 24, 497–516.

Hobbs, B.F., 1980. A comparison of weighting methods in power plant siting. Decision Sciences 11, 725–737.

Jankowski, P., Andrienko, N., Andrienko, G., 2001. Map-centered exploratory approach to multiple criteria spatial decision making. International Journal of Geographical Information Science 15 (2), 101–127.

Jihade Keshavarzi organization, 2009. (Personal communication).

Kalogirou, S., 2002. Expert systems and GIS: an application of land suitability evaluation. Computers, Environment and Urban Systems 26 (2–3), 89–112.

Khuzestan Soil and Water Research Institute, 2009. (Personal communication).

Kontos TD, Komilis DP, Halvadakis CP, 2003. Siting MSW landfills on Lesvos island with a GIS-based methodology. Waste Manag Res 21:262–278.

Kontos TD, Komilis DP, Halvadakis CP, 2003. Siting MSW landfills on Lesvos island with a GIS-based methodology. Waste Manag Res 21:262–278.

Ligtenberg, A., Bregt, A.K., van Lammeren, R., 2001. Multi-actor-based land use modelling: spatial planning using agents. Landscape and Urban Planning 56 (1–2), 21–33.

Lukasheh AF, Droste RL, Warith MA, 2001. Review of expert system (ES), geographical information system (GIS), decision support system (DSS) and their application in landfill design and management. Waste Manag Res 19:177–185.

Malczewski, J., 1999. GIS and Multicriteria Decision Analysis, Wiley, New York.

McBratney A. B. and Odeh I. O. A. 1997. "Application of Fuzzy sets in soil science: Fuzzy logic, Fuzzy measurements and Fuzzy decisions." Geoderma 77: 85-113.

McHarg, I.L., 1969. Design With Nature, Wiley, New York.

Miller, W., Collins, W.M.G., Steiner, F.R., Cook, E., 1998. An approach for greenway suitability analysis. Landscape and Urban Planning 42 (2–4), 91–105.

Ministry of Energy (2006). The comprehensive study of the Shavur plain soil (by Tak- e- Sabz consuting Engineers) (in Farsi).

Moreno, D., Seigel,M., 1988. A GIS approach for corridor siting and environmental impact analysis. GIS/LIS'88. Proceedings from the third annual international conference, San Antonio, Texas 2, 507–514.

O'Sullivan, D., Unwin, D.J., 2003. Geographic Information Analysis, Wiley, Hoboken, NJ.

Rossiter, D. G,. 2004. Technical Note: Statistical methods for accuracy assesment of classified thematic maps. Enschede, the Netherlands. Available: http://www.itc.nl/personal/rossiter/pubs/list.html#pubs_m_R. Accessed January 25, 2007, ITC: 46.

Saaty TL, 1977 A scaling method for priorities in hierarchical structures. J Math Psychol 15:234–281.

Saaty, T. L. and L. G. Vargas, 2001. Models, methods, concepts, and applications of the analytic hierarchy process. Boston etc., Kluwer Academic.

Sanchez Moreno, J.F. 2007. Applicability of knowledge-based and Fuzzy theory-oriented approaches to land suitability for upland rice and rubber, as compared to the farmers' perception. International Institute for Geo-Information Science and Earth Observation, Enschede, the Netherlands. 133 pp.

Sener B, Suzen ML, Doyuran V, 2006. Landfill site selection by using geographic information systems. Environ Geol 49:376–388.

Siddiqui M, Everett JM, Vieux BE, 1996. Landfill siting using geographical information systems: a demonstration. J Environ Eng 122:515–523.

Store, R., Kangas, J., 2001. Integrating spatial multi-criteria evaluation and expert knowledge for GIS-based habitat suitability modelling. Landscape and Urban Planning 55 (2), 79–93.

Sys, C., van Ranst, E., Debaveye, J., 1993 Land Evaluation, part Ш : crop requirements. International Training Center for post graduate soil scientists. Ghent university, Ghent. 199 P.

Xiao, N., Bennett, D.A., Armstrong, M.P., 2002. Using evolutionary algorithms to generate alternatives for multiobjective site-search problems. Environment and Planning A 34 (4), 639–656.

Zadeh, L.H., 1965. Fuzzy sets. Information and Control 8, 338–353.

KNOWLEDGE DISCOVERY FROM MINING THE ASSOCIATION BETWEEN H5N1 OUTBREAKS AND ENVIRONMENTAL FACTORS

Y. L. Si [a, b*], T. J. Wang [a], A. K. Skidmore [a], H. H. T. Prins [b]

[a] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7500AA, Enschede, The Netherlands – (yali, tiejun, skidmore)@itc.nl
[b] Resource Ecology Group, Wageningen University, 6708 PB Wageningen, The Netherlands – herbert.prins@wur.nl

Commission II

KEY WORDS: H5N1, wild birds, environmental factors, GIS

ABSTRACT:

The global spread of highly pathogenic avian influenza H5N1 in poultry, wild birds and humans, poses a significant panzootic threat and a serious public health risk. An efficient surveillance and disease control system requires a deep understanding of their spread mechanisms, including environmental factors responsible for the outbreak of the disease. Previous studies suggested that H5N1 viruses occurred under specific environmental circumstances in Asia and Africa. These studies were mainly derived from poultry outbreaks. In Europe, a large number of wild bird outbreaks were reported in west Europe with few or no poultry infections nearby. This distinct outbreak pattern in relation to environmental characteristics, however, has not yet been explored. This research demonstrated the use of logistic regression analyses to examine quantitative associations between anthropogenic and physical environmental factors, and the wild bird H5N1outbreaks in Europe. A geographic information system is used to visualize and analyze the data. Our results indicate that the H5N1 outbreaks occur in wild birds in Europe under predictable environmental conditions, which are highly correlated with increased NDVI in December, decreased aspect and slope, increased minimum temperature in October and decreased precipitation in January. It suggests that H5N1 outbreaks in wild birds are strongly influenced by food resource availability and facilitated by the increased temperature and the decreased precipitation. We therefore deduce that the H5N1 outbreaks in wild birds in Europe may be mainly caused by contact with wild birds. These findings are of great importance for global surveillance of H5N1 outbreaks in wild birds.

1. INTRODUCTION

The global spread of highly pathogenic avian influenza (HPAI) H5N1 in poultry, wild birds and humans, poses a significant panzootic threat and a serious public health risk. Since July 2005, the H5N1 virus was detected outside of Asia, appeared in Russia, and then arrived in Romania in October (Gilbert et al. 2006a). In 2006, the disease became pandemic around the Black Sea region, the Mediterranean region, Western Europe and Africa. Recent studies (Kilpatrick et al. 2006;Si et al. 2009) suggested that both wild birds and domestic poultry have played roles in the global dispersion of the H5N1 virus. An efficient surveillance and disease control system requires more understanding of their spread mechanisms, including environmental factors responsible for the outbreak of the disease. In regard to the association with environment, the movement of wild birds generally depends on the availability of food and shelter, which is influenced by physical environmental factors. The movement of domestic poultry mainly relies on anthropogenic environmental factors, which is related to human activities.

Few efforts have been made to investigate the influence of environmental factors on H5N1 outbreaks. The transmission of the HPAI H5N1in Nigeria and West Africa has been linked to differences in plant phenology and land-surface reflectance (Williams et al. 2008). In Southeast Asia the predictable factor of H5N1 activity was identified based on free-range duck farming and rice-paddy cultivation (Gilbert et al. 2008). And in mainland China, the minimal distance to the national highway, precipitation and the interaction between minimal distance to the nearest lake and wetland are important predictive variables for the risk of H5N1 (Fang et al. 2008). While the land-use pattern, occurrence of seasonal wetlands, practices of backyard poultry and animal husbandry, and density of human population were identified as risk factors contributed to the transmission of the H5N1 virus in Indian subcontinent (Adhikari et al. 2009). These studies, mainly based on poultry outbreaks, revealed that both anthropogenic and physical environmental factors influence the disease transmission, suggesting that wild birds may facilitate the H5N1 spread in poultry. However, the environmental factors influencing the outbreak of H5N1 in wild birds are poorly understood. Wild bird H5N1 outbreaks may be caused by contact with infected wild birds if the disease pattern is highly correlated with physical environmental factors. Otherwise wild birds may be victims as a result of contact with infected domestic poultry if the disease pattern is more related to anthropogenic environmental factors. In this regard, the detection of key risk factors influencing disease outbreaks in wild birds may help to understand to what extent wild birds can be regarded as the major cause of the H5N1 outbreaks in wild birds.

Surveillance in wild birds tend to be more challenging than in poultry as the population and the movement of wild birds are not clearly known. Derivation of the key environmental indicators influencing the H5N1 spread in wild birds is of great importance for H5N1 surveillance and control. Different from Asia and Africa, in Europe large number of H5N1 outbreaks in wild birds were reported, and with few or no infections in

[*] Corresponding author

poultry nearby (Si et al. 2009). This distinct pattern provides an opportunity to explore the underlying spread mechanism and the key environmental factors for the disease outbreak.

This study, carried out at the European scale, aims to identify the key risk factors contributing to the outbreak of H5N1 in wild birds. The study will contributes to the knowledge about the spread mechanism of H5N1 virus in wild birds and help in taking necessary precautionary measures against any future disease outbreaks.

2. METHODS AND MATERIALS

2.1 Data

Data on avian influenza H5N1 outbreaks in wild birds in Europe, consisted of 808 confirmed events reported form July 18, 2005 to October 9, 2008, were provided by EMPRES-i: a global animal health information system of FAO's Emergency Prevention Programme for Transboundary Animal Diseases (http://empres-i.fao.org/empres-i/home). Environmental data were categorized into anthropogenic and physical environmental subsets, corresponding to two disease spread agents (i.e. poultry and wild birds). Table 1 shows the environmental data sets used for this study. The anthropogenic environmental data sets include urban areas, cities, metropolis, roads, major roads and railways, population density, poultry density. The physical environmental data sets involve DEM, Global Lakes and Wetlands Database, Ramsar sites, potential evapotranspiration, aridity index, precipitation, minimum and maximum temperature and remotely sensed data layers (i.e., monthly NDVI) from 2005 to 2008. Poultry density was classified into anthropogenic environmental factors as it is closely related to human activities. The topographic, wetlands, climatic and NDVI data were selected to stand for physical environmental factors because these conditions are closely related the availability of food and shelters for wild birds.

2.2 Data pre-processing

The duplicate occurrences of the wild bird H5N1 outbreaks from same localities were discarded, resulting into 320 unique geographic coordinates. A total of 296 locations in 2005 and 2006 were assigned as H5N1 presence data for training the model, and additional 24 locations in 2007 and 2008 were utilized for the validation of the resulting output. To facilitate the use of the logistic regression model, 592 pseudo absent points (i.e. two times the H5N1 presence) were randomly generated within the maximum geographic ranges of H5N1 outbreaks from 2005 to 2008, with a minimum distance of 10 km which equals to the size of the surveillance area in Europe (Pittman and Laddomada 2008). All absence of H5N1 was located outside the 10 km radius buffer of the presence of H5N1 during this period. Figure 1 shows the distribution of presence and pseudo absence of H5N1 outbreaks in wild birds from July 2005 to December 2006.

GIS layers detailing the distance to the nearest urban area, city, metropolis, road, major road, railway, lake and wetland and Ramsar site were generated with a spatial resolution of 1 km. The minimum distances were then extracted from the distance layers for all presence or absence locations. An aspect and a slope layer were also generated based on DEM. Using a smooth algorithm in TIMESAT (Jonsson and Eklundh 2004), the monthly NDVI of 2006 were calculated based on monthly NDVI layers from 2005 to 2008. Furthermore, using a zonal statistical calculation method, a 10 km mean buffer zone (equals to the size of the surveillance area in Europe) was calculated for the variables: population density, poultry density, DEM, aspect, slope, potential evapotranspiration, aridity index, precipitation, minimum temperature, maximum temperature and NDVI. A total of 63 environmental variables were generated for this study, comprising 10 anthropogenic environmental variables and 53 physical environmental variables respectively (Table 2).

2.3 Statistical analysis

Statistical analyses were performed using the Statistical Package (SPSS Inc, Chicago, IL, USA). Univariate analyses were conducted to examine the effect of each variable separately, reporting odds ratios (OR), 95% confidence intervals (CIs) and *P*-value in the output. Variables yielding nonsignificant changes in log-likelihood were excluded from the further analysis. The multicollinearity and autocorrelation were as well assessed by examining the Variance Inflation Factor (VIF) and Moran’s I respectively. Collinearity is present when VIF for at least one independent variable is large. A smaller negative value of Moran’s I indicate an increased negative autocorrelation and a smaller positive value reveals an decreased positive autocorrelation; no autocorrelation was detected when the value equals to zero. A subset of variables for identifying key environmental factors were then prepared base on the concern of odds ratios, collinearity and autocorrelation. A stepwise multiple logistic regression with forward entry mode was carried out using the subset of variables and entering each variable accounting for the highest change in the model log-likelihood. This procedure was repeated until no additional significant variable could be added, using a decision rule of P < 0.01 for entry and P > 0.05 for removal. And odds ratios, 95% confidence intervals (CIs) and *P*-value were reported in the output of the multivariate analysis. The performances of the models were assessed by the Hosmer-Lemeshow goodness of fit test.

Based on the predictive model derived from the multivariate logistic regression analysis, a risk map of wild bird H5N1 outbreaks was generated for a spatial resolution of 1 km. The risk map was then classified into four levels (i.e., very high, high, medium and low risk) for the validation using independent samples (i.e., H5N1 outbreaks in wild birds from 2007 to 2008). A 10 km buffer zone was generated for each validating occurrence and the mean risk value under each buffer zone was calculated. The percentage of buffer zones occurring on the high risk areas was then calculated to evaluate the accuracy of the predictive map.

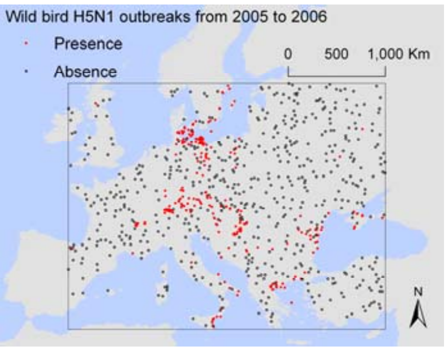


Figure 1. Distribution of presence and pseudo absence of H5N1 outbreaks in wild birds from 2005 to 2006 in Europe

Category	Environmental data sets	Format	Resolution	Data Producer & Source
Anthropogenic environmental data	Urban areas	Polygon	-	ESRI
	Cities	Polygon	-	ESRI
	Metropolis	Polygon	-	ESRI
	Roads	Polyline	-	ESRI
	Major roads	Polyline	-	ESRI
	Railways	Polyline	-	ESRI
	Population density in 2005	Raster	5 km	CIESIN, FAO, CIAT
	Poultry density in 2005	Raster	6 km	FAO
Natural environmental data	DEM	Raster	1 km	WORLDCLIM
	Global Lakes and Wetlands Database	Raster	0.5 km	WWF,ESRI, CESR
	Ramsar sites	Point	-	Wetlands International
	Mean annual potential evapotranspiration	Raster	1 km	CGIAR-CSI
	Mean annual aridity index	Raster	1 km	CGIAR-CSI
	Mean monthly precipitation	Raster	1 km	WORLDCLIM
	Mean monthly maximum temperature	Raster	1 km	WORLDCLIM
	Mean monthly maximum temperature	Raster	1 km	WORLDCLIM
	Monthly MODIS NDVI	Raster	1 km	NASA

Table 1. Environmental data sets used in generating variables for the analysis of H5N1 outbreaks in wild birds in Europe

Category	Environmental variables	Abbreviation	Unit
Anthropogenic environmental variables	Distance to the nearest urban area	Urban	km
	Distance to the nearest city	City	km
	Distance to the nearest metropolis	Metro	km
	Distance to the nearest road	Road	km
	Distance to the nearest major road	Mjroad	km
	Distance to the nearest railway	Railway	km
	Population density in 2005	Popden	p/km^2
	Poultry density in 2005	Poultryden	p/km^2
Natural environmental Variables	Distance to the nearest lake and wetland	GLWD	km
	Distance to the nearest Ramsar site	Ramsar	km
	DEM	DEM	m
	Aspect	Aspect	degree
	Slope	Slope	degree
	Mean annual potential evapotranspiration	Mapet	mm/km^2/year
	Mean annual aridity index	Maaridity	-
	Mean monthly precipitation	PrecJan to Dec	mm
	Mean monthly minimum temperature	TminJan to Dec	degree*10
	Mean monthly maximum temperature	TmaxJan to Dec	degree*10
	Monthly MODIS NDVI	NDVIJan to Dec	-

Table 2. Summary of environmental variables used in the analysis of H5N1 outbreaks in wild birds in Europe

3. RESULTS

The univariate analysis demonstrated that part of physical environmental variables significantly affect the spread of H5N1 in wild birds (Table 3), while no anthropogenic variables were significantly associated with H5N1 outbreaks in wild birds. Positive association were found with the minimum temperature variables, cold-season maximum temperature variables and winter NDVI variables. Negative association were detected with DEM, aspect, slope and most precipitation variables. Among the associated environmental variables for the H5N1outbreaks in wild birds, NDVI in December revealed the strongest influence, with an odds ratio of 12.915 reported.

Seven physical environmental variables (i.e., aspect, slope, mean monthly precipitation in January, mean monthly minimum temperature in February and October, monthly MODIS NDVI in March and December) were selected as the inputs of the multivariate stepwise logistic regression. All selected variables showed a low to medium degree of autocorrelation and the VIF values are lower than 10 (Table 4). Multivariate logistic regression demonstrated that five variables, aspect, slope, mean monthly minimum temperature in October, mean monthly precipitation in January and monthly NDVI in December, were significantly associated with H5N1 outbreaks in wild birds (Table 4). H5N1 outbreaks in wild birds tend to occur in areas with increased NDVI in December, decreased aspect and slope, increased minimum temperature in October and decreased precipitation in January. The regression model adequately fits the data, evaluated by the Hosmer and Lemeshow goodness test (X^2 = 8.596, P = 0.378).

Variables	Univariate analysis		
	OR	95% CIs	*P*-value
DEM	0.998	0.998 ~ 0.999	<0.001
Aspect	0.99	0.985 ~ 0.995	<0.001
Slope	0.843	0.778 ~ 0.913	<0.001
PrecJan	0.988	0.982 ~ 0.995	0.001
PrecFeb	0.988	0.980 ~ 0.996	0.002
PrecMar	0.99	0.982 ~ 0.998	0.016
PrecApr	0.987	0.978 ~ 0.997	0.009
PrecAug	1.006	1.001 ~ 1.011	0.031
PrecOct	0.992	0.985 ~ 0.998	0.014
PrecDec	0.989	0.983 ~ 0.996	0.001
TminJan	1.012	1.009 ~ 1.016	<0.001
TminFeb	1.012	1.008 ~ 1.016	<0.001
TminMar	1.016	1.011 ~ 1.021	<0.001
TminApr	1.016	1.010 ~1.023	<0.001
TminMay	1.015	1.009 ~ 1.022	<0.001
TminJun	1.018	1.011 ~ 1.024	<0.001
TminJul	1.016	1.010 ~ 1.022	<0.001
TminAug	1.017	1.011 ~ 1.023	<0.001
TminSep	1.019	1.013 ~ 1.025	<0.001
TminOct	1.019	1.013 ~ 1.025	<0.001
TminNov	1.016	1.011 ~ 1.021	<0.001
TminDec	1.014	1.010 ~ 1.019	<0.001
TmaxJan	1.009	1.005 ~ 1.012	<0.001
TmaxFeb	1.009	1.005 ~ 1.012	<0.001
TmaxMar	1.01	1.006 ~ 1.014	<0.001
TmaxApr	1.006	1.001 ~ 1.011	0.013
TmaxSep	1.006	1.002 ~ 1.011	0.003
TmaxOct	1.008	1.004 ~ 1.012	<0.001
TmaxNov	1.008	1.004 ~ 1.012	<0.001
TmaxDec	1.009	1.005 ~ 1.012	<0.001
NDVIJan	5.599	1.958 ~ 16.010	0.001
NDVIFeb	5.437	1.897 ~ 15.585	0.002
NDVIMar	6.415	1.962 ~ 21.210	0.002
NDVINov	10.243	2.595 ~ 40.430	<0.001
NDVIDec	12.915	3.576 ~ 46.646	<0.001

Table 3 Significant variables associated with H5N1 outbreaks in wild birds reported by the univariate logistic regression analysis

A predictive risk map of H5N1 presence in wild birds in Europe was generated based on the predictive model derived from the logistic regression analysis. The validating samples were overlapped on the risk map, showing 79% of H5N1 outbreaks in wild birds (19/24) appeared in the predictive high and very high risk areas (i.e., predictive risk > 0.4) (Figure 2).

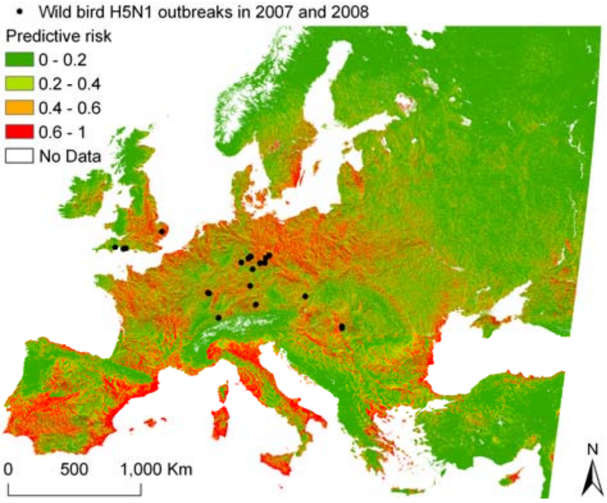


Figure 2. Predictive risk map of H5N1 outbreaks in wild birds in Europe

4. DISCUSSION

The results presented in this paper highlight two main findings: (i) H5N1 outbreaks in wild birds in Europe occur under consistent and predictable environmental circumstances. The key environmental factors influencing the presence of H5N1 outbreaks in wild birds are an increased NDVI in December, decreased aspect and slope (i.e., flat areas with low relief), increased minimum temperature in October and decreased precipitation in January. (ii) H5N1 outbreaks in wild birds in Europe may be mainly caused by contact with infected wild birds as the spatial distribution of wild bird H5N1 outbreaks strongly correspond to most physical environmental factors, but wild bird H5N1 outbreaks are not related to any anthropogenic environmental factors.

The slope and aspect showed consistently negative association with wild bird H5N1 outbreaks in Europe. And consistently positive association with NDVI and temperature during the winter was found, when large number of wild birds overwintering and staging in Europe. As wetlands, rivers, canals, ponds and irrigated networks are concentrated in lowlands, flat plains, deltas, and coastal areas (Gilbert et al. 2006b; Gilbert et al. 2008), the combination of increased winter NDVI and increased winter temperature with flat areas indicates a circumstance of increased food resources for wild birds. The availability of food may influence the movement of wild birds and hence influence the H5N1 outbreaks in wild birds. However, the risk of H5N1 infections is not increased in areas closer to wetlands. One explanation for this result may be the extensive distribution of wetlands in Europe.

This study also highlights that climatic factors significantly contribute to H5N1 outbreaks in wild birds. In agreement with a recent finding in mainland China (Fang et al. 2008), precipitation was found here to be negatively associated with the risk of H5N1 outbreaks in wild birds, possibly because the lower precipitation leads to a higher concentration of birds in limited suitable habitats, and therefore increased opportunities for contact. Areas with higher temperatures tend to have higher risk of disease outbreaks because temperature can stimulate the viral activity. This explains why both monthly minimum and maximum temperature showed positive associations with H5N1 outbreaks.

The univariate logistic analysis in this study demonstrated that the risk of H5N1 infections in wild birds in Europe is not influenced by anthropogenic environmental factors but only by physical environmental factors, indicating that these outbreaks may be mainly caused by wild birds contact with infected wild birds. This is supported by the previous findings that wild birds are capable of excreting abundant viruses (e.g., in their faeces) before and after the onset of clinical signs (Keawcharoen et al. 2008), or even asymptomatically (Chen et al. 2006), and they are suspected to spread the virus over either long and short distances (Si et al. 2009).

Based on the five principal environmental factors identified by multivariate logistic analysis, a predictive risk map was generated, targeting important surveillance areas in Europe. Independent validation of the risk map showed a moderate accuracy of the predictive model, suggesting the key environmental factors identified are consistently affecting the outbreak of H5N1 in wild birds (Peterson and Williams 2008). As temperature and precipitation data are available and of good quality, improved food availability data could be used to enhance the accuracy of risk prediction on H5N1 outbreaks in wild birds in Europe.

	Multivariate analysis				Moran I	
Selected	B	Odds ratio	95% CIs	*P*-value	*P*<0.01	VIF
Aspect	-0.010	0.991	0.985 ~ 0.996	0.001	0.090	1.307
Slope	-0.140	0.869	0.780 ~ 0.969	0.011	0.350	1.502
PrecJan	-0.036	0.965	0.954 ~ 0.976	<0.001	0.360	1.822
TminOct	0.031	1.032	1.024 ~ 1.039	<0.001	0.470	4.272
NDVIDec	5.098	163.765	30.264 ~ 886.171	<0.001	0.510	1.799
Constant	-1.683	0.186	-	0.006	-	-

Table 4 Selected variables by multivariate stepwise logistic regression model for H5N1 outbreaks in wild birds in Europe from 2005 to 2006. Moran's I and VIF indicate lower level of autocorrelations and collinearity among variables

5. CONCLUSIONS

The key environmental factors influencing the presence of H5N1 outbreaks in wild birds in Europe are increased NDVI in December, decreased aspect and slope, increased minimum temperature in October and decreased precipitation in January. It suggests that H5N1 outbreaks in wild birds are strongly influenced by the availability of food resources, and facilitated by the increased temperature and the decreased precipitation. We therefore deduce that H5N1 outbreaks in wild birds in Europe may be mainly caused by contact with infected wild birds. These findings are of great importance for global surveillance of H5N1 outbreaks in wild birds.

REFERENCE

Adhikari, D., Chettri, A. and Barik, S. K. 2009. Modelling the ecology and distribution of highly pathogenic avian influenza (H5N1) in the Indian subcontinent. *CURRENT SCIENCE*, 97(1), pp. 73-78.

Chen, H., Li, Y., Li, Z., Shi, J., Shinya, K., Deng, G., Qi, Q., Tian, G., Fan, S., Zhao, H., Sun, Y. and Kawaoka, Y. 2006. Properties and Dissemination of H5N1 Viruses Isolated during an Influenza Outbreak in Migratory Waterfowl in Western China. *J Virol*, 80(12), pp. 5976–5983.

Fang, L.-Q., de Vlas, S. J., Liang, S., Looman, C. W. N., Gong, P., Xu, B., Yan, L., Yang, H., Richardus, J. H. and Cao, W.-C. 2008. Environmental Factors Contributing to the Spread of H5N1 Avian Influenza in Mainland China. *PLoS ONE*, 3(5), pp. e2268.

Gilbert, M., Chaitaweesub, P., Parakamawongsa, T., Premashthira, S., Tiensin, T., Kalpravidh, W., Wagner, H. and Slingenbergh, J. 2006b. Free-grazing Ducks and Highly Pathogenic Avian Influenza, Thailand. *Emerg Infect Dis*, 12(2), pp. 227-234.

Gilbert, M., Xiao, X., Domenech, J., Lubroth, J., Martin, V. and Slingenbergh, J. 2006a. Anatidae Migration in the Western Palearctic and Spread of Highly Pathogenic Avian Influenza H5N1 Virus. *Emerg Infect Dis* 12(11), pp. 1650-1656.

Gilbert, M., Xiao, X., Pfeiffer, D. U., Epprecht, M., Boles, S., Czarnecki, C., Chaitaweesub, P., Kalpravidh, W., Minh, P. Q., Otte, M. J., Martin, V. and Slingenbergh, J. 2008. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *PNAS*, 105(12), pp. 4769-4774.

Jonsson, P. and Eklundh, L. 2004. TIMESAT – a program for analyzing time-series of satellite sensor data. *COMPUT GEOSCI*, 30, pp. 833-845.

Keawcharoen, J., van Riel, D., Amerongen, G. v., Bestebroer, T., Beyer, W. E., van Lavieren, R., Osterhaus, A. D. M. E., Fouchier, R. A. M. and Kuiken, T. 2008. Wild Ducks as Long-Distance Vectors of Highly Pathogenic Avian Infl uenza Virus (H5N1). *Emerg Infect Dis*, 14(4), pp. 600-607.

Kilpatrick, A. M., Chmura, A. A., Gibbons, D. W., Fleischer, R. C., Marra, P. P. and Daszak, P. 2006. Predicting the global spread of H5N1 avian influenza. *PNAS*, 103(51), pp. 19368-19373.

Peterson, A. T. and Williams, R. A. J. 2008. Risk Mapping of Highly Pathogenic Avian Influenza Distribution and Spread. *Conserv Ecol*, 13(2), pp. 15.

Pittman, M. and Laddomada, A. 2008. Legislation for the Control of Avian Influenza in the European Union. *Zoonoses and Public Health*, 55, pp. 29-36.

Si, Y., Skidmore, A. K., Wang, T., de Boer, W. F., Debba, P., Toxopeus, A. G., Li, L. and Prins, H. H. T. 2009. Spatio-temporal dynamics of global H5N1 outbreaks match bird migration patterns. *Geospat Health*, 4(1), pp. 65-78.

Williams, R. A. J., Fasina, F. O. and Peterson, A. T. 2008. Predictable ecology and geography of avian influenza(H5N1) transmission in Nigeria and West Africa. *Trans R Soc Trop Med Hyg* 102, pp. 471-479.

ACKNOWLEDGEMENTS

This research is sponsored by the China Scholarship Council (CSC) and the International Institute for Geo-Information Science and Earth Observation (ITC), the Netherlands. The authors thank FAO's Emergency Prevention Programme for Transboundary Animal Diseases (EMPRES-i) for providing the H5N1 outbreak data.

SPATIAL OBJECT RECOGNITION VIA INTEGRATION OF DISCRETE WAVELET DENOISING AND NONLINEAR SEGMENTATION

Zhengmao Ye, Habib Mohamadian

College of Engineering, Southern University
Baton Rouge, LA 70813, USA
zhengmaoye@engr.subr.edu, mohamad@engr.subr.edu

KEY WORDS: Spatial Imaging; Pattern Recognition; Wavelet Denoising; Nonlinear Segmentation; Entropy; Mutual Information

ABSTRACT:

Spatial digital image analysis plays an important role in the information decision support systems, especially for regions frequently being affected by hurricanes and tropical storms. For the aerial and satellite imaging based pattern recognition, it is unavoidable that these images are affected by various uncertainties, like the atmosphere medium dispersing. Image denoising is thus necessary to remove noises and retain important signatures of digital images. The linear denoising approach is suitable for slowly varying noise cases. However, the spatial object recognition problem is essentially nonlinear. Being a nonlinear wavelet based technique, wavelet decomposition is effective to denoise blurring spatial images. The digital image can be split into four subbands, representing approximation (low frequency feature) and three details (high frequency features) in horizontal, vertical and diagonal directions. The proposed soft thresholding wavelet decomposition is simple and efficient for noise reduction. To further identify the individual targets, nonlinear K-means clustering based segmentation approach is proposed for image object recognition. The selected spatial images are taken across hurricane affected Louisiana areas. In addition to evaluate this integration approach via qualitative observation, quantitative measures are proposed on a basis of the information theory, where the discrete entropy, discrete energy and mutual information, are applied for the accurate decision support.

1. INTRODUCTION

Spatial image processing has many potential applications in the fields of ground surveillance, weather forecasting, target detection, environmental exploration, and so on. The remote taken images will be affected by various factors, such as atmospheric dispersions and weather conditions, thus spatial images contain diverse types of noises, both slowly varying or rapidly varying ones. Discrete wavelet denoising can be designed to eliminate noises presented in images so as to preserve the characteristics across all frequency ranges. It involves three steps, that is, linear wavelet transform, nonlinear thresholding and linear inverse wavelet transform. Using discrete wavelet transform, a digital image can be decomposed into the approximation component and detail components (horizontal, vertical, diagonal). The approximation component will be further decomposed. Information loss between two successive decomposition levels of approximations will be represented in detail coefficients [1-3, 5-6]. The essence of fractal-based denoising in the wavelet domain has been used to predict the fractal code of a noiseless image from its noisy observation. The cycle spinning is incorporated into these fractal-based methods to produce enhanced estimations for the denoised images [7]. The new image denoising method based on Wiener filtering for soft thresholding has been proposed. It shows a high and stable SNR (signal to noise ratio) gain for all noise models used. This process leads to an improvement of phase images when real and imaginary parts of wavelet packet coefficients are filtered independently [8]. Two techniques for spatial video denoising using wavelet transform are used: discrete wavelet transform and dual-tree complex wavelet transform. An intelligent denoising system is introduced to make a tradeoff between the video quality and the time required for denoising. The system is suitable for real-time applications [9].

Image segmentation is a main step towards automated object recognition systems. The quality of spatial images is directly affected by atmospheric medium dispersion, pressure and temperature. It emphasizes necessity of image segmentation, which divides an image into parts that have strong correlations with objects to reflect the actual information being collected [1-3]. Spatial information enhances quality of clustering. In general, fuzzy K-means algorithm is not used for color image segmentation and not robust against noise. In this case, integration of discrete wavelet denoising and nonlinear K-means segmentation provides a suitable solution. Spatial information can be incorporated into the membership function for clustering of color images. For optimal clustering, gray level images are used. The spatial function is the summation of the membership function in the neighborhood of each pixel under consideration. It yields more homogeneous outcomes with less noisy spots. Image segmentation refers to the process of partitioning a digital image into multiple regions. Each pixel in a region is similar with respect to specific characteristic, like color, brightness, intensity or texture. [10-12]. To minimize the effects from medium dispersing, K-means clustering is critical for image processing. It is used to accumulate pixels with similarities together to form a set of coherent image layers. For K-means clustering, optimization can be implemented via the control algorithms such as the nearest neighbor rule or winner-take-all scheme. Nonlinear K-means clustering is presented here for image segmentation [10-14].

To objectively measure the impact of technology integration of image denoising and image segmentation, metrics of the discrete entropy, discrete energy, relative entropy and mutual information can be introduced to evaluate all the measuring outcomes of image processing integration [4].

2. DISCRETE WAVELET TRANSFORM

Two spatial source images were taken in State of Louisiana regions, which are frequently affected by hurricanes. The first image shows the spatial view of New Orleans and the second image shows the spatial view of Baton Rouge. The source images are contaminated by noises. The objective is to identify diverse types of targets involved. Image processing technology integration is proposed, where the nonlinear wavelet denoising is applied at first and nonlinear K-means clustering is used for target identification.

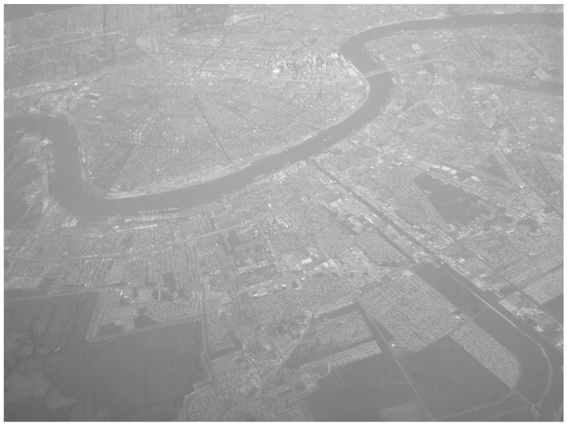


Fig.1 Source Spatial Image of New Orleans Areas

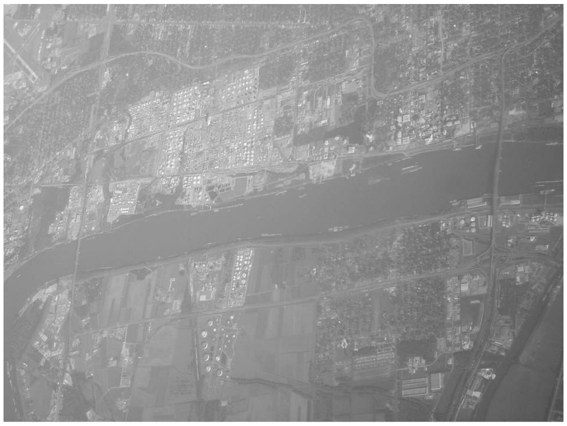


Fig.2 Source Spatial Image of Baton Rouge Areas

Discrete wavelet transform uses a set of basis functions for image decomposition. In a two dimensional case, four functions will be constructed: a scaling function $\varphi(x, y)$ and three wavelet functions $\psi^H(x, y)$, $\psi^V(x, y)$ and $\psi^D(x, y)$. Four product terms produce the scaling function (1) and separable directional sensitive wavelet functions (2)-(4), resulting in a structure of quaternary tree. Here the scaling function and wavelet functions are all determined by Haar Transform.

$$\varphi(x, y) = \varphi(x)\varphi(y) \tag{1}$$

$$\psi^H(x, y) = \varphi(y)\psi(x) \tag{2}$$

$$\psi^V(x, y) = \varphi(x)\psi(y) \tag{3}$$

$$\psi^D(x, y) = \psi(x)\psi(y) \tag{4}$$

The wavelets measure variations in three directions, where $\psi^H(x, y)$ corresponds variations along columns (horizontal), $\psi^V(x, y)$ corresponds to variations along rows (vertical) and $\psi^D(x, y)$ corresponds to variations along diagonal direction. The scaled and translated basis functions are defined by:

$$\Phi_{j,m,n}(x, y) = 2^{j/2}\, \varphi(2^jx - m, 2^jy - n) \tag{5}$$

$$\psi^i_{j,m,n}(x, y) = 2^{j/2}\, \psi^i(2^jx - m, 2^jy - n),\ i=\{H, V, D\} \tag{6}$$

where index i identifies the directional wavelets of H, V, and D. Given the size of image as M by N, the discrete wavelet transform of the function f(x, y) is formulated as:

$$w_\varphi(j_0,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\varphi_{j_0,m,n}(x,y) \tag{7}$$

$$w^i_\psi(j,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\psi^i_{j,m,n}(x,y) \tag{8}$$

where i={H, V, D}, j_0 is the initial scale, the $w_j(j_0, m, n)$ coefficients define the approximation of f(x, y), $w^i_\psi(j, m, n)$ coefficients represent the horizontal, vertical and diagonal details for scales $j >= j_0$. Here $j_0 = 0$ and select $N + M = 2^J$ so that j=0, 1, 2,…, J-1 and m, n = 0, 1, 2, …, $2^j - 1$. The f(x, y) can also be obtained via inverse discrete wavelet transform. Discrete wavelet decomposition and thresholding will both be applied in discrete wavelet transform.

Discrete wavelet transform is implemented as a multiple level transformation, where two level transformation is implemented in context. The decomposition outputs at each level include: the approximation, horizontal detail, vertical detail and diagonal detail. Each of them has one quarter size of its original image followed by downsampling by a factor of two. The approximation will be further decomposed into multiple levels while the detail components will not be decomposed. Information loss between two immediate approximations is captured as the detail coefficients. For the denoising using discrete wavelet transforms, only wavelet coefficients of the details at level one will be subject to thresholding, while the approximation components at the level one and higher levels will stay the same for image reconstruction.

In a thresholding process, the selection of the threshold is critical. Soft thresholding is selected instead of hard thresholding, which will shrink nonzero wavelet coefficients towards zero. Considering that a small threshold produces a good but still noisy estimation while in general, a big threshold produces a smooth but blurring estimation, thus the median value stem from the absolute value of wavelet coefficients at each wavelet decomposition level is selected. The shrinkage function of soft thresholding is formulated at each decomposition level as (9), where THR is the median threshold value based on wavelet coefficients. x is the input signal and f(x) is the nonlinear signal after thresholding.

$$f(x)= \mathrm{sgn}(x)(|x| - THR) \quad (9)$$

Using wavelet denoising, two revised images are generated which represent the intrinsic geographical information of two biggest cities of State of Lousiana (Figs. 3-4). These two denoised images will be further analyzed by nonlinear K-means clustering.

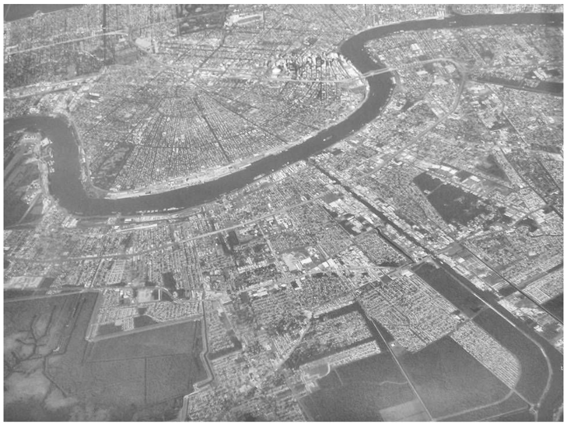


Fig.3 Denoised Image of New Orleans Areas

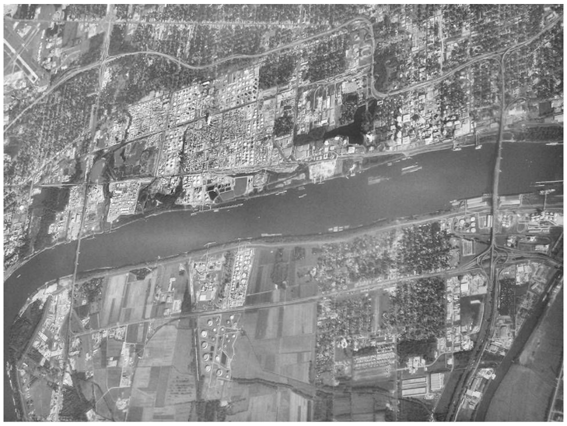


Fig.4 Denoised Image of Baton Rouge Areas

3. NONLINEAR K-MEANS SEGMENTATION

In image nonlinear segmentation, four clusters are proposed for partitioning. Centers of each cluster represent the mean values of all data points in that cluster. A distance metric should be determined to quantify the relative distances of objects. Both Euclidean and Mahalanobis distances are major types of distance metrics. Computation of the distance metrics is based on the spatial gray level histograms of digital images. The Mahalanobis metric distance has been applied, which is formulated as (10), where X_A is the cluster center of any layer X_A, s is a data point, d is the Mahalanobis distance, the K_A^{-1} is the inverse of the covariance matrix.

$$d=(s-X_A)^T K_A^{-1} (s- X_A) \quad (10)$$

K-means clustering assigns each object a space location, which classifies data sets through numbers of clusters. It selects four cluster centers and points cluster allocations to minimize errors. Optimal statistical algorithms are applied for classification, which are categorized as threshold based, region based, edge based or surface based. The distances of any specific data point to several cluster centers should be compared for decision making. For each individual input, winner-take-all competitive learning (11-12) is applied so that only one cluster center is updated. Images will thus be decomposed into four physical entities. In fact, the winner-take-all learning network classifies input vectors into one of specified categories according to clusters detected in the training dataset. All points are eventually allocated to the closest cluster. Learning is performed in an unsupervised mode. Each cluster center has an associated weight that is listed as w's. The winner is defined as one whose cluster center is closest to the inputs. Thus, this mechanism allows for competition among all input responses, but only one output is active each time. The unit that finally wins the competition is the winner-take-all cluster, so the best cluster center is computed accordingly.

$$w_{ij}x =\min(w_i x) \text{ for } j = 1, 2, 3, 4; \; i = 1, 2, 3, 4 \quad (11)$$

$$w_{i1} + w_{i2} + w_{i3}+ w_{i4} = 1 \text{ for } i = 1, 2, 3, 4 \quad (12)$$

Assume the cluster center S wins, the weight increment of S is computed exclusively and then updated according to (13), where α is a small positive learning parameter and it decreases as the competitive learning proceeds.

$$\Delta w_{ij} = \alpha(x_j - w_{ij}), \text{ for } j = 1, 2, 3, 4; i = 1, 2, 3, 4 \quad (13)$$

The K-means clustering outcomes of two city images are shown in Figs. 5-12, where objects of the highway, river, building and grass lawn are major features in 4 clusters.

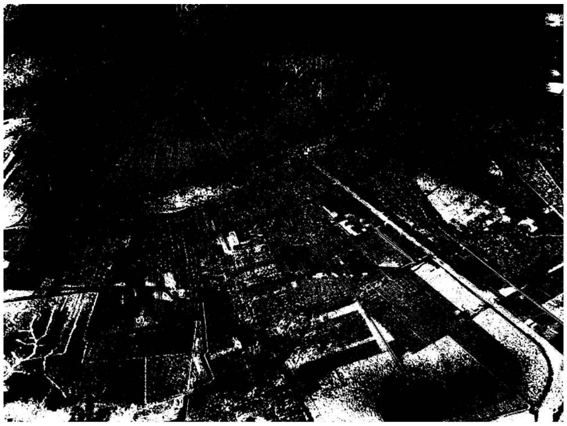


Fig.5 K-means Clustering #1 of New Orleans Areas

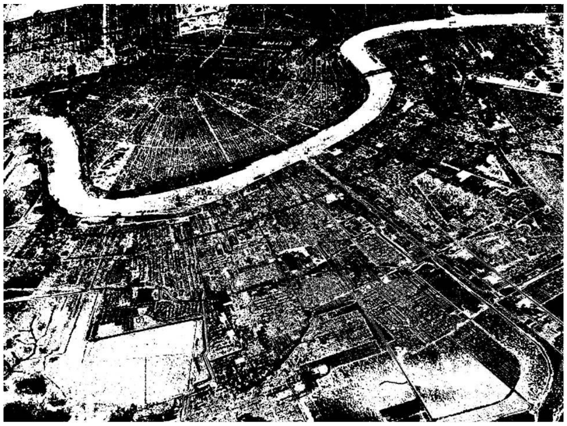


Fig.6 K-means Clustering #2 of New Orleans Areas

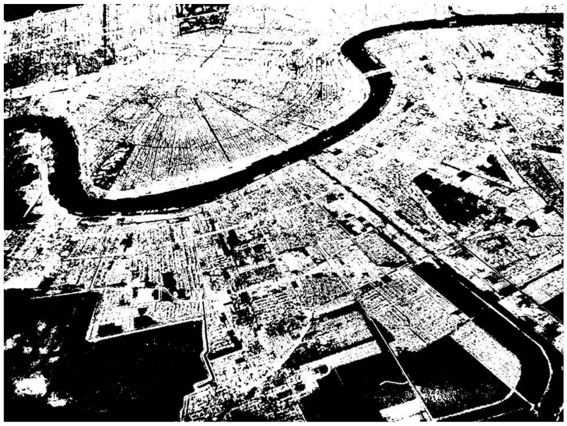


Fig.7 K-means Clustering #3 of New Orleans Areas

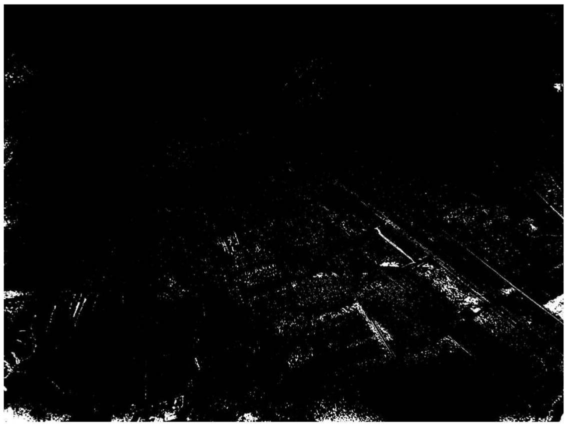


Fig.8 K-means Clustering #4 of New Orleans Areas

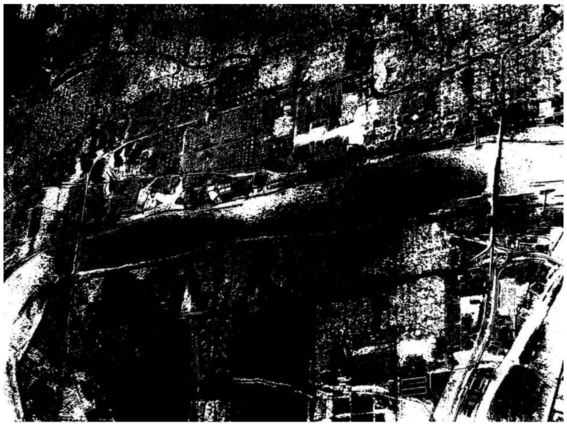


Fig.9 K-means Clustering #1 of Baton Rouge Areas

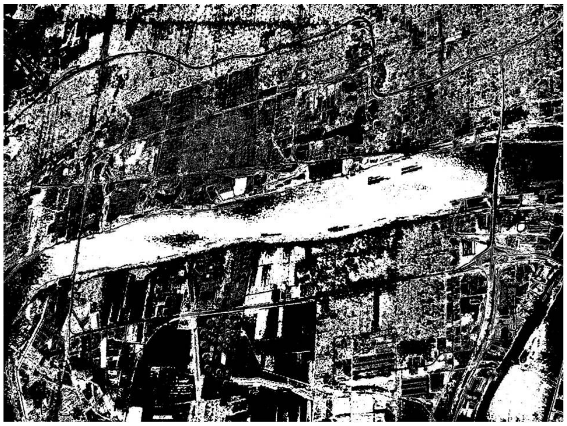


Fig.10 K-means Clustering #2 of Baton Rouge Areas

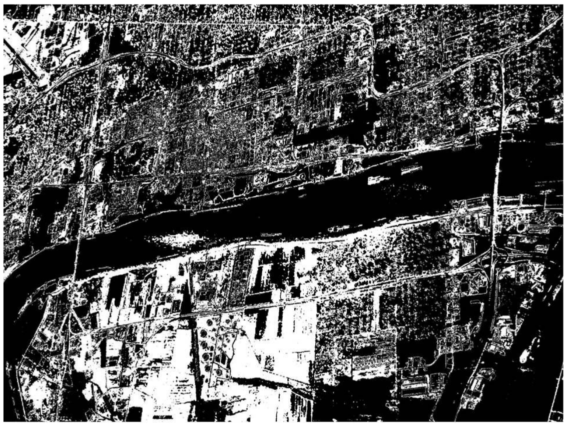


Fig.11 K-means Clustering #3 of Baton Rouge Areas

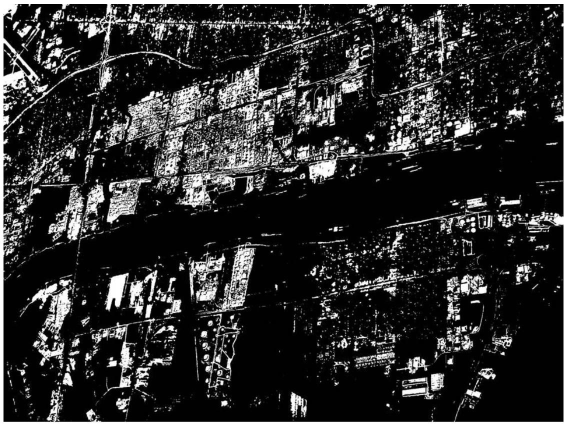


Fig.12 K-means Clustering #4 of Baton Rouge Areas

4. QUANTITATIVE ANALYSIS

4.1 Histogram and Probability Functions

For a M by N digital image, occurrence of the gray level is described as the co-occurrence matrix of relative frequencies. The occurrence probability function is then estimated from its histogram distribution.

4.2 Discrete Entropy Analysis

The discrete entropy is a measure of information content, which represents the average uncertainty of the information source. The discrete entropy is the summation of products of the probability of the outcome multiplied by the logarithm of inverse of probability of the outcome, taking into account of all possible outcomes {1, 2, …, n} as the gray level in the event {x_1, x_2, …, x_n}, where p(i) is the probability at the gray level i, which contains all the histogram counts. The discrete entropy H(x) is formulated as (14) and all corresponding results are shown in Table 1.

$$H(x)=\sum_{i=1}^{k} p(i)\log_2 \frac{1}{p(i)} = -\sum_{i=1}^{k} p(i)\log_2 p(i) \qquad (14)$$

Table 1 Discrete Entropy of Images

Discrete Entropy	Image A (N.O.)	Discrete Entropy	Image B (B.T.R)
Source Image	6.5630	Source Image	6.7279
Denoised Image	6.9670	Denoised Image	7.2252
Cluster 1	1.4650	Cluster 1	2.2084
Cluster 2	3.1182	Cluster 2	3.1886
Cluster 3	2.9009	Cluster 3	3.0486
Cluster 4	0.5474	Cluster 4	2.0423

4.3 Discrete Energy Analysis

The discrete energy measure indicates how the gray level elements are distributed. Its formulation is shown in (15), where E(x) represents the discrete energy with 256 bins and p(i) refers to the probability distribution functions at different gray levels, which contains the histogram counts. For any constant value of the gray level, the energy measure can reach its maximum value of one. The lower energy corresponds to larger number of gray levels and the higher one corresponds to smaller gray level numbers. The discrete energy of the source, denoised and segmented images are shown in Table 2.

$$E(x)=\sum_{i=1}^{k} p(i)^2 \qquad (15)$$

Table 2 Discrete Energy of Images

Discrete Energy	Image A (N.O.)	Discrete Energy	Image B (B.T.R)
Source Image	0.0122	Source Image	0.0112
Denoised Image	0.0090	Denoised Image	0.0077
Cluster 1	0.6705	Cluster 1	0.4802
Cluster 2	0.2684	Cluster 2	0.2609
Cluster 3	0.2788	Cluster 3	0.2938
Cluster 4	0.8965	Cluster 4	0.5411

4.4 Relative Entropy Analysis

Assuming that two discrete probability distributions of the digital images have the probability functions of p(i) and q(i). The relative entropy of p with respect to q is defined as the summation of all the possible states of the system, which is formulated as (16). The relative entropies of the source, denoised and segmented images are shown in Table 3.

$$d=\sum_{i=1}^{k} p(i)\log_2 \frac{p(i)}{q(i)} \qquad (16)$$

Table 3 Relative Entropy of Images

Relative Entropy	Source Image	Denoised Image A	Source Image	Denoised Image B
Cluster 1	0.0705	0.2944	0.0255	0.1938
Cluster 2	0.0882	0.2835	0.0294	0.1994
Cluster 3	0.0905	0.2846	0.0569	0.2665
Cluster 4	0.0512	0.2707	0.0719	0.3043
Denoised Image	0.3167		0.3088	

4.5 Mutual Information Analysis

Another metric of the mutual information I(X; Y) should also be discussed, which is used to describe how much information one variable tells about the other variable. The relationship is formulated as (17).

$$I(X;Y)=\sum_{X,Y} p_{XY}(X, Y)\log_2 \frac{p_{XY}(X, Y)}{p_X(X)p_Y(Y)}=H(X)-H(X|Y) \qquad (17)$$

where H(X) and H(X|Y) are values of the entropy and conditional entropy; p_{XY} is the joint probability density function; p_X and p_Y are marginal probability density functions. It can be explained as information that Y can tell about X is the reduction in uncertainty of X due to the existence of Y. The mutual information also represents the relative entropy between the joint distribution and product distribution. Calculated mutual information outcomes among the source, denoised and segmented images are indicated in Table 4.

Table 4 Mutual Information Between Images

Mutual Information	Source Image	Denoised Image A	Source Image	Denoised Image B
Cluster 1	5.0980	5.5020	4.5194	5.0167
Cluster 2	3.4447	3.8487	3.5392	4.0365
Cluster 3	3.6621	4.0661	3.6793	4.1765
Cluster 4	6.0156	6.4196	4.6855	5.1828
Denoised Image	0.4040		0.4973	

From Table 1 and Table 2, the denoised images cover more useful information than source images and each individual image cluster covers partial information. From Table 1 to Table 4, the quantitative values between the segmented images and original images can be set as measures for target detection when more clusters will be generated. Each cluster will actually represent certain type of objects that need to be identified. This image processing integration approach has been successfully applied to spatial object recognition issues.

5. CONCLUSIONS

This article has presented the outcomes from integration of image processing technologies. Image denoising can be used to maintain the energy of the images and reduce the energy of noises. Being a nonlinear approach, wavelet denoising has advantages of dealing with highly nonlinear spatial images. Using a set of wavelet bases, the wavelet coefficients can be thresholded to reduce the influence from noises. Wavelet denoising has been used to remove noises without distorting important features of images. Image segmentation can be used to identify objects from images. It classifies each image pixel to a segment according to the similarity in a sense of a specific metric distance. To reduce blurring effects of the spatial images stem from atmospheric media, nonlinear region K-means segmentation has been presented for image segmentation, where the competitive learning rule is applied to update clustering centers with satisfactory results. To evaluate the roles of wavelet denoising and nonlinear segmentation approaches, quantitative metrics are proposed. Several information measures of the discrete energy, discrete entropy, relative entropy and mutual information are applied to indicate the effects of integration of two image processing approaches. These methodologies could be easily expanded to other image processing techniques for diverse types of potential practical implementations.

6. REFERENCES

[1] R. Gonzalez, R. Woods, “Digital Image Processing,” 3rd Edition, Prentice-Hall, 2007

[2] R. Duda, P. Hart, D. Stork, “Pattern Classification,” 2nd Edition, John Wiley and Sons, 2000

[3] Simon Haykin, “Neural Networks – A Comprehensive Foundation”, 2nd Edition, Prentice Hall, 1999

[4] David MacKay, "Information Theory, Inference and Learning Algorithms", Cambridge Univ Press, 2003

[5] Z. Ye, H. Cao, S. Iyengar and H. Mohamadian, "Medical and Biometric System Identification for Pattern Recognition and Data Fusion with Quantitative Measuring", Systems Engineering Approach to Medical Automation, Chapter Six, pp. 91-112, Artech House Publishers, ISBN978-1-59693-164-0, October, 2008

[6] Z. Ye, H. Mohamadian and Y. Ye, "Information Measures for Biometric Identification via 2D Discrete Wavelet Transform", Proceedings of the 2007 IEEE International Conference on Automation Science and Engineering, pp. 835-840, Sept. 22-25, 2007, Scottsdale, Arizona, USA

[7] M. Ghazel, G. Freeman, and E. Vrscay, “Fractal-Wavelet Image Denoising Revisited”, IEEE Transactions on Image Processing, Vol. 15, No. 9, September, 2006

[8] J. Lorenzo-Ginori and H. Cruz-Enriquez, “De-noising Method in the Wavelet Packets Domain for Phase Images”, CIARP 2005, pp. 593 – 600, Springer-Verlag

[9] R. Mahmoud, M. Faheem and, A. Sarhan, "Intelligent Denoising Technique for Spatial Video Denoising for Real-Time Applications", 2008 International Conference on Computer Engineering & Systems, pp. 407-12, 2008

[10] Z. Ye, Y. Ye, H. Mohamadian and P. Bhattacharya, "Fuzzy Filtering and Fuzzy K-Means Clustering on Biomedical Sample Characterization", Proceedings of the 2005 IEEE International Conference on Control Applications, pp. 90-95, August, 2005, Toronto, Canada

[11] Z. Ye, J. Luo, P. Bhattacharya and Y. Ye, "Segmentation of Aerial Images and Satellite Images Using Unsupervised Nonlinear Approach", WSEAS Transactions on Systems, pp. 333-339, Issue 2, Volume 5, February, 2006

[12] M. Jaffar, N. Naveed, B. Ahmed, A. Hussain, A. Mirza, "Fuzzy C-means Clustering with Spatial Information for Color Image Segmentation", Proceedings of the 2009 International Conference on Electrical Engineering, 6 pp., April 9-11, 2009, Lahore, Pakistan

[13] Z. Ye, “Artificial Intelligence Approach for Biomedical Sample Characterization Using Raman Spectroscopy”, IEEE Transactions on Automation Science and Engineering, Volume 2, Issue 1, pp. 67-73, January, 2005

[14] Z. Ye, H. Mohamadian, Y. Ye, "Discrete Entropy and Relative Entropy Study on Nonlinear Clustering of Underwater and Arial Images", Proceedings of the 2007 IEEE International Conference on Control Applications, pp. 318-323, October, 2007

INFORMATION MINING FROM REMOTE SENSING IMAGERY BASED ON MULTI-SCALE AND MULTI-FEATURE PROCESSING TECHNIQUES

X.M. Yang [a, *], W. Cui [b], J.M. Gong [a], T. Zhang [a]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, CAS, 100101 Beijing, China - (yangxm, gongjm, zhangt)@lreis.ac.cn
[b] School of Electronic Information Engineering, Beihang University, 100084 Beijing, China - breaky@sohu.com

Commission VI, WG VI/4

KEY WORDS: Image Segmentation, Information Mining, Remote Sensing, Coastal Zone, Multi-scale, Multi-feature

ABSTRACT:

The paper attempts to present an information extraction approach in terms of image segmentation based on an object-oriented algorithm for high-resolution remote sensing images. The hierarchy frame and multi-features of the remote sensing image understanding and processing method are put forward. Firstly we extract various internal features of relatively homogeneous primitive objects using an image segmentation algorithm based on both spectral and shape information. Secondly, those primitives are analyzed to ascertain an optimal object by adopting certain feature rules, such as the traditional feature of the spectrum, shape, texture, spatial relation etc. Results from this research indicate that the model is practical to realize and the extraction accuracy of the coastal information is significantly improved compared to traditional approaches. Therefore, this study provides a potential way to serve our highly dynamic coastal zones for monitoring, management, development and utilization.

1. INTRODUCTION

In recent years, with the development of remote sensing and data storage technique, a great number of spatial data are generated every day, much of which is remote sensing image data. However, the use efficiency of the huge quantities of remote sensing image data is still low. It is very difficult for people to process with thousands of image data and find out knowledge from them. As the researches of date mining, information retrieval, multi media database and other correlative field have rapid progress, it is become possible to manage and analyze large amounts of remote sensing images and find out useful information in different applications.

It is hard to effectively utilize the current segmentation and classification algorithms in remote sensing image processing. The fundamental problem is that the current image analysis approach is different from human vision, and it is hard for a computer to process scene segmentation and image understanding in multi-scale space or to utilize the background knowledge or prior knowledge to eliminate disturbances as human being does. For mankind, segmentation of a scene is first based on the large-scale, that is access the large target or background first, and the corresponding contour. On this basis, scene details or sub-targets are then focused gradually.

The paper attempts to present an information extraction approach in terms of image segmentation based on an object-oriented algorithm for high-resolution remote sensing images. An aim of our research is to establish the hierarchy frame and an identification system of “pixel-primitive-object”, then to carry experiments on extraction of micro-scale coastal zone features, e.g., tidal flat, water line, sea wall, and mariculture pond.

2. FLOW OF REMOTE SENSING DATA MINING

In this paper, the hierarchy frame and multi-features of the remote sensing image understanding and processing method are put forward. Remote sensing images can be divided into simple images and complex images, and the same image can be divided into simple region and complex region. In this section, we consider indicators of spectral statistical measures, geometric feature of geo-objects, and spatial scale to do the complexity description for images, and we process the wide area segmentation and scene partition with the help of the computed complexity measure (Gao, 2010; Yang, 2009). This helps to understand images from a macro perspective, and rapid segmentation can be processed in different regions based on the different complexity measures, further reduce the solution space of information retrieval (Cardaci, 2005; Mario, 2005; Song, 2005). In a scale large enough to carry out large region segmentation, the segmentation can be rough and global, aiming at providing priori knowledge for detailed segmentation. For example, based on spectrum, shape and other features, research area can be divided into water body area, artificial target area, vegetation area, mountain area, etc.(Yang, 2009).

A novel framework of image understanding and computing based on multi-scale and multi-feature is developed in this paper. As shown in Figure 1, the framework consists of five steps, which are image complexity description, big area rough division, multi-scale fine segmentation, feature primitive merging and classification, and feature primitive and target mapping.

* Corresponding author: Xaiomei Yang, E-mail: yangxm@lreis.ac.cn.

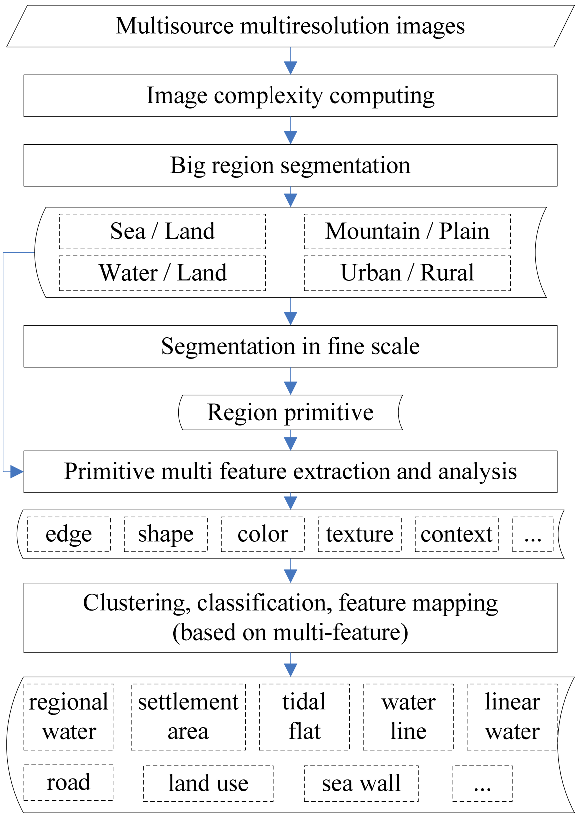


Figure 1. Framework chart of remote sensing data mining

3. IMAGE SEGMENTATION

Based on multi-scale image segmentation, we design some algorithms as follows.

3.1 Fast watershed segmentation

Using the primary image segmentation, we can obtain second plots, which is also called divisional sub-primitive. More details about Fast watershed segmentation see the Hill's paper (2003).

3.2 Fast and repeatable merger

Merging the segmentation results at the sub-primitive level, we can achieve the final divisional plots and finish the entire process of image segmentation.
Figure 2 shows the detail flow of image segmentation. In the process of merger, the difference indexes between different plots have spectral merger cost and shape merger cost, and the latter includes the weighted combination of shape compactness index and smoothness index. When the merger cost excesses the square of certain scale parameter set by the program, the terminative flag will appear and the merger algorithm will be stopped. If different scale parameters are set in the program, we will realize the process of multi-scale image segmentation. The algorithm efficiency we tested has fulfilled the needs of application.

4. FEATURE PRIMITIVE MEASUREMENT

Feature primitive measuring is a process of object expression for latent knowledge of primitives resulting from the image segmentation. In addition to the traditional feature of the spectrum, object expression rules also include shape, texture, spatial relations etc.(Yang, 2009).

4.1 Spectrum feature

(1) Spectrum statistical features
Mainly includes some statistical index such as mean, variance, histogram, and so on.
(2) Spectrum computational features
Mainly includes arithmetic operations between different bands spectrum in the same images, such as NDVI, etc.

4.2 Shape feature

Mainly includes area, perimeter, principal axle direction, and so on. Shape features focus on parameter representation, which comprising size invariance and rotation invariance, which describe sub-matrix and border-matrix or sub-function and turning-function by using Fourier function. This article only lists shape characteristics which are experimental related.

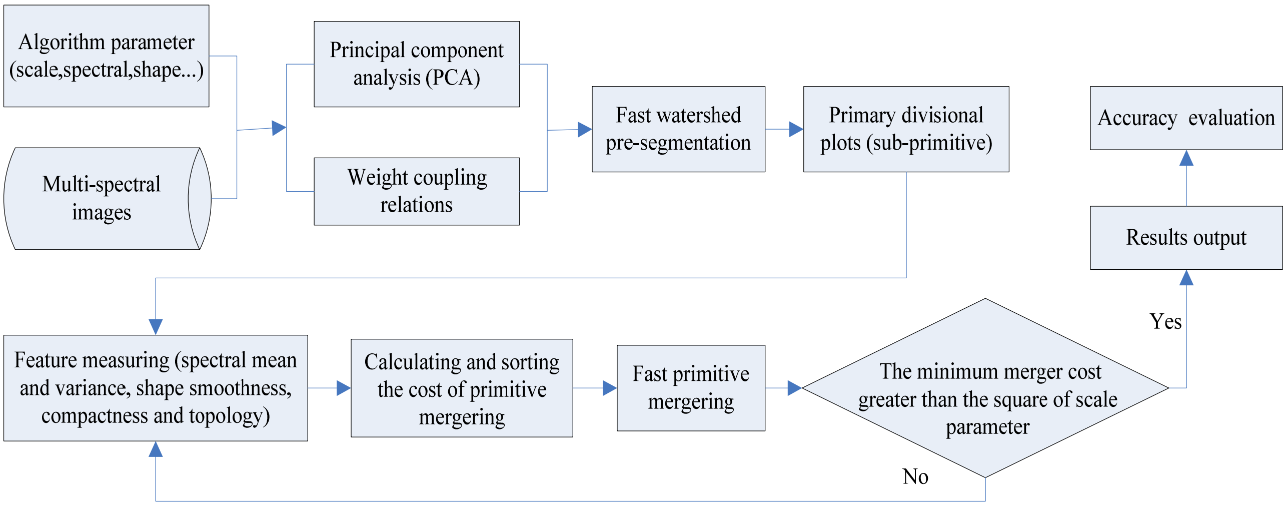

Algorithm parameter (scale,spectral,shape...)
Multi-spectral images
Principal component analysis (PCA)
Weight coupling relations
Fast watershed pre-segmentation
Primary divisional plots (sub-primitive)
Accuracy evaluation
Results output
Yes
Feature measuring (spectral mean and variance, shape smoothness, compactness and topology)
Calculating and sorting the cost of primitive mergering
Fast primitive mergering
The minimum merger cost greater than the square of scale parameter
No

Figure 2. Flowchart of image segmentation

4.3 Texture feature

Texture can be used to describe the grey value distribution features for images. The image texture can be different kinds, like wavelet texture, GABOR filter texture, LBP texture operator, and so on (Yang, 2006). The texture extraction method by Gray Level Co-occurrence Matrix is a classical statistical analysis method, and is also recognized as an image texture analysis method presently (Li, 2006). This article describes the image gray level distribution by the Gray Level Co-occurrence Matrix.

4.4 GIS spatial relationship

Mainly construct the topological relationship between elements, as well as elements spatial orientation information and "XOR" of geographical space knowledge and so on. For example, take DEM as a constraint to distinct terrain features, and take the distance away from the water sideline as a constraint to extract features and so on.

5. EXPERIMENTS

As shown in Table 1, aiming at application of image data mining method which have been discussed above, this article extracts target information concerning typical coastal objects such as water line, sea wall, tidal flat and mariculture pond. And it executes information extraction and result verification through different choice of parameter according to different object features.

Target	Parameters
Water line	DN, area, edge feature
Sea wall	Brightness, ratio of length to width
Tidal flat	DN, area and distance to water area
mariculture zone	Hue, squareness
……	……

Table 1. Main feature parameters of coastal targets

5.1 Water Line Information

As shown in Figure 30, this research selects multispectral SPOT data with resolution of 10m,. Firstly, the image is divided into several patches, among which small apertures and shadow are filtered based on the area. Secondly, seawater and land can be identified preliminarily by means of peak-valley iteration within histogram and threshold selection. Finally, the water line is extracted according to adjacent boundary between seawater and land. The result is demonstrated in figure 4 and 5.

5.2 Tidal Flat Information

Tidal flat is the marsh immerged by sea water, formed by the iterative influence of tide under. This kind of unstable resource of water and soil is often influenced by and changes with scouring and silting.

As shown in figure 6, the experimental data represents a zone around the coastal borderline between water and continent. The types of ground objects include seawater, submarine beach, tidal flat, mariculture pond, estuary, vegetation and residential area. The color of water appears blue in the pseudocolor synthetic image and varies with water depth and sediment amount; while vegetation appears red. Generally, the tidal flat is composed of silt or mud, several tidal channels, and sometimes sparse salt-tolerant vegetation. In pseudocolor synthetic image, the tidal flat is apparently an uneven French grey silt-zone, which distributes merely alongshore and within the bays.

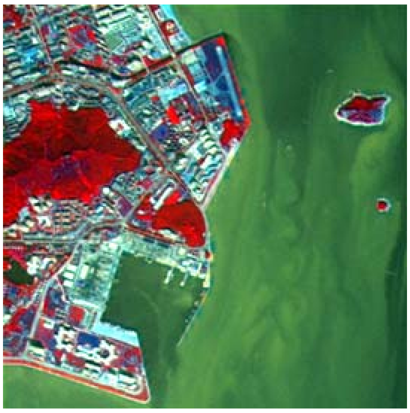


Figure 3. Original image (R,G,B=NR,R,G)

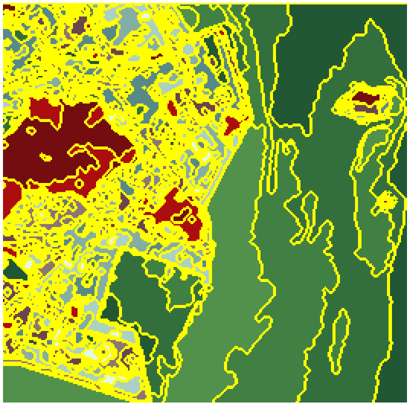


Figure 4. Segmentation result

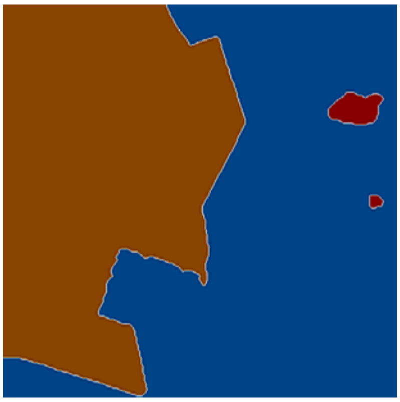


Figure 5. Water line extraction result

In this research, firstly, image segmentation is carried out and the results shown in Figure 7. Secondly, based on the mean of object and feature of area according to the 4th near-infrared band, seawater can be extracted. Thirdly, tidal flat is identified based on its hue and distribution features. Finally, we further the classification of tidal flat through analysis of the brightness, area and distance from seawater. The extraction result of tidal flat is shown as Figure 8.

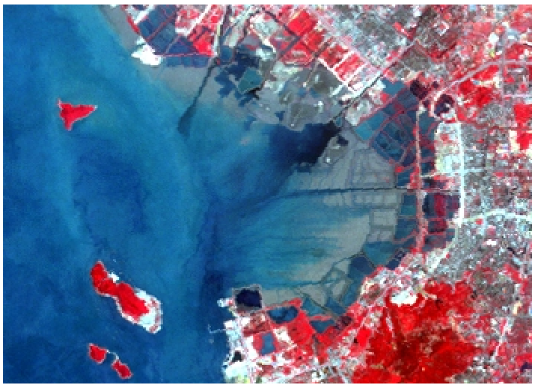


Figure 6. Original image

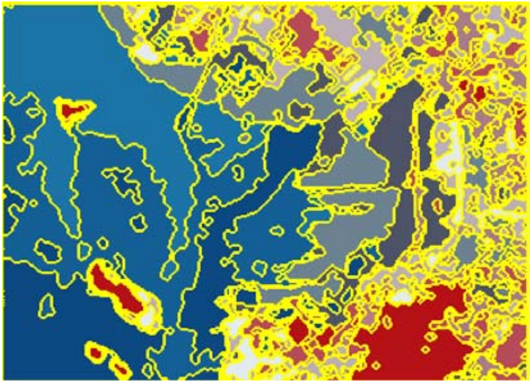


Figure 7. Object boundary display

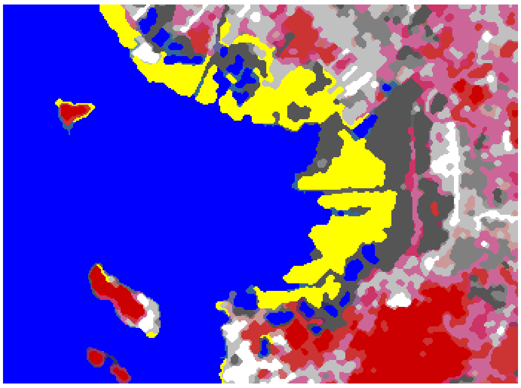


Figure 8. Extraction of tidal flat

6. CONCLUSION

Object-oriented methods of image analysis could conquer the localization of poor precision and present great superiority in the extraction of coastal moderate-and-higher resolution RS information. The tests showed that the identifying method of pixel-primitive-object can fulfill the demand of automatic interpretation of coastal information and enhance the precision of information recognition

In view of the abundance of ground objects' geometric and texture information in high-resolution coastal RS image, along with poor spectral information, the avail information of image is increasing but the noise and useless information are also increasing which complicate the relation among target ground objects. Therefore, on basis of analysis of coastal image and ground objects' internal feature, we should combine them with more geoscience knowledge and make comprehensive use of various information and expertise to increase the universality of segmentation algorithm and extraction methods in different situations.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (No. 40971224) and the National High Technology Research and Development Program of China (No.2008AA121706 and No. 2009AA12Z148).

REFERENCES

Cardaci M，Di Gesu' V，Petrou M，Tabacchi M, 2005. On the valuation of Images Complexity: A Fuzzy Approach. *International Workshopon Fuzzy Logic and Applications*, Crema, Italy, pp.305-311.

Gao ZY, Yang XM, Gong JM, Jin H, 2010. Research on image complexity description methods. *Journal of image and graphics*, 14(12), pp.129-135.

Hill P.R., Canagarajah C.N., and Bull D.R, 2003. Image segmentation using a texture gradient based watershed transform. *IEEE Trans. on Image Processing*, 12(12), pp.1618-1633.

Li J, Liu X, Li H, 2006. Extraction of texture feature and identification method of land use information from SPOT5 image. *Journal of remote sensing*, 10(6), pp.926-931.

Mario I，Chacon M，Alma D，Corral S, 2005. Image Complexity Measure: a Human Criterion Free Approach. *2005 Annual Meeting of the North American Fuzzy Information Processing Society*, Detroit, America, pp.241-246.

Song Xuefeng, 2005. Survey and Prospect on the Science of Complexity. *Complex Systems and Complexity Science*, 2(1), pp.10-17.

Yang XM, Lan RQ, Luo JC, 2006. Quantizing and analyzing the feature information of coastal zone based on high-resolution remote sensing image. *Acta Oceanologica Sinica*, 25(6), pp.33-42.

Yang Xiaomei, Gong Jianming, Gao Zhenyu, 2009. The research on extracting method of microscale remote sensing information combination and application in coastal zone. *Acta Oceanologica Sinica*, 31(2), pp.40-48.

MINING TIME SERIES DATA BASED UPON CLOUD MODEL

Hehua Chi [a], Juebo Wu [b, *], Shuliang Wang [a, b], Lianhua Chi [c], Meng Fang [a]

[a] International School of Software, Wuhan University, Wuhan 430079, China - hehua556@163.com
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China - wujuebo@gmail.com
[c] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China – lianhua_chi@163.com

KEY WORDS: Spatial Data Mining, Time-series, Cloud Model, Prediction

ABSTRACT:

In recent years many attempts have been made to index, cluster, classify and mine prediction rules from increasing massive sources of spatial time-series data. In this paper, a novel approach of mining time-series data is proposed based on cloud model, which described by numerical characteristics. Firstly, the cloud model theory is introduced into the time series data mining. Time-series data can be described by the three numerical characteristics as their features: expectation, entropy and hyper-entropy. Secondly, the features of time-series data can be generated through the backward cloud generator and regarded as time-series numerical characteristics based on cloud model. In accordance with such numerical characteristics as sample sets, the prediction rules are obtained by curve fitting. Thirdly, the model of mining time-series data is presented, mainly including the numerical characteristics and prediction rule mining. Lastly, a case study is carried out for the prediction of satellite image. The results show that the model is feasible and can be easily applied to other forecasting.

1. INTRODUCTION

With the rapid development of spatial information technology, especially spatial data acquisition technology, spatial database has become the data basis of many applications. Through the spatial data mining or knowledge discovery, we can obtain the general knowledge of geometry from spatial databases, including spatial distribution, spatial association rules, spatial clustering rules and spatial evolution rules, which can provide a powerful weapon for making full use of spatial data resources [1, 2]. As in the real world, most of the spatial data is associated with the time; more and more researchers have started to pay attention to the time series data mining. Time series model mining plays an important role in data mining.

Time series forecasting has always been a hot issue in many scientific fields. The study of time series data includes the following important aspects: trend analysis, similarity search, sequential pattern mining and cycle pattern mining of time-related data, time series prediction and so on. G. Box et al. divided the sequence into several sub-sequences through a moving window, and then discovered characteristic change pattern following the way of association rules by making use of clustering to classify these sub-sequences as a specific pattern of change [3]. J. Han et al. used the data mining technology to study cycle fragments and part of cycle fragments of the time series in the time series database, in order to discover the cyclical pattern of time series [4]. In the research [5], the authors proposed cyclic association rule mining. The literature [6] proposed calendar association rule mining. R. Agrawal et al. had given a series of sub-sequence matching criterion [7]. Two sequences were considered similar when existing a sufficient number of non-overlapping and similar sub-sequences timing right between them. Since the literature [8] published the first research paper about piecewise linear fitting algorithm for sequence data, relevant research has received extensive attention [9, 10, 11]. This simple and intuitive linear fitting representation used a series of head and tail attached linear approximation to represent time series. In combination with cloud model, W. H. Cui et al. proposed a new method of image segmentation based on cloud model theory to add uncertainty of image to the segmentation algorithm [12]. X. Y. Tang et al. presented a cloud mapping space based on gradation by the cloud theory aiming to solve the problem of land use classification of RS image [13]. K. Qin proposed a novel way for weather classification based on cloud model and hierarchical clustering [14]. In applications, time series analysis has been widely used in various fields of society, such as: macro-control of the national economy, regional integrated development planning, business management, market potential prediction, weather forecasting, hydrological forecasting, earthquake precursors prediction, environmental pollution control, ecological balance, marine survey and so on.

The time series mining research has received some progress, but there are still some shortcomings. For example: time series should be smooth and normal distribution. This paper presents a prediction model based on cloud model. This model describes the features of time series data through three numerical characteristics. The time series data, such as images, are standardized as the cloud droplets after the pre-processing, and then are executed cloud transform through backward cloud generator to get three numerical characteristics. Similarly, we can calculate number of sample data sets to obtain a series of cloud numerical characteristics. Through curve fitting, we can mine the prediction rules to achieve data prediction.

[*] Corresponding author. wujuebo@gmail.com.

2. THE CLOUD MODEL

2.1 The cloud model

Cloud model is proposed by D. Y. Li in 1996 [15]. As the uncertainty knowledge of qualitative and quantitative conversion of the mathematical model, the fuzziness and randomness fully integrated together, constituting a qualitative and quantitative mapping.

Definition: Let *U* be a universal set described by precise numbers, and *C* be the qualitative concept related to *U*. Assume that there is a number $x \in U$ that randomly stands for the concept *C* and the certainty degree of *x* for *C*, which is a random value with stabilization tendency and meets:

$$\mu : U \to [0,1] \quad \forall x \in U \; x \to \mu(x)$$

where the distribution of *x* on *U* is defined as cloud land. Each *x* is defined as a cloud drop. Figure 1 shows the numerical characteristics {*Ex En He*} of the cloud.

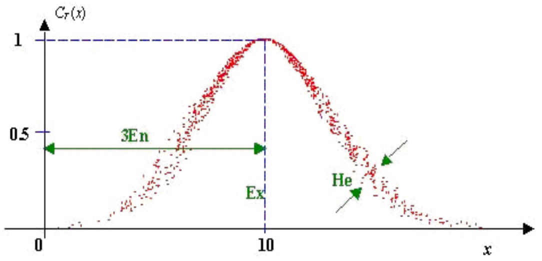


Figure 1. The numerical characteristics {*Ex En He*}

In the cloud model, we employ the expected value *Ex*, the entropy *En*, and the hyper-entropy *He* to represent the concept as a whole.

The expected value *Ex*: The mathematical expectation of the cloud drop distributed in the universal set.

The entropy *En*: The uncertainty measurement of the qualitative concept. It is determined by both the randomness and the fuzziness of the concept.

The hyper-entropy *He*: It is the uncertainty measurement of the entropy, i.e., the second-order entropy of the entropy, which is determined by both the randomness and fuzziness of the entropy.

2.2 Backward cloud generator

Backward cloud generator is uncertainty conversion model which realizes the random conversion between the numerical value and the language value at any time as the mapping from quantitative to qualitative. It effectively converts a certain number of precise data to an appropriate qualitative language value (*Ex, En, He*). According to that, we can get the whole of the cloud droplets. The more the number of the accurate data is, the more precise the concept will be. Through the forward and reverse cloud generator, cloud model establishes the interrelated and interdependent relations. The algorithm of backward cloud generator is as follows:

Input: Samples x_i and the certainty degree $C_T(x_i)$ *(i=1, 2…, N)*

Output: the qualitative concept (*Ex, En, He*)

Steps:

(1) Calculate the average value of x_i, i.e., $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$, the first-order absolute central of the samples $M_1=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\overline{X}\right|$, sample variance $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2$

(2) $Ex=\overline{X}$

(3) $En=\sqrt{\frac{\pi}{2}}\times M_1$

(4) $He=\sqrt{S^2-En^2}$

3. TIME-SERIES MINING MODEL BASED ON CLOUD MODEL

Using cloud model theory, this section proposes the framework of time-series data mining, and gives concrete steps. The time-series data can be described by the three numerical characteristics of cloud model.

3.1 Time-series data mining framework based on cloud model

The main idea of time-series data mining framework based on cloud model is as follows:

Firstly, extract the experimental data from the time-series databases and pre-process the data to obtain cloud droplets.

Secondly, make use of backward cloud generator algorithm to extract the numerical characteristics {*Ex En He*} of the cloud droplet.

Thirdly, extract the numerical characteristics {*Ex En He*} of the each cloud droplet and get all groups of the numerical characteristics {*Ex En He*}.

Finally, make use of the fitting algorithm to do the fitting of all groups of the numerical characteristics {*Ex En He*}, in order to achieve forecast according to the curve.

The framework of time-series data mining and the main flow based on cloud model are shown in Figure 2, where gives the specific description of this framework.

3.2 Data pre-processing

The main purpose of data pre-processing is to eliminate irrelevant information and to recovery useful information. The time-series data from time-series database is rough, and it is necessary to pre-process data in order to carry out the next step. The object of cloud model is cloud droplets. We have to turn the time-series data into cloud droplets as the input of the cloud model.

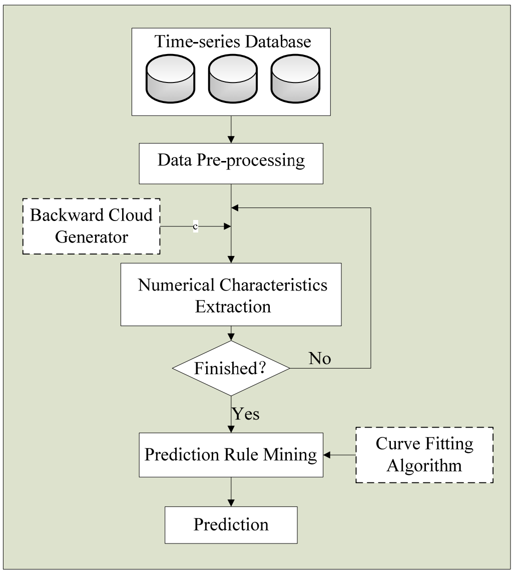


Figure 2. The framework of time-series data mining based on cloud model

3.3 Numerical Characteristic Extraction

Through the numerical characteristic extraction of the time-series data based on cloud model, we can obtain the numerical characteristics sets of time-series data. These characteristics are direct descriptions of the time-series data on behalf of the internal information. Time-series data can come from time-series database and can also change over time. We can also make use of backward cloud generator to extract numerical characteristics. The algorithm is referred as section 2.2.

The numerical characteristics {*Ex En He*} of the cloud reflect the quantitative characteristics of the qualitative concept.

Ex (Expected value)**:** The point value, which is the most qualitative concept in the domain space, reflects the cloud gravity centre of the concept.

En (Entropy)**:** Entropy is used to measure the fuzziness and probability of the qualitative concept, reflecting the uncertainty of qualitative concept. The entropy of time-series data can reflect the size range of the cloud droplet in the domain space.

He (Hyper entropy)**:** The uncertainty of the entropy-entropy's entropy reflects the cohesion of all the cloud droplets in the domain space. The value of the hyper entropy indirectly expresses the dispersion of the cloud.

By the numerical characteristic extraction of the time-series data based on cloud mode, we can get the numerical characteristics of a large number of cloud droplets. These numerical characteristics extracted from the time-series data describe the overall features, as well as the trend of development and change over time. The next step is the curve fitting of the numerical characteristics to obtain prediction rules.

3.4 Prediction Rule Mining

The process is mainly to identify the prediction curves. We regard the numerical characteristic extraction of the time-series data based on cloud model as the sample set of a curve fitting, and get the fitting curve by the curve-fitting method. Due to the differences of the time-series data, we should choose the proper curve-fitting algorithm in the specific application. For example, by using the least square method. The (*Ex En He*) of time-series data is effectively fitted to achieve the prediction rule mining.

According to causality between prediction objects and factors, we can achieve the prediction. There are many factors associating with the target. We must choose the factors having a strong causal relationship to achieve the prediction. There are two factors variable X and the dependent variable Y to express the prediction target. If they are the linear relationship between X and Y, then the relationship is $Y=a+bX$. But if they are the nonlinear relationship between X and Y, then the relationship maybe is $Y=a+b1X+b2X2+$**...** or others. The a and b in the formula are the agenda coefficients. They can be valued by the statistical or other methods. The time-series for short term prediction is more effective. However, if it is used for long-term prediction, it must also be combined with other methods. After getting the prediction curves, we can carry out the further work, namely, to predict future trends of the time-series data.

3.5 Prediction

The time series is a chronological series of observations. By the previous step, we can get one or more fitting curves. These curves are as predictable rules. By these prediction rules, we can carry out time series data prediction. According to the knowledge of the function and time parameters, we can obtain relationship value of fitting curve function. Making use of function relationship, we can predict and control the problem under the given conditions in order to provide data for decision-making and management. After getting mathematical model of long-term trends, seasonal changes and irregular changes by historical data of time series, we can use them to predict the value T of long-term trends, the value S of seasonal changes and the value I of irregular changes in the possible case. And then we calculate predicted value Y of the future time series by using these following models:

Addition mode $T + S + I = Y$
Multiplicative model $T \times S \times I = Y$

If it is hard to get predicted value of irregular changes, we can obtain the predicted value of long-term trends and seasonal changes by putting the multiplier or sum of the above two as predicted value of time series. If the data itself have not seasonal changes or not need to forecast quarterly sub-month data, the predicted value of long-term trends is the predicted value of time series, that is, $T = Y$. In such way, by predicting the rules curves, we can predict the time series data to discover the laws of development of things and make a good decision-making for us.

4. A CASE STUDY

In order to verify the validity of the time-series data mining based on cloud model, we make an experiment about the satellite images in this section. The experiment makes use of the real-time satellite cloud image from the Chinese meteorological. Through the analysis of historical data, we can obtain the prediction rules to carry out the prediction of weather trends.

4.1 Data acquisition

In this study, data sources come from the satellite cloud data of Chinese meteorological. The website: http://www.nmc.gov.cn/. These data are real-time dynamic. The satellite makes the real-time photography, and then returns the data to earth. We choose the three kinds of data as the data source, China and the Western North Pacific Sea Infrared Cloud, China Regional Vapour Cloud, China Regional Infrared Cloud. These data have the following characteristics: Time-series property, synchronization property and consistent view property.

We collect a total of 6 months of the historical data to make the experiments. After data collection, we select the 1600 images of China and the Western North Pacific Sea Infrared Cloud, 1800 images of China Regional Vapour Cloud and 2700 images of China Regional Vapour Cloud, as the sample sets of the experiments. Part of the sample sets are shown in Figure 3, Figure 4 and Figure 5.

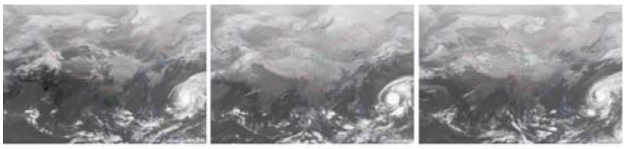


Figure 3: China and the Western North Pacific Sea Infrared Cloud

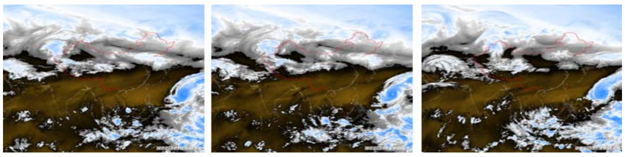


Figure 4: China Regional Vapour Cloud

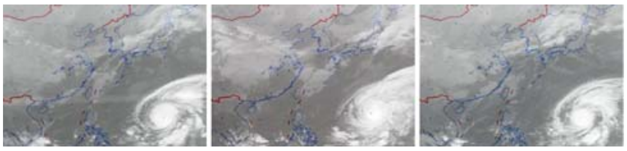
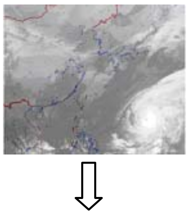


Figure 5: China Regional Infrared Cloud



$$\begin{bmatrix} 125 & 114 & 117 & 124 & 130 & 128 & 123\ldots \\ 124 & 115 & 119 & 127 & 128 & 125 & 125\ldots \\ 130 & 116 & 123 & 128 & 127 & 125 & 125\ldots \\ 132 & 119 & 124 & 125 & 124 & 124 & 124\ldots \\ 130 & 122 & 125 & 122 & 121 & 121 & 124\ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\ \ldots \end{bmatrix}$$

Figure 6: Pixel matrix

4.2 Data pre-processing

In order to handle easily, it should do pre-processing for satellite cloud images before mining the knowledge. The main purpose is to extract the image pixel value. First, satellite images needs to be transformed to the same size. Then, the images should be converted into grid with extracting the corresponding pixel value. The size of the grid is set depending on image content. Through the above steps, a pixel matrix can be obtained. The value of each matrix can be as a cloud droplet in cloud transformation. Figure 6 shows a satellite image corresponding to the value of the pixel matrix, that is, cloud droplets.

4.3 Image Feature Extraction

This step is primarily to do satellite imagery feature extraction and get a collection of three numerical characteristics based on cloud model features. Input data are the pixel values obtained from the pre-processing step. Each pixel value corresponds to the backward cloud generator among the cloud droplets. The implementation process can refer to Section *2.2*, and the main function is to achieve the following:

Calculate expectation value:

```
ImageExp=SelectAveImage(Images, Num);
```

Calculate entropy:

```
ImageEn=CalculateStdImage(Images, ImageExp, Num);
```

Calculate hyper-entropy:

```
ImageHe=CalculateReVarianceImage(1, ImageEn);
```

The data sets of cloud features of satellite images can be generated in this step, which can be as the data source for prediction curve fitting.

4.4 Curve Fitting Prediction

Through the backward cloud generator, we have a sample set of the numerical characteristics based on cloud model. We can get the curve fitting by the sample sets. In this study, we use the least square curve fitting method. For each type of data, we can get three prediction curves, that is, the expected value curve, the entropy curve and the hyper entropy curve. Main function is:
Define the solving functions of the polynomial fitting coefficient; *x*, *y* as the input data; n as the numbers of fitting:

```
function A=nihe(x,y,n)
```

Measure the length of data:

```
m=length(x);
```

Generate the *X* matrix:

```
X1=zeros(1,2*n);
…
X2=[m,X1(1:n)];
X3=zeros(n,n+1);
…
X3(j,:)=X1(j:j+n);end
X=[X2;X3];
Y=zeros(1,n);
…
Y=[sum(y),Y];Y=Y';
```

Obtain fitting coefficient vectors *A*:

```
A=X/Y;
```

4.5 Prediction analysis

Through the above steps and each type of satellite images, we can get three different projection curves. We look them as the prediction rules to forecast the future satellite cloud.

In this study, we use *80%* of the sample sets as the training sets and another *20%* of the data as the prediction reference value. By comparing the prediction values with actual value, we can get the model prediction accuracy rate. The results are shown in table *1*. It shows the results of the different types of satellite cloud through the prediction model, the average accuracy rate of three kinds of satellite cloud is *88.13%* and meets projection demands. The results show that the projection of satellite cloud is feasible.

Location	Trainin g data	Test data	Projection data	Accuracy
China and the Western North Pacific Sea Infrared Cloud	1280	320	360	88.7%
China Regional Vapour Cloud	1440	360	360	86.3%
China Regional Infrared Cloud	2160	540	540	89.4%

Table 1: the results of the satellite cloud projection

5. CONCLUSIONS

Most of the spatial data have the time dimension, and will change over time. Time series space contains the time dimension in spatial association characteristics and can get time series space association rules through time series space association rule mining. This paper presented a method of time series data rules mining, which played an important significance for getting time series space association rules and doing the practical application by time series space association rules. First, by the backward cloud generator and the three numerical characteristics of cloud model, this model described the features of time series data. Second, the digital features of a series of sample sets were obtained as the training sample sets. Then, by these feature points, the rule curve fitting was predicted for obtaining predictive models. Finally, the time series data were predicted by the forecasting rules. The experimental results showed that the method is feasible and effective. Through the data standardization processing, this method can be extended to multiple applications. Time series data mining is an interdisciplinary science and the further research including the following aspects:

(1) Other research of curve fitting method.

(2) The combination of research of cloud model and other algorithms.

(3) The visualization studies of time series and time sequence similarity studies.

6. ACKNOWLEDGEMENTS

This paper is supported by National 973 (2006CB701305, 2007CB310804), National Natural Science Fund of China (60743001), Best National Thesis Fund (2005047), and National New Century Excellent Talent Fund (NCET-06-0618).

7. REFERENCES

[1] D. R. Li, S. L. Wang, W. Z. Shi et al. On Spatial Data Mining and Knowledge Discovery [J]. Geomatics and Information Science of Wuhan University, 2001, 26(6): 491-499.

[2] S. Shekhar, P. Zhang, Y. Huang et al. Trends in Spatial Data Mining. In: H.Kargupta, A.Joshi(Eds.), Data Mining: Next Generation Challenges and Future Directions[C]. AAAI/MIT Press, 2003, 357-380.

[3] G. Box, G. M. Jenkins. Time series analysis: Forecasting and control, Holden Day Inc., 1976.

[4] J. Han, G. Dong, Y. Yin. Efficient mining of partial periodic patterns in time series database. In Proc. 1999 Int. Conf. Data Engineering (ICDE'99), pages 106-115, Sydney, Australia, April 1999.

[5] B. Ozden, S. Ramaswamy, A. Silberschatz. Cyclic association rules. Proceedings of the 15 th International Conference on Data Engineering. 1998, 412-421.

[6] Y. Li, P. Ning, X. S. Wang et al. Discovering calendar-based temporal association rules. Data&Knowledge Engineering, 2003, 44: 193-218.

[7] R. Agrawal, K. I. Lin, H. S. Sawhney et al. Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. Proceedings of the 21th International Conference on Very Large Data Bases, 1995, 490-501.

[8] T. Pavlidis, S. L. Horowitz. Segmentation of plane curves. IEEE Trans. Comput. 23, 1974, 860-870.

[9] S. Park, S. W. Kim, W. W. Chu. Segment-based approach for subsequence searches in sequence databases. In Proceedings of the Sixteenth ACM Symposium on Applied Computing, 2001, 248-252.

[10] K. B. Pratt, E. Fink. Search for patterns in compressed time series. International Journal of Image and Graphics, 2002, 2(1): 89-106.

[11] H. Xiao, Y. F. Hu. Data Mining Based on Segmented Time Warping Distance in Time Series Database [J]. Computer Research and Development, 2005, 42(1): 72-78.

[12] W. H. Cui, Z. Q. Guan, K. Qin. A Multi-Scale Image Segmentation Algorithm Based on the Cloud Model. Proc. 8 th spatial accuracy assest. in natural resources, World Academic Union, 2008.

[13] X. Y. Tang, K. Y. Chen, Y. F. Liu. Land use classification and evaluation of RS image based on cloud model. Edited by Liu, Yaolin; Tang, Xinming. Proceedings of the SPIE, 2009, Vol.7492, 74920N-74920N-8.

[14] K. Qin, M. Xu, Y. Du et al. Cloud Model and Hierarchical Clustering Based Spatial Data Mining Method and Application. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2, 2008.

[15] D. R. Li, S. L. Wang, D. Y. Li. Spatial Data Mining Theories and Applications. Science Press, 2006.

SEMANTIC AUGMENTATION OF GEOSPATIAL CONCEPTS: THE MULTI-VIEW AUGMENTED CONCEPT TO IMPROVE SEMANTIC INTEROPERABILITY BETWEEN MULTIPLES GEOSPATIAL DATABASES

M. Bakillah[1]*, M.A. Mostafavi[1], J. Brodeur[2]

[1]Centre de Recherche en Géomatique, Université Laval, Québec, Canada, G1K 7P4
[2]Centre d'information topographique de Sherbrooke, 2144 King, Canada
Mir-Abolfazl.Mostafavi@scg.ulaval.ca
Jean.Brodeur@RNCan-NRCan.gc.ca

Commission II, WG II/IV

KEY WORDS: Geospatial concept, View paradigm, Rule mining, Knowledge representation, Semantic Interoperability

ABSTRACT:

Semantic interoperability is a key issue for the meaningful sharing of geospatial data between multiples geospatial databases. It requires the establishment of semantic mappings between concepts databases' ontologies. Semantic mappings can be discovered only when semantics is explicit. However, existing concepts' definitions are not always sufficient to represent the semantic richness of geospatial concepts. In addition, semantics may be implicit, refraining from using it during semantic mapping process. This paper, proposes a new representation for geospatial concepts, called Multi-View Augmented Concept (MVAC), which takes into account these drawbacks. We propose a method to generate a MVAC, based on: (1) extraction of the different views of a concept that are valid in different contexts, and (2) augmentation of a concept with implicit dependencies between its features based on rule mining theory. We believe that the proposed approach will play an important role to improve the quality of the semantic interoperability between multiple geospatial databases since it takes into account the implicit semantic relations between different concepts.

1. INTRODUCTION

Semantic interoperability is a major research topic for ensuring data sharing and reuse among heterogeneous systems [Bian and Hu 2007]. It is the knowledge-level interoperability that provides cooperating databases with the ability to resolve differences in meanings of concepts [Park and Ram 2004]. Resolving those differences requires that meaning is available to machines into an explicit representation so it can automatically be processed during semantic mapping, that is, the discovering of semantic relations between concepts of different ontologies. However, current semantic mapping approaches rely on poor concepts' definitions that are not suitable for representing all the richness of geospatial concepts. For example, not considering explicitly the semantics of spatial and temporal properties of a concept reduces its expressivity. In addition, it may contain implicit knowledge that can be inferred from existing knowledge. The structure of the concepts is also important. Considering a concept as a bag of features is not sufficient. To address these problems, we propose a new representation of geospatial concepts, the Multi-View Augmented Concept Model (MVAC) (presented in section 3), and a method to generate MVAC representation (presented in section 4). In this method, we add two additional layers to the definition of the concept. First, we extract the different views it can have in different contexts, and then, we augment it with dependencies between its features. The contribution of the MVAC model is to improve semantic interoperability with a concept that has richer semantics, and a structure that allow discovering semantic relations between concepts of different ontologies that were hard to discover with traditional, lexical-based semantic mapping approaches. This paper is organized as follows. In section 2, we review related work on definition of concepts. In section 3, we propose the MVAC model. In section 4, we propose the MVAC generation method. In section 5, we discuss with a case study how the MVAC can help to improve semantic interoperability. In section 6, we conclude this paper.

2. RELATED WORK ON THE DEFINITION AND REPRESENTATION OF GEOSPATIAL CONCEPTS

Knowledge representation is the problem of encoding the knowledge that human have about the reality, in such a way that it supports reasoning [Kavouras and Kokla 2008]. A knowledge representation is not a complete and perfect picture of the reality; but an abstraction of a portion of reality that is relevant in a domain. Knowledge representation is a fundamental issue for improving semantic interoperability because it supports knowledge sharing (between humans and between machines). The theoretical basis of knowledge representations depends on the different theories of the concept. Cognitively, concepts are mental representations of a category [Medin and Rips 2005], and a category denotes a set of real world entities that have similar properties [Kavouras and Kokla 2008]. It is very difficult to give a framework that would guide the assignment of properties to concepts in a universal way, even if such attempts were made [Bennett 2005]. The choice of a concept's properties depends on the intended purpose [Tomai and Kavouras 2004]. In the geospatial domain, proposed definitions of the concept aim at identifying the special

* Corresponding author: mohamed.bakillah.1@ulaval.ca

properties of geospatial objects. Kavouras and Kokla [2008] define a concept with a term, a set of semantic elements (properties and relations) and their values. This is similar to the definition of the concept in Schwering and Raubal [2005] where concepts are defined by properties (represented as dimensions in a conceptual space) and property values (represented as values of those dimensions). Kavouras and Kokla have identified features such as: *purpose*, *agent*, *shape*, *size*, *location*; *frequency*, *duration*, *is-a*, *part-of* relations; *relative position relations* (upward, downward, behind, etc.); *proximity*, *direction* and *topological relations* (adjacency, connectivity, overlap, etc.). Rodriguez and Egenhofer [2003] have classified features as attributes, functions (representing what is done to or with an object) and parts (structural component of an object). This classification aims at facilitating the separate manipulation of each type of properties in the context of semantic similarity assessment. Brodeur and Bédard [2001] give another set-based definition of concepts. They proposed a definition based on the four-intersection model of Egenhofer [1993]. A concept has an interior, defined by its intrinsic properties (e.g. identification, attributes, attribute values, geometries, temporalities), and a boundary, defined by its extrinsic properties (e.g. relationships and behaviours). The whole set of intrinsic and extrinsic properties forms the *context*. Keßler et al. [2007] argue that the context has two components: the internal context specifies the domain of application and the external context is a set of rules that allows to modify the concept in different circumstances. Bennett [2005] has attempted to provide a generic definition of the concept. He proposes that properties of an object may be classified as physical (including geometry and material properties); historical (how the object came into existence; the events it has undergone, etc.); functional, including static and dynamic functions; conventional properties (related to the *fiat* nature of objects). Bennett mentions "objects that exhibit one property, will very often also exhibit another property", but he does not explicit further those types of dependencies between properties. A first problem with the above approaches is that they define the concept as unstructured set of features. However, features are related through dependencies. For example, position of a moving object depends on time, the value of an object's temperature depend on the value of its altitude, etc. However, if those dependencies are not stated in the concept's definition, it may be possible to discover implicit dependencies by looking in the instances of the concept. A second problem is that in most of the definitions, spatial and temporal properties are not explicit but merged into other classes of properties. This makes the separate manipulation of spatial or temporal properties difficult. Most approaches define properties with their name and range of values, for example, "geometry of house" is a "polygon". This is not sufficient to understand the exact semantics of this spatial property. The polygon may represent "roof of house" or "foundation of house". Spatial and temporal properties have to be described in a more explicit manner. Finally, there are different ways to define a concept depending on the context [Parent *et al*. 2006]. Several researchers have investigated the multi-view paradigm for concepts and propose modelling views in geospatial databases [Bédard and Bernier 2002; Parent et al. 2006] and in ontologies [Bhatt et al. 2006; Wouters et al. 2008]. Beside the strict representation issues, multiples views of a same concept can also provide multiple ways to achieve semantic interoperability. However, existing representation of geospatial concepts tend not to include this paradigm explicitly, nor to demonstate its usefullness in semantic interoperability.

3. THE MULTI-VIEW AUGMENTED CONCEPT (MVAC) MODEL

The new definition of the concept we propose is intended to address the drawbacks of concept definitions identified above, an its contribution is a more rich and structured definition as a basis for improved semantic interoperability. The MVAC adds two additional layers to the original definition of a concept: a set of views valid in different contexts, and a set of dependencies between features of the concept (Fig. 1).

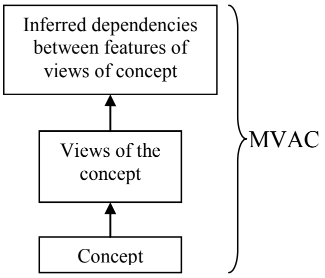


Figure 1. MVAC Model

At the first level, a concept, denoted by c, is defined as: c = <n(c), {p(rp)}, {r(rr)}, {spatial_d(rsd)}, {temporal_d(rtd)}>, where:

- n(c) is the name of the concept;
- {p(rp)} is the set of properties of the concept. The set of possible values of a property, called the range and denoted rp, is given in brackets.
- {r(rr)} is the set of relation that c has with other concepts. rr represents the range of the relation r, that is, the set of concepts c is linked with through relation r.
- {spatial_d(rsd)} is a set of properties, called spatial descriptors, which role is to describe the spatiality of the concept. For example, the concept watercourse could have the spatial descriptor geo-entity (axis of watercourse), meaning that the line geometry representing the watercourse corresponds to the axis of the watercourse. The range od spatial descriptor is denoted rsd.
- {temporal_d(rtd)} is a set of properties, called temporal descriptors, which role is to describe the temporality of the concept. The range of temporal descriptors is denoted rtd. For example, the concept watrecourse may have temporal descriptor waterlogged period (average flooded period) which means that the waterlogged period correspond to the average time the watercourse is flooded overs years.

We provide an example for the concept “watercourse”:

c = <watercourse, {flooding, tourism, transport}, {water level(low, medium, high), category(intermittent, stable), spatial extent(polygon, moving polygon), function(navigation, skating, evacuation area), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, flooded area, frozen area)}{waterlogged period(average flooding period)}>

This concept may represent different realities in different contexts. For each context, we want to create a view that can be used in that context. In a previous work (Bakillah et al. 2009) we have stated that the view paradigm support ontology reuse, by selecting only parts of a concept that are relevant in a given context. We have defined views as the result of inference over logic rules. We precise that views are inferred from rules on context. A view of a concept is a selection of its features that are valid in a given context. The context represents a given real world situation, for example, a disaster. A view is defined as:

View(c): Context(Name of context) → <{p(rp_v)}, {r(rr_v)}, {spatial_d(rsd_v)}, {temporal_d(rtd_v)}>

This expression means that in the named context, the concept c takes its value for a property, a relation or a descriptor in a restricted range rp_v, rr_v and rsd_v, rtd_v respectively. For example, two possible views of the concept watercourse are:

Context(flooding) → function (watercourse, evacuation area)
Context(tourism) → function(watercourse, [navigable, skating])

Meaning that in the context of a flooding, the watercourse has the function of evacuation area to allow boats rescuing people. A view is a spatial view when the condition is imposed on a spatial property, a spatial relation (topology, proximity, orientation) or a spatial descriptor:

Spatial View: Context(Name of context) → spatial property (concept, value of spatial property)
Spatial View: Context(Name of context) → spatial relation (concept, range of spatial relation)
Spatial View: Context(Name of context) → spatial descriptor (concept, value of spatial descriptor)

A view is a temporal view when the condition is imposed on a temporal property, a temporal relation or a temporal descriptor:

Temporal View: Context(Name of context) → temporal property (concept, value of temporal property)
Temporal View: Context(Name of context) → temporal relation (concept, range of temporal relation)
Temporal View: Context(Name of context) → temporal descriptor (concept, value of temporal descriptor)

Besides views, dependencies between features can be inferred to semantically augment a concept. Dependencies express that a first feature's values are related to a second feature's values. For example, property "temperature" depends on property "altitude". We formalize dependencies with rules head → body. The body in the rule is a consequence of the head. Here are examples of thematic, spatial and temporal rules respectively:

Altitude(land, low)→ FloodingRisk(land, high)
Width(watercourse, larger than 7m)→ Geometry(surface)
Flooding frequency(land, more than twice a year)→ Status(land, periodically waterlogged).

Dependencies are rarely represented. However, they may be implicit in the concept's instances; for example, cities with similar values of average temperature have similar values of altitude. The concept, the views and the augmented dependencies form the MVAC:

c^{MVA} = < n(c), {p(c)}, {r(c)}, {spatial_d(c)}, {temporal_d(c)}, {v(c)}, {ctx}, {dep(c)} >

where {v(c)} is the set of views, {ctx} is a set of different contexts for the concept, and {dep(c)} is the set of augmented dependencies. The methodology that will augment a concept to a MVAC is composed of two main methods, a view extraction method, and a method to discover dependencies.

4. MVAC GENERATION METHOD

We have developed this method to transform a concept into a MVAC. The method integrates view extraction paradigm, mining rules techniques and ontology reasoning principles. Fig. 2 shows the MVAC generation method. It consists of two phases: 1) the *view extraction* phase, 2) the *augmentation* phase. The method takes as input an ontology with original concepts as defined in section 3. The first step involves the user in specifying the context extraction rules.

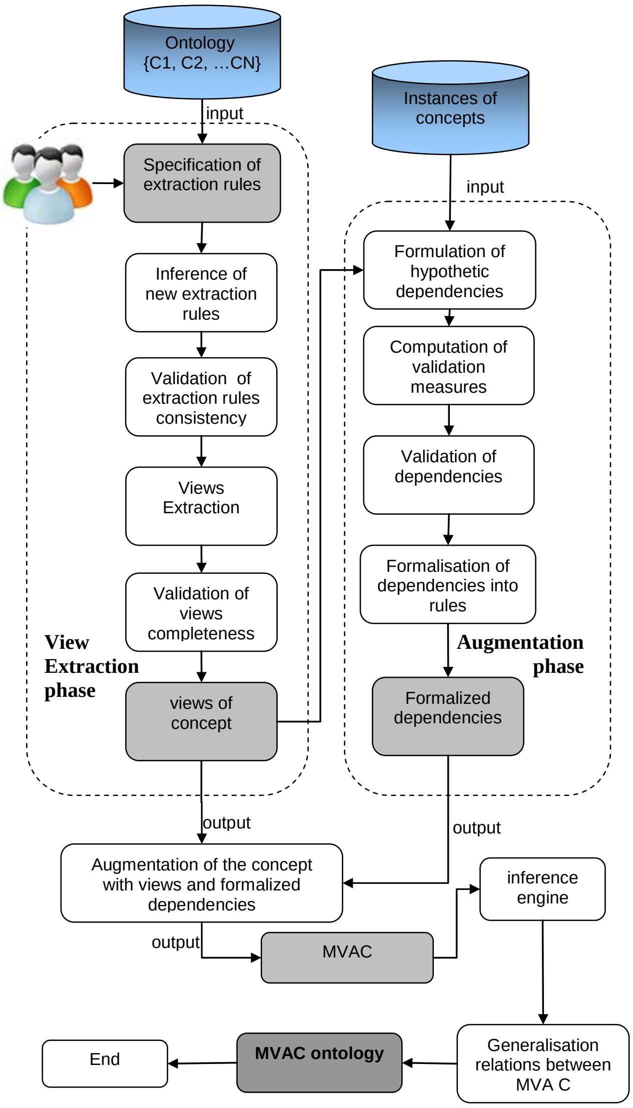

Ontology
{C1, C2, ...CN}
input
Instances of concepts
Specification of extraction rules
input
Inference of new extraction rules
Formulation of hypothetic dependencies
Computation of validation measures
Validation of extraction rules consistency
Validation of dependencies
Views Extraction
Formalisation of dependencies into rules
Validation of views completeness
View Extraction phase
Augmentation phase
views of concept
Formalized dependencies
output
output
Augmentation of the concept with views and formalized dependencies
inference engine
output
MVAC
End
MVAC ontology
Generalisation relations between MVA C

Figure 2. MVAC and Ontology Generation Method

Step 1. Specification of Context Extraction rules. This step requires interaction between users and the view extraction algorithm. The users specify with context rules the values of the properties, relations and descriptors of a concept that are valid in a context. For example, considering the concept "watercourse" with properties "depth" and "category of watercourse", the user specifies their possible values in the context of dryness:

Context(dryness) → water level(watercourse, low) (rule 1)
Context(dryness)→ category of watercourse (watercourse, intermittent) (rule 2)

The contexts of a concept are inferred from those rules.

Step 2. Inference of new Extraction rules. Having a set of extraction rules on contexts of the concept, we verify if new extraction rules can be inferred by combining them. We also use other existing rules that are part of the ontology, and which represent the knowledge of domain experts. This is a way of reusing the existing knowledge to produce new one. The inference of new extraction rules (1) takes as input the extraction rules specified in step 1, plus the rules that are part of the ontology, (2) send them to an inference mechanism, (3) produces new inferred rules, and (4) restart the cycle from (1) to (3) until no new rules are inferred. The inference mechanism determine that if the body of a rule implies the head of a second rule, then the head of the first rule implies the body of the second rule. For example, consider a rule saying that intermittent watercourse are represented with moving polygon: Category of watercourse(watercourse, intermittent) → geometry(watercourse, moving polygon). From this rule and the ones that were specified by the user in step 1, we can infer the following new rule: Context(dryness) → geometry(watercourse, moving polygon) (rule 3). New inferred rules are added to the set of rules that will be used to extract views of the concept.

Step 3. Validation of extraction rule consistency. Before using those rules to extract the views of a concept, we verify if the inferred rules are correct, that is, if they are consistent with the reality. In this case, the reality corresponds to the instances of the concept, which are representation of real world objects stored in the database. To verify is the rules are correct, we assess the consistency between the rules and the instances. Consistency can be defined as the degree of consistency of the data with respect to its specifications (Mostafavi et al. 2004). In our context, data corresponds to instances whereas specifications correspond to rules (since rules define the semantic). Therefore, a rule is consistent if the instances of the concept verify this rule. For example, if we have a rule Context(dryness) → water level(watercourse, low), we verify if instances of the concept "watercourse" which have the context "dryness", also have "low water level". To determine whether an extraction rule is consistent enough, we propose a ratio that will compare the number of instances that respect the rule (denoted with |verifying instances|) with the total number of instances which have for context the one indicated in the rule (denoted with |targeted instances|). Only those rules that have a sufficient degree of consistency are used for view extraction:

$$\text{Degree of consistency } = \frac{|\text{verifying instances}|}{|\text{targeted instances}|} \qquad (1)$$

Step 4. View Extraction. View extraction, as we have defined in (Bakillah et al. 2009), includes two main steps, the extraction of partial views and the merging of partial views. First, in the extraction of partial views, each extraction rule is applied to the concept to create the subconcept that will always respect this rule. For example, for the concept watercourse defined in section 3, applying rule 1 gives the following partial view:

Partial view: Context(dryness) → <watercourse, {water level(low), category(intermittent, stable), spatial extent(polygon, moving polygon), function(navigation, skating, evacuation area), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, flooded area, frozen area)}{waterlogged period(average flooding period)}>

This partial view imposes a restriction only on the values of property "water level". In the second step of the view extraction, all partial views that pertains to a same context and that are non contradicting are merged into a single view. This is the partial view merging process. For example, merging partial views generated by rule 1 to 3 would lead:

view: Context(dryness) → <watercourse, {water level(low), category(intermittent), spatial extent(moving polygon), function(navigation, skating, evacuation area), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, flooded area, frozen area)}{waterlogged period(average flooding period)}>

During the view extraction, relations between views of a concept and other concepts of the ontology are inherited from the definition of the concept when it applies; for example, the above view is linked to the concept "waterbody" with the spatial relation "connect".

Step 5. Validation of view completeness. When all views of a concept are created, we verify if they are complete, that is, the union of all views of the concept result in the concept itself. The restricted range of a property p_i (or relation R_i, descriptor d_i) in a view v_j is r_{ij}. The view completeness can be validated if the following generic expression is verified: $c = < n(c), \{p_1(r_{11} \cup r_{12} \cup r_{13} ...), ... p_n(r_{n1} \cup r_{n2} \cup r_{n3} ...) \}, \{R_1(r_{11} \cup r_{12} \cup r_{13} ...), ... R_n(r_{n1} \cup r_{n2} \cup r_{n3} ...) \}, \{d_1(r_{11} \cup r_{12} \cup r_{13} ...), ... d_n(r_{n1} \cup r_{n2} \cup r_{n3} ...) \}>$, that is, by taking, for all features of the concept, the union operator on the restricted ranges of all views of the concept. The next steps are about augmenting the concept (with its views) with implicit dependencies.

Step 6. Formulation of possible dependencies. Possible dependencies are dependencies that have to be verified against data. For every view of a concept, our method formulates dependencies that express relations between each pair of their features (properties, relations or descriptors). Those dependencies are expressed as rules. For example, for a concept "watercourse" with properties "state (frozen, unfrozen)" and "function (skating, navigable)", we can have:

"If *state* of watercourse = frozen, then *function* = skating"
"If *state* of watercourse = frozen, then *function* = navigable"
"If *state* of watercourse= unfrozen, then *function* = skating"
"If *state* of watercourse= unfrozen, then *function* = navigable"

"If *function* of watercourse = skating then *state* = frozen"
"If *function* of watercourse = skating then *state* = unfrozen"
"If *function* of watercourse = navigable then *state* = frozen"
"If *function* of watercourse = navigable then *state* = unfrozen"

Because the number of possible dependencies may be high, they can be classified (the first series being classified as "function depends on state" rules, and the second as "state depends on function" rules) so that the user can reject the ones that seems non-verifiable. Once we have formulated a set of possible dependencies, we have to validate which ones are true among instances of a view.

Step 7. Computation of rule validation measures. For each rule expressing a possible dependency, we determine the values of two measures that will help to determine if we can retain it as a valid dependency. Those measures, which are *support* and *confidence*, are adapted from the rule-mining domain, which aims at finding correlations between items in datasets (Ceglar and Roddick, 2006). The support measure how many instances respects either the head (Ihead) or the body(Ibody) of a rule, with respect to the total set of instances (Itotal), and the confidence measures how many instances respect the body of the rule among those that respect the head of the rule:

$$\text{Support} = \frac{|\text{Ihead} \cup \text{Ibody}|}{|\text{Itotal}\ |} \tag{2}$$

$$\text{Confidence} = \frac{|\text{Ibody}|}{|\text{Ihead}|} \tag{3}$$

Step 8. Validation of dependencies. For the validation of dependencies, we choose those dependencies for which support and confidence are satisfying. Those measures complete each other since a high confidence but a low support means while this rule is usually respected, it is not frequent in the instance set, so it may be less interesting.

Step 9. Formulation of dependencies into rules. If the rule checked in step 4 is determined to be true, then it is added to the definition of the view in a form: Feature 1(concept, value of feature 1) → Feature 2(concept, value of feature 2).
Now that views and dependencies are extracted, the concept's definition is rewritten with those new elements. However, relations between views and augmented concepts need to be re-computed to form the MVA ontology.

Step 10. The inference of Relations. Views needs to be linked together by generalisation/specialisation relations to create the MVA ontology. Those links are established between the different views of a same concept, and between views of different concepts. Generalisation is when the instances of a first view /concept include all instances of a second view/concept. To perform this task, we can, for example, express MVACs with OWL-DL language and use subsumption-reasoning mechanism provided by reasonign engines. For example, if we have the following view:

View1: Context(dryness) → <watercourse, {water level(low), category(intermittent), spatial extent(moving polygon), function(non navigable, skating), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, frozen area)}{waterlogged period(average flooding period)}>,

it would generalise the following view:

view2: Context(dryness in summer) → <watercourse, {water level(low), category(intermittent), spatial extent(moving polygon), function(non navigable), state(unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse,)}{waterlogged period(average flooding period)}>

which represents a smaller number or real world objects. Therefore, views can be categorised within the MVA ontology.

5. CASE STUDY

Having defined the MVA model and a method to generate it from an existing concept, we aim to show with the following examples that the MVAC can help to improve semantic interoperability. Consider the user of a geospatial database which ontology contains the following concept "watercourse":

C1: <**watercourse**, {water level(low, high), spatial extent(polygon, moving polygon), function(navigable, non navigable}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, waterlogged area)}>

Suppose that this user search a network of geospatial databases for "watercourses" in the context of "dryness".

Consider the concept "stream" which is included in the ontology of another database of the network.

C2: <**stream**, {depth(low, high), spatial extent(surface, moving surface), role(navigable, non navigable)}, {Meet(Lake)}, {geo-entity(bed of watercourse, waterlogged area)}>

First, with no views being defined, and therefore no contexts being specified, we are unable to find if "stream" and "watercourse" can be in a similar context of "dryness". With a lexical matching approach, we would however find pairs of synonyms: "watercourse" ↔ "stream", "polygon"↔ "surface", "connect" ↔ "meet", "waterbody" ↔ "lake", "function" ↔ "role". With semantic mapping rules such as those that were presented in (Bakillah et al. 2009), we would find that "watercourse" overlap "stream", but note that we would be unable to identify that water level corresponds to depth since those properties are not lexically related. Now consider that we employ the MVA generation method we have developed and we build MVACs for "watercourse" and "stream". Suppose we have extracted two views for the concept watercourse, corresponding to contexts dryness, and flooding:

MVAC1: Watercourse
View1(watercourse): Context(**dryness**) → {water level(low), spatial extent(polygon), function(non navigable}}, {Connect(Waterbody)}, {geo-entity(bed of watercourse)}>
View2(watercourse): Context(**flooding**) → <watercourse, {water level(high), spatial extent(moving polygon), function(navigable)}, {Connect(Waterbody)}, {geo-entity(waterlogged area)}>.
In addition to the following dependencies being extracted for "watercourse":
{(d1:water level(watercourse, low)→ function(watercourse, not navigable), (d2:water level(watercourse, high)→ function(watercourse, navigable)}

For the concept "stream", we have for example extracted:

MVAC2: Stream
View1(stream): Context(**lack of rain**) → <**stream**, {depth(low), spatial extent(surface), role(non navigable)}, {Meet(Lake)}, {geo-entity(bed of watercourse)}>
View2(stream): Context(**rain season**) → <**stream**, {water level(high), spatial extent(moving surface), role(navigable)}, {Meet(Lake)}, {geo-entity(waterlogged area)}>
And the following dependencies:
{d3:(depth(stream, low)→ role(stream, not navigable), (d4:depth(stream, high)→ function(stream, navigable)}.

We show how the MVAC will enable to improve answering to the user query by detecting implicit matches using the structure of the MVAC. After having deduced the lexical matches indicated above, comparing the different dependencies of C1 and C2, we find that d1 has the same structure as d3, and d2 the same structure as d4, which allow proposing the following match: Water level↔Depth. We were able to find this match only because we augment the concept with dependencies that brings a richer structure. Comparing the contexts of the different views of "watercourse" and "stream" from a lexical-based approach does not allow finding that "lack of rain" corresponds to "dryness". However, if we compare the definitions of **View1**(stream) and **View1**(watercourse), knowing the previous matches, we find that **View1**(stream) is equivalent to **View1**(watercourse), which allow to propose the following match: Context(**lack of rain**) ↔ Context(**dryness**). This allows the user finally to retrieve "stream" as a concept similar to "watercourse" in the context of dryness. This example shows that augmenting the concept with new structures (views and dependencies) can help to match concepts, contexts or features of concepts that seems dissimilar, and supports improving semantic interoperability between geospatial databases.

6. CONCLUSIONS

In this paper, we have argued that for improving semantic interoperability approaches, one main problem is the poor definition of concepts. This is especially true regarding the geospatial domain where concepts are defined by spatial and temporal features, in addition to multiple contexts and implicit dependencies between features. To address this issue, we have proposed the Multi-View Augmented Concept Model (MVAC), and a MVAC generation approach that includes a view extraction and semantic augmentation methods. We have shown that with the MVAC, we can improve semantic interoperability because we can discover more semantic relations between concepts of different ontologies. Therefore, the MVAC can play an important role in a global semantic interoperability approach designed for ad hoc networks where ontologies of databases are very heterogenous, such as in disaster management and in environmental and health domains. The future research will consider the MVAC as a basis for such an approach, with the goal of developing a semantic interoperability approach that is adapted to the MVAC model, since the quality of semantic interoperability depends on the ability of the semantic mapping approach to consider all the characteristics of the input concepts (Bakillah et al. 2008).

References

Bakillah, M., Mostafavi, M.A., Brodeur J., Bédard, Y., 2009. Sim-Net: a View-Based Semantic Similarity Model for ad hoc Network of Geospatial Databases. *Transactions in GIS*, 13(5-6), 417-447.

Bakillah, M., Mostafavi, M. A., Brodeur, J., and Bédard, Y., 2008. *Elements of Semantic Mapping Quality*. Boca Raton, FL, CRC Press, 37–45.

Bhatt, M., Flahive, A., Wouters, C., Rahayu, W., Taniar, D., 2006. MOVE: A distributed Framework for Materialized Ontology Views Extraction. *Algorithmica*, 45(3), 457-481.

Bédard, Y., E. Bernier & R. Devillers, 2002. La métastructure VUEL et la gestion des représentations multiples. *Généralisation et Représentation multiple*, Hermès, 150-162.

Bennett, B., 2005. Modes of concept definition and varieties of vagueness. *Applied Ontology*, 1(1), 17–26.

Bian, L., Hu, S., 2007. Identifying Components for Interoperable Process Models using Concept Lattice and Semantic Reference System. *International Journal of Geographical Information Science*, 21(9), 1009-1032.

Brodeur, J., Bédard, Y., 2001. Geosemantic Proximity, a Component of Spatial Data Interoperability. International Workshop on "Semantics of Enterprise Integration" ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications, Tampa Bay, Florida.

Ceglar, A. and Roddick, J. F., 2006. Association mining. *ACM Computing Surveys*, 38(2), 5.

Egenhofer, M., 1993. A Model for Detailed Binary Topological Relationships. *Geomatica*, 47, 261-273.

Kavouras, M., Kokla, M., 2008. *Theories of geographic concepts*. CRC Press, Taylor & Francis Group.

Keßler, C., Raubal, M., Janowicz, K., 2007. The Effect of Context on Semantic Similarity Measurement. *On the Move to Meaningful Internet Systems 2007,* 1274-1284.

Medin, D.L., Rips, L.J., 2005. Concepts and Categories: Memory, Meaning and Metaphysics, Concepts and Categorization. *The Cambridge Handbook of Thinking and Reasoning*. Ed. K.J. Holyoak and R.G. Morrisson, 37-72. Cambridge, UK: Cambridge University Press.

Mostafavi, MA., Edwards G., Jeansoulin, R., 2004. Ontology-based Method for Quality Assessment of Spatial Data Bases. In: *Proceedings of ISSDQ'04*, GeoInfo Series, Bruck/Leitha, Austria, 49-66.

Parent, C., Spaccapietra, S., Zimanyi, E., 2006. The MurMur Project: Modeling and Querying Multi-Representation Spatiotemporal Databases. *Information Systems*, 31(8) 733-769.

Park, J., Ram, S., 2004. Information systems interoperability: what lies beneath? *ACM Transactions on Information Systems*, 22(4), 595-632.

Rodriguez, A., Egenhofer, M., 2003. Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15 (2), 442–456.

Schwering, A., and Raubal M., 2005. Measuring Semantic Similarity Between Geospatial Conceptual Regions. Proceedings of the First International Conference on Geospatial Semantics, Mexico City, Mexico.

Tomai, E., Kavouras, M., 2004. From "Onto-GoeNoesis" to "Onto-Genesis": The Design of Geographic Ontologies. *GeoInformatica*, 8(3), 285-302.

Wouters, C., Dillon, T.S., Rahayu, W., Meersman, R., Chang, E., 2008. Extraction Process Specification for Materialized Ontology Views. T.S. Dillon et al. (Eds.): *Advances in Web Semantics I*, LNCS 4891, 130–175.

Acknowledgements

This research supported by an operating grant from Natural Sciences and Engineering Research Council of Canada (NSERC).

GRAPH BASED RECOGNITION OF GRID PATTERN IN STREET NETWORKS

Jing Tian [a,b]*, Tinghua Ai [a,b], Xiaobin Jia [a,b]

[a] School of Resource and Environment Science, Wuhan University, 129Luoyu Road, Wuhan, 430079,China-(yutaka-2010@163.com, tinghua_ai@tom.com, jiaxiaobin_123@126.com)
[b] Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129Luoyu Road, Wuhan, 430079, China

KEY WORDS: Spatial information Sciences, Cartography, Generalization, Pattern, Networks, Vector

ABSTRACT:

Pattern recognition is an important step in map generalization. Pattern recognition in street network is significant for street network generalization. A grid is characterized by a set of mostly parallel lines, which are crossed by a second set of parallel lines with roughly right angle. Inspired by object recognition in image processing, this paper presents an approach to the grid recognition in street network based on graph theory. Firstly, the bridges and isolated points of the network are identified and deleted repeatedly. Secondly, the similar orientation graph is created, in which the vertices represent street segments and the edges represent the similar orientation relation between streets. Thirdly, the candidates are extracted through graph operators such as finding connected component, finding maximal complete sub-graph, join and intersection. Finally, the candidate are evaluated by deleting bridges and isolated lines repeatedly, reorganizing them into stroke models, changing these stroke models into street intersection graphs in which vertices represent strokes and edges represent strokes intersecting each other, and then calculating the clustering coefficient of these graphs. Experimental result shows the proposed approach is valid in detecting the grid pattern in lower degradation situation.

1. INTRODUCTION

Ideas of scale and pattern are central to the process of interpretation in the geosciences. The patterns that are evident at any given scale will have specific causes and consequences (Mackaness, 2007).

What is pattern and pattern recognition? In tradition, pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. Depending on the application, these objects can be images or signal waveforms or any types of measurements that need to be classified (Theodoridis and Koutroumbas, 2004). The “pattern” refers to objects. From the geosciences experts’ view, patterns can be organized into partonomic hierarchies reflecting relations between parts and subparts (Mackness and Edwards, 2002). In summary, pattern can be an object, object cluster, property with an object or object cluster, and relation between objects.

Pattern recognition techniques have been applied to machine vision, character recognition, and computer-aided diagnosis. These applications use images as information source and benefit from recognition of specific pattern in images. In the raster mode, the main tasks of pattern recognition are object detection and feature extraction. Here, street must be extract from image and this work is the focus. In vector mode, however, the elementary objects points, lines and polygons already exist in spatial database. Here, streets already exist in geographical spatial database and the recognition of relations and structures among streets is our interest. Traditional pattern recognition pays little attention to the latter.

In recent years, pattern recognition of spatial cluster object has come to assume increasing importance in the map generalization field. The relation between pattern and map generalization is that pattern recognition is the start point of map generalization, and pattern maintaining is the aim of the map generalization. Maintaining patterns after generalization is not only the generalization guideline but also the generalization result.

Identifying and extracting the spatial patterns of street networks is significant for network generalization. Maintaining the pattern of the street network after generalization is the basic principle in selection of streets. The pattern here refers to the structure of street network. However, most street network generalization methods pay no attention to principle. Zhang (2004) emphasized “modeling the patterns as input to generalization algorithm”.

Perceptual organization refers to the ability of human beings to form meaningful groupings of low level tokens detected in an image (Sarkar and Boyer, 1994). Sarkar and Boyer (1994) present a hierarchical approach to perceptual organization for detecting structure among tokens. Tokens are perceptual objects which are to be organized. The tokens may be points, straight lines, curves, parallel lines and rectangle, etc. The tokens at each level have specific attributes, and then some relation graphs whose nodes represent tokens and edges represent the association among tokens is created using these attributes. For example, proximity graph is a graph whose links join nodes represent tokens having points close together. Tokens at high level are extracted from the tokens at low level by standardized graph operations.

Inspired by Sarkar and Boyer’s idea (1994), this paper presents an approach to recognition of grid pattern in the street network. A grid is characterized by a set of roughly parallel lines, which are crossed by a second set of parallel lines with roughly perpendicular angle. A set of parallel lines consists of line segments having proximity and similar orientation. The parallel lines and the grid are extracted by graph theoretic operators such as finding connected component, finding maximal complete sub-graph, join and intersection.

* Jing Tian. e-mail: yutaka-2010@163.com; phone: 86-13628636229.

The rest of this paper is organized as follows. Section 2 reviews related work on the recognition of patterns in street network. Section 3 introduced concepts and definitions in graph theory which are used in this paper. Section 4 presents the approach to the recognition of grid pattern in detail. An experiment is conducted in order to test the approach and advantages and disadvantages are discussed in Section 5. Finally, section 6 concludes the paper and offers several avenues for future work.

2. RELATED WORK

2.1 Main patterns in street network

Street networks consist of a large amount of streets which intersect with each other in different mode. The patterns can be classified according to different criterions. From geometric view, In Zhang's research (2004), discernable patterns in road networks are star-like, grid-like and irregular pattern. Marshall (2005) proposed that main patterns are linear, treelike, radial, cellular and hybrid forms. Heinzle et al (2006) suggested the typical patterns are strokes as type of linear forms, grids as type of cellular forms, star as type of radial forms and circular roads as type of cellular forms. It is important to note that the grid pattern in Heinzle et al (2005) only consists of streets which are parts of grid. The grid is defined in local sense. From topological view, there are three important patterns: regular, small-world and random (Watts and Strogatz, 1998).If criterion is the distribution of the node degrees in the network, the random and scale-free pattern can be used in such classification. Xie and Levinson (2007) defined four typical connection patterns in road networks. They are Ring and Web form in circuit networks and Star and Hub-and Spoke form in branching networks. Circuit is defined as a closed path that begins and ends at the same vertex. Branching networks consist of sets of connected lines without any complete circuit.

2.2 Characteristics and Quantification Measures of Patterns

A grid is characterized by a set of roughly parallel streets, which are crossed by a second set of parallel streets with roughly perpendicular angle. The main quantification measure is distribution of street direction from statistical view. A star is characterized by a "fuzzy center" degree from which streets radiate or converge. If the fuzzy center is clustered to one intersection, the connectivity degree of it is high compared with other intersections. A ring is approximate a circle. There are rings in different sizes: roundabouts, rings around the hub-like junction, rings surrounding a city center. The ring can be described by shape index: compactness, convexity, and geometric moments. To check whether a network is a small-world, characteristic path length and clustering coefficient are proposed. For scale-free network, degree and degree distribution are used. The degree of a node is the number of edges incident with node. The degree distribution is defined as the probability of nodes with specific number of links. Porta et al (2006) found the clustering coefficient of grid pattern is extremely small. It indicates that the geometric pattern and topological pattern of street network have intrinsic relation.

2.3 Methods of Recognition of Patterns

For grid pattern, Heinzle et al (2005) select so called CRS nodes, at which four edges intersect, as starting points, investigate the histogram of the edge directions, and then examine the polygon regarding its similarity to the neighbor polygons. Yang et al (2010) proposed a method for identifying grid pattern. Their method first builds a node-edge topology, generates polygons from these edges, and then calculates a set of parameters, and finally identifies grid via a multi-criteria decision method. These two methods for grid recognition are both polygon-based structural recognition approach.

For star pattern, Heinzle et al (2005, 2007) take any node in the graph as potential center point, use the Dijkstra algorithm to compute the single source shortest path from this node to all other nodes, and then use a circle with a certain radius to search the center. The fuzzy center is determined by a cluster algorithm. The problem of this approach is, as they pointed out, deficiency of knowledge about the search radius and the range of the configurations.

For ring pattern, Heinzle et al (2006) use convex hull peeling techniques to determine the approximate position for the city center, neighbouring polygons are aggregated as long as they are not separated by a major stroke subsequently, and then some candidates are sought, the geometric moments curvature and convexity are used to evaluate the candidates. The problem of this approach is the huge uncertainty to evaluate whether the candidate is a ring.

It is easy to check whether a network is a small-world network. The small-world has a small characteristic path length as random graphs and a much larger coefficient than random graph. For scale-free networks, it is easy to check whether the degree distribution follows a power law.

2.4 Summary

We can conclude from analysis above: ⑴ There are much work on classification and qualitative description of geometric pattern in street network, but little work on how to quantify the characteristics and how to detect geometric patterns. ⑵ It is easy for people to recognition of geometric patterns cognitively and difficult to formalize them, whereas it is difficult for people to recognition of topological patterns and easy to formalize them. The association of various measurements has not been studied.

This paper proposes a method for detecting grid pattern. The major difference between our proposed method and the methods of Heinzle et al (2005) and Yang et al (2010) is that our method detects grid patterns by street segments grouping.

3. CONCEPTS AND DEFINITION IN GRAPH THEORY

Street network is usually modeled using a graph, where vertices represent intersections and edges represent street segments. This section introduces some basic graph concepts and definitions which will be used in the next two sections. For a more complete introduction to graph theory, readers can refer for example to Diestel (2006).

A **graph** of G consists of a finite set of **vertices** (or nodes, or points) and a finite set of **edges** (or links, or lines). A graph is often denoted as $G(V,E)$ where $V=\{v_1,v_2\cdots v_n\}$ and $E=\{v_i,v_j\}$. Two vertices v_i , v_j of G are **adjacent**, if v_iv_j is an edge of G .The number of vertices of a graph G is its **order**. G is **undirected graph** if E is unordered set. A **complete graph** $G=(V,E)$ is one in which each vertex is connected to every other vertex. Let

$G=(V,E)$ and $G'=(V',E')$ be two graphs. If $V'\subseteq V$ and $E'\subseteq E$, then G' is a **subgraph** of G. A **maximal complete subgraph** is a complete subgraph of G that contains as many vertices of G as possible. A non-empty graph is called **connected** if any two of its vertices are linked by a path in G. A maximal connected subgraph of G is called **connected component** of G.The degree of a vertex is the number of edges at this vertex. A vertex of degree 0 is **isolated**. A vertex of a connected graph is an **articulation vertex** if its removal would create a separate graph. An edge of a connected graph is a **bridge** if deleting it would create a separate graph. We set $G\cap G':=(V\cap V',E\cap E')$, where $\cap$ is **intersection** operation. **Join** is an operation between two graphs that linking every vertex of one graph to every vertex of the other graph.

A **relation graph** is a graph in which the vertices represent entities and the edges represent specific relationships among entities. The structural representation of a street network proposed by Jiang and Claramunt (2004) is a relation graph, where vertices represented named streets and edges represent intersection relation among streets. The **street segment intersection graph** is a graph whose vertices represented street segments and edges represent intersection relation among street segments. The **similar orientation graph** is a graph whose vertices represented street segments and edges represent similar orientation among street segments. The orientation of the street segment is defined by its azimuth angle, whose range is [0°, 180°). The range of angular separation between two street segments is [0°, 90°]. The **street intersection graph** is a graph whose vertices correspond to entire streets and edges correspond to intersection relation among entire streets. The **entire street** consists of street segments which satisfy "good continuation" rule (Thomson and Richardson, 1999).

4. RECOGNITION OF GRID PATTERN IN STREET NETWORKS

4.1 Characterization of Grid

If the element considered is line, a grid is characterized by a set of roughly parallel streets, which are crossed by a second set of parallel streets with roughly perpendicular angle. The perpendicular angle is no compelling characteristic of a grid. If the element considered is polygon, a grid is composed of a set of polygons with similar shape and connect mode. Similar to viewpoint of the Heinzle et al (2005) and Xie and Levinson (2007) , we define grid as a closed structure consist by a set of mostly parallel lines, which are crossed by a second set of parallel lines, not include other lines. Hence, for the recognition of a grid, the key questions arise: ⑴ How can we find the two sets of parallel lines? ⑵ How can we evaluate whether the result is grid or not?

In the perceptual organization, Sarkar and Boyer (1994) chose length of two lines, distance perpendicular to the smaller line from its midpoint to larger line and angular separation to describe parallelism. The value of these parameters' threshold is decided using statistical method and experiments. By formulating the perceptual characteristics leading to human recognition of parallelism in terms of several distinctive forces, Ip and Wong (1997) developed a force-driven model as a new optimization strategy to perform correspondence establishment between points in the matching curves. To detect parallel segments after coupling, they first make sure that the orientations fall within an allowable limit. From analysis above, we can conclude that the angular separation between two segments is the most important factor for recognition of parallelism.

Porta et al (2006) found the clustering coefficient of grid pattern is extremely small. It indicates that the grid pattern and clustering coefficient have intrinsic relation. In fact, transforming a perfect grid to street intersection graph (vertices represent entire streets and edges represent intersection relation), the value of clustering coefficient of this graph is 0. The value in Porta et al's experiment is not 0 because there exists some disturb streets shown in Figure 1. The grid pattern referred by Porta et al (2006) is in global sense. The clustering coefficient can be used to evaluate the grid.

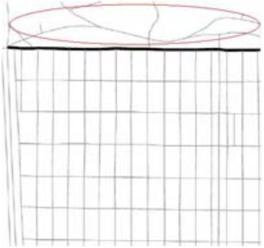


Figure 1. Streets influencing clustering coefficient (modified from Porta et al, 2006)

4.2 The Approach to Recognition of Grid Pattern

The core idea is that grids are composed of two groups of parallel streets. There exist relations among line segments in each group and among line segments in two groups. The angle of every line segment in the same group is similar, including the parallelism and collinearity. The line segments in one group is intersected at least one line segments in another group. Hence, we first take angular similarity as compelling condition. Through the graph-theoretic operations such as finding connected component, finding maximal complete subgraph, intersection and join, the candidates are sought. Finally, the street intersection graph is formed and clustering coefficient is computed to evaluate the candidates. The logic of this idea is that all parallel lines would be found and grouped through their angular separation first and the final result is the subset of these parallel line segments if a grid exists. The candidate is filter through graph-theoretic operations stage by stage. The steps are described in more detail in the following:

Step 1: Data preprocessing. The street network is organized as graph in which the vertices represent street intersections and edges represent street segments, i.e. primal graph. Inspired by Xie et al (2007), identify the bridges, delete bridges and isolated vertices until there are no bridges and isolated vertices in the graph. The reason is that a grid does not contain any bridges and isolated vertices. This step is to clean and refine the street network. Figure 2 shows a simple example. Deleting bridge 13 produces Figure2 (b), removing isolated vertex produces Figure2 (c), and Figure2 (d) is street segment intersection graph of Figure2(c).

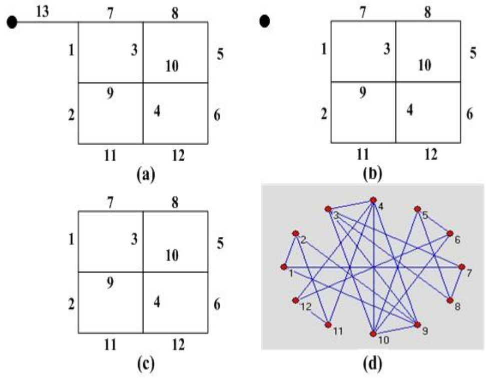


Figure 2. Simple Example ((d) visualization by Pajek)

Step 2: Street grouping. The similar orientation graph is formed in which the vertices represent street segments and edges represent street segments having similar orientation. Figure 3 shows a similar orientation graph of Figure 2(c) with two connected component. There exist no grid or the angular separation is setting too large, if there are only one connected component in similar orientation graph. For example, if the angular separation is 90°, of course there is only one connected component. The intention of this step is that the street segments which are likely to form the grid are extracted, maybe including a lot of street segments disappear in final results.

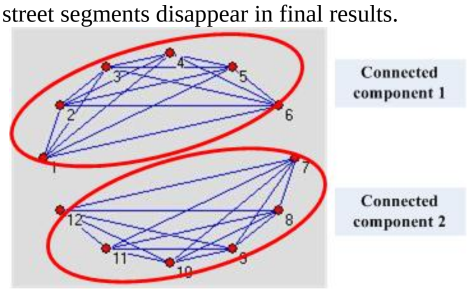


Figure 3. Similar orientation graph with two connected component (visualization by Pajek)

Step 3: Forming sets of parallel streets. For each connected component of similar orientation graph, their maximal complete subgraph is found. The idea of this step is that the orientation of every line segment in one set is similar to each other, including the parallelism and collinearity. This is based on the truth "line segments' angle in one group of parallel lines is similar to each other." This step eliminates the streets which are not parts of a grid. Figure 4 shows the maximal complete graph of two connected component in Figure 3. The Maximal complete graph is the same as the connected component because the connected component self is complete graph. See section 5.1 for general case.

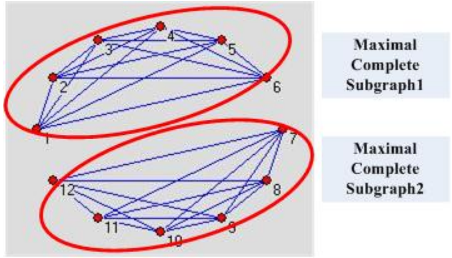


Figure 4. Maximal complete subgraph (visualization by Pajek)

Step 4: Linking two sets of parallel streets to form candidates. Select two maximal complete subgraph, "Join" them, then "intersection" with the street segment intersection graph. Figure 5 shows the "Join" and "Intersection" operation.

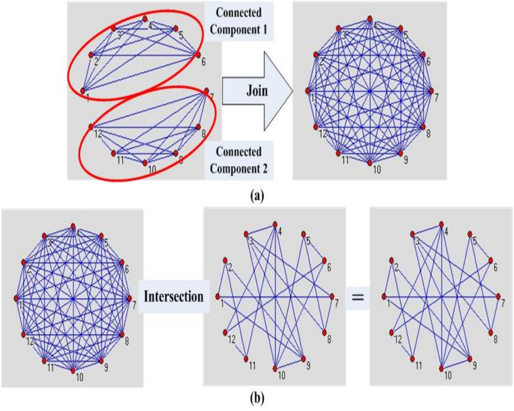


Figure 5. (a) "Join" and (b) "Intersection" (visualization by Pajek)

Step 5: Data post-processing. Delete the bridges and isolated vertices until there are no bridges and isolated vertices in the street network after processing by step 4. This step considers the closed property of a grid. Figure 2 (c) shows a candidate of Figure 2(a). See Figure 10 and Figure 11 for a complex example.

Step 6: Evaluation. For each candidate, the street segments are linked to form entire streets using "good continuation" rule (Thomson et al 1999), then the street intersection graph is created. If the order of the graph is less than 6, then delete the candidate. The reason is that the minimal grid contains 6 entire streets. The clustering coefficient of this graph is computed. The candidate is written out as final result if its clustering coefficient value is 0. Figure 6(a) shows the entire street of Figure 2(c). Figure 6(b) shows the street intersection graph of Figure 6(a).

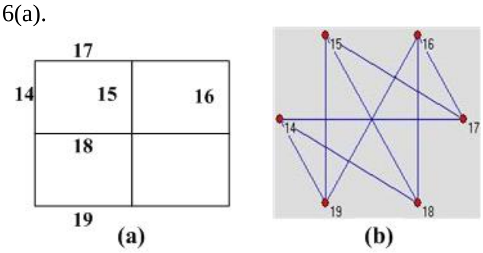


Figure 6. (a) Entire streets and (b) Street intersection graph (visualization by Pajek)

5. EXPERIMENTS AND DISCUSSION

5.1 Experiments

The experiments are conducted on Windows XP platform. The function of creating maximal complete sub-graph is performed using the Graph Theory Toolbox (Iglin, 2003) associated with Matlab (version 7.1). The clustering coefficient is computed using Pajek (Batagelj et al, 1997). Other functions are implemented in Visual C++ (Version 6.0) and integrated into a map generalization system named DoMap. The experiment data is digitized from Wang et al (1993) using ArcGIS (version 9.2). The experimental data is show in Figure 7. The number is the street segment label.

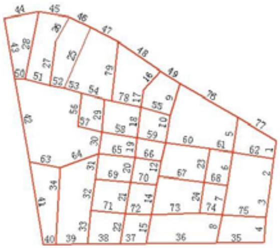


Figure 7. Experimental data

The angular separation threshold is 20°. The 20° is due to the accumulative effect of street segment. Then the similar orientation graph is formed and its connected components are found. The first component includes street segments: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 41, 42, 43, 56, 79.The maximal complete subgraph includes street segments: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 56, 79.The corresponding street segment is shown in Figure 8.

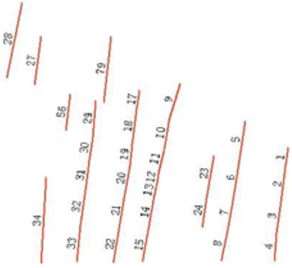


Figure 8. Street segments corresponding to first maximal complete subgraph

The second component includes street segments: 35, 36, 37, 38, 39, 40, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77,78.The maximal complete subgraph includes street segments:35, 36, 37, 38, 39, 40, 45, 50, 51, 52, 53, 54, 55, 57, 58, 59, 60, 61, 62, 63, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 78.The corresponding street segment is shown in Figure 9.

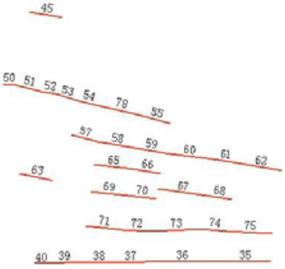


Figure 9. Street segments corresponding to second maximal complete subgraph

We next "Join" the two maximal complete subgraph, then "intersect" with street segment intersection graph created from experimental data in Figure 7. The result is shown in Figure10.

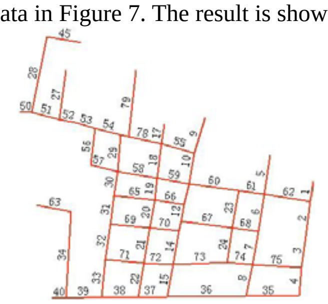


Figure 10. Street segments after Join and Intersection

A candidate is produced after deleting bridges and isolated points until there are no bridges and isolated vertices, as show in Figure 11.

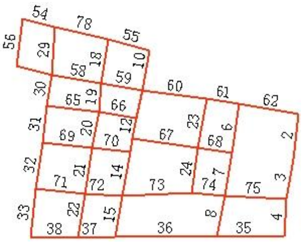


Figure 11. A Candidate

The candidate is processed according to "good continuation" rule to form entire streets, as show in Figure12 (a).The street intersection graph is created (Figure12 (b), and clustering coefficient of this graph is 0.The candidate is written as final result.

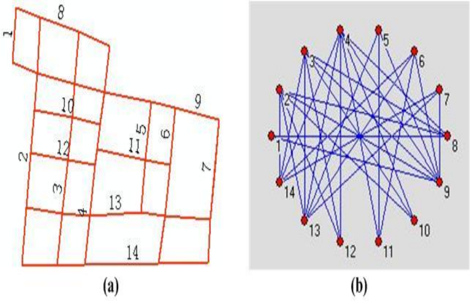


Figure 12. (a)Entire streets and (b)street intersection relation graph (visualization by Pajek)

5.2 Discussion

The experimental results show the proposed approach is valid in detecting the grid pattern in regular and lower degradation situation. The advantages of the proposed approach are: ⑴It uses graph-theoretic operations to represent the human recognition knowledge. i.e. finding a set of roughly parallel streets, which are crossed by a second set of parallel streets with roughly perpendicular angle. ⑵The only parameter is angular separation threshold. In principle, the parameter is set less than 10°in regular case and is set less than 20° in lower degradation cases, which is perceptually satisfactory.

The disadvantages and limitations are: ⑴It is incapable of distinguishing the different grid. The approach proposed by Heinzle et al (2005) can do this task; ⑵The proposed approach requires the initial street network data must be composed of straight line segments. Therefore, we digitize the data into line segments. ⑶Cases similar to Figure13 will be accepted as recognition result by the approach. In fact, these cases are bad results. It should define other evaluation method to exclude these exception cases.

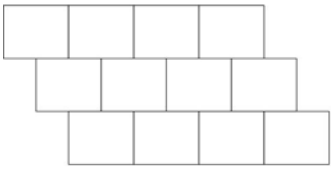


Figure 13 One exception case

6. CONCLUSIONS AND FUTURE WORK

This study is significant for answering the key questions in map generalization on the one hand and extends the research field in pattern recognition on the other hand. This paper presents a graph based approach to recognition of grid in street networks. This approach takes the geometric features as necessary condition. Through the graph-theoretic operations such as finding connected component, finding maximal complete subgraph, intersection and join, the candidates are produced. Finally, the street intersection graph is formed and clustering coefficient is computed. The experiments illustrates that this graph based approach can be used as an effective method for recognition of grid patterns. Compared with the point-polygon based structural recognition approach, this one is line base structural recognition approach.

There are some issues deserves further research. ⑴ The experiment is only conducted on simple data. Experiments on more complex data need to be conducted. The evaluation measures to exclude the exception case need to be developed.⑵ The comparison with other method to solve this problem need to be done to find the best method to a specific kind of cases.⑶ Recognition methods for other patterns in street network need to be developed.

ACKNOWLEDGEMENT

This research is supported by the National High-Tech Research and Development Plan of China under the grant No.2007AA12Z209, and the National Science Foundation, China under the grant No. 40971242. Special thanks go to an anonymous reviewer for constructive comments that substantially improve quality of the paper.

REFERENCES

Batagelj, V, Mrvar, A. 2008. Pajek: Program for Analysis and Visualization of Large Networks. Version 1.22.

Diestel, R. *Graph Theory (3rd Edition)*. 2006. Berlin, Springer-Verlag.

Heinzle, F, Ander. K. H, Sester, M. 2005. Graph Based Approaches for Recognition of Patterns and Implicit Information in Road Networks. In: Proceedings of 22nd International Cartographic Conference, A Coruna, Spain.

Heinzle, F., Ander, K. H., Sester, M. 2006. Pattern Recognition in Road Networks on the Example of Circular Road Detection. In: Raubal.M, Miller.H.J, Frank.A.U, Goodchild.M.F.(Eds): Geographic Information Science, GIScience2006, Munster, Germany, LNCS, vol. 4197, pp.253-267.

Heinzle, F., Ander, K. H. 2007. Characterising Space via Pattern Recognition Techniques: Identifying Patterns in Road Networks. In: Mackaness W A, Ruas A, Sarjakoski L T(Eds): *Generalisation of Geographic Information: Cartographic Modelling and Applications*, Elsevier, pp.233-253.

Iglin, S. 2003. grTheory－Graph Theory Toolbox, http://www.mathworks.com/matlabcentral/fileexchange/4266 (accessed 30 Sep. 2009).

Ip, H., Wong, W. H. 1997. Detecting Perceptually Parallel Curves: Criteria and Force-Driven Optimization. *Computer Vision and Image Understanding*, 68(2), pp.190-208.

Jiang, B., Claramunt, C. 2004. A Structural Approach to the Model Generalization of urban Street Network. *GeoInformatica*, 8(2), pp.157-173.

Mackaness, W., Beard, K. 1993. Use of Graph Theory to Support Map Generalization. *Cartography and Geographic Information Systems*, 20(4), pp.210-221.

Mackaness, W., Edwards, G. 2002. The Importance of Modelling Pattern and Structures in Automated Map Generalization. In: Joint ISPRS/ICA workshop Multi-Scale Representation of Spatial Data, Ottawa, Canada.

Mashall, S. 2005. *Streets and Patterns*. New York: Spon Press.

Muller, J. C., Weibel R, Lagrange J P, Salgé F. 1995. Generalization: state of the art and issues. In: Muller J C, Lagrange J P, Weibel R (Eds), *GIS and generalization: Methodology and practice*. London, Taylor & Francis, pp.3-17.

Porta, S., Crucitti, P., Latora, V. 2006. The network analysis of urban streets: A dual approach. *Physica A*, 369, pp.853-866.

Sarkar, S., Boyer, L. 1994. A computational Structure for Preattentive Perceptual Organization: Graphical Enumeration and Voting Methods. *IEEE Transactions on man and cybernetics*, 24(2), pp.246-266.

Theodoridis, S, Koutroumbas, K. 2009. *Pattern Recognition (4th Edition)*. Amsterdam, Elsevier.

Thomson, R. C., Richardson, D. E. 1999. The "Good Continuation" Principle of Perceptual Organization Applied to the Generalization of Road Networks. In: Proceedings of 19th International Cartographic Conference, Ottawa.

Wang, J.Y. etal. 1993. *Principles of Cartographic Generalization*. Beijing, Surveying and Mapping Press.

Watts, D. J., Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(4), pp.440-442.

Xie, F., Levinson, D. 2007. Measuring the Structure of Road Networks. *Geographical Analysis*, 39, pp.336-356.

Yang, B. S., Luan, X. C., Li, Q. Q. 2010. An adaptive method for identifying the spatial patterns in road networks. *Computers, Environment and Urban Systems*, 34(1) pp.40-48

Zhang, Q N. 2004. Modeling Structure and Patterns in Road Network Generalization. In: ICA workshop on Generalization and Multiple Representation, Leicester.

ESTIMATION OF MODEL ERROR OF LINE OBJECTS

Wenzhong Shi[a] ; Sio-Kei Cheong [b]; Eryong Liu[a, c]

[a]Department of Land Surveying and Geo-Informatics,The Hong Kong Polytechnic University, Hong Kong

[b]Department of Land Management, Renmin University of China, China

[c]School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, China

Commission II

KEY WORDS: model error, truncation error, lines, positional error, geographic information science

ABSTRACT:

This paper presents a method for estimating the model error, specifically, a truncation error, of lines. The generic method for describing truncation error for $n-$ dimensional spatial features is developed first based on the numerical analysis theory. The methods for describing truncation error of the lines in three-dimensional and two-dimensional space, which are the two frequently used cases, are then derived as special cases of the generic method. This research is a further development of earlier work on line error modelling, which were mainly focusing on the propagated error rather than model error.

1. INTRODUCTION

Line feature error is one of the fundamental issues in the areas of modeling uncertainties in spatial data and spatial data quality control. Placing unrealistically high trust in the accuracy of data in GIS may mislead and seriously inconvenience GIS users.

Normally error of lines can be classified as the following categories: (a) model error which is related to the interpolation models for a line, and (b) propagated error which is related to the error of the original nodes that composite the line. The first type error can, in many cases, become a dominant part of the overall line feature error. The second type of error has been widely researched in the past, while study on model error of lines in GIS are relatively less reported. This research intend to investigate the model error of the lines, as a further supplement to the realy studies on propagated error of lines.

Straight line features in GIS include line segments, polylines, polygons, with the latter being regarded as a special representation of a polyline, in which the start point is identical to the end point. Line segments are also a special representation of a polyline – the number of the component line segments being equal to one. Hence, in this study, a polyline is taken as the representative feature for straight line feature error. A third-order curve is also taken as an examplefor truncation error modeling.

2 METHOD FOR DESCRIBING TRUNCATION ERROR OF STRAIGHT LINE FEATURES IN AN $n-$ DIMENSIONAL SPACE

Straight line features in geographic information science can include line segments, polylines, and polygons. The polyline is taken as a generic form of a straight line feature in the following discussion.

2.1 Definition of a polyline in an $n-$ dimensional space

A polyline in $n-$ dimensional space is composed of vectors: $Q_{n1}=[x_{11},x_{12},\cdots,x_{1n}]^T$, $Q_{n2}=[x_{21},x_{22},\cdots,x_{2n}]^T,\cdots,$ $Q_{ni}=[x_{i1},x_{i2},\cdots,x_{in}]^T,\cdots,$ $Q_{nm}=[x_{m1},x_{m2},\cdots,x_{mn}]^T$, where $i=1,2,\cdots,m$ and m is the number of composition points (as shown in Figure 1). The polyline is represented by the line segments $Q_{n1}Q_{n2},Q_{n2}Q_{n3},\cdots,$ $Q_{ni}Q_{n,i+1},\cdots,Q_{n,m-1}Q_{nm}$.

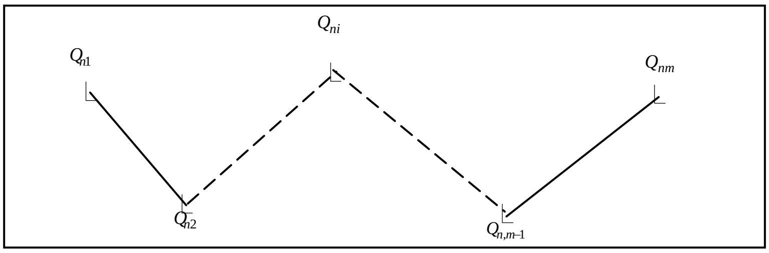


Figure 1. An example of a polyline in $n-$ dimensional space

Any point on the line segment $Q_{ni}Q_{n,i+1}(i=1,2,\cdots,m-1)$ is then represented by

$$Q_{nir}=\begin{bmatrix}x_{i1r}\\x_{i2r}\\\vdots\\x_{inr}\end{bmatrix}=\left(1-r\right)Q_{ni}+rQ_{n,i+1}$$

$$=\begin{bmatrix}(1-r)x_{i1}+rx_{i+1,1}\\(1-r)x_{i2}+rx_{i+1,2}\\\vdots\\(1-r)x_{in}+rx_{i+1,n}\end{bmatrix}\triangleq\begin{bmatrix}p_{i1}(r)\\p_{i2}(r)\\\vdots\\p_{in}(r)\end{bmatrix} \qquad (1)$$

$\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right)$

The true polyline of $Q_{n1}Q_{n2}\cdots Q_{ni}\cdots Q_{nm}$ can be represented by

$$\phi_{nir}=\begin{bmatrix}\mu_{i1r}\\\mu_{i2r}\\\vdots\\\mu_{inr}\end{bmatrix}=\begin{bmatrix}f_{i1}(r)\\f_{i2}(r)\\\vdots\\f_{in}(r)\end{bmatrix}\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right) \qquad (2)$$

2.2 The truncation polyline error in an $n-$ dimensional space

Some assumptions about the functions $f_{ij}\left(r\right)\left(i=1,2,\cdots,m;\right.$ $\left.j=1,2,\cdots,n;r\in[0,1]\right)$ need to be made in order to estimate interpolation error – a model error.

It is assumed that $f_{ij}(r)\in C^2[0,1]$, the following is then obtained

$$\left|R_{ij}(r)\right| = \left|f_{ij}(r)-p_{ij}(r)\right| \le \frac{1}{2}M_{2ij}r\left(1-r\right)$$
$$\le \frac{1}{8}M_{2ij} \triangleq T_{ij} \qquad (3)$$

Where $M_{2ij}=\max\limits_{r\in[0,1]}\left|f_{ij}''(r)\right|$.

Hence, the polyline truncation error can be represented by

$$\begin{bmatrix}T_{i1}\\T_{i2}\\\vdots\\T_{in}\end{bmatrix}\left(i=1,2,\cdots,m-1\right).$$

Next, truncation error models for straight line features in three-, two- dimensional spaces are derived based on the generic truncation error model for straight line features in n -dimensional space. These two truncation error model cases may be frequently used in real world geographic information systems.

3 THE TRUNCATION ERROR MODEL FOR STRAIGHT LINE FEATURES IN THREE-DIMENSIONAL SPACE

3.1 Definition of a polyline in three-dimensional space

A polyline in three-dimensional space is composed of vectors: $Q_{31}=\left[x_{11},x_{12},x_{13}\right]^T$, $Q_{32}=\left[x_{21},x_{22},x_{23}\right]^T,\cdots$, $Q_{3i}=\left[x_{i1},x_{i2},x_{i3}\right]^T,\cdots$, $Q_{3m}=\left[x_{m1},x_{m2},x_{m3}\right]^T$, where $i=1,2,\cdots,m$ and m is the number of composition points.

Any point on the line segment $Q_{3i}Q_{3,i+1}(i=1,2,\cdots,m-1)$ is then represented by

$$Q_{3ir}=\begin{bmatrix}x_{i1r}\\x_{i2r}\\x_{i3r}\end{bmatrix}=\left(1-r\right)Q_{3i}+rQ_{3,i+1}$$

$$=\begin{bmatrix}(1-r)x_{i1}+rx_{i+1,1}\\(1-r)x_{i2}+rx_{i+1,2}\\(1-r)x_{i3}+rx_{i+1,3}\end{bmatrix}\triangleq\begin{bmatrix}p_{i1}\\p_{i2}\\p_{i3}\end{bmatrix} \qquad (4)$$

$\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right)$

The true polyline of $Q_{31}Q_{32}\cdots Q_{3i}\cdots Q_{3m}$ can be represented by

$$\phi_{3ir}=\begin{bmatrix}\mu_{i1r}\\\mu_{i2r}\\\mu_{i3r}\end{bmatrix}=\begin{bmatrix}f_{i1}(r)\\f_{i2}(r)\\f_{i3}(r)\end{bmatrix}\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right) \qquad (5)$$

3.2 The truncation error of a polyline in three-dimensional space

Some assumptions about the functions $f_{ij}(r)(i=1,2,\cdots,m;$ $j=1,2,3;r\in[0,1])$ need to be made, in order to enable the estimation of the interpolation error.

It is assumed that $f_{ij}(r) \in C^2[0,1]$, giving

$$\begin{aligned}&\left|R_{ij}(r)\right| = \left|f_{ij}(r) - p_{ij}(r)\right| \le \frac{1}{2} M_{2ij} r(1-r)\\ &\le \frac{1}{8} M_{2ij} \triangleq T_{ij}\end{aligned} \tag{6}$$

where $M_{2ij} = \max\limits_{r\in[0,1]} \left|f_{ij}^{''}(r)\right|$.

Hence, the polyline truncation error can be represented by

$$\begin{bmatrix} T_{i1} \\ T_{i2} \\ T_{i3} \end{bmatrix} (i = 1,2,\cdots,m-1).$$

4 THE TRUNCATION ERROR MODEL FOR STRAIGHT LINE FEATURES IN TWO-DIMENSIONAL SPACE

4.1 Definition of a polyline in two-dimensional space

A polyline in two-dimensional space is composed of vectors: $Q_{21} = [x_{11}, x_{12}]^T$, $Q_{22} = [x_{21}, x_{22}]^T, \cdots$, $Q_{2i} = [x_{i1}, x_{i2}]^T, \cdots$, $Q_{2m} = [x_{m1}, x_{m2}]^T$, where $i = 1,2,\cdots,m$ and m is the number of composition points.

Any point on the line segment $Q_{2i}Q_{2,i+1} (i = 1,2,\cdots,m-1)$ is then represented by

$$\begin{aligned}Q_{2ir} &= \begin{bmatrix} x_{i1r} \\ x_{i2r} \end{bmatrix} = (1-r)\, Q_{2i} + rQ_{2,i+1}\\ &= \begin{bmatrix} (1-r)x_{i1} + rx_{i+1,1} \\ (1-r)x_{i2} + rx_{i+1,2} \end{bmatrix} \triangleq \begin{bmatrix} p_{i1}(r) \\ p_{i2}(r) \end{bmatrix}\\ &\left(i = 1,2,\cdots,m-1,\ \ r \in [0,1]\right)\end{aligned} \tag{7}$$

The true polyline of $Q_{21}Q_{22} \cdots Q_{2i} \cdots Q_{2m}$ can be represented by

$$\phi_{2ir} = \begin{bmatrix} \mu_{i1r} \\ \mu_{i2r} \end{bmatrix} = \begin{bmatrix} f_{i1}(r) \\ f_{i2}(r) \end{bmatrix} (i = 1,2,\cdots,m-1,\ \ r \in [0,1]) \tag{8}$$

4.2 The truncation error of a polyline in two-dimensional space

Some assumptions about the functions $f_{ij}(r) (i = 1,2,\cdots,m;$ $j = 1,2; r \in [0,1])$ need to be made in order to be able to estimate the interpolation error.

It is assumed $f_{ij}(r) \in C^2[0,1]$, giving:

$$\begin{aligned}&\left|R_{ij}(r)\right| = \left|f_{ij}(r) - p_{ij}(r)\right| \le \frac{1}{2} M_{2ij} r(1-r)\\ &\le \frac{1}{8} M_{2ij} \triangleq T_{ij}\end{aligned} \tag{9}$$

Where $M_{2ij} = \max\limits_{r\in[0,1]} \left|f_{ij}^{''}(r)\right|$.

Hence, the polyline truncation error can be represented by

$$\begin{bmatrix} T_{i1} \\ T_{i2} \end{bmatrix} (i = 1,2,\cdots,m-1).$$

5 THE TRUNCATION ERROR MODEL FOR A CURVE LINE IN AN $n-$ DIMENSIONAL SPACE

5.1 Define a curve line in an $n-$ dimensional space

The curve line is formed by the n -dimensional vectors: $Q_{n1} = [x_{11}, x_{12}, \cdots, x_{1n}]^T$, $Q_{n2} = [x_{21}, x_{22}, \cdots, x_{2n}]^T, \cdots$, $Q_{ni} = [x_{i1}, x_{i2}, \cdots, x_{in}]^T, \cdots$, $Q_{nm} = [x_{m1}, x_{m2}, \cdots, x_{mn}]^T$, where $i = 1,2,\cdots,m$ and m is the number of composition points (as shown in Figure 2).

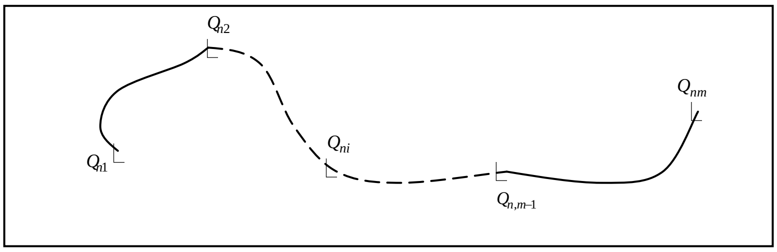


Figure 2. An example of a curve line in n -dimensional space

Any point on the curve line $Q_{ni}Q_{n,i+1} (i = 1,2,\cdots,m-1)$ is then represented by

$$Q_{nir} = \begin{bmatrix} x_{i1r} \\ x_{i2r} \\ \vdots \\ x_{inr} \end{bmatrix} = \begin{bmatrix} a_{i1}r^3+b_{i1}r^2+c_{i1}r+d_{i1} \\ a_{i2}r^3+b_{i2}r^2+c_{i2}r+d_{i2} \\ \vdots \\ a_{in}r^3+b_{in}r^2+c_{in}r+d_{in} \end{bmatrix} \square \begin{bmatrix} p_{i1}(r) \\ p_{i2}(r) \\ \vdots \\ p_{in}(r) \end{bmatrix} \qquad (10)$$

$$\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right)$$

The true curve line of $Q_{n1}Q_{n2}\cdots Q_{ni}\cdots Q_{nm}$ can be represented by

$$\phi_{nir} = \begin{bmatrix} \mu_{i1r} \\ \mu_{i2r} \\ \vdots \\ \mu_{inr} \end{bmatrix} = \begin{bmatrix} f_{i1}(r) \\ f_{i2}(r) \\ \vdots \\ f_{in}(r) \end{bmatrix} \qquad (11)$$

$$\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right)$$

5.2 The truncation error of a curve line in $n-$ dimensional space

For simplicity, it can be assumed that the cubic interpolation is piecewise cubic Hermite interpolation.

Some assumptions about the functions $f_{ij}(r)(i=1,2,\cdots,m;$ $j=1,2,\cdots,n;r\in[0,1])$ need to be made in order to be able to estimate the interpolation error.

It is assumed $f_{ij}(r)\in C^4[0,1]$, giving

$$\begin{aligned} &\left|R_{ij}(r)\right| = \left|f_{ij}(r)-p_{ij}(r)\right| \\ &\le \frac{1}{4!}M_{4ij}r^2\left(1-r\right)^2 \le \frac{1}{384}M_{4ij} \square T_{ij} \end{aligned} \qquad (12)$$

Where $M_{4ij}=\max\limits_{r\in[0,1]}\left|f_{ij}^{(4)}(r)\right|$.

Hence, the truncation error of a curve line can be represented by

$$\begin{bmatrix} T_{i1} \\ T_{i2} \\ \vdots \\ T_{in} \end{bmatrix}(i=1,2,\cdots,m-1).$$

6 THE TRUNCATION ERROR FOR A CURVE LINE IN A THREE-DIMENSIONAL SPACE

6.1 Definition of a curve line in three-dimensional space

The curve line is formed by the three-dimensional vectors: $Q_{31}=[x_{11},x_{12},x_{13}]^T$, $Q_{32}=[x_{21},x_{22},x_{23}]^T,\cdots$, $Q_{3i}=[x_{i1},x_{i2},x_{i3}]^T,\cdots$, $Q_{3m}=[x_{m1},x_{m2},x_{m3}]^T$, where $i=1,2,\cdots,m$ and m is the number of composition points.

Any point on the curve line $Q_{3i}Q_{3,i+1}(i=1,2,\cdots,m-1)$ is then represented by

$$Q_{3ir} = \begin{bmatrix} x_{i1r} \\ x_{i2r} \\ x_{i3r} \end{bmatrix} = \begin{bmatrix} a_{i1}r^3+b_{i1}r^2+c_{i1}r+d_{i1} \\ a_{i2}r^3+b_{i2}r^2+c_{i2}r+d_{i2} \\ a_{i3}r^3+b_{i3}r^2+c_{i3}r+d_{i3} \end{bmatrix} \square \begin{bmatrix} p_{i1}(r) \\ p_{i2}(r) \\ p_{i3}(r) \end{bmatrix} \qquad (13)$$

$$\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right)$$

The true curve line of $Q_{31}Q_{32}\cdots Q_{3i}\cdots Q_{3m}$ can be represented by

$$\phi_{3ir} = \begin{bmatrix} \mu_{i1r} \\ \mu_{i2r} \\ \mu_{i3r} \end{bmatrix} = \begin{bmatrix} f_{i1}(r) \\ f_{i2}(r) \\ f_{i3}(r) \end{bmatrix}\left(i=1,2,\cdots,m-1,\ \ r\in[0,1]\right) \qquad (14)$$

6.2 The truncation error of a curve line in three-dimensional space

For simplicity, it is assumed that the cubic interpolation is piecewise cubic Hermite interpolation.

Some assumptions about the functions $f_{ij}(r)(i=1,2,\cdots,m;$ $j=1,2,3;r\in[0,1])$ need to be made in order to be able to estimate the error of interpolation.

It is assumed that $f_{ij}(r)\in C^4[0,1]$, giving

$$\begin{aligned} &\left|R_{ij}(r)\right| = \left|f_{ij}(r)-p_{ij}(r)\right| \le \frac{1}{4!}M_{4ij}r^2\left(1-r\right)^2 \\ &\le \frac{1}{384}M_{4ij} \square T_{ij} \end{aligned} \qquad (15)$$

Where $M_{4ij}=\max\limits_{r\in[0,1]}\left|f_{ij}^{(4)}(r)\right|$.

Hence, the curve line truncation error can be represented by

$$\begin{bmatrix} T_{i1} \\ T_{i2} \\ T_{i3} \end{bmatrix}(i=1,2,\cdots,m-1).$$

7 THE TRUNCATION ERROR MODEL FOR A CURVE LINE IN A TWO-DIMENSIONAL SPACE

7.1 Definition of a curve line in two-dimensional space

The curve line is formed by two-dimensional vectors: $Q_{21}=[x_{11},x_{12}]^T$, $Q_{22}=[x_{21},x_{22}]^T,\cdots$, $Q_{2i}=[x_{i1},x_{i2}]^T,\cdots$, $Q_{2m}=[x_{m1},x_{m2}]^T$, where $i=1,2,\cdots,m$ and m is the number of composition points.

Any point on the curve line $Q_{2i}Q_{2,i+1}(i=1,2,\cdots,m-1)$ is then represented by

$$Q_{2ir} = \begin{bmatrix} x_{i1r} \\ x_{i2r} \end{bmatrix} = \begin{bmatrix} a_{i1}r^3+b_{i1}r^2+c_{i1}r+d_{i1} \\ a_{i2}r^3+b_{i2}r^2+c_{i2}r+d_{i2} \end{bmatrix} \square \begin{bmatrix} p_{i1}(r) \\ p_{i2}(r) \end{bmatrix} \qquad (16)$$

$$\left(i=1,2,\cdots,m-1,\quad r\in[0,1]\right)$$

The true curve line of $Q_{21}Q_{22}\cdots Q_{2i}\cdots Q_{2m}$ can be represented by

$$\phi_{2ir} = \begin{bmatrix} \mu_{i1r} \\ \mu_{i2r} \end{bmatrix} = \begin{bmatrix} f_{i1}(r) \\ f_{i2}(r) \end{bmatrix} \left(i=1,2,\cdots,m-1,\quad r\in[0,1]\right) \qquad (17)$$

7.2 The truncation error of a curve line in two-dimensional space

For simplicity, it is assumed that the cubic interpolation is piecewise cubic Hermite interpolation.

Some assumptions about the functions $f_{ij}(r)(i=1,2,\cdots,m;\ j=1,2; r\in[0,1])$ need to be made in order to be able to estimate the error of interpolation.

It is assumed that $f_{ij}(r)\in C^4[0,1]$, giving

$$\left|R_{ij}(r)\right| = \left|f_{ij}(r)-p_{ij}(r)\right| \le \frac{1}{4!} M_{4ij} r^2 (1-r)^2$$
$$\le \frac{1}{384} M_{4ij} \square T_{ij} \qquad (18)$$

Where $M_{4ij} = \max\limits_{r\in[0,1]} \left|f_{ij}^{(4)}(r)\right|$.

Hence, the curve line truncation error can be represented by

$$\begin{bmatrix} T_{i1} \\ T_{i2} \end{bmatrix} (i=1,2,\cdots,m-1).$$

8 CONCLUSIONS AND DISCUSSION

To provide a full picture about the overall error of a line, we need to quantify (a) the error of the model that is used to interpolate the line, and (b) the propagated error from the error of the component nodes of the line. This paper presented a research on estimation the error of the interpolation model of the lines, which is a step further to the early studies on propagated error of lines.

A method for estimate the model error, truncation error, of the lines have been developed in this study based on the approximation theory. A line feature in GIS can be either interpolated by linear or nonlinear functions, our research find out that of the model error of a linear interpolated line is larger than that from a nonlinear interpolation method, such as a cubic interpolation method. Such as analysis result can be used as a reference for the selection interpolation methods for lines..

By integration of the analysis result of model error in this study, and the propagated error of the lines in the early study, we can have better estimation on the overall error of lines.

A further extension of this study will be to investigate truncation error of other interpolation methods, such as hybrid interpolation method.

ACKNOWLEDGMENT

The work presented in this paper is supported by, Hong Kong Research Grants Council (Project No.: PolyU 5276/08E), The Hong Kong Polytechnic University (Project No.: G-YX0P G-YG66 G-YF24 1-ZV4F G-U 753), and the National Natural Science Foundation, China (Project No. 40629001).

COMPARISON AND UNCERTAINTY ANALYSIS IN REMOTE SENSING BASED PRODUCTION EFFICIENCY MODELS

Rui Liu [a], Jiu-lin Sun [a, *], Juan-le Wang [a], Min Liu [b], Xiao-lei Li [c], Fei Yang [a]

[a] State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, No. 11A, Datun Road, Chaoyang District, Beijing 100101, PR China - (liur, sunjl, wangjl, yangfei)@lreis.ac.cn

[b] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, No. 11A, Datun Road, Chaoyang District, Beijing 100101, PR China - liusimin122@163.com

[c] School of Civil Engineering and Architecture, Chongqing University of Science and Technology, Chongqing 400042, PR China - li.xiaolei8@gmail.com

KEY WORDS: NPP, Model, PEM, Remote Sensing, Accuracy, Uncertainty, Comparsion

ABSTRACT:

The remote sensing based Production Efficiency Models (PEMs), springs from the concept of "Light Use Efficiency" and has been applied more and more in estimating terrestrial Net Primary Productivity (NPP) regionally and globally. However, global NPP estimats vary greatly among different models in different data sources and handling methods. Because direct observation or measurement of NPP is unavailable at global scale, the precision and reliability of the models cannot be guaranteed. Though, there are ways to improve the accuracy of the models from input parameters. In this study, five remote sensing based PEMs have been compared: CASA, GLO-PEM, TURC, SDBM and VPM. We divided input parameters into three categories, and analyzed the uncertainty of (1) vegetation distribution, (2) fraction of photosynthetically active radiation absorbed by the canopy (*fPAR*), (3) light use efficiency (ε), and (4) spatial interpolation of meteorology measurements. Ground measurements of Hulunbeier typical grassland and meteorology measurements were introduced for accuracy evaluation. Results show that a real-time, more accurate vegetation distribution could significantly affect the accuracy of the models, since it's applied directly or indirectly in all models and affects other parameters simultaneously. Higher spatial and spectral resolution remote sensing data may reduce uncertainty of *fPAR* up to 51.3%, which is essential to improve model accuracy. We also figured out a vegetation distribution based on Maximum value of light use efficiency (ε^*) and ANUSPLIN method for spatial interpolation of meteorology measurement is also an effective way to improve the accuracy of remote sensing based PEMs.

1. INTRODUCTION

Terrestrial net primary productivity (NPP), defined as the rate of atmosphere carbon uptake by vegetation through the process of net photosynthesis minus dark respiration(Ruimy *et al.* 1994), is the central-related variable summarizing the interface between plant and other processes(Field *et al.* 1995). NPP is sensitive to the environmental factors and is highly various in space and time. Thus, estimating NPP more precisely is a key to understanding the terrestrial carbon cycle.

There are two methods available to estimate terrestrial NPP: (1) extrapolating field measurement for local NPP to the biosphere through a vegetation map; (2) modeling plant productivity at the biosphere level(Ruimy *et al.* 1994). Since direct observation or measurement of NPP is unavailable on a global scale, the modeling method has been widely accepted. There are three main types of productivity model: (1) statistical model: estimating NPP by meteorology measurement and experimental parameters, regardless of physiological and ecological characteristics of vegetation, such as: Miami(Lieth 1972), Thornthwaite(Lieth *et al.* 1972), Chikugo(Zenbei UCHIJIMA *et al.* 1985) and Zhou Guang-sheng (Zhou *et al.* 1995); (2) process model: based on plant physiological ecology principles, estimating NPP by simulating process of photosynthesis. This model has been widely used in local areas, such as: CENTURY(Parton *et al.* 1993), CARAIB (Warnant *et al.* 1994, Nemry *et al.* 1996), KGBM (Kergoat 1998), SILVAN (Kaduk *et al.* 1996), CEVSA(Cao *et al.* 1998), TEM(McGuire *et al.* 1995) and BIOME-BGC (Running *et al.* 1993). (3) production efficiency models (PEMs). The "Light Use Efficiency (ε)" concept (Monteith 1972) has been adopted to decompose into independent parameters such as incoming solar radiation, radiation absorption, and conversion efficiency. The main PEMs include CASA(Potter *et al.* 1993, Field *et al.* 1995), GLO-PEM(Prince 1991, Prince *et al.* 1995), SDBM(Knorr *et al.* 1995), VPM(Xiao *et al.* 2004a, Xiao *et al.* 2004b) and TURC(Ruimy *et al.* 1996).

Along with the increasing availability of remote sensing measurement, (1) most parameters can be obtained by remote sensing data, and (2)with easy access to regional data, reducing errors caused by interpolation is possible, thus the remote sensing based PEMs has been applied more and more to estimatting terrestrial NPP.

However, it is very difficult to evaluate the accuracy of the models for two reasons: (1) acquisition of direct observation is

[*] Corresponding author. *Tel.*: +86-10-64889266; +86-10-64889045; *Fax.*: +86-10-64889062
E-mail addresses: sunjl@lreis.ac.cn (Jiu-lin Sun) liur@lreis.ac.cn (Rui Liu)

unavailable on a regional or global scale, and (2) different models come from different data sources and handling methods, impossible to determine which one is more accurate.

Under the scientific sponsorship of the IGBP, such a model intercomparison has been carried out at the Potsdam Institute for Climate Impact Research(Cramer *et al.* 1999, Ruimy *et al.* 1999). Result shows that global NPP estimates (Fig.1) vary greatly between different models. However, since no field data are available to validate the models, we have no way to determine which result is closer to "true value".

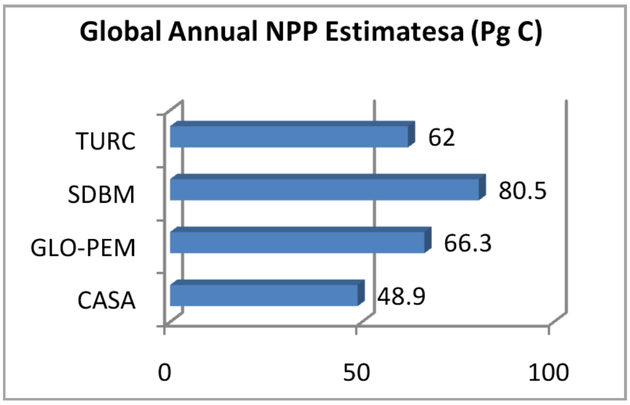


Figure 1. Comparison of global NPP estimations of PEMs

In this study, we focus on the methods of improving the accuracy of remote sensing based on PEMs rather than determining the better one. The input parameters of PEMs are divided into three categories according to acquisition source and the analysis of of each category's influence is performed. We compare the 5 models by taking different data sources as input for each parameter and combining with ground measurement in Hulunbeier and Tibet in order to determine the uncertainty of parameter. The purpose of this study is, first, to identify the most influential input parameters in NPP estimation among the 5 models, and second, to seek access to improvement in model accuracy.

2. PARAMETERS ANALYSIS IN PEMs

PEMs develops from the concept of light use efficiency: NPP has a strong linear relationship in ideal environment with light use efficiency (ε) and absorbed photosynthetically active radiation (*APAR*): NPP=$\varepsilon \cdot APAR$.

APAR is calculated from global solar radiation and fraction of photosynthetically active radiation absorbed by the canopy (*fPAR*) which can be obtained from remote sensing data, and ε is regarded as a conversion scale of *APAR* to NPP, as a result of the interaction of environmental constraints and based on "Maximum value of light use efficiency (ε^*)". These models include various parameters (Tab.1) and focus on different levels of mechanism with inhibition process taken into account. The parameters can be divided into three categories (Tab.2) according to acquisition source.

2.1 Remote Sensing Data

PEMs uses remote sensing data to acquire the land surface condition, especially vegetation type and *fPAR*. In the simulation process, remote sensing data mainly provide the following three types of information:

Model	Influenced by:
CASA	NPP=$f\,(\varepsilon^*, R_S, fPAR, T, EET, PET\,)$
GLO-PEM	NPP=$f\,(\varepsilon^*, R_S, fPAR, T, VPD, SW, R_A\,)$
TURC	NPP=$f\,(\varepsilon^*, R_S, fPAR, R_A, T\,)$
SDBM	NPP=$f\,(\varepsilon^*, R_S, fPAR, CO_2\,)$
VPM	NPP=$f\,(\varepsilon^*, R_S, fPAR, T, W, P_L\,)$

R_S: Solar radiation R_A: Plant autotrophic respiration
PET: Potential evapotranspiration *EET*: Estimated evapotranspiration
P_L: Leaf phenology *T*: Temperature *W*: Water capacity

Table 1. Parameters of PEMs

Model	Vegetation Distribution	Satellite *fPAR*	Meteorological measurements	Plant Physioecology	Other satellite data
CASA	×	×	R_S,*T*,*EET*,*PET*	ε^*	
GLO-PEM		×		ε^*,R_A	R_S,*T*,*VPD*,*SW*
TURC	×	×	R_S,*T*	ε^*,R_A	
SDBM		×	R_S,CO_2	ε^*	
VPM	×	×	R_S,*T*	ε^*,P_L	*W*

Table 2. Parameters classification of PEMs

1. Vegetation Distribution Information: According to different spectral characteristics and temporal variation, the accurate and real-time vegetation distribution on the earth can be obtained through appropriate classification algorithm.
2. Vegetation Index: Spectral reflectance of vegetation is influenced by vegetation type, species composition, vegetation cover, chlorophyll content, plant water and so on. Vegetation index is a comprehensive performance of the spectral reflectance, which bears a strong relationship with NPP estimates.
3. Vegetation growth environment information: The environmental factors can be obtained by means of remote sensing in recent years, such as temperature, precipitation, soil moisture and other relevant information, though further study is to be made on the applicability and accuracy .

GLO-PEM is unique among the 5 models because all variables about climate and vegetation distribution are derived from remote sensing data.

2.2 Meteorology Measurement

The process of plant growth appears responsive to environmental conditions. Therefore, the formation of vegetation NPP depends on the regional light, heat, water conditions and so on, as well as the biome production capacity(Zhou *et al.* 1995). Climate factors' control over vegetation productivity is not only present in the vegetation diversity, but also in photosynthesis inhibition. The meteorology measurements in the models such as radiation, temperature, precipitation are all obtained from meteorology stations except GLO-PEM.

2.3 Plant Physiological Data

The plant physioecology mainly concerns how and to what extent the plant growth responds to the environmental factors such as: increasing CO_2 concentration, ultraviolet radiation enhancement, temperature change, sunlight irradiation and the enlargement of salty habitats. All of these factors are closely associated with the process of global climate change.

ε is a fundamental element in PEMs. It was used for the conversion of *APAR* to biomass, and affected by many environmental factors. Each model illustrates its own approach of simulating the process in which environmental factors influence mode.

3. UNCERTAINTY OF INPUT PARAMETERS

Obviously enough, parameters are highly similar in these models. The main differences lie in (1) the way of obtaining and applying vegetation distribution, (2) the way of obtaining *fPAR* and ε and (3) the use of meteorology factors. Thus, it is important for improving model accuracy to analyze uncertainty of each parameter.

3.1 Vegetation Distribution

Vegetation distribution is considered to be the most important determinant of carbon storage, uptake and release from the terrestrial biosphere, and it affects model accuracy mainly in two ways:

3.1.1 Applying Vegetation Distribution: All remote sensing based PEMs assumes the world is covered by vegetation. CASA and VPM uses an actual vegetation distribution from remote sensing including human land use. SDBM and GLO-PEM does not use a vegetation map directly, but the parameters via remote sensing such as temperature , vapour pressure deficit and *APAR* also explain the vegetation cover change. Only TURC uses potential vegetation regardless the human land use. A comparison between actual vegetation data set and potential vegetation data set shows that the human land use and agriculture affect up to 40% of NPP estimate in temperate mixed forests and deciduous forest on a global scale(Ruimy *et al.* 1999). Regionally the simulated NPP with land use constraint in the south portion of NSTEC was about 65% of that without land use constraint(Gao *et al.* 2003). On the other hand, a better classification accuracy has testified its improvement for NPP estimate(Zhu *et al.* 2006).

3.1.2 Determination of other parameters: Vegetation distribution is applied directly or as an intermediate variable to determine the precision of other parameters such as ε^*, R_A, P_L and *EET*. In most models, these parameters are assumed constant or determined by the vegetation maps. However, these plant physiological-related parameters apparently depend on the vegetation type with the classification accuracy taken into account.

It is impossible for us to know exactly how much vegetation distribution is affected, but we definitely know that a real-time, more accurate vegetation distribution can significantly affect the accuracy of the models.

3.2 Remote Sensing Based *fPAR*

fPAR, a significant parameter for calculating *APAR*, truly reflects the status of vegetation canopy's absorption of photosynthetically active radiation, and has a direct impact on the uncertainty of PEMs. Remote sensing provides a means to estimating *fPAR* globally. Most models (CASA, GLO-PEM, SDBM and TURC) utilizes Normalized Difference Vegetation Index (*NDVI*) to obtain *fPAR* (apply different algorithms), while VPM uses Enhanced Vegetation Index (*EVI*) (Tab.3).

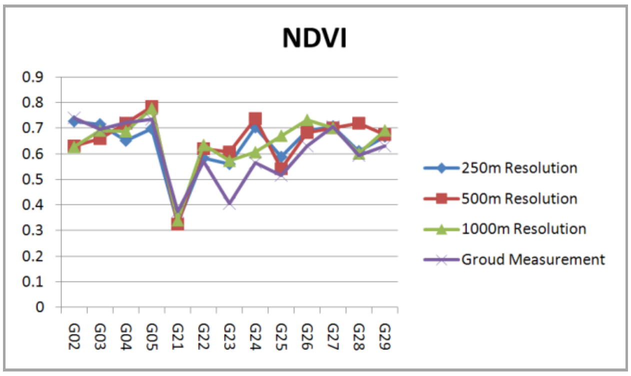
CASA	$fPAR$=min{$(SR-SR_{\min})/(SR_{\max}-SR_{\min})$, 0.95}
GLO-PEM	$fPAR=(SR-SR_{\min})(fPAR_{\max}-fPAR_{\min})/(SR_{\max}-SR_{\min})$
TURC	$fPAR= -0.1914+2.186\cdot NDVI$
SDBM	$fPAR= -0.025+1.25\cdot NDVI$
VPM	$fPAR=1\cdot EVI$

$SR=(1+NDVI)/(1-NDVI)$

Table 3. *fPAR* estimation in PEMs

However, some researches indicate that there are limitations in application of *NDVI*, such as (1) tending to be saturated in well-vegetation cover area(Wang *et al.* 2003), and (2)which is sensitive to the soil structure in the low vegetation cover area(Huete *et al.* 1994). One possible solution is using *EVI* which introduced the blue-ray band aiming to reduce atmosphere effect(Huete *et al.* 1994, Huete *et al.* 1997). Although *EVI* was used in VPM for estimating forest NPP, and considered superior in grassland NPP estimation over *NDVI*(Kawamura *et al.* 2005), the further application in different environment on a global scale is still necessary.

It remains unsolved which *fPAR* is more accurate in these models since we cannot have *fPAR* from ground measurement. However, we can improve the precision by using higher spatial resolution *NDVI* and *EVI*. All the models calculate the *fPAR* with a 8*km* resolution *NDVI* derived from NOAA/AVHRR. For the ease of comparison, we use MODIS *NDVI* and *EVI* data (obtained at 2009-07-28 from USGS) with 1*km*, 500*m* and 250*m* spatial resolution, and ground measured spatial data (obtained at 2009-08-02 by FieldSpec® HandHeld Spectroradiometer) from Hulunbeier grassland (Fig.2). The calculation is strictly in accordance with the MODIS algorithm. It should be known that we are not saying the ground measured data is "true", but it is of great reference value. The relative error (Fig.3) shows that the better spatial resolution brings the higher precision. In some extra point, relative error from the 1*km* resolution can reach up to 51.3%.



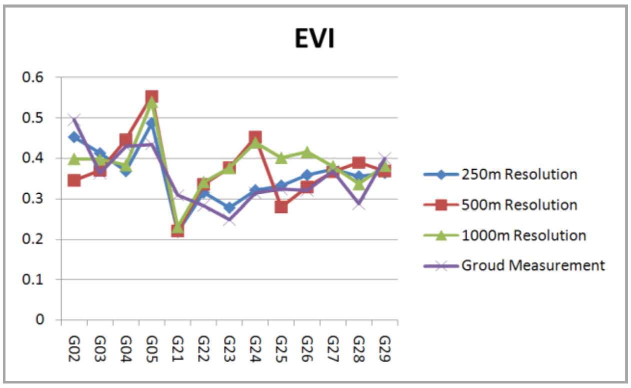


Figure 2. Comparison of *NDVI* and *EVI*

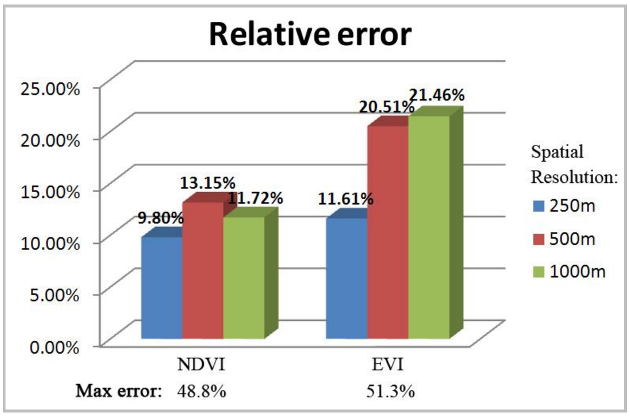


Figure 3. Relative error of *NDVI* and *EVI*

3.3 Estimating Maximum Value of Light Use Efficiency

The light use efficiency (ε) determines the capability that the plants capture and transform environmental resources to dry matter production, and fluctuate with the environmental index such as temperature, moisture, soil, nutrition, plant ontogeny, etc. (Prince 1991). In remote sensing based PEMs, these affections present as constraints of Maximum Value of Light Use Efficiency (ε^*) ranging between 0 and 1. Therefore, $\varepsilon^*$is influential for NPP estimates.

In early researches, ε^* is empirically derived as a conservative quantity(Monteith 1972). CASA takes it as 0.389$gC{\cdot}MJ^{-1}$. However, several researches indicate that ε^* varies due to different vegetation types. And in view of its importance, there is controversy about the value range. Ruimy believes it ranges between 0.108 and 1.580 $gC{\cdot}MJ^{-1}$ and GLO-PEM adopts it between 0.2 and 1.2 $gC{\cdot}MJ^{-1}$. In Guangdong Province of China, the result shows that ε^* range between 0.69 and 1.05 $gC{\cdot}MJ^{-1}$ (Peng *et al.* 2000). Since ε cannot be measured directly, studies about determination of ε generally fall into two types: (1) simulating the plant growth process with the principle of plant ecology, and (2) remote sensing retrieval through PEMs and ground measured NPP. The latter one seems more feasible as plant ecology simulation can hardly be extended to a global scale. By means of remote sensing retrieval, there are methods for improving the precision of PEMs:

1. ε^* should not be considered as a conservative quantity, for it shows difference between different biome.
2. A more accurate and real-time vegetation map could be used which shows the biome distribution.
3. More accurate remote sensing data are uesd for retrieval and plant ecology data have been counted for a more accurate ε^*.

3.4 Spatial Interpolation of Meteorology Measurements

Based on different models, climate factors affect the NPP estimation working as the inhibition of ε^* (Tab.4). However, they are hardly obtained directly through remote sensing data with high accuracy.

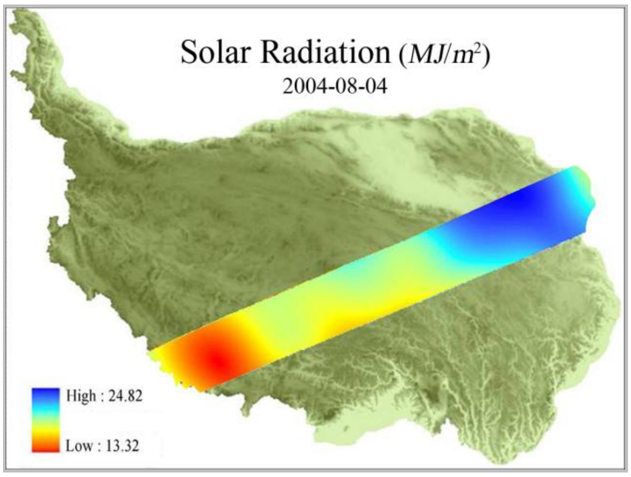
In most cases, the regional and global meteorology distribution are based on station measurement and spatialized by means of interpolation. Therefore, interpolation precision is important in improving accuracy of NPP estimates(Price *et al.* 2000, Wong *et al.* 2003). The precision of meteorology interpolation is mainly influenced by: (1) site's latitude and longitude, (2) site's elevation and (3) regional terrain.

Model	Influenced by
CASA	R_S, T, EET, PET
GLO-PEM [a]	R_S, T, VPD, SW
TURC	R_S, T, R_A
SDBM	R_S, CO_2
VPM	R_S, T, W

[a] All factors in GLO-PEM are obtained from remote sensing data

Table 4. Climate factors in PEMs

A comparisive study of several interpolation method shows that the multiple regression equation had a better performance by introducing elevation and location(Collins 1995). Lin's research demonstrated that Gradient Plus Inverse-distance-squared (GIDS) method is better than others in reflecting temperature change with elevation(Lin *et al.* 2002), but researchers also proved that ANUSPLIN method is better than GIDS(Price *et al.* 2000, Feng 2004). The 88 meteorology station data in Tibet was spatialized using ANUSPLIN method (Liu 2008) combining with $1km{\times}1km$ DEM. The result (Fig.4) proved to be more accurate than other methods.



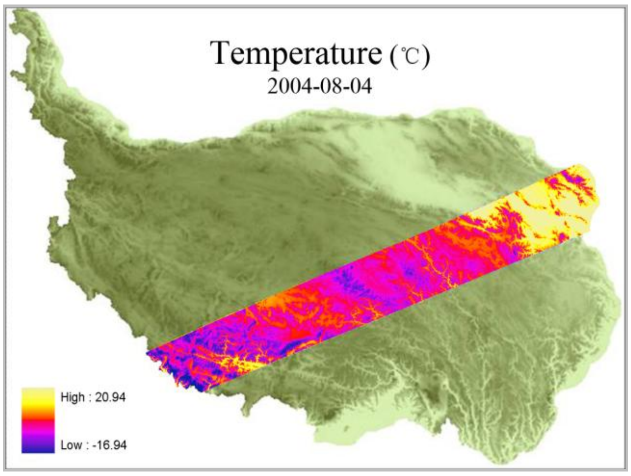


Figure 4. Meteorology interpolation in Tibetan transact

4. CONCLUSION AND DISCUSSION

Remote sensing based PEMs takes good use of the "light use efficiency" theory and adopts several approaches to estimate NPP. Each approach has a theoretical basis for its own. It is not possible to tell which model is "better" since none of them is perfect and cannot be verified on a global scale. NPP estimate varies greatly among models in a comparative research and we have no way to know which result is closer to the "true value". However, there are ways to improve the model accuracy.

Vegetation distribution is the fundamental element among all parameters and has been used directly or indirectly in all models. Apparently, actual vegetation distribution performs much better than potential vegetation distribution while human land use has a great impact on the NPP estimation. Meanwhile, vegetation distribution determines the accuracy of the application of other parameters to a large extent. The uncertainty of vegetation distribution is caused by: (1) inconsistent and ambiguous vegetation types, (2) time inconformity from classification time to current time, (3) mixed pixel of different vegetation types and (4) inconsistent scaling. The developing remote sensing data and techniques provide the possibility to these uncertainties. Overall, a real-time and accurate vegetation map is helpful in greatly improving the accuracy of the PEMs.

The vegetation index is close to photosynthesis by means of determining the *fPAR*. The NOAA/AVHRR *NDVI* is the most common data for NPP estimate. But the 8*km* or lower spatial resolution caused large errors because of mixed pixel. Advanced sensors with better spectral and spatial resolution can provide more accurate *fPAR*. Our experiment in Hulunbeier shows that the better resolution brings higher precision, especially in mixed pixels which have 51.3% relative error.

New vegetation index was introduced for NPP estimate. Although *EVI* proves better to perform vegetation status than *NDVI*, there are still questions about the interrelationship between *fPAR* and *EVI*. It is regarded as a potential solution and needs more research work.

ε^* varies greatly among literatures for there is no convincing method for ground measuring or evaluation. Since ε^* depends on vegetation types, we can enhance the model accuracy by (1) using more accurate vegetation map and (2) combining remote sensing retrieval with plant ecology. Nowadays, ε^* can be obtained precisely by measuring fluxes of CO_2 over whole canopies, and analyzing the relationship between CO_2 exchange and photon flux density. It is generally possible to extract ε^* more representatively though the uncertainty remains to be developed.

Being as constraints of ε^*, meteorology measurement and ε^* determine the plant conversion efficiency together. Spatial interpolation method determines the accuracy of spatialized meteorology data. The chosen method should involve all relating factors, including location, elevation and terrain. Our research shows that the ANUSPLIN method can improve accuracy.

Here, we developed methods to improve parameter accuracy, though the overall accuracy improvement for the model still remains unquantified. Meanwhile, although we are not sure about the interrelation response mechanism between each parameter, it is possible to estimate NPP at a higher precision through applying more accurate parameters.

ACKNOWLEDGEMENTS

We greatly appreciate the support of Nature Science Founding of China (40771146, 40801180) and Open Research Funding Program of LGISEM.

REFERENCE

Cao, M.K. and Woodward, F.I., 1998, Dynamic responses of terrestrial ecosystem carbon cycling to global climate change. *Nature*, 393, pp. 249-252.

Collins, F.C., and P.V. Bolstad. 1996. A comparison of spatial interpolation techniques in temperature estimation. *In: Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling*, Santa Fe, New Mexico, January 21-25, 1996. Santa Barbara, California: National Center for Geographic Information Analysis (NCGIA).

Cramer, W., Kicklighter, D.W., Bondeau, A., Moore, B., Churkina, C., Nemry, B., Ruimy, A., Schloss, A.L. and Participants Potsdam, N.P.P.M.I., 1999, Comparing global models of terrestrial net primary productivity (NPP): overview and key results. *Global Change Biology*, 5, pp. 1-15.

Feng, X.F., 2004, Simulating Net Primary Productivity and Evaportranspiration of Terrestrial Ecosystems in China using a Process Model Driven by Remote Sensing. *Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences.*

Field, C.B., Randerson, J.T. and Malmstrom, C.M., 1995, Global Net Primary Production - Combining Ecology and Remote-Sensing. *Remote Sensing of Environment*, 51, pp. 74-88.

Gao, Q., Li, X.B. and Yang, X.S., 2003, Response of Vegetation and Primary Production in Northern-South Transect of Eastern China to Global Change Under Land Use Constraint. *Acta Botanica Sinica*, 45, pp. 1274-1284.

Huete, A., Justice, C. and Liu, H., 1994, Development of vegetation and soil indices for MODIS-EOS. *Remote Sensing of Environment*, 49, pp. 224-234.

Huete, A., Liu, H., Batchily, K. and Van Leeuwen, W., 1997, A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment*, 59, pp. 440-451.

Hutchinson, M., 2003, ANUSPLIN version 4.2 user guide. Centre for Resource and Environmental Studies, *Australian National University, Canberra, Australia.*

Kaduk, J. and Heimann, M., 1996, A prognostic phenology scheme for global terrestrial carbon cycle models. *Climate Research*, 6, pp. 1-19.

Kawamura, K., Akiyama, T., Yokota, H., Tsutsumi, M., Yasuda, T., Watanabe, O., Wang, G. and Wang, S., 2005, Monitoring of forage conditions with MODIS imagery in the Xilingol steppe, Inner Mongolia. *International Journal of Remote Sensing*, 26, pp. 1423-1436.

Kergoat, L., 1998, A model for hydrological equilibrium of leaf area index on a global scale. *Journal of Hydrology*, 213, pp. 268-286.

Knorr, W. and Heimann, M., 1995, Impact of Drought Stress and Other Factors on Seasonal Land Biosphere CO2 Exchange Studied Through An Atmospheric Tracer Transport Model. *Tellus Series B-Chemical and Physical Meteorology*, 47, pp. 471-489.

Lieth, H., 1972, Modeling the primary productivity of the world. *Nature and Resources*, 8, pp. 5-10.

Lieth, H. and Box, E., 1972, Evapotranspiration and primary productivity; CW Thornthwaite Memorial Model. *publications in climatalogy*, 25, pp. 37-46.

Lin, Z.H., Mo, X.G., Li, H.X. and Li, H.B., 2002, Comparison of Three Spatial Interpolation Methods for Climate Variables in China. *ACTA GEOGRAPHICA SINICA*, 57, pp. 47-56.

Liu, M., 2008, Study on Estimation and Uncertainty of Terrestrial Ecosystem Productivity Based on RS and GIS. *Nanjing Normal University*.

McGuire, A.D., Melillo, J.M., Kicklighter, D.W. and Joyce, L.A., 1995, Equilibrium responses of soil carbon to climate change: Empirical and process-based estimates. Journal of *Biogeography*, 22, pp. 785-796.

Monteith, J.L., 1972, Solar-Radiation and Productivity in Tropical Ecosystems. *Journal of Applied Ecology*, 9, pp. 747-766.

Parton, W.J., Scurlock, J.M.O., Ojima, D.S., Gilmanov, T.G., Scholes, R.J., Schimel, D.S., Kirchner, T., Menaut, J.C., Seastedt, T., Moya, E.G., Kamnalrut, A. and Kinyamario, J.I., 1993, Observations and Modeling of Biomass and Soil Organic-Matter Dynamics for The Grassland Biome Worldwide. Global *Biogeochemical Cycles*, 7, pp. 785-809.

Peng, S.L., Guo, Z.H. and wang, B.S., 2000, Use of GIS and RS to Estimate the Light Utilization Efficiency of the Vegetation in Guangdong, China. *Acta Ecological Sinica*, 20, pp. 903-909.

Potter, C.S., Randerson, J.T., Field, C.B., Matson, P.A., Vitousek, P.M., Mooney, H.A. and Klooster, S.A., 1993, Terrestrial Ecosystem Production - A Process Model-Based on Global Satellite and Surface Data. *Global Biogeochemical Cycles*, 7, pp. 811-841.

Price, D., McKenney, D., Nalder, I., Hutchinson, M. and Kesteven, J., 2000, A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agricultural and Forest Meteorology*, 101, pp. 81-94.

Prince, S., 1991, A model of regional primary production for use with coarse resolution satellite data. *International Journal of Remote Sensing*, 12, pp. 1313-1330.

Ruimy, A., Dedieu, G. and Saugier, B., 1996, TURC: A diagnostic model of continental gross primary productivity and net primary productivity. *Global Biogeochemical Cycles*, 10, pp. 269-285.

Ruimy, A., Kergoat, L., Bondeau, A. and Intercomparison, P.P.N.M., 1999, Comparing global models of terrestrial net primary productivity (NPP): analysis of differences in light absorption and light-use efficiency. *Global Change Biology*, 5, pp. 56-64.

Ruimy, A., Saugier, B. and Dedieu, G., 1994, Methodology for the estimation of terrestrial net primary production from remotely sensed data. *Journal of Geophysical Research-Atmospheres*, 99, pp. 5263-5283.

Running, S. and Hunt, E., 1993, Generalization of a forest ecosystem process model for other biomes, BIOME-BGC, and an application for global-scale models. *Scaling physiological processes: leaf to globe*, pp. 141-158.

Wang, Z.X., Liu, C. and Alfredo, H., 2003, From AVHRR-NDVI to MODIS-EVI: Advances in Vegetation Index Research. *Acta Ecological Sinica*, 23, pp. 979-987.

Warnant, P., Francois, L., Strivay, D. and Gerard, J.C., 1994, Caraib: A Global-Model of Terrestrial Biological Productivity. *Global Biogeochemical Cycles*, 8, pp. 255-270.

Xiao, X., Hollinger, D., Aber, J., Goltz, M., Davidson, E., Zhang, Q. and Moore, B., 2004a, Satellite-based modeling of gross primary production in an evergreen needleleaf forest. *Remote Sensing of Environment*, 89, pp. 519-534.

Xiao, X., Zhang, Q., Braswell, B., Urbanski, S., Boles, S., Wofsy, S., Moore, B. and Ojima, D., 2004b, Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote Sensing of Environment*, 91, pp. 256-270.

Zenbei UCHIJIMA and SEINO, H., 1985, Agroclimatic Evaluation of Net Primary Productivity of Natural Vegetation (1) Chikugo Model for Evaluating Net Primary productivity. *J. Agr.Met*, 40, pp. 343-352.

Zhou, G.S. and Zhang, X.S., 1995, A Natural Vegetation NPP Model. Acta Phytoecologica Sinica, pp. 193-200.

Zhu, W.Q., Pan, Y.Z., He, H., Yu, D.Y. and Hu, H.B., 2006, Maximum Value of Light Use Efficiency Simulation of China's Typical Vegetation. *CHINESE SCIENCE BULLETIN*, 51, pp. 700-706.

SAMPLING SURVEY OF HEAVY METAL IN SOIL USING SSSI

A. H. MA[a], J. F. Wang[b]*, K. L. Zhang[a]*

[a] School of Geography and Remote Sensing Sciences, Beijing Normal University, Beijing 100875,
China, maaihua@mail.bnu.edu.cn, keli@bnu.edu.cn
[b] Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences,
Beijing 100101, China, wangjf@lreis.ac.cn

KEY WORDS: Soil sampling, Spatial dependency, Stratified Sampling, Sampling accuracy, variance, SSSI

ABSTRACT:

Much attention has been given to sampling design, and the sampling method chosen directly affects the sampling accuracy. The development of spatial sampling theory has lead to the recognition of the importance of taking spatial dependency into account when sampling. This text uses the new Sandwich Spatial Sampling and Inference (SSSI) software as a tool to compare the relative error, coefficient of variation (CV), and design effect with five sampling models – simple random sampling, stratified sampling, spatial random sampling, spatial stratified sampling, and sandwich spatial sampling. The five models are simulated 1000 times each with a range of sample sizes from 10 to 80. SSSI includes six models in all, but systematic sampling is not used here because the sample positions are fixed. The dataset consists of 84 points measuring soil heavy metal content in Shanxi Province, China. The whole area is stratified into four layers by soil type、hierarchical cluster and geochronology, and three layers by geological surface.The research shows that the accuracy of spatial simple random sampling and spatial stratified sampling is better than simple random sampling and stratified sampling because the soil content is spatially continuous, and stratified models are more efficient than non-stratified models. Stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. Based on spatial stratified sampling, sandwich sampling develops a report layer composed of the user's final report units, allowing the user to obtain the mean and variance of each report unit with high accuracy. In the case of soil sampling, SSSI was a useful tool for evaluating the accuracy of different sampling techniques.

1. INTRODUCTION

Much progress has been made in developing tools to assess estimation accuracy and the relevancy of the chosen model, such as spatial autocorrelation precision (Haining R, 2003; Cressie N,1991; Wang J. *et al.*, 2002; Li L. et al. 2005) or error variance (Jinfeng Wang *et al.*, 2009). Uncertainty in a spatial sampling survey can be reduced with high-quality information, improvement of statistical estimation accuracy, and increasingly precise geo-statistical software. However, the choice of a relevant sampling model is essential to the quality of the analysis and what will be learned in the study.

Many different sampling methods exist, and selecting one relevant to a specific problem is always an important choice for the scientist, who wants to have the most valid statistical inference analysis possible. The user can decide to study the problem with, for instance, a simple random sampling model, which involves selecting a fixed number of units from a population with equal chances for each, or with a stratified sampling model to assess the relation between samples and a stratified information layer (William G. Cochran. 1977). This choice really depends on the problem, on the objective of the study, or on the available dataset. Having some methods to assess the fit of the sampling model with the sample dataset would be useful for the scientist, in order to improve the statistical analysis and refine results.

Two approaches for matching sampling effort to accuracy are chosen: a classical approach, which ignores spatial dependence between observations; and a geo-statistical approach, which accounts for spatial dependence (R. D. Hedger et al, 2001). Accounting for spatial dependency decreases the standard error when the environmental variable is spatially correlated (R. B. Haining, 1988; J. F. Wang, 2002).

Sandwich Spatial Sampling and Inference (SSSI) is a professional spatial sampling and statistical inference tool. It can be used for sampling design and statistical inference in environmental, resource, land, ecological, social, and economic sciences and practices (http://www.sssampling.org/). The objective of this study is to evaluate the accuracy of different sampling methods by comparing several indicators, such as relative error, coefficient of variation (CV), and design effect, using SSSI. Because the sampling model and number of samples (or sample size) are the two most important factors determining the sampling efficiency (D.M. Chen, 2009), this paper will compare four models with a range of sample sizes from 10 to 80. The dataset is a survey of heavy metal content in soil in Shanxi province, China.

2. MATERIAL AND METHODS

2.1 Data and site description

The investigation and environmental sampling was carried out in 2002 and 2003. Two counties, Zhongyang and Jiaokou, in

*corresponding authors

Shanxi Province were selected as the research area. There are seven administrative villages in each county and each has a high incidence of birth defects. Samples were collected using administration villages as units. Sample sites were considered consistent with birth defect epidemiological survey sites, so there exist some gaps within the data set, such as in the western region. Soil samples were collected far away from the soil disturbance plot, which is affected by residences and road, to ensure the representativeness of local environmental characteristics as much as possible. Samples were collected in most villages, with 84 points in all. Well-mixed cultivated land surface soil samples were acquired with a shovel 10 to 20 cm deep. The samples were dried, cleaned of organic debris and impurities, ground with a mortar, sifted with a 100 mesh aperture, and weighed at about 0.1 g soil 3 copies. They were then mixed with 3 ml of Glass Acid, 1 ml perchloric acid, and 1ml hydrofluoric acid and placed into an oven for baking and nitrification for 5 hours. Finally, they were put on an electric heating board for 1 to 2 hours, then mixed with water to give a uniform volume of 10 ml for testing. An ULTIMA inductively coupled plasma spectrometer was used to measure sixteen elements in the soil: Al，As，Ca，Cu，Fe，K，Mg，Mo，Na，Ni，Pb，Se，Sn，Sr，V，and Zn. In this paper, Mo contents are used as sampling data.

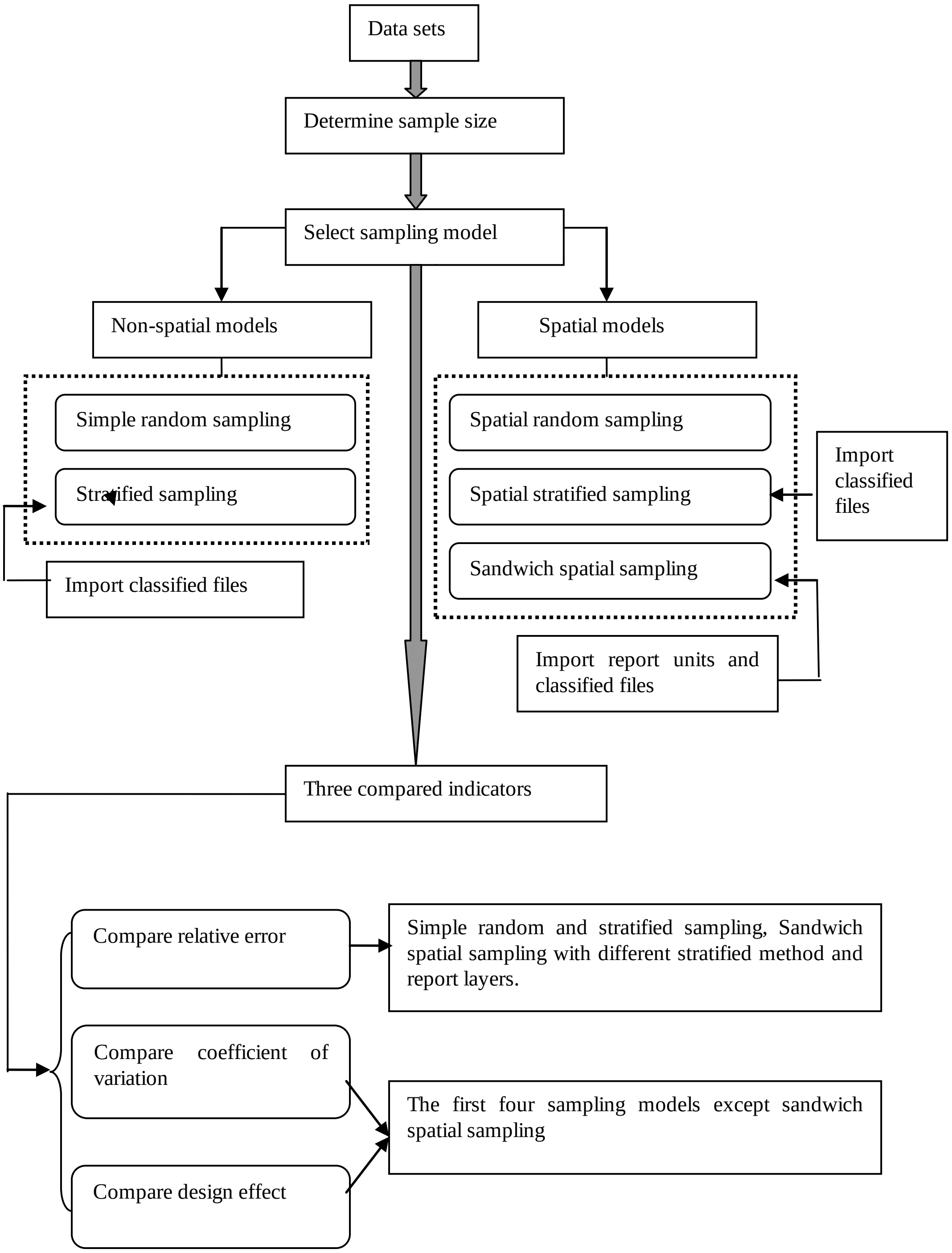

Data sets
Determine sample size
Select sampling model
Non-spatial models
Spatial models
Simple random sampling
Stratified sampling
Import classified files
Spatial random sampling
Spatial stratified sampling
Sandwich spatial sampling
Import classified files
Import report units and classified files
Three compared indicators
Compare relative error
Compare coefficient of variation
Compare design effect
Simple random and stratified sampling, Sandwich spatial sampling with different stratified method and report layers.
The first four sampling models except sandwich spatial sampling

Figure 1 a flow chart of step of this study

Heavy metal contents in the soil depend primarily on the natural levels of soil-forming parent materials and the absence of external contaminants. There are five main soil types in pedogenesis: cinnamon, chestnut cinnamon, loessial soil, regosols soil and Chao soil. In this way, the area is stratified by soil type and geological surface. It can also be stratified geochronologically because elements were introduced at different times. There are eight layers, but in order to keep the sample size reasonable in each layer, it is stratified into four layers. The fourth stratification method is hierarchical clustering accomplished with SPSS.

2.2 Sampling method and algorithm

Five models in SSSI are used: simple random sampling, stratified sampling, spatial random sampling, spatial stratified sampling, and sandwich spatial sampling. The first two models take the classical approach, and the last ones take the geo-statistical approach. The use of different sampling designs and methods may affect the accuracy of the analysis of samples collected in the same area (Paolo de Zorzi *et al.*, 2008). Here, three indicators are selected to compare the efficiency of different sampling models. To get the three indicators, the models are simulated 1000 times each with a range of sample sizes from 10 to 80. Figure1 is a flow chart for the steps which can comprehend the study:

2.2.1 Relative error

Relative error compares the difference between sample mean and its true mean, so the estimated relative error is defined as:

$$\text{relative error} = \frac{\text{abs}(\bar{y}\text{-}\bar{Y})}{\bar{Y}} \qquad (1)$$

where $\bar{y}$ = sample mean

$\bar{Y}$ = observable population mean

The former is estimated by the different models and the latter is acquired by counting the entire observable metal contents.

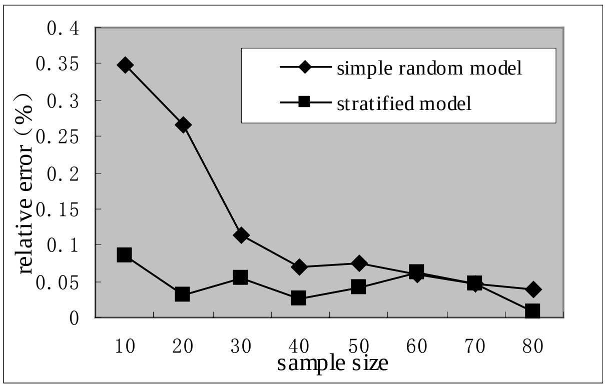


Figure 2 compare relative error between simple random sampling and stratified sampling

Figure 2 compares d_r between simple random sampling and stratified sampling. The sandwich spatial sampling model is the newest method in the SSSI software. It has the same accuracy to the spatial stratified sampling, but it refers to report layers, which can be in any unit, for example, a county border, provincial boundary, watershed, or artificial grid (http://www.sssampling.org/). Table 1 shows the values of the whole relative error of report layers.

report layer	stratified by soil type	stratified by geochronology
Administrative villages	0.180045	0.085048
grids	0.066463	0.052151

Table 1 relative error of two kinds of report layer

2.2.2 Coefficient of variation

The coefficient of variation is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

$$\text{coefficient of deviation} = \frac{\text{Standard deviation}}{m} \qquad (2)$$

$$\text{Standard deviation} = \frac{\sum_{i=1}^{n}(x_i)\ \bar{x}^{\,2}}{n\text{-}1} \qquad (3)$$

where m = sample mean

x_i = value of element content

n = sample size

In this paper it compares the dispersion of each estimator from its true value during the 1000 simulations for each sampling method and sample size (D. M. Chen, 2009).

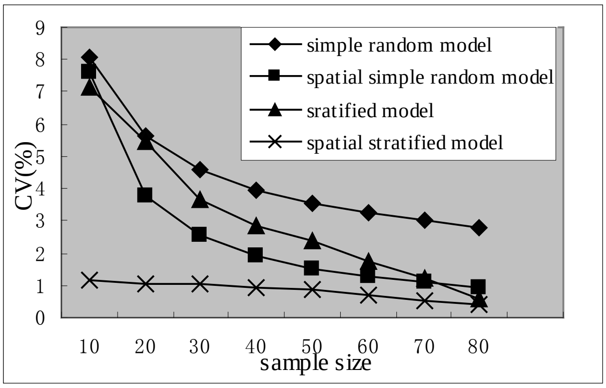


Figure 3 Mo (stratified by soil type)

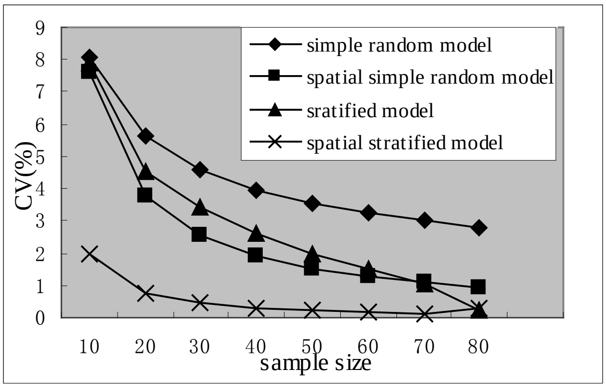


Figure 4 Mo (stratified by geological surface)

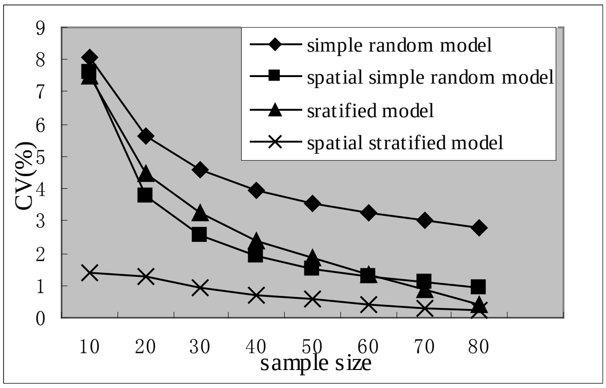


Figure 5 Mo (stratified by Hierarchical cluster)

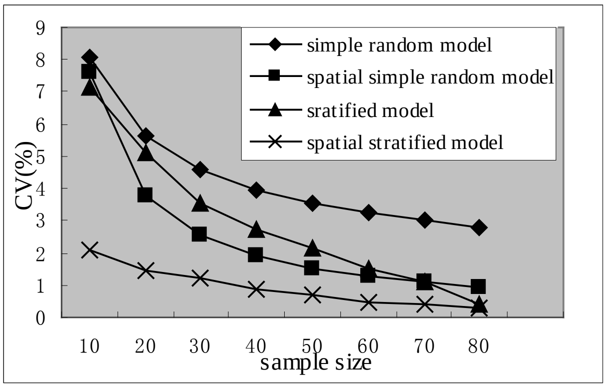


Figure 6 Mo (stratified by geochronology)

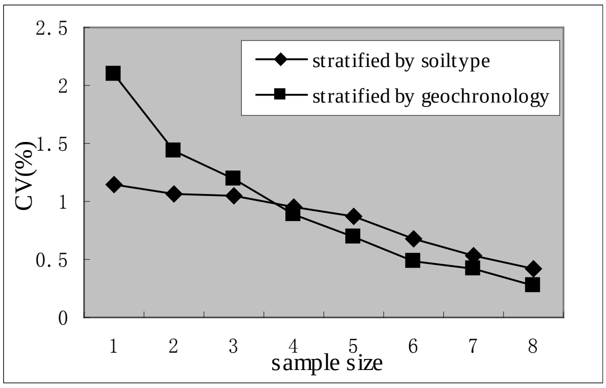


Figure7 Mo compare stratified method soil type and geochronology

Figures 3 through 6 show the CV from four sampling models of Mo elements, stratified by soil type, geological surface, hierarchical cluster, and geochronology. And figure 7 shows which stratified method is more efficient.

2.2.3 Design effect:

Design effect is the ratio of variance of the estimate obtained from the (more complex) sample to the variance of the estimate obtained from a simple random sample of the same number of units (William G. Cochran. 1977). This indicator also compares the efficiency of different sampling models.

3. RESULTS AND DISCUSSION

Although sample data is spatially distributed, but if only compare spatial models, maybe it is not enough, so five models are chosen. Spatial random sampling is the same to simple random sampling in evaluating sampling mean, and spatial stratified sampling is also the same to stratified sampling in it, so only compare relative error between simple random sampling and stratified sampling, and because importing report units in sandwich spatial sampling, each report unit also has its relative error, so compare the whole relative error of report layers. And sandwich spatial sampling has the same accuracy to the spatial stratified sampling in coefficient of variation and design effect, so the last two indicators only compare four sampling models.

3.1 Analyses of the relative error

The simple random sampling model is a method in which each sample has an equal chance of being selected. Stratified sampling divides the whole area into groups based on a special factor. This factor affects the distribution of the research object; there is little variation inside the layer but great variation between layers. Figure2 shows that as the sample size increases, the relative error decreases in the simple random sampling model. With smaller sample sizes, the simple random sampling model has lower accuracy than the stratified sampling model, and the interval is large. When the area is stratified by soil type, each sample is representative in every layer, although sample sizes are small. With larger sample sizes, the stratified sampling model fluctuates within a certain range, but is more accurate than the simple sampling random model from 10 to 80. Therefore, the stratified sampling model has higher sampling efficiency than the simple random sampling model and results in a more representative sample.

Table 1 presents two kinds of report units: administrative villages and grids. Under normal circumstances, the layer created with the stratified sampling model is called the knowledge layer and the number of knowledge layers is less than the number of report layers. The mean and variation of each report layer are obtained with SSSI, and using Formula (1), each report layer’s relative error can been calculated. We can find the whole relative error by combining them. The values of the table are less than 10%, except for administrative villages stratified by soil type.

3.2 Analyses of the coefficient of variance

Figures 3 though 6 show the coefficient of variance (CV) of Mo for the simple random sampling model, spatial random sampling model, stratified sampling model, and spatial stratified sampling model. CV is the ratio of spatial variation to the sample mean. For small sample sizes, the estimated sample mean is less accurate than for large sample sizes. The CV decreases with an increase in the sample size for all sampling models. But the curve fluctuation ranges are different for each sampling model. The spatial stratified sampling model has the highest precision, indicated by its curve below the others on the graphs, and the simple random sampling model has the lowest precision. The spatial random sampling model has higher precision at the middle sample sizes about from 15 to 70. We can conclude that the heavy metal contents in soil have spatial autocorrelation and the precision of the different sampling models is impacted more by the spatial dependency than by the areal heterogeneity for most sample sizes.

Taking spatial dependency into account, the nearer the samples are to one another, the more similar their attributes and the greater their spatial correlation is. Spatial mean variation is calculated from the samples with the following formula:

$$\operatorname{var}\left\{\overline{Z}-Z(A)/Z\right\}=\frac{1}{n}\left[\sigma^2-E\{C(X,Y)\}\right]=\frac{\sigma^2}{n}-\frac{E\{C(X,Y)\}}{n} \qquad (4)$$

where σ^2 = population variation

X 、 Y = random variables subject to uniform distribution in the area A

$C(X,Y)$ = the covariance of random variables X and Y

Classical sampling mean variance is σ^2/n, and spatial sampling mean variance is the above formula with a decrease of $E\{C(X,Y)\}/n$ from classical sampling (Wang J F et al., 2009).

Spatial stratified sampling models take area heterogeneity and spatial dependency into account, not only to keep samples representative, but also to calculate sample autocorrelation. The sampling mean variance is smaller than any other sampling models, but the simple random sampling model does not account for area heterogeneity and spatial dependency, so the efficiency is lower. The stratified sampling and spatial random sampling models account for only area heterogeneity or spatial dependency, so their efficiency is between spatial stratified sampling models and simple random sampling models. For small sample sizes, the stratified sampling model is better than the spatial random sampling model because it shows the dependence of samples is not strong and the samples are more representative. With larger sample sizes, the samples' dependence becomes stronger. Although samples are more representative with the stratified sampling model, the impact degree is not larger than spatial dependence. When the sample size is about 60 to 70 (almost the population size) the spatial dependence reaches its maximum. It tends toward a certain level value, so there are some limitations for spatial dependence at small and large sample sizes.

As shown in Figures 3 through 6, the accuracy of the simple random and spatial random sampling models is the same. There is a difference in the stratified sampling model and spatial stratified sampling model because of different stratification methods. Selecting the best stratification method is important for estimating sampling precision. If the stratification is poor, its accuracy could be lower than simple random sampling. Figure 7 compares the CV of two stratification methods. In the graph, stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. At the smaller samples, the sample in one layer is more close in space by stratified by soil type than by stratified by geochronology, so the samples are more homogeneous in the same knowledge layer. When areas are more homogeneous, a single sample can be more representative, leading to high sampling efficiency. But at the larger sample sizes, geochronological stratification is better, possibly because the Mo content has been more strongly affected by geochronology. When the sample size reaches a certain degree, this effect is manifested.

3.3 Analyses of the Design effect

models sample sizes	Srs	StrRs (1#)	SStrs (1#)	StrRS (2#)	SStrs (2#)
10	0.945	0.891	0.143	0.899	0.228
20	0.664	0.964	0.187	0.798	0.133
30	0.557	0.801	0.230	0.745	0.100
40	0.485	0.719	0.241	0.661	0.070
50	0.430	0.664	0.244	0.556	0.059
60	0.392	0.536	0.209	0.469	0.048
70	0.361	0.402	0.177	0.351	0.043
80	0.338	0.203	0.152	0.092	0.096

Table 2 design effect of spatial simple sampling, (spatial) stratified sampling by soil type and geological surface

models sample sizes	StrRs (3#)	SStrs (3#)	StrRS (4#)	SStrs (4#)
10	0.931	0.172	0.889	0.261
20	0.788	0.228	0.901	0.252
30	0.710	0.203	0.776	0.281
40	0.606	0.178	0.692	0.225
50	0.523	0.156	0.612	0.193
60	0.405	0.127	0.474	0.150
70	0.293	0.106	0.359	0.137
80	0.150	0.087	0.155	0.101

Table 3 design effect of (spatial) stratified sampling by hierarchical cluster and geochronology

Srs: spatial random sampling;
StrRs: stratified random sampling;
SStrs: spatial stratified sampling model
1#: stratified by soil type;
2#: stratified by geological surface;
3#stratified by hierarchical cluster;
4#: stratified by geochronology.

Tables 2 and 3 show the ratio of mean variance for different sampling models to the mean variance from the simple random sampling model. The values in the tables are all less than 1, so the simple random sampling model (Srs) is the least efficient and spatial stratified sampling model (SStrs) is the most efficient. Spatial autocorrelation slightly impacted the sampling accuracy.

The spatial stratified sampling model has higher accuracy than any other sampling models. Samples within the same knowledge layer may be very far apart and even separated from the other layer in space. Therefore, spatial stratified sampling keeps the variation small in the same knowledge layer and large between knowledge layers. It requires balancing the samples in the same layer together in space. In these four stratification methods, the second meets the requirement, so from Tables 2 and 3, 2# has lower values at most sample sizes except at 10 and 80, and 3# has higher accuracy at 10 and 80 samples. Samples in the same knowledge layer in space have high spatial dependence, and $E\{C(X,Y)\}/n$ also increases. But at 10 and 80, the spatial dependence shows its limitations. Hierarchical clustering stratifies samples by their Euclidean distance squared, so the properties of each layer is very consistent within the sample compared to other stratification methods, but they are not together in space, resulting in high sampling efficiency outside the limitations of space.

4. CONCLUSIONS

In this paper, we have looked at both spatial and non-spatial models in the five kinds of sampling models we compared. Of these, sandwich spatial sampling and spatial stratified sampling had the highest efficiency, and the simple random model had the lowest efficiency. The efficiency of spatial random sampling and stratified sampling is dependent on the sample size. The stratified sampling model has higher efficiency than spatial random sampling at large and small sample sizes, but lower efficiency at middle sizes. The efficiency is also up to the stratification method; different methods can affect the sample distribution at the same sample size. Spatial stratified sampling takes spatial dependence into account and keeps samples in the same knowledge layer together in space, so stratification by geological surface has higher efficiency than other stratification methods. By comparing two kinds of stratification methods, by soil type and geochronology, we see that stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. Sandwich spatial sampling has the same efficiency as spatial stratified sampling for the sample mean and spatial mean variation, but it can calculate these values for each report layer. These report layers can be divided arbitrarily to meet the user's needs.

There are two kind of sample distribution: sample layout and without layout, in this paper, without sample layout in the circumstances, we compared the sampling accuracy in different sample sizes with different sampling models, analyzing and discussing the reasons. Because the sample distribution is not very satisfactory, there are some gaps in the data set, so the accuracy of the estimates may not be completely satisfactory, but the study fills gaps in the comparison of current spatial sampling, which is important and can inspire readers to understand the present sampling models and methods. Future work should experiment with relatively uniform sample distributions and also deal with sample with layout using this study method.

ACKNOWLEDGEMENT

The authors would like to thank Professor Chaoyang Wei and Tongbin Chen for their helpful discussions and constructive suggestions, and also appreciate the assistance of Luke ldrisk, Antoine Soullié and Hexiang Bai in the programming.

REFERENCES

Cressive N., 1991. *Statistics for spatial data*. Wiley, New York.

DongMei Chen, Hui Wei., 2009. The effect of spatial autocorrelation and class proportion on the accuracy. *ISPRS Journal of Photogrammetry and Remote Sensing* 64, pp. 140-150.

Haining, R.P., 1988. Estimating spatial means with an applicatio n to remote sensing data. *Communication Statistics – Theory and Methodology*, 17(2), pp. 537-597.

Haining, R.P., 2003. Spatial data analysis: theory and practice. *Cambridge University press*, Cambrdge.

J. WANG, J. LIU, D. ZHUAN, L. LI and Y. GE., 2002. Spatial sampling design for monitoring the area of cultivated land. int. j. *remote sensing*, 23(2), pp. 263–284.

Jinfeng Wang, Robert Haining, Zhidong Cao. 2009. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. DOI ： 10.1080/13658810902873512.

Jinfeng Wang, chengsheng Jiang, Lianfa Li, Maogui Hu., 2009. *Spatial Sampling and Statistical Inference* (in chinese). Beijing: Science Press. pp. 17-20

Li L, Wang J, Liu J., 2005.Optimal decision-making model of spatial sampling for survey of Chinas land with remotely sensed data. *Sci China Ser D* 48(6), pp. 752-764.

Paolo de Zorzi, Sabrina Barbizzi, Maria Belli, Renzo Mufato, Giuseppe Sartori, Giulia Stocchero., 2008. Soil sampling strategies: Evaluation of different approaches. *Applied Radiation and Isotopes* 66, pp. 1691 – 1694.

Wang Guanyu, Pan Mao, Liu Xida, Liang Haihua., 1992. On the relationship between the concentrations of elements in soil and the types of soil-forming parental material in Shandong province, China. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 28(4), pp. 475-485.

William G. Cochran., 1977. *Sampling Techniques*. Third edition, Professor of Statistics, Emeritus Harvard Universtiy. pp. 85-88.

Zhang Zhuodong , Zhang Keli , Wu Jilei , Zhang Ting , Zheng Xiaoying.2008. Spatial information analysis of characteristics of soil trace element contents in high incidence areas of birth defects. Journal of Zhejiang University (Agric1 & Life Sci1) 34 (6),pp. 684～691. DOI: 10. 3785/ j. issn. 100829209. 2008. 06. 015.

http://www.sssampling.org/

ATTRIBUTE UNCERTAINTY MODELING IN LUNAR GIS DATA

P. Weiss [a, *], W.Z. Shi [b], K.L. Yung [a]

[a] Industrial and Systems Engineering Dept., The Hong Kong Polytechnic University, Kowloon, Hong Kong, People's Republic of China - weiss@cppm.in2p3.fr, mfklyung@inet.polyu.edu.hk
[b] Dept. for Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, People's Republic of China – lswzshi@polyu.edu.hk

KEY WORDS: Uncertainty, GIS, Spatial accuracy, Modelling, Correlation, Geography, Moon

ABSTRACT:

A novel methodology to evaluate uncertainties in lunar element abundances is presented. Contrary to most terrestrial applications, lunar GIS data cannot be verified by in-situ measurements due to the limited number of ground control points and their reduced spatial extend. This investigation evaluates the uncertainty in lunar element abundance measurements without the use of ground-checks but by statistical evaluation and comparison of datasets. We find that major elements (Oxygen, Iron, Aluminium, Titanium, Silicon and Magnesium) show distinct correlations between each other. This allows calculating the abundance of an element by deriving its value through a correlation law with another element. By using this method, we can verify remote measurements of the above mentioned geochemical components, and identify regions on the Moon where these correlation laws do not apply. These derivations can be explained by i) an erroneous measurement or ii) by an exotic mixture of elements in the lunar soil. Based on these considerations, conclusions can be drawn about the attribute uncertainty of geochemical measurements in the lunar soil. A special observation of this work was that most theoretically obtained values fit well to the measured ones. High derivations however appear mainly in the Near Side lunar mare regions where the correlation model does not fit.

1. INTRODUCTION

Our celestial neighbour, the Moon, has shifted back into the focus of scientific interest. Various spacecraft are planned to be sent to the Moon in the coming decades. The objective of these international missions is the cartography of the lunar surface, the identification of resources for in-situ resource utilisation (ISRU), and, ultimately, the return of man to the Moon (Landis, 2001). Consequently the precision of remotely sensed data needs to be considered for the choice of future landing sites. The attempt of translating the physical reality into a digital model leads to uncertainties. Since it will not be possible to eliminate those completely, they need to be represented and taken into account in any Decision Support System that is based on such GIS data. Uncertainties "*must always be a notation that provides the language for reasoning and allows decision-makers to evaluate the potential presence of errors in GIS data*" (Zhang and Goodchild, 2002). The challenge in uncertainty evaluation of remote measurements is to quantify the accuracy of this data at the local scale. In terrestrial GIS applications, this process is termed "validation". Validation of remotely sensed data is achieved by analytical inter-comparison to ground control checks, reference data or model outputs (Justice et al., 2000). This can be done by the collection of in-situ measurements, low altitude photography by aircraft or the comparison to independent, but identical in content, satellite observations. In the case of the Moon, very little ground-control points exist which can be used to validate orbital measurements. Extrapolating this data to the whole surface of the Moon would lead to errors, since their correlation to orbital measurements is rather low, as it was shown elsewhere (Weiss et al., 2009).

We describe efforts to model attribute uncertainties in geographic data of the Moon, where no or little in-situ measurements are available. The methodology that is used, evaluates the uncertainty by analysing correlations between orbital measurements of different elements. The uncertainty is estimated based on the deviation of the measured value from the though-correlation-expected one. The methodology was applied in a first work on the distribution of Oxygen in the lunar soil (Weiss et al., 2009). In this paper, we extend the model to six other major elements in the lunar soil: Silicon, Titanium, Aluminium, Iron, Magnesium and Calcium.

This paper is structured in two parts: In the first chapter, we will review different factors that increase the uncertainty in remote measurements of the lunar surface. We will demonstrate and discuss the problem of using the few in-situ measurements available as "ground-truth" for remote measurements. In the second part, we will present the method to estimate attribute uncertainty in lunar GIS data by using a correlation law between different elements. This methodology will be applied to estimate the uncertainty of the element measurements mentioned above. The result is a chart of the lunar surface with regions where the measured value does not fit to the theoretically derived one. We will conclude this work by studying the spatial distribution of the identified regions and the possible origin of the deriving values.

2. UNCERTAINTY IN LUNAR GIS DATA

2.1 Sources of uncertainty in lunar surface data

In GIS data, uncertainties can be categorized following their consequence, namely i) positional uncertainties, ii) attribute uncertainties or iii) temporal uncertainties. The latter play a rather minor role in the context of lunar geography; the processes that form or modify the Moon's surface are slow, and the frequency of observations low. However, temporal

[*] Corresponding author.

uncertainty need to be taken into consideration in the photogrammetry of the lunar surface since changes of the lightening conditions between the lunar day and night can lead to different topographical interpretations. Positional and attribute uncertainties are considered of high importance in selenographic data. The low data quantity, coverage, and scale, make it today very difficult to establish precise models of the Moon's geology and chemistry. In the coming phase of lunar exploration, much effort is focusing on the measurement of the chemical composition of the lunar surface. Uncertainties in these measurements, the attribute of the spatial data, might lead to misinterpretations and faulty decisions.

Attribute uncertainties can originate in system limitations, mission limitations or have a target specific origin. System limitations include all parameters that are related to the hardware of the remote sensing satellite: its measurement capabilities, its resolution, its temperature and the Internal Orientation Parameters (IOP) of the instrument. Mission limitations include the spacecraft's ephemeris, orbital parameters and specific events during the mission time: The satellite's altitude, orbital velocity and Exterior Orientation Parameters (EOP) influence the capability of the instrument to perform measurements on the surface. Solar Particle Events (SPE) and the Sun's relative position further influence the measured data. Third bodies' influence decreases the accuracy of the satellite's position determination, and therefore increases the uncertainties of its measurement data. Finally, the characteristics of the target can further determine the precision with which the surface is analyzed: terrain morphology, surface roughness, surface albedo, surface temperature and MASCONS can bring variations in the remote measurement's precision. Figure 1 resumes schematically the factors mentioned above.

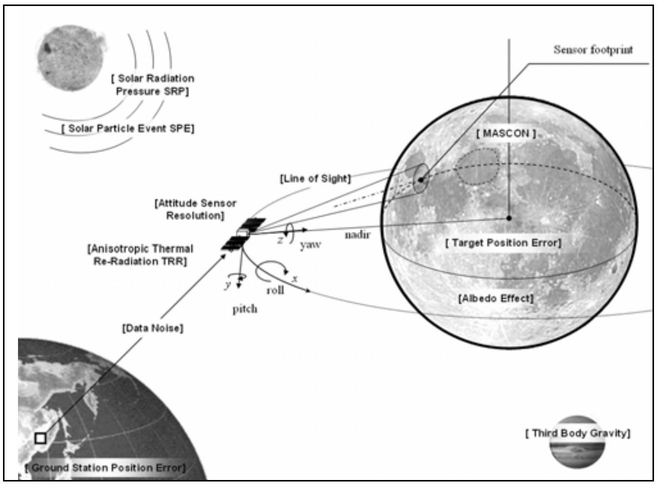

[Solar Radiation Pressure SRP]
[Solar Particle Event SPE]
Sensor footprint
[MASCON]
[Line of Sight]
[Attitude Sensor Resolution]
[Anisotropic Thermal Re-Radiation TRR]
z
yaw
nadir
[Target Position Error]
y
x
roll
pitch
[Albedo Effect]
[Data Noise]
[Third Body Gravity]
[Ground Station Position Error]

Figure 1. Parameters that influence the attribute uncertainty

Modelling the uncertainty of remote measurements as a function of the above mentioned parameters becomes therefore a complex problem. It requires a perfect knowledge of all hardware parameters, the mission's ephemeris and eventual surface influences.

An alternative is to compare the remote measurement with ground-control checks to search for potential derivations. The problem is, however, that there are very little in-situ measurements available, as it will be shown in the following section.

2.2 The dilemma of the missing ground-checks

The Moon is certainly -apart from planet Earth- the most analysed object in the Solar System. It remains, however, a largely undersampled surface: Astronauts and robots, from several US and Russian missions, brought back 382kg of Moon dust and rock (Vaniman et al., 1991). The landing sites of these missions are limited to the equatorial region of the Near Side. We dispose today of no in-situ measurement of the poles or the Far Side. The surface samples are not only limited in their spatial extend, but also in their geological context, where they were taken: Most missions aimed at (safer) landing sites in the lunar Mare regions. Actually only the Apollo 16 mission brought back samples from a site that is considered as Highland region. Figure 2 gives an overview of the different landing sites where lunar soil was returned to Earth. The number of samples and weight is given for the Apollo missions. In an earlier work it was shown that the returned samples do not correlate well with orbital measurements of element abundances (Weiss et al., 2009). This fact makes it difficult to extrapolate remote data to a local scale; both measurements do not fit together. Several reasons can be stated which might explain this divergence: The rocks that were returned by astronauts and robots did eventually not represent the bulk of the surface material. While the first manned mission to the Moon, Apollo 11, recovered a majority of lunar regolith, all consequent missions brought back increasing numbers of solid rock (Vaniman et al., 1991). The collected samples were eventually chosen since they draw the attention of the astronauts or the mission controllers. But those might represent a rather exotic lithology compared to the location. A second potential explanation is that the footprint of the remote measurements largely exceeds the exploration range of the manned missions (Vaniman et al., 2002; Berezhnoy et al., 2006). The recovered samples therefore represent only a small fraction of the area covered by the satellite's instruments.

However, Apollo 17 in-situ measurements were used to extrapolate the geological data to unsampled areas in the same region by using Clementine UV-VIS remote measurements (Jolliff, 1997). The author notes good correlations between the Iron oxide values of the Apollo samples and the spectral parameters of the Clementine data. Elphic and co-workers (1998) developed algorithms to derive the content of Iron oxide and Titanium dioxide lunar wide from Clementine spectral data. Despite these efforts, it is obvious that Apollo and Luna samples cannot serve as global “ground-truth” for the validation of orbital measurements of the lunar surface chemistry. Other techniques are needed to gather the ground-truth.

2.3 A geochemical model based on element correlation

The basic idea behind the following concept is that similar soils, which underwent similar formation processes, should exhibit similar chemical mixtures. A soil that has a specific abundance of one element should therefore have a specific mixture with other elements. The knowledge about the abundance of one element can therefore be used to derive the abundance of another element, if both correlate.

This method allows deriving the composition of a soil by knowing only some parts of its elemental composition. It furthermore allows evaluating the probability of justness of remote measurements by inter-correlating different elemental datasets. The advantage of this methodology is that it can work without Ground Control Points. Nor does it need independent datasets of the same kind (the same measurements by different satellites). Both are very limited for the lunar surface.

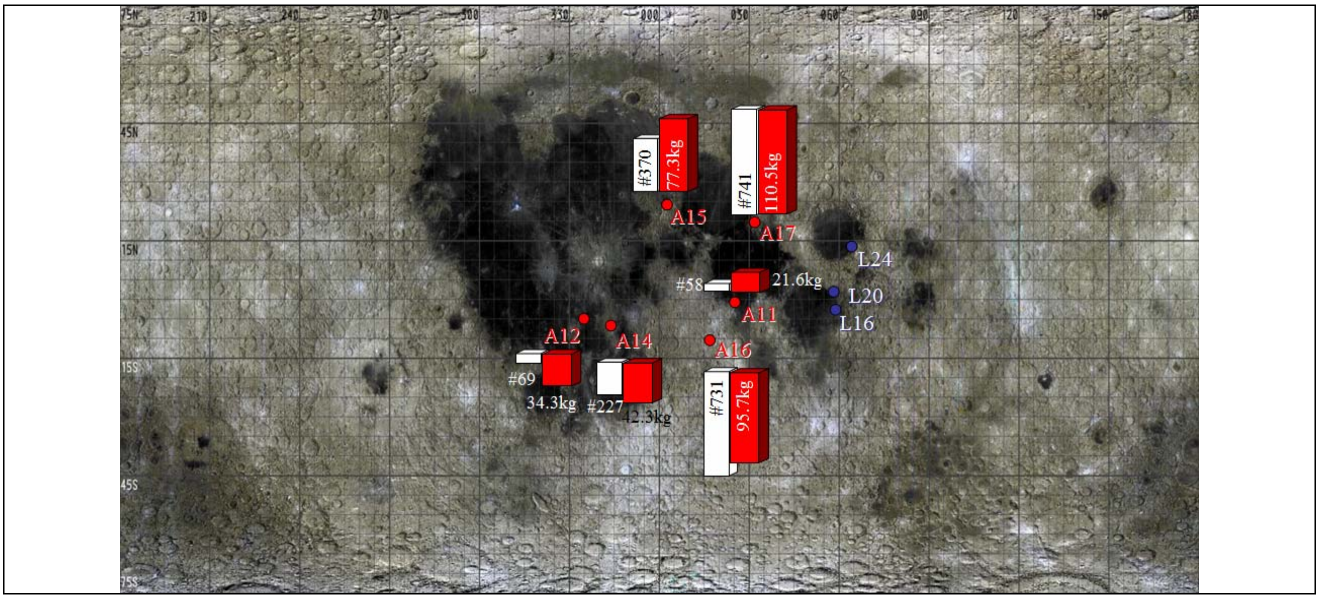


Figure 2. Positions on the Moon where in-situ samples have been returned to Earth. (A) Apollo mission and (L) Luna missions. The number of samples and the overall weight is stated in the columns.

The main condition of this approach is that there exists a correlation rule between two chemical elements. The correlation between the elements x and y can be analysed through the fitting function given in (1), with its parameters α, the intercept and ß, the slope,

$$y = \alpha + ßx \qquad (1)$$

where the intercept is

$$\alpha = \left(\sum_{i=1}^{n} y_i - ß \sum_{i=1}^{n} x_i\right) / \, n\text{`} \qquad (2)$$

The slope determines if there is a correlation between the two elements:

$$ß = \frac{L_{xy}}{L_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (3)$$

If β is larger than zero, then there is a positive correlation between both elements. If β is smaller than zero, then there is a negative correlation between the elements. And if β is equal to zero, then there is no or no simple correlation between both elements.

The above considerations can be applied to the measurements done by the Lunar Prospector mission (Prettyman et al., 2002). Measurements of Oxygen, Titanium, Silicon, Aluminium, Iron, Calcium and Magnesium are available in 5° data products from the PDS Geosciences Node. Figure 3 shows one of the correlation plots, namely the one of Oxygen and Iron, as measured by the Lunar Prospector. The solid line in the middle of the diagram is the regression line as determined through (1). The dotted lines show the one-sigma and two-sigma borders of this correlation function. It was shown by Weiss et al., (2009) that Calcium shows no simple correlation to the other elements. This element was therefore excluded from the further considerations.

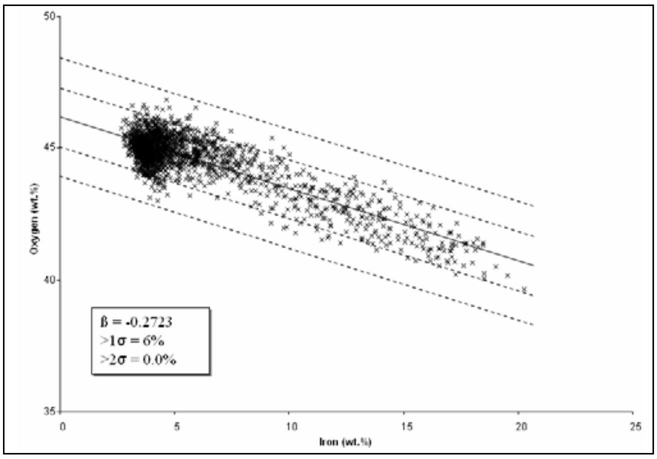


Figure 3. Negative correlation between Oxygen and Iron.

A theoretical abundance value of the above stated elements can now be calculated by using their correlation between each other. The theoretical values are then compared with measured ones at their specific position. Figure 4 shows the measured abundance of Oxygen (top) with the theoretically derived one (bottom). The yellow boxes mark surface cells that show derivations larger than one sigma between both values.

Surface cells that show a deviation of one or two sigma were flagged in the GIS following to (4).

$$\frac{|\ \alpha + ßx - y'\ |}{\sigma} = \begin{cases} < 1 \rightarrow \text{no dispersion} \\ \geq 1 \rightarrow \text{flagged 1} \\ \geq 2 \rightarrow \text{flagged 2} \end{cases} \qquad (4)$$

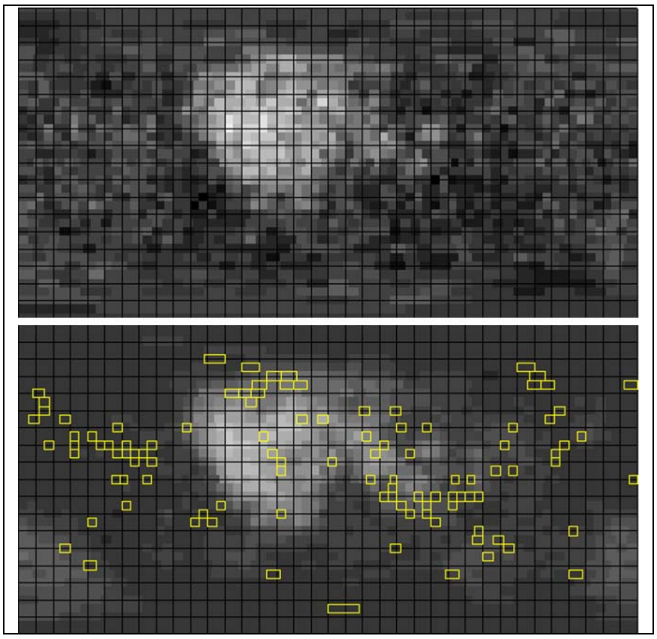


Figure 4. (Top) Oxygen content of the lunar soil as measured by the Lunar Prospector mission. (Bottom) Oxygen value as derived as a function of the measured Iron content.

All six elements (Calcium was excluded) were compared in this way with each other. Table 1 summarizes the percentage of fitting surface cells.

As a first result is that the majority of the theoretical values fit well to the measured ones. This offers the possibility to derive more precise charts of specific elements by deriving those from higher resolution data of other elements. Oxygen can, for example, be a quite precise indicator of Titanium and Iron. Iron as function of Titanium can even be predicted within one sigma in 98% of the cases. Oxygen, as one of the main ISRU elements could be derived by measuring the Iron content of the surface since in our study 94% of the calculated Oxygen values fitted to the measured ones.

In the following, however, we will study the spatial distribution of the values that do not fit and try to find possible explanation to this phenomenon.

Table 1: The input elements (left column) were used to derive the output elements (first row). The percentage shows the number of correct values within one sigma deviation.

└→ Output element							
	O	Si	Ti	Al	Fe	Mg	Ca
O	■	89%	93%	83%	94%	76%	69%
Si	89%	■	90%	72%	87%	80%	71%
Ti	92%	87%	■	84%	98%	80%	77%
Al	84%	74%	88%	■	90%	82%	71%
Fe	94%	86%	97%	90%	■	85%	88%
Mg	80%	83%	87%	82%	89%	■	82%
Ca	77%	76%	85%	72%	84%	83%	■
↑ Input element							

3. FROM CORRELATION TO UNCERTAINTY

3.1 Spatial distribution of outlying surface cells

In the previous chapter, a method was presented to derive element abundance values as function of a correlation to another element. It was shown that this method leads to a majority of fittings (within one sigma). However, some surface cells show large deviations between the theoretically calculated value and the measured one. These cells are defined in the following as uncertain, because their values do not follow the correlation rule. We will discuss in 3.2 different explanations for these outlying values.

It is now interesting to study the spatial distribution of these outliers. Figure 5 shows three example charts of Titanium, Iron and Aluminium. The intensity of red colour indicates the number of deviations. If the cell is marked fit, then all five input elements delivered output correlations that fitted within the one sigma limit. If the cell is marked in deep red, then four of the output values, as function of the input element, derived. Two types of cells can be identified: single occurrences of outliers and in regions clustered cells. The Iron and Titanium show a clear concentration of the outliers in the Mare regions of the Moon (especially in the Oceanus Procellarum).

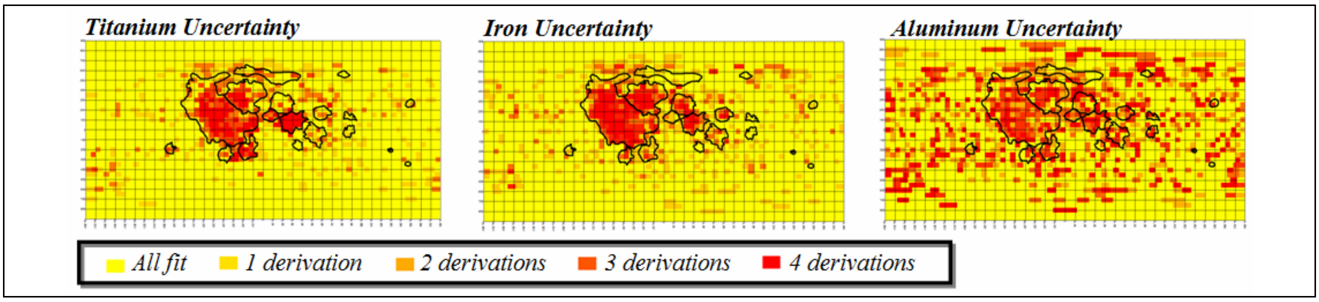


Figure 5. Uncertainty charts for Titanium, Iron and Aluminium. The charts show a concentration of the derivation of values in the Oceanus Procellarum region. This fact offers the possibility to refine the models by separating the values into two categories.

3.2 Possible explanations

We find two explanations for the occurrence of deriving abundances: i) Either the measured value is in deed erroneous or ii) the region shows an exotic mixture of chemistry due to a special formation process.

In the case of the Mare regions it is more likely that the latter case applies. The above method used the totality of the abundance measurements of the lunar surface. However, in lunar geology, there is a clear separation between the Mare regions on the Moon and the Highland regions. While our method seems to deliver good results for the Highland regions, its deriving values are mainly located in the lunar Mare.

3.3 Conclusions and further work

A method was presented to evaluate the uncertainty in lunar element abundance values by comparing different element measurements done by the same remote sensing satellite.

The paper discussed one challenge in the exploration of the lunar surface by observation methods, namely the lack of sufficient ground truth data to validate the remote measurements. The little in-situ data that was gathered by manned and robotic missions is insufficient in quantity and spatial extend to serve as model for the global surface.

A novel method was developed to derive the chemistry of the lunar surface by correlating different elements with each other. The fact that some elements show a correlation allows to evaluate the uncertainty in the data products. If the theoretical value corresponds to the measured one, then there is a high probability that the remote measured value is correct. We showed that the majority of the surface cells fit well to this rule. However, large deviations occur in the Near Side Mare regions. A possible direction of further work is therefore to separate the Mare regions from this model and to develop an own correlation law for the Mare regions of the Moon.

4. ACKNOWLEGEMENTS

The authors would like to acknowledge The Hong Kong Polytechnic University for the support of this study. The datasets used here were produced by the Lunar Prospector mission and are disclosed by NASA and the Washington University in St.Louis on the website of the PDS Geoscience Node (http://pdsgeosciences.wustl.edu/missions/lunarp/reduced_special.html). We would like to express our thankfulness to the team that developed these products and made them freely available to the scientific community.

The authors would like to thank two unnamed reviewers for their helpful input and encouraging comments.

5. REFERENCES

Berezhnoy, A.A., Hasbe, N., Kobayashi, M., Michael, G., and Yamashita, N., 2006. Petrologic mapping of the Moon using Fe, Mg and Al abundances, *Advances in Space Research,* 37, pp. 45-49.

Elphic, R.C., Lawrences, D.J., Feldman, W.C., Barraclough, B.L., Maurice, S., Binder, A.B., and Lucey, P.G. 1998. Lunar Fe and Ti abundances Comparison of Lunar Prospector and Clementine data, *Science*, 4, 281, pp. 1493-1496.

Joliff, B.L., 1997. Clementine UV-VIS multi-spectral data and the Apollo 17 landing site: What can we tell and how well?, *28th Lunar and Planetary Science Conference*, Abstract #1770, Available online at www.lpi.usra.edu/meetings/lpsc97/pdf/1770.PDF (accessed 90th July 2008).

Justice, C., Belward, A., Morisette, J., Lewis, P., Privette, J. and Baret, F., 2000. Developments in the 'validation' of satellite sensor products for the study of the land surface, *International Journal of Remote Sensing*, 21, pp. 3383-3390.

Landis, G.A., 2001. Materials refining on the Moon, *Acta Astronautica*, 60, pp. 906-915.

Prettyman, T.H., Feldman, W.C., Lawrence, D.J., McKinney, G.W., Binder, A.B., Elphic, R.C, Gasnault, O.M., Maurice, S., Moore, K.R., 2002. Library least squares analysis of Lunar Prospector gamma ray spectra, *33rd Lunar and Planetary Science Conference*, Abstract # 2012.

Vaniman, D., Dietrich, J., Taylor, G.J., and Heiken G., 1991. Exploration, samples, and recent concepts of the Moon, In Lunar sourcebook a user's guide to the Moon, Editors G.H. Heiken, D.T. Vaniman, B.M. French, Cambridge University Press, Cambridge USA.

Vaniman, D., Lawrence, D., Gasnault, O., and Reedy R., 2002. Extending the Th-FeO sampling range at Apollo 14: Under the footprint of Lunar Prospector, Abstract No 1404, Proceedings of the 33rd Lunar and Planetary Science Conference, Houston, USA.

Weiss, P., Shi, W.Z., Yung, K.L., 2009. Attribute uncertainty modelling in lunar spatial data, *International Journal of Remote Sensing* (in press).

Zhang J. and Goodchild M., 2002. Uncertainty in Geographical Information, Taylor and Francis, London, 2002.

ASSESSMENT OF EXTENSIONAL UNCERTAINTY MODELED BY RANDOM SETS ON SEGMENTED OBJECTS FROM REMOTE SENSING IMAGES

X. Zhao [a, b], X. Chen [a, c, *], L. Tian [a], T. Wang [b], A. Stein [b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University,
Luo Yu road 129, China – (cxl, tianliqiao)@lmars.whu.edu.cn
[b] International Institute for Geo-Information Science and Earth Observation, Hengelosestraat 99, Enschede,
Netherlands – (xzhao, tiejun, stein)@itc.nl
c The Key Lab of Poyang Lake Ecological Environment and Resource Development, Jiangxi Normal University,
Nanchang, Jiangxi, China

Commission II, WG II/4

KEY WORDS: accuracy assessment, uncertainty, random set, image segmentation, Poyang Lake

ABSTRACT:

A newly developed random set model has been applied to model the extensional uncertainty of a wetland patch. The objective of this research is to explore the corresponding variables collected on the ground for validating the uncertain image objects and to report the quality of the random set modeling. The independent samples t-test and a correlation analysis have been used to identify the main variables, whereas the overall accuracy and Kappa coefficients quantify the quality of the random set model. The results show that significant correlations exist among covering function, *Carex* coverage and NDVI. This suggests that the covering function of the random set can be quantified and interpreted adequately by NDVI derived from satellite images and *Carex* coverage measured in the field. In addition, the core set of the random set has an overall accuracy of 85 percent and a Kappa value equal 0.54, being higher than the median set and support set. We conclude that the random sets modeling of uncertainty allows us to perform an adequate accuracy analysis.

1. INTRODUCTION

Conventional pixel-based and object-based classification approaches generate maps with exclusive categories. These hard classifications are designed for mapping discrete objects and clear bounded land cover, but they are not appropriate for mapping continuous landscapes in nature such as wetlands (Woodcock and Gopal 2000). The limitation of hard classification to represent transition zones and uncertain boundaries has been a motivating factor for the development of alternative approaches based on uncertainty handling theories such as fuzzy set theory (Zadeh 1965) and random set theory (Matheron 1975; Cressie 1993). A number of soft classification and uncertainty modeling methods have been developed for classifying and representing nature landscapes such as beach (Van de Vlag and Stein 2007) and grassland (Zhao et al. 2009a), and for modeling dynamic phenomena such as fire spread (Vorob'ov 1996) and flooding (Stein et al. 2009a).

The accuracy assessment of soft classification results, however, remains a theoretical and practical difficulty. Foody (2002) gave overviews on accuracy assessment issues and current challenges. For the pixel-based soft classification, such as fuzzy classification, some research managed to perform validation by fuzzy confusion matrix which was extended from conventional assessment approach (Woodcock and Gopal 2000). Accuracy assessment of extracted image objects by object-based classification or segmentation has bigger challenge (Zhan et al. 2005), especially when objects are uncertain (Stein et al. 2009b). The main difficulty is that the information about uncertainty represented in the results does not always have corresponding objects in the field. This is because even on the ground, due to vague and unambiguous boundaries, the delineation of uncertain objects like a city may be impossible. In addition, detailed reference data which is critical for validating soft classification and uncertainty modeling results, are often unavailable, especially when the historical field data was only collected for hard classification.

Extensional uncertainty refers to the uncertainty in identifying the geometric elements that describe the spatial extent of the object (Molenaar 1998). Zhao et al. (2009b) have applied the random set model to represent extensional uncertainty of extracted image object from Landsat TM image in 2004. The accuracy of the random set model derived from a historical image is difficult to assess. Since the synchronous data with detailed information are of a significant importance for assessing model uncertainty, a clean and almost synchronous HJ-1A image with comparable spatial resolution (i.e., 30 m) two days before the survey was acquired. Moreover, the sampling plan was specially designed for investigating zonation pattern of wetland grassland and for accuracy assessment of random set model before ground survey. The objective of this research is twofold: (1) to explore the corresponding measurable variables collected on the ground for validating the uncertain image objects modeled by random sets, (2) to quantify the quality of the random set modeling results.

* Corresponding author.

2. METHODS

2.1 Study Area and Image Pre-processing

The study area PLNNR (29°05′ - 29°18′ N, 115°53′ -116°10′ E) is located to the northwest of the Poyang Lake in the JiangXi province, central China. Nine lakes in the PLNNR are connected to the Poyang Lake during the high water levels in summer and disconnected when water levels are low in spring, autumn and winter. All kinds of wetland vegetation are blooming in spring and serve as important habitats and forages for spring migrating birds. In the summer flooding time, grasses growing at high elevations (e.g., *Miscanthus*) remain stand above the water, whereas sedges (e.g., *Carex*) and submerged aquatic species (e.g., *Potamogeton*) at lower elevations are flooded beneath. When the flood is gone and winter migration birds arrive in autumn from late September, only few kinds of vegetation like *Carex* start to turn green again and thrive until winter come. Other vegetation communities become senescent in autumn and dead in winter. When sedges at the lower elevations are shooting up gradually, different kinds of birds forage on leaves and rhizomes of young sedges and rhizomes of submerged aquatic species in different elevation zones. Taller sedges also provide bird's habitat and shelter (Wu and Ji 2002). Wetland grassland has biggest width of zonation and also largest area in Banghu, therefore, our ground survey was carried out surrounding Banghu lake.

HJ-1A/B satellites were launched on Sept 6, 2008 from China. One scene of a HJ-1A image on November 24, 2009 with 30m resolution was applied and downloaded from the China Centre for Resource Satellite Data and Applications (CRESDA). As a second level product, radiometric correction and systematic geometric correction have been done before downloading and the image is under UTM WGS84 projection. A topographic map of scale 1:10,000 in PLNNR was used as the geographic reference data for more accurate geometric correction. The root mean squared error of the geometric correction was less than 10 m. The Normalized Difference Vegetation Index (NDVI) was calculated for the HJ image:

$$\text{NDVI}=\frac{\rho_{nir}-\rho_{red}}{\rho_{nir}+\rho_{red}} \qquad (1)$$

where ρ_{red} and ρ_{nir} stand for the spectral reflectance measurements acquired in the red and near-infrared band, respectively.

2.2 Ground Survey

From October 26th until November 6th, 2009, a ground survey was carried out around Banghu lake to investigate the zonation pattern of wetland grassland. Although it is assumed that for a random set modeling, some versions of random sampling preferably stratified unaligned random sampling be used rather than a systematic sampling. But in order to verify the gradual change of vegetation and restricted by accessibility and costs, four typical transects have been designed, starting from the river bank and perpendicularly crossing different zonations until they reach the lake bank of Banghu (Fig. 1). These four transects have 27, 24, 14, 8 sample plots respectively, so that 73 plots have been sampled in total.

Sample plots with size of 30m * 30m are distributed evenly along transects, and were fixed by measuring tape. The location of each sample plot was measured by GPS at the centre of the square, with the accuracy less than 10 m. Within the plots, the following variables were recorded: land cover types, vegetation types (communities or species), vegetation height, percent vegetation cover and bird signs such as drops and falling feather. Furthermore, five or eight subplots (1m * 1m) were also established within some 30m * 30m plots to measure the less homogeneous plots and took averaged variables. The spectral characteristics of typical vegetation types were measured within 1m * 1m subplots using SVC field-portable spectroradiometer. For each reading of the spectroradiometer, around 30 separate measurements were taken, which were then averaged for each 1m * 1m subplot.

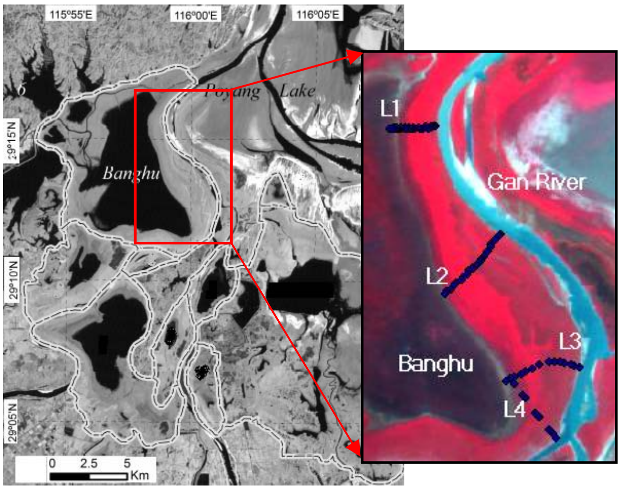


Fig. 1. Nine lakes in Poyang Lake National Nature Reserve (dish line on the left figure) and sample plots along four transects (L1-L4) on false colour HJ image (blue dots on the right figure).

2.3 Image segmentation and extensional uncertainty modelling

Random set theory has been employed as a foundation for the study of randomly varying populations and randomly varying geometrical shapes (Stoyan and Stoyan 1994) and thus applied in our previous research for modelling extensional uncertainty of image objects (Zhao et al. 2009a). In this research, our target object is the *Carex* patch which is located on the east bank of Banghu (Fig. 1). Since *Carex* is the dominant green vegetation in October, we use NDVI image as input for the region growing segmentation. Based on a threshold range (minimum and maximum pixel values), the region growing algorithm expands from a small seed combining connected pixels within pre-specified limits (Russ 2007). Vegetation patches with vague boundaries are sensitive to the setting of parameters in the region growing algorithm. Therefore, by slightly changing the parameters, e.g., under normal distribution with a small value of variance, and segmenting iteratively, a set of resulting objects established a random set (Zhao et al. 2009a; b). The image object with extensional uncertainty thus was modeled by the generated random set.

The procedure is as follows: Firstly, the growing seed was selected inside the central interior part of the target object. Then we initialized the parameters, i.e., upper and lower threshold. For each parameter, a random number was generated from a normal distribution with the initialized parameter as the mean and a preset variance. Fourthly, start the segmentation and obtain object O_i as one sample of the random set. Iterating the above steps several times and a set of resulting objects establish a random set. Several characteristics of random set can be used to describe the extensional uncertainty of objects. For example, n polygons resulting from n times segmentation are samples of the random set, denoted as $O_1, \ldots, O_n$. The probability that pixel $x \in R^2$ occupied by the random region can be determined by $\Pr_{\bigcup_{i=1}^{n}} O_i(x)$. An estimator of the covering function of random set Γ is obtained as:

$$\Pr_{\Gamma}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{O_i}(x), x \in R^2 \qquad (2)$$

Where $I_{O_i}(x)$ is the indicator function of $O_i(x)$. The covering function can be interpreted as the probability of the pixel x on space R^2 being covered by the random set. All the pixels with covering function equal to or larger than p construct a p-level set of the random set. The 0-level set, 0.5-level set and 1-level set are called support set, median set and core set respectively. In practice, we can adjust the support set to 0.05-level set and core set to 0.95-level set, in order to avoid extreme outliers. Further theoretical details about random set models and technical details about segmentation can be found in our previous work (Zhao et al. 2009a; b).

After analyzing the field data, we found that pixels with pure *Carex* have maximum NDVI values around 0.6, whereas 50 percent *Carex* coverage corresponds to NDVI values around 0.45. Since plots with NDVI larger than 0.6 are dominated by *Carex*, to simplify the random set generation procedure, we fixed 0.7 as the maximum threshold in region growing segmentation and do not apply randomization on it. On the other hand, we selected 0.45 as the initialized minimum threshold, and generated 100 random numbers from normal distribution with mean equal to 0.45 and σ^2 equal to 0.1. Finally, we placed the growing seed at the location of one sample plot where 100 percent *Carex* was recorded, and used the 100 randomized threshold intervals to obtain 100 objects and modelled them as a random set Γ .

2.4 Accuracy assessment

In order to validate the modelling result, all the 73 sample plots are used for accuracy assessment. The overall accuracy (OA), producer accuracy (PA), user accuracy (UA) and kappa coefficient are derived from error matrix. Each of these provides a different summary of the information contained in the error matrix. A widely applied kappa z-test (Congalton et al. 1983) is also used to test for statistically significant differences in accuracy of outputs. Independent samples t-tests in SPSS software were adopted to determine if the mean value of the *Carex* coverage is different for sample plots which are included and excluded by the median set. Moreover, correlation between *Carex* coverage, NDVI and covering function value were quantified by regression models.

3. RESULTS

3.1 Ground survey

From the river bank to the lake bank, transects L2-4 are approximately 2000 m long, whereas L1 is approximately 1200 m. Three different zones occur along all the four transects. The first zone is on the river bank. Flowered *Miscanthus* of 1-2 m height appears at high elevations near the river bank, often mixed with *Cynodon*, *Carex*, *Polygonum*, and human planted poplar. Some of the flowered *Miscanthus* also has green leaves of lower height, or some shorter *Miscanthus* is not flowered, we call them green *Miscanthus* in this paper. If large number of cattle graze on *Carex*, as for example in L2, the *Miscanthus* is cut to approximately 0.5 m. The second zone is *Carex* dominant zones, which across approximately 500 m horizontal distance. The height of *Carex* ranges from 0.3 to 0.6 m, and they thrive, with very high density. As forward to the lake bank, the height and density of *Carex* are decreasing and *Polygonum*, *Artemisia* and *Eleocharis* appear and mixed with it. The third zone is near the lake bank where elevation changes gradually. The indicators of low elevation are high soil moisture and plant like *Cardamine* and young *Carex*. The farthest place we reached on the lake bank in L1 is covered by 10 cm shallow water and dead *Potamogeton* and *Vallisneria* beneath. We found birds' drops and feathers frequently and the birdcall is very clear. On the bank of L2-4, we found wet soil, *Cardamine*, dead and dry *Potamogeton* and *Vallisneria* covered on soil, and shooting up *Carex* with very low density.

Figure 2 shows dominant vegetation types and their coverage at 27 sample plots along transect L1. As illustrated in the legend, the length of the bar indicates the percent coverage of vegetation which is averaged from 5 or 8 subplots. The average height of each vegetation type can be read from the centre point of the bar according to the scale on the left. For the three zonations we categorized above, samples from 1 to 5 belong to the first zone, and samples from 20 to 27 belong to the third zone. Samples from 6 to 12 are relatively homogenous and belong to the second zone, whereas samples from 13 to 17 are in transition area. From the NDVI values extracted from HJ image at corresponding sample plots, we found the NDVI achieve the peak around 0.6 at sample 8 until sample 12, where either homogenous *Carex* plots appear or *Carex* mixed with green *Miscanthus*. As the *Carex* goes shorter forward the lake, the coverage of wet soil and dead *Potamogeton* increase and NDVI values reduce to 0.2 and remain stable.

3.2 Corresponding variable in the field

In the ground survey, we measured as many variables as we can in 1m * 1m subplots, including land cover type, vegetation species, percent coverage, height and spectral curve, to detail the field information for accuracy assessment. We compares field measured NDVI values of vegetation, to which different vegetation types, coverage and heights contribute differently. First of all, plots covered by lower percent vegetation coverage will have lower NDVI. The cases of *Carex* with coverage 100, 70, 30, 5 percents show obviously decreasing NDVI. Secondly, vegetation with different heights may have the same NDVI. For example, pure *Carex* plots with height between 0.4 and 0.6 meter are have the same NDVI value 0.89. The possible reason is *Carex* extremely thrive at that heights and with very high growing density, which cause NDVI be fully saturated. Thirdly, for cases where NDVI does not achieve saturation, e.g. four *Miscanthus* plots with 100 percent coverage, the height of the plant shows its impact on NDVI. We measured both the height of flowered part and green leaves part of *Miscanthus*. The flowered parts of *Miscanthus* are dry in autumn, thus having low NDVI around

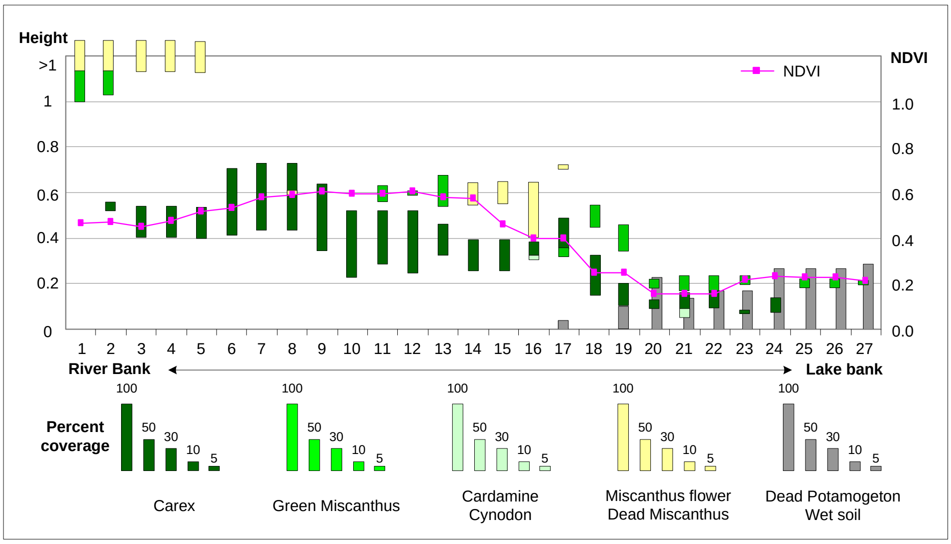


Fig. 2. Types of dominant vegetation, their percent coverage and heights along transect L1 are compared with NDVI extracted from corresponding pixels at field samples

0.25. Since the NDVI is not saturated when flowered part exist, the NDVI values reduce as heights of their green leave part decreasing. Last but not least, vegetation growing density also influences the NDVI value. For example, *Artemisia* is kind of plant which also thrive in autumn, but with very low proportion and usually mixed with *Carex*. For plots which are fully covered by *Carex* and *Artemisia* at the same height of 0.4 meter, *Artemisia* has lower NDVI of 0.79 compared with 0.89 for *Carex*. The possible reason is that coverage percent only reflects the proportion of the projected canopy of vegetation on the ground, so that *Carex* with high density has larger NDVI than *Artemisia*.

According to the results above, we found that *Carex* coverage is an outstanding variable which is closely related to NDVI, may act as the corresponding variable of covering function derived from the random set model. For the other variables, they are closely related with each other when contributing to NDVI value and can not be act as independent variable.

3.3 Extensional uncertainty modelled by random sets

Main Characteristics of random set Γ, including the covering function, the support set, the median set, the core set and the variance, were estimated. The contours of support, median illustrated in Fig. 3a indicate the possible spatial extension of this *Carex* patch with above 0.05 probabilities and above 0.5 probabilities respectively. The pixels with covering function value larger than 0.95 are enclosed by the contour of core set which almost ensure the pixel belonging to *Carex* patch. The differences between the spatial extension of support set and core set indicate the extensional uncertainty of *Carex* patch. The higher uncertainty corresponds to higher variance of random set in Fig. 3b.

3.4 Linking covering function to *Carex* coverage

By plotting the field data and estimated covering function, the values of the covering function (CF) were compared with NDVI and percent coverage of *Carex* (PCC) for all the sample plots, among which in transect L1 are shown in Fig. 4. We found that these three curves have similar trends along the transect L1. They have peak values approximately from sample 6 to 12 and low values at the two ends. Both curves of CF and NDVI have relatively gentle slope from sample 1 to 7, compared with steep slope from sample 15 to 27. But PCC decreases rapidly from sample 6 to sample 1, making a steep slope at the left end of the curve. This mismatch can be explained by the mixture of green *Miscanthus* with *Carex* (Fig. 2), which also contribute to the NDVI. At the right end of the curves, although PCC has values around 20 percent, but due to the low height and low density of young *Carex* (Fig. 2) contributing little to NDVI value, the CF shows 0 values from sample 18 to sample 27.

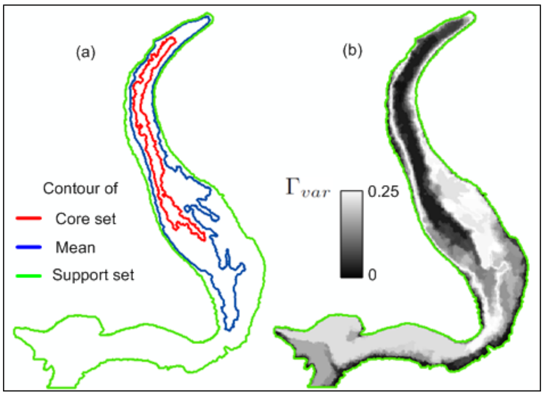


Fig. 3. Extracted object and its extensional uncertainty described by concepts from random set theory: (a) support set, median and core set (b) variance

T-test was then used to explore the relationship between median set and *Carex* coverage. The null hypothesis is that the mean value of the *Carex* coverage of samples which included by the median set is equal to the mean value of the *Carex* coverage of samples which excluded by the median set. The two-tailed p value associated with the test p = 0.000 which is smaller than 0.05, then we reject the null hypothesis. That implies that there is sufficient evidence to conclude that samples included and excluded by the median set have different *Carex* coverage.

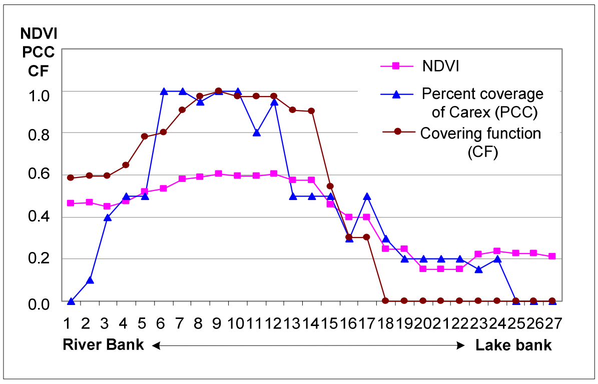


Fig. 4. Percent coverage of *Carex*, NDVI and covering function at sample plots compared along transect L1

Table 1 highlights that the correlations between covering function and *Carex* coverage were observed with the R^2-values ranging from 0.46 to 0.71, whereas correlations between covering function and NDVI have higher R^2-values ranging from 0.82 to 0.97. For the total 73 samples, although *Carex* coverage has lower R^2-value with covering function than with NDVI, but it still explains 54% of variation in covering function and the relationship is significant at 0.01 confidence level. This relationship suggests that covering function of the random set can be quantified and interpreted adequately by either NDVI from image or *Carex* coverage from the field data. The relationship between *Carex* coverage and NDVI with 0.56 R^2-value suggests that NDVI is closely related to *Carex* coverage at the time the image acquired. This result also supports our previous decision that using NDVI image as input to extract *Carex* patch.

Table 1. R^2-values of the correlation relationships between covering function (CF) and percent coverage of *Carex* (PCC) and NDVI for four transects separately and in total

Transect	L1	L2	L3	L4	L1-L4
CF-PCC	0.67	0.51	0.71	0.46	0.54
CF-NDVI	0.97	0.91	0.82	0.91	0.93
PCC-NDVI	0.63	0.62	0.63	0.48	0.56

3.5 Accuracy assessment of the extracted uncertain object

In order to validate the uncertain object, percent coverage of *Carex* recorded for each sample plot were used to group testing samples. The accuracy assessment was applied to the support set by comparing with all the samples where *Carex* appears, to the median set by samples where above 50 percent of area is dominated by *Carex*, and to the core set by samples which are 95 percent covered by *Carex*. In each error matrix, number of samples belongs to tow classes: presence of *Carex* and absence of *Carex* are indentified. Table 2 details the mapping accuracy of the support set, the median set and the core set by OA, PA, UA and kappa coefficient derived from error matrix. The highest overall accuracy was achieved by the core set with OA of 85 percent and kappa 0.54. According to (Mather 1999), the core set and the median set have moderate kappa value, whereas the support set has poor kappa value. By further looking at each class, we find that presence and absence of *Carex* has high PA and UA for support set and core set respectively, which indicate that these two classes are reliable in support set and core set respectively. Presence of *Carex* has high UA and low PA in core set, showing that there is more area of *Carex* in the field than is indicated by the core set. Absence of *Carex* has both low PA and UA in support set, because 12 out of 16 samples which are classified as absence correspond to presence in the field data. The possible explanation for the unsatisfied accuracy is that the grouping criteria of making testing samples for the support set is not appropriate. We get supporting evidences from sample 18 to sample 24 in Fig. 4. These samples have 0 covering function, and not belong to the support set, but they still have 20 percent *Carex* at low height and with low density. This result suggests that the support set is not sensitive to the *Carex* coverage lower than 20 percent.

A kappa z-test for pair-wise comparison in accuracy shows that there was significant difference between the support set and other sets, but no significant difference between the core set and the median set. The results suggest the quality of the core set and the median set is significantly higher than that of the support set.

Table 2. Comparison of OA, UA, PA and kappa coefficients for the core set, the median and the support set. C1 indicates class presence of *Carex* and C2 for class absence of *Carex*.

Core set	PA (%)	UA (%)	OA (%)	Kappa
C1	47	90	85	0.54
C2	98	84		
Median set				
C1	76	85	77	0.52
C2	78	66		
Support set				
C1	82	93	78	0.22
C2	50	25		

4. CONCLUSION AND DISCUSSION

In this research, we applied the random set model for representing uncertain boundary of a *Carex* patch, and perform accuracy assessment on the modelling results. We found that *Carex* coverage can be the corresponding variable collected on the ground, by which the covering function of random sets can be quantified and interpreted adequately. The core set of the random set has higher accuracy than the median set and the support set.

Significant correlations were found among covering function, *Carex* coverage and NDVI, which suggests that covering function of the random set can be quantified and interpreted adequately by NDVI derived from image and *Carex* coverage measured in the field. For the other variables, such as vegetation types, height and density and coverage, they also influence the modelling result, and should be considered together. These variables, however, may belong to different scales, such as nominal (e.g. vegetation type), ordinal (e.g. big or small density) and ratio scale (e.g. height and coverage). So they are difficult to integrate into one general variable which might be better match with the covering function of random set.

The quality of the random set models was assessed quantitatively by OA, UA, PA and the kappa coefficients.

The accuracy of core set is better than that of the median set and much better than that of the support set. Since the support set is not sensitive to the young *Carex* with coverage less than 20 percent, inappropriate criteria for grouping test samples might be the reason for its poor accuracy. This result suggests that the random set model has a better performance on the high coverage area and criteria for validating the support set should be determined not only based on the coverage. Moreover, it supports that *Carex* coverage cannot be the only variable fully explaining the covering function and other variables such as height should be considered especially when the coverage is low.

The accuracy of random set model applied in this study is just moderate according to the assessment report. Several reasons could contribute: on one hand, the parameters in the region growing segmentation algorithm need further adjustments. On the other hand, the field data which are often referred as the ground truth may not perfectly match with modeling results. In this study, more factors such as heights and density of vegetation should be integrated with vegetation coverage as united reference data for accuracy assessment.

ACKNOWLEDGEMENTS

This Research is funded by the 973 Program (Grant No. 2009CB723905), Sino-Germany Joint Project (Grant No. 2006DFB91920), National Key Project of Scientific and Technical Supporting Programs (2007BAC23B05) and NSFC project (Grant No. 40721001). The authors would like to express their many thanks to CRESDA for providing HJ image and Mr. Lian Feng, Mr. Xiaobing, Cai and Mr. Shoujing, Yi for their assistants in the ground survey.

REFERENCES

Congalton, R. G., R. G. Oderwald and R. A. Mead, 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering & Remote Sensing*, 49, pp. 1661-1668.

Cressie, N. A. C., 1993. Statistics for spatial data(eds.). Wiley-Interscience, pp. 725-803.

Foody, G. M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80, pp. 185-201.

Mather, P. M., 1999. Land cover classification revisited. *Adances in remote sensing and GIS analysis*. P. M. Tate and N. J. Tate (eds.). Wiley, Chichester, pp. 7-16.

Matheron, G., 1975. *Random sets and integral geometry*. New York: Wiley.

Molenaar, M., 1998. *An Introduction to the Theory of Spatial Object Modeling for GIS*. Taylor & Francis.

Nguyen, H. T., 2006. *An Introduction to Random Sets*. Chapman Hall.

Pal, N. R. and S. K. Pal, 1993. A review on image segementation techniques. *Pattern Recognition*, 26, pp. 1277-1294.

Russ, J. C., 2007. *The Image Processing Handbook*. Taylor & Francis.

Stein, A., P. Budde and M. Z. Yifru, 2009a. Stereology for Multitemporal Images with an Application to Flooding. *Research Trends in Geographic Information Science*. G. Navratil (eds.). Springer-Verlag, pp. 135-150.

Stein, A., N. A. S. Hamm and Y. Qinghua, 2009b. Handling uncertainties in image mining for remote sensing studies. *International Journal of Remote Sensing*, 30, pp. 5365-5382.

Stoyan, D. and H. Stoyan, 1994. *Fractals, Random Shapes and Point Fields*. John Wiley&Sons.

Van de Vlag, D. E. and A. Stein, 2007. Incorporating uncertainty via hierarchical classification using fuzzy decision trees. *IEEE Transactions on geoscience and remote sensing*, 45(1), pp. 237-245.

Vorob'ov, 1996. Random set models of fire spread. *Fire technology*, 32(2), pp. 137-173.

Woodcock, C. and S. Gopal, 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, 14(2), pp. 153-172.

Wu, Y. and W. Ji, 2002. *Study on Jiangxi Poyang Lake National Nature Reserve*. China Forestry Publishing House Beijing.

Zadeh, L. A., 1965. Fuzzy sets. *Information and Control*, 8, pp. 338-353.

Zhan, Q., M. Molenaar, K. Tempfli and W. Shi, 2005. Quality assessment for geo-spatial objects derived from remotely sensed data. *International Journal of Remote Sensing*, 26(14), pp. 2953-2974.

Zhao, X., A. Stein and X. Chen, 2009a. Application of random sets to model uncertainties of natural entities extracted from remote sensing images. *Stochastic Environmental Research and Risk Assessment*. 10.1007/s00477-009-0358-3

Zhao, X., A. Stein and X. Chen, 2009b. Quantification of Extensional Uncertainty of Segmented Image Objects by Random Sets. *IEEE Transactions on geoscience and remote sensing*. (under review)

STUDY ON THE DATA QUALITY MANAGEMENT AND THE DATA QUALITY CONTROL—A CASE STUDY OF THE EARTH SYSTEM SCIENCE DATA SHARING PROJECT

Chongliang Sun*, Juanle Wang

Institute of the Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101
suncl@lreis.ac.cn

Commission II, WG VI/4

KEY WORDS: Scientific data, Sharing, Data Quality, Management, Control

ABSTRACT:

With the rapid accumulation of scientific data, there has formed mass data storage, which is ascending in every minute. So it becomes a very desiderate problem to be solved that how to manage the mass data effectively and to control the data quality scientifically. This paper analyzes several current problems in the field of scientific data management based on the author's experience on the data quality management of the Earth System Science Data Sharing Project, such as (1) the data quality problem is still not solved, (2) the absence of data living period design in data quality management, (3) the data quality management strategy always not that clear, etc. And according to the data management regulation home &abroad, the paper talks about an elementary method combining with the data classification and the quality management and some solving measures with a case study on the Earth System Science Data Sharing Project and it's soil vector data with name of *the distribution map of the national soil in a scale of 1:4,000,000(1980s)*. At last, the paper takes a prospect of the data management & quality control problem.

1. INTRODUCTION

With the globalization of information and economy, scientific data has become into the vital strategic resource for the national economy and society development. And it plays a key role in the national economy, social development, national security, and the public services, etc.

Under the condition of network's rapid development, especially the embedded work on the scientific data sharing network, the scientific data has been into massive data. And how to manage these data effectively, and control the data quality efficiently to improve the level on the scientific data production, processing, storage, sharing, and use, has been the hotspot. The paper analyzes the problem based on the work experience of data management, and put up with a resolved method to provide a way for the improvement of the scientific data sharing efficiency.

2. THE CURRENT PROBLEMS

For many years, due to the uncertainty of the scientific data[1], and the traditional data management system, it result in the data resources separated management, isolated using. So the data quality evaluation system has not formed yet. Thereby, it is difficult to exchange and share data, which limit the data use abroad in the field of inter-discipline, inter-branch, inter-territory, and the inter-industry.

Scientific data quality management refers to much content, large scope, and many domains, which include data quality itself, as well as the quality problem during the data producing process.

The paper analyzes several main problems of the data quality management, based on the long term work on the data quality control. There are, (1) the data quality problem is still need to be solved, (2) the absence of data living period design in data quality management, (3) the data quality management strategy always not that clear, etc.

3. ANALYSIS OF DATA QUALITY MANAGEMENT

Data quality refers to the accordance between the data and the impersonality state which described by the data during the process of data production, processing, storage, sharing, and data use. And to the scientific data, it can be described as the matchup degree during the above processes. The scientific data quality indexes include the data authenticity, completeness, and self-consistency. The data process quality include the transfer quality, storage quality, use quality, etc[2]. For the spatial data, the quality indexes mainly consist of data completeness, data logical consistency, data position accuracy, property accuracy, temporal accuracy and some narration of the data.[3]

Data quality management materializes in the stages of data production, storage, transfer, sharing, and the use, during which the management problem could be paid more attention in practice. And the data quality management should meet the need in the following aspects, such as the data format, data completeness, and data readability, etc.

Data format: Scientific data has the special storage format, such as the vector format, grid format, and property format, etc, which can be selected by the data characteristic.

Data completeness: scientific data's completeness mainly refers to whether the data cover the range of concrete entity

* Corresponding author. E-mail: suncl@lreis.ac.cn.

entirely or not[3]. The scientific data stored by a special format should be a complete entity followed by special request.

Data veracity: the scientific data, as an integral individual, should be read by the given software, and recognize the given parameters exactly, which include the accurate position information, temporal information, and the theme information, etc. For example, the soil vector data in geo-science, should be opened by the geo-software, such as the ArcGIS software or the SuperMap software, and read the several parameters from the property sheet.

In general, according to the different data type, we could divide the quality evaluation elements into two kinds, that is the Quantitative elements, and Non-quantitative elements. And the Quantitative elements include,

Completeness: superabundance, and scarcity.

Logical consistency: concept, codomain, format, topology.

Position accuracy: absolute, relative, grid.

Temporal accuracy: temporal survey accuracy, temporal consistency, temporal correctness.

Theme accuracy: classification correctness, non-quantitative property accuracy, quantitative property accuracy.

The non-quantitative elements include the data-use object, purpose, data log, etc.

The highly efficient management of scientific data depends on the high-quality management method. In this paper, to manage the data comprehensively, we introduce the concepts of completeness, superabundance and scarcity, which differ from ISO 19113 and ISO 19114, but only used in this paper and no need to elaborate them. Based on the above cognition, we analyze the management request deeply according to the management demand, and put up with a resolving method for the quality management. That is to analyse the living period of data management, and implement them into the management the data quality.

4. THE LIVING PERIOD OF DATA FOR THE DATA QUALITY MANAGEMENT

As well as any other process, the scientific data has its own living periods, which include the production, storage, use, even to the finish. For the different period, there exists different data quality discipline, so the data quality control flow would be separated into the following parts, there are data production period, data process period, data storage period, data sharing period, and the data use period. The quality control measures should be in all of the above period. The logic frame of the data quality management could be described by the fig1.

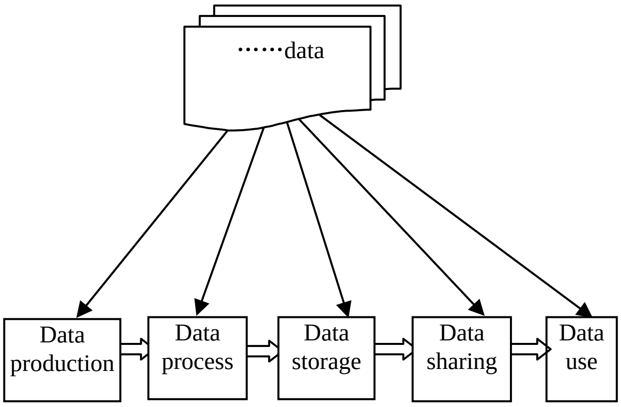


Figure 1 the logical stream of data quality management

5. THE IMPLEMENTATION OF THE DATA QUALITY MANAGEMENT STRATEGY

Given to the above analysis, the physical operation of the data quality management should include the two parts, there are the data quality evaluation method and the evaluation indexes control.

5.1 The evaluation method and flow

Due to the several types of the scientific data, and every kind has its own characteristic. So they need many evaluation methods. For the vector data, the proper method to data quality evaluation is the method based on classic mathematics. While for the property data, the right method may be the one based on the statistic mathematics, the chaos maths[4], and rough set theory[5]. But in general, the evaluation flow would be in the same control chart, and the main evaluation content consists of the followings, which are the data quality evaluation method, the evaluation operator, the evaluation date, and the evaluation result, etc.

5.2 Data quality evaluation indexes control

According to the scientific data characteristic, the data type could be divided into the vector data, grid data, property data, and the metadata during the sharing period, see fig2. So the data quality evaluation indexes should include the index of the different data types. There are the vector data evaluation index, grid data evaluation index, property data evaluation, and the metadata evaluation index, etc.

Vector/grid data indexes include the following factors, such as the storage format, project parameter, data coding, data scale, polygon closing, data document, data name criteria, data version, the last evaluation rank.

Property data evaluation indexes include the following factors, such ad the storage format, field inspection, record inspection, data document inspection, data name criteria, data version, the last evaluation rank.

Metadata evaluation indexes include the following factors, such as the metadata ID, metadata name criteria, the abstract standard, the responsibility information, the key words evaluation, and the data services mode(off-line/on-line).

The data types should be as the fig 2.

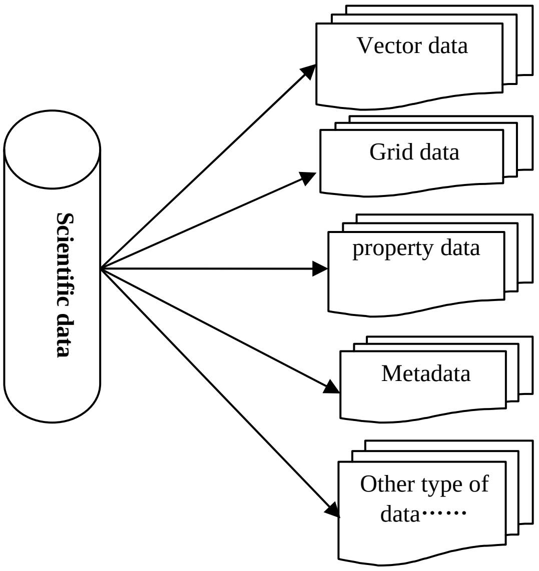


Fig 2 the classification of the scientific data

The relationship among the data evaluation indexes of different data type can be described by the fig3.

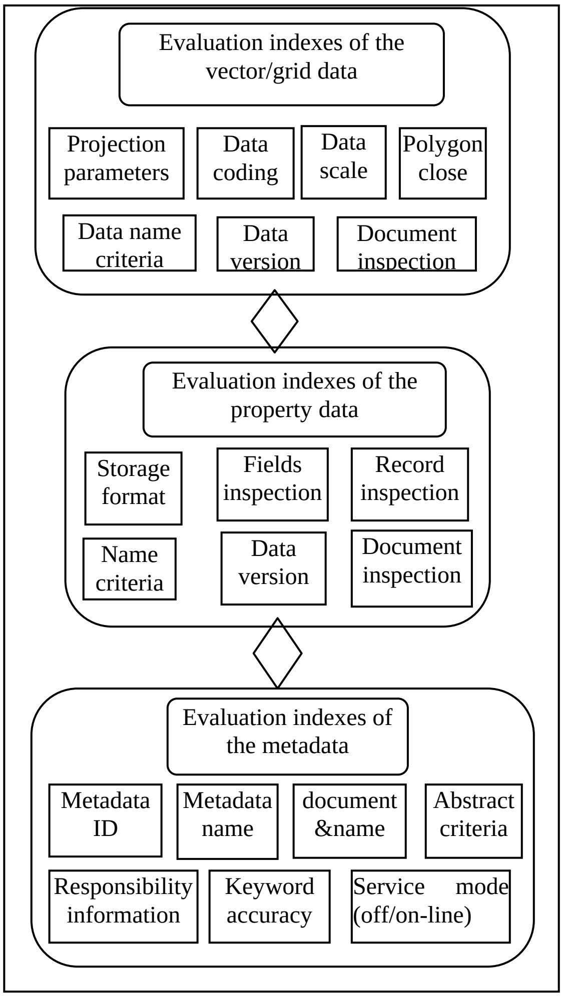


Fig 3 control index chart of the data quality evaluation

The evaluation result could be calculated by the following formula,

$$R= Value_i*Weight_i \qquad (1)$$

where $Value_i$ = evaluation value, ranging from 0-100;
$Weight_i$= evaluation weight, ranging from 0-1.

The value and weight is valued with the criterion of Data quality management guidelines from the institute of Geographic sciences and natural resources research, CAS. The result was divided into the following four classifications, as in the table 1.

Table 1 the table of evaluated value result and their ranking classification for data evaluation

value	90~100	75~89	60~74	<60
ranking	perfect	good	qualification	disqualification

6. CASE STUDY

The scientific data sharing project was put forward by the Ministry of Science and Technology(MOST), China, in 2003 to embark the data sharing work, and this project was transferred into function period in 2006. As the key one in the project, the Earth System Science Data Sharing Platform(ESSDSP) has owned more than 46000 registered users, and more than 24TB scientific data. For so large amount of data, how to manage them efficiently is a crucial problem in the construction of the platform, especially for the data management and etc.

On the above method this paper talk about, the ESSDSP data also divides into the following types, such as the vector data, grid data, property data, and the metadata, etc. According to the respective evaluation indexes, we select a soil vector data for a case study, and the data name is *the distribution map of the national soil in a scale of 1:4,000,000(1980s).* And table 2 is the soil data evaluation index value and the weight.

Table 2 the table of index value and the weight for vector data evaluation

Index	Projection parameter	Data coding	Data scale	Polygon close
value	100	100	100	98
Weight	0.3	0.15	0.1	0.2
Index	Document inspection	Name criteria	Data version	Evaluation ranking(R)
value	80	70	100	94.6
Weight	0.1	0.1	0.05	

With the above formula 1, this soil data evaluation result R is 94.6. So this soil vector data belongs to the perfect classification according to table1, and its quality control is also perfect.

7. CONCLUSIONS AND PERSPECTIVE

From this study, we can see in the current cognition level, this data evaluation method is properly for the data management, and could deal the data efficiently, which make a favourable base for the scientific data sharing project.

The future work would also be focused on the study of data management methods. And the conjunction of this method and the other method is also important for us.

References from Books:

Shi Wenzhong, 2005. *Principle of Modelling Uncertainties in*

Spatial Data and Analysis. The Science Press, Beijing, pp. 3-7.

Awrejcewicz J, 1989. Bifurcation and chaos in simple dynamical systems. Singapore: World Scientific.

Pawlak Z, 1982. Rough Sets. International Journal of Computer and Information Sciences, 11(5):341-356.

References from Other Literature:

The scientific data management group of the MOST, 2006, the data quality management discipline of the Scientific data sharing projects. Beijing, China.

The institute of Geographic sciences and natural resources research, CAS, 2009, Data quality management guidelines. Beijing, China.

References from websites:

China KDD, "About the data quality." http://www.dmresearch.net/research/shujucangku/2008/0805/124049_2.html (accessed 6 Nov. 2009)
Baidu. "data quality control of the spatial data". http://baike.baidu.com/view/125958.htm (accessed 28 Sep. 2009)

Acknowledgements and Appendix (optional)

The paper was funded by earth system science data sharing program(2005DKA32300); the National Natural Science Foundation of China(40801180、40771146); Open Research Funding Program of LGISEM; the National 863 plans projects(2006AA01A120); the forward project for youth researcher of igsnrr: study on the intelligent search for geo-metadata based on geographic ontology.

RESEARCH ON VISUALIZED DATA QUALITY CONTROL METHODS OF GROUND OBJECT SPECTRUM IN YANZHOU MINING AREA

Jun-fu Fan [a, *], Min Ji [a], Ting Li [a], Zhuo Li [a]

[a] Geomatics College of Shandong University of Science and Technology, 579 Qianwangang Road Economic & Technical Development Zone, Qingdao, China, 266510

KEY WORDS: Ground Object Spectrum, Data Quality Control, Cluster Analysis, Box-and-Whisker Plots, Gross Error Detection

ABSTRACT:

Errors or outliers are prone to be made on account of various accidental factors or system errors in the observation process of ground object spectrums. It is necessary to carry on some rigorous gross error detection and quality control measures on field spectroscopy data before which is conducted to further spectral analysis. To this end, in this paper, in accordance with measured data of several typical crops in Yanzhou mining area, a theory of cluster analysis for field spectroscopy data quality controlling was proposed and 4 different cluster methods included Statistical distance, Aitchison distance, Pearson's correlation coefficient and Multidimensional Vector Cosine were used in the gross error visualized detection. For the common characteristic bands of different spectrum data, the goal of visualized detection and identification of outliers was achieved by means of the statistical method of box-and-whisker plots. Outliers which were identified can be getting rid of in the use of several self-developed graphic interactive controls based on GDI+ technology. The theory proposed in this paper provided effective quality assurance for in-depth spectroscopy analysis.

1. INTRODUCTION

Study on spectral characteristics of ground objects is an important part of modern remote sensing technology. It is not only the accordance of sensor design and band selection, but also the basis for interpretation of remote sensing data analysis (Edward J. Milton, et al., 2009). The accuracy of field spectral measure results is affected by many factors, such as measure time, instrument FOV, observation geometry, solar azimuth and altitude angle, atmospheric environmental factor, etc (He Ting, et al., 2003). Therefore, the raw data of field spectral observations need to go through rigorous data quality control process to identify and get rid of records that contain errors or outliers before the whole dataset are used for in-depth spectral analysis. This paper on 2 groups of spectral data got from Yanzhou mining area as examples, Self-consistent accuracy (SCA) calculation method is used to evaluate the stability of the spectrum instrument, a theory that cluster analysis for field spectroscopy data quality controlling is proposed and 4 different cluster methods include Statistical distance, Aitchison distance, Pearson's correlation coefficient and Multidimensional Vector Cosine are used for gross error visualized detection in a same batch of spectral dataset and the pros and cons of them are discussed too. For some characteristic bands of spectral data we implemented the detection and identification of errors and outliers in a visual way by using the box-and-whisker plots. The GDI+ technology is used to draw plots automatically based on the results of 4 different cluster analysis methods and box-and-whisker plots models. The goal of visualized gross error detection and outlier identification on field spectroscopy data was achieved and thus provided effective quality assurance for in-depth spectroscopy analysis.

2. SPECTRAL CHARACTERISTICS AND QUALITY CONTROL METHODS

The spectral data of ground objects got by same spectral instruments in the same conditions ought to have same or similar characteristics, such as spectral resolution, band width and curve shape features, which can be used as the basis for the classification and anomaly detection. Only when the instruments are in a relatively stable state, do the measured data have the availability. The goal of evaluating the stability of spectral instruments can be reached by calculating self-consistent accuracy (SCA) of the data.

Clustering usually refers to grouping data or objects into a number of classes or clusters. Data records or objects in the same cluster have high similarity and different data records or objects in different clusters low (Kaufman, L., et al., 1990). The goal of cluster analysis is collecting data on the basis of similarity to classification, and can identify the one contains large differences in a group of similar dataset. Therefore, for data records which might contain gross errors and outliers in a batch of spectral datasets got in a same period of time can be found by using cluster analysis methods. The clustering results do not contain any detail information about data errors and outliers and some other measures are needed for further inspection and viewing. As a statistical method which have the characteristic of robustness of the median and quartile, box-and-whisker plots can provide detail information on changes in data range and extreme values (Wang Jian, et al., 2002). Whether data with full or specify bands scope can use box-and-whisker plots to find outliers with detail information and view the whole comparison of data records.

We evaluated the stability of spectral instruments by calculating self-consistent accuracy (SCA). Cluster analysis methods were used to detect gross errors and outliers. Box-and-whisker plots were used as further measure to compare the data records to troubleshoot out with errors and extreme values. The results of

[*] Corresponding author: Jun-fu Fan; E-mail address: yeahgis@yeah.net.

cluster analysis and box-and-whisker plots were visualized by self developed controls based on GDI+ technology.

3. ALGORITHM PRINCIPLE AND APPLICABILITY ANALYSIS

3.1 Methods on Spectral Instruments Stability Testing

To check the stability of spectral instruments, a certain number of repeated observations in the same conditions should be taken. The repeat measured data records can be tested and checked by calculating the mean square error of them to assess the accuracy and stability of the spectral instruments.

$$\varepsilon_j = \pm\sqrt{\frac{\sum_{i=1}^{n}\delta_{ij}^2}{n}}, (j = 1,2,\cdots,m) \tag{1}$$

$$\delta_{ij} = x_{ij} - x_i, (i = 1,2,\cdots,n; j = 1,2,\cdots,m) \tag{2}$$

$$x_i = \frac{\sum_{j=1}^{m} x_{ij}}{m}, (i = 1,2,\cdots,n) \tag{3}$$

$$\varepsilon = \pm\sqrt{\frac{\sum_{j=1}^{m}(\sum_{i=1}^{n}\delta_{ij}^2)}{m\times n}} \tag{4}$$

In Eq. (1-4), *i* is band number, *j* is data record (curve) number, *m* is the count of data records (curves), *n* is the count of bands. ε_j is the mean square error of curve *j*, x_{ij} is the reflectance value of curve *j* on band *i*, x_i is the average reflectance value of all curves on band *i*, δ_{ij} is the difference of x_{ij} and $x_{j,}$ and ε is the total self-consistent accuracy (T-SCA) of all curves. The results of Eq. (3) and Eq. (4) got from raw spectral data can be used as indicators of the stability of instruments. Similar curves have similar values and the smaller the better.

3.2 Cluster Analysis Methods on Gross Error Detection

There are 2 similarity measurements between the observational data records which are processed by number normalization, the distance and similarity coefficient (Yu Xiu-lin, et al., 2002). Supposed that, within the selected feature bands range there are *p* sampling points of a reflectivity curve. According to distance-based methods, curves can be regarded as points in a *p*-dimensional space, and the distance is defined in the space, points with short distance in a certain range fall into same classes and the ones with long distances fall into different classes. The methods based on similarity coefficient get the categories by calculating the similarity coefficient between data records (curves). The closer the absolute value of a similarity coefficient to 1 between curves, the more similar of them. Similarity coefficient must close to 0 if they are different with each other.

3.2.1 Classification Based on Distance Measurement: For a given spectral instrument, if the selected characteristic bands scope contains *p* sampling points, reflectance curves can be seen as a series of points in *p*-dimensional space. This space is a simplex space of non-Euclidean space, but also can be approximated as Euclidean space. Then the degree of similarity between two observational data records (curves) can be measured by the distance between the two points in the *p*-dimensional space. Two methods as below are used to calculate the distance in this paper.

a) Statistical Distance

The Euclidean distance equation as below:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{5}$$

where
p = sampling points count
i, j = Number of 2 data records (curves)
x = Coordinates in *p*-dimensional space
d_{ij} = Euclidean distance

Euclidean distance is a commonly used method in cluster analysis, but there are some shortcomings of its own. One is that Euclidean distance is related to the dimensions count of the statistic index, but there is no such problem for spectral reflectance curves. The other one is that the Euclidean distance does not take into account of the correlation between the various indicators. Some effective approaches must be taken to revise this problem (Yu Xiu-lin, et al. 2002). We introduced the method of weighted index variance to achieve the purpose. Eq. (6) is the improvement of Eq. (5).

$$\begin{gathered} d^s{}_{ij} = \sqrt{\sum_{k=1}^{p}\frac{(x_{ik} - x_{jk})^2}{s_{kk}}}, (k = 1,2,\cdots,p) \\ s_{kk} = \frac{1}{n}\left[(x_{1k} - \bar{x}_k)^2 + \cdots + (x_{nk} - \bar{x}_k)^2\right] \end{gathered} \tag{6}$$

where
S_{kk} = Variance of the No. *k* index
n = Spectral reflectance curves count
x = reflectance values
$d^s{}_{ij}$ = The statistical distance between curve *i* and *j* in *p*-dimensional space

The $d^s{}_{ij}$ equals to Euclidean distance when $S_{11}=S_{22}=\cdots=S_{kk}=\cdots=S_{pp}$. Euclidean distance can only be used in the situation of the indicators have the same deviation and equal contribution to the distance, variance weighted statistical distance is not subjected to this restriction and the results of practical application are better.

b) Aitchison Distance

Aitchison Distance is a calculation method which is used to measure the distance between objects defined in simplex space. A very good natural example is the distribution of probabilities

$P_1+\cdots+P_d=1$ for an event with *d* possible outcomes (Vêncio RZ, et al, 2005). The vector cluster, having *p* continuous sample points formed by *n* spectral reflectance curves which are observed in the same situations, can be seen as *n* points in non-normalized *p*-dimensional simplex space. To measure physical distance between objects in astronomical scales one should not use the regular Euclidean distance as shown in Eq. (5) but rather use proper relativistic distance measurements. This complication aroused because our world is not a Euclidean world. It is meaningless to calculate distance between 2 objects using the Euclidean distance model in a simplex and non-Euclidean space. John Aitchison proposed a meaningful calculation model to measure the distance between objects, such as *u* and *U*, in the simplex space (Aitchison, J., 1986).

$$d_{uU} = \sqrt{\sum_{j=2}^{p}\sum_{i=1}^{j-1}\left[\ln(\frac{u_i}{u_j}) - \ln(\frac{U_i}{U_j})\right]^2} \tag{7}$$

where d_{uU} = The Aitchison distance between *u* and *U*
i, j = Number of sample points (bands' No.)
p = Count of sampling points
$u_{i/j}$, $U_{i/j}$= The reflectance value of curve *u* & *U*

Aitchison distance can be converted to the equivalent form of Euclidean distance by the transformation from simplex space to Euclidean space, the transformation methods may include stretching, expanding, or inflation. Aitchison distance is a complexity theory dealing with distance problems in simplex spaces. It provides a method for distance measuring and reasonable classification basis derived from series of mixed data in cluster analysis applications (Aitchison, J., 2001). The advantage of Aitchison distance are higher reliability than statistical distance because the former can show the topological space relationships between objects based on the method of algebraic topology. The disadvantage of it is the lower computing efficiency than other distance methods. We used the Aitchison distance to cluster analysis for gross error detection and obtained satisfactory results.

3.2.2 Classification Based on Similarity coefficient: Similarity coefficient describes the similarity degree between samples. We used two different similarity coefficient methods included multidimensional vector cosine and Pearson's correlation coefficient to calculate the Similarity coefficient.

a) Multidimensional Vector Cosine

Cosine of the angle between multidimensional vectors is inspired by similar figures (Yu Xiu-lin, et al. 2002). For a spectral reflectance curves cluster contains *n* curves, if the length of the curves is not the major object of study, multidimensional vector cosine represents a kind of similarity coefficient which can show the corresponding similarity on the shape between the curves involved.

$$\cos\theta_{ij} = \frac{\sum_{\alpha=1}^{p} x_{i\alpha}x_{j\alpha}}{\sqrt{\sum_{\alpha=1}^{p} x_{i\alpha}^2 \cdot \sum_{\alpha=1}^{p} x_{j\alpha}^2}}, (0 \le \cos\theta_{ij} \le 1)$$

(8)

where $\cos\theta_{i,j}$ = The cosine value between vector *i* and *j*
i, j = Number of vectors (spectral curves)
p = Count of sampling points (vector dimensions)
x = The reflectance value

In normal circumstances, spectral curves have the same sampling points in the same scope of bands and as a result, vectors based on the curves are in the space with same dimension. But if spectral instruments with different bands width or spectral resolution, spectral curves got by the instruments may have different count of sampling points and the vectors based on these curves have different dimensions, distance-based methods can not be used for the classification of such spectral curves. In this case, vectors with fewer dimensions ought to be interpolated to add dimensions as well as sampling points, or the ones with more dimensions should discard some to make all vectors having the same dimensions. The interpolation or discard process can reduce the impact given by raw data on the classification results, one serious problem is that the process on raw data may make errors or outliers in raw data amplified or neglected. The effect of this method on the gross error detection and outlier identification is not as good as that of the methods based on distance but we kept and used it as an ancillary method for cluster analysis in this paper.

b) Pearson's Correlation Coefficient

Pearson's correlation coefficient is known as the best method of measuring the correlation, because it is based on the method of covariance (Li Xi-qiang, et al., 2008). It gives information about the degree of correlation as well as the direction of the correlation. Pearson's correlation coefficient, *r* in Eq. (9), is a covariance-based theory about correlation measurement. It not only gives the correlation between samples, but also shows the direction relevance. Different classes can be divided by comparing a series of similar spectral curves' correlation coefficient and in turn data records or curves with gross errors or outliers can be identified.

$$r = \frac{\sum_{\alpha=1}^{p} x_{i\alpha}x_{j\alpha} - \frac{1}{p}\sum_{\alpha=1}^{p} x_{i\alpha} \cdot \sum_{\alpha=1}^{p} x_{j\alpha}}{\sqrt{\Delta}}$$

$$\Delta = \left[\sum_{\alpha=1}^{p} x_{i\alpha}^2 - \frac{\left(\sum_{\alpha=1}^{p} x_{i\alpha}\right)^2}{p}\right]\left[\sum_{\alpha=1}^{p} x_{j\alpha}^2 - \frac{\left(\sum_{\alpha=1}^{p} x_{j\alpha}\right)^2}{p}\right] \tag{9}$$

where i, j = Number of curves
p = Count of sampling points
x = The reflectance value

3.3 Box-and-Whisker Plots for Error and Outlier Viewing

Box-and-whisker plots, also known as schematic plots, can provide the information about variation range and extreme values of data (Wang Jian, et al., 2002). This statistic method is used for gross error detection and outlier identification because the significant character of robustness of the median and quartile. The box of a box-and-whisker plot represents the 50% values in the most middle of a data record. The upper and bottom edge of the box represent the value at 75% and 25% location of a data record which sorted from small to large, called the upper and lower quartile. The values of upper and bottom whisker are the max and min values which are smaller than max outlier limit and larger than min outlier limit in a data record. The max outlier limit is the value of upper quartile plus the interquartile range or *IQR*, similarly, the min outlier limit is the value of lower quartile minus *IQR*. The value of *IQR* is the absolute value of difference between the upper and lower quartile. All values larger than the max outlier limit or smaller than the min outlier limit are regarded as outliers (D.L. Massart, et al., 2005).

$$IQR = \left|Q_3 - Q_2\right|$$
$$E_{\max} = Q_3 + 1.5 \times IQR;\quad E_{\min} = Q_2 - 1.5 \times IQR \tag{10}$$

where IQR = The interquartile range
Q_2 / Q_3 = The lower/upper quartile
E_{max}/E_{min} = The max/min outlier limit

3.4 Pros and Cons

Cluster methods can find out the one which contain errors or outliers in a series of similar spectral curves. Practical applications on spectral data found that the effect of distance-based methods is superior to the methods based on the similarity coefficient. The Aitchison distance is more excellent than other methods on the detection results, but it is on the cost of calculation efficiency.

4. APPLICATION EXAMPLES AND VISUALIZATION

We select field spectral data got by a series of parallel experiments of 2 kinds of crops in Yanzhou mining area, winter wheat (*95128*, 4-April-2009, 7 curves, coded as *95128-1~7*) and single cropping rice (*Xiushui 110*, 13-September-2009, 9 curves, coded as *A002~9* and *A110*) as our experimental data. The results of classification, data plots and box-and-whisker plots are drawn on a self-developed graphic control based on GDI+ technology using C# language.

There is one data record contains measurement noises, the codes of them were *95128-1* and *A110*, in each of the 2 data groups. The spectral curve plot of each data record is shown in Figure 1.

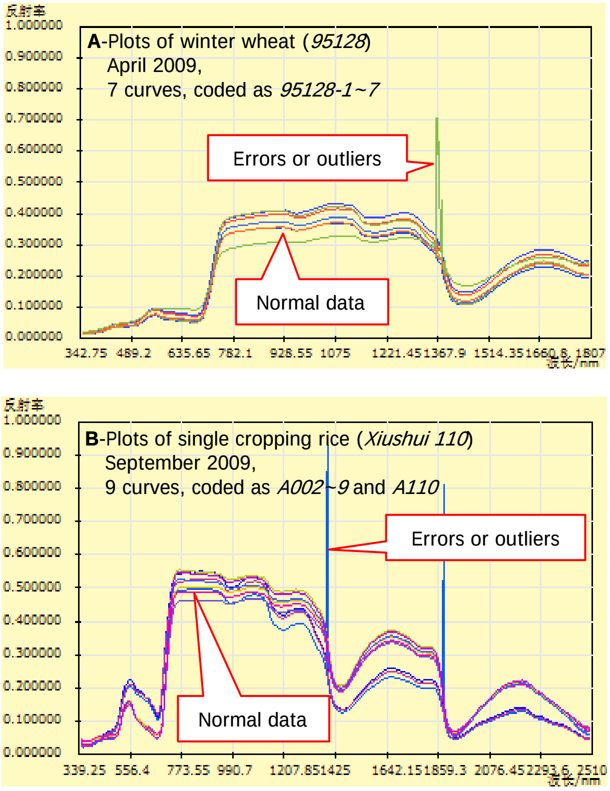


Figure 1. Plots of the experimental data

We calculated the self-consistent accuracy (SCA) of the 2 groups of spectral curves and the results are shown in table 1 and table 2.

Curve No. of winter wheat	Calculate start wavelength (nm)	Calculate end wavelength (nm)	SCA
95128-1	350	1800	±0.0446
95128-2	350	1800	±0.0194
95128-3	350	1800	±0.0144
95128-4	350	1800	±0.0104
95128-5	350	1800	±0.0297
95128-6	350	1800	±0.0212
95128-7	350	1800	±0.0176
T-SCA			±0.6344

Table 1. SCA results of winter wheat (*95128*)

Curve No. of single cropping rice	Calculate start wavelength (nm)	Calculate end wavelength (nm)	SCA
A002	1067	2189	±0.0258
A003	1067	2189	±0.0291
A004	1067	2189	±0.0178
A005	1067	2189	±0.0127
A006	1067	2189	±0.0169
A007	1067	2189	±0.0138

A008	1067	2189	±0.0342
A009	1067	2189	±0.0306
A110	1067	2189	±0.0503
T-SCA			±0.6665

Table 2. SCA results of single cropping rice (*Xiushui 110*)

In table 1 and table 2, each of the SCA value of the spectral curves is small and it represented that the spectral instruments in a stable state. Even though, we can see that the SCA values of *95128-1* and *A110* are relatively greater than others in their groups, this shows that there may be errors or outliers in the data of the curves.

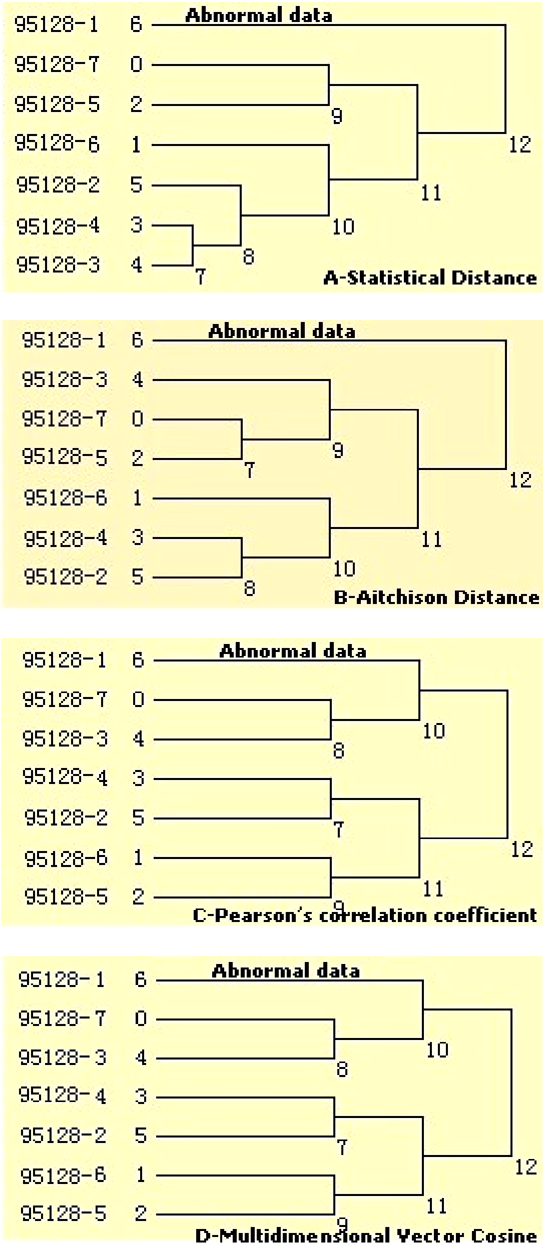


Figure 2. Cluster analysis plots of winter wheat (*95128*)

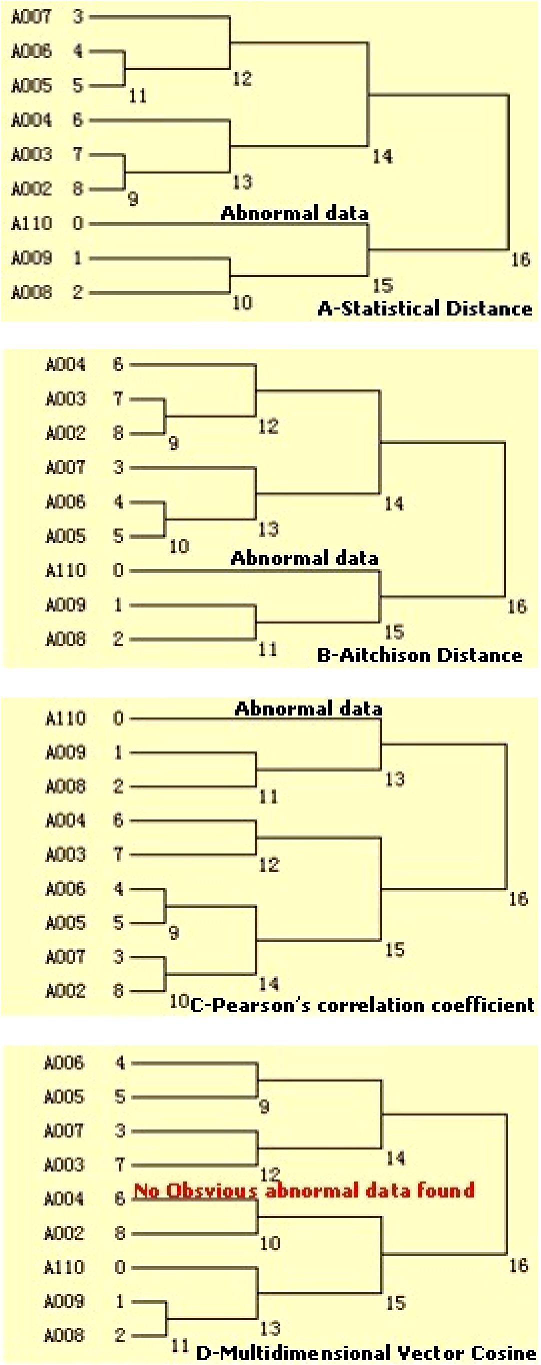


Figure 3. Cluster analysis plots of single cropping rice (*Xiushui 110*)

As shown in figure 2 and figure 3, spectral curves with errors or outliers can be identified by cluster analysis methods. Compared with spectral classification methods based on similarity coefficient, distance-based methods are more sensitive to abnormal data and give better results. If there are a finite number of sharp noise points in a spectral record (curve), the classification method based on multidimensional vector cosine may be not able to find out the abnormal curve contains errors or outliers. But this method can be used at the fuzzy classification analysis of ground object spectrum. Once the abnormal data record is identified in a series of similar data records, the detail information about the outliers or errors can be shown by box-and-whisker plots as in the figure 4.

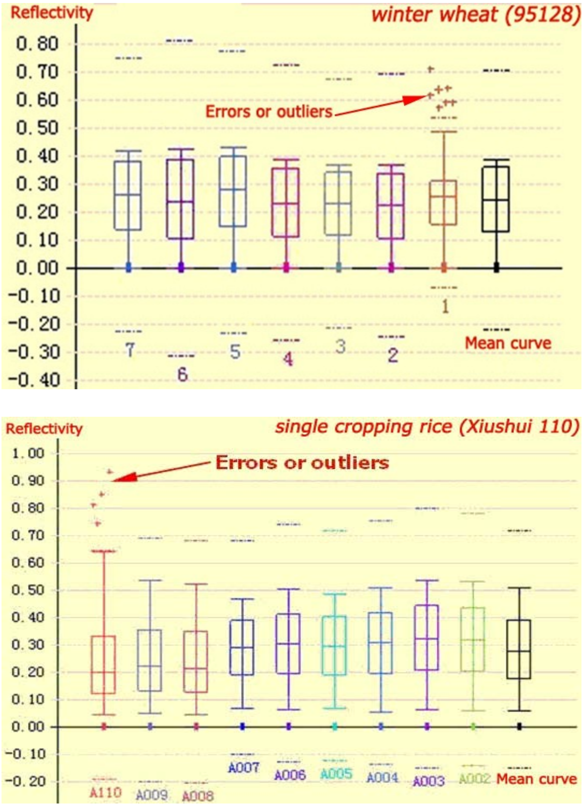


Figure 4. Box-and-whisker plots of the experimental data

Errors and outliers in a data record can be found in the box-and-whisker plots as shown in figure 4. There are several sharp noise points in the *95128-1* and *A110* spectral curve. The abnormal data records must be discarded or take some smoothing measures before being used for in-depth spectral analysis.

5. CONCLUSION

The precision of field spectroscopy data is affected by various factors and it is difficult to give out the priori statistics information of them. It is a venture that adopts some traditional gross error detection methods blindly. The process of studying on the statistical properties of research data according to the actual situation before appropriate methods are selected for data quality controlling is considered necessary and essential. This paper on 2 groups of field spectroscopy data, one was winter wheat and the other was single cropping rice. We used self-consistent accuracy (SCA) model to evaluate the stability of the spectrum instrument, proposed the theory that cluster analysis can be used for field spectroscopy data quality controlling and 4 different cluster methods were used for gross error visualized detection. For the characteristic bands of spectral data we implemented gross error detection and outlier identification in a visual way by the use of box-and-whisker plots. The GDI+ technology is used to draw plots automatically based on the calculate results of 4 different cluster analysis and box-and-whisker plots models. The practical application results show that abnormal data can be identified and removed before in-depth analysis is taken on. The goal of visualized gross error detection and outlier identification for field spectroscopy data got in Yanzhou mining area is achieved and thus provide quality assurance on spectrum data for in-depth spectroscopy analysis.

REFERENCES

Aitchison, J., 1986. *The Statistical Analysis of Compositional Data*. Wiley, NY.

Aitchison, J., 2001. *Algebraic Methods in Statistics and Probability: Contemporary Mathematics Series.* American Mathematical Society, Providence, Rhode Island, pp. 1-22.

D.L. Massart, J. Smeyers-Verbeke, X. Capron, et al., 2005. Visual Presentation of Data by Means of Box Plots. *LCGC Europe*, 18(4), pp. 215-218.

Edward J. Milton, Michael E. Schaepman, Karen Anderson, et al., 2009. Progress in field spectroscopy. *Remote Sensing of Environment*, 113(S1), pp. S92-S109.

He Ting, Liu Rong, Wang Jing., 2003. The Influences Factors on Field Spectrometry. *Geography and Geo-Information Science*, 19(5), pp. 6-10.

Kaufman, L. & Rousseeuw, P.J., 1990. *Finding Groups in Data*. Wiley, NY.

Li Xi-qiang, Wang Di, Lu She-ming, et al., 2008. Study of Fingerprint Spectra of Tobacco Flavor with Pearsonion Correlation Coefficient and the UPLC. *FINE CHEMICALS*, 25(5), pp. 475-478.

Vêncio RZ, Varuzza L, de B Pereira CA, et al., 2007. Appendix - Simcluster: clustering enumeration gene expression data on the simplex space, London, United Kingdom. http://www.biomedcentral.com/content/supplementary/1471-2105-8-246-s1.pdf (accessed 18 Aug. 2009)

Wang Jian & Jin Feng-xiang., 2002. Box-and-Whisker Plots and Correlation Model Method for Data Quality Control. *Journal of Shandong University of Science and Technology (Natural Science)*, 21(2), pp. 55-58.

Yu Xiu-lin & Ren Xue-song., 2002. *Multivariate Statistical Analysis.* China Statistic Press, Beijing, pp. 61-69.

APPENDIX

This article was sponsored by Chinese National Programs for High Technology Research and Development (2009AA12Z147); the Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping (2009B14); the Key Laboratory of Advanced Engineering Surveying of SBSM (TJES0805).

A HIERARCHICAL QUALITY-DEPENDENT APPROACH TOWARD ESTABLISHING A SEAMLESS NATIONWIDE TOPOGRAPHIC DATABASE

S. Dalyot, A. Gershkovich, Y. Doytsher

Mapping and Geo-Information Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
Tel. (972)48292361, Fax (972)48295708
(dalyot, doytsher)@technion.ac.il, sarielg@t2.technion.ac.il

KEY WORDS: Spatio-temporal data modeling and integration, Spatial data quality and certainty, ICP, Nationwide topographic database, DTM

ABSTRACT:

Nationwide geospatial databases in general and topographic ones in particular are today one of the most common infrastructure for mapping and other geo-related tasks. These databases are designated to establish an adequate, continuous and if possible homogeneous representation of our natural environment. New data acquisition technologies, which present high accuracies and resolution levels that were not known until recently, yield rapid and frequent updating of existing nationwide databases. This enables the generation of a multi-source mosaiced database that is multi-quality as well, i.e., introducing varied accuracies within its coverage area. Simultaneous analysis, such as integration, of two or more of these nationwide databases will evidently present multi-scale spatial inconsistencies. These are a function of various factors, among them the different levels of accuracy within each database. Common height integration mechanisms will not suffice here. This paper presents a framework for dealing with the problems and considerations in utilizing topographic databases that are quality derived while trying to give a solution to the existing geometric ambiguities. A conceptual new algorithmic approach is detailed, which relies on a hierarchical modeling mechanism that is designated for extracting the existing varied-scale discrepancies in order to produce a common geospatial framework. Moreover, designated quality-derived constraints are implemented in the process to ensure that accuracy is preserved. This novel approach proved to be accurate while producing seamless topographic database that retained the level of detailing and accuracies presented in the source databases, as well as local trends and morphology.

1. INTRODUCTION

The emergence of nationwide geospatial databases is an evident progression. These seamless databases, such as Orthophoto layers or varied scale Digital Terrain Models (DTM), are an essential requirement for establishing an efficient and computerized management of our environment. The assumption is that they constitute a unique, constant, uniform, reliable, seamless and - as much as possible - homogeneous mapping and Geographic Information (GI) infrastructure (National Research Council, 1990). As such, these databases serve as basis for a wide variety of research and analyses capabilities, as well as many commercial applications. Many national mapping agencies, as well as private companies and public agencies, are involved today in establishing this type of infrastructure (Parry and Perkins, 2000). Forming a reliable nationwide geospatial databases is a growing need, mainly in developing regions.

DTM databases provide up-to-date and detailed representation of the topographical variations in the earth's surface. Until recently, these databases were produced via traditional technologies and techniques, such as photogrammetric means from aerial and satellite imagery or cartographic scanning of existing analogue topographic contour maps. As a result, a nationwide DTM would usually show constant level-of-detailing (LOD), resolution and accuracy. However, local successive updates of the DTM might result in damaging its homogeneous structure. Furthermore, new data acquisition technologies, such as Airborne Laser Scanning (ALS) systems or Interferometric Synthetic Aperture Radar (IfSAR), present today high accuracies and resolution levels that were not known until recently. This intensifies the fact that an updating process performed on an existing DTM with new dense and accurate data will result in varied accuracies and LOD within its coverage area, i.e., loosing the database's homogenous nature (Hovenbitzer, 2004; Hrvatin and Perko, 2005). It can be described as if the nationwide DTM is a mosaiced database composed of patches, each acquired by a different technology, via a different technique and usually on a different period of time. Respectively, each patch presents different level of accuracy, such that an accuracy polygon map for the nationwide DTM is introduced, as depicted in Figure 1.

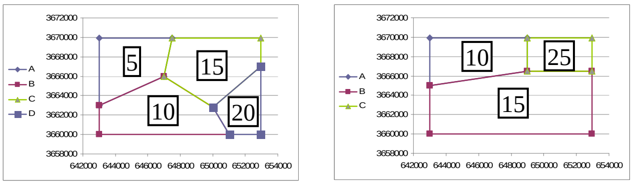


Figure 1. Scheme of two accuracy polygon maps of two nationwide DTMs; accuracy value is depicted in meters

Utilizing simultaneously several nationwide seamless DTMs for various mapping and GI applications requires the existence of continuous and contiguous terrain relief models. For example, integration is required when these models represent different zones within a larger region and a continuous nationwide terrain relief representation is required. Furthermore, for applications, such as line of sight, visibility maps, Orthophoto production - to name a few, utilizing models that are discontinuous will eventually lead to incorrect outcome. Inconsistencies between the databases are a function of various factors, such as production techniques, time of data-acquisition, LOD, datum framework - to name a few (Lee and Chu, 1996; Wang and Wade, 2008). This reflects semantically on the representation and position of the databases' described entities, thus geometric discrepancies are evident. These discrepancies affect data-certainty, for example when morphologic comparison or change detection process is at hand. Though the utilized databases for the geo-related task are geographically registered to a certain coordinate reference system, i.e., geo-referenced, these factors lead to the presence of global-systematic and local-random errors (Hutchinson and Gallant, 2000). It is evident, then, that each nationwide DTM utilized for an analysis task may present different levels of accuracy, which generally coincide to an area produced by a certain technology and/or via a certain technique, quantified via the accuracy polygon maps. As a result, not only does ambiguity exists regarding the heights required for the geo-related analysis carried out - but also the corresponding relative accuracies needed to be utilized in that process.

Several researches were carried out to solve the framework inconsistencies as well as the data-structure and data-uncertainty problems when the task of integrating different nationwide DTMs is at hand. Still, the majority of these researches handle the DTMs data as already geo-referenced, thus dealing only with the height inconsistencies of the DTMs and its quality - and not the complete geo-spatial mutual inter-relations that exist between them (Hahn and Samadzadegan, 1999; Frederiksen et al., 2004; Podobnikar, 2005).

This paper outlines a novel framework for dealing with the problems and considerations in utilizing seamless topographic DTMs that are quality-dependent. A hierarchical modeling mechanism is generated, in which the varied-scale discrepancies are monitored in order to enable a common geospatial framework that is datum-dependent free. Moreover, designated algorithms responsible for acquiring the correct position-derived accuracy from the quality polygons that are given for each nationwide DTM are integrated into this hierarchical modeling mechanism. This is vital in order to preserve the spatially varying quality and trends exist in the different DTMs, and hence, as in the case of an integration process, achieve a uniform, free of gaps and seamless nationwide DTM. This approach becomes essential in cases where no arranged and seamless mapping is available while the integration of topographic databases from different sources is crucial.

2. ALGORITHM OUTLINE

The hierarchical integration process of two (or more) homogenous DTMs where each has a single constant accuracy was proposed in the work of Dalyot and Doytsher (2008). This research presented a hierarchical modeling and integration mechanism that utilizes complete and accurate sets of different-scale data-relations that exist within the DTMs mutual coverage area. The use of these data-relations enabled precise modeling of the DTMs, i.e., extracting a mutual reference working frame (schema). Thus, the generation of an integrated unified and seamless DTM was achieved. A short review of this mechanism and its main stages is given here:

- **Pre-integration**, i.e. global rough registration, whereas choosing a common schema (framework) of both DTMs is carried out (thus solving the datum ambiguities exist between both DTMs). This is achieved while implementing the Hausdorff distance algorithm that registers sets of selective unique homologous features (objects) exists in both DTMs' skeletal structure. The skeletal structure of each DTM is identified via a novel topographical interest point identification mechanism;
- **Local matching** that is based on geometric and morphologic schema specifications analyses. This is carried out while implementing the Iterative Closest Point (ICP) algorithm with designated constraints for nonrigid surfaces matching. This stage is essential for achieving precise reciprocal modeling framework between the two databases, i.e., localized transformation quantification;
- **Reverse engineering integration** schema, which uses the matching modeling relations evaluated in the local matching stage and the data that exists in both DTMs, i.e., enabling data fusing. Obtaining an enhanced and accurate terrain representation is achieved.

Still, the existence of accuracy polygon map for each DTM requires certain considerations within the proposed hierarchical mechanism.

2.1 Pre-Integration

Data accuracy derives the certainty of a correct positioning of a topographical interest point. Consequently, the rough estimation of a mutual global registration value of both DTMs through the Hausdorff distance algorithm takes into account this factor through a weighting process on the participating points.

2.2 Local Matching

The registration value extracted in 2.1 gives the required information regarding the 'global' reciprocal working reference frame. Thus, the implementation of an adequate autonomous ICP matching process on homologous corresponding local data frames divided from the complete mutual coverage area is feasible. An independent and separate matching of small frames is more effective in monitoring and modeling the local random incongruities and trends (and consequently prevents local minima solution). The ICP algorithm is based on coupling up pairs of counterpart points (from each DTM frame that participates in the matching process) that are considered as the nearest ones exist. Thus, the estimation of the rigid body transformation that aligns both models 'best' is attained. This 'best' transformation is applied to one model while the procedure continues iteratively until convergence is achieved.

ICP matching is accomplished via Least Squares Matching (LSM) of a goal function, which measures the squares sum of the Euclidean distances Γ between the surfaces, depicted in Equation 1.

$$\sum \| \Gamma \| = \min \tag{1}$$

Monitoring errors (Γ) is achieved by minimizing the goal function, i.e., extracting the best possible correspondence between the frames. The geometric goal function is defined by a spatial transformation model between the DTM frames, and is described by a general 6-parameter 3D similarity transformation model, depicted in Equation 2.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_f = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} + R \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}_g \tag{2}$$

where, $[x, y, z]^T_f$ denote the target DTM, $[t_x, t_y, t_z]^T$ denote the 3D translation vector, *R* denote the 3D orthogonal rotation matrix, and, $[x, y, z]^T_g$ denote the source DTM. *R* is a function of the three rotation angles $[\omega, \varphi, \kappa]$.

It is important to remember that the rotation magnitude is in respect to the center of mass of each frame. Thus, $[x, y, z]^{T\,0}_f$ and $(x, y, z)^{T\,0}_g$, which denote the center of mass for each counterpart frame, are subtracted from the original coordinates before transformation is carried out.

In order to perform Least Squares estimation, i.e., linearization, Equation 2 is expanded using the Taylor series, of which only the linear terms are retained (2nd and higher orders are omitted). Consequently, each observation formula is related to a linear combination of the 6-parameters, which basically are variables of a deterministic unknown (Besl and McKay, 1992). This model is written as a matrix notation in Equation 3.

$$-e=A\cdot x-l \tag{3}$$

where, *A* is the design matrix (derivatives of the 6 unknown parameters), *x* is the unknown 6-parameter vector $\{dt_x, dt_y, dt_z\ d\omega,\ d\varphi,\ d\kappa\}^T$, and, *l* is the discrepancy vector that is the Euclidean distance between the corresponding DTMs' elements, i.e., frames data points: *f(x, y, z) - g(x, y, z)*.

The Least Squares solution gives as the generalized Gauss-Markov model the unbiased minimum variance estimation for the parameters, as depicted in Equation 4.

$$\begin{aligned} x &= \left(A^T\cdot P\cdot A\right)^{-1}\cdot\left(A^T\cdot P\cdot l\right)\\ v &= A\cdot x-l\\ \sigma_0^2 &= \frac{\left(v^T\cdot P\cdot v\right)}{n-u} \end{aligned} \tag{4}$$

where, *x* denotes the solution vector of the 6-parameters transformation, *v* denotes the residuals vector of surface observations, σ_0^2 denotes the variance factor, *n* denotes the number of observations, and, *u* denotes the number of (unknown) transformation parameters in the model, i.e., $u=6$.

Due to the fact that both nationwide DTMs are actually nonrigid bodies, several aspects are to be considered:

- Nationwide DTMs represent different data-structures - namely LOD and resolution - implying that the existence of homologous points for the ICP process is not at all explicit;
- Nationwide DTMs that were acquired on different times (epochs) will surely represent different surface topography and morphology (either natural or artificial activities);
- Data and measurement errors can reflect on the position certainty of points in relatively large scale.

To ensure convergence of the ICP process as well as to assure that the nearest neighbor search criteria is achieved correctly and fast between two homologous local frames, three geometric constraints are implemented in the ICP process - outlined in Equations 5. These constraints verify that each of the counterpart paired-up point is the closest one exists, as well as having the same relative topography surroundings. It is worth noting that these constraints are suitable for grid-space. Though not very common, Triangulated Irregular Network (TIN) structure of topographic databases does exist (mainly in areas acquired by ALS technology). With slight modifications, these equations can fit TIN characteristics as well.

$$\begin{aligned} Z_i^g &= \frac{h_1}{D}\cdot X_i^g+\frac{h_3}{D}\cdot Y_i^g+\frac{h_4}{D^2}\cdot X_i^g\cdot Y_i^g\\ Z_i^g &= -\frac{h_4\cdot y_f^{'}}{D^2}\cdot X_i^g+\frac{h_3}{D}\cdot Y_i^g+\frac{h_4}{D^2}\cdot X_i^g\cdot Y_i^g+\left(z_f^{'}-\frac{h_3\cdot y_f^{'}}{D}\right)\\ Z_i^g &= \frac{h_1}{D}\cdot X_i^g-\frac{h_4\cdot x_f^{'}}{D^2}\cdot Y_i^g+\frac{h_4}{D^2}\cdot X_i^g\cdot Y_i^g+\left(z_f^{'}-\frac{h_1\cdot x_f^{'}}{D}\right) \end{aligned} \tag{5}$$

where h_1 to h_4 are calculated from the height of local DEM grid cell corners: Z_1 to Z_4 ($h_1=Z_1-Z_0$, $h_2=Z_2-Z_0$, $h_3=Z_3-Z_o$, $h_4=h_2-h_1-h_3$); *D* denotes the database's grid resolution; *g* and *f* denotes the source and target databases; (X^g_i, Y^g_i, Z^g_i) denotes the paired-up nearest neighbor in *g*; and, ($x_f^{'}$, $y_f^{'}$, $z_f^{'}$) denotes the transformed point from dataset *f*.

Here, the assumption that each coupled-up point in every observation equation within the matching process has a different accuracy. This can be depicted as if each counterpart point 'falls' within certain polygon in the accuracy polygon map associated with the nationwide DTM. Thus, the accuracy polygon maps of both DTMs are taken into consideration during the ICP implementation. Instead of giving each row (*i*) in the design matrix (*A*) of size (*n* by 6) the same weight (where $i\in n$), a different weight is given to each row, which is derived from the accuracy polygons each of the points falls in. For example: if point *a* from DTM *f* ($a\in f$) falls in a polygon with accuracy value of *Acc_1*, and its corresponding counterpart point *b* from DTM *g* ($b\in g$) falls in a polygon with accuracy value of *Acc_2*, then their relative weight in the matching process for that frame is derived by these accuracy values, as depicted in Equation 6 (*Acc_0* denotes the accuracy of a unit-magnitude weight). The more accurate the polygon is (smaller value of *Acc*), the higher the weight value is. Hence, more accurate coupled-up points will have higher influence on the ICP process, producing a more reliable solution that characterizes correctly the given data and its quality. Thus, a weight matrix *P* (p_{ii}) can be added to the linear approximation depicted earlier in Equation 4.

$$Weight_i=\frac{Acc\_0}{\sqrt{(Acc\_1)^2+(Acc\_2)^2}} \tag{6}$$

Each matching set includes 6-parameters transformation model that best describes the relative spatial geometry of the mutual homologous frames that were matched. Since this process yields better localized registration definition, it ensures matching continuity on the entire area (as opposed to matching the entire data in a single matching process). These registration sets can be described as elements stored in 2D matrix: each set is stored in the cell that corresponds spatially to the homologues frames it belongs to. This data structure contributes to the effectiveness of the integration process.

2.3 Integration

Integration is achieved via a "reverse engineering" mechanism that utilizes the quantified correspondence between the two nationwide DTMs. This spatial correspondence is expressed by the sets of transformation, or registration parameters, which are stored in a 'registration matrix', where the values in each cell express the modeling between two matched frames. A "reverse engineering" mechanism implies that each height in the integrated DTM is calculated independently and regardless to the other values. For each position in the integrated DTM a weighted height average is calculated based on the complete spatial relations between the DTMs (stored in the matrix) and their corresponding heights. The integrated DTM can be depicted as if it exists in the space between the two source DTMs. Thus, a two-way transformation from all nodes (planar position) of the integrated DTM to each of the source DTMs while utilizing the spatial relations is implemented. Because two heights are obtained via the process (two sources) the weight of each of the two heights is derived from the corresponding accuracy polygon it falls in, thus a weighted average process is carried out. (For further reading the reader is kindly referred to Dalyot and Doytsher, 2008).

2.4 Smoothed Polygon Map Establishment

Each DTM presents internal varying accuracies - along with existing accuracy differences among the DTMs. Still, the integrated DTM has to present seamless terrain relief, regardless of abrupt accuracy changes derived from the polygons. This is obtained via the establishment of a new "smoothed" accuracy map that is based on the data exists in the source accuracy polygon map. The "smoothed" map presents gradual accuracies change by implementing a buffer-like process around each source accuracy polygon. A schematic description of this concept is depicted in Figure 2; where there exists continuous values transition from accuracy polygon *A* (turquoise) to accuracy polygon *B* (yellow) along a buffer distance of *D*.

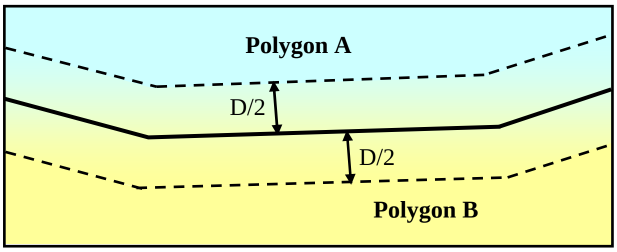


Figure 2. Schematic description of smoothing concept: gradual change from polygon *A* (turquoise) to polygon *B* (yellow). Bold line denotes the original conjoint border of two polygons

This algorithm is based mainly on the polygons' topology and their planar layout within the accuracy polygon map. The input of this algorithm is composed of polygon sets assemble each map, and their corresponding accuracy. An automatic process generates the following additional information:

- Polylines composing each polygon and their corresponding start and end point coordinates;
- Start and end points index, where:
  - 0 denotes point positioned on a map corner;
  - 1 denotes point positioned on the east/west map limits;
  - 2 denotes point positioned on the north/south map limits;
  - 3 denotes an inner point connecting two lines;
  - 4 denotes an inner point connecting three lines;
- The width of the buffer size vertical to two polygon's conjoint line; this value is derived by the difference magnitude of accuracy values of two adjacent polygons.

It is obvious that in the general case a point can connect *n* lines - and not merely three (as index 4 indicates). Still, practically this case is rare where an accuracy polygon map is at hand, so the common topologic cases are considered here. This algorithm suggests the computation of a new accuracy polygon map based on the topology and accuracies presented. New polygons are generated via the smoothing process; each holds gradually changing accuracy values. More specifically, the process generates new trapeze and triangle shaped polygons, such as the example depicted in Figure 3. An accuracy map presenting four polygons connected by two points (*11* and *22*) is transformed into a new accuracy map presenting two new triangles and five new trapezes (along with the four 'reduced' original polygons).

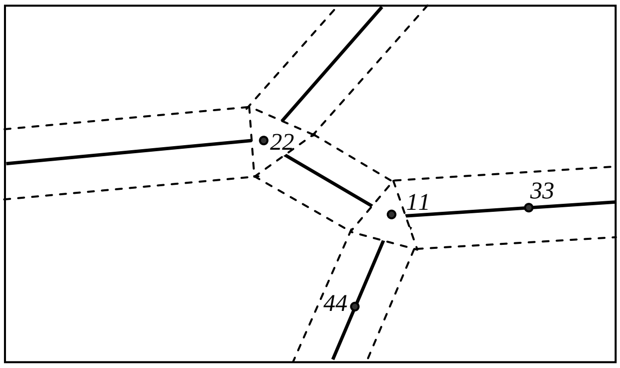


Figure 3. Schematic representation of new smoothed accuracy polygon map. Bold lines depict original polylines; dashed lines depict computed polylines (whereas the bold ones are erased in the new generated accuracy map)

Due to the fact that many possible topologies exist (and the size limit of this paper), only the most complicated one is described. Consider point *11* (indexed 4) connected to three other points: *22*, *33*, and *44* (depicted in Figure 3). For each connecting line two accuracies exist: along the left and right sides: *Acc_L* and *Acc_R*, correspondingly. For each connecting line the azimuths are calculated, as well as the azimuth values from point *11* to points *20*, *30*, and *40*, which are calculated using the buffer distance (in case the buffer size is a constant value for all polygons, these points lie on the angles' bisectors), as depicted in Figure 4. Consequently, points *20*, *30*, and *40* can be computed via geometric and trigonometric functions utilizing azimuths values, the known points' coordinates, and the given buffer distance *D*. A triangle is formed by these new points, where the accuracies corresponding to each of these points is also known.

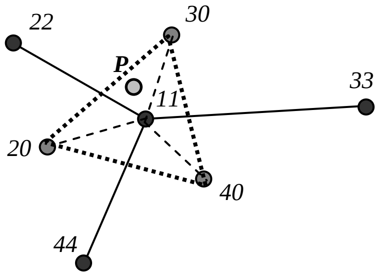


Figure 4. Formation of a triangle shaped new polygon (index 4) storing gradual accuracy values

With this, for each point within the formed triangle the accuracy calculation that corresponds to point with position value of *P* (depicted in Figure 4) is feasible. Let *P* have planar coordinates of (x_P, y_P), thus utilizing triangular coordinates can be implemented, as depicted in Equation 7. Similar process is carried out for the trapeze shaped polygons: instead of using triangular coordinates the utilization of linear transition along the buffer direction between polylines edges is implemented.

$$2S = \begin{vmatrix} 1 & 1 & 1 \\ x_{20} & x_{30} & x_{40} \\ y_{20} & y_{30} & y_{40} \end{vmatrix}$$

$$\begin{vmatrix} t_{20} \\ t_{30} \\ t_{40} \end{vmatrix} = \frac{1}{2S} \cdot \begin{vmatrix} (x_{30} \cdot y_{40} - x_{40} \cdot y_{30}) & (y_{30} - y_{40}) & (x_{40} - x_{30}) \\ (x_{40} \cdot y_{20} - x_{20} \cdot y_{40}) & (y_{40} - y_{20}) & (x_{20} - x_{40}) \\ (x_{20} \cdot y_{30} - x_{30} \cdot y_{20}) & (y_{20} - y_{30}) & (x_{30} - x_{20}) \end{vmatrix} \cdot \begin{vmatrix} 1 \\ x_P \\ y_P \end{vmatrix} \tag{7}$$

$$Acc\_P = Acc\_20 \cdot t_{20} + Acc\_30 \cdot t_{30} + Acc\_40 \cdot t_{40}$$

3. EXPERIMENTAL RESULTS

The proposed quality-dependent hierarchical mechanism was tested on several DTM databases; two of them are depicted in Figure 5. One was generated via satellite photogrammetric means (top), while the other was produced based on vectorization of 1:100,000 contour maps (bottom). Data of both databases was acquired on different times, where both cover the same area that is approximately 100 sq km.

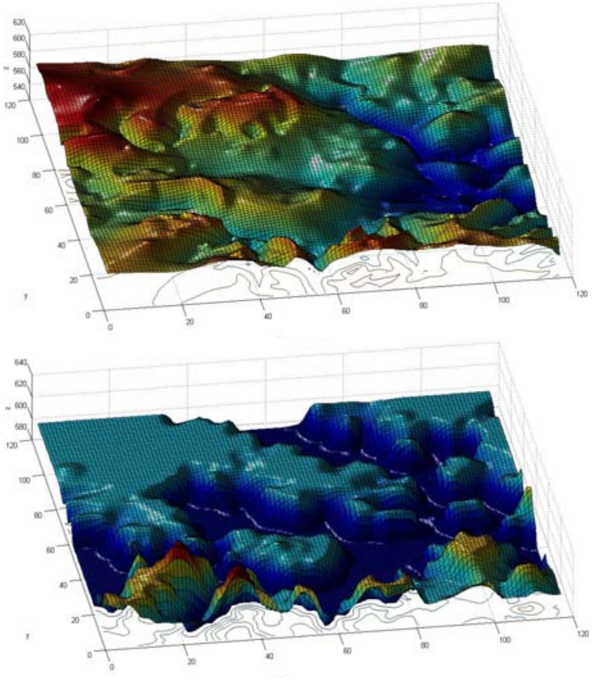


Figure 5. Two DTM databases generated via different observation technologies and on different times

Several experiments evaluating the proposed concept were carried out, of which two are presented here. On the first experiment two synthetic generalized accuracy polygon maps were produced, which showed abrupt accuracy changes and large values differences - depicted in Figure 6. This experiment is aiming to validate that inner morphology is maintained and no discontinuities exist while "moving" between neighboring accuracy polygons; accuracies chosen enabled emphasizing this.

The outcome of implementing the proposed concepts is an integrated DTM topography that is continuous with no data holes - depicted in Figure 7. Moreover, inspecting the representation closely clearly shows that the eastern area is basically a copy of the topography that exists in the 1st source DTM (as can be seen by the underlying contour lines, which are nearly the same). This is due to the fact that within this area the accuracies values are 5m from one map and 30m from the other. This translates to weighted heights average magnitude of 1:36. The western area is generated basically by an averaging process derived by the corresponding accuracy polygons that have the same weight magnitude in the process.

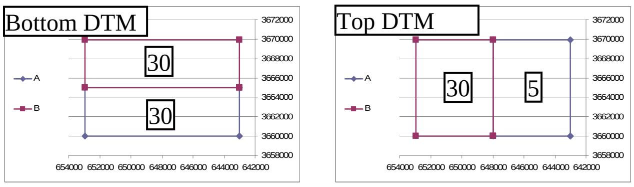


Figure 6. Generalized accuracy polygon maps; accuracy value is depicted in meters

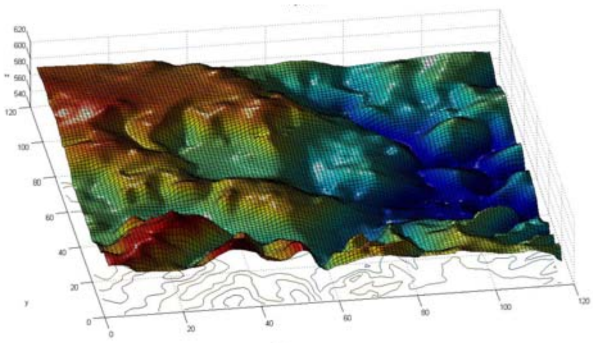


Figure 7. Integrated DTM generated in the first experiment

For the second experiment real accuracy polygon maps are utilized, describing realistic nationwide DTMs accuracies ranging between 5 - 25m, depicted in Figure 1 (left map relates to the top DTM, while the right map to the bottom DTM).

Figure 8 (left) depicts contour representation of the accuracy values of the left accuracy polygon map after the proposed smoothing process, where two triangles and five trapezes were generated. It is clear that there are no visible accuracy discontinuities - accuracy transition is constant and smooth - thus presenting a qualitative and reliable smoothing process. Figure 8 (right) depicts contour representation of the weight values used in the reverse-engineering integration process in respect to one source DTM heights. These values take into account both smoothed accuracy polygon maps generated, resulting in a continuities weighing. The contour representation resembles the geometry and topology of both accuracy polygon maps (from Figure 1), resembling a superposition of both maps, with no contour discontinuities or abrupt value changes. Consequently, the generated DTM, which is depicted in Figure 9, shows continuous and uniform topography while preserving inner and mutual morphology - as well as local trends.

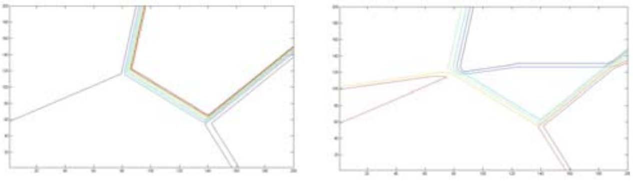


Figure 8. Contour representation of smoothed accuracy polygon map generated (left); Contour representation of the weight values used in the integration process (right)

Emphasizing the reliability of the proposed mechanism, a DTM was generated that is the outcome of the straight-forward height averaging integration mechanism - depicted in Figure 10. The height averaging integration mechanism, which utilizes the source accuracy maps, shows abrupt topography changes and discontinuities (denoted by dashed circles); as well morphology that is not natural in respect to those presented in the source DTMs. The proposed hierarchical concept, on the other hand, shows continuous topography and morphology preservation. As the proposed hierarchical concept takes into consideration the complete multi-scale geospatial inter-relations, the averaging process ignores theme and relates to the heights alone.

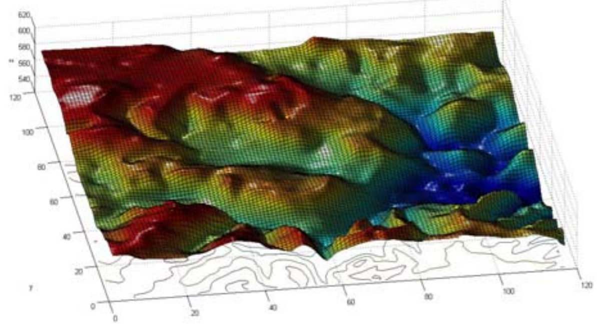


Figure 9. Integrated DTM generated in the second experiment

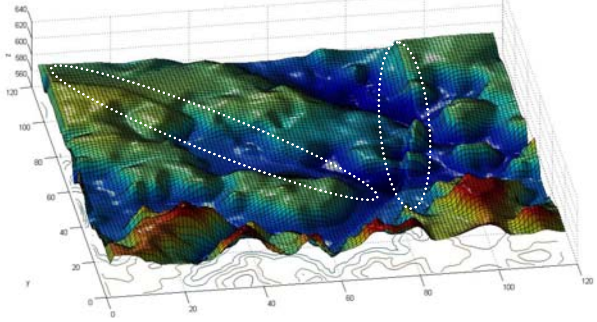


Figure 10. Integrated DTM generated by the common height averaging integration mechanism

4. CONCLUSIONS

The importance of establishing and maintaining nationwide DTM databases was discussed. Different data acquisition technologies, as well as DTM generation techniques and algorithms, derive its inner accuracy, as well as its LOD and resolution - to name a few. These factors influence mutual geometric ambiguities that exist among DTMs representing the same coverage area. A straight-forward integration mechanism can not answer the multi-scale spatial inconsistencies and multi-accuracies issues that might exist in order to produce a qualitative solution.

Novel approach is introduced that ensures the preservations of all existing mutual local correlations and inter-relations between the DTMs - instead of coercing a singular global one. This aims at retaining all topologic and morphologic inter-relations. Moreover, the varied accuracies exist in nationwide DTMs within their coverage area - depicted as an accuracy polygon maps - are taken into consideration. Utilizing the mutual accuracies in the process is important to ensure a continuous terrain relief representation despite the existence of abrupt accuracy changes - within a DTM and between DTMs - as was proved in the experiments that were carried out. The terrain relief representation of the integrated DTM is unified and continuous; it preserves inner geometric characteristics and topologic relations (morphology); it introduces more accurate modeling results of the terrain than any of the original surfaces individually by selecting the significant data out of the two available sources; thus, preventing representation distortions. Moreover, it is important to note that this approach has no dependency on the source DTMs resolution, density, datum, format and data structure. It presents a step toward integrating wide coverage terrain relief data from diverse sources and accuracies into a single and coherent DTM, thus enabling the creation of a seamless and homogenous nationwide DTM.

REFERENCES

Besl, P.J., and McKay, N.D., 1992. A method for registration of 3-D shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 14, No. 2, pp. 239-245.

Dalyot, S., and Doytsher, Y., 2008. A hierarchical approach toward 3-D geospatial data set merging. In: *Representing, Modeling and Visualizing the Natural Environment: Innovations in GIS 13*, CRC Press, London, pp. 195-220.

Frederiksen P., Grum J., and Joergensen L.T., 2004. Strategies for updating a national 3-D topographic database and related geoinformation. In: *Proc. of ISPRS XXth Congress*, WG II/IV, Istanbul, Turkey.

Hahn, M., and Samadzadegan, F., 1999. Integration of DTMs using wavelets. In: *Int. Arc. of Photogrammetry and Remote Sensing*, Vol. 32, Part 7-4-3 W6, Valladolid, Spain.

Hrvatin, M., and Perko, D., 2005. Differences between 100-meter and 25-meter digital elevation models according to types of relief in Slovenia. In: *Acta geog. Slovenica*, 45-1: pp. 7–31.

Hovenbitzer M., 2004. The digital elevation model 1:25.000 (DEM 25) for the federal republic of Germany. In: *Proc. of ISPRS XXth Congress*, WG II/IV, Istanbul, Turkey.

Hutchinson, M.F., and Gallant, J.C., 2000. Digital elevation models and representation of terrain shape. In: *Terrain Analysis: Principles and Applications*. John Wiley and Sons, Inc., New York.

Lee, J., and Chu, C. J., 1996. Spatial structures of digital terrain models and hydrological feature extraction. In: HydroGIS: Application of Geographic Information Systems in Hydrology and Water Resources Management, IAHS Publications, No. 235, pp. 201-206.

Parry, R. B., and, Perkins, C. R., 2000. World Mapping Today. University of Reading, Department of Geography, Cartographic Unit, UK, 2nd edition.

Podobnikar T., 2005. Production of integrated digital terrain model from multiple datasets of different quality. In: Int. J. of Geographical Information Science, Vol. 19, No. 1, pp. 69-89.

Spatial Data Needs: The Future of the National Mapping Program, 1990. National Research Council, Mapping Science Committee, National Academy Press, Washington, D. C.

Wang, Y., and Wade, S. E., 2008. Comparisons of three types of DEMs: a case study of the Randleman reservoir, N. Carolina. Southeastern Geographer, Vol. 48, No. 1, pp. 110-124.

INFLUENCE POWER-BASED CLUSTERING ALGORITHM FOR MEASURE PROPERTIES IN DATA WAREHOUSE

Min Ji [a, *], Fengxiang Jin [a], Ting Li [a], Xiangwei Zhao [a], Bo Ai [a]

[a] Geomatics College, Shandong University of Science and Technology, 579 Qianwangang Road, Economic & Technical Development Zone, Qingdao, China, 266510 – jimin@sdust.edu.cn

KEY WORDS: Influence Power, Hierarchical Tree, Neighbor Function Clustering, Data Mining, Gravitational Clustering, Nature Clustering

ABSTRACT:

The data warehouse's fact table can be considered as a multi-dimensional vector point dataset. In this dataset, each point's measure property can be transformed as the influence power against its neighbor points. If one point's measure is larger, it would have more influence power to attract its neighbor points, and its neighbors would have a trend to be absorbed by this point. Being inspired by the Gravitational Clustering Approach (GCA), the paper introduces a new method named IPCA (Influence Power-based Clustering Algorithm) for clustering these vector points. The paper first defines several concepts and names the local strongest power points as Self-Strong Points (SSPs). Using these SSPs as the initial clustering centers, IPCA constructs serials of hierarchical trees which are rooted by these SSPs. Because there are only a few SSPs left, by using each SSPs' influence power, the paper adopts the neighbor function clustering method to define the clustering criteria function, and gives the detail clustering procedure of SSPs. IPCA follows the nature clustering procedure at the micro-level, with a single scan, it can achieve the initial clustering. From the experiment result, we can see that IPCA not only identifies different scale clusters efficiently, but it also can get arbitrary shape clusters easily.

1. INTRODUCTION

Clustering is of fundamental importance in machine learning and data mining (Kundu, 1999). The principles of clustering include partitioning, hierarchical based, grid based, and model based (Guha, 2000. Dutta, 2005). Most of these clustering algorithms use the distance to measure the similarity between two vector points. Sometime this can partition a big nature cluster into several sub-clusters. The Gravitational Clustering Approach (GCA) (Giraud, 2005. Mohayaee, 2003. Chen, 2005) quotes the universal gravitation principle, and calculates the attraction force between two points to measure whether they merge or not. GCA doesn't restrict the clustering radius (Jiang, 2005). According the gravity power, GCA can partition vector points into a few super-spheres which have different radius, and can hold the nature clustering procedure.

As you know, the power of the gravity between two objects is determined by their qualities. If one object has more quality, it would have more power to attract its neighbor objects, and the neighbors would have a trend to be absorbed by this object. For the measure property in data warehouse's fact table, we can consider it as the purpose or result of data observation or statistics. It will have more roles and be more important than other dimensional attributes in the clustering procedure. The fact table can be considered as a multi-dimensional vector dataset, and one record is a vector point. According to the GCA idea, the magnitude of one point's measure property will determine its influence power against its neighbor points or its dependent direction. Based on this, we propose a new clustering algorithm –Influence Power-based Clustering Algorithm (IPCA). IPCA can achieve the initial clustering with a single scan. It starts from the micro-level, and can satisfy the nature clustering procedure. The rest of the paper is organized as follows; Section 2 defines several core concepts and discusses their usage in IPCA. Section 3 describes the algorithm's whole clustering idea, and defines the SSPs' clustering procedure which is referenced the neighbor function clustering method. Section 4 gives the calculation steps and Section 5 describes a simple by using a multi-dimensional fishery dataset. The result shows that IPCA is efficient, and can identify different-scale clusters or arbitrary shape clusters.

2. BASIC CONCEPTS OF IPCA

With the long-term observation and statistics of the natural and social phenomena, the people has accumulated a large number of multi-dimensional datasets, such as biological population distribution data, natural resource survey data, economic development statistics data, etc. All these data are integrated respectively in data warehouse's multi-dimensional fact tables (Marc, 1997. Coliat, 1996. Han, 2000). In the fact table, there are two kinds of attributes: dimensions and measures. We can classify these dimensions as classification dimension, order dimension, temporal dimension, spatial dimension, and so on. All these dimensions construct a multi-dimension space, and divide the space as a multi-dimension cube collection. Each record in the fact table corresponds one of these cubes. While measure properties correspond to the observation value or the statistics in the cube, and represent the purpose and results of data acquisition. They should have more roles in the clustering procedure. In order to descript more clearly for the follow sections, we gives these follow concepts.

The Influence Power is used to measure the important degree of a multi-dimension point. For the multi-dimension dataset in data warehouse, it refers to the measure property of one point. According to GAC idea, if one point has a higher measure value,

[*] Min Ji , Associate Professor of Shandong University of Science and Technology, Research Fields: Spatial Data Organization and GIS Application Developing, Email: jimin@sdust.edu.cn, jamesjimin@hotmail.com

it will have a higher influence power to attract its neighbor points, and it will have more possibilities to become a cluster center. So, we use the magnitude of one point's measure property as the influence power, and also normalize it by using Eq. (1).

$$I_i = \frac{p_i - P_{\min}}{P_{\max} - P_{\min}} \qquad (i = 1,2,\cdots,n) \qquad (1)$$

where n = the number of all points
p_{i} = the measure value of the ith point
$P_{\max}$ = the maximum measure value
$P_{\min}$ = the minimum measure value
$I_{\mathrm{i}} \in [0,1]$

Grid Neighbor Points (GNPs) refer to those points that have one unit distance with the current point along each dimension in the multi-dimensional cube collection (also called as grid matrix). Fig. 1 shows a three-dimensional cube, the red point at the center is the current point that we will examine its influence power, and all the black points are its GNPs.

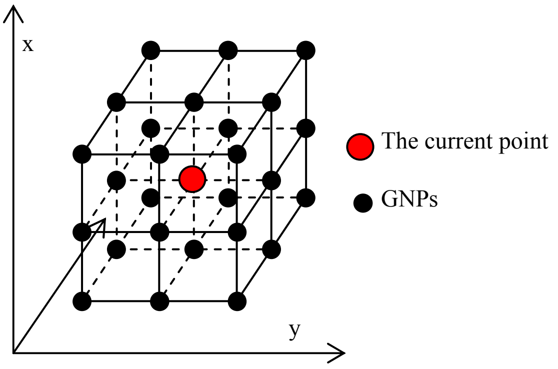


Figure 1. The relationship of current point and its GNPs

The Maximum Connection Strength point (MCS point) is the point among all GNPs of the current point which has the strongest influence power. If there are several GNPs which influence power are larger than the current point's, the current point will have a trend to be absorbed by these GNPs. Which GNP should the current point belong to? We should follow the principle of playing up to those in power. And then the current point should connect to the GNP which has the strongest influence power. That will be the maximum connection strength point.

The Strongest Power Connection direction (SPC direction) is the current point's attaching direction that connects to its MCS point.

Self-Strong Point (SSP) is the point that has the maximum influence power at local area. If one point's influence power is higher than any one's power of its GNPs, it will become the local strongest power point. Because it has no MCS point, we call it as self-strong point. There is a special situation. If one point has no GNPs, in another words, it is an isolate point, we also call it as SSP. SSPs can be used as the initial clustering centers.

SuBsidiary Points (SBPs) refer to those points that are not SSPs. Each SBP has its own MCS point, and will connect to its MCS point. But one SBP's status is not everlasting, it also can be a MCS point, and can have its own SBPs.

Natural Clustering Hierarchy tree (NCH tree) is the core concept. NCH tree is composed by a SSP, MCS points and SBPs. The root node is the SSP. Each parent node corresponds to a MCS point, and the SSP is the top MCS point. Each MCS point at least has one child node, and this child is its SBP. One SBP also can be a MCS point, and can have its own SBPs, so there forms one to many hierarchical relationships between MCSs and SBPs. One NCH tree can be seen as an initial cluster.

3. THE IPCA CLUSTERING ALGORITHM

3.1 The Initial Clustering

According to the description of Section 2, the measure property can be as a key index for estimating one multi-dimension point's important degree. If one point's measure value is higher, it will have more influence power to attract its GNPs, and it will have more possibilities to become a cluster center. Based on this common knowledge, by calculating each point's influence power, we can get many local strongest power points (SSPs), and can select these SSPs as the initial clustering centers. SSPs can be as the first level of MCS points, and the next step is just continuously expanding MCS points' SBPs. Because one SBP can also be a MCS point, only after each point has been scanned, and has found its MCS point, the initial clustering would accomplish. After this procedure, we will get a few NCH trees. The number of NCH trees just equals the SSPs'. If the condition is good enough, we would get the final clustering result. But for a general condition, we should go another deeper step.

After the initial clustering procedure, each NCH tree represents a sub-cluster. Our follow issue is just how to measure the similarity between these sub-clusters. Because each NCH tree has a SSP, we only measure the similarity between SSPs, we will get the answer. Section 3.2 just describes SSPs' clustering procedure.

3.2 SSPs' Clustering Algorithm

In accordance with the whole clustering idea, the final clustering process is mainly reflected in SSPs' clustering procedure. The quality of SSPs' clustering will have a direct impact on the quality of the final clustering result. Inspired from the neighbor function clustering method (Sun, 2001), we proposed the follow influence power-based SSPs' neighbor function clustering method.

Firstly, we give a distance threshold $\mathrm{D}_{\max}$ which is calculated by numerical dimensions of these SSPs. Within one SSP's $\mathrm{D}_{\max}$ rang, there forms a SSPs' sub-group. Each SSP in the sub-group will have an attraction force on this current point. We can use Eq. (2) to calculate the force.

$$E_{oj} = I_j / D_{oj} \qquad (j = 1,2,\cdots,m) \qquad (2)$$

where m = other SSPs within current SSP's D_{max} rang
I_j = the influence power of the jth SSP
D_{oj} = the distance between current SSP and its jth SSP

According to E_{oj} values, we sort them descending and there forms the influence neighbor order between current SSP and its surrounding SSPs. Here, we give another concept – the Influence Neighbor Points (INPs), which are those SSPs within current SSP's D_{max} rang.

For any two SSPs P_i and P_j, if P_i is P_j's Ith INP, we designate I as the influence neighbor coefficient that P_i puts on P_j, denoted as E (i, j) = I; Similarly, we can get E (j, i) = J. Then, we define Eq. (3) as the influence neighbor function between P_i and P_j.

$$\alpha_{ij} = E(i,j) + E(j,i) - 2 = I + J - 2 \tag{3}$$

If P_i and P_j are the first INP for each other, then $\alpha_{ij} = 0$. This shows that the smaller the α_{ij}, the greater the attraction between two points, and the more possibility to cluster together. Suppose the number of SSPs is N, then $\alpha_{ij} \leq 2N - 4$. If the distance between P_i and P_j exceeds D_{max}, we require $\alpha_{ij} = 2N$. In order to avoid SSP's self-loop clustering, we also require $\alpha_{ii} = 2N$.

In SSPs' clustering process, if P_i and P_j can merge together, we can claim that P_i and P_j are connected to each other. In order to show the loss for the connection, we introduce the SSP Connection Loss concept. According to Eq. (3), we can use α_{ij} as the SSP Connection Loss.

If SSPs can merge together, we define the connection loss in this cluster as $\Sigma\alpha_{ij}$. If there are c clusters: ω_p, p = 1,2, ..., c, we can define the total connection loss among these clusters as Eq. (4).

$$L_w = \sum_{p=1}^{c} \sum_{\substack{P_i \in \omega_p \\ P_j \in \omega_p}} \alpha_{ij} \tag{4}$$

In order to describe SSPs' internal expanding degree, we define the inner maximum connection loss in one cluster as Eq. (5).

$$\alpha_{pm} = \underset{\substack{P_i \in \omega_p \\ P_j \in \omega_p}}{MAX} [\alpha_{ij}] \qquad p = 1,2,..., c \tag{5}$$

In order to describe the similarity between clusters, we define the connection loss between clusters as Eq. (6).

$$\gamma_{pq} = \underset{\substack{P_i \in \omega_p \\ P_j \in \omega_q}}{MIN} [\alpha_{ij}] \qquad p,q = 1,2,..., c; p \neq q \tag{6}$$

In order to determine the quality of the previous iterative clustering results, we define the minimum connection loss between cluster ω_p and all the other clusters as Eq. (7).

$$\gamma_{pk} = \underset{\substack{q \\ q \neq p}}{MIN} [\gamma_{pq}] \qquad q = 1,2,..., c \tag{7}$$

If $\gamma_{pk} > \alpha_{pm}$ *and* $\gamma_{pk} > \alpha_{km}$, it shows that the previous clustering result is successful, otherwise the clusters should merge together. The merge condition is as $\gamma_{pk} \leq \alpha_{pm}$ *or* $\gamma_{pk} \leq \alpha_{km}$.

In order to describe the total loss between all the clusters, we define Eq. (8) as the loss-cost function between clusters.

$$\beta_p = \begin{cases} (\alpha_{pm} - \gamma_{pk}) + (\alpha_{km} - \gamma_{pk}) & if\, \gamma_{pk} > \alpha_{pm}\, and \gamma_{pk} > \alpha_{km} \\ \alpha_{pm} + \gamma_{pk} & if\, \gamma_{pk} \leq \alpha_{pm}\, and \gamma_{pk} > \alpha_{km} \\ \alpha_{km} + \gamma_{pk} & if\, \gamma_{pk} > \alpha_{pm}\, and \gamma_{pk} \leq \alpha_{km} \\ \alpha_{pm} + \alpha_{km} + \gamma_{pk} & if\, \gamma_{pk} \leq \alpha_{pm}\, and \gamma_{pk} \leq \alpha_{km} \end{cases} \tag{8}$$

In Eq. (8), the first case is a reasonable clustering result; β_p is negative, indicating that the loss is negative. The other cases are unreasonable, there has a need to merge clusters, β_p is positive, indicating there has some connection loss.

Based on Eq. (8), we define the total connection loss among all the clusters as Eq. (9).

$$L_B = \sum_{p=1}^{c} \beta_p \tag{9}$$

The final goal of SSPs' Clustering is to enable γ_{pk} as large as possible, and enable α_{pm} as small as possible, so we define the clustering criterion function for SSPs as Eq. (10).

$$J_L = L_W + L_B \quad \rightarrow \quad Min \tag{10}$$

4. IPCA CLUSTERING PROCEDURE

Based on the description and definition of the earlier sections, we experimented with a three-dimension dataset whose record number is 2187. We present serials of efficient steps for obtaining the finial clusters by using each point's measure property. All the steps are as follows.

Step 1. Find the maximum measure value P_{max} and the minimum value P_{min} from the entire dataset.

Step 2. Normalize the measure value for each record by using P_{max} and P_{min}, and get each point's influence power I_i.

Step 3. According to each point's influence power, select SSPs and construct NCH trees. From the 2178 points, we got 153 SSPs.

Step 4. Sort these SSPs descending by using their influence power, and construct the SSPs' dataset $\{P_1, P_2, \ldots, P_N\}$.

Step 5. Calculate the Euclidean inverse distance matrix ***D***, the element in ***D*** is as $D_{ij} = 1/\ d\ (P_i, P_j)$, where d (P_i, P_j) is the Euclidean distance between the *i*th SSP and the *j*th SSP, $i \neq j$. Set the main diagonal element as 0.

Step 6. According to the data characters, set the distance threshold D_{max}. Set the matrix element's value as 0 if its value is smaller than $1/D_{max}$.

Step 7. Construct the influence power vector $\boldsymbol{I} = (I_1, I_2, \ldots, I_N)$.

Step 8. Calculate the attraction degree matrix $\boldsymbol{M}=\boldsymbol{DI}^T$.

Step 9. Calculate the influence neighbor coefficient of each non-zero element in matrix ***M***, and form the influence neighbor coefficient matrix ***H***.

Step 10. According to the non-zero element in matrix ***H***, calculate the influence neighbor function matrix ***L***, where the element $L_{ij} = h_{ij} + h_{ji}$-2. Set all zero elements as 2*N*. L_{ij} represent the connection loss if two SSPs merge together.

Step 11. Select these first-M SSPs as the clustering centers. According matrix ***L*** to determine whether there are any SSPs which have minimum influence neighbor function value (min-value) for each other in these first-M SSPs, if existing, then merge these SSPs together. After that, determine whether the remaining SSPs have min-value with these first-M SSPs, if existing, then merge them to their corresponding clusters, otherwise, they will be consider as a single cluster. These single clusters maybe outliers, if one single cluster doesn't have SBPs, we should delete it from the cluster collection.

Step 12. According Eq. (5-7) to calculate $\gamma_{pk}(p=1,2,\cdots,c)$、$\alpha_{pm}$、$\alpha_{km}$, if $\gamma_{pk} \le \alpha_{pm}$ *or* $\gamma_{pk} \le \alpha_{km}$, merge ω_p and ω_k into one cluster, and then repeat this step. This procedure will iterative until Eq. (10)'s J_L value is not diminished

Step 13. After accomplish SSPs clustering procedure, finally add each node of each NCH trees to its corresponding cluster and finish the whole clustering procedure.

5. EXPERIMENTS

Based on the previous clustering procedure, by using 2178 records in one marine fishery data warehouse's fact table, we got 12 natural clusters, which include 7 big clusters, 2 small clusters and 3 outlier clusters, just as shown in Table 1. From Table 1, we can see that cluster 10, 11, 12 only have one SSP , and do not have any SBPs, they are far from any clusters, obviously they are outliers, we should delete them from the finial clustering result. Cluster 6 and cluster 9 only have a small number of SSPs and their total point numbers are also far less than other clusters. They can be considered as small-size clusters. Cluster 2 is the biggest-size cluster. It has 40 SSPs, and 722 points, more than 30 times of cluster 6 or cluster 9. This demonstrated that IPCA method can identify multiple clusters with different scales very clearly. Fig. 2 illustrates the clustering distribution in three-dimensions which are x-dimension, y-dimension, and T-dimension. The x, y values represent the vector point's coordinate position in our flat space, and the T-value represents the point's happening time in the temporal space. From the three-dimensional clustering distribution map, we can see that there are very large changes in these clusters' shapes. They are not spherical, not liner. They can be arbitrary shapes. This will be an excited character than the traditional clustering algorithm.

No.	SSPs Num.	Total Num.	Measure Property	Cluster Center		
				x	y	T
1	25	482	1661.6	146	41	47
2	40	722	1619	154.5	43.5	35
3	15	239	1002	166.5	41.5	29
4	23	317	771.5	159.5	44	38
5	18	145	751.3	161.5	41.5	31
6	2	24	600.6	151.5	41	47
7	11	130	437.9	-175	41	27
8	10	95	219.8	178	41	28
9	6	21	34.7	171.5	38.5	24
10	1	1	8.7	166	38	20
11	1	1	8.1	178.5	40.5	39
12	1	1	7	180	36.5	26

Table 1 The clustering result table of 2178 vector points

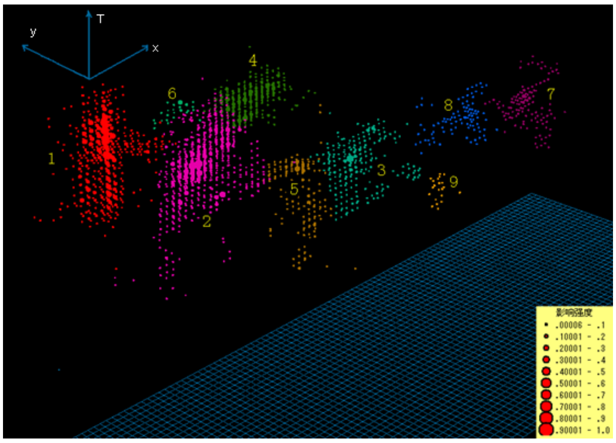


Figure 2 The 3D distribution map of clustering result

6. CONCLUSION

This paper presents the IPCA algorithm, a new clustering algorithm based on objects' influence power against each other, which is inspired from the gravity theory in physics. This method is particularly suitable for handling multi-dimensional huge data collections in data warehouse's fact table. Because measure properties are observation values or statistic results, they should have more roles in data mining procedure. IPCA just uses the measure property of each multi-dimensional record to measure its attraction force on its neighbor records in the multi-dimensional space. Only for a single scan, this algorithm can get the initial clustering result and construct many hierarchy trees which are rooted by self-strong points which have the strongest influence power in the local area. If the condition is good enough, we can get the finial clusters. IPCA follows the natural clustering process, and the experimental results also show that it can identify any size clusters and arbitrary shape clusters efficiently. These two characters will make it as a new member in the data clustering analyze family.

REFERENCES

S. Kundu, 1999. Gravitational Clustering: a new approach based on the spatial distribution of the points, Pattern Recognition, 32, pp. 1149-1160.

S. Guha, R. Rastogi, and K. Shim, 2000. Rock: A robust clustering algorithm for categorical attributes, Information systems, 25(5), pp. 345-366.

M. Dutta, A. Kakoti Mahanta, A. K. Pujari, 2005. QROCK: A quick version of the ROCK algorithm for clustering of categorical data, Pattern Recognition, 26, pp. 2364-2373.

J.A. Garcia, J. Fdez-Valdivia, F.J. Cortijo, and R. Molina, 1995. A dynamic approach for clustering data, Signal Processing, 44,pp. 181-196,

C. Giraud, 2005. Gravitational clustering and additive coalescence, Science Direct, 115, pp. 1302-1322.

R. Mohayaee, L. Pietronero, 2003. A cellular automaton model of gravitational clustering, Science Direct, 323, pp. 445-452.

C.Y. Chen, S.C. Hwang, Y.J. Oyang, 2005. A statistics-based approach to control the quality of subclusters in incremental gravitational clustering, Pattern Recognition, 38, pp. 2256-2269.

S.Y. Jiang, Q.H. Li, 2005. Gravity-based clustering approach, Journal of Computer Applications, 2, pp. 285-300.

Jiawei Han, M. Kamber, 2000. DATA MINING: Concepts and Techniques, Morgan Kaufmann Publishers, pp. 100-150.

Marc G., 1997. A Foundation for Multi-dimensional Databases, In Proc. of the 23rd VLDB Conference, pp. 106-115.

Coliat G, 1996. OLAP, relational, and multi-dimensional database system, ACM SIGMOD Record, 25(3), pp. 64-69.

J.X. Sun, 2001. Modern Pattern Recognition, National University of Defense Technology, pp. 40-43.

APPENDIX

This article was sponsored by Chinese National Programs for High Technology Research and Development (2009AA12Z147); the Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping (2009B14); the Key Laboratory of Advanced Engineering Surveying of SBSM (TJES0805).

NORMALIZING SPATIAL INFORMATION TO IMPROVE GEOGRAPHICAL INFORMATION INDEXING AND RETRIEVAL IN DIGITAL LIBRARIES

Damien Palacio and Christian Sallaberry and Mauro Gaio

LIUPPA, Université de Pau
avenue de l'Université
64000 PAU, FRANCE
damien.palacio@univ-pau.fr, christian.sallaberry@univ-pau.fr, mauro.gaio@univ-pau.fr

KEY WORDS: Geographic Information Retrieval, Spatial Information, Normalization, Geographic Information Combination

ABSTRACT:

Our contribution is dedicated to geographic information contained in unstructured textual documents. The main focus of this article is to propose a general indexing strategy that is dedicated to spatial information, but which could be applied to temporal and thematic information as well. More specifically, we have developed a process flow that indexes the spatial information contained in textual documents. This process flow interprets spatial information and computes corresponding accurate footprints. Our goal is to normalize such heterogeneous grained and scaled spatial information (points, polylines, polygons). This normalization is carried out at the index level by grouping spatial information together within spatial areas and by using statistics to compute frequencies for such areas and weights for the retrieved documents.

1 INTRODUCTION

The digitization of printed literature is currently making significant progress. The Google Books Library Project, for instance, aims at creating digital representations of the entire printed inventory of libraries. Other initiatives specialize in the legacy literature of specific domains, such as medicine or cultural heritage (Sautter et al., 2007). For instance, libraries or museums are now offering their electronic contents to a growing number of users.

While some projects only aim at creating digital versions of the text documents, domain-specific efforts often have more ambitious goals (Sautter et al., 2007). For example, to maximize the use of the contents, text documents are annotated and indexed according to domain-specific models. The Virtual Itineraries in the Pyrenees[1] (PIV) project[2] consists in managing a repository of the electronic versions of books (histories, travelogues) from the 19th and 20th centuries. It appears that the contents present many geographic aspects (Marquesuzaà et al., 2005). This kind of repository is quite stable (few suppressions and modifications, regular inserts of documents) and not too large. Therefore, the cost of a back-office refined semantic aware automated indexing is reasonable (Gaio et al., 2008).

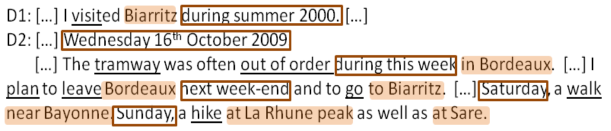
Although well-known search engines still deliver good results for pure keyword searches, it has been observed that precision is decreasing, which in turn means that a user has to spend more time in exploring retrieved documents in order to find those that satisfy his information needs (Kanhabua and Nørvåg, 2008). One way of improving precision is to include a geographical dimension into the search. We consider the generally accepted hypothesis that Geographical Information (GI) is made up of three components namely spatial, temporal and thematic. A typical textual sample is: "Fortified towns in the south of the Aquitaine basin in the 13th century." To process this textual unit, we claim that each of its three components (spatial, temporal and thematic) should be treated independently, as is put forth by (Clough et al., 2006). This can be done by making several indexes, one per component, as is advised by (Martins et al., 2005). In this way, one can limit the search to one criterion and easily manage the indexes (e.g., to allow adding documents to the corpus). So, our approach consists in processing components independently, in order to better combine them later on. It contributes to the field of Geographic Information Retrieval (GIR) as defined by (Jones and Purves, 2006).

The current version of the PIV platform is comprised of three independent process flows: spatial (Gaio et al., 2008), temporal (Le Parc-Lacayrelle et al., 2007) and thematic (Sallaberry et al., 2007). For example, Figure 1 illustrates automatic annotations resulting from such process flows: spatial information is highlighted, temporal information is outlined and the thematic one is underlined. Figure 2 illustrates the richness and accuracy of the resulting specific indexes: i.e., the PIV computes geometric representations of spatial information, time intervals corresponding to temporal information and lists of terms corresponding to thematic information. Experiments (Sallaberry et al., 2007) demonstrate the effectiveness of these indexes within specific spatial, temporal or thematic information retrieval scenarios. Two important problems were pointed out during these experiments: 1-results scoring does not integrate spatial features or temporal features frequency within documents: e.g. we are looking for "Biarritz," D1 and D2 will have the same weight even if D1 contains only "Biarritz" spatial feature whereas D2 contains "Biarritz" spatial feature and many other ones; 2-merging results within a geographic information retrieval process remains a challenge (Visser, 2004): as each index is built with one dedicated approach, as well as each document relevancy calculation formula is based on different methods (which correspond respectively to spatial, temporal or thematic criteria), how to combine spatial, temporal and thematic specific relevancy scores of the retrieved documents?

We propose to normalize each geographic indexing criteria. It consists on rearranging geographic information within a uniform representation form: we represent geographic information within spatial tiles (spatial areas), temporal tiles (calendar intervals) and thematic tiles (concepts) and compute each tile evocation frequency in the documents. Then, we apply statistic formulae generally used for plain-text information retrieval to compute relevancy scores for each resulting document.

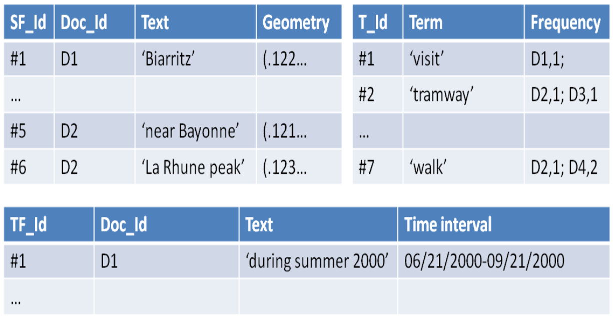
[1] Mountains of the south west of France

[2] Part of this project is supported by the Greater Pau City Council and the MIDR media library



D1: [...] I visited Biarritz during summer 2000. [...]
D2: [...] Wednesday 16th October 2009
[...] The tramway was often out of order during this week in Bordeaux. [...] I plan to leave Bordeaux next week-end and to go to Biarritz. [...] Saturday, a walk near Bayonne. Sunday, a hike at La Rhune peak as well as at Sare.

Figure 1: Example of automatically annotated textual documents



SF_Id	Doc_Id	Text	Geometry
#1	D1	'Biarritz'	(.122...
...			
#5	D2	'near Bayonne'	(.121...
#6	D2	'La Rhune peak'	(.123...

T_Id	Term	Frequency
#1	'visit'	D1,1;
#2	'tramway'	D2,1; D3,1
...		
#7	'walk'	D2,1; D4,2

TF_Id	Doc_Id	Text	Time interval
#1	D1	'during summer 2000'	06/21/2000-09/21/2000
...			

Figure 2: Example of spatial, temporal and thematic indexes

This approach proposes (1) frequency parameter integration within the relevance scoring algorithms and (2) geographic data normalization within a new level of spatial, temporal and thematic indexes. Moreover, we propose to produce different granularity level indexes (for example, spatial administrative segmentations: cities, counties, countries) in order to parse the indexes best suited to the grain of each query.

Merging results provided by such hybrid querying criteria would only make sense if such normalized indexes were homogeneous as well as if the relevance calculation formulae were similar. That is why the next section presents a spatial normalization approach, which we will later apply to the temporal and thematic aspects.

The paper is organized as follows. Section 2 briefly outlines the textual process flow indexing geographic information within the PIV prototype. Section 3 describes related works and our proposals for the creation of new indexes through spatial normalization. Section 4 details the proposed model for computing spatial relevance and describes experiments we carried out to evaluate these propositions. Finally, section 5 and 6 discuss our future perspectives and conclude.

2 TEXTUAL PROCESS FLOW LEADING TO SPATIAL NORMALIZATION

A document textual content processing sequence is usually composed of four main steps: (a) "tokenization" splits the document into smaller blocks of text, (b) lexical and morphological analysis carries out recognition and transformation of these blocks into lexemes, (c) the syntactic analysis, based on grammar rules, allows links between lexeme to be found, finally, (d) the "semantic" step carries out a more specific analysis allowing meaningful lexeme groupings to be interpreted.

As explained hereinafter, our data processing sequence is quite different. This spatial information process flow is described in Figure 3. Steps 1 to 4 are detailed in (Gaio et al., 2008). This approach was developed and experimented within the PIV project:

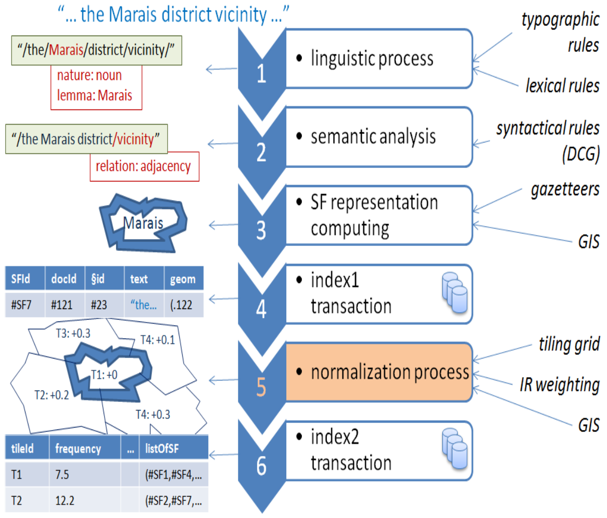


Figure 3: Spatial information process flow

1. After a classical textual tokenization preprocessing sequence and according to (Baccino and Pynte, 1994) we adopt an active reading behavior, that is to say sought-after information is a priori known. A marker of candidate spatial token locates spatial named entities using typographic and lexical rules (involving spatial features initiator lexicons). Then, a morphosyntactic analyzer associates a lemma and a nature with each candidate token (e.g. "Marais", noun).
2. A semantic analyzer marks candidate Absolute Spatial Features (ASF, e.g. "Marais district") first and candidate Relative Spatial Features (RSF, e.g. "Marais district vicinity") next thanks to a Definite Clause Grammar (DCG). For instance, syntagms of composed nouns (i.e. "Marais district," "Emile Zola street," "Wild Chamois peak") are brought together and spatial relationships (adjacency, inclusion, distance, cardinal direction) are tagged (Egenhofer, 1991).
3. ASF are validated and geolocalized thanks to external and/or internal gazetteers (IGN French Geographic Institute resources, Geonames resources and contributive hand-craft local resources). Then expressions containing RSF are built from pointed out ASF: embedded spatial relationships (e.g. adjacency: "vicinity") are interpreted and corresponding geometries are computed.
4. Only validated spatial features are retained. Thus we get a spatial index describing each SF with the corresponding geometry, text, paragraph and document. This first level of index supports IR scenarios: query/index overlapping geometries are computed and scored relevant textual paragragraphs are returned.
5. SF are grouped, weighted and mapped into a set of segmented grids. We propose different grained tiling grids: regular and administrative grids (district, city, county, ...). We use information retrieval TF.IDF formulae (Spärck Jones, 1972) to compute spatial tiles' frequencies and weight them.
6. Finally, we get a spatial index describing each tile with the corresponding frequency and SF (geometry, text, paragraph and document). This second level of index supports new IR capabilities: a query is mapped to the more convenient grid and query/index overlapping areas are computed and relevance scoring algorithms integrate each tile frequency. This promotes textual paragraphs centered on the required SF only. Moreover, it allows different querying strategies: for example, thin-scale queries are compared to district grids and large-scale ones are compared to country grids.

A GIS supports spatial operations of all the previous stages. This paper focuses on the spatial information normalization process (stage 5). It describes statistical IR approaches integration in such a process. An experiment compares different index tiling grids and IR statistical formulae to validate our propositions.

3 SPATIAL INFORMATION GATHERING FOR SPATIAL NORMALIZATION

3.1 Related works

One of the most popular models developed in textual-based information retrieval research is the vector space model (Salton and McGill, 1983). Using a vector space model, the content of each document can be approximately described by a vector of (content-bearing) terms (Cai, 2002). An information retrieval system stores a representation of a document collection using a document-by-term matrix (Table 1), where the element at (i, j) position corresponds to the frequency of occurrence of term i in the jth document (Manning et al., 2008, Cai, 2002).

$$\begin{array}{c c} & \begin{array}{cccc} T_1 & T_2 & \dots & T_t \end{array} \\ \begin{array}{c} D_1 \\ D_2 \\ \vdots \\ D_n \end{array} & \begin{pmatrix} w_{11} & w_{21} & \dots & w_{t1} \\ w_{21} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \dots & w_{tn} \end{pmatrix} \end{array}$$

Table 1: Document-by-T matrix within the vector space model

The vector space model can support selecting and ranking of documents by computing a similarity measure between a document and a query or another document (Salton and McGill, 1983). There are obvious advantages and disadvantages of using vector space model in retrieving geographical information. Vector space model is well accepted as an effective approach in modeling thematic subspace and it allows spatial information to be handled the same way as thematic information (Cai, 2002). (Cai, 2002) proposed to manage place names within a vector space model. Place names are integrated as independent dimensions in a vector space model, whereas in fact, they are points (or regions) in a two-dimensional geographical space. In order to improve such a keyword-based search method, (Cai, 2002) proposed to integrate proper ontologies of places as promoted by (Jones et al., 2001).

Our approach is different as it extends such a term-based matrix to a tile-based matrix. In the vector space model, all the objects (terms, spatial tiles, temporal tiles, thematic tiles (concepts)) can be similarly represented as vectors. This paper proposes to gather SF into spatial tiling grids to compute a similar document-by-tile matrix (Figure 1), where the element at (i, j) position corresponds to the frequency of occurrence of spatial tile i in the jth document.

On the one hand, current spatial oriented research works distinguish:

- spatial generalization: defined as spatial features selection, displacement and/or simplification processes (Zhang, 2005, Zhou and Jones, 2004, Zhou et al., 2000, Glander and Döllner, 2007);
- spatial normalization: defined as an image registration process estimating and applying warp-fields (Robbins et al., 2003);
- spatial summarization: defined as spatial features aggregation / combination into larger features (i.e. cell-based structure) (Rees, 2003).

On the other hand, information retrieval oriented research works define normalization as a stemming process of words in order to gather and weight them (Spärck Jones, 1972, Li et al., 2002). So, what we call normalizing spatial information, in the following section, means spatial information (representations computed from textual documents) gathering into spatial tiles in order to weight them according to frequency computations.

The originality of the approach described in the following section consists in:

a) the proposition of different granularity level spatial indexes: administrative and/or regular grids;

b) the adaptation of effective full-text IR technics in order to process such indexes.

3.2 Spatial Gathering for normalization

First, we detail the spatial normalization process (stage 5 Figure 3) leading to the index2 (stage 6 Figure 3). Then, we briefly explain how we take advantage of this normalized index within an IR process.

3.2.1 Information Indexing. Our approach consists in gathering spatial information into a unique type of spatial representation: the tile. So we divide space by attaching each detected SF to tiles. It is similar to the lemmatisation process, for which each term is attached to a lemma. Two segmentation types are possible. The first concerns regular tiles (i.e., segmentation into rectangular tiles of the same size — see Figure 4). It is similar to truncation [3]. The second concerns administrative tiles (i.e., segmentation into cities for example — see Figure 5). It is similar to lemmatisation [4]. To calculate a tile frequency, one just has to count the number of SF that intersect it, while keeping in mind that a SF can intersect several tiles.

For illustration purposes, let's go back to the example in Figure 1 and 2. If we choose to use regular segmentation (Figure 4), we obtain the tiles index shown in Table 2. In this table, several scenarios are presented. First, SF #5 intersects two tiles (T2 and T3); so the discrete frequency of both of them is incremented by 1. Moreover tile T2 is intersected by two SF (#1 and #5); consequently it has a weight of 2.

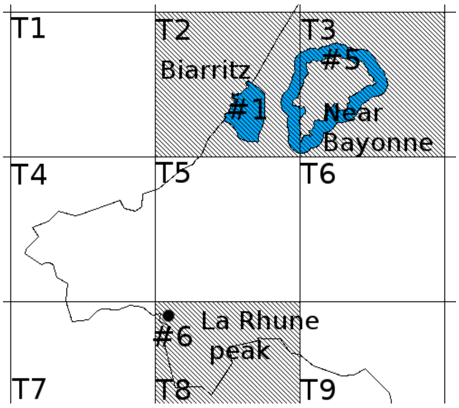


Figure 4: Part of thin-grained SF obtained in index1 projected on a segmentation by regular grid

id$_t$	id$_{sf}$ list	discrete frequency	continuous frequency
T1	[]	0	0
T2	[#1;#5]	2	0.15
T3	[#5]	1	0.20
...			

Table 2: Spatial index2 with regular tiles (phase 6 - Figure 3)

One should note the granularity problem of the spatial information that is being processed, and the proportionality issue between

[3] e.g. for word "forgotten" the truncation returns "forgott"

[4] e.g. for word "forgotten" the lemmatisation returns "forget"

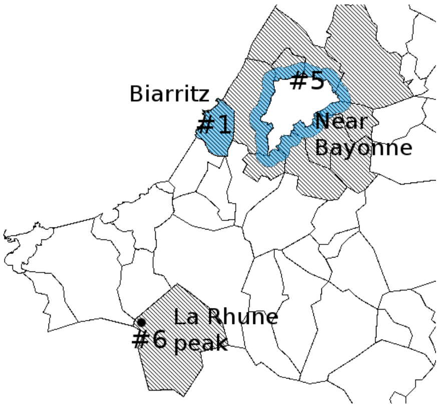


Figure 5: Part of thin-grained SF obtained in index1 projected on an administrative segmentation (cities)

Discrete frequency	$freq_t = \sum freq_{sf}$
Continuous frequency	$freq_t = \sum freq_{sf} * \frac{Ar_{sf,t}}{Ar_t} * \frac{1}{NbTiles_{sf}}$

Table 3: Frequency formulae ($Ar_{sf,t}$: SF area on tile t, Ar_t: tile t area, $NbTiles_{sf}$: number of tiles intersected by the SF)

its representation (SF) and a tile as well as the size of their overlapping area. Indeed, one may wonder whether a SF's area that only covers a small part of a tile should have as great a weight as a SF's area that covers most of the area of the same tile? Thus, we suggest two frequency calculation (see Table 3). Indexing may be discrete; so, for a given document unit, a tile frequency is incremented by 1 every time there is an intersection with a SF (Table 2 column 3). We have also considered a continuous indexing approach. According to the ratio overlay SF/tile, a tile frequency is incremented by a value between 0 and 1 (Table 2 column 4).

These indexes are intended to support spatial IR. This involves weighting the results. Consequently, we use regular IR formulae and carry out experiments on such indexes of spatial tiles.

3.2.2 Information Retrieval. We propose 4 IR formulae (Table 4). TF, TF.IDF and OkapiBM25 are dedicated to discrete weighting. They are widely used in full-text IR (Manning et al., 2008, Savoy, 2002). In our context, the TF formula avoids reducing the weight of overly frequent tiles. Nevertheless, classical frequency does not take spatial specificity like granularity into account. That is why we decided to apply TF onto continuous frequency, and we call this approach TFc.

4 EXPERIMENTS

Our hypothesis is that the segmentation must be adapted to the SF type of the corpus and of the queries. We propose to use the granularity of the query to choose the best suited index. For complex queries, composed of different grained SF, we propose to define a default well suited index. So here we are looking for what segmentation and weighting formula give the better results for our corpus.

Our approaches are based on thin-grained spatial data. But spatial evaluation campaigns like GeoCLEF (Mandl et al., 2007) do not give accurate resources (like polygons) and do not handle French documents. That's why we realize our experiment on our cultural heritage digital library.

10 French books were indexed. In order to compare our propositions to manually sorted methods, we chose a sample of 1,019 document units, corresponding to 1,028 SF (902 ASF and 126 RSF). Each document unit may contain from 0 to many SF. We submitted 40 queries (the baseline is index1). 15 queries involve an ASF : 5 of each type (small grained like peaks, intermediate grained like cities and large grained like regions). 25 other contain a RSF : 5 of each type (orientation, proximity, union, inclusion, distance). Table 5 shows some examples of relations. We observed that 30% of our ASF are well identified cities, 12% are larger well identified ASF (department, regions), 38% are smaller ASF (peaks, cabins, . . .) and the others (approximatively 20%) have a variable average size.

Tile Frequency (TF)	$W_{t,Du} = TF_{t,Du} = \frac{freq_{t,Du}}{\sum_{i=1}^{n} freq_i}$
TF.IDF	$W_{t,Du} = TF_{t,Du} * IDF_t$ and $\mathrm{IDF}_t = log\left(\frac{NDu}{NDu_t}\right)$
OkapiBM25	$W_{t,Du} = \left(\frac{(k_1+1)*TF_{t,Du}}{(K+TF_{t,Du})}\right)$ and $\mathrm{K} = \mathrm{k}_1 * [(1-b) + \frac{b*n}{advl}]$
TFc	$W_{t,Du} = TFc_{t,Du} = \frac{freqC_{t,Du}}{\sum_{i=1}^{n} freqC_i}$
$freq_{t,Du}$: tile frequency in the document unit, n : number of tiles in the document unit, NDu_t : number of document units with tile t, NDu : number of document units, k_1 = 1.2, b = 0.75, $advl$ = 900, $freqC$: continuous frequency	

Table 4: Weighting formulae, used with index2, for a tile t and a document unit Du

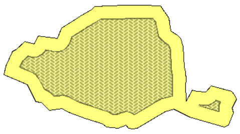
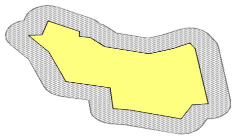
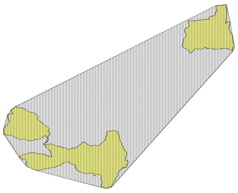
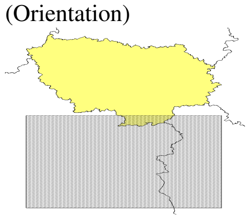
in downtown Paris (Inclusion) 	near Gavarnie (Proximity) 
on Tarbes-Lourdes axis (Union) 	in south of Ile-de-France (Orientation) 

Table 5: Examples of RSF

We tried 6 different indexing segmentations: 3 administrative segmentations (city, department and region) and 3 regular segmentations (grid of 100x100, grid of 200x200 and grid of 400x400). The grid of 200x200 corresponds to the average city size. Finally, we tested all theses segmentations with the 4 weighting formulae presented in last section (TF, TF-IDF, OkapiBM25, TFc).

As we can see in Table 6, for all segmentations, the TFc gives the best results. As we explained in section 3, every classical statistical weighting formulae (TF, TF-IDF, OkapiBM25) use discrete frequency. They give the same weight for a geometry which fills the major part of one tile, and for a geometry which represents a little part of the same tile. On the contrary, the TFc uses continuous frequency and gives a weight depending on the area of overlapping between the tile and the SF's geometry.

Concerning the segmentation, the Table 6 shows that all segmentations give good results excepts department and regions (they are too large so they gather SF which are too far away from each other). For segmentation by regular grid, the one of 200x200 gives the best results. Concerning the administrative segmentation, city segmentation gives the best results. The main explanation is that an important part of the indexed ASF concerns well identified cities. So it confirms our hypothesis that the segmentation must be adapted to the type of the SF contained in the corpus.

Tables 6 and 7 also show that the city segmentation associated to the TFc gives better results than our baseline. Let's take example of Figure 1 to illustrate why we obtain such results. If we consider query "in Biarritz," the relevancy score for D2 on index1 is 1.0 because the text contains the SF "in Biarritz." It does not take into account the other SF. On the other hand, the city segmentation associated to the TFc gives a relevancy score of about 0.17. It computes a lower score to the document unit because it contains other less relevant SF.

MAP	TF	TF-IDF	Okapi	TFc
City Segmentation	0.61	0.61	0.63	**0.70**
Department Segmentation	0.40	0.39	0.40	**0.53**
Region Segmentation	0.40	0.39	0.39	**0.56**
Grid of 100x100	0.59	0.59	0.62	**0.68**
Grid of 200x200	0.61	0.60	0.63	**0.69**
Grid of 400x400	0.63	0.62	0.65	**0.66**

Table 6: Results of experiment on SF with different segmentations and weighting formulae

MAP	Spatial Overlapping
index1 (baseline)	0.62

Table 7: Results of experiment on SF with baseline (index1)

In conclusion, we advise segmentation into cities and the TFc formula (cf Table 6) for cultural heritage digital libraries. This normalization allows one to introduce an initial approximation of the spatial context (weighting a document unit takes into account all the SF it contains).

5 ONGOING AND FUTURE WORK

The PIV platform supports a similar processing sequence producing temporal indexes (Figure 3). It deals with calendar temporal features (CTF) that may be absolute (ACTF) or relative (RCTF) like spatial ones.

Let one consider that the previous text sample involves the following temporal features: CTF1-"the 26th of December", CTF2-"Saturday 29th of December at 2pm", CTF3-"at the beginning of the winter", CTF4"the last days of December 1933". The PIV produces such an index (Table 8, Figure 6).

The PIV temporal information normalization process (ongoing development similar to the spatial normalization process) would return weighted temporal intervals presented in Figure 7. This example illustrates calendar segmentation where each interval represents a week: TF4 intersects weeks W51 and W52. For example the week W52 has a weight of 4 according to the discrete indexing approach.

Currently, we are working on temporal normalization experiment. We aim to propose spatial and temporal criteria combination strategies with geographic IR scenarii.

id$_{ctf}$	text	type	timestamp
ctf1	the 26th of December	actf	(1933-12-26, 1933-12-26)
ctf2	Saturday 29th of December	actf	(1933-12-29, 1933-12-29)
ctf3	at the beginning of the winter	rctf	(1933-12-22, 1934-03-19)
ctf4	the last days of December 1933	rctf	(1933-12-21, 1933-12-31)

Table 8: Index of temporal features in PIV

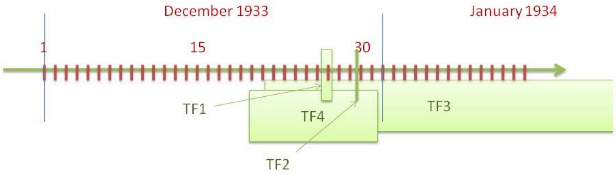


Figure 6: Temporal Index

6 CONCLUSION

The Virtual Itineraries in the Pyrenees (PIV) project consists in managing a repository of the electronic versions of books (histories, travelogues) from the 19th and 20th centuries. The PIV engine automatically annotates, interprets and indexes spatial, temporal and thematic information contained in those documents. Three independent process flows support spatial, temporal and thematic indexing and IR operations.

Two important problems were pointed out during a first campaign of experiments (Sallaberry et al., 2007): 1-results scoring does not integrate spatial or temporal features frequency within documents; 2-merging results within a geographic information retrieval process remains a challenge. The main problem of current geographic IR systems comes from the fact that the index structure and relevancy computation approaches used for space, time and theme are intrinsically different (Visser, 2004).

Our hypothesis is based on a spatial, temporal and thematic tiling of information in order to build higher level indexes and to adapt effective full-text IR technics to process such indexes.

In this paper we propose an approach for normalizing spatial indexes automatically. Such a gathering of spatial features into spatial tiles implies some loss of accuracy. However, as we have different grained indexes, we may select the best suited one during a querying stage. Moreover, experiments point out the effectiveness of our solution: a continuous spatial tiles frequency computation associated to a continuous document units relevancy computation formula gives better results than our baseline dedicated to the weighting of the most relevant SF of a document unit.

As explained before, the aim of this normalization method is to develop a general indexing strategy that is suited for spatial, temporal and thematic information in order to combine such geographic IR results. We are currently working on the evaluation of the effectiveness of this indexation on a larger sample of texts and queries. We are also working to apply normalization methods for the building of normalized temporal and thematic indexes from textual input. Future improvement of the presented approach would be to explore how to combine normalized spatial, temporal and thematic indexes and compute a unique relevancy scoring. Merging results for a geographic IR approach combining such different criteria is a recurring research question nowadays (Martins et al., 2008, Vaid et al., 2005).

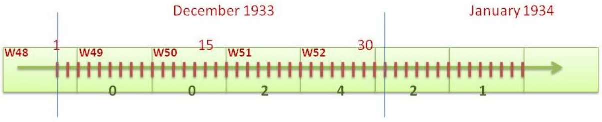


Figure 7: Calendar Segmentation

REFERENCES

Baccino, T. and Pynte, J., 1994. Spatial coding and discourse models during text reading. Language and Cognitive Processes 9, pp. 143–155.

Cai, G., 2002. GeoVSM: An Integrated Retrieval Model for Geographic Information. In: Max J. Egenhofer and David M. Mark (ed.), GIScience, Lecture Notes in Computer Science, Vol. 2478, Springer, pp. 65–79.

Clough, P., Joho, H. and Purves, R., 2006. Judging the Spatial Relevance of Documents for GIR. In: ECIR'06: Proceedings of the 28th European Conference on IR Research, Lecture Notes in Computer Science, Vol. 3936, Springer, pp. 548–552.

Egenhofer, M. J., 1991. Reasoning about Binary Topological Relations. In: Oliver Günther and Hans-Jörg Schek (ed.), SSD, Lecture Notes in Computer Science, Vol. 525, Springer, pp. 143–160.

Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaa, C. and Lesbegueries, J., 2008. A global process to access documents' contents from a geographical point of view. In: Journal of Visual Languages And Computing, Vol. 19number 1, Academic Press, Inc., Orlando, FL, USA, pp. 3–23.

Glander, T. and Döllner, J., 2007. Cell-based generalization of 3D building groups with outlier management. In: Hanan Samet and Cyrus Shahabi and Markus Schneider (ed.), GIS, ACM, p. 54.

Jones, C. B., Alani, H. and Tudhope, D., 2001. Geographical Information Retrieval with Ontologies of Place. In: D.R. Montello (ed.), Conference on Spatial Information Theory - (COSIT 2001), Vol. 2205 / 2001, Springer-Verlag Heidelberg, Morro Bayand California USA, pp. 322–335.

Jones, C. B. and Purves, R., 2006. GIR'05 2005 ACM workshop on geographical information retrieval. SIGIR Forum 40(1), pp. 34–37.

Kanhabua, N. and Nørvåg, K., 2008. Improving Temporal Language Models for Determining Time of Non-timestamped Documents. In: ECDL'08: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, Berlin, Heidelberg, pp. 358–370.

Le Parc-Lacayrelle, A., Gaio, M. and Sallaberry, C., 2007. La composante temps dans l'information géographique textuelle. Revue Document Numérique 10(2), pp. 129–148.

Li, H., Srihari, K. R., Niu, C. and Li, W., 2002. Location Normalization for Information Extraction. In: 19th International Conference on Computational Linguistics (COLING2002)- Howard International House and Academia Sinicaand Taipeiand Taiwan, Association for Computational Linguistics.

Mandl, T., Gey, F. C., Nunzio, G. M. D., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C. and Xie, X., 2007. GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Carol Peters and Valentin Jijkoun and Thomas Mandl and Henning Muller and Douglas W. Oard and Anselmo Penas and Vivien Petras and Diana Santos (ed.), CLEF, Lecture Notes in Computer Science, Vol. 5152, Springer, pp. 745–772.

Manning, C. D., Raghavan, P. and Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York.

Marquesuzaà, C., Etcheverry, P. and Lesbegueries, J., 2005. Exploiting Geospatial Markers to Explore and Resocialize Localized Documents. In: M. Andrea Rodríguez and Isabel F. Cruz and Max J. Egenhofer and Sergei Levashkin (ed.), GeoS, Lecture Notes in Computer Science, Vol. 3799, Springer, pp. 153–165.

Martins, B., Manguinhas, H. and Borbinha, J. L., 2008. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. In: ICSC, IEEE Computer Society, pp. 1–9.

Martins, B., Silva, M. J. and Andrade, L., 2005. Indexing and ranking in Geo-IR systems. In: GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval, ACM, New York, NY, USA, pp. 31–34.

Rees, T., 2003. "C-Squares", a New Spatial Indexing System and its Applicability to the Description of Oceanographic Datasets. In: Oceanography, Vol. 16number 1, pp. 11–19.

Robbins, S., Evans, A. C., Collins, D. L. and Whitesides, S., 2003. Tuning and Comparing Spatial Normalization Methods. In: Randy E. Ellis and Terry M. Peters (ed.), MICCAI (2), Lecture Notes in Computer Science, Vol. 2879, Springer, pp. 910–917.

Sallaberry, C., Baziz, M., Lesbegueries, J. and Gaio, M., 2007. Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study. In: ICEIS'07: Proceedings of the 9th International Conference on Enterprise Information Systems, pp. 190–197.

Salton, G. and McGill, M. J., 1983. Introduction to Modern Information Retrieval. McGraw-Hill.

Sautter, G., Böhm, K., Padberg, F. and Tichy, W. F., 2007. Empirical Evaluation of Semi-automated XML Annotation of Text Documents with the GoldenGATE Editor. In: ECDL'07: Proceedings of the 11th European Conference on Digital Libraries, LNCS, Vol. 4675, Springer, pp. 357–367.

Savoy, J., 2002. Morphologie et recherche d'information. Technical report, Institut interfacultaire d'informatique, Université de Neuchâtel.

Spärck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), pp. 11–21.

Vaid, S., Jones, C. B., Joho, H. and Sanderson, M., 2005. Spatio-textual Indexing for Geographical Search on the Web. In: Claudia Bauzer Medeiros and Max J. Egenhofer and Elisa Bertino (ed.), SSTD, Lecture Notes in Computer Science, Vol. 3633, Springer, pp. 218–235.

Visser, U., 2004. Intelligent Information Integration for the Semantic Web. Springer-Verlag, Heidelberg.

Zhang, Q., 2005. Road Network Generalization Based on Connection Analysis. In: Developments in Spatial Data Handling, Springer Berlin Heidelberg, pp. 343–353.

Zhou, S. and Jones, C. B., 2004. Shape-Aware Line Generalisation With Weighted Effective Area. In: Developments in Spatial Data Handling 11th International Symposium on Spatial Data Handling, Springer, Springer, pp. 369–380.

Zhou, X., Zhang, Y., Lu, S. and Chen, G., 2000. On Spatial Information Retrieval and Database Generalization. In: Kyoto International Conference on Digital Libraries, pp. 380–386.

DESIGN AND CONSTRUCTION OF A GIS-BASED DATABASE FOR MANAGING LANDSLIDES IN MINING AREA

Shanshan Wang [a,] *, Min Ji [a], Xiangwei Zhao [a]

[a] Geomatics College, Shandong University of Science and Technology, 579 Qianwangang Road, Economic & Technical Development Zone, Qingdao, China,266510-wagshan@126.com, (jimin, zhaoxwchina)@sdust.edu.cn

KEY WORDS: Mining Area Landslides, Spatial Data, Entity-Relation Model, Relation Model, Equalization, Uniform Coding

ABSTRACT:

Landslides caused by mining are seriously affecting the construction and sustainable development of mining area. With the characteristics of multi-resource, multi-format, multi-theme and large storage capacity, mining landslide data is difficult to be managed effectively and applied to the evaluation, forecast, prevention and treatment of landsides. Based on mechanism of landslides in mining areas and GIS technology, we analyzed the data in mining area landslide database on different aspects. Landslides and other related objects in mining area were abstracted to relational models by conceptual design, logical design and physical design. A new unique coding method was present according to current specifications and the features of mining areas. Then, we built a GIS-based mining area landslide database using Oracle and ArcSDE. Through program implementation and an experiment, the results shows that the database with rich content is well-structured, satisfying the users who manage mining area landslide information or collect data in the field.

1. INTRODUCTION

The acceleration of modernization in China increases mining in recent decades. Under the influence of open-pit mining, underground mining or other engineering construction, landslide hazard in some mining area becomes increasingly aggravated, making a great amount of landslide data accumulated. In the traditional data management, the data acquisition and updating cycle is long, making the data can not be timely and effectively transformed into information and be shared. Meanwhile, the development of computer technology and spatial technology make it possible to get more and more mining area landslide data of multi-resource, multi-format, multi-theme and large storage capacity. How to manage the data effectively and apply them to the evaluation, forecast, prevention and treatment of landsides in mining area has become a bottleneck to deal with.

GIS started from 1960s, has been applied to the research of landslides for its advantages on data management, visual representation, spatial analysis, virtual reality and integration with decision support system. According to Xie (Xie, 2003), a 3D landslide assessment model was presented and a Grid-based landslide assessment system named 3DSLOPEGIS was developed based on a landslide spatial database. Ulrich Kamp et al developed a spatial database using ASTER satellite imagery and GIS technology to analyze the relationship between earthquakes and landslides in Kashmir earthquake region (Ulrich, 2005). A multi-method approach for the assessment of the stability of natural slopes and landslide hazard mapping was applied to the Dakar coastal region (Fall, 2006). Pece V. Gorsevski et al developed a Spatial-Temporal database model by grid tools of GIS to represent the uncertainty and variability of parameters which caused the landside in Pete King area. (Gorsevski, 2006).

GIS is playing an important role in landslide assessment and forecast, and the construction of GIS-based landslide database is the most foundational part. In this paper we shall discuss how to design and construct a GIS-based mining area landslide database to support data acquisition, information management, mapping and hazard analysis of mining area landslides.

2. GENERAL PLAN OF DATABASE CONSTRUCTION

In order to accelerate the informatization of mining area landslide management, users of the GIS-based mining area landslide database are people who collect landslide data in the field with GPS and PDA, who operate landslide data by desktop applications, who evaluate and forecast landslide stability and who view landslide data comprehensively and make decisions in an effective way.

According to specific demands of different users, we abstracted mining area landslides from the real world to the information world and the machine world by the optimal data model, and turned the objects to operational data. Data in the database should have small redundancy and stable structures during operation, can be shared, expanded and refreshed by users and can be independent from application programs (Li, 2007). Aiming at this target, we evaluated and optimized the database by physical implementation and application running. As shown in Figure 1, owing to abundant spatial data in the database, the theories and technologies of GIS were used during almost the whole process of design and construction of the database, which included data analysis, conceptual design, logical design, physical design, physical implementation and application running.

* Corresponding author. E-mail address: wagshan@126.com; Tel.:13646428546

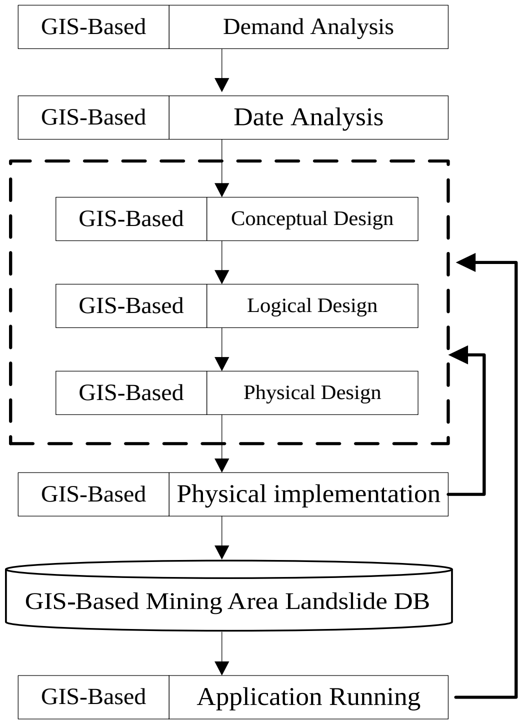


Figure 1. The general plan of database construction

3. DATA ANALYSIS

3.1 Data Content

Data is the blood of a database system, having its cost accounting for 80% of the total cost. It is necessary to analyze the data related to mining area landslides comprehensively and scientifically on the basis of landslide mechanism and the features of mining areas so as to abstract mining area landslides in a reasonable way.

Hazard is the result of a combination of factors, which can be regrouped into breeding environment, inducing factors and bearing body (Song, 2008). For mining area landslides, breeding environment which is quasi-static, contributes to landslide susceptibility and contains foundational geographical environment (terrain, landscape, hydrology，vegetation，etc.) and various geological environment (stratum, geological structure, under water, mechanical parameters for rock and soil, etc.) of the mining area where landslides happen or will happen (Dai, 2002). Inducing factors which reflect the basis of material and energy of landslides, are dynamic variables such as rainfall, earthquakes and mining, tending to trigger landslides in mining areas of a given landslide susceptibility. Bearing body, which reflects vulnerability of the mining area once landslides happen, contains objects that breeding environment and inducing factors act on. Since the above-mentioned factors are various in different mining areas, we should collect landslide data required as comprehensive as possible according to the features of mining area to express landslides sufficiently and satisfy users' daily operation and analysis. Figure 2 shows the data content in mining area landslide database.

In terms of data format, these data can be divided into vector data, raster data, text and other multimedia data. Vector data, that is digital line graph (DLG), reflects the distribution of mining area landslides or their effect factors by points, lines or polygons. Raster data consists of multi-source remote sensing images, digital raster graphs (DRG) and digital terrain models (DTM). Text data is written records accumulated by mining area landslide information collecting of relevant departments. In addition to the above three data, multimedia data also include pictures (photos of mining areas where landslides happen, profile plans of landslides, result charts of landslide stability analysis, etc.) , audios(hazard reports, warning notices, etc.) and videos (animation of landslides, etc.). Vector data and raster data constitute spatial data in mining area landslide database, while text data and other multimedia data constitute attribute data.

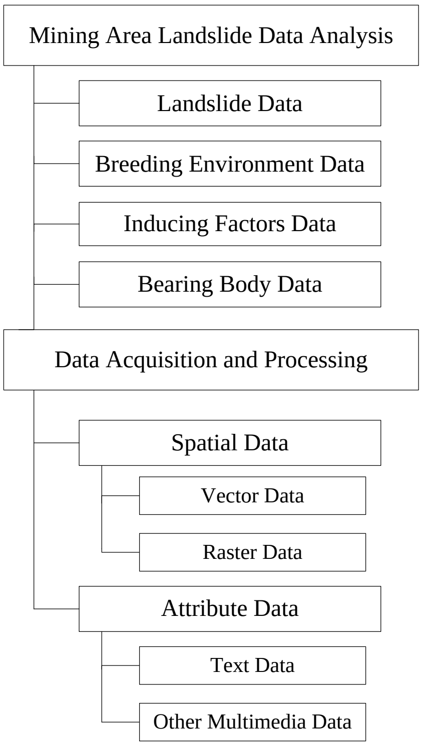


Figure 2. The data content in mining area landslide database

3.2 Data Acquisition and Processing:

We acquired and processed mining area landslide data in different ways according to their different data format.

Vector data was acquired and processed by the following ways:
(ⅰ) scanning and vectoring the accumulated maps of mining area landslides and their effect factors, among which error checking, topology processing, sheet splicing and projection transformation were important;
(ⅱ) creating and editing features on existing layers according to the result of mining area landslide field collecting based on GPS and PDA;
(ⅲ)accepting coordination values of GPS- equipped control stations which were set up on landslides and then representing them on a layer, so as to simulate the motion of landslides dynamically and even forecast the trend of landslides.

Multi-source remote sensing images were acquired by various receivers and stored into database after a series of image pre-processing, such as correction, registering, mosaic, clipping and so on. The DRG were raster-formed graphs of paper maps. DTM were obtained from data of other formats by format conversion, spatial analysis, 3D Analysis and other GIS tools. Some software, such as ArcGIS, Envi, Erdas, were used during spatial data processing.

We get attribute data not only from collecting and entering historical statistics of mining area landslides, but also from creating and editing records on basis of field collecting and landslide analysis.

3.3 Data Organization:

In view of so much spatial data in mining area landslide database, we adopted the top-down physical structure–"Project->Base->File->Layer->Feature" in the database combining the data management concept of GIS. Figure 3 shows the physical structure we adopted.

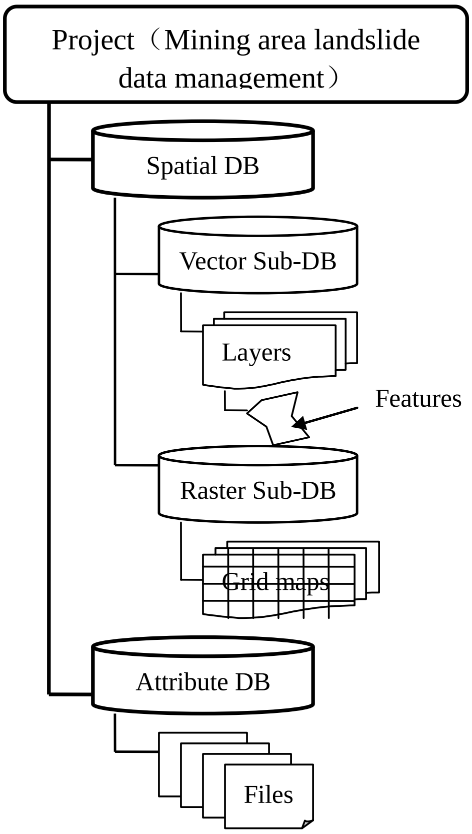


Figure 3. Data organization of mining area landslide database

As shown in Figure 3, a "project", which is the most top-level object of landslide information management in a certain mining area, is a collection of all kinds of data (Zhang, 2001). That means only one project can be built for a mining area. A project contains several bases of different types. Under the control of "project", a "base" is a collection of "files" because of their same data format or logic applications. Hierarchy is a property of bases, that is, a base can consist of several sub-bases. According to different data formats, the mining area landslide database can be divided into spatial database and attribute database, while the spatial database can be divided into vector database and raster database.

A "file", which can be a picture, a text document, a layer or a remote sensing image, is the fundamental unit of data to be managed by applications. "Layer" and "feature" are intended for vector data. In vector database, the data range of a "layer" is the same with that of a "file". Layers express the distribution of landslides and their factors. A "feature" is the smallest data unit, expressing an object by a point, a line or a polygon.

4. DATABASE DESIGN

4.1 Conceptual Design

Conceptual design, which is often considered as the key to the success of a database, abstracts research objects from the real world to the information world. In the conceptual design of mining area landslide database, we synthesized and described mining area landslides and other objects by standing on different users so as to meet their demands.

E-R (Entity-Relation) method, which is often used in conceptual design, abstracted a collection of landslides or other objects to an "entity", a relationship between different objects to a "relation" and a feature of objects to a "property". The conceptual design of mining area landslide database using E-R method included the following two steps:
(ⅰ) designing partial E-R models. Figure 4 shows the partial E-R model of mining area landslides, control stations and the relationship between them.
(ⅱ) synthesizing all the partial E-R models to an overall E-R model.

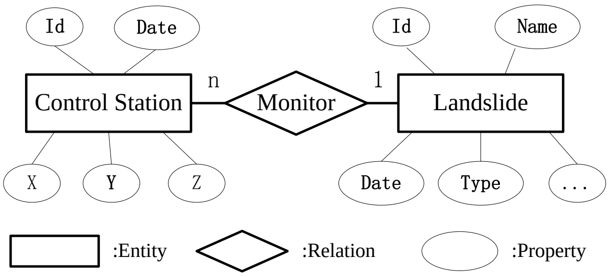


Figure 4. The partial E-R model of mining area landslides and control stations

4.2 Logical Design

At present, the common data models in the field of database are network model, hierarchical model, relational model and object-oriented model. Among the characteristics of simple and flexible structures, editing and updating data conveniently and easy to be maintained, relational model owes its greatest advantage to consistency of description. Relational model is not only the most commonly used data model in database, but also an effective data organization approach to build relationships between spatial data and attribute data. Most of GIS attribute data are organized in relational data models, and even some systems adopt relational database management system to manage spatial data (Zhang, 2007). Therefore, we adopted relational mode in the GIS-based mining area landslide database.

4.2.1 Relational Model Deriving: Since the mining landslide E-R models were independent of any specific data mode, we turned them into logical structures equivalent to relational models firstly. An entity or relation was expressed as a relation table, and their properties were expressed as attributes of relational tables, namely “fields”. As a result, mining area landslides and other objects were abstracted to the machine world.

The relational model corresponding to the partial E-R model of mining area landslides and control stations are:
Landslide: (LSId, LSName, LSDate, LSTye,…)
Control Stations: (CSId, CSDate, CS-x, CS-y, CS-z)
Monitor: (CSId, LSId)

4.2.2 Equalization of Relational Models: The initial relational models we derived seems disordered and confused for the complicated relation between mining area landslides and their breeding environment, inducing factors and bearing body. For example, properties of mining area landslides alone could be expressed as lots of fields in the landslide attribute relational table. Aiming at this issue, we adjusted the relational models by dividing every initial relational table into a group of simpler and more stable relational tables when it was necessary according to the different retrieval frequencies, logical relations and people’s attention of various properties. Then, more logical and orderly relational models were obtained. The process we adjusted the relational models is called equalization of a relational models. It makes a lot of sense to efficiency and security of database systems. Figure 5 shows the relational tables of mining area landslides after equalization.

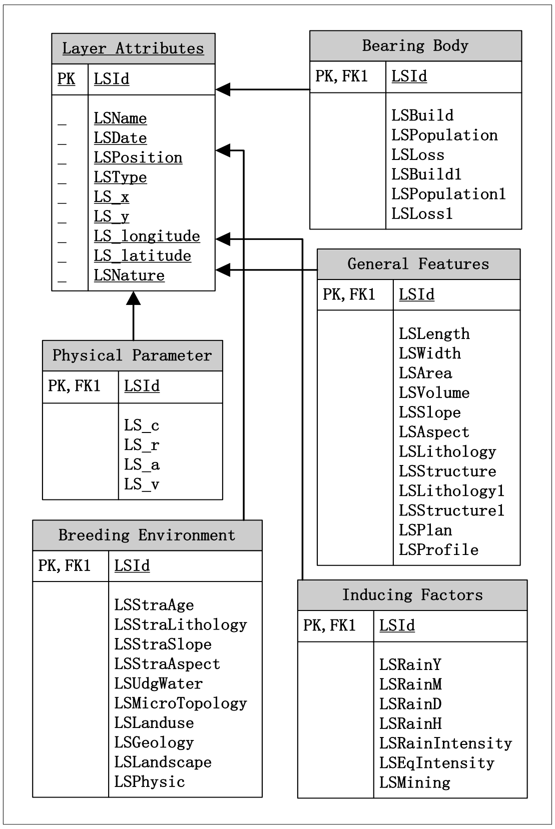


Figure 5. Equalization of the relational model of mining area landslides

As shown in Figure 5, because people care more about basic properties of mining area landslides (name, time of happening, landslide type, and position, etc.), we stored them in the inner layer attribute table, the attributes of which can be retrieved more quickly and conveniently. General features, breeding environment, inducing factors, bearing body and physical parameters stored general properties of mining area landslides in external tables. They were regrouped by the different logical relations between mining area landslides and other objects. The table named general features stored attributes reflecting the development of landslides. The table named physical parameters stored mechanical parameters of stratum which are often retrieved and used in landslide stability assessment on the basis of mechanics (Wu, 2006). Among all the attribute tables of mining area landslides, the table named layer attributes takes the uniform code (LSId) as principal key to define the existence and uniqueness of a mining area landslide, while other tables take it as foreign key. Then, all the attributes established connections with the feature in the mining area layer.

4.2.3 Uniform Coding and Indexing: In order to realize data sharing and improve retrieval efficiency, we created indexes in the mining area landslide database. As the premise, uniform coding for each data organization structure were considered inevitably.

Actually, there are several uniform coding specifications for geological hazard spatial database at hand. But all of them aren’t customized specially for mining area, and cannot organize and manage all the data of a mining area in a big collection. In additional, by these specifications, the spatial characteristics of spatial data and attachment relationships of different data organizations cannot be reflected. Therefore, by referring to and modifying the specifications, we finished uniform coding for mining area landslide database, containing semantic consistency and rich descriptions. Taking “#” as an Arabic numeral and “□” as a letter, uniform coding for “project”, “layer”, “feature” will be shown in Figure 6-8.

(ⅰ)As the top-object in the database, a project takes its Mine Number which is recorded in *Mineral Reserves Registration Statement* as its uniform code. The Mine Number consists of 9 digits. The first 6 digits is the code for the administrative division where the mine located and can be determined on the basis of *GBPT2260 – 2002*, and the last three is the mine’s sequence number in its administrative division. The structure of the uniform code of a project is shown in Figure 6.

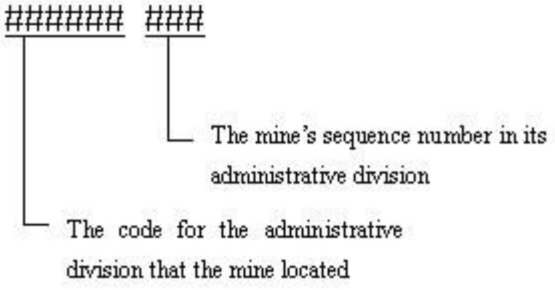

###
The mine’s sequence number in its administrative division
The code for the administrative division that the mine located

Figure 6. Uniform code of a project

(ⅱ) The uniform code of a layer is made up of the uniform code of the project the layer attached to and its own codes. The structure of the uniform code of a layer is shown in Figure 7.

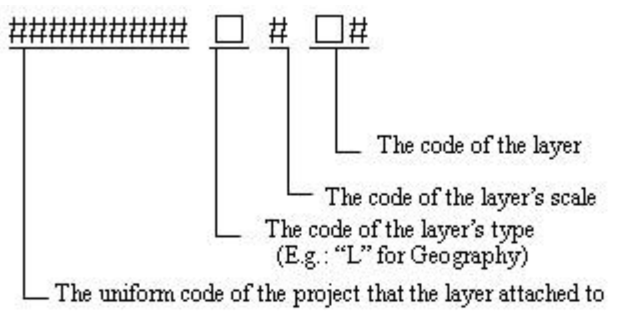

######### □ # □#
The code of the layer
The code of the layer's scale
The code of the layer's type
(E.g.: "L" for Geography)
The uniform code of the project that the layer attached to

Figure 7. Uniform code of a layer

The codes of the layer's type and the layer are determined on the basis of *GB/T13923-92*, which is customized for geological hazard spatial database and accepted by lots of people. The available codes of the layer's scale are: "A"-1:1000000, "B"-1:500000, "C"-1:250000, "D"-1:100000，"E"-1:50000，"F"-1:25000，"G"-1:10000，"R"-1:200000.

(ⅲ) The uniform code of a feature is made up of the uniform code of the layer that the feature attached to and its own code, which is a 4 digits generated by its entering order to the database. The structure of the uniform code of a feature is shown in Figure 8.

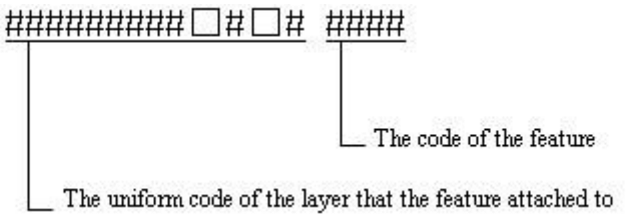


Figure 8. Uniform code of a feature

4.3 Physical Design

Since relational model was adopted in mining area landslide database, a relational database system should be chosen to manage data. We chose Oracle as mining area landslide data management platform considering its characteristics including: (ⅰ) supporting high-performance multi-user transaction; (ⅱ) supporting mass multi-media data, such as binary graphics, audios, videos and multidimensional data; (ⅲ) security and integrity; (ⅳ) supporting distributed-database and distributed-transaction ; (ⅴ) portability, compatibility and connect ability.

In the case of spatial data management, ArcSDE, which is a middleware between application and relational database system, can act as the server to access multi-user geodatabase stored in relational database system and offer an open interface. Considering ArcSDE's advantages of mass data storage, multi-user concurrent access, version management, long transaction, we chose ArcSDE as spatial data management engine in mining area landslide database, storing and managing spatial data and attribute data uniformly and efficiently.

5. DATABASE IMPLEMENTATION AND RUNNING

The implementation of mining area landslide database can be divided into two sections: data loading and programs writing and debugging (Wang, 2004). According to the logical structure of the mining area landslide, we built relational tables in Oracle, and realized the links between the tables as well as the links between attribute records and features by setting up primary keys and foreign keys. In order to facilitate data input, output and management, a mining area landslide database management system was developed by ArcEngine and C#. In addition, a GPS-base mining area landslide field data collecting system running on PDA，through which we can create and edit data in the database, was developed by ArcGIS Server and C#. We carried out an experiment in Yanzhou mining area which is located in Jining, Shandong Province, China. The result was ideal, and showed that the mining area landslide database we designed and constructed met the users who collected data in the field.

6. CONCLUSION

In this paper, the process we designed and constructed the mining area landslide database was discussed. We analyzed the data of the mining area landslide database on aspects of content, acquisition and organization structure on the basis of landside mechanism and GIS technology. By database design, landslides and other objects in mining areas were abstracted from the real world to the machine world and reasonable relational models were obtained. A unique coding method for mining area landslide was present according to current specifications and the features of mining areas. We built the mining area database by Oracle and ArcSDE. The database we designed and constructed was proved to be feasible and satisfy the users who manage mining area landslide information and collect data in the field by application running. In the following work, we will focus on the research and implementation of mining area landslide stability assessment model so as to adjust and maintain the database and satisfy the users who analyze and make decisions on mining area landslides.

ACKNOWLEDGMENT

This work is supported by Chinese National Programs for High Technology Research and Development (2009AA12Z147) and Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping (2009B14). Thanks to Mr. Tengfei Yu, Miss. Lina Zhao and Mr. Hongtao Wang for their helpful comments.

REFERENCES

Dai, F., 2002. Landslide risk assessment and management: an overview. *Engineering Geology*, 64 (1), pp. 65– 87.

Fall, M., 2006. A multi-method approach to study the stability of natural slopes and landslide susceptibility mapping. *Engineering Geology*, 82, pp. 241– 263.

Gorsevski, P.V., 2006. Spatially and temporally distributed modeling of landslide susceptibility. *Geomorphology*, 80, pp. 178－198

He, R., 2008. The Establishment and Application of the Spatial database on the Geological Hazard in NeiKun Railway based on GIS. Master thesis, Southwest Jiaotong University, Chengdu, Sichuan, China.

Hu Y., 2008. Construction of database for earthquake emergency rescue based on ArcSDE. *Journal of Catastrophology*y, 23(2), pp.132-134.

Li Y., 2007. Design and construction of geological hazard database in Chongqing City. *The Chinese Journal of Geological Hazard and Control*,18(1),pp.115-119.

Song, C., 2008. An Approach to Risk Assessment of Mine Collapse. Master thesis, Chinese Aeademy of Geological Sciences, Beijing, China.

Ulrich, K., 2005. GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region. *Geomorphology*, 101, pp. 631–642.

Wang, D., 2004. Designing and planning of the geological hazard information management system supported by the GIS. Master thesis, Central South University, Changsha, Hunan, China.

Wu, J., 2006. *Engineering Geology*. Higher Education Press, Beijing, China, pp. 253-265

Xie, M., 2003. GIS component based 3D landslide hazard assessment system:3DSLOPEGIS. *Chinese Geographic Science*, 13(1), pp. 66-72.

Zhang, Y., 2001. *The Design and Development Of Geographical Hazard Information System*. Geology Publishing House, Beijing, China, pp. 12-16.

Zhang, S., 2007. Design of the spatial database of the mine data based on Oracle Spatial. *Geospatial Information*, 5(2), pp. 91-93.

EXPLORATORY SPATIAL DATA ANALYSIS OF REGIONAL ECONOMIC DISPARITIES IN BEIJING DURING 2001-2007

Xiaoyi Ma *, Tao Pei

State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, 100101 Beijing, China - (maxy, peit)@lreis.ac.cn

KEY WORDS: Regional Economy, The Olympic Games, Exploratory Spatial Data Analysis, GIS, Spatial Autocorrelation, Spatial Heterogeneity

ABSTRACT:

The unbalanced development of regional economics in Beijing has induced the ubiquitous regional economic disparities and may affect the sustainable development of economy and social stability. How to alleviate this unbalanced status has been an important issue which the society and government face. As usual, the hold of Olympics Games may boost the economy of host city, however it is not clear that the event could narrow the regional economic disparities. In order to investigate the influence of the 2008 Olympics Games for the development of regional economics in Beijing, this paper studies the space-time dynamics of economic development of Beijing since the year when Beijing won the bid (the year of 2001) using exploratory spatial data analysis (ESDA), and based on the county level and measure index of the regional per capita gross domestic product (GDP). The results show not strong evidences of global spatial autocorrelation, but clear evidences of local spatial autocorrelation and spatial heterogeneity in the distribution of regional per capita GDP. Driven by the bid winning of 2008 Olympics, the total economic disparity in Beijing had not been alleviated, and the situation got even more complicated. Since the economic increasing-speeds of Changping and Shijingshan Districts were significantly lower than their some neighbouring regions, a new centre-surrounding polarization scheme was gradually replacing the North-South polarization scheme in Beijing from 2001 to 2007.

1. INTRODUCTION

The unique history foundation and location condition cause the unbalanced development of regional economics in Beijing. This unbalanced status induces the ubiquitous regional economic disparities and may affect the sustainable development of economy and social stability. How to alleviate this unbalanced status has been an important issue which the society and government face. As history had proved that, the hold of Olympics may boost the economy of host city. The 2008 Olympic Games provided Beijing with an opportunity to successfully promote the economic development and accelerate the completion of major infrastructure upgrades in public service industry, transportation and other sectors. According to the research report of Beijing Municipal Statistic Bureau, the total investment for the 2008 Olympic Games from 2002 to 2007 was about 351 hundred million dollars, which efficiently promoted the growth in GDP of Beijing. In 2006, the Beijing’s per capita gross regional product reached 6,210 dollars exceeding the target of 6000 dollars at the end of 2008, two years ahead of previously predicted. However, whether or not the Olympic Games could lessen the regional economic disparities is unclear.

In order to investigate the influence of the 2008 Olympic Games for the development of regional economic disparities in Beijing, this paper studies the space-time dynamics of economic development of Beijing since the year 2001 when Beijing won the bid, using exploratory spatial data analysis (ESDA). ESDA emphasizes the significance of spatial interactions and geographical location in the studies of regional economic development. By identifying spatial autocorrelation and spatial heterogeneity, the economic performance can be characterized according to time. Therefore, ESDA is a powerful tool for revealing the development of regional economic disparities. Several previous studied have been implemented on this issue focusing on the EU regions. Le Gallo and Ertur (2003), López - Bazo et al. (2004) applied ESDA to study the distribution of regional per capita GDP in Europe. Ertur and Koch (2006) investigated the case affected by the enlargement of the European Union from 1995 to 2000. In addition, there are some studies have assessed the regional economic disparities using ESDA in China, including the work of Pu et al. (2005) on the Jiangsu province and Qiu et al. (2009) on the Huaihai economic zone of China. However, few of them revealed the space-time dynamics of regional economics caused by important historical events or national policy guides in China.

Consequently, this study, combining ESDA with GIS technology, investigates the development of regional economics after 2001, and attempts to explore the possibility that important historical events or national policy guides may associate with change in spatial patterns of regional economic disparities over time. Our method is based on the county level and measure index of the regional per capita gross domestic product (GDP).

2. MATHODOLOGY AND DATA

2.1 Exploratory spatial data analysis (ESDA)

Exploratory spatial data analysis (ESDA) is a set of techniques to describe and visualize spatial distributions, discover patterns of spatial association (spatial clustering or hot spots), identify atypical observations (outliers), and suggest different spatial regimes or other forms of spatial heterogeneity (Anselin, 1994;

* Corresponding author. Email: maxy@lreis.ac.cn.

1996; 1999). Central to ESDA is the measure of global and local spatial autocorrelation.

The measure of global spatial autocorrelation is usually based on Moran's ***I*** statistic, expressed as:

$$\boldsymbol{I} = \frac{\boldsymbol{n}}{\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{w}_{ij}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})}{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{w}_{ij}} \qquad (1)$$

where ***n*** is the number of regions in the study area, $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are the attribute values for region ***i*** and ***j***, $\overline{\boldsymbol{x}}$ is the mean of the attribute value for the ***n*** regions, and $\boldsymbol{w_{ij}}$ is the spatial weight matrix, which is usually defined as a bivariate symmetric matrix ***W*** to express the spatial contiguity relationship of regions. This paper adopts the spatial weight matrix based on the distance between the centroids of regions, that is, the corresponding element $\boldsymbol{w_{ij}}$ =1 if the distance between regions ***i*** and ***j*** is within the threshold distance and zero otherwise.

The Moran's ***I*** statistic is based on the measure of covariance $(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})$, which is used to assess the similarity of specified attribute values of variable ***X*** (Getis and Ord, 1996). The inference of significance can be based on the permutation approach (Anselin, 1995).

Moran's I statistic can be visualized as the slope in the Moran scatter plot (Anselin, 1995), which plots the spatially lagged variable (***Wy***) on the original variable (***y***), here the ***y*** are in deviations from their mean. The four different quadrants of the Moran scatter plot represent the four types of local spatial association respectively. The upper right (HH) and the lower left (LL) quadrants represent positive spatial autocorrelation, that is, a region with a high (low) value surrounded by regions also with high (low) values. The upper left (LH) and lower right (HL) quadrants correspond to negative spatial autocorrelation, i.e., a region is surrounded by regions with dissimilar values.

As one of the local spatial autocorrelation tools, the Moran scatter plot can identify local spatial clustering with high or low values, and indicate what patterns of spatial association make more contribution for the results of global spatial autocorrelation statistic, and visualize atypical observations. It makes up the major shortcoming of global spatial autocorrelation statistic, which can only indicate overall clustering. However, the Moran scatter plot does not indicate significance.

Local indicators of spatial autocorrelation (LISA) defined by Anselin (1995) provides a way to assess significant local spatial patterns, besides it can be used similarly to the Moran scatter plot.

The local Moran's statistics is a familiar LISA formulated as the follows:

$$\boldsymbol{I}_i = \frac{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})}{\boldsymbol{m}_0}\sum_j \boldsymbol{w}_{ij}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) \quad \text{with} \quad \boldsymbol{m}_0 = \sum_i \frac{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^2}{\boldsymbol{n}} \qquad (2)$$

where ***n*** is the number of regions in the study area, $\boldsymbol{x_i}$ is the attribute values for region ***i***, $\overline{\boldsymbol{x}}$ is the mean of the attribute value for the ***n*** regions, and $\boldsymbol{w_{ij}}$ is the spatial weight matrix. The inference of significance is based on the permutation approach (Anselin, 1995). The results of the local Moran's statistics can be classified into four types of spatial association, and visualized in a LISA cluster map (Anselin, 1995).

2.2 Data and Background

The data for exploratory spatial data analysis has two types:

1. Regional per capita GDP and vital statistics data over the period 2001–2007 extracted from the Beijing area statistical yearbooks.
2. The administrative map of Beijing got from the administrative map of China with the scale of 1:1,000, 000 as shown in Figure 1.

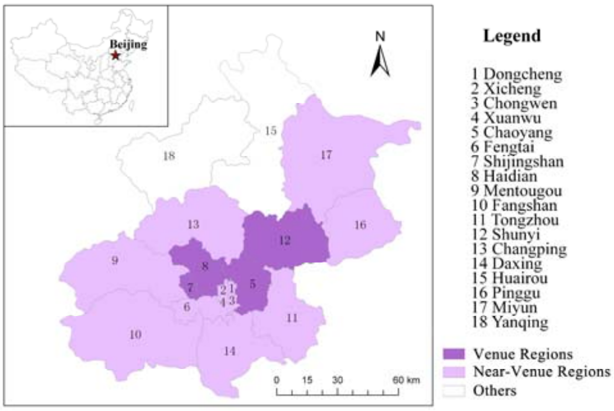


Figure 1. The administrative map of Beijing, China

The venues of Beijing Olympic were mainly distributed into four areas, one central area and three partitions. The central area is located in the Olympic park in Chaoyang district, and the three partitions are the university area, the western community, the northern scenic tourist area, located in Haidian, Shijingshan, and Shunyi respectively. Figure 1 depicts the administrative map of Beijing with the purple area as venue regions and the light purple areas as the regions adjacent to venue regions according to the spatial weight matrix.

The theory is that venue and near-venue regions would be positively impacted by the presence of the Olympic Games. The per capita GDP growth of near-venue regions might be more than those regions not close to Olympic events. Therefore, some changes in spatial patterns of regional economic disparities would be expected to be observed due to the influence of the Olympic Games. However, because of the neglect of spatial interactions and geographical location, the expected effects of the Olympic Games could be overestimated or underestimated by using the traditional statistics analysis. Therefore, in the next section, ESDA is adopted to investigate the influence of the Olympic Games for the development of regional economic, and reveal the spatial pattern of regional disparities in Beijing.

3. RESULTS AND ANALYSIS

In this paper, both global and local spatial autocorrelation analyses were carried out by using the GeoDa software package

(Anselin et al., 2006). Figure 2 shows the global Moran's ***I*** statistic for per capita GDP from 2001 to 2007 using the spatial weight matrix with the threshold distance about 36,000 meters. The values of global Moran's ***I*** statistic show a downward trend from 2001 to 2007 and no significant positive or negative spatial autocorrelation, except for 2002. The results suggest that the locations of regions with high or low per capita GDP are almost random, and the spatial pattern has a developing trend of negative spatial autocorrelation since 2005.

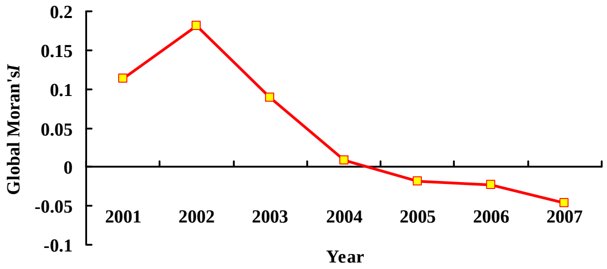


Figure 2. The chart of Moran's I statistics for per capita GDP of Beijing's eighteen districts or county, 2001-2007

The Moran's ***I*** is a global statistic, and gives a single result to assess the spatial association pattern for the entire study area. Therefore, it cannot identify the regions contributing more to the results of global spatial autocorrelation, and detect the hot spot or atypical localizations. In order to solve these problems, the Moran scatter plot and LISA are adopted.

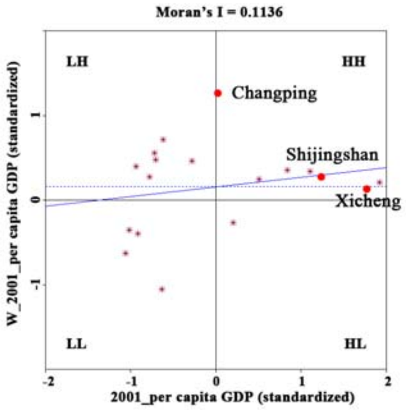


Figure 3. Moran scatter plot for per capita GDP, 2001

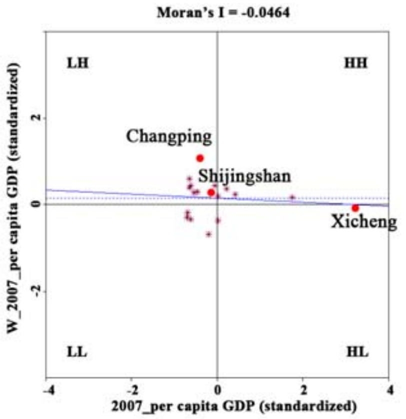


Figure 4. Moran scatter plot for per capita GDP, 2007

Figure 3 and 4 are the Moran scatter plots of per capita GDP for the initial year 2001 and final year 2007, respectively. It can be seen that 11 regions were characterized by positive spatial association (7 regions in quadrant HH, 4 in LL) and 7 regions were characterized by negative spatial association (6 regions in quadrant LH, 1 in HL) in 2001. The regions of local spatial clusters with high values contributed more to the global spatial autocorrelation. However, 10 regions presented negative spatial association (8 regions in quadrant LH, 2 in HL) in 2007, which dominated the global spatial autocorrelation. The main change of spatial pattern between the two years concerned three regions. In 2001, Xicheng, Changping and Shijingshan districts had similar high values of per capita GDP to their neighbours' and they were in quadrant HH. However in 2007, Xicheng located in quadrant HL indicating its higher level of economic development than their neighbours'. Contrarily, Changping and Shijingshan moved to the quadrant LH, which suggested their lower level of economic development comparing to their neighbours'.

The analysis of Moran scatterplots seems to be contradictory to the theory. Shijingshan District, as one of the venue regions, should have positive impact on the economic development. However, we can observe that it was relatively poorer than surrounding regions in 2007. Moreover, the other venue and near-venue regions have not obvious change in spatial pattern.

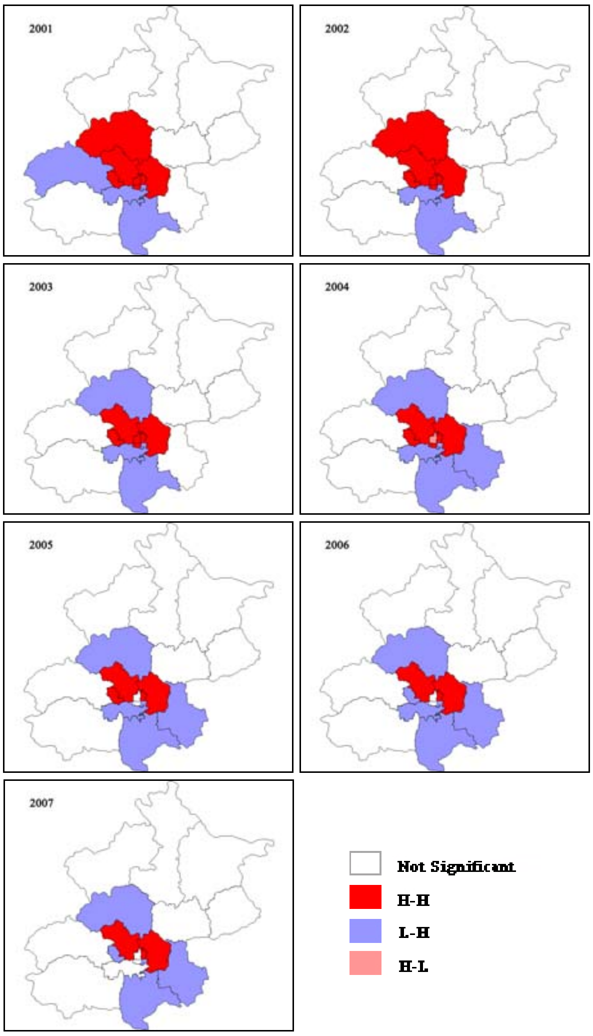


Figure 5. The LISA cluster map for per capita GDP of Beijing's

Figure 5 is the LISA cluster map for per capita GDP from 2001 to 2007. The colored regions are significant with $\boldsymbol{p}$ = 0.05 for each year. The map indicated the two main patterns of spatial association, that is, clusters of high values surrounded by high values (HH) and clusters of low values surrounded by high values (LH). Since 2003, Changping and Shijingshan had moved one after the other from HH to LH region, which caused a new centre-surrounded polarization scheme to gradually replace the North-South polarization scheme.

It is worth stressing the following interesting result: Although the Shijingshan is one of the venue regions, and Changping district is adjacent to both Haidian and Chaoyang (two venue regions), they belong significant to the LH quadrant lagging behind surrounded regions. It suggests the 2008 Olympic Games did not benefit the economic development of Shijingshan, and promote the efficient growth in per capita GDP of Changping district even adjacent to venue regions.

This result could be explained that the bid winning of 2008 Olympics accelerated the proceedings of Suburbanization of Residence in Beijing. The large bedroom communities such as Tiantongyuan and Huilongguan were built in Changping district, which led to the sharp increase of population migration from the city centre to Changping district. Although the 2008 Olympic Games promoted the growth in GDP of Changping district, its per capita GDP is much lower than neighbours' due to the increase of population. In addition, Shijingshan district also moved from the quadrant HH to the quadrant LH following Changping in 2006. An efficient cause should be the policy-based limiting production of the Capital Iron and Steel Company. Due to the 2008 Olympic Games, Beijing accelerated the improvement of environment pollution by decreasing the production of the Capital Iron and Steel Company (Wan, 2006), which led to the lower GDP increasing in Shijingshan than some surrounding areas.

4. CONCLUSIONS

This paper attempted to investigate the influence of the 2008 Olympic Games for the development of regional economic disparities in Beijing since the year 2001 when Beijing won the bid. By combining ESDA with GIS technology, the spatial association patterns of regional per capita GDP were analyzed from 2001 to 2007.

The results showed not strong evidences of global spatial autocorrelation, but clear evidences of local spatial autocorrelation and spatial heterogeneity in the distribution of regional per capita GDP. Although the 2008 Olympic Games boost the economic development of Beijing, the total economic disparity in Beijing had not been alleviated.

Moreover, the results of the Moran scatterplot and the LISA highlight that a new centre-surrounding polarization scheme was gradually replacing the North-South polarization scheme in Beijing from 2001 to 2007.

ACKNOWLEDGEMENTS

This study was funded through support from a grant (Project Number: 2006AA120106) from the High-Tech Research and Development Programme of China, a grant from the State Key Laboratory of Resource and Environment Information System (Project Number: 088RA400SA).

REFERENCES

Anselin, L., 1994. Exploratory spatial data analysis and geographic information systems. *New Tools for Spatial Analysis*. Eurostat, Luxembourg, pp. 45–54.

Anselin, L., 1995. Local indicators of spatial association—LISA. *Geographical Analysis*, 27, pp.93–115.

Anselin, L., 1996. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *Spatial Analytical Perspectives on GIS*. Taylor & Francis, London, pp. 111–125.

Anselin, L., 1999. Interactive techniques and exploratory spatial data analysis. *Geographical Information Systems: Principles, Techniques, Management and Applications*. Wiley, New York, pp. 251–264.

Anselin, L., Syabri I., Kho Y., 2006. GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*, 38, pp. 5－22.

Ertur C., Koch W., 2006. Regional disparities in the European Union and the enlargement process: an exploratory spatial data analysis, 1995-2000. *Annals of Regional Science*, 40, pp. 723-765.

Getis A., Ord J., 1996. Local spatial statistics: an overview. *Spatial analysis: modeling in a GIS environment*. Cambridge: Geoinformation International, pp. 261-277.

Le Gallo J., Ertur C., 2003. Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980–1995. *Regional Science*, 82, pp. 175–201.

López-Bazo E, Vayá E, Artís M., 2004. Regional externalities and growth: evidence from European regions. *Regional Science*, 44, pp. 43–73.

Pu Y., Ge Y., Ma R., et al., 2005. Analyzing regional economic disparities based on ESDA. *Geographical Research*, 24(6), pp. 965－974.

Qiu F., Zhu C., Tong L., et al., 2009. Spatial analysis of economic disparities of county level in Huaihai Economic Zone. *Scientia Geographica Sinica*, 29(1), pp. 56-63.

Wan X., 2006. Report on the implement of national economic and social development plan in 2006 and draft in 2007 of Shijingshan district. http://huigu.bjsjs.gov.cn/affair/guihuajh/8a8481d21b533410011b584631130012.html

STUDY ON DATA INTEGRATION AND SHARING STANDARD AND SPECIFICATION SYSTEM FOR EARTH SYSTEM SCIENCE

Wang Juanle, Sun Jiulin

Institute of Geographic Sciences and Natural Resources Research, State Key Lab of Resources and Environment Information System, Chinese Academy of Sciences, 100101, Beijing, China - (wangjl, sunjl)@igsnrr.ac.cn

KEY WORDS: Earth system science, Geosciences data sharing, Standards and specifications, Geosciences Facility

ABSTRACT:

Earth System Science (ESS) has evolved into a new historic stage beyond Earth Science. ESS takes the whole earth systems and their interactive actions among spheres as its objective. This causes that its development need lots of multi-disciplinary, multi-sources, multi-type, integrated geosciences data resources support. According to this requirement, Data Sharing Network of Earth System Science (DSNESS) was established at the beginning of China Scientific Data Sharing Program (SDSP) launched in 2002. Data sharing in DSNESS need standards and specifications environment. Basic data sharing concept model are studied firstly for the requirement. According to 4 principles designed, ESS data sharing standards and specifications system and its relationship among related international and domestic standards system are studied. Designed standards and specifications system includes 4 main classes, i.e., mechanism and rules class, data management standards and specification class, platform development specification class, data service specification class. Total 18 standards, rules and specifications have been drawn up. Through almost 6 years research and application, nowadays all of these 18 standards and specifications have been used successfully in the distributed data sharing network of earth system science which include 1 general center and 13 sub-centers. In the near future, standards and specification environment of DSNESS will be further development two directions oriented, one direction is basic concept reference model research, and another direction is data fusion and assimilation standards and specifications study.

1. INTRODUCTION

Geosciences have evolved into Earth System Science stage in 21st century (Liu Dongsheng, 2004). Because Earth System Science (ESS) takes the whole earth systems and their interactive actions among spheres as its objective, it has series typical features, such as global system, whole temporal and spatial scale, multi-disciplinary integration, high technology application, global collaboration, and so on. Usually, its research approach includes observation, understanding, simulation and global environment change prediction (Zhou Xiuji, 2004. Huang Dingcheng, etc., 2005). Its development need lots of multi-disciplinary, multi-sources, multi-type, integrated geosciences data resources support according to its disciplinary features and characteristics. While, most of these geosciences data are distributed in different research agencies, group or even personal scientists' database. How to integrate and share these data in Earth System Science research community is a challenge for ESS development.

According to this requirement, China has begun to take actions for scientific data sharing since 1994. Through about 10 years' preparation, China Scientific Data Sharing Project (SDSP) was launched in 2002 (Xu Guanhua, 2003). Data Sharing Network of Earth System Science (DSNESS) is one of the first 9 pilot projects in SDSP (Sun Jiulin, 2003). It has been one of long term data sharing platforms in National Science & Technology Infrastructure (NSTI) since 2005. This project will provide Earth System Science data sharing platform for scientists. Its objective is to enhance and support the Earth System Science research and science & technology innovation through integration and sharing of distributed scientific data in institutes, universities, individuals, government funded research projects, and international organization or data centers. At present, DSNESS' data resources mainly focus on 'Land surface and human-land relationship' research Area.

2. STANDARD AND SPECIFICATION REQUIREMENTS

Most of the Earth System Science data are produced in the processing of science and technology research activities, especially from the science and technology research projects funding by government. These data are different with the data derived from the long term operational system by varies agencies or departments, such as meteorology observation data, earthquake monitoring data and so on.

National Science Foundation of USA make a clear definition for research data and other data in its report, "National science board: Long-lived digital data collections: enabling research and education in the 21st century" published in sep. 2005. In this report, data collections are divided into 3 types, i.e., research data collections, resource or community data collections, reference data collections. Research data collections are the products of one or more focused research projects and typically contain data that are subject to limited processing or curation. They may or may not conform to community standards, such as standards for file formats, metadata structure, and content access policies. Quite often, applicable standards may be nonexistent or rudimentary because the data types are novel and the size of the user community small. Research collections may vary greatly in size but are intended to serve a specific group, often limited to immediate participants. There may be no intention to preserve the collection beyond the end of a project.

Resource or community data collections serve a single science or engineering community usually. These digital collections often establish community-level standards either by selecting from among preexisting standards or by bringing the community together to develop new standards where they are absent or inadequate.

According to the data collection definition of NSF, ESS data belong to typical research data. Integration and sharing of ESS research data need standard and specification environment.

3. STANDARD AND SPECIFICATION CONCEPT MODEL

Earth System Science data sharing standard and specification environment concept model is shown in figure 1. As left side of figure 1 showed, data derived from “project A” is only used in small user group inner project A. These data will not be shared for other users. This make these data is difficult to be found and accessed. And even some users find these data luckily, they can’t understand and use them easily, because these data have different catalogue classification, content structure, concept semantic, data format etc. Similar with this, project B’s data will only be used by the scientists in project B. other users can’t acquire and use them without the data sharing environment. This leads large number of geosciences data can’t be used by more scientists, and geosciences data can’t play important role for the geosciences development.

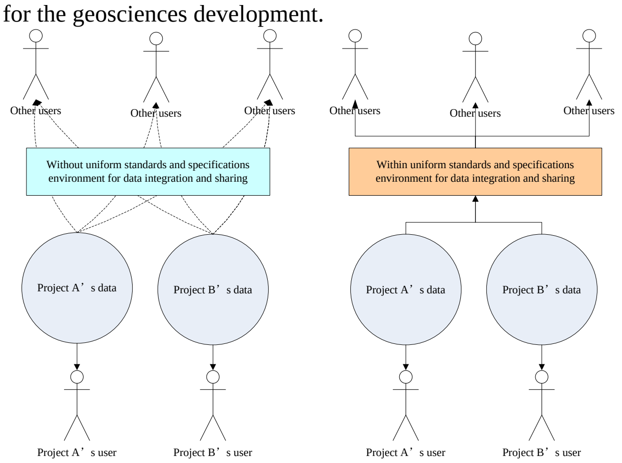


Figure 1. Geosciences data sharing environment

As shown in right side of figure 1, if we establish a data sharing standard and specification environment for Earth System Science, project A and project B’s data can be searched, browsed, accessed and acquired with uniform standards for more scientists.

But it is not difficult to establish this environment. Concretely, there are 3 main problems needed to be studied at present, i.e., how to establish the ESS data sharing mechanism? How to integrate and share multi-disciplinary data? How to make data easily for user find and access?

4. BASIC PRINCIPLES FOR DATA SHARING STANDARD AND SPECIFICATION ENVIRONMENT

Through analysis, there are 4 basic principles should be paid more attention to for Earth System Science data sharing standard and specification environment construction.

4.1 Corresponding principle

Standard and specification should correspond with their inside system and outside systems. First of all, all the standards and specifications should avoid conflict inside the Earth System Science data sharing system. At the same time, standard and specification of Earth System Science should correspond with its out side system. For example, Earth System Science data sharing is one of important components of NSTI in China, so it should keep consistency with the standards and specifications in NSTI system. Also, these standard and specification should keep consistency with the international standards, such as ISO 19115 metadata standard, etc.

4.2 Easy access principle

The final objective of Earth System Science data sharing is to provide an easy collection, management and dissemination environment. Let the user easily access all the Earth System Science data is a basic principle for the whole standard and specification system. Under this principle, Earth System Science data should show perfect and easy understand data catalogue for users firstly.

4.3 Reference model instruction principle

Geosciences data have a complex process from the data collection to dissemination. There should be a reference model to instruct the data sharing standard and specification environment. For example, ISO geography information reference model is a good example for Earth System Science standard and specification environment.

4.4 Software implementation principle

Data sharing need information and technology support. At present, distributed network is a popular tool for data sharing. Based on Earth System Science standard and specification concept model, its logical and physical model should also be researched. This principle will contribute to the software developers develop data collection and management tools, which is a good method for the standard and specification implementation in science community.

Based on the basic principle 1 and 2, Earth System Science data sharing standards and specifications should keep consistency with national and international standard systems. As figure 2 shows, Earth System Science standard and specification system should reference the ISO 19100 geography information series standards, Open Geospatial Consortium (OGC) interoperability technology standards, China national basic geography information standards system, China e-government data sharing standards system, NSTI basic standards system, SDSP basic standards system, etc.

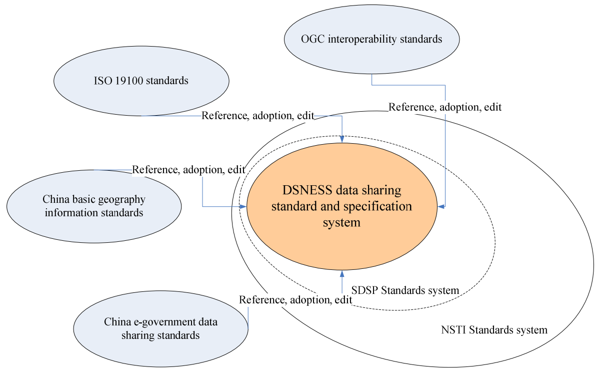

OGC interoperability standards
ISO 19100 standards
Reference, adoption, edit
Reference, adoption, edit
Reference, adoption, edit
China basic geography information standards
DSNESS data sharing standard and specification system
SDSP Standards system
Reference, adoption, edit
China e-government data sharing standards
NSTI Standards system

Figure 2. Orientation of standard and specification system for Earth System Science data sharing

5. SYSTEM ARCHITECTURE

Based on the principles and orientation of standards and specifications system, DSNESS established this initial architecture through almost 6 years of study and practice. The general architecture includes 4 class standards and specifications (shown in figure 3). They are mechanism and rules class, data management standards and specification class, platform development specification class, data service specification class.

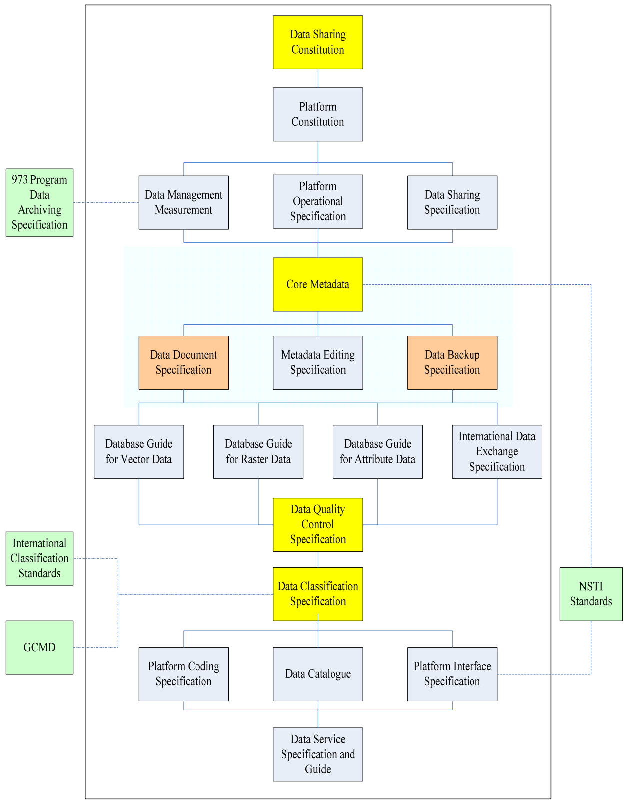

Data Sharing Constitution
Platform Constitution
973 Program Data Archiving Specification
Data Management Measurement
Platform Operational Specification
Data Sharing Specification
Core Metadata
Data Document Specification
Metadata Editing Specification
Data Backup Specification
Database Guide for Vector Data
Database Guide for Raster Data
Database Guide for Attribute Data
International Data Exchange Specification
Data Quality Control Specification
International Classification Standards
Data Classification Specification
NSTI Standards
GCMD
Platform Coding Specification
Data Catalogue
Platform Interface Specification
Data Service Specification and Guide

Figure 3. Standard and specification system structure for Earth System Science data sharing

5.1 Mechanism and rules class

ESS standard and specification mechanism and rules class include data sharing Constitutions, platform implementation measurement, platform operational management rules, data sharing rules and guides. These rules keep consistency with the national data sharing law and rules. For example, it will under the law of "Science and Technology Progress Law of China", "Specification of Scientific Research Program data archiving in Resources and Environment fields of National Key Basic Research Program" (Wang Juanle, etc., 2009), etc.

Data sharing constitution is the core mechanism of DSNESS. This constitution's core idea is to call for those geosciences research institutes, university, data centers or organizations, and personal scientists joining into this union. Data sharing organization architecture, union members' rights and responsibilities and duties are designed in this union. There are about 20 fixed members in this union at the beginning stage.

5.2 Data management standard and specification class

Data management standard and specification class includes metadata standard, metadata editing specification, data document specification, data backup specification, international data collection and exchange specification, data quality control specification, database design specification for vector, raster and attribute data, etc. It also keeps consistency with the national specifications. For example, DSNESS metadata standard is coordinated with NSTI core metadata standard and ISO 19115 geography information standard.

In these specifications of data management, metadata standard is the core one for data integration and sharing standards. Metadata, known as "data about data", describes data content and related dissemination information. Geosciences metadata, usually presented in XML format, represents who, what, when, where, why and how about a piece of geosciences data or other resource. Metadata possesses many benefits in data archive, assessment, management, discovery, transfer, and distribution. In the geosciences data context, metadata has the following usages: discovery of resources, Evaluation of resources, Use of resources, Contract between the user and the provider.

There are object level metadata and collection level metadata in DSNESS. Object level metadata describes a single entity. Collection level metadata describes a series or a group of entities. Geosciences metadata typically includes the following:

- Identification: including dataset title, citation, abstract, purposes, and keywords, etc.
- Quality: including positional accuracy, data completeness and consistency, etc.
- Spatial reference: including coordination system, spatial extent and temporal coverage of the data, etc.
- Distribution information: including distributor of and options for obtaining the data set, such as the format of the resource, the URL to download the data or access the Web services, etc.

Core metadata standard of DSNESS include 188 metadata elements, including 22 core elements.

5.3 Platform development specification class

Platform development specification class includes data classification specification, software development and coding specification, software interface specification, etc. Earth System Science data sharing classification specification is the core specification in this system. DSNESS designed the time scale classification, spatial scale classification, data catalogue classification and related coding system. This classification is the basic for data management and access. It is also the basic for the platform software development. According to the basic principles, DSNESS data classification specification is consistency with china's disciplinary classification and code standard (GB/T13745-92) and Global Change Master Directory.

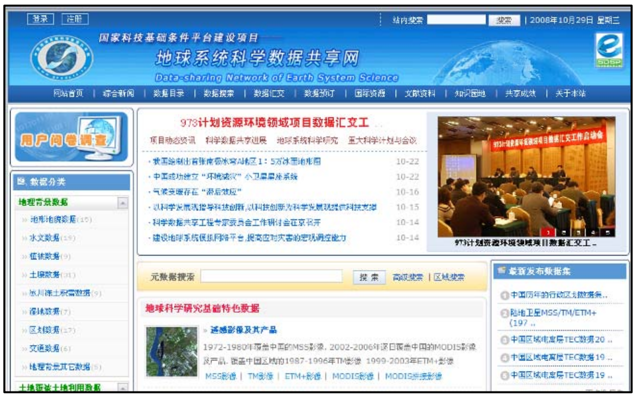
5.4 User service specification class

It includes user service specification and data service guides. These specifications ensure perfect user services for data sharing.

6. IMPLEMENTATION AND APPLICATION

At present, about 18 standards and specifications in this architecture have been drawn up. All of them have been used in DSNESS till now. Based on the organization of DSNESS, one general center and 13 distributed sub-centers all use these specifications and standards.

Under this standard and specification environment, more than 18TB data resources have been collected and shared in this network. Till to the end of 2009, almost 46000 users have resisted in the platform, 24.58 TB data have been downloaded for public. Figure 4 shows the website homepage of DSNESS (http://www.geodata.cn).



973计划资源环境领域项目数据汇交工 ..

- 我国绘制出首张南极冰穹A地区1：5万冰面地形图 10-22
- 中国成功建立"环境减灾"小卫星星座系统 10-22
- 气候变暖存在"滞后效应" 10-16
- 以科学发展观指导科技创新,以科技创新为科学发展观提供科技支撑 10-15
- 科学数据共享工程专家委员会工作研讨会在京召开 10-14
- 建设地球系统模拟网络平台,提高应对灾害的宏观调控能力 10-14

973计划资源环境领域项目数据汇交工 ..

1972-1980年覆盖中国的MSS影像，2002-2006年逐日覆盖中国的MODIS影像及产品，覆盖中国区域的1987-1996年TM影像 1999-2003年ETM+影像

1. 中国历年的行政区划数据集..
2. 陆地卫星MSS/TM/ETM+ (197 ..
3. 中国区域电离层TEC数据20 ..
4. 中国区域电离层TEC数据19 ..
5. 中国区域电离层TEC数据19 ..

Figure 4. Website home page of DSNESS

7. DISCUSSION

Earth System Science data sharing need a long term processing and its standard and specification environment should be revised and fulfilled continually. Face to the near future, there are two directions that these standards and specifications should be developed urgently. One is DOWN direction and another is UP direction. Under the DOWN direction, the basic concept model of standard and specification environment should be researched and studied deeply. Under the UP direction, different stage's concrete standards and specifications according to the data sharing life cycle should be paid more attention to, especially for the data product specifications, dataset fusion and assimilation specifications, dataset service specifications, etc.

8. REFERENCES

Liu Dongsheng, 2002. Global change and sustainable development science. *Earth Science Frontiers*, 9(1), pp. 1-9.

Zhou Xiuji, 2004. Some coginitions on earth system science. *Advance in Earth Science*, 19(4), pp.513-515.

Huang Dingcheng, Lin Hai, Zhang Zhiqiang, 2005. *Strategy Study on Earth System Science Development*. Beijing: Meteorology Publishing House.

Xu guanhua, 2003. Advance for enhance China's science and technology innovation capacity by data sharing, *China Basic Science Research*, (1), pp. 5-9.

Sun Jiulin, Shi Huizhong, 2003. Construction of Earth System Science Data Sharing Network in China. *China Basic Science Research*, (1). pp. 76-82.

Sun Jiulin, Lin Hai, 2009. *Earth System Study and Scientific Data*. Beijing: Science Press.

National Science Foundation. Long-lived Digital Data Collections: Enabling Research and Education in 21st Century , September 2005

Wang Juanle, Sun Jiulin, 2007. Development of China WDC Systems for Data Sharing, *China Basic Science Research*, (2), pp. 36-40.

Wang Juanle, Yang Yaping, Zhu Yunqiang, etc. 2009. Data Archiving Progress and Data Types Analysis of National Basic Research Program of China （973 Program） on Resource and Environment Field. *Advance in Earth Science*, 24(8), pp. 947-953.

NSF Board, 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, http://www.nsf.gov/pubs/2005/nsb0540

9. ACKNOWLEDGEMENTS

The research was funded by MOST (2005DKA32300); NSFC (40801180); Open Research Funding Program of LGISEM; Forward project for youth researcher of IGSNRR; Informationization Project of CAS. Also thanks Prof. Teng Mianzhen, You Songcai, Chen Shenbin and other colleagues' contribution in this research.

ESTIMATION OF IMPRECISION IN LENGTH AND AREA COMPUTATION IN VECTOR DATABASES INCLUDING PRODUCTION PROCESSES DESCRIPTION

JF. Girres [a], A. Ruas [a]

[a] COGIT Laboratory, Institut Géographique National,73 Avenue de Paris, 94165 Saint-Mandé, FRANCE
(jean-francois.girres, anne.ruas)@ign.fr

Commission VI, WG VI/4

KEY WORDS: Spatial data quality, Geometric imprecision, Production processes, Measurements, Data usage

ABSTRACT:

This paper presents a research on the estimation of the impact of geometric imprecision on basic measurements (length, area) in vector databases, in order to generate relevant information for decision making. The goal consists in the elaboration of a model allowing a non-expert user to evaluate the geometric imprecision of its dataset, using data analysis as well as description of production processes (such as digitising or generalisation). We suppose that these processes induce variable contribution to errors in a dataset, and are exposed to spatial heterogeneity according to the geographical context. This model lays on a knowledge base, based on measurements, contextual indicators and additional information on the dataset production. In order to evaluate a dataset's geometric imprecision impact without any reference, decision rules are under development, using the knowledge base coupled with hypothesis on the influence of the geographical context on production processes. Experimentation on a road network illustrates the respective impact of production processes in the final length measurement error. Possibilities to communicate this impact following a particular usage are also evocated.

1. INTRODUCTION

Since the three past decades, a significant number of research has been conducted on spatial data quality: the description of causes and consequences of errors in spatial databases (Chrisman, 1984; Burrough, 1986), the development of models to describe and visualise error and uncertainty (Goodchild et al., 1992; Hunter and Goodchild, 1996; Fisher, 1999), the development of error propagation models (Heuvelink, 1998) and applications to communicate the impact of spatial data quality for decision making (Devillers, 2004; Ying He, 2008).

In the same period, a global evolution occurred in the production and usages of geographic information: democratization of GIS tools, development of the GIS community in a large variety of users, and more recently the apparition of Volunteered Geographic Information (Goodchild, 2007) allowing to transform anybody in a sensor and a distributor of geographic information. In this context, issues related to spatial data quality and its communication to the final user become relevant.

The COGIT laboratory has been involved in researches on spatial data quality since years (Vauglin, 1997; Bonin, 2002; Olteanu, 2008) and looks for the development of methods and models allowing users to estimate the quality of spatial databases, and its impact on basic measurements. Since 2009, a PhD is conducted in the COGIT Lab, on the conception of a model to evaluate geometric imprecision in vector databases, in order to communicate its impact for decision making.

This paper proposes to expose the approach chosen to elaborate this model, focusing preliminarily on the context and the objectives of this research. A description of the approach to build the model is presented afterwards, illustrated by a practical example, before concluding.

2. CONTEXT AND OBJECTIVES OF THE STUDY

2.1 Spatial data quality

Concepts of spatial data quality

ISO defines *Quality* as the "totality of characteristics of a product that bear on its ability to satisfy stated and implied needs" (Oort, 2005). This definition of spatial data quality includes in fact two sub concepts (Devillers and Jeansoulin, 2006), as presented in Figure 1: *internal quality*, which can be described as the "ability to satisfy specifications" defined by the database producer, and *external quality*, also known as "fitness for use", which corresponds to the needs of the database users.

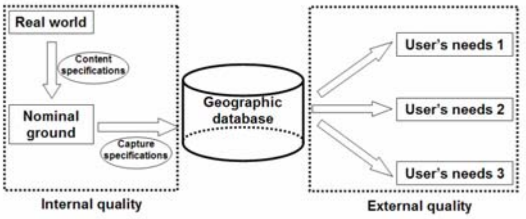


Figure 1. Internal and external quality

Specifications of geographic databases usually satisfy a part of both producer and users concerns. Unfortunately, it appears impossible to integrate all requirements involved by specific usages in the specifications, because they are endless.

Indeed, it is possible to establish a hierarchy of main usages (positioning, or length and area computation) because both concepts of internal and external quality are not completely disconnected. Also, tools and methods have to be provided to the final user in order to communicate the risk involved for a particular usage and to avoid misuses.

Elements of spatial data quality

To evaluate spatial data quality, different characteristics, called “elements” are distinguished. The ISO norm differentiates the following elements (Kresse and Fadaie, 2003): *geometric accuracy, attribute accuracy, completeness, logical consistency, semantic accuracy, lineage and temporal accuracy*.
(Oort, 2005) identified eleven elements in five influent sources publicised since two decades, integrating also *usage*, *variation in quality*, *meta-quality* and *resolution*.

Imprecision, Inaccuracy and Error

Our research concerns the element “geometric accuracy”, but focuses in particular on *geometric imprecision* and its impact on measurements.
Geometric imprecision is defined as the limitation on the granularity or resolution at which the observation is made, or the information is represented (Worboys, 1998a). It has to be differenced to *inaccuracy and error*, defined as the deviation from true values (Worboys, 1998b). For instance, we can estimate the geometric precision in positioning using the Root Mean Square Error (RMS Error).

2.2 Impact of geometric imprecision on measurements

We can consider that the impact of geometric imprecision on length computation can be illustrated by the formula bellow (1):

$$L_{comp} = L_{ref} + \Delta L \qquad (1)$$

where L_{comp} is the compared dataset’s length
L_{ref} is the reference dataset’s length
ΔL is the length variation
L_{ref} is more accurate than L_{comp}

As evocated before, we admit that the use of the RMS Error is well adapted to evaluate geometric imprecision in term of positioning, and looks enough to fit this use. But is it suitable to estimate the impact of geometric imprecision on measurements? To answer this question, a first experimentation using an error-simulation model, following a random law and parameterised with the RMS Error of the dataset, has been performed.

A sample of BDCARTO® road network dataset (RMS Error = 20 m, from specifications) is compared to a simulated BD TOPO® road network dataset (using the RMS Error of the BDCARTO). As presented in the Figure 2, the use of this error-simulation model does not represent faithfully the reality of the exposed example of a road network.

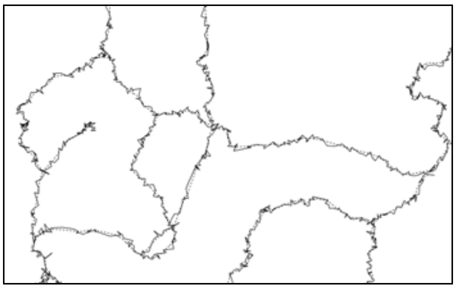


Figure 2. Unrealistic error-simulation on a road network (in plain black) following a random law

Moreover, comparisons of lengths show that this method is not adapted at all to measure the length: The total length of BDCARTO® road network is 67.1 km, compared to the 129.2 km of the simulated BDTOPO® road network, which is totally unrealistic (initial length of BDTOPO® is 65.9 km). This example shows that RMS Error is not suitable to evaluate the impact of geometric imprecision on length measurement.

The problem is quite complex to resolve and we suppose that the comprehension of causes of errors has to be introduced, using description of the different processes potentially considered as sources of errors. The formula below presents the contribution of production processes in the final length deviation (2):

$$\Delta L = \sum_{i=1}^{n} \Delta P \qquad (2)$$

where ΔP is the variation caused by a production process

We also suppose that local variations of geographical context can affect the contribution of each process generating the final “aggregated error”, because observations show that ΔL presents spatial heterogeneity in the entire dataset. If adding these values is pessimistic, at least it gives a boundary value.

In this context, processes potentially considered as sources of errors have to be understood and modelled, according to them sensitivity to the geographical context.

2.3 Causes of errors in basic measurements

Main sources of errors, in vector databases have been introduced by (Burrough, 1986). We focus here on five sources of geometric errors impacting the computation of length or area: digitizing errors, polygonal approximation, projection system and georeferencing, terrain modelling and generalisation.

Digitizing errors

Digitizing error is generated by the operator during the process of construction of geographic objects (Figure 3). It corresponds to the position uncertainty of each vertex of a vector object.

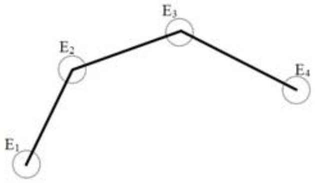


Figure 3. Digitizing error

Digitizing error is a random and independent error, modelled statistically by a probability distribution function (Gaussian Law). Its impact on the length computation of a polygonal line $E_1E_2\ldots E_n$ is modelled by the standard deviation bellow (3).

$$\sigma(e) = \sqrt{1 + 2 \sum_{2 \le i \le n-1} \sin^2 \frac{\theta_i - \theta_{i-1}}{2}} * \varepsilon_q \qquad (3)$$

where $\theta_i - \theta_{i-1}$ is the angle between consecutive vectors $E_{i-1}E_i$ and E_iE_{i+1} and ε_q is the digitizing precision

Properties of the Gaussian law give a confidence interval of 99.73% between -3σ(e) and 3σ(e).

Polygonal approximation

The polygonal approximation of curves generates a negative and systematic error (Figure 4) on lengths and areas. For a polyline, this error can be estimated by the difference between the polygonal length and the computed length of the curve.

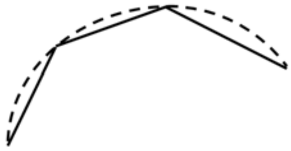


Figure 4. Polygonal approximation of curves

Projection system and Georeferencing

Representations using map projections generate distortions in the representation of the earth surface, and therefore in the computation of lengths. The scale error, defined as the difference between the distance on the map (particular scale) and the distance on the ellipsoid (Principal scale), is used to evaluate the impact of projection on length computation.

In the same time, the georeferencing of the data support (satellite or aerial imagery, maps…) can provide a systematic error in the dataset, after digitising. Parameters like translation, rotation and homothetic transformation have to be estimated.

Terrain modelling

Computation of lengths and areas in two dimensions are systematically smaller than using altitudinal information. Even if the altitude is not provided in the dataset, it can be extracted from Digital Terrain Model. Because the impact of the terrain can be important (especially in mountainous areas), differences have to be estimated to inform the final user.

Generalisation

If a dataset is produced using a map, effects of generalisation also generate errors which impact length and area computations. For instance, road may be translated, sinuous road are smoothed, some bends are removed, or houses may be enlarged and translated to facilitate visualisation.

As illustrated in Figure 5, several types of errors can be modelled by effects of generalisation process: anamorphous, translation, smoothing, and exaggeration.

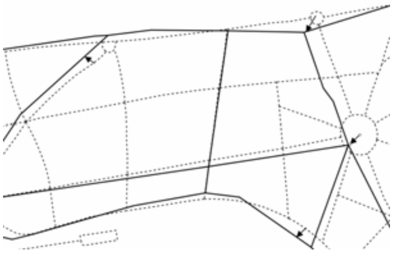


Figure 5. Effects of generalisation on road networks digitizing between BDCARTO® (in plain black) and BDTOPO®

Information on the potential impact of generalisation needs to be provided, by automatic detection, or using user's knowledge.

3. APPROACH

This ongoing research has multiple objectives. The first one is to understand the contribution of each production process in the final aggregated error, according to a particular geographical context. The second one consists in combining appropriate indicators to model geometric imprecision's impact on measurements. Thus, we intend to build a system, based on either reference datasets or hypothesis, related to knowledge on production process and on data.
This system supposes the construction of a knowledge base and decision rules, using preliminarily comparison of datasets.

3.1 Construction of knowledge base and decision rules

As exposed in Figure 6, in a first configuration, comparisons between databases DB1 and DB2 are performed in order to estimate deviations Δ in term of position, length and area. The computed rule base calculates uncertainty (on length or surface) based on measurements on data (data and context) and process by means of machine learning.

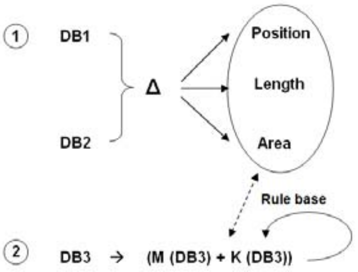


Figure 6. Comparisons between datasets to build the rule base

Then, in a second configuration, for a new dataset DB3 (with no reference dataset), estimation of deviations are computed by means of the rule base, the knowledge on the process (K(DB3)) and measures on the data (M(DB3)). Rule base computation R for a database DB3 is defined as the function (4) below:

$$R\ (DB3) = f\ (\Delta,\ M\ (DB3),\ K\ (DB3)) \qquad (4)$$

where Δ are measurements computed using comparisons
M (DB3) is the estimation of deviations
K (DB3) is knowledge on data

Rule base computation supposes to formulate a set of hypotheses. For instance, we can suppose that effects of generalisation are stronger in urban area, effects of terrain are stronger in mountains… Hypothesis formulation involves knowledge on both production process and data. This supposes to collect information provided by the user, and computed using appropriates measurements and contextual indicators (developed in Section 4) integrated in a model (Section 5). Experimentation on a mountainous road illustrates the impact of production processes on length measurements in Section 6.

3.2 Knowledge on production process

Prior the computation of indicators, the user should provide a set of normalised information about the datasets. Most of them are contained in the metadata, but among them, information like the processes used to create the dataset (generalisation or not) or the scales of production and usage have to be provided. This additional information also deals with the confidence on the data sources and processes, and the possible usage.

4. INDICATORS AND MEASUREMENTS

This part presents measurements performed using datasets comparison, but also contextual and shape indicators used to compute the rule base.

4.1 Measurements based on comparisons

To compute measurements, a preliminary phase of data matching of homologous objects is performed. This is realised automatically using algorithms developed by (Mustière and Devogèle, 2008) for linear objects, and (Bel Hadj Ali, 2001) for polygonal objects. Both of them are implemented in the GeOxygene library (Bucher et al., 2009).

Each type of primitives supposes the computation of adapted measurements. For points, indicators of precision (defined as the "fluctuations of a data series around its mean") and accuracy (defined as the "fluctuations of a data series around the nominal value") are computed using respectively standard deviation and RMS Error, for X and Y coordinates and deviations (using Euclidian distance). These indicators allow to evaluate the potential bias of the dataset, possible impact of the georeferencing. For polylines, curvilinear abscissa difference, Hausdorff distance and Average distance (Figure 7a) are performed. For polygons, Area difference, Hausdorff distance and Surface distance (Figure 7b) are computed.

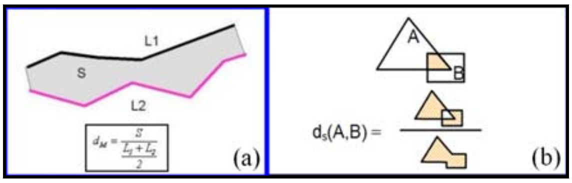


Figure 7. Computation of average (a) and surface distance (b)

4.2 Shape indicators

Shape indicators are computed for linear and polygonal objects. They provide important information to compute rule base. For polylines, indicators of granularity (smallest segment's length, average segment's length) and Sinuosity index (Plazanet et al., 1998) are for instance computed. For polygons, indicators of concavity and compactness are also provided. These measurements don't represent an exhaustive list. Further indicators will be integrated to complete the knowledge on data, as far as contextual indicators.

4.3 Contextual indicators

As exposed in hypotheses, we consider that the weight of the different processes potentially considered as sources of errors, is exposed to variations according to the geographical context. Contextual indicators are produced to characterise the geographical configuration of the dataset and its internal heterogeneity (terrain, density of objects...) in order to determine large areas, like mountains, urban or rural areas. These indicators are produced using the dataset itself, but also external datasets (DTM, networks…). Indicators of neighbourhood are computed in order to evaluate the potential effects of generalisation. For each object belonging to a dataset to evaluate, its neighbourhood population has to be determined. We consider that if this population is important, the object is exposed to effects of generalisation. Distances between objects can provide useful information to determine the scale of generalisation and its potential impact.

5. EVALUATION MODEL

To estimate the impact of geometric imprecision on classical measurements in vector databases, we propose an evaluation model (under development) based on three steps:

- Step 1: Evaluation of a dataset using rule base
- Step 2: Communication of geometric imprecision impact on measurements
- Step 3 : Rule base reinforcement

5.1.1 Evaluation of a dataset using rule base

Two configurations can arise for a user involved in the evaluation of its dataset (Figure 8):

- the user has reference dataset
- the user has no or few reference datasets

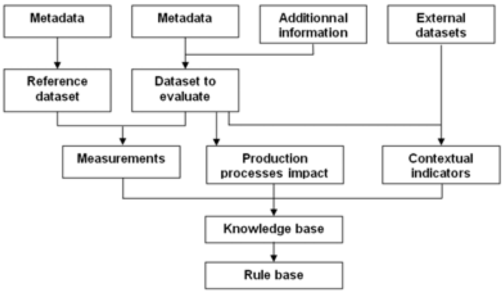


Figure 8. Creation of knowledge base and decision rules

If the user has reference dataset, he can perform comparisons using measurements presented in Section 2, combined with knowledge on production processes and on data. If the user has no, or few reference datasets (like a DTM), geometric imprecision impact on measurements is auto-estimated using decision rules. The rule base is elaborated using previous comparisons integrating knowledge on production processes and on data. Section 6 presents an example of comparison, estimating the respective impact of each production process on length measurement in a mountainous area. Samples of comparisons in different geographical contexts will be performed to elaborate the knowledge base.

Dataset evaluation provides a raw result, not really understandable for the user. In order to fit the use, and communicate clearly the impact of geometric imprecision on measurements, usages have to be taken into account.

5.2 Communication of geometric imprecision impact on basic measurements

Communication of the geometric imprecision impact to the final user involves the introduction of profiles and levels of usage, in order to adapt results to particular usage contexts. Various profiles of users have to be determined, in order to adapt the evaluation in a comprehensive talk. In the same way, thematic example will be used to adapt this communication to particular usage. The goal is to propose as possible to furnish sensitive information to the final user.

5.3 Rule base reinforcement

The last step deals with the reinforcement of the rule base, using validation, or not, of prior evaluations. This revision will be performed manually at the beginning, but we plan to provide a system able to modify rule base according to validated results.

6. EXPERIMENTATION

First experiments are realised to illustrate production processes impact on length computation in linear vector databases, in order to create the rule base. The example focuses on a road network extraction in the mountainous region of Grenoble (France) in two databases: The BDTOPO® and BDCARTO®, produced by IGN, the French National Mapping Agency.

6.1 Presentation of the datasets

The IGN BDTOPO® is a topographic database, of metric positional precision, captured using photogrammetric restitution and ground surveys. The IGN BDCARTO® is a cartographic database, captured using 1:50000 IGN maps and SPOT satellite imagery. Its average positional precision is around 20 meters.
The experimentation focuses on an extraction of road network, the D112 (Figure 9), modelled by polygonal lines in both databases. The projection system used is the RGF Lambert93.

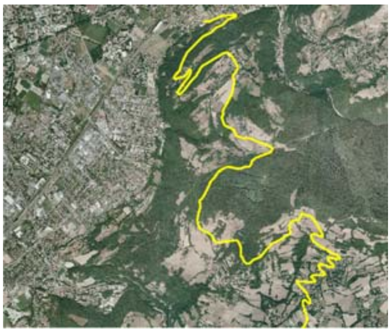


Figure 9. Localization of the road D112, in Grenoble's suburb

Using a classical GIS measurement tool, lengths of the D112 are 12.86 km for BDTOPO® and 12.62 km for BDCARTO®.

6.2 Components of length computation error

The different components of error are exposed hereafter, in order to model them impact in term of length measurement.

Impact of the projection system

Prior to the evaluation of causes of errors (section 2.3), the impact of projection system has to be taken into account.
In the example, the mean scale factor of the road is -0.67 m/km. In consequence, the total length of the road is underestimated of 8.6 meters for BDTOPO® and 8.4 meters for BDCARTO®.

Impact of digitizing error

To model the impact of digitizing error in the measurement of length, the digitizing precision ε_q is defined by the rate between sensibility of the capture (0,1 mm) and the digitizing scale (1:10000 for BDTOPO® and 1:50000 for BDCARTO®). The impact of digitizing error on the total length of the road is expressed by the standard deviation σ(e) in the Table 1.

Dataset	Length	ε_q	σ(e)	3σ(e)
BDTOPO	12,86 km	1 m.	4,9 m.	+/-14,7 m.
BDCARTO	12,62 km	5 m.	30,8 m.	+/-92,4 m

Table 1. Estimation of the impact of digitizing error

Impact of polygonal approximation

Considered as a curve object, the polygonal approximation involves a negative error on the road D112's length. The error b expresses the error in length computation.

Dataset	Length	Corrected Length	Error b
BDTOPO	12,86 km	12,89 km	+35,7 m.
BDCARTO	12,62 km	12,73 km	+112.4 m.

Table 2. Estimation of the impact of polygonal approximation

Impact of the terrain

The BDTOPO® road network is provided with altimetry, what is not the case for the BDCARTO®. In order to assign altitudes for each objects vertex of BDCARTO®, the BDALTI® is used. Computation of lengths using altitudes provides important differences, in comparison with a simple 2D computation, as shown in Table 3.

Dataset	Length 2D	Length 2D5	Difference
BDTOPO	12,86 km	12,95 km	+89,8 m.
BDCARTO	12,62 km	12,70 km	+80,6 m.

Table 3. Estimation of the impact of terrain

6.3 Discussion

The last contribution to consider in the length computation error is the one provided by the generalisation process. Detecting generalisation in a dataset is an ongoing task. This contribution to the final error is more complex to model, as it provides different effects on the objects shapes (such as simplification, enlargement of curves, bends removal).

Nevertheless, we can compute the corrected distance of the road D112 on both BDTOPO® and BDCARTO® datasets (Table 4). We suppose we can aggregate the errors modelled previously.

Dataset	Total Error	Length Min	Length Max
BDTOPO	134.1m.(+/-14,7)	12,97 km	13.00 km
BDCARTO	201.4m.(+/-92,4)	12.73 km	12.91 km

Table 4. Computation of the corrected maximum and minimum lengths by addition of errors

For the BDTOPO®, which it is not a generalised dataset, the addition gives a corrected distance of 12,99 km (+/-14,7 m.) where the most important part of the error is provided by not taking account of the terrain. For the BDCARTO®, the addition of errors gives a corrected distance of 12,82 km (+/- 92,4 m.). If we use the maximum value of the corrected length (12,91 m., which is close to the corrected length of the BDTOPO®), the error reaches 300 m, with an important impact of the polygonal approximation of curves.

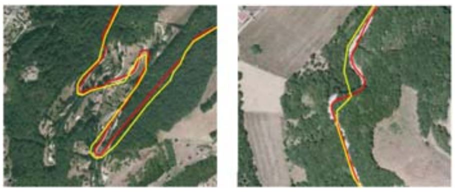


Figure 10. Example of generalisation impact on BDCARTO® road network (in yellow)

As we know that the BDCARTO® is captured using generalised 1:50000 IGN maps, as exposed in Figure 10, the impact of generalisation can be important and have to be estimated.

Thus, the example of the D112 well illustrate that the different components of error provide different impacts on the computation of length. This road has been voluntarily chosen because of its mountainous configuration, which exaggerates impacts of the terrain or also generalisation. In comparison, the same computation of errors performed in a region of plain provides results significantly different. For example, for a road of 2,96 km, the total error is 2,5 m (addition of impacts of projection system, polygonal approximation and terrain), with a digitizing error uncertainty of +/-7,18 m. This result illustrates that the impact of contributions of the final length error is completely different according to the geographical context.

Nevertheless, estimations performed show that integrating knowledge on production processes can help to understand the components of the error in a dataset and to estimate their impact in the length computation error.

7. CONCLUSION AND PERSPECTIVES

This paper presents an overview of an ongoing research on the conception of a model to evaluate geometric imprecision impact on classical measurements, and its communication to the final user. The integration of knowledge on production processes constitutes the original aspect of this work as we assume it provides variable contributions to the final error according to the geographical context, impacting the computation of length and area. Experimentation performed on a road network illustrates the respective impact of each production process in the length measurement error. In perspectives, the elaboration of the model will suppose to integrate measurements, contextual and shape indicators with additional information in order to constitute a knowledge base. Validation of hypothesis and rule base represents the core of the model, as far as the understanding of the combination of production processes impact in the final error. Finally, the communication of results constitutes the ultimate step to attend.

REFERENCES

Bel Hadj Ali, A., 2001. Qualité géométrique des entités géographiques surfaciques, Application à l'appariement et définition d'une typologie des écarts géométriques, PhD Thesis., Marne-la-Vallée University, France, 210 pp.

Bonin, O., 2002, Modèle d'erreur dans une base de données géographiques et grandes déviations pour des dommes pondérées ; application à l'estimation d'erreurs sur un temps de parcours, PhD Thesis, Paris 6 University, France, 147 pp.

Bucher, B., Brasebin, M., Buard, E., Grosso, E. and Mustière, S., 2009. GeOxygene: built on top of the expertness of the French NMA to host and share advanced GI Science research results. *Proceedings of International Opensource Geospatial Research Symposium 2009* (OGRS'09), 8-10 July, Nantes (France).

Burrough, P., 1986. Principles of Geographical information system for Land Ressources assessment. Oxford University Press, 193 pp.

Chrisman, N., 1984. The role of quality information in the long term functioning of a geographic information system. *Cartographica*, 21(2-3), pp. 79-87.

Devillers, R., 2004. Conception d'un système multidimensionnel d'information sur la qualité des données géographiques, PhD Thesis, Laval University, Québec, Canada and Marne-la-Vallée University, France, 157 pp.

Devillers, R., and Jeansoulin, R., 2006. *Fundamentals of Spatial Data Quality.* ISTE, London, 312 pp.

Fisher, P., 1999, Models of uncertainty in spatial data. In: *Geographical information systems: principles, techniques, management and applications*, John Wiley and sons, London, Vol. 1, pp. 191-205

Goodchild, M., Sun, G. and Yang S., 1992. Development and test of an error model for categorical data. *International journal of geographical information system*, 6(2), pp. 87-103

Goodchild, M., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research, 2*, pp. 24-32

Heuvelink, G., 1998, Error Propagation model in environmental modelling with GIS. Taylor and Francis, London, 127 pp.

Hunter, G. and Goodchild, M., 1996. A new model for handling vector data uncertainty in GIS. *Journal of the urban and regional information systems association*, 7(2), pp. 11-21

Kresse, W. and Fadaie, K., 2004. *ISO standards for geographic information.* Springer, Berlin, 322 pp.

Mustière S., Devogele T., 2008. Matching networks with different levels of detail, *GeoInformatica*, 12 (4), pp. 435-453

Olteanu A., 2008. A Multi criteria fusion approach for geographical data. In *Quality Aspects in Spatial Data Mining*, Taylor and Francis, pp 45-56

Oort, P., 2005. *Spatial Data Quality: from description to application*, PhD Thesis, Wageningen University, The Nederlands, 132 pp.

Plazanet C., Bigolin, N., Ruas, A., 1998. Experiments with learning techniques for spatial model enrichment and line generalization. *GeoInformatica* 2(3), pp. 315-333

Vauglin F., 1997, Modèles statistiques des imprécisions géométriques des objets géographiques linéaires, PhD Thesis, Marne-la-Vallée University, France, 286 pp.

Worboys, M., 1998. Computation with imprecise geospatial data. *Computers, Environment and Urban Systems* 22(2), pp. 85-106

Worboys, M.F., 1998. Imprecision in finite resolution spatial data. *Geoinformatica,* 2(3), pp. 257-280

Ying He, 2008. *Spatial Data Quality Management*, PhD Thesis, University of New South Wales, Sydney, Australia, 188 pp

DIGITAL CHART CARTOGRAPHY: ERROR AND QUALITY CONTROL

D. WU [a, b, c, d] *, H. HU [d] , X.M. YANG [b], Y.D. ZHENG [d], L.H. ZHANG [d]

[a] Yantai Institute of Coastal Zone Research, CAS, Yantai 264003, China, wudiok2468@sina.com
[b] Key State Lab of Resources and Environmental Information system, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China, wud@lreis.ac.cn
[c] South China Sea Institute of Oceanology, CAS, Guangzhou 510301, China
[d] Dalian Naval Academy, Dalian 116018, China

KEY WORDS: Tetrahedron model, Quality Control, User, Digital Chart, Cartography, Practicability

ABSTRACT:

This paper puts forward a novel concept of tetrahedron model of chart cartography error analysis and quality control which is necessary for introducing user to the former triangle model of error analysis and quality control, i.e. Reality-Producer-Chart-User. For the human being, high quality does not necessarily represent high accuracy, but represent a concept that what is fit for use. Classical cartography error theory coming from physics and mathematics will no longer meet the requirement of users, so the new concept deriving from communication and cognition sciences will compensate for the research. New model should adopt thought science which is from humanities. In tetrahedron model, producer refers to people who provide data, who manipulate data and who check data. User is considered as a quality controlling factor which has equal status with producer. Data quality control has changed from “Check-Driving” to “User-Driving”. The words for quality assessment change from error or uncertainty to practicability or applicability. Reality here refers to ocean and coastal zone. From a new viewpoint, the chart has been redefined as chart set. And characters of chart differing from terrain map and the specific quality assessment indexes have been discussed. Some innovative ideas of the research work were as follows: 1) Establishing a new tetrahedron model of quality control. The relationship between error and quality assessment was analyzed and practicability was adopted to assess the quality of chart product. 2) Defining the chart set. After changing from paper chart to electronic chart or digital chart, chart product become a set of digitizing collecting, storing and renewing data. 3) Building up the user assessment system of the chart product, which would perform real-time chart quality control.

1. ACTUALITY OF CHART CARTOGRAPHY ERROR THEORY

The series of steps of chart cartography, e.g. chart editing and designing, and generalizing (including selection, simplification, combination and replacement) will inevitably influence chart quality. How we estimate the degree of those operations? Is the final chart product fit for use? How is the dependability or reliability of the product? All the questions require scientific assessment of chart quality.

The error research of digital chart data has for a long time adopted error and uncertainty theory of spatial data. Because these are part of spatial data, and have many commonness. Now the important norm when evaluate the chart data quality is the error or accuracy of spatial data of chart. The description of error or accuracy is provided by the producer. However, the specific chart cartography process and special user of chart are not deeply discussed. And the reliability and practicability of a chart could not be provided to the user. These lead to re-examine the quality assessment system and more reasonable and efficient quality control model should be built up. Based on the actuality of error theory, a new quality assessment model was established. Then we re-defined the chart set in the digitized society. Finally we built up a user assessment system to perform a real-time quality control.

1.1 Error theory

Error, accuracy, uncertainty, quality, the four words are all about the assessment of geographic spatial data, and each has its own emphasis. The essential of error, accuracy, uncertainty, quality is all about the spatial data application system. Study on the error, accuracy, uncertainty, quality theory is carried out at home and abroad. For example, Li Deren, Liu Dajie, Chen Jicheng, Shi Wenzhong and Michael Goodchild, Giles Foody, Gerard Heuvelink, D. Griffith all study on spatial data error and establish the error analysis system, including the error source, error discrimination and measurement method, error propagation model, error administration and some minimizing methods to errors that influencing spatial data quality (Li D., 2002; Liu D., 1999; Cheng J., 2004; Shi W., 2005; Goodchild M., 1989). Traditional mathematical statistics methods are the base to the error analysis theory system. But the classic mathematical statistics theory must be revised and reinforced according to spatial data manipulating characters. A lot of progress has achieved since 1990s, for example, GIS spatial elements uncertainty model establishing methods (Heuvelink), error propagation model of spatial data processing and analysis(Liu,1999), remotely sensed image error and its impact and its error index establishment (Foody, ;Arnoff, 1985). But there are also many areas that need to be researched, for

* Corresponding author. Tel.: +86-10-64889132
E-mail address: wudiok2468@sina.com

example, the correlation between uncertainty study and spatial data quality control has little compactness. In fact the study of uncertainty focuses on only quality assessment but not on the quality control. So the uncertainty theory and practice should be emphasized on the practice of geographic information system, i.e. the quality control of GIS producing practice.

As M.F. Goodchild said in the keynote presentation in the symposium of Accuracy 2008, "…We need to ask a series of questions, beginning with what should spatial accuracy assessment mean in a world in which everyone is a potential user of geospatial data?" This is a very different perspective from the traditional one of the past 14 years, when it was possible to believe that the results of spatial accuracy assessment were of concern only to small elite of geospatial professionals (Goodchild M., 2008).

1.2 Chart cartography error theory

Chart spatial data has accuracy problems in every procedure from original acquiring, digitizing, generalization and examination. Error or uncertainty is adopted to describe the data quality. Accuracy assessment and error propagation model establishment toward a digital chart production are more complicated than normal survey adjustment and accuracy estimation. For the one thing data source of charting spreads various types of data from different area and different collecting methods. For another, different from survey data which have rigid geometric relationship and are easy to follow their error propagation process, but operations in chart cartography are more complicated. Aside from spatial information, there are attribute information in the operation objects. So the error propagation model of these operation objects is hard to build up. Take positional information for example, considering different scales, different years and different projections of a point position, the error propagation of the position is difficult to follow.

Although the automatic charting producing provides fast and efficient pattern, the quality reliability assessment is not provided. And literatures about digital charting and application have few articles related to error propagation of geographical spatial data, and the quality of input data and output data have no estimation. Though an integrated chart spatial data error analysis theory has not yet been established, there are many articles about the confronted problems and solving methods about theory establishment(Zeng, 2004; Sun, 2004; Li, 2007). These dispersive achievements are important to establish an integrated charting error analysis theory.

2. NEW TETRAHEDRON MODEL OF QUALITY CONTROL

2.1 Triangle model

The quality assessment of chart cartography is defined in a triangle model, i.e. Reality-producer-Chart, a closed loop, see Figure 1.

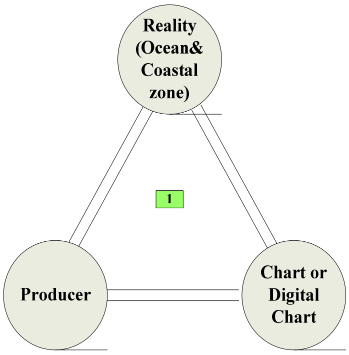


Figure 1 Triangle Model of Quality Control

There are broad meanings of producer, referring to those who provide data, who manipulate data and who check data. Data in different phases will be consider as different product (later will be discussed in the third part of chart set). These people will control the data of different phases subsequently and the contents of error analysis are as the following four part.

1) Error analysis of cartography data preparing

(a) Chart cartography raw material (e.g. sounding data) error analysis and accuracy index;

(b) Chart cartography historical data (e.g. scanning terrain map) error analysis and accuracy index;

2) Error propaganda and accumulate in chart designing and editing

(a) Error analysis in chart sheet designing and mathematics base deciding, the emphasis is laid on to the accuracy of ground control points and cartographic grid.

(b) Error analysis in chart contents selecting and geographic elements expressing, emphasis is laid on to the accuracy of the drawing of various elements and the reasonability of these relations.

(c) Error analysis in cartographic data processing, there are correcting from new material, old material transferring, adjoining the land and sea area, and data copying.

3) Error and uncertainty analysis in cartographic generalization

(a) Integrated analysis of the randomicity of location of spatial data and their and the fuzzy character of attribute data.

(b) Quality assessment of automatic generalization including generalizing process model, generalizing arithmetic and the rationality, maturity of rules, the degree of intelligentizing and the other assessment models.

4) Error propaganda and accumulate in other chart producing
Quality management of plate distribution, publishing, etc.
Figure2 give a concrete process of quality control.

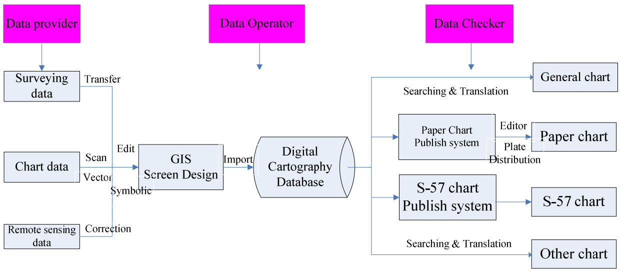


Figure 2 People in digital chart quality correct process

2.2 Tetrahedron model

From the above analysis of digital chart quality control process, the concept model is based on a triangle plane relationship. As the user should be one of the elements of quality control, the former triangle model should be developed to tetrahedron model including six edges or six key relations concerning quality, thus a new quality control concept model based on tetrahedron could be set up, see Figure 3.

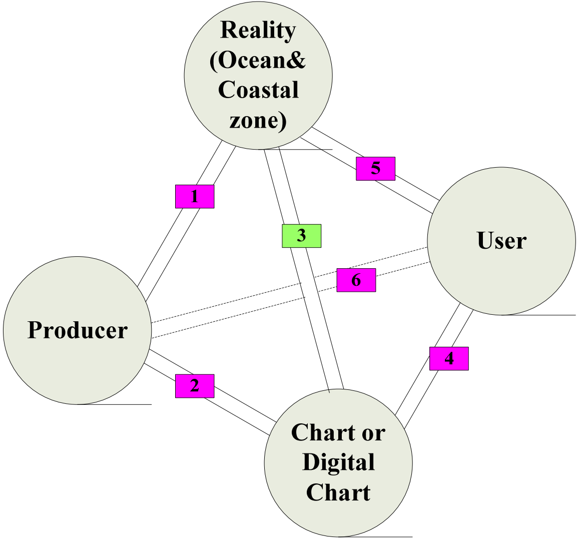


Figure 3 Tetrahedron Model of Quality Control

Here one point must be made clear that quality is determined by the buyer or the seller. In the "Client is God" society, the answer is quite clear. There are few productions like chart, merging arts and sciences together, the quality assessment of chart will not only refer to location, attribute accuracy, logic consistence and time precision, but refer to more important elements about the convenience, simplicity and fastness of chart use. "People-oriented" is an all around standard in this digitizing information era (GAO, 2004).

What the triangle model missed is just the user. For one side, the spatial cognition, visualizing thought when reading a chart, the feeling about virtual space of the user should be considered as influence elements of chart cartography quality, for the other side, the degree of simplicity and fastness should also be regarded as indexes of quality assessment. The former standard of high quality refers to high accuracy. But the viewpoint of equalling high quality with truth of reality has inherent contradiction in concept (Andrew U.F, 2008). And the ontological analysis reveals the necessity to separate the ontology (reality) proper from the epistemology (data). The ordinary language approach clearly identifies the two conflicting interpretations of "data quality": the viewpoint of the producer and the contrasting viewpoint of the consumer (Timpf et al, 1996). In order to make an overall quality assessment, user should be considered in the process of use. So the data quality does not only refer to accuracy but also refer to practicability or applicability. We aim to make an integrated and scientific error analysis and quality assessment as to set up a tetrahedron model of quality control.

2.3 Integrated Error theory based on Tetrahedron

Traditional error theory focused on the first three relationships. There are abundant research works in RS, GIS spatial data error analysis (Liu, 1999; Cheng, 2004; Ge, 2006). But the 4th, 5th and 6th relationships are lack of attention. The error and quality analysis with only three relationships are partial and local. Only taking consideration of all the six relationships can we wholly master the error accumulation and propagation theory.

In fact, the missing of three edges is the missing of "user". Error accumulation consists of not only the first three relations: e.g., the survey error in Reality-Producer relationship, and the scanning and digitizing error in Producer-Chart relationship, and the geometric rectification error in Reality-Chart relationship. These three kinds of errors are the main analysis objects of chart cartographic error theory. After introducing tetrahedron model, the user spatial cognition, the visualizing thinking of readers and the error influence effects of virtual spatial cognition should be taken into consideration. People factors that influence the error accumulation system including two aspects: one is that the subjective factors of man can influence the producing process and then lead to error accumulation, the other is that the degree of convenience and simplification of reading for ordinary people can be made an index of chart quality assessment.

The ultimate purpose of error analysis is to establish mathematics statistics equations and models in order to acquire high quality. Take a new perspective of the six edges of the tetrahedron, the error analysis is no longer traditional rigid, linear or static. The 4th, 5th and 6th relationships are related to thinking science, cognitive science and non-linear science, thus, should adopt different research patterns. The characteristic of this study is to integrate the six edges into a comprehensive error theory. The theory framework and conceptual model of each relation are showed in Table 1.

Table 1 Cartography tetrahedron uncertainty conception model

ITEM	Representation	People action	Error describing	Uncertainty theory
1 P-R	Producer-Reality	Cognition	Distance, angle distortion	Probability theory
2 P-C	Producer-Chart	Design	Scanning error, Topology, logic identify	Mathematical statistics
3 R-C	Reality-Chart	Ontology	Error matrix	Ontology
4 U-C	User-Chart	Reading	Fuzzy logic	Fuzzy theory
5 U-R	User-Reality	Practise	Psychophysical action	Psychology theory
6 U-P	User-Producer	Exchange	Communications	Information Science

3. DIGITAL CHART QUALITY ASSESSMENT ELEMENT

Digital chart is produced when chart integrated to computer science. In the new information era, charts are known as marine charts, hydrographical charts, or admiralty charts. They differ somewhat from bathymetric charts which are virtually topographic maps of the ocean floor. But digital charts are still objective reflection of the ocean and coastal zone, and are not necessarily the reflection of paper charts. Digital charts still solve the three contradictions because of scale, map loading and plane limits, i.e. the contradiction of ellipsoid and plane, the finite and infinite, the ordered and disordered. The movement and balance of the contradictions decide the scientific characters of chart.

The same as terrain map, charts have strict mathematics base, simple and efficient symbol system and scientific cartographic generalization. Charts have vivid characters of their own that form the specific style and tradition. The foremost characters exist in their cartographic purpose and their expressive objects. The objects of chart cartography are ocean and coastal zone. The most difference between ocean and land is the covered water. In different parts of ocean and sea, there are different depth, different temperature, salinity, density and transparency. There are also dynamic phenomena due to astronomy, weather and crustal movement, the ever-lasting movement of sea water, the vertical movement of tide, the parallel movement of tide current, ocean current and sometimes tsunami, wave and vortex. When we perform hydrographical survey, we hardly use optics instruments, we often use acoustics instruments. So the instruments, survey methods, the output production, error and accuracy are very different from land terrain surveying. The main product of land surveying are graph materials, while hydrographical surveying produces notes paper, notes tape and text data. So the cartographic process of chart and that of terrain map is of great difference. For example, multi-scanning sonar produced a depth black sheet by scanning the sea bed. In the style sheet, the arrangement of depths is determined by the surveying methods, i.e., four neighbour depth points form a rhombus, the direction of long diagonal line of the rhombus is parallel to the shoreline, and the direction of short diagonal line of the rhombus is vertical to the shoreline. So the differences of surveying methods lead to differences of generalization methods.

Besides these, the difference is bigger still in the content and expression of chart. It should therefore be useful to examine briefly the special features of charts.

1) Projection

Mercator is the most used projection because straight lines on the map are lines of constant azimuth on the ocean. It is convenient for navigation.

2) Scale

Coastal charts are generally in the l: 50 000 to l: 300 000 scale range, while port charts are at larger scales. Such charts are rarely in series and scale is determined by the area covered and the chosen paper size.

3) Datum

The datum for sounding reduction is not mean sea level, but some special lowest normal low water that is fit for navigation.

4) Sheets

Sheets are divided along the seashore, or along the navigation course. Adjoining sheets often have large area of overlap which will be convenient to navigate.

5) Serial number

In order to adapt to the adjoined sheets, charts have definite serial number system.

6) Detail

The detail shown on a chart is carefully selected so that everything useful to the mariner is shown and no more. The topography of the coastline is fully shown but inland only prominent objects visible from the sea, e.g. hills, radio masts, high buildings, chimneys, etc.

On or below water level are shown: (a) aids to navigation: lights, buoys, and beacons; (b) dangers: rocks, wrecks, and cables (c) currents and tidal streams: direction and velocity; (d) nature of the bottom: sand, mud, rock, etc. (e) routes and limits; (f) soundings and approximate isobaths. Information useful to mariners which is not suitable for printing on charts is compiled in book form, such a book being called a 'Pilot' (or Sailing Directions). It gives information about ports and harbours and navigation along coasts, climate, and many other relevant details.

7) Units

Again, because navigation is based on astronomy, the basic unit of distance is one minute of latitude, called a nautical mile. The average value of l852 m is an international nautical mile.

Depths of water are in metres (0.l m in shallow water) for all new work although there are many older charts showing fathoms (= 6 ft). The metre is also replacing the foot for all heights above sea level.

Horizontal scale bars show nautical miles, cables (l cable = 0.l nautical mile), metres, and feet; vertical bars shown feet, metres, and fathoms.

4. DESIGN FOR QUALITY CONTROL SYSTEM AND DISCUSSION

4.1 Understanding of quality

The charts or more specific nautical charts, have different elements concerning about quality assessment from terrain map. These elements, though objective, are still selective factors that influence the purpose of user. Chart quality descriptions represent the viewpoint of the chart producer and are not very helpful for the potential chart user to decide if it is “fit for use”. So before the quality assessment process, some work should be performed to decide what the quality really means. Purely from a management point of view on the concept of quality, a well-known American quality management expert, Dr. Juran J.M holds the point of view that has been widely recognized: it believed that product quality is applicability of the product, that is, the product can be in what degree met the user needs.

This definition has two meanings, namely, "the user’s requirements" and "satisfaction." People use the product, they make certain requirements and these requirements of the products are often affected by the time, the location, the objects, and social environment and market competition factors. Changes in these factors make people put forward different requirements of the same product quality. Thus, the quality is not a fixed concept, it is dynamic, changing and developing over time, place and users and with the development of society and technology, the definition of quality will be constantly updated and enriched.

The requirements of the product from users are reflected in the product performance, economic characteristics, service characteristics, environmental characteristics and psychological characteristics and so on. Therefore, the quality is a comprehensive concept. It does not require the technical characteristics of the higher the better, but the pursuit the best combination of performance, cost, quantity, namely, the so-called most applicability or practicability.

4.2 Advantages when concerning user

Study of practicability and reliability of chart cartography product has not attracted attention from the documentations. A complete error analysis theoretical system of spatial data of chart has not yet formed more concentrated focus on the study of location uncertainty, attribute uncertainty and so on. But for spatial data products of chart, errors, accuracy of the production process given are not enough to guide the user to practice. So the real user satisfaction survey is a proof of chart of acceptable quality. The user satisfaction is equivalent with the product reliability and suitability.

Learn from the marketing strategy of most commodities, user session will be taken into account. The assessment of practicability has very important and far-reaching significance:

1) Improve chart product competitiveness. To provide appropriate evaluation indicators is becoming a necessary condition of entering the international market. If chart does not provide reliable indicators, it is difficult to survive in the increasingly competitive society.

2) Improve the theory of chart cartography error and quality control. For the special area of cartographic charts and special users, uncertainty and reliability are closely related. Reliability studies can contribute to the research of uncertainty and further refinement.

3) From quality evaluation based on user feedback information, establish the corresponding mapping model, then the problem

can be found in order to chart correction, easy to carry out quality control.
4) Reduce the risks of failure; minimize the occurrence of quality control problems. This can improve the reliability of chart products, but also can reduce the man labour of data maintenance, then to reduce production costs.

5. CONCLUSIONS

As the world's rapid economic development and industrial competitiveness have sharpened, and the seller's market began to change to the buyer's market. Man has been growing attention to quality. The well-known American Quality Management scientist Dr. Juran predicted that the 21st century will be the century of quality. Quality will be the most effective and peaceful weapon to occupy the market and will become a powerful driving force of social development.
For the quality assessment of digital charts cartography, a tetrahedron quality control model was established, but still need further exploration, and there are a range of issues required lots of experiments and calculations, further study should be gradually explored and resolved:
1) Spatial data error and uncertainty theory is getting more sophisticated, but that only applies to the general GIS and remote sensing data, the special nature of chart data are still not involved, it is necessary to extend the theoretical model of error and quality control and change from the former study of error and uncertainty to user-led evaluation of the suitability.
2) As the chart cartographic data with GIS spatial data in general have both similarities and particularities. Some statistical data need to take into account such as surveys on the navigators, engineers of marine resource investigation or other specific chart users.
3) Man is the most critical part of charts suitability evaluation, but the study of people is the most difficult part at present. This requires multi-disciplinary research, such as cognitive psychology, behavioural science, reliability engineering, management and engineering disciplines.

6. ACKNOLEDGEMENT

This study was supported by the National Natural Science Foundation of China (No. 40971224) and the National High Technology Research and Development Program of China (No.2008AA121706 and No. 2009AA12Z148).

References:

[1] Arnoff, S., 1985. The minimum Accuracy Value as Index of Classification Accuracy. *Photogrammetric Engineering and Remote sensing*, 51(1), pp.593-600.

[2] Andrew, U.F, 2008. Data Quality - What can an Ontological Analysis Contribute? In: *The 8th International Symposium on Spatial Accuarcy Assessment in Natural Resources and Environmental Sciences, v1: Spatial Uncertainty* , World Academic Press., pp.393-397.

[3] An M., Zhang G., Tao D, 2006. The Cognizing Elements on the Map Spatial Relationship. *Journal of Zhengzhou Institute of Surveying and Mapping*, 23(6):436-439.

[4] Chen, S., Yue, T., Li, H. 2000. Studies on Geo-information Tupu and its application. *Geographical research*, 19(4),pp.337-343.

[5] Chen J., Zhou C,. Cheng W., 2007. Area error analysis of vector to raster conversion of areal features in GIS. *Acta Geodaetica et Cartographica sinica*, 36(3):344-350.

[6] Cheng J., Guo, H., Shi, W., 2004. *Uncertainty of remote sensing data*. Beijing, Science Press, pp.2-23.

[7] Casti Emanuela., 2005. Toward a Theory of Interpretation: Cartographic Semiosis. *Cartographia*, 40(3), pp.1-16.

[8] D.Griffith.,2008. Spatial Autocorrection and Random Effects in Digitizing Error. *In: The 8th International Symposium on Spatial Accuarcy Assessment in Natural Resources and Environmental Sciences, v1:Spatial Uncertainty* ,World Academic Press, pp.94-102.

[9] Deng, H., 2006. *A study of Automated cartographic generalization based on design for quality*. PhD thesis, Information Engineering University, China.pp5-11.

[10] Gao J., 2004. Cartographic Tetrahedron: Explanation of Cartography in the Digital Era. *Acta Geodetic et Cartographica sinica*, 33(1), pp.6-11.

[11] Ge Y., Liang Y.,Ma J,.Wang J, 2006. Error propagation model for registration of remote sensing image and simulation analysis. *Journal of Remote Sensing*, 10(3):299-305.

[12] Goodchild, M.F., Gopal, S., 1989. *The accuracy of spatial databases*, Taylor&Francis, pp.3-18.81-90.

[13] Goodchild, M.F.,2007. Towards user-centric description of data quality, Keynote presentation, *International Symposium on Spatial Data Quality*, Eenschede, Netherlands.

[14] Goodchild, M.F.,2008. Spatial accuracy2.0. In: *The 8th International Symposium on Spatial Accuarcy Assessment in Natural Resources and Environmental Sciences, v1:Spatial Uncertainty* ,World Academic Press, pp.1-7.

[15] Gilmartin, P.P., 1981. The Interface of Cognitive and Psychophysical Research in Cartography. *Cartographia*, 18(3), pp.9-20.

[16] Heuvelink, B.M, Burrough, P.A, and Stein A., 1989. Propagation of Error in Spatial Modelling with GIS. *International Journal of GIS*, 3, pp.303-322.

[17] Li, D., and Yuan, X., 2002. *Error processing and reliability theory*. Wuhan University Press, pp. 1-7.

[18] Li, J., Sun, W., 2007. Study on the digital chart producing system and the process of quality control. *Hydrographic surveying and charting*, 27(1), pp.74-77.

[19] Liu, D., Shi, W., Tong, X., 1999. *Accuracy analysis and quality control of GIS spatial data*. Shanghai, Shanghai science and technology archive press, pp.12-21.

[20] Ma A., 2000. *On geographical Science and Geographical Information Science*. Wuhan,Wuhan publishing house, 261-281

[21] Shi,W.,2005. *Principle of modelling uncertainties in spatial data and analysis*. Beijing, Science Press, pp.10-37.

[22] Shi,W., 2008. From Uncertainty Description to Spatial Data Quality Control. In: *The 8th International Symposium on Spatial Accuarcy Assessment in Natural Resources and*

Environmental Sciences, v2:Accuracy in Geomatics ,World Academic Press., pp.412-417.

[23] Stehman, S.V., 2008. Sampling Designs for Assessing Map Accuracy. In: *The 8th International Symposium on Spatial Accuarcy Assessment in Natural Resources and Environmental Sciences, v2:Accuracy in Geomatics* ,World Academic Press, pp.8-15.

[24] Sun, W., Sun, Q., Zhang B., Shen J., 2004. Design and Implementation of quality checking system for chart publishing. *Hydrographic surveying and charting*, 24(3), pp.40-43.

[25] Timpf, S., M.Raubal and W.Kuhn, 1996. Experiences with Metadata. 7th International Syposium on Spatial Data Handling, SDH'96, Delft, the Netherlands (August 12-16, 1996), IGU

[26] Wood, Denis., John Fels., 1992. *The Power of Maps.* New York City: Guillford Press, pp. 1-7.

[27] Zhang, J., Yang,Y., 2009. Analysis on the status of Spatial Data Uncertainty Research. *Geospatial Information*, 7(3), pp.4-8.

[28] Zeng, Y., 2004. *Research on spatial data quality control and evaluation technique system*. PhD thesis, University of Wuhan, China.pp101-115.

CHARACTERIZATION AND DETECTION OF BUILDING PATTERNS IN CARTOGRAPHIC DATA: TWO ALGORITHMS

Xiang Zhang [a,b,*]**, Tinghua Ai** [b]**, Jantien Stoter** [c,d]

[a]ITCt University of Twente,he Netherlands
xzhang@itc.nl
[b]School of Resource and Environment Sciences, Wuhan University, China
tinghua_ai@tom.com
[c]Delft University of Technology, the Netherlands
[d]Kadaster, Apeldoorn, the Netherlands
j.e.stoter@tudelft.nl

KEY WORDS: Pattern Recognition, Building Pattern, Map Generalization, Delaunay Triangulation, Minimum Spanning Tree, Algorithm, Graph Theory

ABSTRACT:

Building patterns are important features in applications like automated generalization and spatial data mining. Many previous work has however focused on a few specific patterns (i.e. collinear pattern), while many others are less discussed. This paper proposes a comprehensive typology of available building patterns through the study of existing maps, and discusses their characteristics. This typology includes collinear, curvilinear, align-along-road, grid-like and unstructured patterns. Two algorithms are presented to detect align-along-road and unstructured building patterns, which are tested against a topographic dataset of the Netherlands.

1 INTRODUCTION

Building patterns are important features in urban and rural areas. The automated detection of visually significant building patterns is required for applications like automated map generalization, automated evaluation of generalized output, semantic enrichment of spatial databases, and spatial data mining. For example, collinear patterns have been extensively investigated (Boffet and Rocca Serra, 2001, Christophe and Ruas, 2002) in order to simplify and typify building groups. An approach to detect higher-level semantics like terraced house (Lüscher et al., 2009) made use of detected building alignments; the detection approach is however not widely applicable as the buildings in their case was topologically adjacent, which is not commonly the case. Therefore, a more comprehensive view of which building patterns are available, and a generic approach to detect and characterize them are required.

As for the detection techniques, a remarkable and comprehensive investigation has been made on the use of minimum spanning trees (MST) in the field of pattern recognition (Zahn, 1971). After being successfully applied to some classical clustering problems for point sets, this technique was applied to detect building clusters for generalization purposes (Regnauld, 1996). Nevertheless, no interesting building patterns is recognized with this technique except some general tree-like clusters.

This paper firstly proposes a typology of building patterns by mainly studying existing topographic maps (Section 2). Then Section 3 develops a Graph-theoretic based approach to the detection and characterization of two common patterns of the typology. The proposed algorithms are implemented and tested in Section 4. This paper ends with conclusions (Section 5).

2 TYPOLOGY AND CHARACTERISTICS OF BUILDING PATTERNS

A typology of building patterns is needed as it formalizes our knowledge on building structures available in geospatial domain. The typology of all building patterns that occurred in the studied maps and previous literature are identified and characterized in Section 2.1, and then we focus specifically on align-along-road pattern, discussing its relationships to other linear patterns (Section 2.2).

2.1 Identifying and characterizing the typology

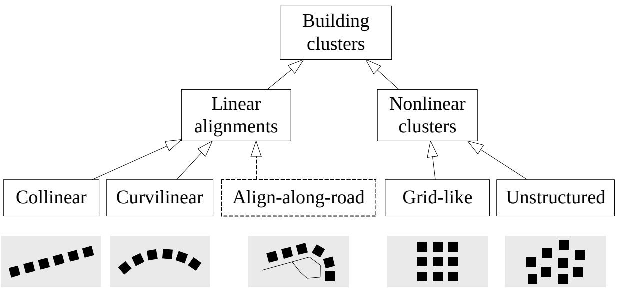


Figure 1: Typology and schematic examples of building patterns

The building patterns discussed in this paper are categorized as low-level, localized visual patterns, because they are important considerations in map generalization. The typology (Figure 1) is structured as follows. We define building patterns at top level as building clusters of spatially proximate objects with similar geometrical (e.g. spacing, size, orientation and shape) and semantic properties, extended from the definition of point patterns (Zahn, 1971). At an intermediate level, we refine the building clusters into linear alignments and nonlinear clusters in terms of 'group shape'. In general, the linear alignments appear to be more elongated and their constituent buildings can be organized by a linear path, while the nonlinear ones appear to stretch in two dimensions. At a finer level, the linear alignments are subdivided into collinear, curvilinear, and align-along-road patterns; the nonlinear clusters consist of grid-like and unstructured patterns.

We identify this typology from several sources. First, Gestalt principle of visual perception (Wertheimer, 1923) is used to define building patterns in general (proximity for general clusters) and specific (e.g. good continuity for linear alignments). As a

[*] Corresponding author. Email: xiangzhangchina@gmail.com

result, this typology is generic in the sense that all visual building patterns are clusters. Second, by comparing with those proposed by others we find that this typology is more comprehensive and generic. As mentioned before, collinear alignment has been acknowledged in the literature; grid-like pattern, though being less studied, has been discussed once for typification (Anders, 2006). He (Anders, 2006) also proposed a typology consisting of linear, circular, grid, star and irregular patterns. His linear pattern is enriched by our collinear and curvilinear alignments; meanwhile, his circular is a special case of curvilinear pattern and the star pattern are a combination of two crossing collinear alignments. We propose a further type (i.e. align-along-road) which integrates the relationship to surrounding roads. The typology proposed in this paper is regarded to be generic as all of the pattern types are confirmed by studying the maps (of the Netherlands, France, and Spain) in the EuroSDR generalization study (Stoter et al., 2009a), and maps of China, examples are shown in Figure 2. Other higher-level or global pattern can be seen as spatial combination of these low-level patterns and the integration of these patterns with other information such as semantics.

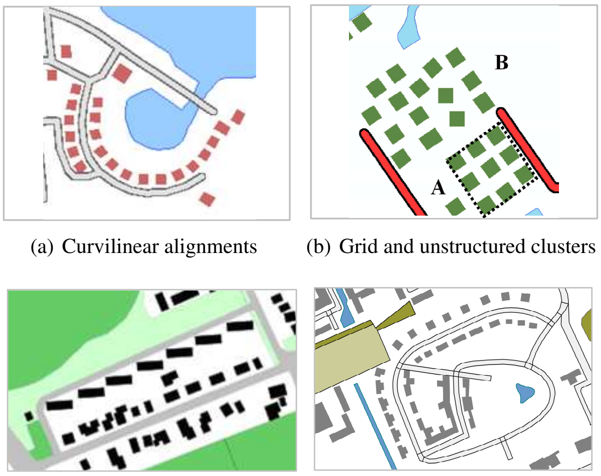
The typology is characterized as follows. According to previous work (Boffet and Rocca Serra, 2001, Christophe and Ruas, 2002, Ruas and Holzapfel, 2003), the homogeneities of general clusters can be realized using standard deviations of all the properties. In our approach, we improve this calculation by adopting the concept of coefficient of variance ($CV = Std/Mean$):

$$Homogeneity(P) = \frac{\sqrt{(p_i - Mean(P))^2}}{Mean(P)} \quad , \qquad (1)$$

where P represents the properties of spacing, size, orientation, shape, and semantics of the cluster; $p_i \in P$ denotes the values measured from or between the cluster's elements.

The use of CV applies to the properties of spacing, size, shape and semantics. This is not because we want to normalize the properties but rather because CV is a dimensionless number. It means that the homogeneities become then relative numbers invariant to the choice of measurement, which is also consistent with our perception about building patterns. For example, if two patterns have the same standard deviation of size, the one with larger mean size (i.e. smaller CV) is more homogeneous in terms of size property. However, CV cannot be applied to orientation as it is meaningless to calculate CV for a cyclic variable. Therefore, the homogeneity of orientation is calculated using standard deviation.

The homogeneities (Equation 1) are common characteristics shared by all types of building patterns. In addition, most of the specific types of building patterns have their own characteristics. In the case of collinear alignments, the patterns are characterized by *straightness* describing the sinuosity degrees of the paths and *main angle* describing the directions of the paths. For curvilinear alignments, *smoothness* of the paths and their *curvature descriptions* should be emphasized. Align-along-road patterns should be attached to the roads along which they are aligned. Besides, it has an extra homogeneity, that is the homogeneity of distances to the aligned road. This property reflects the degree to which the pattern are parallel to the road. For the two nonlinear clusters, unstructured clusters have no specific property while grid patterns can be further characterized by *squareness* and *parallelism*. If we connect proximate buildings in a grid pattern together, we should get two sets of parallel lines, which intersect each other approximately right-angled. All patterns of the typology are illustrated using existing maps (Figure 2(a)-2(d)).



(a) Curvilinear alignments (b) Grid and unstructured clusters

(c) Collinear alignments (d) Align-along-road patterns

Figure 2: Examples of different building patterns as a result of studying existing mapsr

2.2 Align-along-road pattern and its relationship to other linear alignments

As a result of studying existing topographic maps, we find that the align-along-road patterns are one of the most common features which are visually significant in urban and rural structures; we also find that the distinction between align-along-road patterns and another two linear (i.e. collinear and curvilinear) alignments is not always clear. On the one hand, it is common cases that buildings are located near roads and streets, and therefore such collinear and curvilinear alignments (as shown in Figure 2(c) and 2(a)) are also align-along-road patterns. On the other hand, collinear and curvilinear alignments may be parts of align-along-road patterns, because normally align-along-road patterns are not regular in terms of curvatures. This sometimes means that a align-along-road pattern can be segmented into pieces of collinear and/or curvilinear alignments. Of course, there are also situations where collinear and curvilinear patterns are independent of roads.

We therefore limit ourselves to the detection and characterization of two patterns, i.e., the align-along-road and unstructured patterns. Some of the other patterns have been discussed by previous authors. For example, the collinear pattern by (Christophe and Ruas, 2002) and the grid pattern by (Anders, 2006).

3 DETECTION AND AUTOMATIC CHARACTERIZATION METHOD

In this section, we propose two Graph-theoretic based detection algorithms for align-along-road and unstructured patterns. Currently, semantics of buildings (e.g. detached/terraced house) is not available in most topographic datasets (Stoter et al., 2009b), and the semantic information is thus not discussed in this paper. It is nevertheless possible to integrate such information to refine detection results in future work.

This section first introduces the preliminary work which will be used by the subsequent detection and characterization (Section 3.1). Then the detection and automatic characterization for align-along-road (Section 3.2) and unstructured (Section 3.3) patterns are presented.

3.1 Basic computational tools

Refined constrained Delaunay triangulation The constrained Delaunay triangulation (CDT) plays an important role in the following detection and characterization, and hence some fundamental computations based on the CDT are introduced at first.

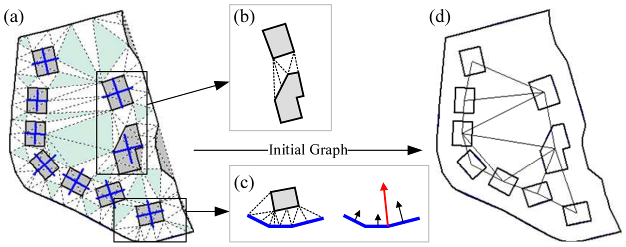


Figure 3: (a) Constrained Delaunay Triangulation takes buildings (grey polygons) and roads (bold lines) as constrained objects; (b) distances defined on the incident triangles between proximate objects; (c) segments of a road along which the building aligns and an illustration of the normal direction of this part of the road (the longest red arrow); (d) the initial proximity graph from (a)

As shown in Fig. 3(a), the CDT is built on buildings and roads, taking their outlines as constrained lines. The CDT is refined by inserting extra points to the constrained lines, and the interval between inserted points is based on the minimal distance between all data points. Two objects are considered as neighbors only when they are connected by edges of the triangles; the proximity relationship between two buildings, and between buildings and roads are explicitly modeled by this structure.

Based on the proximity relationship between buildings, an initial graph (Fig. 3(d)) can be generated from which the Minimum Spanning Tree (MST) can be automatically derived (see next section). Note that in the initial graph, and in the derived MST as well, all edges is actually weighted based on the proximity between building outlines, although for graphic presentations the edges connecting the centroids of buildings (i.e. vertices of the graph) are delineated. This means that the weights stored in the edges are calculated by the nearest distances between building outlines, instead of building centroids.

Several benefits can be drawn from the above CDT. Although being less efficient, it enables a better representation of proximity relationship than building a CDT on the centroids of buildings because it takes the shapes of buildings into account. In addition, because road centerlines are also involved in the CDT, the initial graph is automatically segmented in the sense that no graph edge intersects any dead end within a partition formed by the roads (Fig. 4).

As for the computational efficiency, the theoretical efficiency for the whole detection procedure (including constructing initial graph, deriving MST, detecting and characterizing building patterns) can be compensated for by the refined CDT. For example, the calculation of the nearest distance between two proximate objects can make use of their incident triangles (white triangles in Fig. 3(a), (b) and (c)), enabling a faster computation. That is, for each incident triangle between two objects, a distance (d_i) can be computed from the triangle edge that coincides with one of the objects to the third vertex; the nearest distance between the two objects is the minimum amongst all distances ($Min(\{d_i\})$). The time complexity of this calculation is $O(t)$, where t is the number of incident triangles between two objects; while a trivial computation of the nearest distance between two buildings (as was used in Regnauld, 1996) takes $O(nm)$ time, where n and m are the numbers of the points of the two buildings respectively. This speeds up greatly the nearest distance calculation between two buildings (Fig. 3(b)) and especially between a building and a nearby road (Fig. 3(c)), since t is generally much smaller than nm that can be expected for two spatial objects.

Other required calculations are as follows. First, the wall statistical weighting method (Duchêne et al., 2003) was implemented for computing building orientation; they showed that this method suits well for describing the orientations of typical man-made features (see also the crosses in the buildings in Fig. 3(a)). Second, the normal direction of a portion of a road (the longest arrow in Fig. 3(c)) is computed based on the average of all the segments normal vectors (three shorter arrows in Fig. 3(c)) weighted by segment lengths. The two calculations are crucial for the detection of align-along-road patterns. The last calculation is shape index (AGENT, 1999), which is required to characterize the homogeneity of shape.

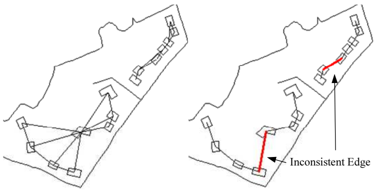


Figure 4: Initial graph, its derived MST and inconsistent edges

Minimum Spanning Trees In graph theory, a *spanning tree* of an undirected graph G is a tree that contains all vertices of G; the weight of a tree is defined as the sum of the weights of all its constituent edges. A *minimum spanning tree* of G is then a spanning tree whose weight is the minimum among all spanning trees of G. Since a graph may not necessarily be connected (Fig. 3a), it has a union of minimum spanning trees for its connected components (Fig. 3b). In this work, the vertices of G represent building features; the edges record their proximity relationships and are weighted by the nearest distances. Prim's algorithm (Prim, 1957) which is more efficient than Kruskal's algorithm (Zahn, 1971) was implemented to derive MST from initial graph.

A core concept for MST in pattern recognition is so-called inconsistent edge proposed by Zahn (1971), where he provided evidence that MST and inconsistent edge are perceptually significant in the point set clustering. In general, an *inconsistent edge* is an edge of MST whose weight is significantly larger than the mean of nearby edge weights on both sides of the edge (e.g. bold edges in Fig. 4). It can be defined as follows:

$$edge_i = \begin{cases} inconsistent, & if\, w_i > I_l \cap w_i > I_r \\ consistent, & otherwise \end{cases} , \quad (2)$$

where I is a measure that can be defined on both left (I_l) and right (I_r) sides of $edge_i$:

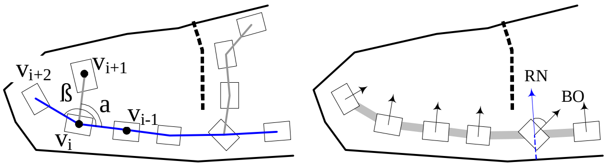
$$I = max\,\{f \cdot mean_{weight},\; mean_{weight} + n \cdot sd_{weight}\} \quad .$$

An edge is said to be inconsistent if its weight exceeds the mean weight ($mean_{weight}$) of its neighboring edges (within p depth) on both sides by n units of standard deviation (sd_{weight}) and further more if it is at least f times as large as both 2.0n weights (Zahn, 1971, p. 82). Given a normal distribution, there is less than 1% chance that an edge weight would exceed $mean_{weight}$ by 3 unit of sd_{weight}. Therefore, $n \geq 3$ is regarded to be statistically significant. In some cases where sd_{weight} approaches to zero, the factor $f > 1$ ensures that the inconsistent edge is still 'outstanding' compared with $mean_{weight}$. A detailed discussion on the parameterization issue refers to Zahn (1971).

However, inconsistent edge alone is not sufficient for detecting building patterns. We did experiments and found that no matter how to adjust and combine the parameters, most of the building patterns do not automatically show themselves up by simply cutting the inconsistent edges off the MST except for some rough clusters. The reason is mainly that the MST-based techniques, which only enforce the proximity principle (Zahn, 1971), do not always lead to a result that resembles human perception. Therefore other principles of perceptual organization must also be integrated for further processing depending on the kind of pattern to be detected. For the detection of collinear or curvilinear patterns, the principle of good continuity is much preferred.

The parameterization of inconsistent edge seems to be not very critical in our work. Since further processing is required for detecting specific patterns anyway, a conservative set of parameters was primarily used in this work (i.e. $p = 2, n = 3, \text{and } f = 2.0$), making sure that some very significant edges are pruned while other less significant ones are kept for further decisions. However, our experiments show also that the use of other sets of parameter values with small variations makes little difference for the final detection results, as the difference caused in the pruning step can be made up for by refined processes (see Section 3.2).

3.2 Align-along-road building pattern



(a) concepts for the detection of align-along-road pattern

(b) concepts for the characterization of align-along-road pattern

Figure 5: Related concepts for detecting and characterizing align-along-road pattern

There is a general Graph-theoretic based framework for the detection of all linear alignments in our work, that is, the detection is achieved by tracing paths in the previously obtained MST, and the trace of the paths has to conform to different sets of constraints appropriate for the detection of each type of patterns. Formally, a *path* is defined here as a sequence of vertices in a MST (e.g. the bottom line going through the buildings in Figure 5(a)); and the *path angle* at vertex v_i is defined by the angle of $vector(v_{i-1}, v_i)$ and $vector(v_i, v_{i+1})$ in that path (Figure 5(a)).

Detection of the align-along-road pattern The basic idea of detecting align-along-road patterns is that a path should be traced from the pruned MST and the buildings on the path should be close enough to a nearby road. Therefore, we formulate the detection of the align-along-road pattern as a path tracing procedure that advocates the same aligned road, proximity, size similarity and path angle constraints (Algorithm 1). This choice of constraints set is justified as follows.

As a result of studying existing maps, we find that shape and orientation are not as dominant as aligned road, proximity, and size in determining a collinear group. First, that buildings align along the same road can be recognized as a align-along road pattern is a fundamental constraint, which should never be violated. Second, most buildings have similar man-made shapes; those with complex shapes are usually much bigger than the ordinary buildings, and thus they can easily be filtered out by size similarity in the first place. In addition, a generic shape measurement for map generalization is still not available; commonly used shape measures (e.g. compactness, shape index) describe specific aspects of shape and thus are not sufficient for our case. Third, the theoretically maximum deviation of building orientation computed with the wall statistical weighting method is 45° (Duchêne et al., 2003), which has much less impact on the perception of align-along-road patterns than proximity constraint. Consequently, we discarded the use of shape and orientation constraints in our detection algorithm of align-along-road pattern in the experiments. Nevertheless, these two aspects are integrated in the characterization of this pattern.

Algorithm 1: Detecting the align-along-road pattern

```
Input: buildings; partition roads; $constraints$
Output: $Collection$ of align-along-road patterns ($AARP$)
1: calculate CDT and then MST for the input data;
2: prune the inconsistent edges from the MST;
3: trace paths in the pruned MST as follows:
foreach $v_0$ = vertex of degree 1 or vertex of degree 3 in MST do
    loop "initialize $AARP$"
        $v_1$ = adjacent vertex of $v_0$;
        if $v_0$,$v_1$ align along the same closer road then
            initialize $AARP$ with $edge(v_0, v_1)$;
            break;
        else
            $v_0$ = $v_1$;
    endloop;
    loop "add new buildings to $AARP$"
        if $v_1$ is of degree 1 then
            add $AARP$ to $Collection$ if $AARP$ contains more
            than 3 buildings;
            break;
        if $v_1$ is of degree 3 then
            select a vertex $v_2$ with which the $edge(v_1, v_2)$
            forms larger path angle with its predecessor edge.
        else
            $v_2$ = adjacent vertex of $v_1$
        check $edge(v_1, v_2)$ w.r.t. the $constraints$ for $AARP$;
        if $edge(v_1, v_2)$ exists then
            add $edge(v_1, v_2)$ to $AARP$;
            $v_0$ = $v_1$; $v_1$ = $v_2$; $v_2$ = NULL;
        else
            add $AARP$ to $collection$ if $AARP$ contains more
            than 3 buildings;
            empty $AARP$; renew $AARP$ with $edge(v_1, v_2)$ if
            the two vertices share a road;
            $v_0$ = $v_1$; $v_1$ = $v_2$; $v_2$ = NULL;
    endloop;
4: combine the detected $AARP$ if possible;
```

The mechanism how the selected constraints work is described as follows. These constraints work on a new edge occurring in the tracing process in 1, determining if the new edge can be added to an existing alignment. Aligned road is firstly checked using the information stored as a result of constructing refined CDT. If buildings connected by the new edge share the same road, then the tracing proceeds; and vice versa. The proximity constraint is based on the idea of inconsistent edge introduced in Section 3.1: if the weight of the new edge is inconsistent concerning the existing pattern, it cannot be added to the existing pattern. Finally, the size similarity constraint ensures that the size contrast between the two buildings at both ends of the new edge should not be too large (i.e. bigger building/smaller building $< size\_contrast$). It has to be mentioned that in the iteration step where there are more than one new edges available (e.g. $edge(v_i; v_{i+1})$ and $edge(v_i; v_{i+2})$ in Figure 5(a)), an edge with larger path angle ($\alpha, \beta \in [0°, 180°]; \beta > \alpha$) should be selected and added to the existing pattern in order to keep good continuity principle (Wertheimer, 1923).

In the step 4 of Algorithm 1, a combination is recommended in the cases of two detected groups approach to each other at their ends. This combination can be done by introducing a connecting edge consisting of two vertices at the proximate ends of both groups, if on the one hand the two groups align along the same road and on the other the distance of the introduced connecting edge is not too long. The parameter showing promising results

```
Algorithm 2: Detecting unstructured clusters
Input: buildings; partition roads [optional]; $postconditions$
Output: $Collection$ of unstructured clusters ($UC$)
1: calculate CDT and then MST for the input data;
2: prune the inconsistent edges from the MST;
3: prune the edges connecting buildings whose size difference $> size\_contrast$; if $FragRatio > 0.5$, the edges should not be pruned;
4: organize connected subgraphs and populate $UC$ with them;
5: filter out those $UC$ that cannot pass $postconditions$, and add remaining $UC$ to $Collection$;
```

for align-along-road pattern (Section 4) is $size\_contrast = 3.2$.

Characterization of the align-along-road pattern The characterization of align-along-road building patterns is by applying Equation 1 to spacing, size, shape, and distance to the aligned road. For these homogeneities, mean values are firstly calculated for the properties and then the homogeneity is computed. It is worth noting that both spacing (i.e. inter distance between buildings) and distance to the aligned road are computed using the nearest distance calculation forth mentioned in Section 3.1.

The calculation of the homogeneity concerning the orientation is described as follows. The orientation is considered to be more regular if the buildings change their orientations right according to the normal directions (e.g. RN in Figure 5(b)) of the local road segments that they align. For each building in the pattern, an angle deviation $\in [0°, 45°]$ is calculated between RN (Normal of Road segments) and BO (Building Orientation), the calculation of the two is presented in Section 3.1; the homogeneity of orientation is then computed from the standard deviation of all the deviations.

3.3 Unstructured clusters

Detection of unstructured clusters Unstructured clusters are also a common feature on topographic maps, especially at larger scales (1:10k-1:50k).

The method (Algorithm 2) detecting this type of building pattern is realized simply through pruning edges that are inconsistent and edges that connect two buildings whose size difference exceeds $size\_contrast$, and grouping the connected subgraphs from the pruned MST. Those subgraphs, however, are just candidates which have to tested against several *postconditions*. t The first *postcondition* is the number of buildings contained. in our experiments, we define that at least three buildings form a cluster or pattern, as only in this case the calculation of mean and standard deviation required by the detection and characterization is meaningful. A second *postcondition* is so-called black-and-white ratio, which is defined as follows:

$$BWRatio = \frac{\sum Area(b_i)}{Area(ConvexHull(UC))} \quad ,$$

where $b_i \in UC$. This *postcondition* protects some wriggling linear alignments from being recognized as unstructured clusters (Figure 6(a)). A final *postcondition* is termed fragmentation ratio ($FragRatio$), which is the ratio between the number of pruned edges and the number of total edges in an initial cluster. This *postcondition* should be applied in step 3 (see Algorithms 2) to prevent clusters from being over-fragmented, as in the cases where small and big buildings are arranged alternately (e.g. Figure 6(b)).

Characterization of unstructured clusters Despite the homogeneity of orientation which is calculated based on absolute orientations of buildings, the homogeneity of spacing, size, shape is exactly the same to those calculated for align-along-road patterns. The algorithms for the detection and characterization were implemented and results are presented in Section 4.

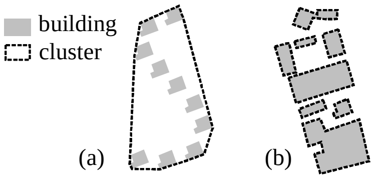


Figure 6: Bad examples of unstructured clusters

4 IMPLEMENTATION AND RESULTS

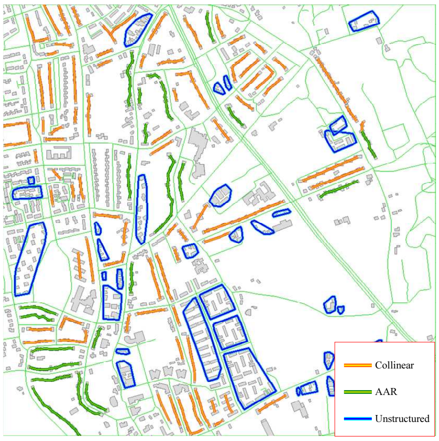
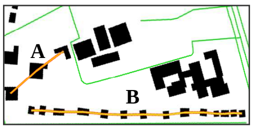


Figure 7: Test case and detection results of collinear, AAR, and unstructured patterns

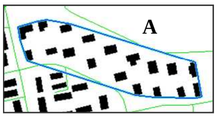


id	spacing	size	orientation	shape
A	0.06	0.32	9.74°	0.43
B	0.52	0.26	5.57°	0.11

Figure 8: Measured characteristics of collinear alignments

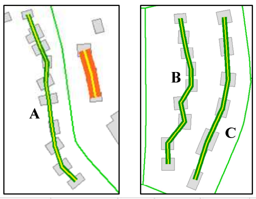
We implemented the two proposed algorithms in an interactive generalization prototype system using C++. In addition, an algorithm detecting collinear building patterns was also implemented. The test case and detection results for collinear, align-along-road (AAR), and unstructured patterns are shown in Figure 7. There are 151 patterns detected from the test dataset, where collinear alignments are 94, unstructured clusters are 34, and align-along-road patterns are 23. It is noticeable that some building groups can be e.g. both AAR and collinear alignments. In this implementation, the final pattern type of a cluster was decided according to the computed characteristics; the pattern type with smallest homogeneity values was assigned to the cluster.

The characterization results are shown in Figures 8, 9, and 10. A general observation is that the measure values for spacing, size, shape, and dis2road (values $\in [0, 1]$) and the values for orientation (values $\in [0°, 45°]$) are consistent with our perception of these detected patterns. For examples, the measured homogeneities of size, orientation, and shape for the collinear patterns in Figure 8 shows that alignment B is superior to A in these aspects; the homogeneous cluster shown in Figure 9 is confirmed



id	spacing	size	orientation	shape
A	0. 22	0. 22	2. 03°	0. 24

Figure 9: Measured characteristics of an unstructured cluster



id	spacing	dis2road	size	ori1	ori2	shape
A	0. 39	0. 13	0. 28	4. 34°	9. 94°	0. 15
B	0. 39	0. 78	0. 15	2. 4°	2. 36°	0. 09
C	0. 07	0. 11	0. 28	2. 14°	12. 95°	0. 18

Figure 10: Measured characteristics of align-along-road patterns

by the measured homogeneities of this unstructured cluster, as all the homogeneity values are relatively low (below a quarter).

For the align-along-road patterns as shown in Figure 10, several observations can be made. First, 'dis2road' column confirms that this characteristic is more homogeneous for A and C than B, as B is aligned along the road to its left. Second, an extra orientation ('ori2') is calculated based on absolute values, in order for the readers to compare it with the orientation change with aligned roads ('ori1'). The results show that 'ori1' is generally less than 'ori2', as we can also observe in Figure 10 that although the individual orientations are rather fluctuated the orientations indeed change according to their aligned roads respectively. This observation confirms that orientation varying according to the aligned road is well suited for characterizing align-along-road pattern.

5 DISCUSSION AND CONCLUSION

This paper proposes a typology of building patterns available in cartographic and geospatial domain based on the study of existing topographic maps and related literatures. In this typology, fundamental visual patterns have been identified and their characteristics been discussed. This is an important step towards formalizing our knowledge on the building patterns in this field. Then, two Graph-theoretic based algorithms are presented in order to detect two pattern types of the typology, namely the align-along-road and unstructured patterns. The detection and characterization method was implemented and tested using Dutch topographic datasets. The results appear to be promising.

The proposed approach is generic in the sense that visually important building patterns can be detected and characterized no matter the buildings are topologically adjacent or not, even if the spatial objects are represented by points, since all these cases can be handled by the refined CDT. This approach can be further applied to detecting the patterns of archipelago, by adjusting some of the constraints or postconditions of the algorithms.

Although we argue in this paper that in most cases collinear and curvilinear alignments can be replaced by align-alogn-road patterns, the characterization of align-along-road patterns is still too general. That is, some important characteristics, like the straightness and main angle of collinear alignments, and smoothness and curvature descriptions of curvilinear patterns, would be lost if those linear patterns were recognized as align-along-road patterns. Therefore, future work will be focus on the detection and characterization of curvilinear alignments.

Further testing of the proposed approach against non-Dutch datasets will be carried out in order to confirm the claimed generality. Also noted that another work on the automated evaluation of building pattern preservation constraint (Zhang et al., 2010) is carrying out based on the detection results reported in this paper.

ACKNOWLEDGEMENTS

The work was partially supported by the National High-Tech Research and Development Plan of China under the grant No. 2007-AA12Z209, and the National Science Foundation, China under the grant No. 40971242. International Institute for GeoInformation Science and Earth Observation (ITC), University of Twente, which funds the first author as a PhD student is gratefully acknowledged.

REFERENCES

AGENT, 1999. Selection of basic measures, deliverable c1. Technical report.

Anders, K.-H., 2006. Grid typification. In: Progress in Spatial Data Handling, pp. 633–642.

Boffet, A. and Rocca Serra, S., 2001. Identification of spatial structures within urban block for town qualification. In: ICC, Vol. 3, Beijing, pp. 1974–1983.

Christophe, S. and Ruas, A., 2002. Detecting building alignments for generalisation purposes. In: D. E. Richardson and P. van Oosterom (eds), Advances in Spatial Data Handling (SDH 2002), Springer Verlag, Berlin, pp. 419–432.

Duchêne, C., Bard, S., Barillot, X., Ruas, A., Trévisan, J. and Holzapfel, F., 2003. Quantitative and qualitative description of building orientation. In: 5th Workshop on Progress in Automated Map Generalization, Pairs.

Lüscher, P., Weibel, R. and Burghardt, D., 2009. Integrating ontological modelling and bayesian inference for pattern classification in topographic vector data. Computers, Environment and Urban Systems 33(5), pp. 363–374.

Regnauld, N., 1996. Recognition of building clusters for generalization. In: M. J. Kraak and M. Molenaar (eds), Advances in GIS Research II (Proceedings of 6th SDH'96, Delft), Taylor & Francis, London, pp. 4B.1–4B.14.

Ruas, A. and Holzapfel, F., 2003. Automatic characterisation of building alignments by means of expert knowledge. In: ICC, Durban, pp. 1604–1515.

Stoter, J., Burghardt, D., Duchene, C., Baella, B., Bakker, N., Blok, C., Pla, M., Regnauld, N., Touya, G. and Schmid, S., 2009a. Methodology for evaluating automated map generalization in commercial software. Computers, Environment and Urban Systems 33(5), pp. 311–324.

Stoter, J., van Smaalen, J., Bakker, N. and Hardy, P., 2009b. Specifying map requirements for automated generalization of topographic data. Cartographic Journal, The 46(3), pp. 214–227.

Wertheimer, M., 1923. Laws of organization in perceptual forms. In: W. D. Ellis (ed.), A Source Book of Gestalt Psychology, Routledge & Kegan Paul, pp. 71–88.

Zahn, C. T., 1971. Graph-theoretical methods for detecting and describing gestalt clusters. Computers, IEEE Transactions on C-20(1), pp. 68–86.

Zhang, X., Stoter, J., Ai, T. and Kraak, M.-J., 2010. Formalization and data enrichment for automated evaluation of building pattern preservation. In: Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science (SDH' 2010). to appear.

FORMALIZATION AND DATA ENRICHMENT FOR AUTOMATED EVALUATION OF BUILDING PATTERN PRESERVATION

Xiang Zhang [a, b,] *, Jantien Stoter [c, d], Tinghua Ai [b], Menno-Jan Kraak [a]

[a] ITC, University of Twente, the Netherlands
{xzhang, kraak}@itc.nl
[b] School of Resource and Environment Sciences, Wuhan University, China
tinghua_ai@tom.com
[c] Delft University of Technology, the Netherlands
[d] Kadaster, Apeldoorm, the Netherlands
j.e.stoter@tudelft.nl

KEY WORDS: Automated Evaluation, Building Pattern Preservation, Pattern Matching, Constraint Formalization

ABSTRACT:

Automated evaluation of generalization output relies to a large extent on that requirements (e.g. specifications, constraints) being formalized in machine-readable formats. Previous studies suggest that the formalization and automated evaluation are relatively easier for legibility constraints (improve the readability of maps) than for preservation constraints (preserving important real-world phenomena). Three major difficulties, i.e., pattern classification and characterization, pattern matching, and constraint formalization, in the automated evaluation of building pattern preservation constraint are analyzed in this paper. A classification of available building patterns is reviewed based on a previous work. In addition, the transition events describing allowed changes for building patterns to preserve during generalization are obtained through the study of existing maps series (from 1:10k to 1:100k). Based on the obtained knowledge on pattern types and acceptable transition events, an approach to automatically match corresponding building patterns at different scales is presented. The methodology proposed is validated by applying it to the interactively generalized data. The result shows promising results and also further improvement in order to apply the method in an overall evaluation to indicate acceptable generalization solutions.

1. INTRODUCTION

In the map generalization process small scale map is generated from a large scale map. This intelligent information management process involves a combination of data reduction and simplification related techniques in order to suppress unnecessary detail. At the same time geographic patterns are emphasized to achieve a clear view of information that resembles the original data as much as possible. Because building patterns are significant for topographic maps, preservation of building patterns is an important cartographic constraint in the generalization process. It aims at keeping important real-world entities by discerning interesting patterns such as urban and rural structures.

This paper studies the automated evaluation of building pattern preservation in generalization. Automated evaluation of generalization output aims to assess (i.e. measure) whether or to what extent the output satisfies the cartographic constraints according to automatically derived indicators (e.g. size and shape). Automated evaluation relies firstly on the formalization of the specifications (i.e. constraints). From related work (Burghardt et al., 2007; Stoter, et al., 2009a) we can conclude that preservation constraints (e.g. on networks, patterns, and spatial distributions) are more difficult to formalize and to evaluate automatically than legibility constraints such as minimum dimension of an object required to distinguish it on the map. In this paper, formalization and data enrichment techniques for the automated evaluation of building pattern preservation are examined. In this evaluation, generalized building patterns are compared with the patterns in the original data to see if the generalized patterns meet the specification of building pattern preservation.

Difficulties in automated evaluation of building pattern preservation are manifold. First, the existing specifications concerning building patterns originally intended for interactive generalization are not easy to formalize since they are specifically meaningful for cartographers. A cartographic constraint says for example that building alignments should be preserved, which can be interpreted by cartographers so that they can apply generalization according to their knowledge or experience. Formal knowledge for computers to measure and characterize the patterns and to describe their change at scale transitions is required for automated evaluation but is not yet available. Therefore, existing specification need to be enriched for automated processes.

Second, building patterns are not stored as database objects in common topographic datasets. Consequently to automatically evaluate generalized datasets on building pattern preservation, the datasets have to be enriched with building pattern objects identified through pattern recognition techniques, or visual inspection..

A third difficulty in automating the evaluation of building pattern preservation is the lack of explicit links between correspondences at different scales. In the context of the evaluation of generalization output, the links are especially necessary for the automated evaluation of preservation

* Corresponding author. Email: xiangzhangchina@gmail.com

constraints (Stoter et al., 2009a) to allow the system to know which pattern objects at different scales represent the same building group in reality.

The links can be created via generalization operators as part of the generalization process (during the process it is still known which representations are generalized into which representations at smaller scales) and via data matching (Hampe et al., 2004). The latter (data matching) is the only choice in our case as the links generated by generalization processes are not available for the datasets to be evaluated.

The following sections focus on addressing the above described difficulties.

2. CLASSIFICATION AND CHARACTERIZATION OF BUILDING PATTERNS

This section reviews a typology of different building patterns which is discussed in detail in our another paper (Zhang et al., 2010), where algorithms to detect them are also presented. A UML class diagram of the proposed typology is shown in Figure 1.

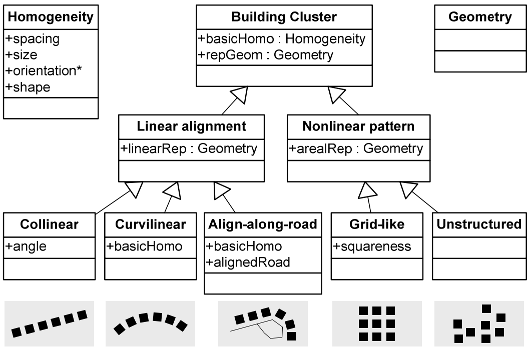


Figure 1. A UML model of the proposed pattern classification and examples of each pattern type

The UML diagram models the patterns of the typology with their representational geometries and characteristics (which can automatically be detected by the algorithms proposed in Zhang et al. (2010)). We distinguish five types of patterns, namely three linear patterns (collinear, curvilinear, and align-along-road alignments) and two nonlinear patterns (grid-like patterns and unstructured clusters). Properties all these types have in common are the representational geometries (denoted as 'repGeom') and the homogeneities (denoted as 'basicHomo') in the UML model. Two types of representational geometries are used:

- Linear representations (denoted as 'linearRep'): Skeletons of linear alignments; the skeletons are generated from the MST of building centers (Zhang et al., 2010);
- Areal representations (denoted as 'arealRep'): Convex hulls of nonlinear patterns; buffers of the above-mentioned skeletons in some cases (see Section 3.3).

The common attribute 'basicHomo' consists of the homogeneity of spacing (between proximate buildings), size, orientation, and shape. These homogeneities can be calculated using the concept of standard deviation, reflecting the regularity of these properties (Ruas and Holzapfel, 2003). Note however that orientation should be implemented differently for curvilinear and align-along-road patterns. This is why these two classes override 'basicHomo' from their parent class. For example, the orientation is homogeneous for a collinear pattern if all buildings in the pattern have the same orientation; while it is homogeneous for a curvilinear pattern only if the orientations of buildings vary right according to the 'path' (Figure 2(2)). Besides, specific characteristics in the model are explained: the 'angle' of a collinear alignment is the main angle of its skeleton; for each align-along-road pattern, a 'alignedRoad' should be assigned; the 'squareness' characterize the degree to which buildings align in rectangles.

It has to be noted that, these characteristics are not used to classify groups of buildings into different patterns; the pattern types are instead detected using different ad-hoc algorithms. The characteristics are calculated after the patterns have been detected, to describe the quality of the detected patterns.

3. METHODOLOGY FOR IDENTIFYING, FORMALIZING AND EVALUATING BUILDING PATTERN PRESERVATION

Now the building patterns are detected from data and characterized by homogeneity measures, the next step is to assess whether the generalized map preserves the homogeneity properties of the initial patterns in an acceptable way. This requires firstly knowledge on which patterns at different scales represent the same real-world entities and how these change at scale transitions (addressed in Section 3.1 to 3.3), and secondly ways to (a) automatically evaluate the constraints and (b) to interpret the quantitative evaluation results (Section 3.4).

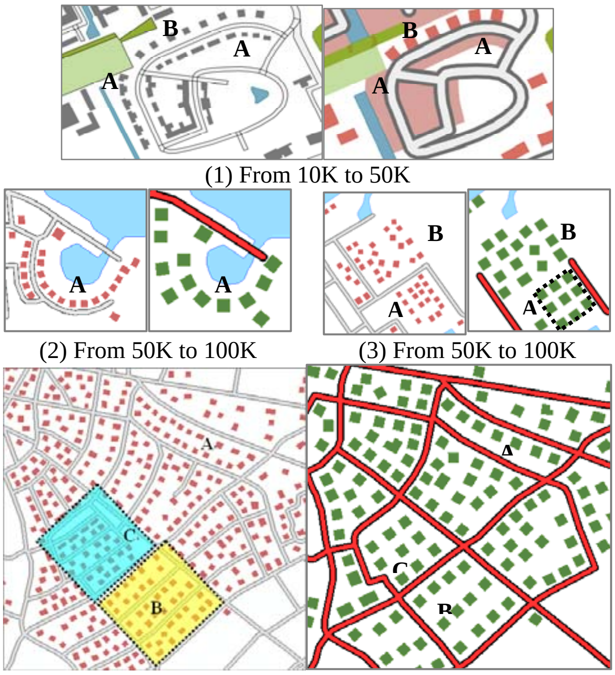
3.1 Identification and formalization of changes of building patterns during generalization

A visual analysis of topographic map series was carried out for different purposes: (1) to understand why and how building patterns change at scale transitions; (2) to enrich the available written specifications for generalizing building patterns. In this analysis we identified the homogeneity classes based on pattern characteristics in the original data and try to match these with patterns in the generalized map to describe and quantify the change of patterns at scale events, i.e. do they diminish, are they preserved or are they transformed?

For this visual analysis two scale transitions (1:10k to 1:50k and 1:50k to 1:100k) in the topographic map series of the Netherlands have been studied. The map at scale 1:10k (supported with an object oriented database, called *TOP10NL)* is the most detailed topographic data, from which map at scale 1:50k (supported with *TOP50NL* database) is interactively generalized. The map at scale 1:100k (supported with *TOP100NL* database) is interactively generalized from 1:50k map. By comparing building patterns at these transitions, the following knowledge on their changes (as used by cartographers in the interactive generalization process) was obtained (see for examples Figure 2).

In general, we observe three forms of pattern changes at scale transitions. First, some building patterns are diminished (undetectable). Second, some patterns are transformed into built-up areas (e.g. 1A in Figure 2). The third important observation is that the group of buildings constituting a building pattern is generalized (i.e. simplified, typified, eliminated etc), resulting in a change of the pattern.

In our research on automated evaluation we specifically focus on tolerated changes to building patterns as a result of these changes, where the cases of building patterns being diminished or transformed into built-up areas are visually qualified as patterns being depressed and they are therefore out of the scope of this paper. Section 3.2 further explores how the patterns are preserved (or actually, which changes are allowed to preserve the pattern), resulting in formalized transition events.



(1) From 10K to 50K

(2) From 50K to 100K

(3) From 50K to 100K

(4) From 50K to 100K

Figure 2. The cases illustrating typical changes of building patterns at scale transitions in map series of the Netherlands' Kadaster

3.2 Formalizing observed transition events of pattern preservation

From closer observations of the above visual analysis of the map series, a set of transition events describing allowed changes for building pattern preservation are formalized in Table 1. Representative cases of each observed event are exemplified in Figure 2. Note that the multiplicity of the transition events is at the pattern level rather than at the building level, indicating one- or many-to-one relationships between the pattern objects. The transition events and their possible causal factors are detailed as follows.

Event 1 shows the simplest transitions (1:1 relationship), where each ungeneralized pattern type matches with a generalized pattern of the same type, because the pattern remains significant after the generalization of buildings. For example, an align-along-road pattern can be matched to another one of the same type (1B in Figure 2). It is the same for curvilinear patterns (2A in Figure 2) and for unstructured clusters (3B in Figure 2).

Event 2 can be observed in situations where an unstructured cluster in a larger scale map is enhanced (becoming a grid pattern) during the generalization (e.g. 3A in Figure 2). Events 3-5 describe the areas where the density of buildings enclosed by roads is relatively too high. Since major roads were kept during the generalization, some of the building patterns are replaced by a row of collinear pattern due to the competition of space and legibility constraints. In events 2-4, 1:1 relationships can be found between initial and generalized patterns (i.e. a pattern object is converted as a whole into another pattern object); while event 5 (4A in Figure 2) exhibits an n:1 relationship (n ≥ 2; several pattern objects are changed into one pattern object). Event 4 describes that a grid pattern can become a collinear alignment but not a curvilinear pattern after the generalization.

Events 6-8 describe the cases where the change of patterns is mainly caused by the removal of streets and roads between and surrounding the patterns. This usually occurs in urban areas where the road network is highly dense and roads are selected during generalization. In such areas building patterns in nearby blocks can be aggregated or typified into bigger ones and the n:1 relationship is usually observed (e.g. 4B in Figure 2). Event 8 is shown by 4C in Figure 2.

Align-along-road patterns are somewhat special, as they rely on aligned roads and are sometimes similar to collinear and curvilinear patterns except their relationships to roads. Event 9 describes a situation where align-along-road patterns change with the removal of the aligned roads, which leads to the change of the patterns into collinear or curvilinear alignments.

All the formulated transition events can be observed for both transitions (i.e. 10k to 50k and 50k to 100k). The observed events formalize the multiplicity relationships and allowed changes from one type of patterns into another to preserve important pattern characteristics. Although these transition events are empirically observed and are subject to further refinement, transitions that are considered as unacceptable (e.g. grid to unstructured) are excluded. As a result, they can be used to further enrich the available written specifications for building pattern preservation to make these suitable for automated generalization and evaluation.

In practice, the distinction between the events may not be as strict, mainly due to the uncertainty in the recognition and generalization of the patterns. A building pattern can be seen as an unstructured cluster, a grid pattern, or multiple rows of linear alignments; an align-along-road type can also be a linear alignment if the road were removed. However, this does not influence the matching because all possible transitions from one type into another are covered by the transition events while impossible ones are excluded. For example, 2A in Figure 2 can be seen as either 'curvilinear to curvilinear' (event 1) or 'align-along-road to curvilinear' transition (event 9); 4B can be described either as three unstructured clusters changing into three rows of linear alignments (event 3) or as six rows of linear alignments changing into one grid pattern (event 6).

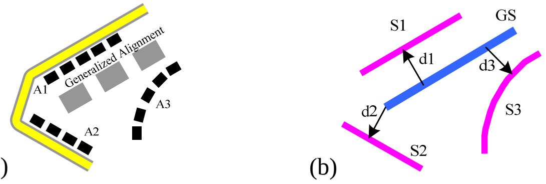
Table 1. Observed transition events of building pattern preservation at scale transitions

Event	**1:10k**	**Multiplicity**	**1:50k**	**Examples shown**
	1:50k		**1:100k**	**in** Figure 2
1.	Each type of patterns	1:1→	The same type	1B; 2A; 3B
2.	Unstructured cluster	1:1→	Grid pattern	3A
3.	Unstructured cluster	1:1→	Linear alignment	4B
4.	Grid pattern	1:1→	Collinear alignment	x
5.	Linear alignments	n:1→	Linear alignment	4A
6.	Linear alignments	n:1→	Grid pattern	4B
7.	Linear alignments	n:1→	Unstructured cluster	x
8.	Nonlinear clusters	n:1→	Nonlinear cluster	4C
9.	Align-along-road pattern	1:1→	Collinear/curvilinear pattern	2A

3.3 Method for automatic matching process

The matching process consists of two sub-processes: geometric and characteristic matching, which is based on two kinds of previously enriched information, namely the pattern descriptions and the knowledge on the transition events. The pattern descriptions include pattern type information, representational geometries and homogeneity properties (Section 2). The representational geometries of the detected pattern objects are used mainly in the geometric matching process. The obtained transition events are in general used as part of the characteristic matching process: when a generalized pattern object is geometrically matched with an initial pattern object, type information concerning the two patterns is checked with respect to the transition events. If the type information is consistent with the transition events, the matching pair is selected as a candidate for further matching; if otherwise the matching of the two fails. In this latter case, one can conclude that the pattern is not preserved during generalization. It is worth noting that the buildings to be evaluated are partitioned by road networks and each step of the automatic matching is restricted to buildings within a partition cell. The technical detail of the process is described as follows.

The geometric matching deals with the similarity between geometries of the patterns to be matched. In the matching of the same type patterns (event 1), the similarity can be measured by distances like nearest distance, Hausdorff distance, and Fréchet distance (Alt and Godau, 1995). To match polygons (nonlinear patterns), another similarity measure (i.e. contrast model) developed by Tversky (1977) is appropriate. In the cases events 3-7 where linear alignments are matched with nonlinear clusters or several linear alignments are matched to one linear alignment, buffers of skeletons should be used instead of skeletons alone for representing linear alignments to improve the matching result. In our first experiment (Section 4), we used nearest distance and the contrast model as similarity measures to match between linear and polygonal representational geometries respectively, for simplicity reasons.



(a) Matching between linear alignments

(b) Geometric representation of the patterns

Figure 3. A scenario illustrating the problem of matching building patterns

The result from geometric matching can be refined by characteristic matching, especially when the result contains several candidates. When matching the patterns of the same type, the characteristics like pattern type, align angle, curvature and aligned road are of great importance. For example, Figure 3a describes a possible case where a generalized collinear alignment (gray buildings in Figure 3a) is (potentially) matched with two collinear patterns (A1 and A2) and a curvilinear one, as the distances (d1, d2 and d3 in Figure 3b) are similar. Such a case is possible due to the displacement of A1 caused by the widened symbol of the road. In this case, pattern type is firstly used to filter out the curvilinear pattern A3, and then align angle is calculated based on the skeletons of the alignments (angle of GS is similar to S1 than to S2), leading to a correct match between the generalized alignment and A1.

3.4 Automated evaluation and interpretation of the evaluation results

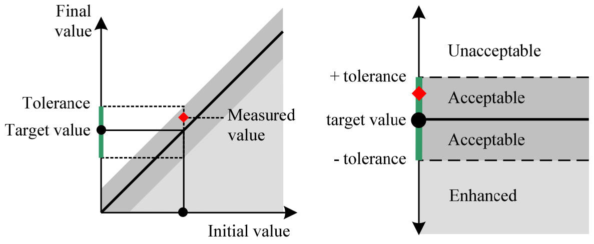
3.4.1 Evaluation by computing changes between matched building pattern objects

After all the required information has been collected, the automated evaluation is carried out by firstly computing characteristics defined in Section 2 for each matched building pattern objects that represent the same real-world building group, and then comparing the measured characteristics between the two pattern objects. The interpretation of the quantitative comparison is described in Section 3.4.2.

To be more concrete, the computed characteristics for both initial and generalized building patterns could be a set of separate values (e.g. homogeneity of space, size, etc.) or an aggregated value (summary of separate homogeneity values). The implementation (Section 4) is based on separate characteristics in order to demonstrate the main idea. Future work will discuss and apply the aggregation of the characteristics. Note also that if matched patterns are of the same type, then all their (common and specific) characteristics are well-matched, and then the evaluation is simply comparing their measured homogeneity values in a pair-wise manner; if they are of different types, only 'basicHomo' is compared. In the cases where n (n≥2) initial patterns are linked to one generalized pattern, the initial homogeneities are computed by weighted average of each initial pattern; the weighting is based on lengths of skeletons for linear patterns, and on areas of convex hulls for nonlinear patterns.

3.4.2 Interpret the quantitative evaluation results

The next key issue of the automated evaluation of building pattern preservation is to decide the acceptable values for the pattern characteristics after the generalization, based on which the quantitative evaluation results can be qualified into a human readable format (e.g. 'acceptable' or 'unacceptable'). According to Bard (2004) an idealized evolution function (Figure 4a) can be specified for each preservation constraint, where target characteristic values should be equal to initial values. In order to be more flexible, an acceptable range (tolerance) is introduced (dark gray areas in Figure 4).



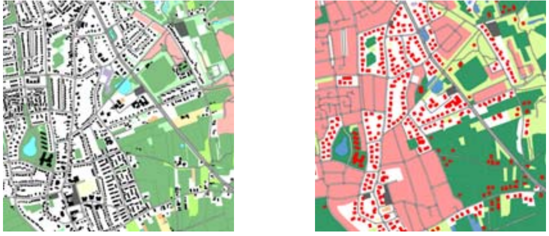
(a) Evolution function (b) Interpretation function

Figure 4. Idealized evolution function and interpretation function of building pattern preservation constraints (modified from Bard, 2004)

In our research, we slightly modify the interpretation function as proposed by Bard (2004) to be more appropriate for the pattern preservation constraint as follows. If a measured homogeneity property falls into the dark gray area in Figure 4b (|MeasuredVal – TargetVal| ≤ tolerance), then this property is considered as being well preserved (marked as ‘acceptable’). This is motivated by the fact that small deviations are tolerated by human eyes. If the value is larger than ‘target value’ by a unit of tolerance, the preservation is regarded ‘unacceptable’; while if the value is less than ‘target value’ by a unit of tolerance (light gray area in Figure 4b), the property is regarded to be ‘enhanced’ rather than unacceptable. This is because the building patterns can be improved by reducing the homogeneity values (i.e. improve the regularities).

4. IMPLEMENTATION AND VALIDATION OF THE METHODOLOGY

This section implements and validates the methodology as presented in Section 3 by applying them to a test case, to show the feasibility of the concepts. The test case consists of two datasets different from the datasets used for the visual analysis (Section 2): one is a Dutch topographic dataset at 1:10k (Figure 5a); another dataset (*TOP50NL*) is interactively generalized from the *TOP10NL* (see Figure 5b). The idea is that the evaluation results from applying the methodology to interactively generalized datasets should indicate that the building patterns detected in the data are preserved sufficiently. The test results of automatic pattern matching (Section 4.1) and of validation the automated evaluation (Section 4.2) are presented.



(a) Extract of *TOP10NL* (b) Extract of *TOP50NL*

Figure 5. Test case for validating our evaluation methods

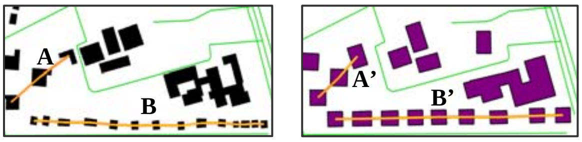
4.1 Results of pattern matching

We applied the method for automatic matching as described in Section 3.3. First, the patterns were detected using the algorithms in Zhang et al. (2010); then the data matching was carried out. In most cases, one partition cell contains only one to two detected patterns, so around 94% of the patterns (48 out of 51) detected in *TOP50NL* were correctly matched with their correspondences in *TOP10NL*. There are 53% of the patterns (68 out of 129) detected in *TOP10NL* mismatched. This is expected since most of the mismatch is caused by buildings transformed into built-up areas (light red areas in Figure 5b); the remaining mismatch is due to the absence of correspondences in *TOP50NL*. In the latter case the building patterns may have been diminished to satisfy other important constraints (e.g. minimum distances). There are also cases where some patterns (e.g. unstructured clusters) were correctly matched with others of different types (e.g. collinear alignments) and where n:1 relationships were correctly created.

4.2 Results of characterizaion and evaluation

The homogeneities were calculated for all matched patterns, and some rather simple examples (1:1 relationship) are illustrated (see Figure 6 and Figure 7). Some of the homogeneities were computed using coefficient of variance (std/mean) rather than standard deviation (std), as is suggested by Zhang et al. (2010). Homogeneity of orientation, however, should be computed by std, as it is a cyclic variable and is thus meaningless to calculate the coefficient of variance for it.

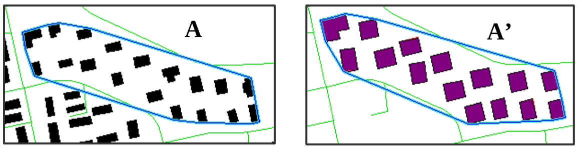
These figures show that the illustrated patterns are enhanced during the generalization and that at the same time the values for the homogeneity of spacing, size, and orientation are generally reduced. These results show that the proposed methods for automated evaluation of building pattern preservation are promising. In order to see if the methods can draw meaningful statement about the whole test case, regression analysis was carried out for all matched patterns. In a second step regression functions indicating the relationships between initial and target values were established concerning the homogeneity of spacing, size, and orientation (Figure 8).



Collinear alignments in *TOP10NL* Collinear alignments in *TOP50NL*

ID	Spacing (std/mean)	Size (std/mean)	Orientation (std)
A	0.06	0.32	9.74°
A’	0.02	0.11	8.99°
B	0.52	0.26	5.57°
B’	0.35	0.09	0.00°

Figure 6. Changes of the characteristics of linear patterns at a scale transition



Unstructured cluster in *TOP10NL* Unstructured cluster in *TOP50NL*

ID	Spacing	Size	Orientation
A	0.22	0.22	2.03°
A’	0.19	0.19	1.17°

Figure 7. Changes of the characteristics of unstructured clusters at a scale transition

These figures confirm that homogeneity values of the patterns are significantly reduced (most of the data points are below the idealized preservation function in Figure 8). This reduction means that the patterns become more homogeneous analysis of the test datasets: most of the preserved building patterns become more regular in terms of spacing, size, orientation, and shape. This is mainly due to the preservation constraint (i.e. preserve or enhance the homogeneity); legibility constraint mainly increases the minimum size and simplifies the shape of the buildings.

However, the figures also show the homogeneity values for orientation are not significantly reduced. We can see that around half of the data points in Figure 8 (leftmost) are above the idealized function, and most of the deviations are less than 5°. This fact can be explained because deviations ranging from 0° to 10° are less detectable for human eyes, and hence it is acceptable to introduce this noise during the interactive generalization. A second reason why some orientation homogeneity values increase is that some patterns detected in *TOP50NL* add new buildings as their elements, and these newly added buildings contribute a lot to the rise of orientation std values. In addition, due to the minimum distance constraint in some cases, some buildings in a pattern rotate themselves, and this also increases the orientation homogeneity. In order to formalize allowed deviations for the homogeneity of orientation, a tolerance should be derived from the training data using statistical analysis.

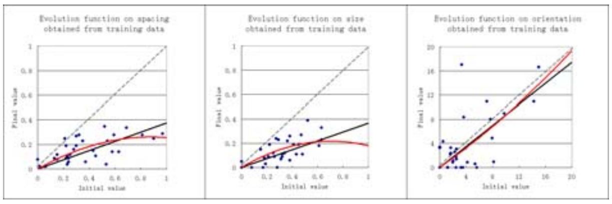


Figure 8. Evolution functions obtained from the test datasets; black line – linear function; red curve – nonlinear function; dotted line – idealized preservation function

5. DISCUSSION AND CONCLUSIONS

This paper studies supported methods and techniques for automated evaluation of the building pattern preservation constraint. The automated evaluation of this constraint evaluation is currently impeded by issues concerning pattern detection, pattern matching, and constraint formalization. We address these issues by firstly summarizing previously obtained results in pattern detection and characterization (Zhang et al., 2010), on which the pattern matching and evaluation are based. The classification and characterization of building patterns formalizes our knowledge on building patterns. The proposed methodology also contributes to the automated evaluation of preservation constraints in general.

Then we find that the matching of building patterns is not straightforward in the sense that a type of pattern may become another type. As a result, acceptable changes (transition events) for building patterns are formulated by studying existing map series (see Table 1). This is the second valuable finding of this work, and thereby we suggest that these events could enrich the written specifications for automated processes. Later, a two-step matching process is proposed based on the enriched pattern descriptions and the transition events.

The presented methodology are implemented and validated by applying them to interactively generalized data. The major conclusion is that the methodology is promising in automated evaluation of building pattern preservation as confirmed by the validation results. The results confirm again that the pattern detection and characterization methods in Zhang et al. (2010) are consistent with our perception; the detection method is applicable to non-Dutch maps as various patterns were successfully detected in maps of France, Spain, and China. The results also indicate that the automatic matching is in itself sufficient to automatically evaluate the constraint in the case of patterns being preserved; it suggests also that a more generic framework and automatic measures would be appreciated to cover the automated evaluation of building patterns being diminished or transformed to built-up areas, although the pattern preservation constraint can be identified visually as being violated for these two cases.

For the interpretation method, the evaluation functions and tolerances in our current approach have to be decided by users and the function forms are still oversimplified as reveal by the test results. A solution to this deficiency is to determine a more adaptive evolution functions from training data. Therefore further research will employ this solution to optimize the evolution functions in order to better interpret the evaluation results.

Future work can integrate the evaluation approach with the evaluation of other constraints to obtain overall values, indicating acceptable generalization solutions. In addition, the current formalized transition events still have some overlaps and could be further optimized.

ACKNOWLEDGMENTS

The work is supported by the National High-Tech Research and Development Plan of China under the grant No. 2007AA12Z209, and the National Science Foundation, China under the grant No. 40971242. International Institute for GeoInformation Science and Earth Observation (ITC), University of twente, which funds the first author as a PhD student is gratefully acknowledged.

REFERENCES

Alt, H. and Godau, M., 1995. Computing the Fréchet distance between two polygonal curves. *Int J of Computational Geometry and Applications*, 5(1-2), pp. 75–91.

Bard, S. and Ruas, A., 2004. Why and How Evaluating Generalised Data? In: *Developments in Spatial Data Handling (SDH'04)* (Springer-Verlag), pp**.** 327-342.

Bobzien, M., Burghardt, D., Petzold, I., Neun, M. and Weibel, R., 2008. Multi-representation Databases with Explicitly Modeled Horizontal, Vertical, and Update Relations. *Cartography and Geographic Information Science*, 35(1), pp. 3-16.

Burghardt, D., Schmidt and S., Stoter, J., 2007. Investigations on cartographic constraint formalisation. *In 10th ICA workshop on generalisation and multiple representation*, Moscow, 2007.

Hampe, M., Anders, K. and Sester, M., 2003. MRDB applications for data revision and real-time generalization. In *Proceedings of the 21st International Cartographic Conference*, pp. 192-202.

Mascret, A., Devogele, T., Berre, I. L. and Hénaff, A., 2006. Coastline matching process based on the discrete Fréchet distance. In: Riedl, A.; Kainz, W. & Elmes, G. A. *(ed.), Proceedings of the 12th International Symposium on Spatial Data Handling*, pp.383–400.

Ruas, A. and Holzapfel, F., 2003. Automatic characterisation of building alignments by means of expert knowledge, ICA, Durban, pp. 1604-1615.

Stoter, J., et al., 2009a Methodology for evaluating automated map generalization in commercial software. *Computers, Environment and Urban Systems*, 33(5), p. 311-324.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, pp. 327–352.

Zhang, X., Ai, T., and Stoter, J., 2010. Characterization and detection of building patterns in cartographic data. *Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science (SDH' 2010)*, Hong Kong, to appear.

A METHOD USING ESDA TO ANALYZE THE SPATIAL DISTRIBUTION PATTERNS OF CULTURAL RESOURCE

Dongying ZHANG [a*], Xiajun MAO [a], Lingkui MENG [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.
yineast@mail.whu.edu.cn, xiajun.mao@gmail.com, lkmeng@whu.edu.cn

Commission VI, WG VI/4

KEY WORDS: Spatial Statistics, Clustered Qualification, ESDA, Spatial Distribution Pattern, Spatial Association, Spatial Autocorrelation, Cultural Geography

ABSTRACT:

The spatial distribution pattern of cultural resource generally manifests clustered qualification to some extent, and we can infer the vicissitudes of the correlative culture by analyzing the changing rule of the spatial distribution pattern of cultural resource. However, the majority of the concerned researches are still at the level of qualitative statistics and thematic map visualization of some general features at present, without taking the characteristic into account. Aiming at making up the shortage, the paper therefore analyzes the spatial distribution of a special type of cultural resource - the ancient Toponym which is defined as the name of the places where the immigrants settled down by using Exploratory Spatial Data Analysis (ESDA) methods. We first made both qualitative analysis and quantitative analysis to confirm the existence of the clustered qualification by computing Global Moran's I and Global Geary's C values, later we made a further study about the similarity of different clustered regions and told them apart. At last, we came to a conclusion that not only the spatial distribution pattern of cultural resource is usually clustered, but also the similarity degree of the clustered regions differs from one another. In addition to revealing the spatial patterns of the Toponym distribution, this paper promotes an explicitly spatial view that has certain methodological implications for the application of spatial statistic methods in Cultural Geography research.

1. INTRODUCTION

In order to increase the scale of the population and develop the local economy in Sichuan Region, the government of Qing Dynasty adopted emigration policy which lasted more than 100 years, it's the well-known legend called "HuGuang Tian Sichuan" policy in the history, which not only affected the population distribution of China strongly, but also promoted the amalgamation of emigration cultural and local cultural directly.

The Toponym has accompanied the history in a long run, it's the cultural patrimony which not only reflect local geographical environment, but track the emigration of the ancient race and their war， it looks like a live fossil that can play the role of the indicator of culture (Zhengxiang Cheng 1992). The Toponym is the visual cultural sight that can not only reflect the distribution of emigration directly, but can be mined for plenty of cultural content (Shangji Situ 1983).

As a special type of cultural resource, the Toponym plays a key role in the research of historical geography, cultural geography and Geographical information science, and its importance has been recently recognized. For example, some historical geographers have statistically analyzed the quantity of the Toponym in every administrative unit during Qing Dynasty in Sichuan Region so as to estimate each part's immigrants and the percentage by their descent, ignoring the spatial interaction between the Toponym. With spatial techniques, GIS experts have already visualized the spatial distribution of the Toponym by using spatial interpolation and obtained clustered areas where people say the same dialect using point-based cluster analysis methods (Fahui Wang 2009). Recent researches have proved the spatial distribution patterns of cultural resources qualitatively. However, we need spatial distribution patterns expressed quantitatively as the spatial correlation ratio to support the economic and cultural researches, the higher ratio the more developed economic and smaller cultural difference. This article takes the ancient Toponym quantity into research, with ESDA to analyze the spatial distribution patterns of cultural resource so as to promote the application of spatial analysis methods in the research of Cultural Geography.

2. MAJOR ANALYSIS ISSUES AND ANALYSIS MEANS

2.1 Major Analysis Issues

Exploratory spatial data analysis (ESDA) has its origins in exploratory data analysis (EDA) which is a term coined by the American statistician John Tukey in the 1970s to describe statistical procedures used by applied statisticians when they

[*] Corresponding author.

were in the first stages of analyzing a new set of data. ESDA can be considered the extension of EDA methods to spatial data. Often, ESDA is used to identify data properties for four purposes as follows: detecting spatial patterns in data, detecting possible data errors (‘outliers’ and ‘spatial outliers’), formulating hypotheses based on the geography of the data and assessing spatial models. According the first law of Geography: ”Everything is related, but things nearby are more related than things far away”, we know that spatial association is inherent in geographic data, when working on spatial data, analyses based on regular statistics are very likely to be misleading or incorrect, There is positive spatial association when high or low values of a random variable tend to cluster in space and there is negative spatial association when geographical areas tend to be surrounded by neighbors with very dissimilar values, all of which consists of spatial patterns, and therefore the main aim of ESDA is for patterns detection. Spatial patterns include three types: collected pattern, randomized pattern and dispersed pattern.

In the analysis of the spatial distribution of the Toponym, the major concerns are to reveal spatial patterns. The distribution of the Toponym is intrinsically spatial and, moreover, space-dependent due to the potential interactions of the long term movement of emigration. The spatial distribution pattern of the Toponym may be a reflection of the vicissitudes of the local culture and the amalgamation of emigration culture and local culture.

2.2 Major Analysis Means

Spatial analysis and techniques for measuring spatial association have been proposed in the literature. Getis and Ord family of Gi(d) statistics (Getis and Ord 1993; Ord and Getis, 1995) and Anselin’s LISA(Local Indicators of Spatial Association) (Anselin 1995) are two basic local statistics of spatial association. Compared to LISA, Gi(d) statistics is more simple in detecting places with unusual concentrations of high or low values (i.e., ‘hot’ or ‘cold’ spots). On the other hand, techniques for spatial heterogeneity include the expansion method (Casetti 1972; Jones and Casetti 1992), the method of spatial adaptive filtering (Foster and Gorr1986; Gorr and Olligschlaeger 1994), the random coefficients model (Aitken 1996), the multilevel modelling (Goldstein 1987), the moving window approach (Fotheringham et al. 1997) and geographically weighted regression (GWR) (Brunsdon et al. 1996; Fotheringham et al. 1997). However, GWR is relatively a simple but effective technique for exploring spatial heterogeneity which allows different relationships existing.

In this article, we mainly use global Moran and Geary, local G-Statistics and LISA methods to analysis the degree of spatial autocorrelation of the Toponym’s attributes data from global regions to local counties.

3. A BRIEF REVIEW OF GISA AND LISA

3.1 GISA

GISA is short for Global Indicators of Spatial Association, it mainly includes Moran’s I indicator and Geary’s C rate.

3.1.1 Moran’s I

$$I=\frac{n\sum\sum w_{ij}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)}{\sum\sum w_{ij}\sum\left(x_i-\bar{x}\right)^2} \tag{1}$$

where x_i, x_j =the observed value at location (i, j)
$\bar{x}$ = the average of the $\{x\}$ over the n locations
w_{ij} = the spatial weight measure defined as 1 if location i and j are adjacent, or else as 0
i = contiguous to location j and 0 otherwise

The expected value and variance of the Moran I for samples of size n could be calculated according to the assumed pattern of the spatial data distribution (Cliff and Ord 1981, Goodchild 1986).

For the assumption of a normally distribution:

$$E_R(I)=\frac{-1}{n-1} \tag{2}$$

$$V_1=\frac{n\left[\left(n^2-3n+3\right)S_1-nS_2+3W^2\right]}{(n-1)(n-2)(n-3)W^2} \tag{3}$$

$$V_2=\frac{k\left[\left(n^2-n\right)S_1-nS_2+3W^2\right]}{(n-1)(n-2)(n-3)W^2} \tag{4}$$

$$\mathrm{var}_R(I)=V_1-V_2-\left[E_R(I)\right]^2 \tag{5}$$

$$z_{Moran}=\frac{I-E_R(I)}{\sqrt{\mathrm{var}(I)}} \tag{6}$$

where
$$W=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij} \tag{7}$$

$$S_1=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij}+w_{ji}\right)^2}{2} \tag{8}$$

$$S_2=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{j=1}w_{ij}+\sum_{i=1}w_{ji}\right)^2 \tag{9}$$

$$k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

(10)

The Moran I is significant and positive when the observed value of locations within a certain distance tend to be similar, negative when they tend to be dissimilar, and approximately zero when the observed values are arranged randomly and independently over space (Goodchild 1986).

3.1.2 Geary's C

Geary' s C statistic is defined as the following:

$$C = \frac{(n-1)\sum\sum w_{ij}(x_i - x_j)^2}{2\sum\sum w_{ij}\sum(x_i - \bar{x})^2} \qquad (11)$$

Where x_i, x_j =the observed value at location (i, j)
$\bar{x}$ = the average of the $\{x\}$ over the n locations
w_{ij} = the spatial weight measure defined as 1 if location i and j are adjacent, or else as 0
i = contiguous to location j and 0 otherwise

The Geary statistic is always positive and asymptotically normal. The hypothesis for the Geary statistic test is that the mean of the Geary statistic is 1 if there is no spatial autocorrelation. A significant and low value (between 0 and 1) indicates a positive spatial autocorrelation while a significant and high value (greater than 1) indicates a negative spatial autocorrelation (Cliff and Ord 1981).

Spatial Pattern	*Geary's C*	*Moran's I*
Clustered Pattern	$0 < C < 1$	*I > E(I)*
Random Pattern	*C* ~= 1	*I ~ = E(I)*
Dispersed Pattern	$1 < C < 2$	*I > E(I)*

Table 1 Three Types of Spatial Pattern

According to Table 1(David W.S. Wong, Jay Lee 2005), the expectation stands for no spatial autocorrelation. To Geary's C, the Expectation is close to 1, that's, when C ~= 1, it illustrates the points' distribution is randomised model, when 0 < C < 1, the points' distribution is Clustered model. We consider there is probably negative spatial autocorrelation when Moran's I is smaller than the Expectation while probably positive spatial autocorrelation when Moran's I greater than the Expectation.

3.2 LISA

LISA is short for Local Indicators of Spatial Association, mainly including G Statistics, local Moran and local Geary. The G statistics (Ord and Getis 1992; Getis and Ord 1994) and LISA (Anselin 1995) provide measures for the experiments of the local spatial association. Local Moran and local Geary statistics, as suggested by Anselin (1995), are alternative local indicators. The local Moran allows for the identification of spatial agglomerative patterns similar to G statistics, while the local Geary allows for the identification of spatial patterns of similarity or dissimilarity. One advantage of the local Moran and the local Geary is that they can be associated with the global statistics (Moran I and Geary C) and can be used to estimate the contribution of individual statistics to the corresponding global statistics.

3.2.1 Local Moran

The local Moran statistic for each observation i is defined as (Anselin 1995)

$$I_i = z_i \sum_j w_{ij} z_j$$

(12)

where the observations z_i and z_j are the correspond deviation of x and $\bar{x}$:

$$z_i = \frac{x_i - \bar{x}}{\delta}$$

(13)

and w_{ij} is mostly row-standardized:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = n$$

(14)

The interpretation of the local Moran is similar to the G statistic (Getis and Ord 1992). A small p-value (such as p <0.05) indicates that location i is associated with relatively high values of the surrounding locations. A large p-value (such as p > 0.95) indicates that location i is associated with relatively low values in surrounding locations.

3.2.2 Local Geary

A local Geary statistic for each observation i may be defined as follows (Anselin 1995)

$$c_i = \sum_j w_{ij}(z_i - z_j)^2 \qquad (15)$$

where S 2 is the same as before, Z_i and Z_j are standardized values, and w_{ij} are the elements of a row standardized spatial weights matrix. If each observation contributes equally to the global statistic, each local Moran or local Geary should be. The individual significance of the C_i statistic can be obtained by the same permutation approach used for Ii outlined above.

The calculation of the pseudo-significance level p-value is similar to that of local Morans. A large p-value (such as $p > 0.95$) indicates a small C_i in extremes, which suggests a positive spatial association (similarity) of observation i with its surrounding observations, while a small p-value (such as $p < 0.05$) indicates a large C_i in extremes, which suggests a negative spatial association (dissimilarity) of observation i with its surrounding observations.

4. DATA AND METHODS

4.1 Data Collection

The Map data of 141 point-shaped counties and boundaries of 26 prefectures of Sichuan Region in China were attained from China Historical geographical information system (CHGIS). CHGIS is a historical GIS database on ancient China from Qin Dynasty to Qing Dynasty. It identifies the location of prefectures, and attempts to draw the boundaries of county--the sub-state administrative units, the four layers are county point file, prefecture point file, the boundaries of the regions file and the boundaries of prefecture file.

The attributes data of 141 counties and 26 prefectures mainly come from a series of 《SichuanXianZhi》from the early days to the middle years of Qing periods, for the mass emigration began at Kangxi Period, ended at Jiaqing Period, and from some history literatures about the Toponym(Yong Lan 1995).

4.2 Data Processing and Analysis

Before the experiment, two important issues about the attribute data must be solved.

On one hand, according to the attribute data (The quantity of the Toponym) of the point layer of 141 counties, the attribute data are different from each other, though some points with these attributes were neighbours. The standardization of the column of the data should be taken because some values in the column may be 0. If the data values keep owing the original ones, it's almost impossible to calculate the true value of the global indicators, for these data appearing randomized more than normalized while the T-value is based on the normalized maximum simulation theory that need the data set to be obeying normal distribution.

On the other hand, an in-depth hotspot analysis is carried out to identify the "hotspot" and "cold spot" areas based on the logarithms (the quantity of Toponym) at "county" level. Different weight matrices are tested: inverse distance weighted, fixed distance bands of 5, 10, 15, 20 and 25 kilometers and so on. The best weight matrix is considered to be a fixed distance band of 15 kilometers. There are three reasons: first, the inverse distance weighted which fits the area observation well is too large because it may encompass all the observations in the study area while a fixed distance band of 5 kilometers is too small to encompass any observation in the study area; second, with reference to the average distance of counties which are adjacent, a fixed distance band of 15 kilometers is more reasonable; third, the experiment using the fixed distance band of 15 kilometres has a better result including relatively more significantly clustered areas.

And it's now possible to carry on the experiment following three steps: calculate the spatial autocorrelation coefficient and variance and expectation and Z value, then evaluate the statistical significance with this value, and then compare the different resulted thematic map and discover the spatial pattern of them.

5. RESULTS AND DISCUSSION

5.1 Global Moran and Geary

Global Moran and Geary (Getis-Ord general G) uses the randomization "z" statistic to evaluate the existence of clusters in the spatial arrangement of the given samples and show the level of significance with the rule that if the "z" statistic value is greater than the key value 1.96 then we consider the significance level of the given samples is 5%, and then if the "z" statistic value being even greater than another key value 2.576 will lead to a higher significance level of 1%.
In the experiment, we normalized the quantity of the Toponym in each of 141 counties and each row of the spatial weight matrix and chosen 15km as the fixed band distance, and calculated Moran's I and Geary's C values in the table below:

	I or *C*	*E(I)* or *1*	*Variance*	*Z Value*
Moran's I	0.0532	-0.0071	0.0002	4.8743
Geary's C	0.5799	1	0.0110	2.7629

Table 2 The Values of Global Moran's I and Geary's C

From Table 2 we see that the "z" statistic value of Moran's I is 4.8743, which is bigger than 2.576, that's the significance level is 1%, showing us that the distribution pattern of the Toponym of Sichuan region after the mass emigration is a high cluster and has high spatial autocorrelation. The Geary's C value in Table 2 also shows the existence of strong spatial autocorrelation in the research area with C value 0.5799 and "z" value 2.7629.

5.2 Hot spot and cold spot analysis（G_i）

Being a composite index, Global Moran and Geary is the measure of the overall clustering of the data, used to evaluate the overall spatial association of the total research area. But it is reasonable for us to consider that the spatial autocorrelation level of different census area is not exactly the same. For this reason, we use a local indicator called Gi (Getis-Ord Gi*) to detect and evaluate the spatial autocorrelation of local census area, high Gi means the census area is a cluster of high ratio while low Gi means the census area is a cluster of low ratio.

During the experiment, we have standardized the quantity of the Toponym in each of 141 counties, standardized each row of the spatial weight matrix, chosen 15km as the fixed band distance, and calculated Local G_i values which are visualized as the point layer, and classified the points in seven kinds which are marked different colours according to Z value range in legend "Couty_Point" in Figure 1, meanwhile, we have standardized the quantity of the Toponym in each of 26 prefectures, and built their spatial weight matrix calculated by the inverse weighted distance, standardized each row of it, and calculated G_i data which are visualized as the area layer, and classified the polygons in seven kinds which are marked different colours according to Z value range in legend "Prefecture_area" in Figure 1 below:

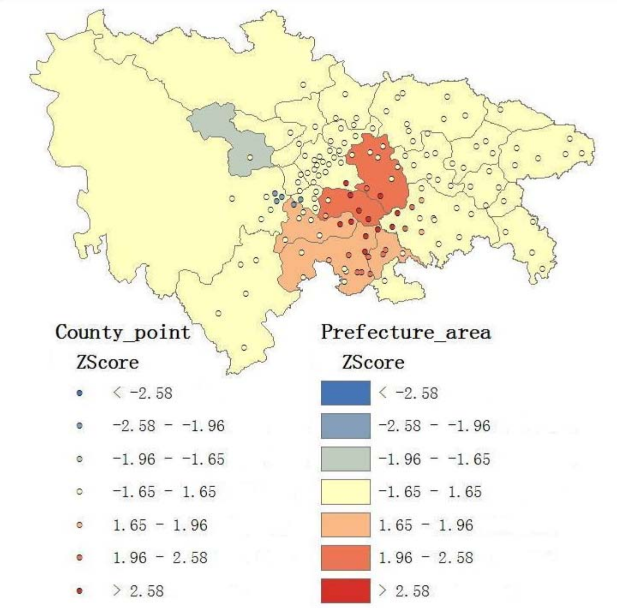


Figure1. Thematic Map Using Local G-Statistic Method

We use a local indicator called G_i (Getis-Ord G_i^*) to detect and evaluate the spatial autocorrelation of local census area, high G_i means the census area is a cluster of high ratio while low G_i means the census area is a cluster of low ratio.

Global Moran and Geary (Getis-Ord general G) uses the randomization "z" statistic to evaluate the existence of clusters in the spatial arrangement of the given samples and reveals the level of significance with the rule that if the "z" statistic value is greater than the key value 1.96 then we consider the significance level of the given samples is 5%, and then if the "z" statistic value being even greater than another key value 2.576 will lead to a higher significance level of 1%.

Figure 1 shows us that the similar points with high ratio in red clustered in the red area, while the points with sub-high ratio in orange clustered in the orange area. That's, the hot spot area of the point layer matches well with the hot spot area of the prefecture polygon layer. And it further revealed that the majority of counties in the three red prefectures lies in the south of Sichuan region had a highly developed emigration culture after the ancient mass emigration.

5.3 Local Moran Analysis

Global spatial autocorrelation analysis yields only one statistic to summarize the total research area. In other words, global analysis assumes spatial homogeneity. As we all know that spatial heterogeneity is one type of spatial association which is a basic feature in geographic researches so that assumption does not hold, then having only one statistic does not make sense as the statistic should differ over space. Furthermore, we can still find clusters at a local level using local spatial autocorrelation if there is no global autocorrelation or no clustering.

During the experiment, we have standardized the quantity of the Toponym in each of 141 counties, standardized each row of the spatial weight matrix, chosen 15 km as the fixed band distance, and calculated Local Moran's I values which are visualized as the point layer, and classified the points in seven kinds which are marked different colours according to Z value range in legend "Couty_Point" in Figure 1, meanwhile, we have standardized the quantity of the Toponym in each of 26 prefectures, and built the their spatial weight matrix calculated by the inverse weighted distance, standardized each row of it, calculated Local Moran's I values which are visualized as the area layer, and classify the polygons in seven kinds which are marked different colours according to Z value range in legend "Prefecture_region" in Figure 1 below:

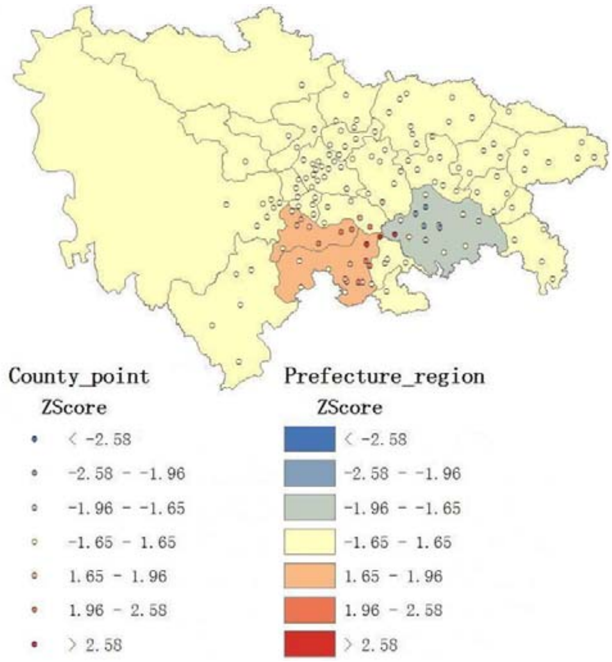


Figure2. Thematic Map Using Local Moran Mean

Local Moran Indicator just evaluates the similarities and dis-similarities of clusters and it cannot tell us whether the degree of the clusters is great or small. In Figure 2, red points denote there are points with similar ratios clustering together and the red region in the south also matches well with the hot spot area in Figure. On the contrary, blue points in the blue areas in the southeast reflect the majority of the counties of the prefecture have dissimilar emigration culture development patterns from one another.

6. CONCLUSION

First, the Global Moran and Geary analysis proved strong spatial autocorrelation in the spatial distribution of the Toponym of Sichuan region which showed that one point may be surrounded by other points with similar attributes to itself in other words, the closer one point to another the more similar the two points are. It could be helpful in the Cultural Geography research work taking spatial association into consideration; Second, the hot spot analysis shows spatial association of the Toponym with hot spot area in red and cold spot area in blue, which further indicates the existence of sub-areas that developed differently over space in the cultural domain. In the emigration cultural research area, We tentatively interpret the hot spot as the most developed emigration culture areas and the cold spot as the most undeveloped; Third, Local Moran analysis illustrates that the most developed culture areas in the cultural research region tend to be a cluster which indicates that the spatial pattern of cultural resources distribution does exist with similar features clustering together.

The application of spatial analysis methods based on ESDA in analyzing the spatial distribution patterns of cultural resources has shown us that it is possible to use spatial statistic methods in cultural domain and proved that geographic location factors and spatial associations could be used in Cultural Geography research. These two thematic maps before-mentioned are illustrations of spatial autocorrelation of the distribution of the ancient Toponyms in Sichuan region after the mass emigration. Spatial statistic methods appeared therefore as a powerful tool to reveal the characteristics of cultural regions in sub-administrative unit (i.e. prefecture) is in relation to those of its geographical environment and yield scientific explanations for spatial distribution patterns.

REFERENCE

Zhengxiang Cheng, 1983. *China Cultural eography • Taiwan's Toponym* — Neolithic cultures.Sanlian Bookstore.

Shangji Situ, 1992. The Historical geographical Research of GuangDong's Toponym. *Comments on Chinese Historical Geography*.

Yong Lan, 1995. The Research on Geographical features of the Distribution of Local Residents and Emigration in Qing period in Sichuan Region. *Comments on Chinese Historical Geography*.

Fahui Wang, 2009. GIS based quantitative methods and their applications. *The Commercial Press*, pp. 51~61, 208~210.

Anselin, L., 1995. Local Indicators of Spatial Association - LISA, *Geographical Analysis* 27, pp. 93-115.

Bao, Shuming and Mark S. Henry. 1996. Heterogeneity issues in local measurements of spatial association. *Geographical Systems*, Vol. Ⅲ, pp: 1-13.

Cliff A. and Ord, J.K. 1973. *Spatial Autocorrelation*. Pion, London.

Cliff, A. D. and Ord, J. K., 1981. *Spatial Processes: Models and Applications*. Pion, London.

Cressie, Noel A, 1993. *In Statistics for Spatial Data*, John Wiley & Sons, Inc. pp. 79-122.

David W.S. Wong, Jay Lee, 2005. *Statistical Alalysis of Geographic Information with Arcview GIS and ArcGIS*, John Wiley & Sons, Inc. pp. 302-367.

Cressie, Noel A. and Hawkins, D.M., 1980, Robust estimation of the variogram, *I.Journal of the International Association for Mathematical Geology*, 12, pp. 115-125.

Getis, Arthur and Ord, J. Keith. 1996. Local Spatial Statistics: An Overview. In Spatial Analysis: Modeling in a GIS Environment, *Geoinformation International.* P. Longley and M. Batty (eds.), Cambridge, UK.

Getis, Arthur and Ord, J. Keith. 1995. The Use of a Local statistic to Study the Diffusion of AIDS from San Franciso, *Regional Science AssociationInternational in Cincinnati*.

Getis, Arthur and Ord, J. Keith. 1992. The Analysis of Spatial Association By the Use of Distance Statistics. *Geographical Analysis*, 24, pp. 189-206.

Goodchild, M. F., Haining, R. P. and Wise, S. 1992. Integrating GIS and spatial data analysis: problemsand possibilities. *In International Journal of Geographical Information Systems* 6(5), pp. 407-423.

LAYOUT OPTIMIZATION OF URBAN UNDERGROUND PIPELINE BASED ON 3D DIGITAL CITY

Jianchun He[a,b], Jinxing Hu[a,*], Qingyuan Tang[a], Shanshan Guo[a,b]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[b] School of Geoscience and Environment Engineering, Central South University, Hunan, China

Commission VI, WG VI/4

KEY WORDS: Underground Pipeline, Layout Optimization, GIS, 3D Digital City, Network Analysis, Buffer Analysis

ABSTRACT:

With the fast development and utilization of underground space resource in cities, the spatial distribution of urban underground pipelines becomes more and more complex. How to efficiently utilize the latest technology, realize underground space resources' management, visualization, layout optimization and mine applications, are challenging issues in current digital city development. Taking into account of the frequently changing and complicated information of underground pipeline, we are facing great difficulty in data management. With the comprehensive integration management requirement of 3D digital city and underground pipeline system, we present layout optimization based on three-dimensional (3D) digital city.

The objective of this paper is to describe design and implementation of underground pipeline 3D visualization and layout optimization based on 3D digital city. Firstly, the automatic conversion from two-dimensional (2D) pipeline data to 3D model is realized and we also integrates management with ground data, such as 3D house, 3D model, DEM, vector data, RS image, etc. which can quickly determine the relative positions of underground pipelines in a more intuitive way. Secondly, in accordance with related national standards, and the ensured accuracy and validity of pipeline data, an underground pipeline layout optimization method is realized, which is on the basis of spatial analysis and principle expanding. It analyzes the pipeline boundary layout in order to guarantee the distribution rationality; also it can ensure optimization of the pipeline layout. Finally, take the data of a city for example, an underground 3D pipeline optimization system based on 3D digital city is developed, which is according to the overall architecture and optimization method. It offers a set of techniques for the pipeline layout optimization and simulation, and provides decision support for urban underground resources management, pipeline planning, and urban planning, etc.

1. INTRODUCTION

Underground pipeline is an important component of urban infrastructure, which is called city's "lifeline". It responds to fulfil energy transference and material transportation, which is the basis of city's survival and development [1]. With fast expansion of cities, urban material flow and energy flow are increasing significantly, and the intensity and density of urban underground pipelines are also dramatically increasing. How to manage underground pipeline dynamically and effectively and to study the pipeline layout optimization method is a hot research issue, especially the study of pipeline layout optimization.

Currently, most management of underground pipelines is manual, or in two dimensional. It has disadvantages on intuitive pipeline visualization, low efficiency and difficult dynamical management, which consequently lead to pipeline accidents frequently. How to meet the requirements of management departments and construction units has become imperative [2]. It is significant to explore layout optimization method of underground 3D pipeline, recognize the optimization principles and provide technical support and decision making for urban underground resources management, pipeline planning and construction of 3D digital city. During the development of current digital city, underground pipeline layout optimization is not only helpful for related department interacts with underground pipeline to facilitate their construction, but also helpful for researchers use latest technology effectively to realize layout optimization of underground space resources and make better management.

2. RELATED RESEARCH

At present, urban underground pipeline system has been focus on visualization, spatial analysis and route optimization of 3D pipeline. With continuous development of technology, people are gradually changing their attention on pipeline from 2D environment to 3D platform. However, the representation of the variability of underground pipeline is still lacked. So, some new method is urgent to deal with the uncertainty and vagueness of the layout of pipeline.

Urban Pipe Network Visualization System (UPNVS) is designed and organized by spatial metadata, which is based on the analysis of the features of urban pipe network and the requirements of the 3D model; 3D spatial data model is the basic of data representation and spatial visualization [1]. Pipeline visualization is always a challenge, UPNVS, which compared to traditional 2D visualization, has clearly reviewed

[*] Corresponding Author: Jinxing Hu; jinxing.hu@sub.siat.ac.cn; phone 86-755-86392373

3D visualization urban underground pipeline is much more intuition.

A new methodology is proposed for automated route selection for the construction of new power lines which is based on geographic information system. It used a dynamic programming model for route optimization. Environmental restrictions are taken into account together with all of the operating, maintenance, and equipment installation costs [3]. It describes the method of GIS spatial analysis which is applied to electric line routing optimization. This methodology is used for generate a new economic lines by the selection of route, at the same time, geographic factors and evaluating the uncertainties is consider to associated with routing costs. So, GIS spatial analysis is identified as the powerful tool to develop automatic reveal of the layout of urban underground pipeline.

According to the characteristics of urban underground pipeline, we investigate the principles of pipeline layout optimization, and by the application of GIS spatial analysis. Pipeline layout optimization is integrated management of urban underground 3D pipeline by using the technology of 3DGIS. As a result, the method of urban underground 3D pipeline layout optimization is put forward in this paper.

3. LAYOUT OPTIMIZATION OF URBAN UNDERGROUND PIPELINE BASED ON 3D DIGITAL CITY

3.1 Flow Chart of Pipeline Layout Optimization

This paper discusses establish of underground 3D pipeline model dynamically. And based on 3D digital city, we propose the layout optimization method of underground pipeline.

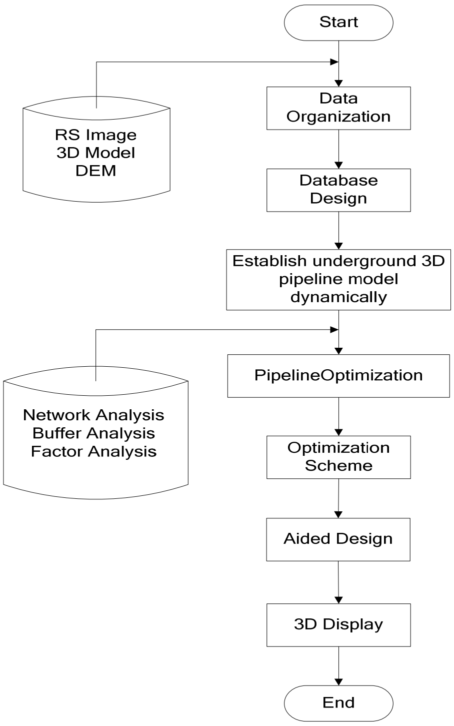


Figure 1. Flow chart of pipeline layout optimization

As show in Figure 1, it describes the process of urban underground pipeline layout optimization. According to relevant requirements, we firstly organize and manage the data of remote sensing image, digital elevation models, and 3D building models, etc. and then import them into 3DGIS system. Secondly, with the data of pipe node, tube-well, and 2D pipeline, etc. We implement underground 3D pipeline dynamically, which is related with ground-level structures, furthermore, locate positions of pipe sections quickly is in a more intuitive way. Thirdly, with the existing road network, we select the starting and ending points of the pipeline, by utilizing GIS spatial analysis to generate several pipeline layout schemes automatically. Fourthly, based on national norms and standards, collision analysis of pipeline layout is put forward, which is used to search for the adjacent pipelines. Besides, horizontal and vertical spacing distances between pipelines and facilities are measured. Finally, according to horizontal and vertical spatial analysis between pipelines and facilities, we consider the factors of the process of underground pipeline layout optimization, propose the solution to lay out underground pipeline optimally, and reveal the results of optimization analysis in 3D environment.

3.2 Data Organization

All the about data of pipe node, well and 2D pipeline should be imported into spatial database according to the industry standards and the design of database. It is worth to mention that in the code scheme and database construction, the underground pipeline data mainly include pipe nodes and pipelines. Pipe nodes are feature points (such as bend, tree, cross, etc.) or accessory points (such as well, valves, etc.) [4]. Pipeline is a line according to certain connection composition of pipe nodes. Models of valve, bend, tree and cross are shown in Figure 2.

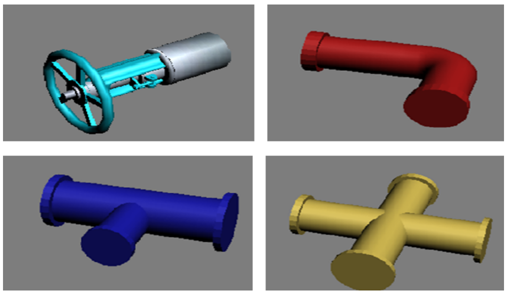


Figure 2. Pipeline valve, bend, tree, and cross models

The attribute information of pipeline include the number of start point and end point, type of pipeline, the depth of pipeline layout and tube's diameter, etc. Spatial entities such as pipe nodes, pipelines and facilities, which is embodied in underground space, are always represented improperly. Understanding the features of underground pipeline profoundly has become an important issue to lay out pipeline rationally. Generally, urban underground pipeline is usually invisible, complex, and potential, etc. the features of underground pipeline can be characterized as follows [5]:

a) The prominent feature of pipeline distribution is invisible. Since most of the pipelines are buried underground, pipeline management demands accuracy and completeness.

b) The distribution of pipeline is complex. There are several different kinds of urban pipelines, and the relationship between them is tight. The pipeline system is so enormous that if some part doesn't work, it will influence the other pipelines around it.

c) With the continuous development of urban, urban underground pipeline is constantly changing, expanding and updating, and the extent and density of urban pipeline layout keep growing. The data of pipelines are also required to update dynamically.

3D visualization is a time consuming and since it requires a large number of spatial information. We must process the data of remote sensing image, digital elevation model and 3D building models into 3DGIS system in the light of relevant requirements. First of all, remote sensing image and DEM are fused in 3D software, and then import the results into 3DGIS system. Moreover, a batch of 3D architectural model is exported in the form of model file. Meanwhile, model conversation is advanced accordingly with modelling software. In addition, 3D architectural model is integrated into 3DGIS system and modify some parameters correspondingly according to remote sensing image.

3.3 Realization of Layout Optimization of Urban Underground Pipeline Based on 3D Digital City

A lot of work has been done on 3D visualization, however, it is still a challenge to establish underground 3D pipeline model dynamically. The establishment of underground 3D pipeline model is based on the pipe node data, well and 2D pipeline, also corresponding to ground-level structures to locate the position of pipe sections quickly in a more intuitive way. 3D visualization of underground pipeline is more suitable for the representation of spatial relationship between all kinds of pipeline clearly, so we can review the connectivity and intersection structured of pipeline quickly and easily. Compared with the traditional 2D visualization, 3D visualization presents complex relationship to technical staff of urban planning departments and construction units, especially improves human-computer interaction further. Based on 3D digital city platform, we propose the method for pipeline layout optimization which enables 3D Digital City as a tool to organize and manage underground pipeline information.

Spatial analysis which describes and represents urban underground pipeline information accurately and completely has been developed for pipeline layout optimization. Currently, GIS spatial analysis is well-known and widely used in two dimension space for practical applications in 2D GIS. With its specific function of extracting, displaying and transferring invisible geographical spatial information, GIS spatial analysis is gradually used in three dimension environment. 3D spatial analysis method is defined as a key technology in underground pipeline layout optimization. Underground 3D pipeline layout optimization method is mainly used to execute basic analysis in the underground pipeline, and collect statistical information and excavate spatial data of pipeline [6]. Exploring the method of underground pipeline layout optimization by using GIS spatial analysis, the management of underground pipeline in digital system is helpful for resource management in urban planning department.

3.3.1 3D Network Analysis

Network analysis is a process of geographic analysis and model optimization deals with geographic network and urban infrastructure network. It discusses state of the network, simulates and analyzes resource flows of network and distribution, then achieves optimization of network structure and resource, etc [7].

Underground pipeline layout is based on different roads and districts. We firstly set starting and ending points according to existing road network and pipelines the topologies though utilizing 3D network analysis, we can obtain several network schemes. Other factors, such as the minimum distance between pipelines and facilitates in national norms, are considered to optimize the allocation of underground pipeline resource.

3.3.2 3D Buffer Analysis

Buffer analysis refers to create region with a certain width around point, line and surface entities automatically [8]. By means of 3D buffer analysis, band area with a certain distance around entities is created, and influence region between entities and the nearby objects can be identified clearly. The analysis is mainly used to create buffer area of pipe nodes or pipelines. Consequently, buildings and facilities within the buffer area are identified, and the results are outputted statistically.

Collision analysis is mainly utilizing buffer analysis of GIS spatial analysis to deal with the pipeline in 3DGIS. Collisions often exist between pipelines, pipelines and buildings, pipelines and facilities. According to the related national norms and standards, such as Code of Urban Engineering Pipeline Comprehensive Planning, collision analysis is used to search for buildings and facilities around pipeline.

3.3.3 Influencing Factors Analysis

On the basis of horizontal and vertical spacing, the impact factors among pipelines, pipelines and buildings, pipelines and facilities are analyzed. Besides, we can calculate horizontal and vertical spacing statistically between pipeline layout and the surrounding pipelines, buildings and facilities. Ultimately, compared with related national norms and standards, we can determine whether the horizontal and vertical spacing are in compliance with them. If not, the scheme is unreasonable. By adjust the positions of pipeline layout dynamically; we can avoid accident between pipelines, construction materials and facilities.

3.4 Optimization Scheme

According to the existing road network, we select start point and end point of pipeline which needs to layout. Then we utilize 3D network analysis to deal with road network, and generate several pipeline layout schemes automatically. 3D buffer analysis (Figure 3.) is used to process collision analysis of pipelines in 3D environment. On the top of that, according to analyzing influence factors of pipeline layout optimization, flow scheme for underground pipeline layout optimization is proposed. And we show optimization result in 3D digital city.

For the purpose of meeting requirements of pipeline layout optimization, making use of the latest technology to manage all types of underground professional pipeline effectively, we can sum up critical factors in the process of collision analysis and put forward layout optimization scheme in 3D space environment. Therefore, by means of 3D spatial analysis method and influence factor analysis, we can apply the results into urban underground resources management, pipeline planning, construction planning and 3D digital city applications.

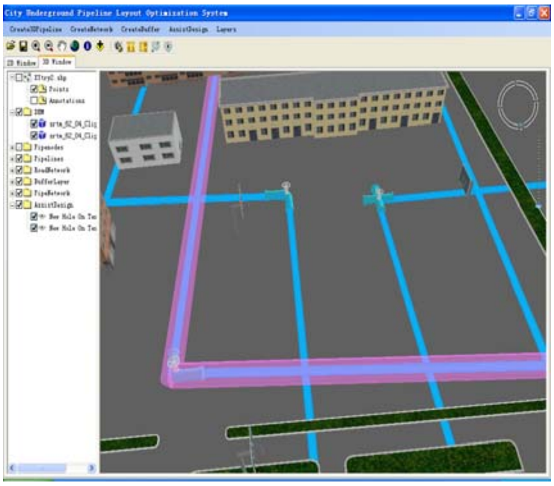
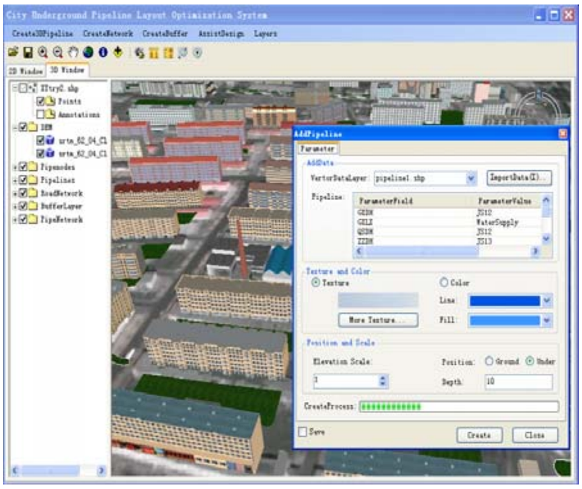


Figure 3. 3D buffer analysis result

3.5 Aided Design

Aided design does not only include adding, deleting pipe node and pipeline, but also include manage pipe node and pipeline dynamically. Based on pipeline layer, we take the number of pipe node and pipeline as identifier, then add and delete the data of pipe node and pipeline into the spatial database. When adding pipe node and pipeline (Figure 4), input various properties information of pipe node and pipeline. At the same time, with the continuous changing and complicated information of underground pipe nodes and pipelines, we can update and manage pipe nodes and pipelines database correspondingly and effectively. Aided design is capable of achieving the management of urban underground pipeline dynamically. Underground pipeline system can monitor and control various changes, so it’s able to update the existing underground pipeline files quickly and ensure truth and accurate reflection of underground pipeline layout.



VectorDataLayer: pipeline1.shp

Pipeline:

ParameterField	ParameterValue
GXDM	JS12
GXLX	WaterSupply
QSDM	JS12
ZZDM	JS13

● Texture ○ Color

Line: 🟦

Fill: 🟦

Elevation Scale: 1

Position: ○ Ground ● Under

Depth: 10

CreateProcess:

□ Save

Figure 4. Add pipeline operation process screenshot

4. EXPERIMENT

According to the process of urban underground 3D pipeline layout optimization described above, we conduct an experimental through the application of spatial analysis, and horizontal and vertical space analysis between pipelines and facilities. We propose a layout optimization solution for underground pipeline and reveal results of optimization analysis in 3DGIS system. With underground 3D space model and pipeline layout optimization analysis, we can adjust pipeline dynamically and lay out pipeline in a more rational way. Ultimately, we can get layout optimization solution for underground pipeline layout (Figure 5).

The method of underground pipeline layout optimization based on the 3D digital city is implemented and the mining of optimization information in various pipelines on the basis of underground pipeline optimization method is realized. It facilitates the effective management of related departments which manages every corner of city's pipeline. The method enhances the layout of urban underground pipeline effectively and creates more convenience for community.

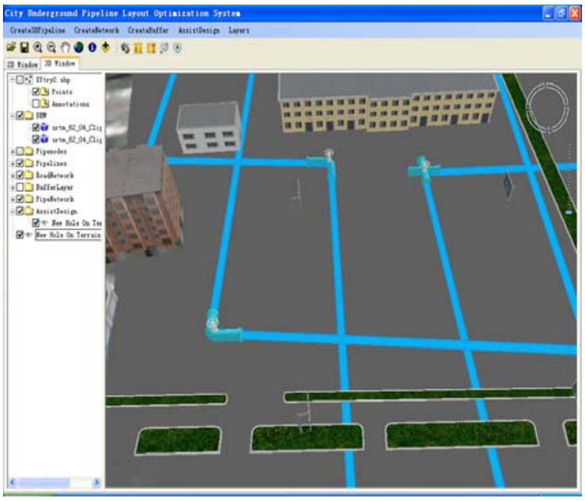


Figure 5. Experiment result

5. CONCLUSION

This paper proposes a method of underground 3D pipeline layout optimization based on 3D digital city. The purpose of optimization method is to obtain urban underground 3D pipeline optimization system, which is associated with characteristics of 3D spatial data and the process of layout optimization. We realize the fusion of various kinds of data and underground 3D pipeline model dynamically, also achieve the data mining of pipeline layout optimization. Consequently, this method is not only helpful for related department interacts with underground pipeline to facilitate their construction, but also helpful for providing technical support and assist decision support for urban management, urban normal operation, and 3D digital city construction, etc.

REFERENCES

[1] Juan Yang, Hao Lin and Yupeng Xiao, 2009. Spatial Data Model for Visualization System of GIS Based Urban Pipeline. International Forum on Information Technology and Application, pp.98-102

[2] Yingdong Li. 2007. The Design and Exploitation of Undergroung-integrated Pipeline Information System of Ganzhou Urban. JiangXi University of Science and Technology, pp.7-18

[3] Claudio Monteiro, et. al. 2005. GIS Spatial Analysis Applied to Electric Line Routing Optimization, IEEE Transactions on Power Delivery, 20(2), pp.934-942

[4] Yong Han, Ge Chen and Haitao Li, 2004. Construction and Implementation of Underground 3D pipeline layout optimization Models for Urban Underground Pipelines, Periodical of ocean university of china, 34(3), pp.506-512.

[5] Jun Gong, Xinzhou Wang, Wenqing Wang and Xiong Zhang,

2005. Research on Urban Underground Pipeline Information System, Geospatial information, 3(3), pp. 9-11.

[6] Chun Liu, Lianbi Yao and Weigang Lei, 2003. Application of Spatial Analysis in Information System of Urban Synthesis Underground Pipeline, Science of Survering and Mapping, 28(4), pp.55-57.

[7] Guo'an Tang and Xin Yang, 2006. ArcGIS Geographic Information System Spatial Analysis Expermental Course, Press of Science, pp. 212-219.

[8] Dong Ma, 2007. Research and Realization of Urban Pipeline Network Information Management System Based on ArcGIS engine, Jiao Tong University, pp.58-59

MULTI-RESOLUTION REPRESENTATION OF DIGITAL TERRAIN AND BUILDING MODELS

Fuan Tsai, Wan-Rong Lin and Liang-Chien Chen

Center for Space and Remote Sensing Research
National Central University
Zhong-li, Taoyuan 320 Taiwan
ftsai@csrsr.ncu.edu.tw

KEY WORDS: Multi-Resolution, Digital Terrain Model, Building Model, Level of Detail, Cyber City

ABSTRACT:

This research develops effective algorithms for multi-resolution representations of three-dimensional (3D) digital terrain and building models to achieve better performance in cyber city applications. The objective is to create multiple levels of detail (LOD) of terrain meshes and polyhedral building models so they can be used efficiently according to viewing parameters and application requirements, while preserving critical features of the datasets. For terrain meshes, a tile-based approach is adopted. A mesh refinement algorithm based on modified quad-tree process is developed to generate multi-resolution representations of each terrain patch. On the other hand, a divide-and-conquer strategy is employed for the generalization of 3D building models to formulate LOD representations of complicated buildings. The idea is to apply generalization in 2D orthographic views of original polyhedral building models and then reconstitute simplified 3D models accordingly. Experimental results with complicated terrain and building datasets demonstrate that the developed LOD algorithms can improve cyber city performance significantly.

1 INTRODUCTION

Terrain and building are fundamental and two of the most important components in cyber city and other three-dimensional (3D) Geographic Information Systems (GIS) implementations. However, the vast amount of data in a large-scale cyber city modeling often poses a great challenge to efficient processing, analysis and visualization, especially in real-time applications. Level of Detail (LOD) is a commonly adopted technique to generate multi-resolution representations of objects in computer graphics and visualization (Luebke et al., 2003). The OpenGIS® CityGML (City Geography Markup Language) encoding standard proposed by OGC (Open Geospatial Consortium) also defines five levels of detail (LOD0~LOD4) for 3D digital city implementations (Gröger et al., 2008). However, the CityGML LOD specification is designed primarily based on functionality and thus may not be adequate in terms of performance consideration, especially for real-time visualization and applications. To address this issue, this paper presents systematic approaches to generate multi-resolution representations of large-scale digital terrain and building models that can be used to improve the performance in data transmission, processing and rendering of a cyber city system.

2 LOD FOR TERRAIN MESHES

As there are usually millions of points and polygons in a large-scale digital terrain model, it is a practical necessity to reduce the data amount for efficient processing and rendering. Level of detail techniques have been proposed for multi-resolution representation of terrain meshes, such as Bin-tree hierarchies (Blow, 2000), Bin-tree regions (Cignoni et al., 2003), geometric clipmaps (Losasso and Hoppe, 2004) and Quad-tree based meshes refinement (Tsai et al., 2006). Among them, Quad-tree based approaches are more suitable for geo-spatial applications, because they can better preserve critical terrain features while reduce the data amount.

A previous study suggested applying Quad-tree simplification on terrain meshes separated into tiles and proposed an adaptive progressive mesh data structure to achieve near real-time rendering frame rate for complicated digital terrain models (Tsai et al., 2006). However, in their algorithms the thresholds for quad-tree subdivision were determined from the difference between the maximum and minimum elevations of a tile and may cause over-sampling or under-sampling in areas. A new thresholding scheme (Tsai and Chiu, 2008) based on view-dependent image-space error metric was proposed to provide better LOD generation of terrain meshes. By calculating the ground sampling distance (GSD) as illustrated in Fig. 1 and Eq. (1) and (2) under different viewing parameters, different thresholds of quad-tree processing can be determined more reasonably according to “view-importance”, thus generating more appropriate LOD datasets.

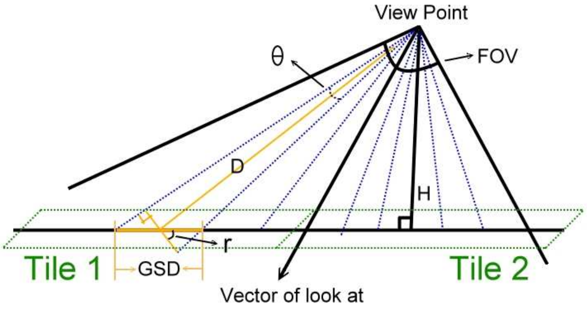


Figure 1: Ground sampling distance (GSD)

$$\theta = \frac{FOV}{\text{pixels per scanline}} \tag{1}$$

$$GSD = \frac{D\tan(\theta)}{\cos(r)} \tag{2}$$

In addition to the thresholding scheme, a nested LOD pre-process was also proposed (Tsai and Chiu, 2008). The idea was to generate an Outer-LOD-Set based on the coarsest level of the primary LOD terrain meshes (Core-LOD-Set). This can further increase the system performance in visualization, especially during the initializing stage of a large-scale application.

The reason of applying tile-based approach is that it is easy to implement view-dependent visualization for reducing the amount of data to process and render. However, a disadvantage in tile-based terrain visualization is the cracks caused by discontinuities (T-junctions) among tile boundaries. A commonly adopted technique to compensate this artifact is applying a pseudo generic texture layer beneath the terrain surface (Pouderoux and Marvie, 2005). Nevertheless, other than not providing true textures, this workaround will not work if the view angle is too low. A mesh refinement procedure was developed to address this issue. Taking Fig. 2 as an example, the procedure to remove T-junctions is listed in Algorithm 1.

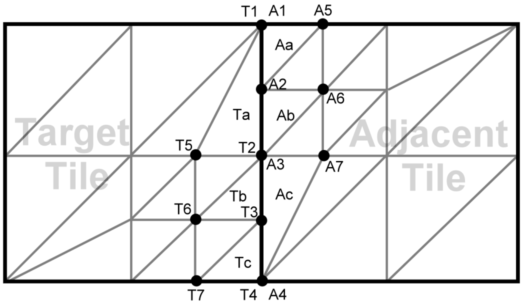


Figure 2: T-junction removal

Algorithm 1 T-junction removal procedure

1. Starting from $T1$ and $A1$, because $T1 = A1$, they remain unchanged.
2. Moving to the next pair, because $T2 \neq A2$, there exists a T-junction.
3. Remove Ta and add two new triangles, $\Delta(T1, T5, A2)$ and $\Delta(A2, T5, T2)$.
4. Continue the process from $T2$ and $A3$ and a new T-junction is found at $T3$.
5. Replace Ac with $\Delta(A3, T3, A7)$ and $\Delta(A7, T3, A4)$.

Combining these algorithms, the proposed tile-based quad-tree processing of terrain meshes can generate multiple LODs of large digital terrain datasets based on different viewing parameters. The algorithms can reduce data to process and render while preserving important terrain features. In addition, the adaptive data structure enables progressive transmission of data streams. All together, they provide high-performance terrain visualization capability for real-time applications.

3 MULTI-RESOLUTION 3D BUILDING MODELS

Unlike terrain data, most 3D building models used in cyber city implementations are not in mesh format. Therefore, mesh-based LOD schemes may not be applied to the process of building models. A few approaches have been proposed to deal with the LOD generation of 3D building models. For example, an algorithm based on mathematical morphology and curvature space of scale-spaces theory was presented to generalize 3D building models (Mayer, 1998). The algorithm was further refined by moving parallel facets toward each other to eliminate protrusion and close the gaps (Forberg, 2007). Another type of approach is segmenting building models into several structural elements and performing generalization on individual building segments, such as applying half-space modeling by cell decomposition and primitive instancing (Kada, 2007) or simplifying 2D projections of 3D geometries but only with linear and neighboring building groups (Anders, 2005).

Most of existing LOD techniques for 3D building models are either computationally expensive or are limited to certain types of buildings. A semi-automatic generalization approach is proposed to provide better multi-resolution representation of complicated building models. The idea is to apply generalization in 2D orthographic views of individual buildings and then reconstruct simplified 3D models accordingly as illustrated in Fig. 3. The process can be repeated with different generalization parameters so multiple levels of details can be created. The principle is similar to Anders (2005), but the procedure and algorithms employed in this study is very different. Anders (2005) utilized a program (CHANGE), which was originally designed to aggregate two-dimensional building ground plans for the generation of topographic maps, to simplify three projections of building groups and then grue them to form 3D block models. This approach is efficient for linear building groups but might not be adequate for complex buildings. On the other hand, the algorithms developed in this study is a divide-and-conquer strategy that is capable of dealing with building models with complicated structures and shapes. In addition, the developed algorithms can also generalize buildings with non-orthogonal façades, which are difficult to handle using existing methods.

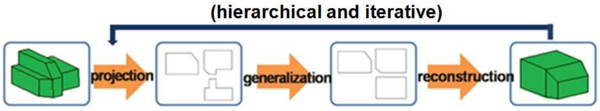


Figure 3: LOD for 3D building models

Before applying the generalization process, a geometric structure analysis is conducted to determine the complexity of each building and its number of levels to generalize. The shape complexity is defined as Eq. (3), where N_r is the number of vertices of the roof polygon and N_C is the number of vertices of the 3D convex hull of all roof structures. A few examples of shape complexity are demonstrated in Fig. 4. Based on calculated shape complexity, necessary levels of detail for individual building models can be determined. For instance, the most complicated model (SC=0.72) in Fig. 4 may require four levels of generalization while the second case (SC=0.3) may need only two levels.

$$SC = \frac{N_r - N_C}{N_r} \tag{3}$$

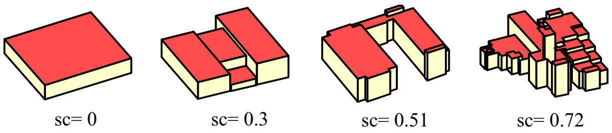


Figure 4: Shape complexity examples

After determining the levels to generalize, at least three orthographic projections of each building are generated and converted into raster formats. Then, a series of morphological operations (including dilation, point in polygon elimination, and erosion) is applied to create raster versions of the orthographic projections for constructing the outlines of building models as demonstrated in Fig. 5. A topology connector, which is modified from the Target Defined Ground Operator (TDGO) (Chen and Lee, 1992), is developed in this study to establish topological relationships among projection points. The modified TDGO performs topological encoding of each pixel according to the surrounding pixels in a 3x3 moving window. Some examples of TDGO encoding are listed in Fig. 6. Applying these operators to examine the generated raster projections, their edges and corner points can be identified correctly as displayed in Fig. 7.

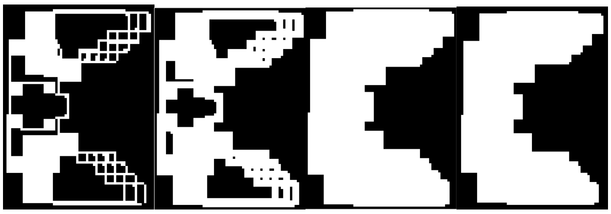


Figure 5: Outline generation of raster projections (left to right: (a) original projection; (b) dilation; (c) point-in-polygon filtering; (d) erosion)

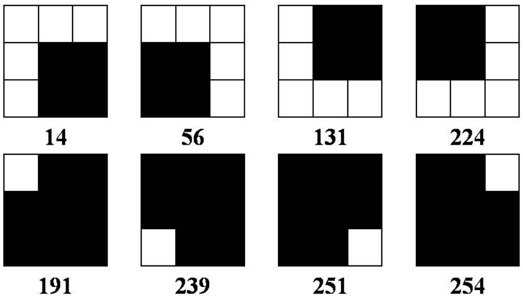


Figure 6: TDGO encoding examples

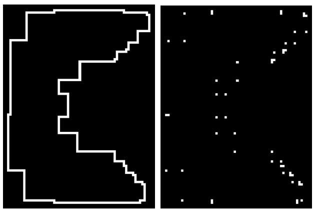
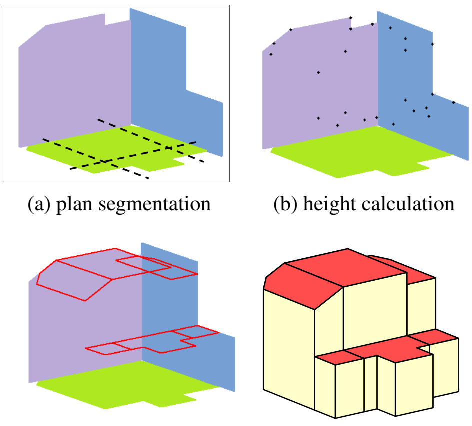


Figure 7: Identified edges and corners

Each orthographic projection is then generalized with convex-concave structure generalization and edge regularization. Convex-concave structure generalization is to detect convex and concave structures and eliminate small structures. Edge regularization is to eliminate short-length edges. In this step, the lengths of edges and the angle between the neighboring edges of a vertex in orthographic views are calculated based on the topological relationship of orthographic views. Concave and convex structures and edges with short lengths are detected by calculated angles and lengths. The areas of convex and concave structures are calculated and compared with a pre-defined threshold to remove minor structures. Edges with lengths smaller than pre-defined thresholds are also eliminated. New junction points for each eliminated structure or edge are calculated accordingly to reshape the orthographic projections.

After generalization of orthographic projections, simplified 3D models can be reconstructed from them as illustrated in Fig. 8. First, according to vertical orthographic views (front and side views), the horizontal orthographic view (top view) is segmented into several pieces. A loop tracing technique is applied to connect segmented pieces as polygons. After collecting all the segmented polygons, the heights of vertices points on segmented plans are calculated according to vertical orthographic views. These elevated points are reconnected in 3D space to shape the simplified 3D model.



(a) plan segmentation (b) height calculation

(c) 3D shaping (d) result

Figure 8: 3D model reconstruction from orthographic projections

For non-planar roof structures, they are detected and segmented in the geometric structure analysis step. As displayed in Fig. 9a, gabled roof structures are simplified according to the generalized plan view. When the ground plan is generalized, the topmost polygons are updated by transformation relationship through back-projection from ground plan to topmost polygons. For barrel roofs, the curve-shaped roof structure is maintained by curve fitting technique, as shown in Fig. 9b and placed on top of the major façades. There are different methods to fit a curve from points (Gallier, 2000). If there are sparse roof vertices, which is usually the case in polyhedral building models, a simple way is to use the two boundary vertices and the highest roof point for a conic arc fitting as illustrated in Fig. 9b. However, the boundary vertices need to be adjusted according to the generalization result of façades. This way, important characteristics such as building height and dimensions can be preserved. For complicated curves, collecting all vertices and performing a spline fitting or by least squares fitting (Coope, 1993) is a more appropriate approach, but requires more computation.

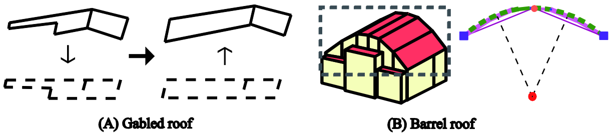
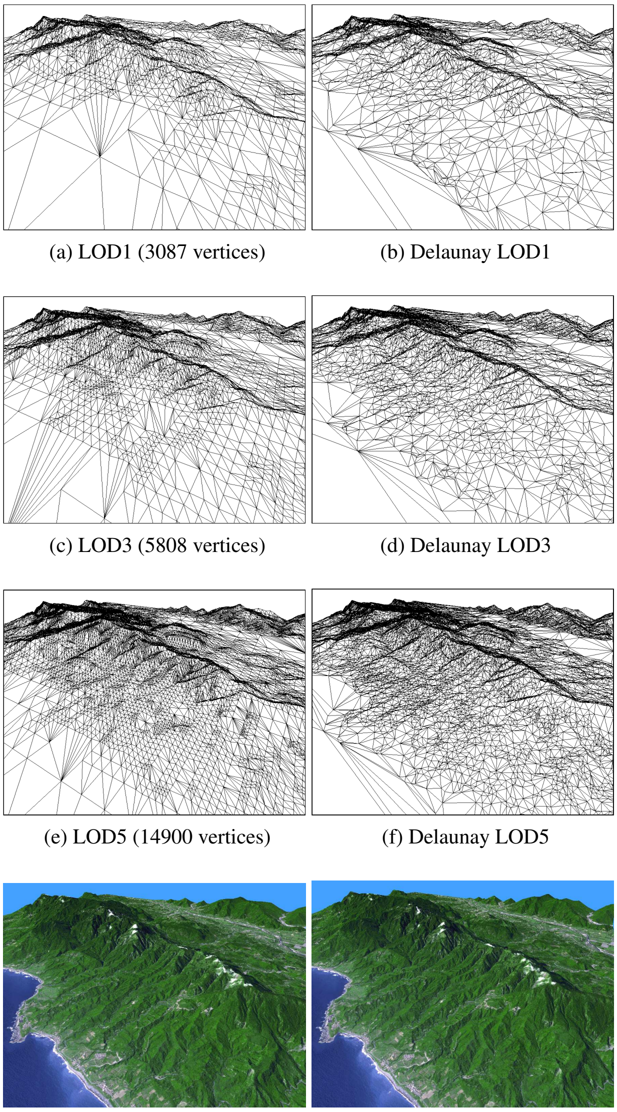


Figure 9: Generalization of roof structures

The described generalization method is highly automated and can deal with a variety of building models. However, in some special cases, additional (interactive) processes may be necessary. For example, for buildings with courtyards, the orthographic views of their inner structures will be detected interactively and then generalized with the generalization process of orthographic views. Another example is buildings with non-planar façades. To keep their curved characteristics, the non-planar façades are segmented from the original building models and generalized similar to the barrel roofs.

4 EXPERIMENTAL RESULTS

The developed multi-resolution representation algorithms for digital terrain and building models were applied to real datasets to validate their performances. Figure 10 shows the LOD1, 3 and 5 of a large terrain mesh and the comparison with Delaunay triangulation. Both the proposed quad-tree based LOD processing and Delaunay triangulation can effective reduce the data in different LOD levels.



(a) LOD1 (3087 vertices)

(b) Delaunay LOD1

(c) LOD3 (5808 vertices)

(d) Delaunay LOD3

(e) LOD5 (14900 vertices)

(f) Delaunay LOD5

(g) Textured LOD1

(h) Textured Delaunay LOD1

Figure 10: Terrain LOD with proposed method (left) and Delaunay triangulation

Although it may appear that Delaunay has better triangulated meshes, the proposed method can also preserve important terrain features in different LODs. (If the meshes are textured, the difference between the two is almost indistinguishable, even in the low resolution LOD1 as displayed in Fig. 10g and h.) More importantly, the data structure of the proposed Quad-tree processing is more organized than Delaunay and makes the rendering more efficient. Taking T-junctions removal as an example, Delaunay will require significantly more efforts to remove T-junctions because the triangles on tile edges are irregular. In addition, for Delaunay triangulations, it will be inefficient to use the "difference vectors" scheme because vertices in different levels of Delaunay-based LOD do not have an "add-on" property. Therefore, it will be difficult, if not impossible, to achieve progressive transmission and adaptive rendering of Delaunay-based LOD tiles and thus inadequate for real-time visualization applications.

Figure 11 shows the effect of T-junction removal. From the figure, the crack at the tile boundary has been repaired with the proposed mesh refinement algorithm and resulting in a seamless terrain scene. The proposed mesh refinement algorithm is very efficient. In previous tests (Tsai and Chiu, 2008), the CPU time used to remove T-junctions at run-time was almost negligible and caused insignificant impact to the rendering performance. Figure 12 shows the comparison of accumulated CPU time for a fly-through of a large terrain (originally 5022 by 9555 grids and partitioned into 10 by 19 tiles) with and without performing T-junction removal while rendering the terrain meshes. It is clear that applying T-junction removal has increased very little rendering time, but the improvement in visual quality is significant as demonstrated in Fig. 11.

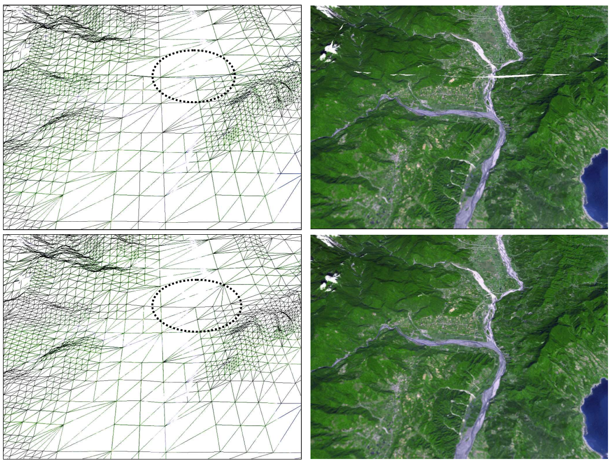


Figure 11: T-junction removal

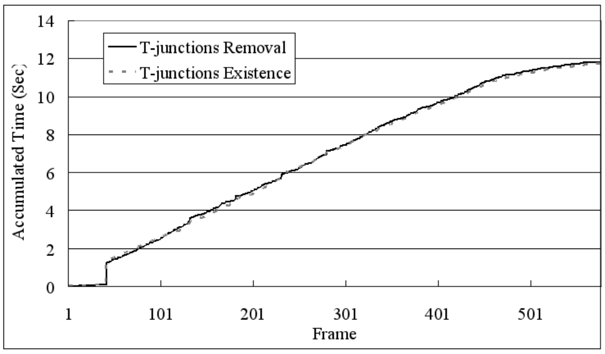
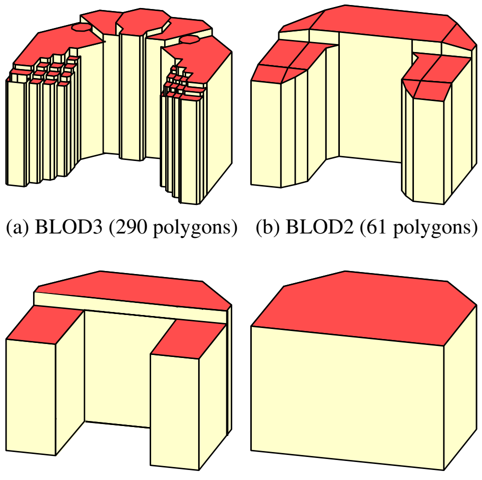


Figure 12: Cumulative rendering time comparison of T-junction removal

The proposed terrain LOD generating algorithms were tested with two large DEM datasets in different scenarios (Tsai and Chiu, 2008). From the tests, the frame rate of rendering can achieve at least 24.5 FPS (frames per second) within two-pixels error on screen display even in a very complicated mountainous terrain. This should be adequate for real-time visualization and applications.

Figure 13 displays an example of generated multiple building level of detail (BLOD) from a fairly complicated 3D building model. In this example four different levels of detail were generated with the generalization algorithms described above. From the figure, it is clear the generalization has effectively reduced the number of polygons from 290 to 61, 19 and 7 subsequently. Although the data (polygon) amount has been reduced significantly in lower-resolution building models, the characteristics of the building has been preserved.



(a) BLOD3 (290 polygons) (b) BLOD2 (61 polygons)

(c) BLOD1 (19 polygons) (d) BLOD0 (7 polygons)

Figure 13: Multiple LOD of building model

The example in Fig. 13 demonstrates that the proposed generalization algorithms and the iterative building level of detail generation procedure are effective for buildings with regular (planar) façades regardless their complexity. Figure 14 and 15 demonstrate generalization results of a few buildings with irregular shapes and structures including non-planar roof structures. These examples indicate that the developed BLOD algorithms are also effective for special building models.

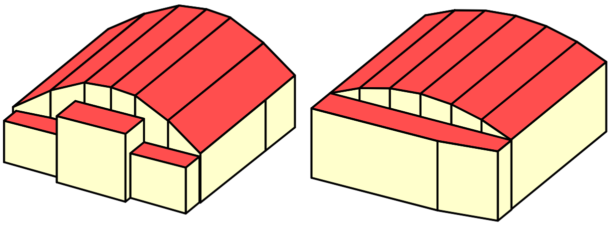


Figure 14: BLOD for special building model with barrel roof

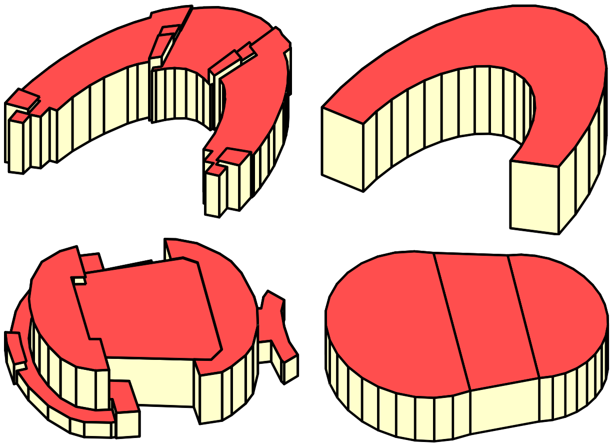
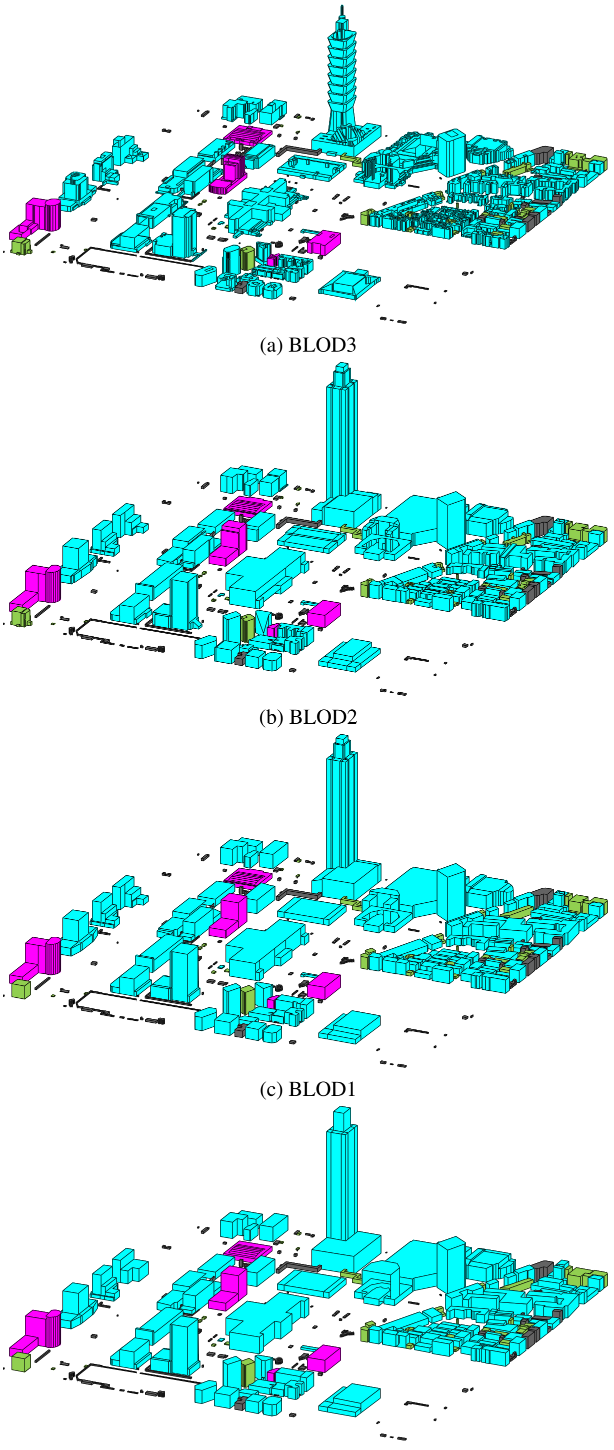


Figure 15: BLOD for building models with special shapes and structures

Applying the developed generalization algorithms to buildings in a city model, multi-resolution representation of the building models can be generated effectively. Figure 16 shows four levels of detail of a business district in Taipei. There are a couple of hundreds buildings with different degrees of complexity and styles in this area. The proposed algorithms generalize them according to determined shape complexity effectively. The generalization reduce the data amount significantly (from 27% to 38% of the original number of polygons as listed in Table 1) but still preserve important geometric characteristics (features) of buildings.



(a) BLOD3

(b) BLOD2

(c) BLOD1

(d) BLOD0

Figure 16: BLOD of a business district in Taipei Taiwan

With the generated multiple levels of detail, a cyber city system can load require building models progressively from BLOD0 to BLOD3 according to different viewing parameters and system or analysis requirements to increase the performance in visualization and analysis.

Table 1: Data reduction rate of BLOD

	BLOD3	BLOD2	BLOD1	BLOD0
Points	61795	22825	17535	16110
Reduction rate		37%	28%	26%
Polygons	14045	5350	4175	3841
Reduction rate		38%	30%	27%

One thing to note is that it seems data reduction is more aggressive when generalizing the models from BLOD3 to BLOD2 than the rest generalization. This might caused by inappropriate thresholding for generalization. However, BLOD3 is the original (most detailed) building model and consists of many minor structures. (This can be observed from the examples presented in previous figures.) Therefore, it is expected to have a significant reduction in terms of point and polygon numbers, but the geometric shapes or characteristics of the models do not deform too dramatically. In addition, the buildings are treated independently, thus building aggregation may seem necessary. Whether to aggregate building groups should depend on the objective of applications. If it is necessary, aggregation should be performed with additional merging process.

5 CONCLUSIONS

This paper presents systematic approaches to create multiple levels of detail for digital terrain and building models. For terrain meshes, a tile-based quad-tree processing with thresholds determined from ground sampling distance under different viewing parameters is suggested to generate terrain LODs. A mesh refinement procedure to correct discontinuities (T-junctions) among tile edges is also presented, which can eliminate discontinuities between adjacent tile meshes effectively and have little impact to the overall rendering performance. For 3D building models, an iterative procedure based on algorithms for generalization on 2D orthographic projections and reconstructing simplified 3D models is proposed. For special structures of building models, additional processes are developed to simplify them but maintain their characteristics. The developed algorithms are effective for individual building models and large building groups with various degree of complexity.

Examples demonstrated in this paper indicate that the developed algorithms can be used for multi-resolution representation of complicated terrain and building models effectively and efficiently. All together, the proposed methods can increase the performance of large cyber city implementation for real-time visualization and applications. More importantly, while reducing the data amount to transmit and process, the proposed multi-resolution representation methods can preserve important geometric and visual characteristics of complicated terrain and building models

ACKNOWLEDGEMENTS

This study was partially supported by the National Science Council of Taiwan under project No. NSC-98-2221-097-MY2.

References

Anders, K., 2005. Level of detail generation of 3d building groups by aggregation and typification. In: Proc. of 22nd International Cartographic Conference, La Coruña, Span.

Blow, J., 2000. Terrain rendering at high levels of detail. In: Proceedings of Game Developers Conference, pp. 903–912.

Chen, L.-C. and Lee, L.-H., 1992. Progressive generation of control frameworks for image registration. Photogrammetric Engineering and Remote Sensing 58(9), pp. 1321–1328.

Cignoni, P., Ganovelli, F., Gobbetti, E., Marton, F., Ponchio, F. and Scopigno, R., 2003. BDAM-batched dynamic adaptive meshes for high performance terrain visualization. In: Proceedings of EUROGRAPHICS, Vol. 22number 3, pp. 505–514.

Coope, I. D., 1993. Circle fitting by linear and nonlinear least squares. Journal of Optimization Theory and Applications 76(2), pp. 381–388.

Forberg, A., 2007. Generalization of 3d building data based on a scale-space approach. ISPRS Journal of Photogrammetry & Remote Sensing 62, pp. 104–111.

Gallier, J., 2000. Curves and Surfaces in Geometric Modeling. Morgan Kaufmann, San Fancisco, CA USA.

Gröger, G., Kolbe, T. H., Czerwinski, A. and Nagel, C., 2008. OpenGIS city geography markup language (CityGML) encoding standard. Technical Report OGC-08-007r1, Open Geospatial Consortium Inc. v. 1.0.0.

Kada, M., 2007. Generalization of 3d building models by cell decomposition and primitive instancing. In: Proc. Joint ISPRS Workshop on Visualization and Exploration of Geospatial Data, Stuttgart, Germany.

Losasso, F. and Hoppe, H., 2004. Geometry clipmaps: Terrain rendering using nested regular grids. ACM Transactions on Graphics 23(3), pp. 769–776.

Luebke, D., Reddy, M., Cohen, J. D., Varshney, A., Watson, B. and Huebner, R., 2003. Level of Detail for 3D Graphics. Morgan Kaufmann Publishers.

Mayer, H., 1998. Three dimensional generalization of buildings based on scale-spaces. Technical report, Dept. of Photogrammetry and Remote Sensing, Technische Universität München, Germany.

Pouderoux, J. and Marvie, J.-E., 2005. Adaptive streaming and rendering of large terrains using strip masks. In: Proceedings of ACM GRAPHITE 2005, pp. 299–306.

Tsai, F. and Chiu, H.-C., 2008. Adaptive level of detail for terrain rendering in cyber city applications. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVII-B4, pp. 579–584.

Tsai, F., Liu, H.-S., Liu, J.-K. and Hsiao, K.-H., 2006. Progressive streaming and rendering of 3d terrain for cyber city visualization. In: Proc. ACRS2006, Ulaanbatar Mongolia.

LOCATION BASED CONTEXT AWARENESS THROUGH TAG-CLOUD VISUALIZATIONS

V. Paelke, T. Dahinden, D. Eggert, J. Mondzech

IKG, Institute for Cartography and Geoinformatics
Leibniz Universität Hannover
Appelstr. 9a, 30167 Hannover, Germany
{paelke, dahinden, eggert, mondzech}@ikg.uni-hannover.de

KEY WORDS: Information Visualization, Context Awareness, Tag-clouds, Location Based Services

ABSTRACT:

People often only see what that they already know or explicitly look for. Thus, while mobile users might be interested in the historic background of their surrounding, its current significance or its relation to specific events, they are likely to miss places of interest if they are not explicitly pointed out. Location based services (LBS) like mobile tourist guides offer a potential technological solution, but the production cost of multimedia content is prohibitive in many cases, limiting the coverage of such services to major tourist areas.

To provide mobile users with information on the spatial context of a location or a route we present an approach that gathers context information from freely available sources like Wikipedia and creates visualizations of this data that provide users with the necessary cues to increase awareness of their spatial context. In our approach we first gather geo-referenced information that is located close to a point or route. In the second step this data is filtered to extract key-words that characterize the environment. These are then rendered as a tag-cloud in the third step. By skimming the tag-clouds a user gets a good impression of the characteristic features of an environment and in essence performs a further filtering step. The user can interactively adjust the level of detail of the visualization or follow up on individual key-words to adjust the presentation to his interests.

By combining web 2.0 technologies and public data sources with filtering and visualization techniques we exploit the browsing capability of humans to provide a service that increases location awareness at arbitrary locations.

The approach makes it easy to author an additional text and it can incorporate the ever increasing amount of available geo-referenced information.

1. MOTIVATION

Advances in mobile computing and wireless communication technology enable the creation of location based services (LBS) on a variety of mobile devices ranging from mobile phones to PDAs and other portable computing devices. Modern Smartphones like Apple’s iPhone, the Android based T-Mobile G1 and similar devices combine GPS based positioning with a digital compass, high resolution displays and high-bandwidth data connections. As the technical prerequisites for location based services become widely available the development of practical services and the creation of the content required for them becomes an increasingly important question.

One possible approach is the development of specific authoring tools that support the creation of location based multimedia content and its integration into user interfaces that consider the constraints of mobile devices. In this paper we consider a complementary approach: Information is gathered from freely available sources like Wikipedia and a combination of automatic filtering and processing techniques with information visualization techniques based on tag-clouds are used to provide the user with a display of the available information.

Location based services like mobile tourist guides provide two main services: They point out potential points of interest in the surroundings of the user and can provide users with detailed background information on these on demand. This functionality is useful because people often only recognize what that they explicitly look for. Because the production effort for content of current location based services is significant such information is often only available for touristic areas, where the high cost of content production can be apportioned on a large number of users.

A possible alternative to content that is explicitly authored for use in a LBS system is to exploit information that is already available, e.g. on websites. To make this information useful for the user two steps are required: First, information has to be spatially selected, so that only information pertaining to the surrounding of the user is used. Second, the information has to be refined into a form that is suitable for easy interpretation by the user.

In the approach that we present in this paper we exploit information from websites like Wikipedia and process it into a visualization that is based on the concept of tag-clouds to provide a presentation that can be quickly browsed by the user. After an introduction of the concept in the following section we discuss related work in section 3. In section 4 we discuss the current implementation. Examples are presented and discussed in section 5. Finally we discuss the experience with the approach and future work.

2. CONCEPT

As illustrated in Figure 1 our system starts either by acquiring the spatial position of the user or by planning a route through the environment. Current Smartphones feature a GPS unit that can be used for the localization, otherwise a more imprecise localization can be derived by identifying wireless network cells in the environment.

Using this position it becomes possible to supply the user with information on his spatial surroundings. The second case - involving route planning - is a bit more complex. Here not only the current surroundings of the user are queried but a spatial buffer around the planned route is used. This can be exploited to aid pedestrian navigation, by using the context data to improve route descriptions.

As the examples in section 5 show landmarks of special interest can usually be identified from the tag-cloud, thus supplying a textual representation of landmarks for pedestrian navigation. The user can interactively influence the processing at this stage both by supplying start and destination information for the route planning.

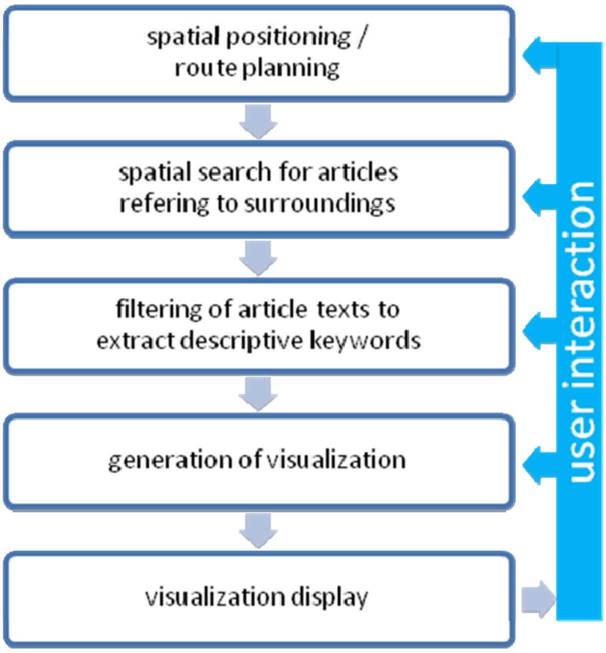

spatial positioning / route planning
spatial search for articles refering to surroundings
filtering of article texts to extract descriptive keywords
generation of visualization
visualization display
user interaction

Figure 1: System processing workflow

The second step of the processing pipeline uses a spatial query to search for spatially referenced articles in the surroundings of the user. This is either a simple proximity query around the user's current location or a buffer operation around the planned route.

To influence the process the user can interactively specify a query radius in which information should be retrieved. It thus becomes possible to adjust the amount of data to the preferences of the user.

At this stage it is also possible to introduce additional information into the process. In some cases it could be desirable to provide explicitly authored or preselected information for specific areas. These are simply added to a spatial database and retrieved in the same fashion as the articles. If desired they can be flagged as additional" and thus prioritized in the following processing steps.

In the third step the text from the retrieved articles is filtered to remove irrelevant words. Usually verbs and articles are less descriptive than nouns and therefore they are removed. Very common words (e.g. the name of the city in which the user is currently located) also add little to the location awareness and should therefore also be removed. The filtering can be adjusted by the user and specific filter lists can also be supplied to suppress unwanted words from the tag-cloud.

In the fourth step a tag-cloud visualization is created from the filtered text. The user can influence this step by adjusting font type and size, layout criteria and the number of tags that should be displayed. A smaller number of tags is easier to browse and more descriptive and therefore the default setting. If a user is especially interested in his surroundings and wants to investigate more deeply or if the information provided in the display is not sufficient the user can easily choose to display more tags in the cloud.

The final step is the display of the tag-cloud. To improve readability the user can zoom in and out and scroll the display if required. This is especially useful for route visualizations. In addition the user can also “drill down” into the tag-cloud to retrieve more detailed information on tags of potential interest. If the user selects a specific tag the system can provide him with a list of the original occurrences in the source articles. Alternatively a web-search for the tag-cloud be initiated. While the tag-cloud provides an overview visualization of information of potential interest, exploiting the user's browsing capability to further filter the information, the possibility to link to the original articles can implement the second function of a LBS, namely to provide detailed information on points of interest in the environment.

A description of the hardware platform and the software used to implement the different processing stages in our prototypical implementation is given in section 4. The implementation concerns an evolutionary prototype where a sophisticated route planning and the interaction options of the tag-cloud are not yet implemented.

3. RELATED WORK

Our work was motivated by previous work on the development and use of conventional LBS using authored content. LBS like tourist guides (Zipf, 2002; Baus et al., 2005; Schilling et al., 2005) have become popular to use in areas where adequate content is available. Unfortunately, the production effort of such content is still very high and commercially successful implementations are usually restricted to top tourist destinations.

Central to our approach is the exploitation of the user's browsing capability using the tag-cloud visualizations. Tag-clouds have become popular through their use on Web 2.0 sites, namely Flickr (Flickr, 2009). The first widely published use of the format is usually attributed to Douglas Coupland who used a similar form of visualization in his book Microserfs (Coupland, 1995). As Viegas and Wattenberg (2008) point out there have been predecessors to this. Of special interest for our approach is the study of Milgram (Milgram, 1976) in which he aggregated the mental maps that people have of a space (in the case of this study Paris) into an aggregate visualization that has a lot of similarity to the tag-clouds produced by our system.

From a cartographic perspective it is also interesting to note the relations between tag-clouds and more structured lists of keywords. Lamantia (Lamantia, 2009) has pointed out the similar relation between structured lists and tag-clouds on one side and maps and cartograms on the other. In recent years the study of tag-cloud visualizations for different purposes has become a research topic in information visualization (Halvey and Keane, 2007; Rivadeneira et al., 2007). A large number of layout algorithms and libraries that implement them have been proposed and studied (for an overview see e.g. Kaser and Lemire, 2007).

Our approach relies on the availability of spatially referenced information in publicly accessible sources like Wikipedia (Wikipedia, 2009). In recent years there have been several attempts to exploit this information for purposes in a Geographic Information Science context, e.g. the GeoSR System (Hecht and Raubal, 2008) used spatially references Wikipedia data to explore semantic relations in spatial data.

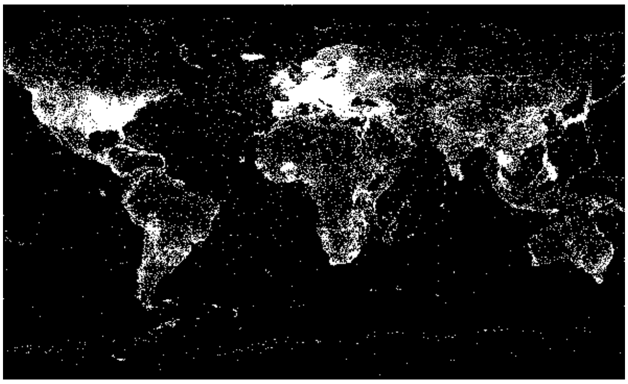


Figure 2: Distribution of Wikipoints of the German Wikipedia

4. IMPLEMENTATION

The implementation of our system demands a client-server approach – a client visualizing the tag-cloud and providing user interaction, as well as a server harvesting the desired data and creating the corresponding tag-cloud information. As mentioned before, this information can be obtained from various sources like Wikipedia, Flickr or any other knowledge repository with geo-referenced articles. Our prototypical implementation utilizes Wikipedia only.

4.1 Client

The tag-cloud client component is implemented on an Android Smartphone using the Android SDK v1.5. The Java-based Android SDK provides the Java 5 compliance level, which facilitates the use of nearly all Java libraries written for standard PC environments. In order to utilize the Java ObjectStream class to send and receive the tag-cloud data, the server component is also written in Java. Furthermore a shared tag-cloud object specification is needed, which is covered by the OpenCloud Java library (http://opencloud.sourceforge.net) used on the client as well as on the server side.

For the purpose of visualization and interaction the MapView and MapOverlay classes, provided by the Android Google API Add-On, are extended. As the name reveals the MapView class provides a map view based on GoogleMaps. The tag-cloud is visualized by extending the MapOverlay class. For resource saving purposes a received tag-cloud is precompiled into a picture object. Various output filters (e.g. alphabetic or score-based ordering) can be applied to the tag-cloud to manipulate the shown tags in the displayed picture. The user interaction is provided by extending the MapOverlay to a MarkRegionOverlay class. The MarkRegionOverlay enables the Smartphone's touch capability to let the user mark a region the tag-cloud is created on.

4.2 Server

As mentioned before the tag-cloud server component is implemented in Java as well. The server component consist of three modules with the corresponding tasks: *network communication*, *harvesting data* as well as *data filtering and tag-cloud creation*.

The network communication module utilizes the Java ObjectStream class to exchange data with the client. The following simple communication protocol is used between the client and server nodes, as shown in Figure 3. After the TCP/IP connection is established, both nodes exchange version information to guarantee compatibly. Following this, the client node transmits the coordinates and region buffer values for the desired tag-cloud *(1)*. After successfully creating the tag-cloud *(5)*, the server sends it back to the requesting client *(6)*. Finally the connection is closed and the server is listing for incoming client connections again.

The second server module is responsible for harvesting the data for the desired location/region. At first the module queries the Wikipedia World gazetteer with the given region properties: latitude, longitude and region buffer *(2)*. In result all article names for this region are returned. Finally the entire content of those articles are harvested using the Wikipedia API (http://de.wikipedia.org/w/api.php) *(3) + (4)*.

The third module, utilizing the OpenCloud library, creates a tag-cloud from the previously harvested Wikipedia data. Various input filter mechanisms are supported by the library. So far the following filters are used in our prototype:

- *blacklist filter*: discarding all words matching a predefined blacklist
- *length filter*: discarding words shorter than 4 letters
- *score filter*: discarding all words with minor occurring-score

The blacklist contains Wikipedia specific patterns, like "edit", "hide" or "contents". They represent common patterns of the Wikipedia user-interface and appear in nearly all articles. This high occurrence results in a high tag score, while the tags are of minor relevance. For this reason the blacklist filter discards these words.

Following the blacklist filter, a length filter is applied. As threshold we choose a length of 4 letters, meaning all tags shorter than 4 letters are removed from the tag-cloud. It's the simplest way to remove most pronouns, verbs and articles from the tag-cloud. As the blacklist patterns, these tags get a high score, while they are of minor relevance. This admittedly restrictive filter is easy to implement, but implies in the drawback of discarding short names and abbreviations, which might be of relevance, as well. Therefore the filtering of pronouns, verbs and so on demands further research. For instance a suitable verb filter, discarding all verbs, could be applied.

The last applied filter is a score filter discarding all words with a minor occurring-score, which means in our case they are of minor relevance. This shrinks the tag-cloud to the highest scored tags and therefore saves valuable bandwidth during the transmission back to the clients Smartphone.

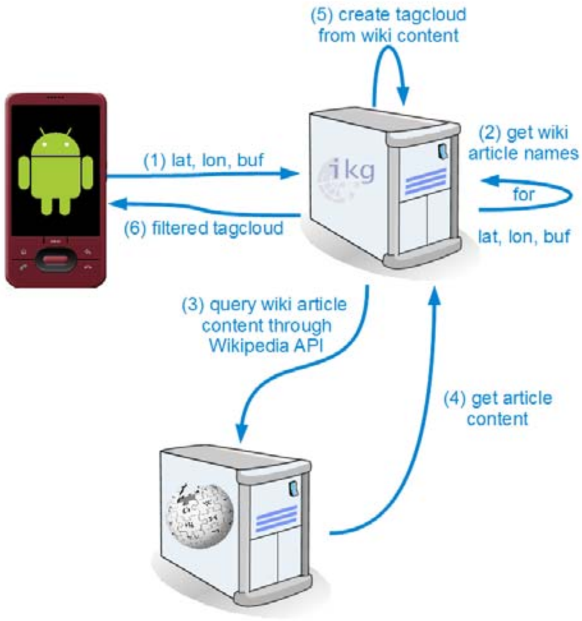


Figure 3: Client - server configuration and dataflow

5. EXAMPLES

As stated before we implemented a prototype on an Android based mobile phone. Figures 4 – 7 are showing some screenshots of the implemented application, demonstrating the user interaction, server communication und tag-cloud visualization parts.

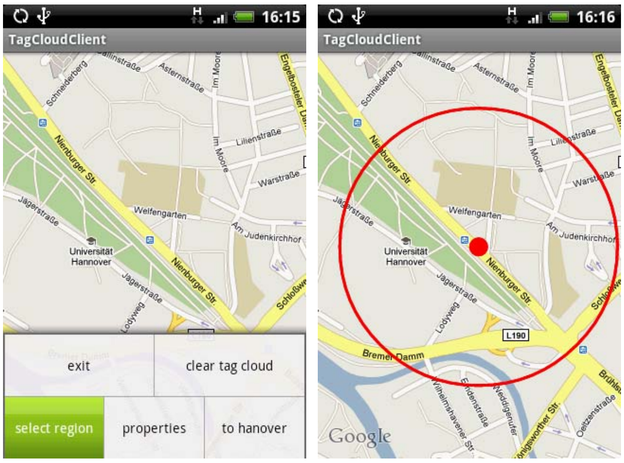
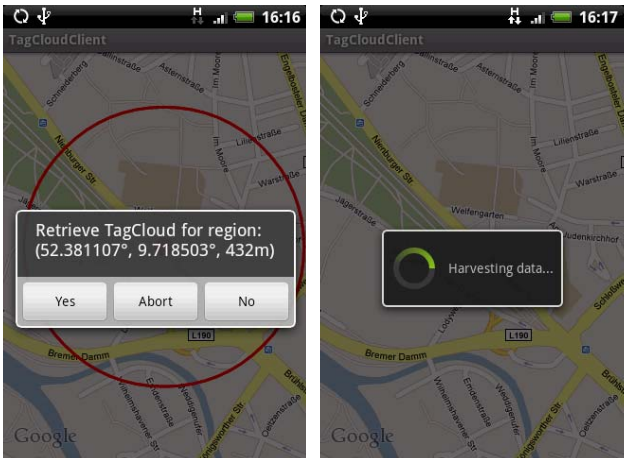


Figure 4: Query user interface (left); Select area dialog (right)

Figure 4 (left) shows an example of a possible user interaction. The user activates the *select region* option menu item. This enables the touch capability, so the user can select a region to build the corresponding tag-cloud.
After the user has marked a region, as shown in Figure 4 (right) the associated coordinates and the region radius are presented, as shown in Figure 5. After the user has confirmed the region, the region values are sent to the server. A progress dialog is shown to indicate the server harvesting the corresponding Wikipedia content which could take some time, depending on the size of the selected region.


Retrieve TagCloud for region:
(52.381107°, 9.718503°, 432m)

Harvesting data...

Figure 5: Harvest tag-cloud data dialog

After the server has finished harvesting the data and building the filtered tag-cloud, the tag-cloud is sent to the client device. The client compiles a picture from the tag-cloud and adds it to the map overlay, as shown in Figure 6.

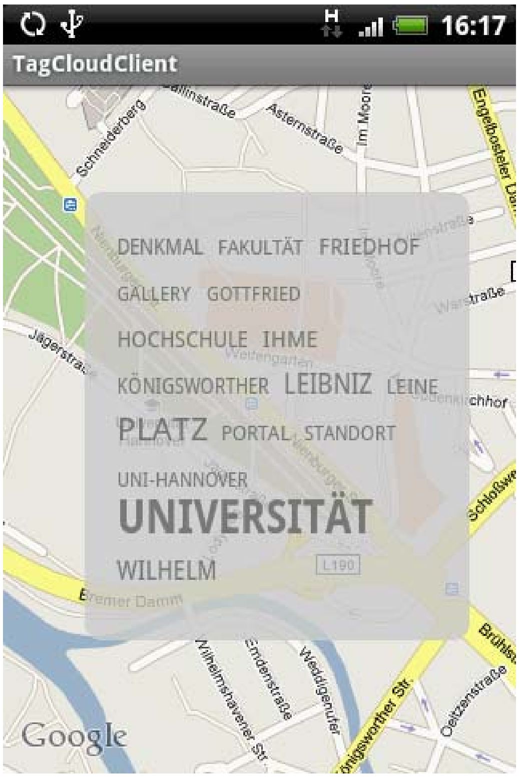


Figure 6: Visualization of generated tag-cloud

As described before the approach is not limited to individual points. Tag-clouds can be calculated along a route (either for decision points along the route, which is desirable if the tag-clouds are to be used to support pedestrian navigation or in a spatial buffer along the route if information on the surrounding area is of primary interest). Figure 7 shows a route visualization with tag-clouds for two decision points.

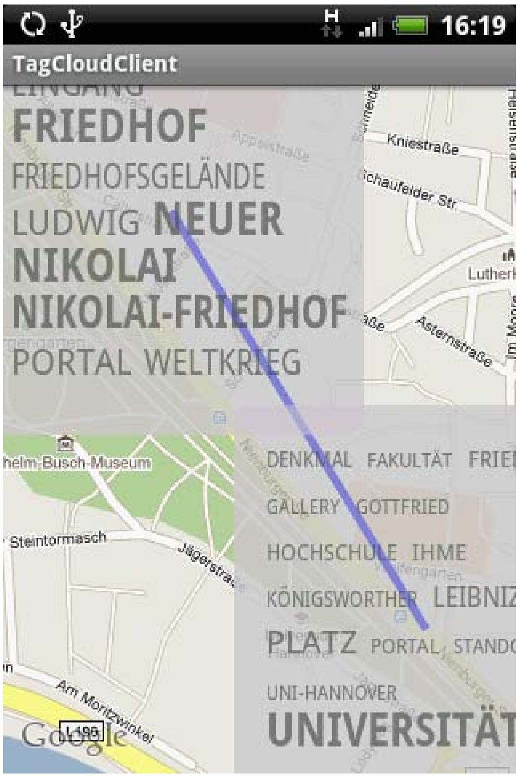


Figure 7: Visualization of route

6. RESULTS AND OUTLOOK

In order to validate the results from the given example, shown in Figure 6, we marked the locations of the presented tags in Figure 8. The tag-cloud from Figure 6 presents, among other, the tags: *Gottfried, Wilhelm, Leibniz, Universität, Uni-Hannover, Hochschule, Friedhof, Königsworther, Platz, Leine* and *Ihme.* Figure 8 shows their corresponding locations in the investigated area.

In contrast to the common Wikipoints overlay known from Google Maps or Google Earth, the generated tag-clouds present additional information, e.g. related events or related objects without an own geo-referenced article. For instance, the tag "Nikolai-Friedhof" (Nikolai-Cemetery) in the upper tag-cloud of Figure 7 is a Wikipoint, since it's the name of the Wikipedia article. This would be shown as a POI in the Wikipoints Google Maps overlay. Furthermore, the visualized tag-cloud presents the tag "Weltkrieg" (World War) as well. This tag would not be shown by the Wikipoints overlay, since it's not a geo-referenced Wikipedia article name for the investigated area. The tag gives the user the hint, that the marked area is related to one of the World Wars or there is a World War related object in that area.

We implemented and refined the tag-cloud visualization system as a client-server system that runs in part on mobile android devices and in part on a PC based server. The development experience using the Android platform was very constructive, especially compared to previous developments conducted for the Windows mobile platform: for instance, using Java on client as well as on server side made the prototype development quick and coherent. Furthermore, the use of an existing tag-cloud specification library (the OpenCloud library) simplified the communication between the client and server nodes significantly, reducing the communication to simple exchanges of serialized Java objects.

Currently no open library for rendering tag-clouds exists for the Android SDK, we therefore implemented a simple renderer for test purposes that lacks many of the advanced visualization facilities provided by current PC based tag-cloud renderers. While this limited the variety of the tag-cloud visualizations used in our prototype it was sufficient to validate the concept itself.

In order to improve the visualization further research on possible tag-cloud visualization schemes aligned to mobile phone displays would be worthwhile.

In addition sophisticated filter techniques are another field of interest. First, filtering words is highly language depended, which means that each language needs its own filter setup. For instance verbs will nearly always need to be discarded, so a language specific verb filter would be a good one to setup. Second a filter needs to be aligned to the source of the used content (e.g. Wikipedia). Each content source needs its own blacklist filter to discard typical source dependent words like *hide* or *edit* in the case of Wikipedia.

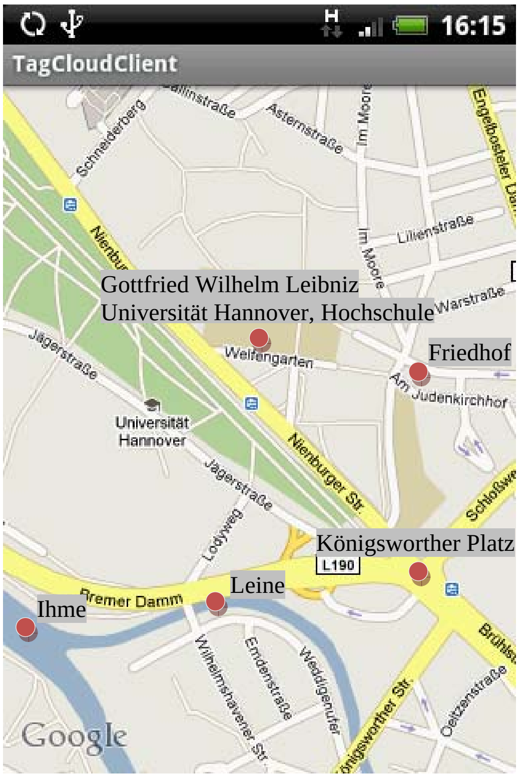


Figure 8: Found tags in the area under investigation

Initial experiences with the system are encouraging. In areas well covered by spatially annotated Wikipedia articles like the central areas of Hannover between 40 and 80 percent of the tags are meaningful landmarks. Since tag-clouds support fast browsing of information this is usually sufficient to provide users with pointers to potential objects of interest. Initial test users got quickly used to the system and were positive about the results delivered. Even test users who knew an area well were often surprised by the detail of information available.

The intention of the first prototype described in this paper has been to validate the viability of idea. In the future we aim to address several areas for refinement:

First, we aim to experiment with more advanced filtering algorithms to increase the signal to noise ratio in the tag-cloud and to extend the system to other languages, notably English.
Second, we intend to extend the data used for harvesting tags to other sources of geo-referenced data beyond Wikipedia, e.g. by incorporating Flickr tags.
We are also working on more refined rendering and adaptive layout strategies for the tag-clouds. Other open questions include the best size, the optimal tag count and the best visualization style to combine the cloud display with maps. Thus, while the use of public data sources seems to be promising to fill the gaps in spatial coverage in current LBS there remain many open questions to be addressed by future work.

7. REFERENCES

Flickr, 2009: http://www.flickr.com/photos/tags/ (accessed 28 Oct. 2009)

Lamantia, J. 2009: http://www.joelamantia.com/ideas/cartograms-tag-clouds-and-visualization/ (accessed 28 Oct. 2009)

Wikipedia, 2009: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates (accessed 28 Oct. 2009)

Coupland, D., 1995. *Microserfs.* Harper Collins.

Zipf, A., 2002. Location aware mobility support for tourists. Trends & Controversies. *IEEE Intelligent Systems*. Nov./Dec. 2002. pp. 57-59.

Baus, J., Cheverst, K. and Kray, C., 2005. A Survey of Map-based Mobile Guides. *Map-based Mobile Services Theories, Methods and Implementations*, Springer, pp.197-213.

Schilling, A., Coors, V. and Laakso, K., 2005. Dynamic 3D Maps for Mobile Tourism Applications. *Map-based Mobile Services Theories, Methods and Implementations*, Springer, pp. 233-244.

Viégas, F.B., Wattenberg, M., 2008. Tag Clouds and the Case for Vernacular Visualization. *ACM Interactions,* Vol. 15, No. 4, pp. 49–52.

Hecht, B., Raubal, M., 2008. GeoSR: Geographically Explore Semantic Relations in World Knowledge. *Proc. AGILE Conf. 2008*, pp. 95-113.

Kaser O., Lemire, D., 2007. Tag-Cloud Drawing: Algorithms for Cloud Visualization, CoRR, .

Milgram, S. *Environmental Psychology: People and Their Physical Settings*, 2nd ed. Holt, Rinehart and Winston, New York, USA, 1976, pp. 104–124.

Halvey, M. J. and Keane, M. T., 2007. An assessment of tag presentation techniques. *Proc. ACM WWW '07*, Banff, Alberta, Canada, May 2007, pp. 1313-1314.

Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R., 2007. Getting our head in the clouds: toward evaluation studies of tagclouds. Proc. ACM CHI '07, San Jose, California, USA, April 2007, pp. 995-998.

AN OPEN-SOURCE WEB ARCHITECTURE FOR ADAPTIVE LOCATION-BASED SERVICES

Gavin McArdle, Andrea Ballatore, Ali Tahir, Michela Bertolotto

School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland
gavin.mcardle@ucd.ie, andrea.ballatore@ucd.ie, ali.tahir@ucd.ie, michela.bertolotto@ucd.ie

KEY WORDS: Location Based Services, GIS, Personalisation, Web Technologies, Adaptive Interfaces

ABSTRACT:

As the volume of information available online continues to grow, there is an increasing problem with information overload. This issue is also escalating in the spatial domain as the amount of geo-tagged information expands. With such an abundance of geo-information, it is difficult for map users to find content that is relevant to them. The problem is intensified when considering Location-Based Services. These services, which are dependent upon a user's geographic location, generally operate on portable devices. These devices have a reduced screen size coupled with a limited processing power and so the need to provide personalised content is of paramount importance. Our previous work has focused on examining techniques to determine user interests in order to provide adapted and personalised map content which is suitable to display on portable devices. In this paper, in order to reduce the processing load on the user's device, a novel client server architecture is employed. The framework is designed using open-source, web-based technologies which monitor user locations and interactions with map content overtime to produce a user profile. This profile is then used to render personalised maps. By utilising the power of web-based technologies in an innovative manner, any operational issues between different mobile devices is alleviated, as the device only requires a web-browser to receive map content. This article describes the techniques, architecture and technologies used to achieve this.

1 INTRODUCTION

Information overload is a ubiquitous problem which is increasingly prevalent in the web domain. The issue is also present in the spatial domain where the quantity of spatially-referenced material has increased (Yang and Claramunt, 2005). Techniques which have been utilised in the predominantly text-based web domain are now being explored in the context of geo-spatial content. Traditional systems permit the user to personalise map content explicitly however such techniques can be time consuming and distorted by subjectivity. Furthermore, these approaches often distract the user from the main task at hand (Wu et al., 2008).

Monitoring interest indicators implicitly in order to obtain an insight into user preferences can alleviate this problem. Implicit profiling involves the system monitoring user interaction with an underlying interface which permits the automatic personalisation of information, and in the context of spatial data, the adaptation of map content. As implicit profiling is unobtrusive, it has a 100% completion rate and does not place any additional overhead on the user. This approach is employed successfully in several web-based systems where user actions such as link clicking, bookmarking and printing act as indicators of interest in a web-page's contents (Kelly and Teevan, 2003). Recently in the spatial domain similar interest indicators have been identified via interactions with map content (Brunato and Battiti, 2003; Mac Aoidh et al., 2007; Weakliam et al., 2005).

The need for personalisation of spatial data is further emphasised when dealing with Location-Based Services (LBS). Such services typically operate on mobile devices which often have limited processing power coupled with a restricted screen size and so there is a requirement to reduce the amount of information presented through personalisation and adaptation of spatial data. There are a multitude of diverse platforms and operating systems used in state-of-the-art mobile phones and developing a suitable LBS which can perform in such varied environments is challenging. Several standalone applications have been conceived (Kupper, 2005), however they are typically restricted to one class of device.

An approach to provide personalised LBS to users in a more accessible way is required. The research presented in this paper utilises open-source, web-based technologies which adhere to internet and Open Geospatial Consortium (OGC) standards to achieve this. Utilising web-based technologies permits a flexible client-server architecture to be implemented. Such an architecture, where processing can be distributed between server and client, reduces the computational load on the mobile device. Furthermore, open-source, web-based technologies are ubiquitous and only require the user device to be equipped with a suitable web-browser thereby alleviating most of the cross platform operational issues while also increasing the accessibility. In this paper the power of the client-server paradigm is combined with recent web technologies and applied in the context of adaptive LBS. Existing techniques which implicitly monitor user behaviour and context to provide personalised map content are demonstrated using this new approach.

The remainder of this paper is organised as follows: section 2 discusses related work in the area of user profiling, spatial recommendation systems and web-based technologies for developing spatial applications. section 3 outlines a sample application including the web-based architecture with a detailed description of the technologies and core algorithm used. Section 4 details the prototype which has been developed to evaluate the approach using a test scenario, while section 5 identifies limitations and directions for future work. In section 6, conclusions on the paper are made.

2 RELATED WORK

Although there is an extensive body of research in web personalisation (Albanese et al., 2004; Middleton et al., 2002), personalisation in the LBS domain remains relatively unexplored. The systems which have emerged use diverse techniques for both profiling and recommending and have not been widely adopted. For example, Hippie (Oppermann et al., 1999) is a personalised location-based content delivery service, tailored specifically for museums. Interests are inferred by analysing the information which the user viewed, in order to recommend content based on their position within the museum. Other researchers focus on interactions at the map level as a means of determining interests. Weakliam et al. (2005) monitor user interactions with the layers within the map and use this information to recommend semantic groups of objects. Likewise, the system described by Mac Aoidh et al. (2007) monitors user interactions with map content, focusing on the mouse position relative to individual map objects as an indicator of interest. This information is used to automatically generate a user profile from which relevant map objects can be recommended in the future. Similarly, PILGRIM (Brunato and Battiti, 2003) is a location-aware system which utilises the physical position of users as an interest indicator by assuming that proximity to objects is relevant in determining interests.

Such LBS recommender and personalisation systems tend to be standalone applications and restricted to a specific device type for a particular domain. These applications fail to take advantage of recent developments in web-based technologies which have been developed for geo-spatial data management and display. These services offer the potential for multi-platform deployment of LBS systems which can personalise information and adapt a map for users.This web setup offers several advantages over the traditional standalone approach. The use of the World Wide Web to give access to geographical information dates from 1993 (Longley et al., 2005). Since then GIS has successfully adapted to the Internet paradigm and has benefited from the momentum generated by the web, and the history of GIS and that of the Internet have become increasingly intertwined. Today, there is a fast-growing number of GIS applications deployed on the Web, ranging from mapping to routing and geographical yellow pages. The reasons for this success are various: the web is an established, widely used platform where accepted standards allow for smooth integration and manipulation of heterogeneous data types. Furthermore, the interactive and exploratory nature of navigating geo-referenced hyperlinked information is appealing for the end users. The shift from standalone GIS to webGIS is apparent. The use of standalone GIS has dramatically decreased in recent years (Longley et al., 2005).

Furthermore, thanks to the growing spread of Internet-enabled and location-aware smartphones, surveyed by Oliver (2009), WebGIS is increasingly becoming portable. In this fast moving context, the term *GIServices* defines the web applications offering distributed access to centralised spatial contents. Recent advances in client-side technologies allow for the development of complex interfaces that can be run on a web browser without installing additional software components thanks to the use of widely adopted standards. These standards have been utilised in numerous pieces of GIS software. As economic factors drive the shift towards open-source technologies (Sui, 2008) many GIS tools have been released under licenses similar to the GNU Public License (GPL). Since their introduction in the 90s, such licenses permit the sharing and distribution of source code and have boosted the development of open-source technologies. The quality of this software, ranging from mapping tools to topology libraries, has been constantly improving and in some cases has matched the functionality offered by commercial competitors. The significance of open-source software usage and development in the geo-spatial community is increasing. Sanz-Salinas and Montesinos-Lajara (2009) have surveyed this open-source ecosystem.

This paper describes an approach which utilises the benefits of open-source, web technologies to define a three-tier component based architecture to produce an adaptive map-based LBS. The resulting system is suitable for use on multiple devices via a standard web browser thus minimising cross platform operational issues. Map adaptation is achieved using techniques discussed by Ballatore et al. (2010) where an algorithm which implicitly monitors user interaction, in order to recommend tailored spatial content, is described. This algorithm is used in a prototype of the new architecture.

3 SYSTEM

The system proposed is an open web platform for spatial personalisation and visualisation. Based on a client-server architecture, the system delegates the task of computing complex algorithms, generating visual output and dealing with costly operations to the server. The clients, on the other hand, send requests to the server and render the output for the user. Within this domain interoperability is crucial. Firstly, input and output spatial data have to be defined in well-known formats. Secondly, the web-based approach minimises the coding overhead necessary to port the application to different mobile and desktop platforms. Last but not least, Open Web Services and an API expose the system functionalities on the Internet and allow external applications to interact with the system. The following section describes in detail the system architecture, while section 3.2 moves on to describe the personalisation algorithm.

3.1 Architecture

The architecture of the system proposed is structured in 3 tiers: client tier, middle tier and data sources. This approach emphasises the independence of the various components of the system, that can be deployed and combined in different contexts. Figure 1 illustrates the interaction among tiers, outlining the main components of the system and their logical position. The following sections describe these tiers in detail.

3.1.1 Client Tier The user interaction with the system takes place in this tier. The user visualises web pages containing spatial information on their device through a common web browser. These web pages display interactive maps and monitor certain actions performed by the user, such as mouse clicks, zoom, etc. Such actions get sent to the *Personalisation and Visualisation Service*. In a typical scenario a web page displays a dynamic map served by the *Map Renderer* and other information from the *Personalisation and Visualisation Service* (both situated in the Middle Tier).

3.1.2 Middle Tier This tier contains the core services and functionalities of the system. A web application hosts the Web Server Pages which constitute the access point to the system for the users and the *Personalisation and Visualisation Service*, in which several web services are deployed and exposed on the Internet.

This *Personalisation and Visualisation Service* tracks the user sessions, handles the user profiles, logs all of the relevant actions performed by the clients and keeps track of the user location when available. The personalisation and visualisation algorithms are also implemented within this service to take advantage of the server-side computational power and full access to

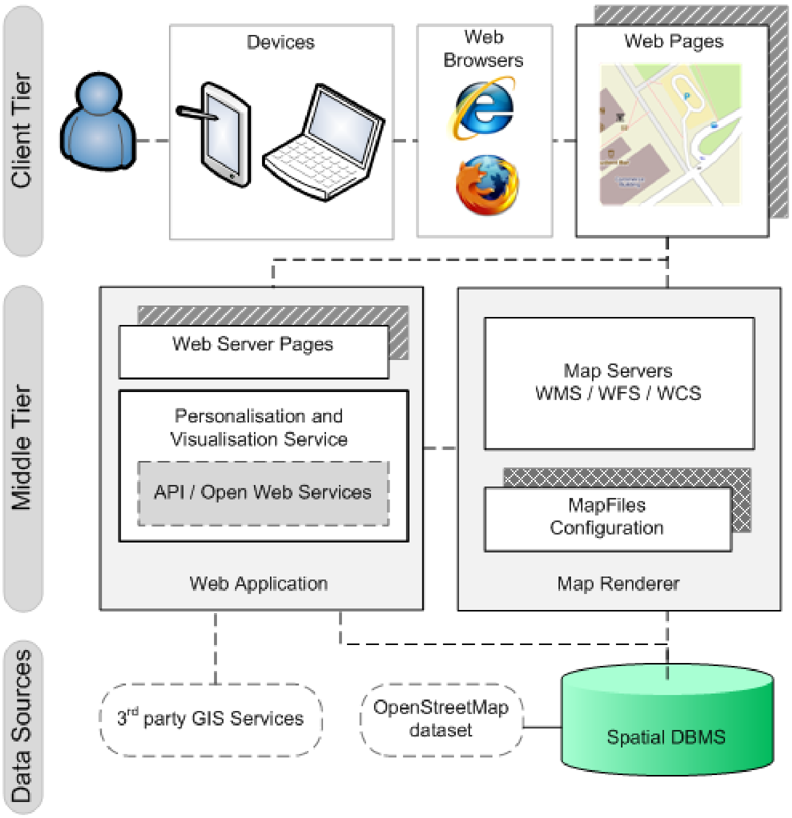


Figure 1: System architecture

the spatial datasets and recorded interaction. The *Personalisation and Visualisation Service* can be accessed by external applications through an API defining functions with parameters and the format of the returned values. The personalisation algorithm described in Ballatore et al. (2010) has been implemented in this service and it is outlined in section 3.2.

The *Web Server Pages* define the dynamic contents that are served to the clients. A Web Server Page can aggregate spatial and non-spatial data from heterogeneous sources, tailored for a specific user. Such a page gets rendered and delivered to the client. The navigation logic among these pages is defined and controlled by the Web Application.

At the same architectural level a *Map Renderer* is deployed. This component is able to load spatial data and to render them according to a dynamic map file configuration, which defines what data have to be rendered and with which visual style. The map file configurations are generated dynamically by the *Personalisation and Visualisation Service* according to the profile of the user receiving the spatial data. The *Map Renderer* hosts several map servers, which are spatial web services specifically designed to provide the clients with spatial information in standard and widely-adopted formats.

Both the *Personalisation and Visualisation Service* and the *Map Renderer* interact with the Data Sources to obtain and store the spatial information needed to execute their tasks.

3.1.3 Data Sources The main function of this tier is to provide the other tiers with spatial information. Its main component consists of a *spatial DBMS*, which stores spatial datasets (e.g. vector maps) and data related to the user profiles. The spatial DBMS is used by the *Personalisation and Visualisation Service* to run the personalisation logic, whilst the *Map Renderer* accesses it to perform the visualisation. The relevant spatial datasets can be imported from an external data provider (e.g. OpenStreetMap, etc) and stored in the spatial DBMS and updated on a regular basis.

Another apparent phenomenon on the Internet is the growing availability of commercial GIServices accessible for free (e.g. Virtual Earth, Google Earth, Google Maps and Cloudmade). These services rely on high-end commercial infrastructures to provide access to various datasets through Web APIs, mostly for routing, geocoding and reverse geocoding. The Web architecture described in this section lets the Web Application request data from such web services for with little development cost.

3.2 Personalisation Algorithm

This personalisation algorithm described by Ballatore et al. (2010) has been implemented in the *Personalisation and Visualisation Service*. The user sees the world as a set of *items*. An item is a point of interest representing a geographical entity (either a point or a polygon) which the user might have interest in. Certain types of item (such as shops, cinemas, etc) can be associated with relevant resources on the Web (such as web sites or web services). These items are displayed on an interactive digital map, rendered by the *Map Renderer*.

The *Personalisation and Visualisation Service* assigns an interest score to each item located within the current *interest radius*. The interest radius is inversely proportional to the current user speed. The interest score α for the item i is calculated with Equation (1) taking into account both historical interactions (*interaction*) and current user routing distance from the item (*proximity*)

$$\alpha_i = P_i P_R + I_i I_R \\ P_R + I_R = 1, \quad P_i \in [0,1], \quad I_i \in [0,1] \tag{1}$$

P_R and I_R are respectively the *proximity ratio* and the *interaction ratio*, meaning the weight the system attributes to each indicator. Those ratios are dynamic and change over time: the more the user interacts with items within the interest radius, the more the scales will be tipped toward P_R to emphasise proximity, and vice versa. Furthermore, the proximity score P_i and the interaction score I_i are normalised between the maximum and minimum score among the items within the proximity score. Given the volatile nature of user interests (Wu et al., 2008), a time decay function based upon the days elapsed since the last interaction with the item i is also applied on $|I_i|$ and $|P_i|$.

When a certain condition occurs in the user context and profile (e.g. user moves to a new area), the *Personalisation and Visualisation Service* triggers an adaptive action. For example, when the user either starts to explore a new area or alters the user profile by interacting with the map, the client requests new *recommended items*. The items having the highest interest score at a given time for a given location are sent to the map.

4 PROTOTYPE

After analysing the requirements for the architecture described in section 3.1, a survey of the available open-source technologies has been carried out to identify a suitable set with which a prototype can be developed. The open-source technologies which have been selected to implement a system prototype are depicted in section 4.1. Section 4.2 then describes a typical session in which a user interacts with the personalised map.

4.1 Technologies

This section describes the main open-source technologies that have been chosen for the implementation of a prototype of the system. This survey has mostly focused on software packages released under a GPL-like license, adopted by active online communities, supporting standard formats and reasonably stable and reliable. The areas surveyed are related to spatial/GIS services, spatial DBMS, spatial object-relational mapping and agile web development.

MapServer (http://mapserver.org) is an open-source geo-spatial web mapping and rendering tool which complies with OGC standards such as Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS) and Geography Markup Language (GML). Developed at the University of Minnesota with help from NASA, it has built-in support for various vector and raster formats. The core functionality of Mapserver is its MapFile, a configuration file defining the raster and vector layers along with their visual styling. In the system proposed in this paper it is possible to tailor a MapFile for a specific user, updating visual style and structure of the map dynamically. The conceptual simplicity of the MapFile is one of the main advantages of MapServer over similar systems. A significant feature of the MapFile is also the direct connection and support for many DBMS such as PostGIS and Oracle Spatial. An extensive description of MapServer can be found in Kropla (2005).

Within the architecture proposed in this paper, MapServer implements the *Map Renderer*.

Grails (http://grails.org) is an open-source web application framework based on the design paradigm *Convention over Configuration* (Richardson, 2009). Grails embraces agile methodologies and naturally complements Java application development, exploiting the features of Groovy, an open-source dynamic language for the Java Virtual Machine. Groovy provides features offered by other dynamic languages such as Ruby, Python or Smalltalk (Koenig et al., 2007). In its default configuration, Grails relies on Hibernate for the object-relational mapping. Grails offers a seamless integration with *Ajax* (Asynchronous JAvascript + Xml), one of the main technologies whose adoption led to the Web 2.0. Thanks to Ajax, web pages can request data from the server asynchronously in the background without breaking the interaction flow to reload the page. The use of Ajax has led to the development of complex web-based interactive interfaces.

Within the system proposed in this paper, Grails implements the Web Application, with its Web Server Pages and the *Personalisation and Visualisation Service*.

PostGIS (http://postgis.refractions.net) is an open-source plugin that adds support for geographic objects to the PostgreSQL object-relational database. PostGIS constitutes a spatial extension for PostgreSQL and can be used as a back-end spatial database for spatial and geographical applications. PostGIS supports the OpenGIS Simple Features Specification for SQL. It has been developed by Refractions Research as a project in open-source spatial database technology. PostGIS is being increasingly adopted as an alternative to costly commercial products such as Oracle Spatial and Microsoft SQL Server Spatial. For this reasons, several open-source GIS projects support it natively. PostGIS includes user interface tools, topology support, data validation, coordinate transformation and programming APIs. Performances of PostGIS on spatial queries have been compared with MySQL, DB2 and Oracle by Zhou et al. (2009).

In the system proposed in this paper, a PostgreSQL with PostGIS constitutes the Spatial DBMS.

Hibernate Spatial (http://www.hibernatespatial.org) is an extension to Hibernate for handling geographic data, developed by the GIS company GeoVise. Hibernate Spatial provides a standardised, cross-database interface to manipulate geographic data. The *Hibernate Query Language* is extended with geo-spatial data types following the OGC Simple Feature Specification. Optimised SQL code is generated for the underlying spatial DBMS, such as Postgis and Oracle Spatial. Hibernate Spatial has been successfully used by Hespanha et al. (2008).

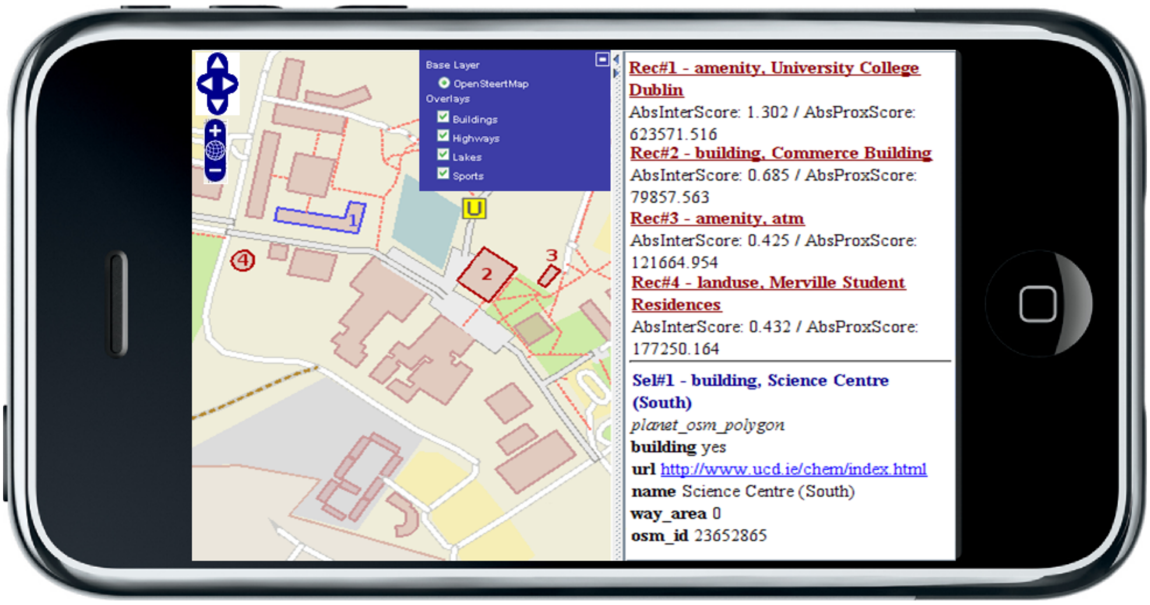
Within the system proposed in this paper, Hibernate Spatial bridges Grails with the Spatial DBMS.

OpenStreetMap (OSM) (http://www.openstreetmap.org) is a collaborative mapping project founded at University College London in 2004. OSM follows the Wikipedia model to create a vector dataset that is free to use, editable and licensed under a new map copyright scheme, similar to the GPL license for software. Although OSM does not have consistent quality assurance procedures, which makes it unsuitable for certain critical applications, its fast-growing free dataset has attracted the attention of the research community over the past 5 years. The success of OSM is largely due to the high cost and the restrictive licensing of most European datasets. The project is described extensively by Haklay and Weber (2008).

Within the system proposed in this paper, a section of the OSM vector map has been stored in the Spatial DBMS.

4.2 User Session

While the system operates with any web-enabled device, the prototype has been implemented for a smartphone. The prototype



Rec#1 - amenity, University College Dublin
AbsInterScore: 1.302 / AbsProxScore: 623571.516
Rec#2 - building, Commerce Building
AbsInterScore: 0.685 / AbsProxScore: 79857.563
Rec#3 - amenity, atm
AbsInterScore: 0.425 / AbsProxScore: 121664.954
Rec#4 - landuse, Merville Student Residences
AbsInterScore: 0.432 / AbsProxScore: 177250.164

Sel#1 - building, Science Centre (South)
planet_osm_polygon
building yes
url http://www.ucd.ie/chem/index.html
name Science Centre (South)
way_area 0
osm_id 23652865

Figure 2: Campus Map on iPhone

consists of an interactive web mapping application to assist users of University College Dublin as they navigate around the campus. The user commences their session by logging into the system, using the smartphone's web browser. The user views the personalised map on the smartphone. This map is personalised and adapted for the user based on their profile built during previous interactions using the personalisation algorithm described in section 3.2. The recommendations are highlighted on the map and detailed information is displayed on a panel to the right. The application, in this case on an Apple iPhone, is shown in Figure 2. A user can zoom in, zoom out, pan and has control over the visibility of layers such as buildings, roads, lakes and amenities. A user can see their position and recommendations on the screen. The recommendations are updated when certain interactions occur. All the user interactions such as map navigation and clicks on items are recorded along with the user location and time so that their profile is continuously updated. A user terminates their session by closing the web browser on the smartphone.

5 DISCUSSION

While there are benefits of the approach described here, it must also be recognised that there are some limitations which must be addressed. This section discusses these issues and how they can be overcome through further development and refinement. Furthermore, details of the evaluation which is required to assess the benefits of the approach are also provided.

The client-server architecture offers many advantages. In particular, it does not require any additional software on the user device. However, there are some limitations with this approach. The client-server architecture has well know bottle-neck issues and it is recognised that beyond a certain number of clients, performance of the server is compromised (Vatsavai et al., 2006). This is potentially troublesome in a geo-spatial application where complex analysis is carried on the server. However, as the power of portable devices improves, there are possibilities to resolve this issue by utilising load-balancing techniques described by Vatsavai et al. (2006), where some computation can be carried out on the user device.

The system described in this paper assumes that the user has a permanent connection to the Internet. This permits a constant stream of information to be exchanged between client and server. However, there are occasions when connection is interrupted or unavailable and contingencies to resolve this limitation are required. One possible solution involves developing client-side caching, whereby events and interactions are stored on the client and transmitted to the server when connection is available. This permits the server to continue user profiling and provide adapted and personalised maps using a complete interaction history. In order to improve map rendering and performance at the interface level, tile caching needs to be implemented on the server-side. Software such as TileCache (http://tilecache.org) works by prefetching map content which is most likely to requested next by the user. This improves performance of the interface and the response time of map interactions.

Now that a stable architecture has been implemented it is possible to use this as a test-bed for future developments. In particular, the algorithm described in section 3.2 can be refined to improve its performance. For example, at present the recommended items are calculated on the interaction history of a single user, however collaborative recommender systems in the web domain benefit from building collective group profiles (Schafer et al., 2007). By defining a user profile similarity function it is possible to cluster users who have similar interests and add new elements to the interest score calculation. The area of group profiling techniques is being surveyed in order to apply them in the geo-spatial domain and extend the algorithm described in this paper. Other techniques such as the introduction of negative scores to map objects is also being explored in the context of this algorithm.

Evaluating the performance of an adaptive LBS is crucial and non-trivial. Currently, a tool to generate user profiles by simulating common usage patterns of the system is being implemented. These profiles represent stereotypical behaviour of certain user categories within the campus prototype that has been developed (e.g. undergraduate students, lecturers, etc.) and will be used as a dataset for future preliminary experiments. These experiments will assist in highlighting both the strong and weak points of the approach and will help refine the design of experiments involving human subjects. Evaluation metrics for recommendations are also being examined. Adomavicius and Tuzhilin (2005) have outlined the advantages and limitations of empirical evaluation through such metrics. These considerations will be taken into account in the next steps of the design and development of the web-based personalisation system.

6 CONCLUSION

This paper presents a web-based system, which uses open-source technologies, to provide personalised and adapted map content to users on portable mobile devices via LBS. A new modular architecture, which comprises of reusable components has been described. The power of the client-server architecture within web-based GIS has been clearly outlined. Using open-source, web technologies, which adhere to OGC standards, removes interoperablily issues which currently exist when developing standalone LBS applications for specific devices. In such cases, the user devices require minimal software to run web based applications. Web technologies also permit the exploitation of existing non-commercial web services within the architecture. For example, external route planning services can easily be incorporated into the design. The system utilises an existing personalisation algorithm to carry out user profiling and map adaptation. The innovative architecture described here provides new opportunities to test this algorithm while also permitting the development of new LBS personalisation techniques.

ACKNOWLEDGEMENTS

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

References

Adomavicius, G. and Tuzhilin, A., 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering 17(6), pp. 734–749.

Albanese, M., Picariello, A., Sansone, C. and Sansone, L., 2004. Web personalization based on static information and dynamic user behavior. In: Proceedings of the 6th annual ACM international workshop on Web information and data management, ACM, pp. 80–87.

Ballatore, A., McArdle, G., Kelly, C. and Bertolotto, M., 2010. RecoMap: An Interactive and Adaptive Map-Based Recommender. In: SAC 2010: Symposium On Applied Computing (in press), ACM.

Brunato, M. and Battiti, R., 2003. PILGRIM: A location broker and mobility-aware recommendation system. In: Pervasive Computing and Communications, IEEE, pp. 265–272.

Haklay, M. and Weber, P., 2008. OpenStreetMap: user-generated street maps. IEEE Pervasive Computing 7(4), pp. 12–18.

Hespanha, J., van Bennekom-Minnema, J., Van Oosterom, P. and Lemmen, C., 2008. The Model Driven Architecture approach applied to the Land Administration Domain Model version 1.1-with focus on constraints specified in the Object Constraint Language. In: FIG Working Week, pp. 14–19.

Kelly, D. and Teevan, J., 2003. Implicit feedback for inferring user preference: a bibliography. ACM SIGIR Forum 37(2), pp. 18–28.

Koenig, D., Glover, A., King, P., Laforge, G. and Skeet, J., 2007. Groovy in action. Manning Publications.

Kropla, B., 2005. Beginning MapServer: Open Source GIS Development (Expert's Voice in Open Source). Apress Berkely, CA, USA.

Küpper, A., 2005. Location-based Services: Fundamentals and Operation. John Wiley & Sons Inc.

Longley, P., Goodchild, M., Maguire, D. and Rhind, D., 2005. Geographical information systems and science. John Wiley & Sons Inc.

Mac Aoidh, E., Bertolotto, M. and Wilson, D. C., 2007. Analysis of implicit interest indicators for spatial data. In: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, ACM, pp. 1–4.

Middleton, S., Alani, H., Shadbolt, N. and De Roure, D., 2002. Exploiting synergy between ontologies and recommender systems. In: 11th International WWW Conference, Semantic Web Workshop, Citeseer, pp. 41–50.

Oliver, E., 2009. A survey of platforms for mobile networks research. ACM SIGMOBILE Mobile Computing and Communications Review 12(4), pp. 56–63.

Oppermann, R., Specht, M. and Jaceniak, I., 1999. Hippie: A nomadic information system. Lecture Notes in Computer Science pp. 330–333.

Richardson, C., 2009. ORM in dynamic languages. Communications of the ACM 52(4), pp. 48–55.

Sanz-Salinas, J. and Montesinos-Lajara, M., 2009. Current Panorama of the FOSS4G Ecosystem. Novática pp. 43–51.

Schafer, J., Frankowski, D., Herlocker, J. and Sen, S., 2007. Collaborative filtering recommender systems. Lecture Notes In Computer Science 4321, pp. 291–324.

Sui, D., 2008. The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. Computers, Environment and Urban Systems 32(1), pp. 1–5.

Vatsavai, R., Shekhar, S., Burk, T. and Lime, S., 2006. UMN-MapServer: A high-performance, interoperable, and open source web mapping and geo-spatial analysis system. Lecture Notes in Computer Science 4197(2006), pp. 400–417.

Weakliam, J., Bertolotto, M. and Wilson, D., 2005. Implicit interaction profiling for recommending spatial content. In: GIS '05: Proceedings of the 13th annual ACM international workshop on Geographic information systems, ACM, pp. 285–294.

Wu, D., Zhao, D. and Zhang, X., 2008. An Adaptive User Profile Based on Memory Model. In: Web-Age Information Management., pp. 461–468.

Yang, Y. and Claramunt, C., 2005. A hybrid approach for spatial web personalization. Lecture notes in computer science 3833, pp. 206–221.

Zhou, Z., Zhou, B., Li, W., Griglak, B., Caiseda, C. and Huang, Q., 2009. Evaluating query performance on object-relational spatial databases. In: 2nd IEEE International Conference on Computer Science and Information Technology, pp. 489–492.

CAMPUSGIS ROUTING
– A WEB-BASED LBS FOR THE UNIVERSITY OF COLOGNE

U. Baaser*, R. Laudien, G. Bareth

Dep. of Geography, University of Cologne, 50923 Koeln, Germany – (u.baaser, rlaudien, g.bareth)@uni-koeln.de

Commission II, WG 7

KEY WORDS: Internet/Web, Orientation, Database, routing, mobile

ABSTRACT:

The CampusGIS of the University of Cologne (http://www.campusgis.de) is designed, developed and established by the GIS & Remote Sensing research group at the Department of Geography. It is a web-based application to provide and visualize general and spatial campus information, e. g. advanced search functions, orientation, routing, and facility management. Several existing relational database systems of the University of Cologne with spatial data were connected within that online GIS environment. The overall design approach is based on this database linkage with spatial data to provide location-based services (LBS) (Baaser et al. 2008). This paper presents the major recent networking improvements based on a topological validated line-network to provide online routing applications for pedestrians and comparable user groups. Therefore, several objects and attributes of the CampusGIS geodatabase were surveyed, digitized and joint to selected objects of the ATKIS-street dataset (ATKIS = Authoritative topographic cartographic information system of the Federal Republic of Germany).

1 INTRODUCTION

Due to the lack of digital information about buildings, institutions, and persons, the CampusGIS of the University of Cologne (http://www.campusgis.de) was designed, developed and established by the GIS & Remote Sensing research group at the Department of Geography. CampusGIS is a web-based application to provide and visualize general and spatial campus information, e. g. advanced search functions, orientation, routing, and facility management. The overall approach of this project is to link different non-spatial databases to spatial entities within an online GIS environment. To be platform independent the internet was chosen as it comes with a digital, forward-looking and ubiquitous technology. Besides the web, GIS technology is used to provide geodata and attribute information about buildings, staff, and institutions in a spatial context.

The intern communication of CampusGIS is based on AJAX (Asynchronous JavaScript + XML) that allows a fast "Web 2.0"-*look and feel* (O'Reilly 2005). The latest release of the system is an ESRI ArcGIS Server Java-based application.

CampusGIS contains the following applications (status quo: December 2009):

- Web-based search engine for staff, buildings and institutions
- Building visualization in 3D – LoD1
- Embedding of CAD-drawings for facility management
- Pedestrian routing (under construction)
- GUI for mobile devices (under construction)
- Pedestrian sight-seeing tours based on several themes, e. g. history or architecture; each with 7-10 locations showing points of interest with background information (under construction)

2 NETWORK ANALYSIS APPROACH

This paper presents the development of the pedestrian routing application which is embedded into the latest version of CampusGIS. This routing application is based on the network analysis approach (ESRI 2009). According to Bill (1999), network analysis with predominantly curve objects are similar-distinguished GIS applications like intersections of polygons or interpolations of point objects. Networks are designed to simulate the flow of goods or data, e. g. water or electricity through a utility network, oil and gas pipelines on a commodity network or traffic on a road network (Curtin 2008). In the presented context the network analysis provides a routing application for pedestrians and comparable user groups at the Campus of the University of Cologne by using the three components:

- Topology
- Modeling a multimodal network
- Calculating the Shortest Route model

3 THE NETWORK DATASET

The CampusGIS geodatabase consists of different datasets: Besides (i) alphanumerical data, (ii) aerial and satellite images, and (iii) georeferenced CAD-drawings, geospatial geometry data is stored as feature classes in a relational file geodatabase, containing (iv) authoritative topographic geodata, and (v) surveyed data of buildings, building-entrances, footpaths and retail trade and services.

3.1 Topology

The digital base map of the system is derived by selected and grouped objects of the ATKIS-dataset (Authoritative topographic cartographic information system of the Federal

Republic of Germany), combined with footpaths, buildings of the university and building-entrances (Baaser et al. 2008).
For verification whether all the footpaths, buildings and entrances are digitized correctly, several rules are implemented in the geodatabase with regard to its topology. A topology describes the connections and relationships between objects. The topology model is based on the mathematical graph theory and describes points as nodes and lines as edges (Bernhardsen 2002). Topology rules are defined to erase topology errors. Compared to geometry data, topology data does not know the distance between two nodes, but their spatial distribution to each other (Bill 1999a).

The CampusGIS routing algorithms use the following implemented topology rules (ESRI 2004):

- Buildings must not overlap
- Entrance-points must be covered by an endpoint of a footpath
- Entrance-points must be covered by the boundary of buildings
- Room-entrance-points must be covered by the boundary of rooms
- Room-entrance-points must be covered by endpoint of a indoor-footpath
- Footpaths must be single part
- Footpaths must not overlap
- Footpaths must not self-overlap
- Footpath must not overlap with streets
- Indoor-footpaths must be single part
- Indoor-footpaths must not overlap
- Indoor-footpaths must not self-overlap
- Indoor-footpaths must not overlap with footpaths
- Indoor-footpaths must not overlap with streets
- Rooms must be covered by building-polygons
- Streets and Footpaths must not have dangles

3.2 Graph theory

A mathematical graph is the basic principle to describe the topology of a network. A graph is connected, if there is a connection from any node to any other node. If two arbitrary nodes are connected through particularly one path, that graph is known as a tree. A graph is planar, if all the paths are at the same level (Fig. 1).

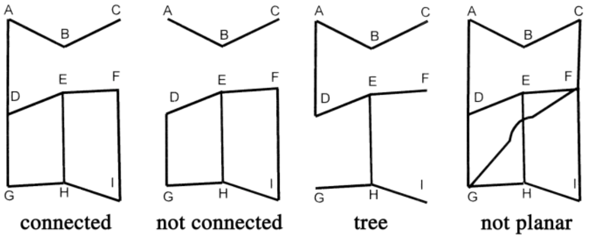


Figure 1. Types of graphs (de Lange 2005, p. 94, modified).

If one point is marked as the starting- and another as the endpoint, the graph is directed. A directed graph is marked with arrows. The edges can be weighted, e. g. according to travel distance or time (Elias 2007).

3.3 The multimodal network CologneMultiNet

Joint modeling of different networks results in a multimodal network (Bartelme 2005). With a multimodal network, it is possible to union numerous topologies, e. g. streets, footpaths, railroads, etc. into one single complex network. This provides the possibility of working with different topologies from different entities. Within this multimodal network the connectivity is defined by connectivity groups.

To meet these requirements with the routing extension of CampusGIS, streets and footpaths are stored in the first connectivity group CG1, tram tracks in CG2, and the tram stations are connection points and therefore stored in both CGs.

Based on this theory, a multimodal network named CologneMultiNet was developed for CampusGIS, containing seven object classes. Every edge is associated into one of the two above mentioned connectivity groups:
CG1: street network
CG2: tram network

Tram stations are point objects connecting the two connectivity groups to a multimodal network. The tram stations allow changing from one network to the other.

Within feature classes the connectivity policy defines how to model line segments as edges: If “end point”-policy is set, line features are split into multiple edges only at their endpoints. With “any vertex“-policy line features become edges at coincident vertices (ESRI 2009). The tram lines become edges only at coincident endpoints, because there is no way to get in and off or change one’s direction except at tram stops. The “any vertex”-connectivity is set to footpaths and streets. At point features the default is to honor the edge source's connectivity policy. As tram stops may be placed at an intermediate vertex, the tram stop-point feature policy will “override” the default behavior of connecting a junction to a given edge.

4 CALCULATION OF THE SHORTEST ROUTE

The CampusGIS-Routing application allows the calculation of user-aware routes: The system provides the user the choice of either pedestrian routes or special routes for handicapped and disabled people. To solve a route, CampusGIS differentiates between algorithms without stairs or without stairs and ramps with an inclination of more than 6%. Based on the network dataset CologneMutiNet “calculate shortest route”-models are generated with ESRI ArcGIS Model Builder. The model for the calculation of routes without stairs and ramps is exemplarily displayed in figure 2: CologneMultiNet comes with the network attributes distance and time of travel for pedestrians as resistance to any line segment. It is the input dataset to the “Make route layer”-tool that creates a new data layer. With the “Add locations”-tool the user is prompted to specify the starting point and the destination. “Add locations”-tool is also used to load the barriers point objects avoiding stairs and steep ramps depending on user’s selection. The tool “solve” generates a network analysis layer by using ESRI’s Network Analyst’s common algorithm of Edsger W. Dijkstra to find routes on a connected, positive weighted and directed graph (Dijkstra 1959, ESRI 2009). Due to the fact that ArcGIS Server does not accept the data type “network analysis layer” as an output parameter, the routes sublayer is selected using the “Select data”-tool. This sublayer is shown as the generated route. Simultaneously a python script is accessed to convert the text directions in HTML (ESRI 2009).

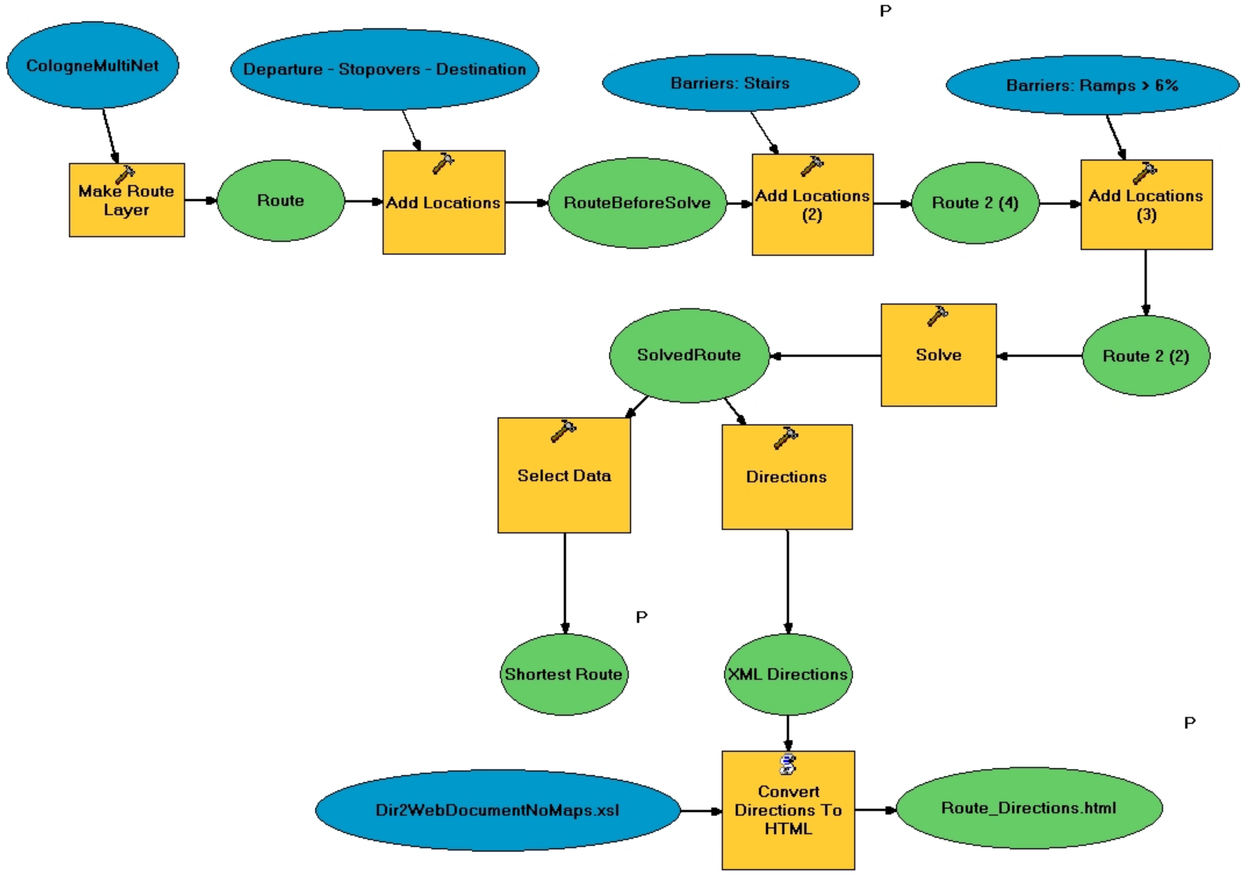


Figure 2. “Calculate Shortest Route for handicapped”-Model

5 RESULTS

The CampusGIS serves several GIS functionalities (Fig. 3). In detail, the following tools are implemented: There are buttons to zoom into a larger scale, into a smaller scale, to full extent, and to Campus extent. With the activated ‘hand’-button it is possible to pan the map in every direction. The arrows switches to the last displayed views. Mouse-click on the identify-button ‘i’ opens a pull down menu. Within that menu, the user is able to select a layer to receive specific information about university buildings, entrances, footpaths, and retail trade and services. A further mouse-click into the map shows alphanumerical information about buildings, footways, entrances, and shops. The ‘field-glass’-buttons open complex search forms for searching for staff, buildings, institutions, or catchwords. The results are shown at the top of the right hand of the map (Baaser et al. 2006a). Additionally there are buttons to call the routing applications for pedestrians (Fig. 3: 1) or disabled persons (Fig. 3: 2, 3). The shortest route-application comes with a form the user is called to specify departure and destination as well as optional stopovers by mouse-clicking into the map (Fig. 3: 4). With the ‘Run’-button the user starts the calculation of the route. The results are displayed: the calculated shortest route is drawn into the map (Fig. 4) and the directions are listed. The database connections and the routing functionalities are implemented in a Java-based web application using Enterprise Java Beans (EJBs). According to Brabec and Samet (2007) the Java environment has emerged as “the platform of choice” for cross-platform internet applications. Hence this application follows that actual approach. With handheld devices like mobile phones or personal digital assistants (PDA), routing and navigation by means of location-based services (LBS) will be commonly used applications in the near future (Hennig et al. 2009).

6 CONCLUSIONS AND OUTLOOK

The integration of routing applications results in a great benefit for the CampusGIS user. More detailed input data and data from different sources could enhance the scope of the described system. For better visualization a 3D-LoD1 model has been generated (Hennig 2008), its implementation is under construction. Routing and navigation in combination with 3D-data is much more user-friendly as it comes closer to the perception of the real world.

Additionally, a graphical user interface for mobile devices will be implemented in the future version. User’s position will be grabbed from a W-LAN-router, a cell-ID of mobile telecommunication network or –this would give the highest accuracy– from an internal GPS receiver. With this accurate input location data CampusGIS will provide navigation in real time.

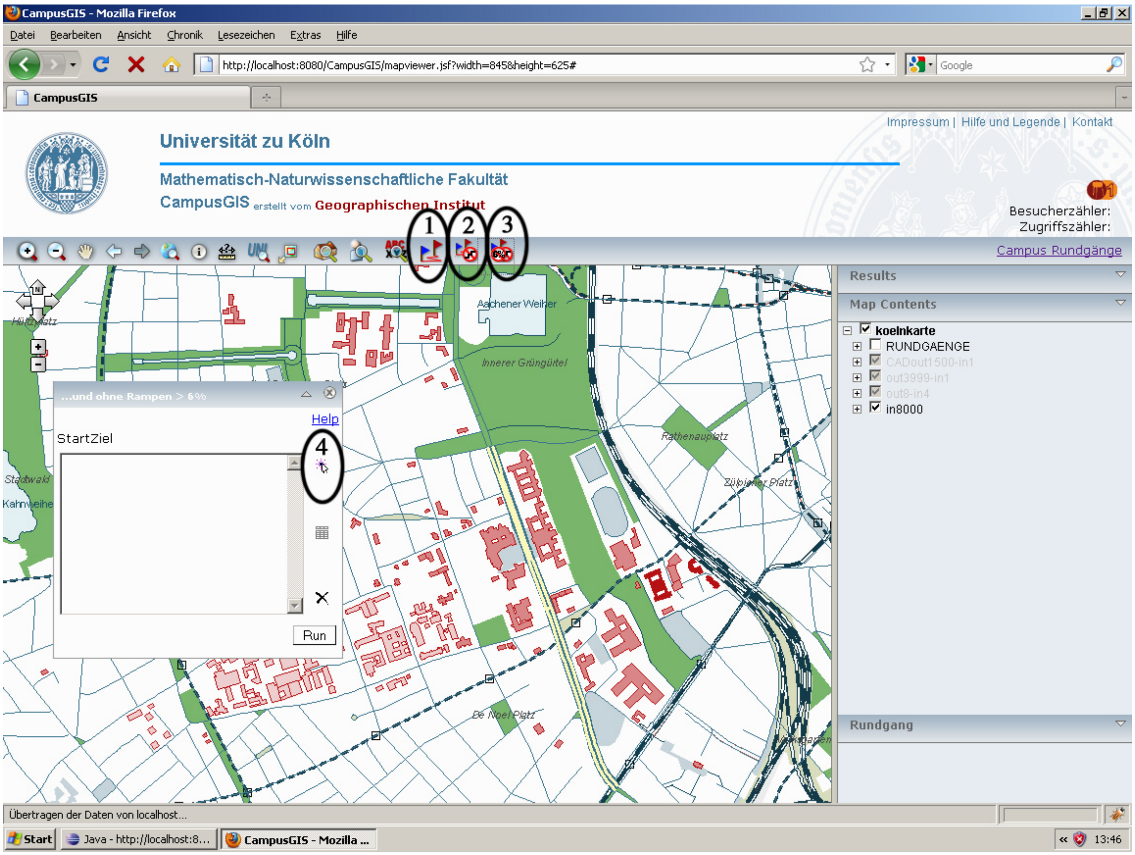
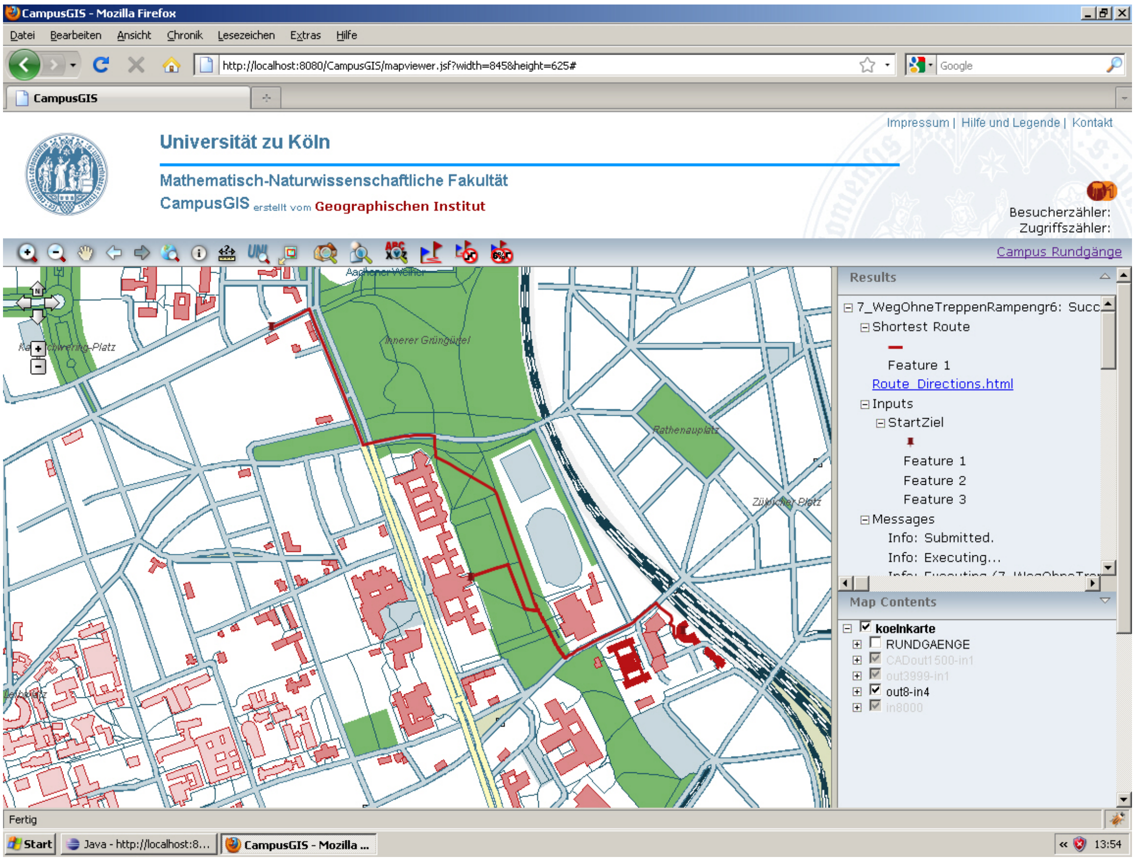

CampusGIS - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://localhost:8080/CampusGIS/mapviewer.jsf?width=845&height=625#
Google
CampusGIS
Impressum | Hilfe und Legende | Kontakt
Universität zu Köln
Mathematisch-Naturwissenschaftliche Fakultät
CampusGIS erstellt vom Geographischen Institut
Besucherzähler:
Zugriffszähler:
Campus Rundgänge
1
2
3
4
...und ohne Rampen > 6%
Help
StartZiel
Run
Aachener Weiher
Innerer Grüngürtel
Rathenauplatz
Zülpicher Platz
Stadtwald
Kahnweiher
De Noel Platz
Results
Map Contents
koelnkarte
RUNDGAENGE
CADout1500-in1
out3999-in1
out8-in4
in8000
Rundgang
Übertragen der Daten von localhost...
Start
Java - http://localhost:8...
CampusGIS - Mozilla ...
13:46

Figure 3. Screenshot CampusGIS-Routing application


CampusGIS - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://localhost:8080/CampusGIS/mapviewer.jsf?width=845&height=625#
Google
CampusGIS
Impressum | Hilfe und Legende | Kontakt
Universität zu Köln
Mathematisch-Naturwissenschaftliche Fakultät
CampusGIS erstellt vom Geographischen Institut
Besucherzähler:
Zugriffszähler:
Campus Rundgänge
Aachener Weiher
Innerer Grüngürtel
Rathenauplatz
Zülpicher Platz
Results
Map Contents
koelnkarte
RUNDGAENGE
CADout1500-in1
out3999-in1
out8-in4
in8000
Fertig
Start
Java - http://localhost:8...
CampusGIS - Mozilla ...
13:54

7_WegOhneTreppenRampengr6: Succ...
Shortest Route
Feature 1
Route_Directions.html
Inputs
StartZiel
Feature 1
Feature 2
Feature 3
Messages
Info: Submitted.
Info: Executing...

Figure 4. Screenshot CampusGIS - calculated route

REFERENCES

Baaser, U.; Gnyp, M. L.; Hennig, S.; Hoffmeister, D.; Köhn, N.; Laudien, R.; Bareth, G., 2006a Online CampusGIS for the University of Cologne: a tool for orientation, navigation and management. In Wu, H.; Zhu, Q. (Edts.): *Geoinformatics 2006: Geospatial Information Technology*. Wuhan, China, 64211L

Baaser, U.; Hennig, S. D.; Aasen, H.; Dornauf, E.; Gnyp, M. L.; Hoffmeister, D.; Köhn, N.; Louwen, B.; Laudien, R.; Bareth, G., 2008 AJAX-based linkage of databases for location-based-services: The Online-CampusGIS of the university of Cologne. In: CHEN, J.; JIANG, J.; NAYAK, S. (Edts.): *ISPRS Congress Proc. Bd. XXXVII, Part B4, Commission IV*. Beijing, China, XXXVII ISPRS Congress, pp. 745–750

Bartelme, N., 2005 *Geoinformatik: Modelle, Strukturen, Funktionen* Springer, Berlin

Bernhardsen, T., 2002 *Geographic Information Systems: An Introduction.* John Wiley & Sons, New York

Bill, R., 1999a *Grundlagen der Geo-Informationssysteme. Bd. 1: Hardware, Software und Daten.* Herbert Wichmann Verlag Heidelberg

Brabec, F.; Samet, H., 2007 Client-Based Spatial Browsing on the World Wide Web. In: *IEEE Internet Computing* Vol. 11 (2007), No. 1, pp. 52–59

Curtin, K. M., 2008 Network Data Structures. In: Kemp, K. (Edt.): *Encyclopedia of Geographic Information Science*. Sage Publications Los Angeles, pp. 314–317

De Lange, N. 2005 *Geoinformatik in Theorie und Praxis*. Springer, Berlin

Dijkstra, E. W. 1959 A note on two problems in connexion with graphs. In: *Numerische Mathematik 1* pp. 269–271

Elias, B., 2007 Pedestrian Navigation - Creating a tailored geodatabase for routing. In: Kaiser, T.; Jobmann, K.; Kyamakya, K.. (Edts.) *WPNC'07: 4th Workshop on Positioning Navigation and Communication* Workshop Proceedings, pp. 41–47

ESRI, 2004 *ArcGIS Geodatabase Topology Rules* Environmental Systems Research Institute, Inc., Redlands, CA.

ESRI, 2009 ArcGIS Desktop 9.3 Help http://webhelp.esri.com/arcgisdesktop/9.3 (accessed 10 Dec. 2009) Environmental Systems Research Institute, Inc., Redlands, CA.

Hennig, S. D. 2008 *Prozessierung von Laserscanndaten zur Erstellung eines 3D-Stadtmodells: CampusGIS-3D* Diplomathesis, Universität zu Köln

Hennig, S. D.; Baaser, U.; Bareth, G., 2009 CampusGIS-3D established with ESRI's product family. In: *Proceedings of ESRI International User Conference 2009*, ESRI UC No. 1110

O'Reilly, T., 2005 What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed 09 April 2008)

LOCATION-AWARE PERSONAL LIFE CONTENT MANAGER AND ITS PRIVACY FUNCTIONS

Hideki Kaji* and Masatoshi Arikawa

Center for Spatial Information Science, The University of Tokyo
5-1-7 Kashiwanoha, Kashiwa-Shi, Chiba 277-8568 Japan
(kaji, arikawa)@csis.u-tokyo.ac.jp

KEY WORDS: Personal Life Content, Location Based Service, Privacy, Information Sharing Levels, Mobile Computing

ABSTRACT:

In this paper, we will introduce a location-aware personal life content manager (laPLCm) and the functions of location privacy for data representation, query, transmission and positioning methods on laPLCm. laPLCm allows users to be reminded and access their own personal life content by spatial keys of the content and the users in their real lives. Examples of personal life content representing past, present and future events are diary, schedule, to-do list, GPS logs, photos and videos which are recorded, created and stored with our daily mobile devices. Personal life content cannot be fully treated on current commercial LBS. Our proposed laPLCm gives a new platform for users to easily generate their own LBS for themselves, their families, friends and colleagues using their personal life content in the form of blog as well as original privacy functions. The functions of setting levels of information sharing to each of persons and groups flexibly realize part of the privacy for data in our system. Personal mapping services introduced in the paper prevent from recoding positions of users in the servers of web mapping providers. Our proposed self-positioning methods are also significant to keep the location privacy for positioning methods. Furthermore, we demonstrate our prototype system based on the architecture of location privacy, and discuss usability, feasibility and sustainability for the system with comparison of present commercial LBS.

1. INTRODUCTION

Location based services (LBS) grow popular and many people use these kinds of services on their mobile phones and other mobile devices with GPS receivers. For example, users can find their positions on maps, search points of interest (POI) around them, generate itineraries of their trips using complex time tables of public transportation, and navigate in the real world (Arikawa et al., 2007). On the other hand, there are many users who fear a lack of security and privacy of their location information (Dobson et al., 2003; Nouwt, 2008). These users think service providers may estimate their activities and movement patterns in life from their location information if they keep using commercial LBS.

On the Internet, people make personal life content, for example records of dairy activities, their opinions for interesting things, to-do lists and schedules as user generated contents like blog, twitter (Twitter, 2009), SNS (Boyd, D. et al., 2007) and video/photo sharing services. These systems have various privacy policies and features for privacy setting so users can control sharing level of their content. One of the reasons why people keep recording on blogs is not only informing other users about author's opinions, but also retrieving them as needed (Nardi et al., 2004). Most of the records, however, might never be accessed in their lives, and many of these personal records include spatial information which can be provided as spatial content in LBS. From this point, we developed and started experiment of blog based location-aware personal life content manager (laPLCm). laPLCm can provide users with their personal information or services based on locations and their personal information which are managed in blog. Also, users can easily launch user-generated LBS for themselves as well as other users on the Internet. On the other hand, privacy factor becomes much more important when LBS have functions to treat personal life content. In this paper, we focus on privacy settings for our proposed laPLCm and our developed prototype system of it.

2. PRIVACY SETTINGS ON SOME SERVICES

In this section, we will introduce treatments of privacy on some Internet services.

2.1 Social Network Service (SNS)

SNS is one of the most popular services on the Web. It provides environments for building online communities. Boyd and Ellison defined that it allows individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system (Boyb, D. et al., 2003). Facebook (Facebook, 2009) is a popular SNS that was designed for communication between students at first. It has simple privacy setting for sharing user's profiles. Users can set

* Corresponding author. This is useful to know for communication with the appropriate person in cases with more than one author.

statuses to each attribute like basic information (e.g. sex, birthday, home town and political view), self introduction (e.g. activity, hobby, music, TV programs) and friend list to limit accessing from other users. Users can choose a restriction status from four options.

- **Everyone**
- **Friends of Friends**
- **Only Friends**
- **Custom**

Authors can add restricted friends if you set the status to **Custom**. Furthermore, authors can set posting ability to their content from other users. This feature is easy to set restriction but it cannot set these statuses to each post separately so it has lack of flexibility

2.2 Google Latitude

Google Latitude (Google Inc., 2009a) provides an environment for sharing one's current location information and messages with Gmail (Google Inc., 2009b) users as a Web service. Its location sharing setting is very simple. Users set some sharing status of his/her location information for each Gmail contact. There are four statuses on acceptance of location information sharing.

- Accept and share back
- Accept, but hide my location
- Do not accept
- Block

Furthermore, users can choose a detail of location information from three levels

- Share best available location
- Share city level location only
- Hide from this friend

Additionally, users can set their location by manual pointing instead of using GPS. This setting feature is very simple, but it may take much time if you have a lot of Gmail contacts.

3. LOCATION-AWARE PERSONAL LIFE CONTNET MANAGER

We developed a prototype system to realize new Location-Aware Personal Life Content Manager (laPLCm). On this system, to create and store personal life content with location information, we adopted blog for a base system of the LBS server. We thought, blog had become popular on the Internet and many people have or had their own blogs. It is an easy way to make personal life content on the Internet. Thus, we decided to utilize location information on a new blog system so that users can create spatial personal life content easily. Furthermore, we selected mobile phones as a platform for LBS client applications. Latest mobile phones equipped with various features such as GPS receiver, motion sensor, digital compass, digital camera and network accessibility that provide a good environment for using spatial personal life content in the real world.

3.1 Architecture of laPLCm

Figure 1 shows the architecture of our laPLCm which is designed as an open platform to realize laPLCm based on protocols of the Internet. The system is constructed with the place enhanced blog and the LBS client on mobile phones. Our LBS server, that is, a place enhanced blog application is coded by Perl as a Web CGI application, thus they are working on Web servers and using HTTPS to communicate with Web browsers and LBS clients. LBS clients connect the interface script to interactively retrieve, and display POIs from the place enhanced blog through user-friendly GUI on the screen of a mobile phone.

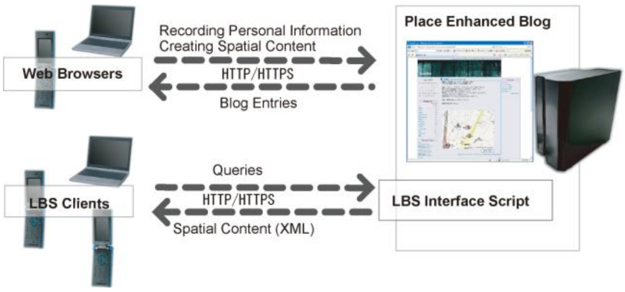


Figure 1. Architecture of pTalk, that is, the name of laPLCm.

3.2 Place Enhanced Blog

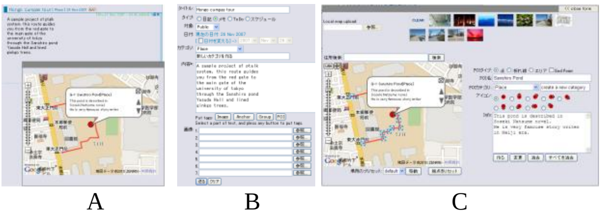


Figure 2. Entry view (A), entry input form (B) and place information input form (C). (Map images: copyright 2009 Google, copyright 2009 ZENRIN)

Place enhanced blog provides users with an extended function of dealing with spatial information such as point of interest (POI) and area of interest (AOI) in addition to general functions of common blog systems such as browsing and managing personal information. Users add place descriptions to their blog entries through blog input interfaces on Web browsers. To create an entry and corresponding place descriptions, users need to use two input forms, one is entry input form a simple input form same as normal blogs. Other one is a place information input form. Users can create multiple place objects on it. When users create a new place object, they point a target place on the map view and fill some fields for descriptions of this object. These place objects are included with this new entry (Figure 2).

3.3 Personal Map Content

On our system, instead of using map images from open global map services, users can use personal map images as base maps on LBS. Arbitrary uploaded images like hand drawn images, photos and captured facility maps are utilized for personal maps. A content using personal maps, personal map content, includes spatial objects same as global map content on our LBS, they are placed by local X-Y coordinates of a personal map. Additionally all spatial objects on our LBS are allowed to make a link to other spatial object on another global/personal map content. These links lead from a global content to a personal content, or represent connections between a personal content and other personal content. Using these links, users are able to go into personal map and go back to global map quickly. Furthermore, it is utilized to represent connections between facility maps of each floor and some buildings. (Figure 3)

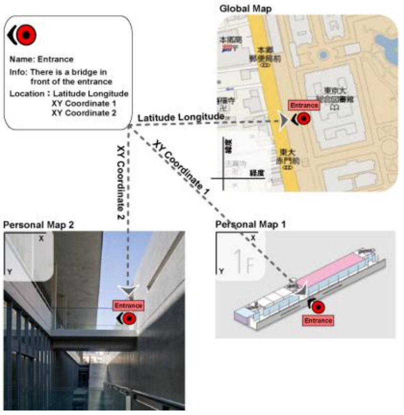


Figure 3. An example of linked POI on several personal maps.

Following two figures (Figure 4, 5) represent sample personal map content. Figure 4, this content uses a station's facility map. The facility map is not north up and it is a little complex to go to the north exit from platforms of super express train. It indicate a route to the north gate, and users can approach to the north gate easily by tracking the route (the red line) on this personal map content, when they arrive at this station by super express train.

Figure 5 represents pedestrian navigation content using sequential photos. To follow links of spatial objects is similar to turning over photos and users look for same landscape in each photo for self-positioning and self-navigation.

In personal map content, users only use self-positioning, but we think if entrance points are easy to recognize, users can adjust their positions by themselves, so they can continue walking in personal map content with our LBS client. Also personal map content does not use latitude-longitude coordinates for positioning, thus users can hide from logs of global map services.

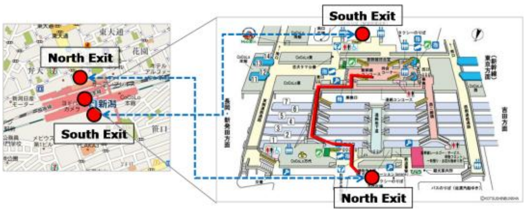


Figure 4. A personal map content using captured facility map that leads travellers from platform of super express to the north gate. (Facility map: (c) KOTSUSHINBUNSHA)

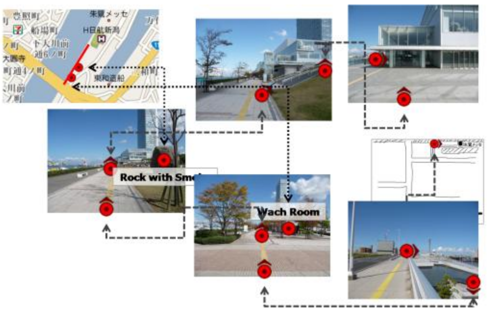


Figure 5. A personal map content by sequential photos. It represents a path from a bridge to a convention hall using sequential photos.

4. CONNECTION BETWEEN WORDS AND PLACE INFORMATION

On our blog, spatial information and a blog entry have a loose relationship normally. Sometimes multiple spatial objects, while a spatial object represents spatial information on a map, can be included in an entry, and blog readers need to find correspondence relationships between a part of the entry text and a spatial object from the context of the text. For setting an explicit relationship between a part of text and a place object and for assigning restriction levels for privacy to a spatial object, we prepared a pair of POI tags, [poi #] and [:poi]. A POI tag is composed of a start tag [poi #] including the number (#) represents a target spatial object, and an end tag [:poi] which consists of a colon and the tag word "poi". Owners can write descriptions with other various tags between the start and end POI tags. When a POI tags are embraced group tags, the target spatial object is assigned a limitation of data access for browsing. Thus, only permitted spatial objects of a reader are displayed on a digital map on a blog site and screen on a LBS client. Using this feature, users can add flexible restrictions of data access control to spatial objects.

On our blog, despite of using angle brackets "< >" for representing tags, it uses square brackets "[ ]". This avoids several security issues, for example, cross site scripting, session hijacking and so on, without any complicated process.

When a user makes mistype of control tags on his/her blog, layout and linking tags, [image] and [poi] don't have serious problem. But if users write wrong tag style on access control tags, group, it has possibilities of serious privacy leak. We think users can avoid it using some inputting and confirming functions. Thus we provide users with tags placing function places tags on head and tail of a selected substring then users need to add attributes to placed tags, and preview function before publishing a new entry on our blog. It will reduce mistypes and problems.

5. PRIVACY SETTINGS ON LAPLCM

5.1 Levels of Group Sharing for Blog Documents in Part

Our place-enhanced blog has sharing group settings same as some normal blogs and SNS. However, those blogs and SNS can set only one sharing level to each record, on our blog, authors can set flexible access restrictions not only whole text of each blog entry but also each word in the blog entry text using these sharing groups. For example, the following texts include access controlled partial texts. A writer can set a restriction level in part of the text for information sharing with all friends. A line headed by "A part" can be shown to friends of "Group A", a line of "B part" can be also accessed by friends of "Group B".

> Today, I went to my office, 30km far from my house, by bicycle.
> *(A part) My office is <u>here</u>*
> It needed 2 hours to reach the office.
> When I returned, I have spent 3 hours on the way because of fatigue and against wind
> *(B part) I'm sorry for late for the dinner party.*

Users on the author’s friend list can show the following text.

> Today, I went to my office, 30km far from my house, by bicycle.
> It needed 2 hours to reach the office.
> When I returned, I have spent 3 hours on the way because of fatigue and against wind

“Goup A” users on the list can show additionally the A part line. So, users of “Goup B” can show the B part line.

We will explain about sharing details and how to set restrictions to each word text. There are four levels and groups for sharing blog entries (Table 1). Sharing levels express of showing of each entry and arbitrary parts of the blog text. Sharing groups are defined as lists of users for sharing. Users can make multiple sharing groups on their accounts.

Table 1. Sharing levels of laPLCm

<table>
<tr><th colspan="5">Sharing group / Sharing level</th></tr>
<tr><td colspan="5">Private</td></tr>
<tr><td>Group1</td><td>Group2</td><td>Group3</td><td>Group4</td><td>…</td></tr>
<tr><td colspan="5">Friend</td></tr>
<tr><td colspan="5"><i>Unlisted</i> / Public</td></tr>
</table>

*Names of sharing level and group are same without Public level

Our proposed system provides enough functions of restrictions for users to access part of text by setting sharing levels. A user can assign all other users to a sharing group. **Group** level is separated into subgroups, and when a user is put in Group level, the user has to be put in a subgroup like “Group1”. Users can make arbitrary subgroups on Group level.

- Unlisted users can only show “Public” level contents.
- Friend level users can access “Public” and “Friend” level contents.
- Group level users can show “Public”, “Friend” and their included sub group level contents. For instance, if U is in Group1, U can show “Public”, “Friend” and “Group1” level contents.
- Private level users can show all content of this author.

When users set restrictions for blog text, first, an Author chooses sharing level of a blog entry. Second, put some pairs of group tag, [group foo] [:group], in the blog text for setting restriction to arbitrary parts of the text. Following text is the sample of usage of group tags.

> (Whole text: Friend)
> Today, I went to my office, 30km far from my house, by bicycle.
> *[group A] My office is here[:group]*
> It needed 2 hours to reach the office.
> When I returned, I have spent 3 hours on the way because of fatigue and against wind
> *[group B] I’m sorry for late for the dinner party.[:group]*

Our blog provide flexible restrictions of blog texts with users using this group tags. A user just make a friend list on the user’s blog account to use this access control.

5.2 Personal Mapping Services

It is an important problem that locations of users are recorded to servers of commercial map providers when web mapping services are used from mobile computing environment with GPS receivers, because the queries to obtain map data around their positions mean that users always inform their positions to web mapping servers. Our proposed personal mapping services prevent from the recording of users' locations as the log of web mapping. The personal mapping services serves as proxy servers of web mapping services in additions to personal map repositories. If the personal mapping servers have cache data of maps of users’ interests which have been fetched from web mapping servers before, maps of their interests are available through the personal mapping services without any accesses to web mapping servers. If the personal mapping servers do not have cache data of maps of their interests to visit in future but their planned routes have already known for users and their navigation scheduler software, map prefetching transactions following the planned routes can be executed to store the cache data of maps of their interests in advance. The prefetching spatial queries can be recorded in web mapping servers, but real-time position tracking is prevented from recording. Also, the LBS clients on mobile devices can use secured private transmission protocol to make spatial queries to and to receiving spatial data from personal mapping servers. For example, cached map data can be transformed by intended differential coordinates as transmission data on the Internet. Furthermore, the spatial data transmissions between the LBS client applications and personal mapping server applications can be asynchronous to hide real-time positions of users.

5.3 Self-positioning

Our developing LBS client applications can cover indoor navigations using indoor map content and self-positioning function. GPS signal is usually not available indoor or is inaccurate. Our proposed self-positioning function allows users to easily set their position with natural computer-human interactions. Network data representing ways of both indoor and outdoor are used for the basis of the positioning. Users’ positions are on the ways. Default speed of users' movements may be three kilo meters per hour. The positions of users are automatically moving on the ways of the displaying map at the constant setting speed which can be easily changed. Also, users can stop, forward, and backward the movement of their positions using user-friendly interactions like a mobile music player. On branch points, users can easily choose their ways by simple selecting operations. The user friendly self-positioning interfaces can be considered tiresome, but many subjects do not feel tiresome in the operation of self-positioning. They felt fun the self-positioning like the operations of computer games. The self-positioning function is important for indoor LBS and navigation from location privacy as well as practical viewpoints for universal navigation systems. If users use the self-positioning, they need not communicate with global online positioning services, including assisted GPS, Wi-Fi Positioning and so no, which can record the real-time positions of users on the servers of global positioning providers.

Navigation on cars and other high speed or imbalance vehicles needs hi accurate positioning and correct information, because drivers need to watch and decide their direction in short time, if their position on navigation systems provide drivers with inaccurate information, it may happen traffic accident. On the other hand, pedestrians have more allowance to check and operate a mobile device to use LBS client, because they can stop walking and step aside any time. This means users can take time to match their position and choose useful information on LBS so we think it suits self-positioning and using unsure content on user generated content.

6. EXPERIMENT IN REAL SITES

We had an experiment to demonstrate the efficiency of our system as a location based service. We tried out the system in a class. Twelve graduate students of the University of Tokyo used our system. Students created town guide content on it, then they experienced these content on each site.

6.1 Creating Personal Spatial Content

Students created spatial content or town guide content as an assignment after a lecture using our system for an hour. The content was a walking guide for an area around Tokyo. Students created some entries including related POIs and lines on the place enhanced blog. POIs represent favourite restaurants, interesting hobby shops, view points and so on. Lines represent some route to walk the area.

Most students created their content in only one night. The target areas and numbers of spatial objects in the content are shown in Table 2. According to time stamps of each entry and access log of our Web server, they did not spend much time for creating their content.

Table 2. Numbers of Entries and Spatial Objects in each of Spatial Content Created by Students

	Route Count	POI Count	Entry Count	Making Time
S1	3	29	3	-
S2	3	27	30	2:40
S3	1	13	14	0:50
S4	3	22	25	4:00
S5	2	20	22	0:40
S6	1	20	21	1:00
S7	0	20	20	3:00
S8	1	7	7	0:30
S9	1	21	22	1:00
S10	2	20	22	1:20
S11	4	20	24	2:00
S12	0	13	13	2:30

*Time is a span from the first entry posted to the last entry posted. S1 did not post entries continuously thus S1's time is not filled.

6.2 Experience on the Sites

Students experienced the spatial content with each other using LBS client application on mobile phones. Current our LBS client application displays multiple points of interests with map images on screen. Users can read description of each point of interest and a text of corresponding entry in a window. If there is one or more dashed poly lines on map images, users can start scrolling maps along a poly line automatically by select and start walking on maps function. To retrieve place information around a user, it uses key word or a latitude and longitude coordinate that it is pointed by center point of the map view or GPS on a mobile phone. Additionally, users set down target users of our blog, target term and radius for searching. On the site, students tracked a route in the content and checked real places of POI described. They can choose arbitrary route and change his/her route to another any time.

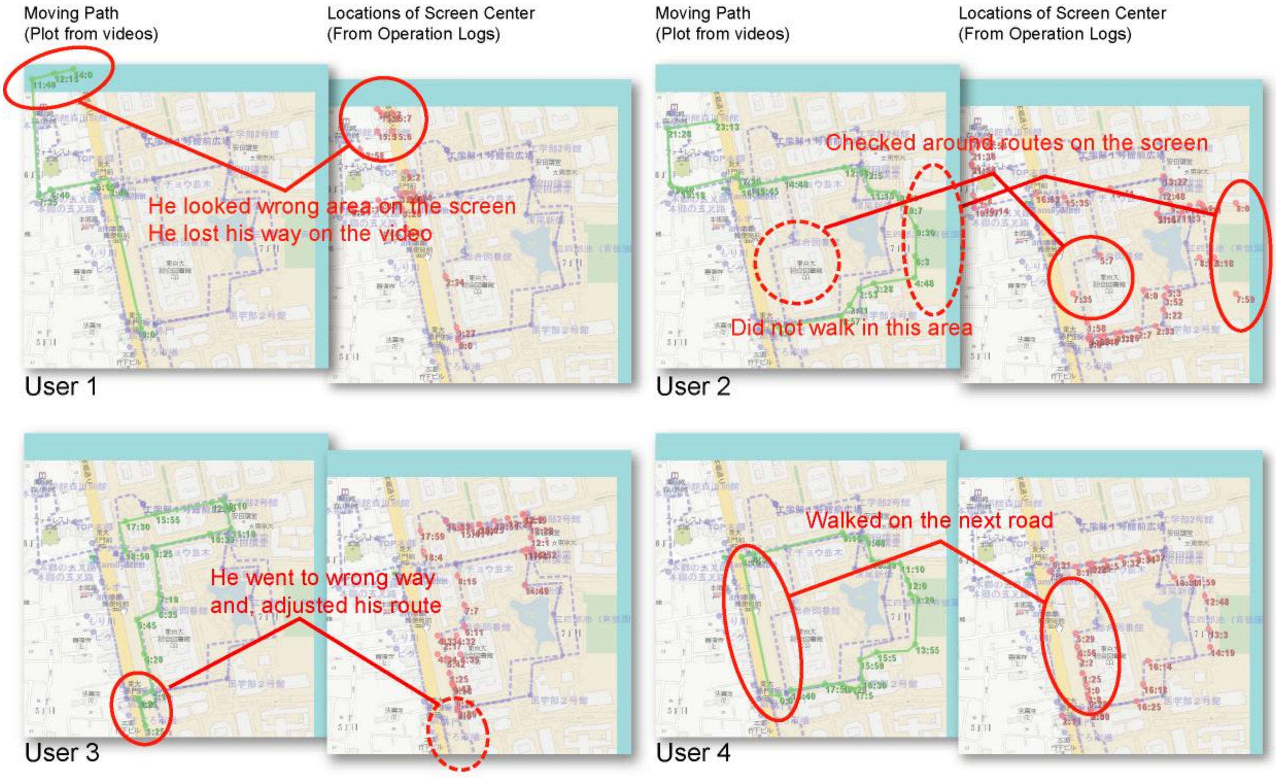


Figure 6. Walked paths and logged points of experiences on the Hongo Campus

Our prototype LBS client does not use GPS functions, thus its users must adjust their current positions on the map displayed the screens of their mobile phones. Some students lost their ways when they experienced the navigations of our LBS. There are some students' feedbacks after the experiment as follows:

- The mobile phone client took time to be getting used to.
- It is interesting to relive the experiences of someone else.
- I was prone to get lost with GPS-less navigation.
- It was easy and useful to create spatial information on the place enhanced blog.

Figure 6 shows some pair of plots of moving paths and plots of locations of screen center on maps that users used a content to guide around Hongo Campus of the University of Tokyo. Usually, these tester students are based on the Kashiwa Campus that is located 30km far from Hongo, and they are not familiar with the Hongo Campus. These maps represent user's matching accuracy and mismatching points in the experiences. On our system, users need to make self-positioning to reduce using GPS thus matching accuracy is an index of usability. Red circles mean characteristic points of user behavior. User2 used the client easily and he could walk in the campus along multiple routes without losing his way, furthermore he checked around routes using map scroll feature. User3, at first, went to wrong direction, but he realized his mistake soon and returned to the correct way. User4 walked on the next road of the target route then he adjusted his position at the junction of these two roads. User1 lost his way and he entered an area that maps of the area were not prepared on our prototype map server. Thus, he could not return to any route.

All students began to check the mobile phone continuously, when they missed their right locations. According to video records of their behavior, we presumed that they were checking their locations and spatial information around them, when they did not operate their mobile phones. Also, according to these points of Figure 6, it was easy for users to match their locations without GPS and other sensors, and it proved that our proposed route based navigation method worked well.

7. CONCLUSION

In this paper, we have proposed the laPLCm framework and reported our implemented prototype system of the laPLCm for using personal life content with reasonable privacy functions. The prototype system provided users with a useful environment to create and use personal life content with spatial information as an extended blog system. Then, we confirmed the effectiveness of our proposed laPLCm. Useful interfaces and easy operations of place enhanced blog allows trial users to create many blog entries as a spatial content on the blog in a short time, after only a simple short lecture for using our system. Flexible and simple privacy setting for personal spatial content is important to treat personal location information and spatial content on LBS. Our proposed location privacy functions may enable users to keep location privacy in their daily lives.

It is not enough that only one commercial LBS contains all personal spatial information. Therefore, it is important to integrate between laPLCm and commercial LBS. For instance, when a user walks around Akihabara downtown in Tokyo, this system retrieves user's to-buy list in personal records and related shops' information on the commercial services, then this information is displayed on his/her mobile device synchronized with places of the users. Pushing users' past blog entries to each user can make them remember past forgotten memories and clarify their present situation from the life-span viewpoint. We have proposed a place enhanced blog for laPLCm in this paper. LBS are generally developed on the commercial telecommunication network services, and are usually not open in technical and use senses. Our proposal of place enhanced blogs is open on the platform of the Internet. Individuals can create and modify their own services by themselves using Web browsers and special software on mobile phones.

Easy manual self-positioning methods of both choosing routes of sidewalk networks and controlling walking mode such as walking, stopping, returning and running are significant to keep the location privacy for positioning methods. GPS and other positioning sensors can be also used with the manual positioning methods, but they are used in the pull style, not in the push style to allow users to be aware of the location of ourselves recorded by global positioning providers.

References from Journals:

Boyd, D., Ellison, N., 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), pp. 210-230.

Nouwt, S., 2008, Reasonable Expectations of Geo-Privacy? *SCRIPTed – A Journal of Law*, 5(2), pp. 375-403.

References from Other Literature:

Nardi, A. B., Schiano, J. D., Gumbrecht, M., Swartz, L. 2004. Why we blog, *Communications of the ACM*, Vol. 47, Issue 12, pp. 41-46.

Dobson, J. E., Fisher, P. F., 2003. Geoslavery, *IEEE Technology and Society Magazine*, Vol. 22, Issue 1, pp. 47-52.

References from websites:

Moons, T., 1997. Report on the Joint ISPRS Commission III/IV Workshop "3D Reconstruction and Modeling of Topographic Objects", Stuttgart, Germany. http://www.radig.informatik.tu-muenchen.de/ISPRS/WG-III4-IV2-Report.html (accessed 28 Sep. 1999)

Facebook Inc., 2009, Facebook, Palo Alto, United States. http://www.facebook.com/ (accessed 08 Dec. 2009)

Google Inc., 2009a, Google Latitude, Mountain View, United States. http://www.google.com/latitude/ (accessed 08 Dec. 2009)

Google Inc., 2009b, Gmail, Mountain View, United States. http://mail.google.com/ (accessed 08 Dec. 2009)

Twitter Inc., 2009, Twitter, San Francisco, United States. http://twitter.com/ (accessed 08 Dec. 2009)

ROUTING WITH MINIMUM NUMBER OF LANDMARKS

Jun Luo[*] and Rong Peng[†] and Chenglin Fan[‡] and Jinxing Hu[§]

Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, China

KEY WORDS: Landmark, algorithms, routing

ABSTRACT:

Routing problem has been studied for decades. In this paper, we focus on one of the routing problems: finding a path from source to destination on road network with the guidance of landmarks. People use landmarks to identify previously visited places and reoriented themselves in the environment. When people give direction instructions for other people, they also like to refer to landmarks. In this sense, we want to find a path such that it visits as many landmarks as possible but also the distance of the path is as short as possible. However, in some situations, the wayfinder may not want to see as many landmarks as possible along the way. For example, the wayfinder drives a car from source to destination. He probably doesn't want to use many landmarks to guide his driving since it's not convenient to switch from one landmark to the other landmark frequently. But he still want to have at least one landmark to be seen at any point along the way. Therefore, the problem becomes: find a path P from s to t such that the driver can see at least one landmark at any point along P and the number of landmarks the driver can stick to is minimized. There are two cases: (a) The same landmark in different road segments counts twice. (b) The same landmark in different road segments counts once. We give the optimal solutions for those two problems by using modified Dijkstra's shortest path algorithm and modified Bellman-Ford algorithm.

1 INTRODUCTION

Routing problem has been studied for decades. The first routing problem could be traced back to 1959 when Dantzig and Ramser proposed vehicle routing problem (Dantzig and Ramser, 1959). Routing problem is very important not only in the field of transportation, but also in the field of logistics, distribution, TCP/IP networks, wireless sensor networks and so on (Golden et al., 2008, Campbell et al., 1997, Campbell et al., 2002, Huitema, 1995, Al-Karaki and Kamal, 2004). The classic routing problem is the shortest path problem that is given the road network and finding the shortest path from source to destination.

In this paper, we focus on one of the routing problems: finding a path from source to destination on road network (Car and Frank, 1993, Gaisbauer and Frank, 2008). A good wayfinding system provides precise and enough indicators of where the wayfinder current location is and how to get to the destination from his/her current location. One of the important indicator for human wayfinding is landmark (Jacob et al., 1999, Lovelace et al., 1999, Peebles et al., 2007, Raubal and Winter, 2002, Elias, 2003, Michon and Denis, 2001).

Landmarks are defined as entities that are salient and easily distinguishable from their surrounding background. People use landmarks to identify previously visited places and reoriented themselves in the environment. When people give direction instructions for other people, they also like to refer to landmarks. In this sense, we want to find a path such that it visits as many landmarks as possible but also the distance of the path is as short as possible. However, in some situations, the wayfinder may not want to see as many landmarks as possible along the way. For example, the wayfinder drives a car from source to destination. He probably doesn't want to use many landmarks to guide his driving since it's not convenient to switch from one landmark to the other landmark frequently. But he still want to have at least one landmark to be seen at any point along the way.

Suppose we already know the road network N with n nodes, m land marks $M_i (i = 1, ..., m)$, the road segments that could be seen by M_i are painted by color $C_i (i = 1, ..., m)$ (see Figure 1). Actually, the road network should be directed since the same landmark could be seen on one direction at one place but couldn't be seen

[*] jun.luo@sub.siat.ac.cn

[†] rong.peng@siat.ac.cn

[‡] cl.fan@sub.siat.ac.cn

[§] jinxing.hu@sub.siat.ac.cn

on the other direction at the same place. However, this will not change the complexity of our algorithms. Therefore we assume the road network is not directed in the remaining part of this paper. Suppose the road network can be segmented. Each segment has unit length. Each color covers either the whole segment or none of the segment. Then we can count how many colors cover each road segment.

Given source and destination s, t in N, the problem becomes:

Problem 1 **MinLandmarks**. *Find a path P from s to t such that the driver can see at least one landmark (or can be associated with at least one color) at any point along P and the number of landmarks (or colors) the driver can stick to is minimized. There are two cases: (a) The same landmark in different road segments counts twice. (b) The same landmark in different road segments counts once.*

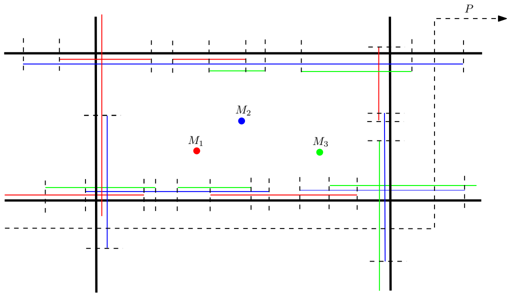


Figure 1: Illustrating the two cases for problem 1. For the first case, the number of times the driver sees the landmark M_1 (red color) along path P is 3 while in the second case, that is 1.

Shortest path problem can be divided into two categories: single-objective shortest path problem (SSPP) and multi-objective shortest path problem (MSPP). The objective of SSPP is only one: minimize the distance from source to destination. Although the objective of **MinLandmarks** problem is not to minimize the distance from source to destination, it can be converted to single-objective shortest path problem. There are abundant algorithms for single-objective shortest path problem from the classic Dijkstra's algorithm to the latest evolutionary algorithm (Cormen et al., 2001, S. Baswana and Neumann, 2009).

2 MINIMIZE THE NUMBER OF LANDMARKS

In this section, we solve the **MinLandmarks** problem defined in section 1. Again, we use the color scheme introduced in section 1 such that each landmark is assigned a color and the road segments that are visible to that landmark are also painted by the color of that visible landmark. Furthermore, we insert *virtual nodes* on two endpoints of each colored road segments. We call the original nodes of the road network N are *real nodes* (see figure 2). We use *nodes* to refer both *virtual nodes* and *real nodes*. If the virtual node is coincide with real node, then the virtual node is deleted. There are two observations of this new graph.

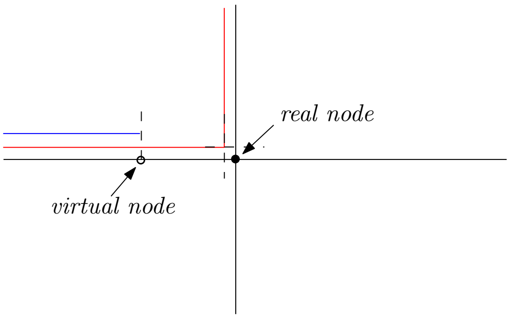


Figure 2: Illustration of *virtual nodes* and *real nodes*.

Observation 1 *The colors cover the edge between* nodes *do not change.*

Observation 2 *Without loss of generality, we assume the endpoints of the different color road segments are not in the same places. Therefore, for a virtual node, there are only two edges adjacent to it and the number of colors covering those two edges is only different by one.*

Since one landmark could be seen by different disconnected road segments, the different road segments with same color could be seen more than once along a certain path. Depending on how we count the same color road segments, it leads to two different objective functions thus leads to two different algorithms.

2.1 The same landmark in different road segments counts twice

In this section, we discuss the scenario that along some path if we see the same landmark on two disconnected road segments, the same landmark is treated as two different landmarks. In other words, we count the landmark or the color twice.

Our algorithm is similar to Dijkstra's single source shortest path problem. The input is the graph $G = (V, E)$,

```
Algorithm MIN-LANDMARK-TWICE($G, w, s$)
INITIALIZE-SINGLE-SOURCE($G, s$)
$S \leftarrow \emptyset$
$Q \leftarrow V[G]$
while $Q \neq \emptyset$
    do $u \leftarrow$ EXTRACT-MIN($Q$)
        $S \leftarrow S \cup \{u\}$
        for each vertex $v \in Adj[u]$
            do RELAX($u, v, w$)
```

```
Algorithm INITIALIZE-SINGLE-SOURCE($G, s$)
$C1[s], C2[s] \leftarrow \emptyset$
$counter[s] \leftarrow 0$
for each node $u \in V \backslash s$
    do $C1[u], C2[u] \leftarrow \emptyset$
        $counter[u] \leftarrow \infty$
```

where G is the road network with virtual nodes, V are the nodes and E are the edges between nodes. However, the weight w of each edge $(u, v) \in E$ is not the distance between u, v. The weight w is actually the list of colors that cover edge (u, v). For each node u, we maintain two sets $C1, C2$ and one integer variable $counter$. $C1, C2$ store the colors that are visible just before node u. $counter$ records the minimum number of colors needed to cover the optimal path from s to u so far.

Actually, $C1[v]$ denotes the latest possible colors people stick to on the optimal path from source to node v, the colors in $C1[v] \cup C2[v]$ are the colors covering the edge (u, v). We should stick to one of the color in $C1[v]$ for the edge (u, v). Choosing which one to stick depends on which color lasts the longest but we can not decide that up to node v unless there is only one color left in $C1[v]$. The colors in $C2[v]$ are the colors might be used as guidance color later and will be moved to $C1[v]$ only after $C1[v]$ is empty. Of course, some road segments may not be covered by any colors. In that case, we may not find a path satisfying our requirement. We consider this case in our algorithm.

The algorithm MIN-LANDMARK-TWICE looks exactly the same as Dijkstra's algorithm. However, the three subroutines are different. The first two are straightforward. We explain the third subroutine RELAX(u, v, w) in detail.

```
Algorithm EXTRACT-MIN($Q$)
output the node $u \in Q$ such that $counter[u]$ is the minimum
```

```
Algorithm RELAX($u, v, w$)
1.  $C1' \leftarrow C1[u] \cap w$
2.  $C2' \leftarrow w - C1'$
3.  if $w = \emptyset$
4.      then exit
5.      else if $C1' = \emptyset$
6.              then $counter' \leftarrow counter[u] + 1$
7.                   $C1' \leftarrow C2'$
8.                   $C2' \leftarrow \emptyset$
9.           if $counter' < counter[v]$
10.             then $counter[v] \leftarrow counter'$
11.                  $C1[v] \leftarrow C1'$
12.                  $C2[v] \leftarrow C2'$
13.          if $counter' = counter[v]$
14.             then $C1[v] \leftarrow C1' \cup C1[v]$
15.                  $C2[v] \leftarrow C2' \cup C2[v] - C1[v]$
```

In line 1, the colors appear in both $C1[u]$ and w are assigned to the temporary color list $C1'$ because we want to stick to the colors in $C1[u]$ as long as possible for traveling edge (u, v). In line 2, the colors appear in w but not in $C1[u]$ are put into another temporary color list $C2'$ because currently we don't need to stick to those colors but we probably need them later. Line 3 to 6 deal with the situation if $C1'$ is empty. That means all colors in $C1[u]$ disappear in w. We have to switch to the colors in w and the counter needs to increase by one. Line 7 to 10 just substitute $counter[v], C1[v], C2[v]$ with the new values $counter', C1', C2'$ if the new counter is smaller than the old one. If the new counter is the same as the old one, line 11 to 13 just merge $C1[v], C2[v]$ with the new values $C1', C2'$ and take off the colors appearing in updated $C1[v]$ from updated $C2[v]$. The correctness of line 7 to 13 is given as follows:

Lemma 1 *For a path passing through nodes v_1, v_2, v_3 consecutively, $counter[v_3]$ is larger than $counter[v_2]$ only by one at most, that means $counter[v_2] \leq counter[v_3] \leq counter[v_2] + 1$.*

Proof. Suppose the color lists of edge (v_1, v_2) and (v_2, v_3) are w_1 and w_2, if w_1 and w_2 are the same, which could happen when v_2 is a real node, then the color we stick to on w_1 can be still used on (v_2, v_3), thus $counter[v_3] = counter[v_2]$. If w_1 and w_2 are different, then if the color we stick to in w_1 is still appear in w_2, thus we can continue to use that color, that means $counter[v_3] = counter[v_2]$. Otherwise we just stick to a new color in w_2, that means $counter[v_3] = counter[v_2] + 1$. Thus the lemma follows. ∎

Lemma 2 *For a node v, we only need to keep the smallest $counter[v]$ and corresponding $C1$ and $C2$.*

Proof. If v is a virtual node, from observation 2, we know there are only one incoming edge of v, then there is only one value of $counter[v]$.

If v is a real node, there could be multiple incoming edges. Without loss of generality, we assume there are two incoming edges, (u_1, v) and (u_2, v) and one outgoing edge (v, x). Let the color lists for (u_1, v), (u_2, v),(v, x) be w_1, w_2, w_3 respectively and $counter[v]_1, counter[v]_2$, $counter[x]_1, counter[x]_2$ be counter numbers from u_1, u_2 to v and then to x respectively. From lemma 1, we know $counter[v]_1 \leq counter[x]_1 \leq counter[v]_1 + 1$ and $counter[v]_2 \leq counter[x]_2 \leq counter[v]_2 + 1$. There are three cases:

- $counter[v]_1$ and $counter[v]_2$ are different by one such that $counter[v]_2 = counter[v]_1+1$(if $counter$ $[v]_1 = counter[v]_2 + 1$, the proof is symmetric). In worst case, $counter[x]_1 = counter[x]_2$ that means the colors we could stick to on edge (u_1, v) (which are actually the colors in $C1[v]$) are disappear and we have to switch to the colors on edge (v, x). According to the algorithm RELAX, $C1[x]_1 = w_3 \supseteq C1[x]_2$. Therefore, $C1$ will be empty earlier if the path goes through u_2. Because the color counter increases only when $C1$ is empty, the path going through u_1 is better than the path going through u_2.
- $counter[v]_1$ and $counter[v]_2$ are different more than one. Suppose $counter[v]_2 \geq counter[v]_1+2$. According to the proof of above case, the counter for the path going through u_1 will never be larger than the counter for the path going through u_2. Thus we only need keep smaller one.
- $counter[v]_1 = counter[v]_2$. According to the algorithm RELAX, $C1[v]_1$ and $C1[v]_2$ are merged into $C1[v]$. Suppose if we keep $C1[v]_1$ and $C1[v]_2$ separately and $C1[v]_1$ becomes empty first, then $C1[v]$ becomes empty when $C1[v]_2$ becomes empty. That means we implicitly follow the optimal path going through $C1[v]_2$.

∎

The proof of the correctness of our algorithm is the same as that of Dijkstra's algorithm except we use $counter[v]$ instead of $d[v]$. The running time of our algorithm is also similar to that of Dijkstra's algorithm except in relax step, the transactions of intersection and union of two lists $C1$ and $C2$ take extra $O(m)$ time where m is the number of landmarks. So the total running time is $O(|V|^2 + |E|m) = O(n^2m)$.

Theorem 1 *For the road network with n vertices and m landmarks, we can find an optimal path from s to t in $O(n^2m)$ time for the* **MinLandmarks** *problem if the same landmark counts twice when it is seen twice at two disconnected road segments.*

2.2 The same landmark in different road segments counts once

For this problem we use different data structures. Original input is the road network graph $G = (V, E)$ with each edge covered by different color road segments. We augment G to $G' = (V, E')$ as follows: for each edge $(u, v) \in E$, we compute all the combinations of colors that could cover the whole edge. Each combination is represented by an edge from u to v (see figure 3). Then each edge (u, v) in G could be augmented to many edges in G' and each edge in G' is associated with a color list $C[u, v]_i$ where $1 \leq i \leq k$ and k is the number of combinations of colors that could cover the whole edge (u, v). Let $|C[u, v]_i|$ denote the number of colors in the list and $(u, v)_i$ denote the ith augmented edge of (u, v).

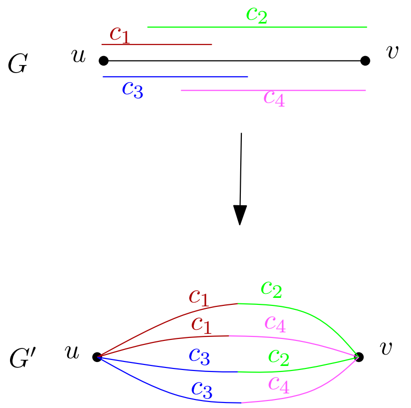


Figure 3: Augmentation from G to G'.

This problem is more complicated than counting twice case and we can not use Dijkstra's algorithm. This is because the optimal path from s to t passing through node v may not consist of the optimal path from s to v and the optimal path from v to t. For example, in figure 4, the optimal path from s to t consists of P_1 and P_3 which are covered by three colors c_1, c_2, c_3. While the optimal path from s to v is P_2 which is covered by two colors c_4, c_5 and the optimal path from v to t is P_4 which is covered by two colors c_6, c_7. To solve this problem, it seems that we have to record all the color lists for all possible paths form s to v. Fortunately, we don't need to do that. We can use the Bellman-Ford algorithm with different data structure. Let $C_i[v]$ be the ith color list for node v that could cover some paths from s to v.

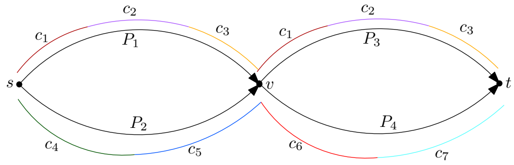


Figure 4: The example shows that the optimal path from s to t passing through node v may not consist of the optimal path from s to v and the optimal path from v to t.

Lemma 3 *If the optimal path from s to v is $p =< v_0, v_1, ..., v_k >$ where $v_0 = s$ and $v_k = v$, the color list corresponding p is $C_{opt}[v]$, then after kth passes over the edges of G' in* MIN-LANDMARK-ONCE, *$C_{opt}[v]$ is one of the color lists of v.*

Proof. Actually, $C_{opt}[v] = C[v_0, v_1]_{i_1} \cup C[v_1, v_2]_{i_2} \cup ... \cup C[v_{k-1}, v_k]_{i_k}$ where $C[v_{j-1}, v_j]_{i_j}$ is the color list for the i_jth augmented edge of edge (v_{j-1}, v_j). After first pass over the edges of G', we can get the list $C[v_0, v_1]_{i_1}$ for v_1 and after second pass over the edges of G', we can get the list $C[v_0, v_1]_{i_1} \cup C[v_1, v_2]_{i_2}$ for v_2, and so on. We can get $C[v_0, v_1]_{i_1} \cup C[v_1, v_2]_{i_2} \cup ... \cup C[v_{k-1}, v_k]_{i_k}$ for v_k. Thus the lemma follows. ∎

After the algorithm MIN-LANDMARK-ONCE, we can get color lists $C_1[v], C_2[v], ..., C_k[v]$ for each node v. According to lemma 3, we know that $C_i[v]$ is $C_{opt}[v]$ if the number of colors of $C_i[v]$ is the smallest among all color lists of v. However, we can not report the optimal path from s to v since we does not provide any backtrack scheme in the algorithm and data structure. Actually, we only need to add a pointer for each color list $C_i[v]$ of v. This pointer points to the color list $C_j[v']$ that generates $C_i[v]$. To report the optimal path from s to t, we first find $C_{opt}[t]$ and get the pointer for $C_{opt}[t]$. Suppose the pointer points to $C_i[v]$, then the predecessor of t is v and we backtrack the path from $C_i[v]$ recursively until we reach $C[s]$.

```
Algorithm MIN-LANDMARK-ONCE ($G'$,$C[u,v]$,$s$)
    INITIALIZE-SINGLE-SOURCE($G', s$)
    for $i \leftarrow 1$ to $|V(G')| - 1$
        do for each edge $(u, v) \in E(G')$
            do RELAX($u, v, C[u, v]$)
```

```
Algorithm INITIALIZE-SINGLE-SOURCE($G, s$)
    for each node $u \in V$
        do $C[u] \leftarrow \emptyset$
```

```
Algorithm RELAX($u, v, C[u, v]$)
    for each color list $C_{[}u]$ of $u$
        do $C[v] = C[u] \cup C[u, v]$
```

The running time of this algorithms is $O(|V(G)| \cdot |E(G')| \cdot l \cdot m) = O(n^3 lm)$ since each edge relaxation needs $O(l)$ times of two color lists union operations and each union takes $O(m)$ time where l is the maximum number of color lists for one node. The running space is $O(|V(G)| \cdot l \cdot m) = O(nlm)$ since each color list needs $O(m)$ space.

Theorem 2 *For the road network with n vertices and m landmarks, we can find an optimal path from s to t in $O(n^3 lm)$ time and $O(nlm)$ space for the* $\mathbf{MinLandmarks}$ *problem if the same landmark counts once when it is seen twice at two disconnected road segments, where l is the maximum number of color lists for one node.*

3 DISCUSSION

In this paper we presented $O(n^2 m)$ time algorithm for the $\mathbf{MinLandmarks}$ problem when the same landmark counts twice and $O(n^3 lm)$ time and $O(nlm)$ space algorithm when the same landmark counts once where n is the number of vertices of road network and m is the number of landmarks and l is the maximum number of color lists for one node. For the latter case, we know that the running time and space of the optimal algorithm are all related to l. In worst case, $l = O(C_1^m + C_2^m + ... + C_m^m) = O(m^m)$. If m is constant, that optimal algorithm is a polynomial time and space algorithm. Otherwise, that is exponential time and space algorithm which is unacceptable. Thus it's worth investigating whether there exists an approximation algorithm with polynomial running time and space of n, m.

The other interesting open problem is to finding a path such that it visits as many landmarks as possible but also the distance of the path is as short as possible. This problem is much harder than $\mathbf{MinLandmarks}$ problem and actually is a multiobjective shortest path problem which has been proved to be NP-complete. Therefore proposing an approximation algorithm for this problem is also a challenge.

REFERENCES

Al-Karaki, J. N. and Kamal, A. E., 2004. Routing techniques in wireless sensor networks: a survey. IEEE Wireless Communications 11(6), pp. 6–28.

Campbell, A., Clarke, L. and Savelsbergh, M., 2002. Inventory routing in practice. In: D. V. P. Toth (ed.), The Vehicle Routing Problem, SIAM monographs on discrete mathematics and applications, pp. 309–330.

Campbell, A., Clarke, L., Kleywegt, A. and Savelsbergh, M., 1997. The inventory routing problem. In: T. Crainic and G. Laporte (eds), Fleet Management and Logistics, Kluwer Academic Publishers, pp. 95–113.

Car, A. and Frank, A. U., 1993. Hierarchical street networks as a conceptual model for efficient way finding. In: Proceedings of the EGIS'93, pp. 134–13913.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C., 2001. Introduction to Algorithms, Second Edition. The MIT Press.

Dantzig, G. and Ramser, J., 1959. The truck dispatching problem. Management Science 6(1), pp. 80–91.

Elias, B., 2003. Extracting landmarks with data mining methods. In: COSIT 2003: Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, pp. 375–389.

Gaisbauer, C. and Frank, A. U., 2008. Wayfinding model for pedestrian navigation. In: Proceedings of the 11th AGILE International Conference on Geographic Information Science 2008, pp. 59–68.

Golden, B., Raghavan, S. and Wasil, E., 2008. The Vehicle Routing Problem: Latest Advances and New Challenges. Springer.

Huitema, C., 1995. Routing in the Internet. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jacob, R., Marathe, M. and Nagel, K., 1999. A computational study of routing algorithms for realistic transportation networks. J. Exp. Algorithmics 4, pp. 6.

Lovelace, K. L., Hegarty, M. and Montello, D. R., 1999. Elements of good route directions in familiar and unfamiliar environments. In: COSIT '99: Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, Springer-Verlag, London, UK, pp. 65–82.

Michon, P.-E. and Denis, M., 2001. When and why are visual landmarks used in giving directions? In: COSIT 2001: Proceedings of the International Conference on Spatial Information Theory, Springer-Verlag, London, UK, pp. 292–305.

Peebles, D., Davies, C. and Mora, R., 2007. Effects of geometry, landmarks and orientation strategies in the drop-off orientation task. In: COSIT 2007: Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, pp. 390–405.

Raubal, M. and Winter, S., 2002. Enriching wayfinding instructions with local landmarks. In: GIScience '02: Proceedings of the Second International Conference on Geographic Information Science, Springer-Verlag, London, UK, pp. 243–259.

S. Baswana, S. Biswas, D. D. T. F. P. K. and Neumann, F., 2009. Computing single source shortest paths using single-objective fitness functions. In: Proceedings of ACM International Workshop on Foundations of Genetic Algorithms (FOGA 2009).

INDIVIDUAL DIFFERENCES IN THE TOURIST WAYFINDING DECISION MAKING PROCESS

M. F. Abdul Khanan [a, *], J. Xia [b]

[a] Faculty of Geoinformation Science and Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai,
Johor, Malaysia – m.abdulkhanan@postgrad.curtin.edu.au
[b] Department of Spatial Science, Curtin University of Technology, GPO Box U1987, Perth,
Western Australia 6845, Australia – c.xia@curtin.edu.au

KEY WORDS: Individual differences, Tourism, Wayfinding, Decision making, Physical movement, GIS, GPS

ABSTRACT:

Wayfinding is an important aspect that should be considered by tourist park managers when allocating resources and facilities to aid tourists navigating their way through a park. This paper discusses the influence of individual differences such as age, gender, travel group, and familiarity with the environment towards the wayfinding decision making process and physical movement. A case study was conducted where respondents' movements were traced using GPS receivers to look at the movement patterns and at the end, they were asked to complete a questionnaire particularly to determine the individual differences. The spatial and attribute data were analysed using ESRI ArcGIS Tracking Analyst and ET Geowizards. SPSS 17 were used for statistical purposes. As a result, in terms of decision making, specific landmarks and strategies were noticed for specified age, gender, travel group, and familiarity with the environment. Furthermore, correlations were found between gender and familiarity with the environment with the physical movement such as direction, distance, and arrival time. This paper highlights the need for tourist managers to understand that tourists use different methods of wayfinding and produce different results based on different individual differences and that management should provide complementary materials to assist in wayfinding.

1. INTRODUCTION

1.1 Tourism and wayfinding

Tourism management is a critical issue. The major challenges in regards to tourism is the diversity of users competing for the same resource, and the need to balance these multiple objectives while maintaining a positive tourism experience (O' Connor, Zerger, and Itami 2005). The key to overcome the issue, according to them is by understanding tourist behaviour. One of the sub-areas in the discussion would be on tourism navigation and wayfinding (Walmsley and Jenkins 1992).

Wayfinding is a cognitive psychological process for finding a pathway from an origin to a specified destination (Xia 2008). It is a complex process and will be different for individuals depending on the purpose of the trip or in response to external environmental conditions (Golledge 1999).

1.2 Decision making and physical movement of wayfinding

This paper defines the wayfinding process into two areas which are decision making and physical movement. Decision making will influence physical movement during wayfinding. People referred to two items in terms of decision making during wayfinding. They are landmarks and wayfinding strategies. According to Sorrows and Hirtle (1999), a landmark is a distinct object that people referred to help memorise and distinguish routes, and locate themselves in terms of their destination. Examples of landmarks is signboard.

People are different in the strategies they use when navigating through an environment, from noting landmarks to using a map or spatial layout of the environment (Passini 1984). Generally, in terms of spatial knowledge, there are two strategies commonly used which are egocentric and allocentric wayfinding strategies (Gramann et al. 2005). Wayfinding also deals with physical movement. Physical movement concerns the location and arrival time during wayfinding (Xia 2007). It also can be elaborated into direction and speed, duration, and the mode of movement (Xia 2008).

1.3 Individual differences

Individual differences are the ways in which people differ in their behaviour. Individual differences could directly or indirectly influence wayfinding (Xia 2008). Psychological research has shown that individual differences exist for spatial task performance in the laboratory (Malinowski and Gillespie 2001). The primary aim of this research is to ascertain if tourist wayfinding behaviours correlate with individual differences.

2. ARE THERE DIFFERENCES IN THE WAY WE WAYFIND?

2.1 Age

Previous research has shown that older adults do not perform as well as younger adults on a variety of spatial tasks, including those requiring information about specific environmental layout (Kirasic 2000) and forming cognitive maps or surrounding environment (Iaria et al. 2009).

2.2 Gender

Gender differences can influence wayfinding performances. A tenacious stereotype is that males are more efficient (Chebat, Gelinas-Chebat and Therrien 2008). Males have better knowledge of geographical maps and draw better maps (Harrell, Bowlby, and Hall-Hoffarth 2000), which is usually attributed to

* Corresponding author.

the fact that men are more socialised with maps (Lawton 1994). Some researchers have found that men are more efficient at finding destinations (Malinowski 2001).

Another implication of gender differences is that women and men may differ in strategies for finding a destination where females are more likely to adopt the egocentric strategy and the males are more likely to adopt the allocentric strategy (Chen, Chang and Chang 2008). Another possible implication of gender differences is that women and men may differ in the way they feel about performing tasks that appear to require a sense of direction (Lawton and Kallai 2002). Women show a higher level of anxiety than men such as trying a new shortcut without the aid of a map or figuring out which way to turn when emerging from a parking garage (Lawton 1994).

2.3 Travel group

Individual differences between types of travel groups may be an important factor in wayfinding behaviours and strategies because of the various influences each member may have on decisions (Xia, Packer, and Dong 2009). The difference between various travel groups can be observed in the usage of landmarks. Individuals are less likely than other travel groups to navigate using signposts, while couples are more likely than other types of groups to do so (Xia, Packer, and Dong 2009).

2.4 Familiarity with the environment

Familiarity with the environment does influence strategy choice in directed wayfinding tasks (Holscher et al. 2007). Xia et al. (2008) has found that the type of landmark used was related to the familiarity that tourists have with the site. Linear landmarks such as pathways were used more often by those tourists that are either totally familiar or have never visited the site (Xia 2008). In another research by Xia, Packer, and Dong (2009), it has been discovered that the more tourists that are familiar with the environment, the less chance they use landmarks.

3. METHODOLOGY

3.1 Theoretical framework

The methodology has looked into the relationship between individual differences with wayfinding. Four individual differences will be involved which are age, gender, travel group, and familiarity with the environment. Wayfinding is divided into two components which are decision making and physical movement (Xia 2007). In making decisions during wayfinding, people depend on strategies and landmarks.

There are three components in the physical movement of wayfinding which are spatial, temporal and spatio-temporal. Spatial involves location, temporal involves time, while spatio-temporal involves both location and time (Xia 2007). Spatial elements include direction, location, and distance. Temporal elements include arrival time and duration while spatio-temporal element only includes speed (Xia 2008).

3.2 Case study area

The Koala Conservation Centre (KCC) is centrally located on Phillip Island. It was established in 1991 to protect koalas from cars and dogs and provide close viewing opportunities for tourists. The KCC is composed of six hectares of enclosed woodland (Woodland Bush), a half hectare koala viewing area that includes two boardwalks, a nine hectare plantation and visitor centre (Xia et. al. 2008). A further seven hectares is available for expansion of the woodland habitat (Reed 2000). The KCC features a treetop boardwalk so tourists can see the koalas at close range. They can also walk through eucalyptus bush to see more koalas. There are on average 120,000 tourists who visit the KCC each year (Hallahan and Bomford 2005). Figure 1 shows the map of KCC.

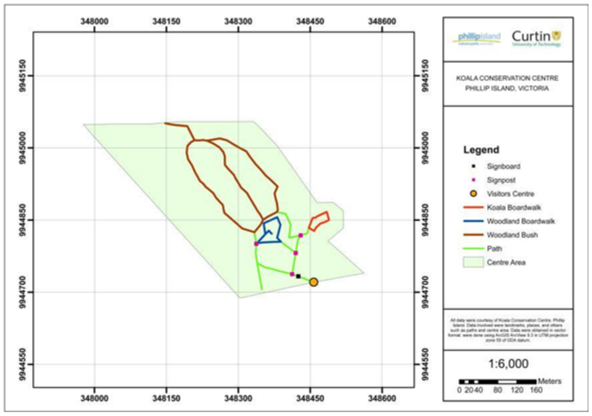

348000
348150
348300
348450
348600
9945150
9945000
9944850
9944700
9944550
phillipisland
Curtin
University of Technology
KOALA CONSERVATION CENTRE
PHILLIP ISLAND, VICTORIA
Legend
Signboard
Signpost
Visitors Centre
Koala Boardwalk
Woodland Boardwalk
Woodland Bush
Path
Centre Area
All data were courtesy of Koala Conservation Centre, Phillip Island. Data involved were landmarks, places, and others such as paths and centre area. Data were obtained in vector format, were done using ArcGIS ArcView 9.3 in UTM projection zone 55 of GDA datum.
1:6,000
Meters
0 20 40 80 120 160

Figure 1. Koala Conservation Centre map

3.3 Data gathering

The researcher has used secondary data obtained from previous research by Xia et al. (2008). A case study was conducted by Xia at KCC from January 17 to 20 in 2005. For the case study, 124 tourists, six group tour guides and two rangers were questioned using a random intercept method, given GPS receivers and tracked using a method similar to that adopted by Arrowsmith, Zanon, and Chhetri (2005). The participants were then interviewed after their visit. GPS surveys enabled spatio-temporal movements to be ascertained whilst interviews and questionnaires provided demographic data and wayfinding methods employed by the participants (Xia 2008).

3.3.1 Individual differences classifications

Participants were interviewed using questionnaires. They were asked to determine their individual differences classifications. The researcher has classified all four individual differences based on the details below:

1. Age – 18 to 34 as young, 35 to 54 as middle and 55 or above as old.
2. Gender – male and female.
3. Travel group – individuals, couples, relatives and friends and groups.
4. Familiarity with the environment – very familiar, familiar but not sure, first visit and follows interest and first visit and follows landmarks.

3.3.2 Landmarks and wayfinding strategies classifications

Besides determining their individual differences, participants were also asked to verify their landmarks and wayfinding strategies used based on classifications below:

1. Landmarks – signboard, signpost, track surface, vegetation type, follow the crowd and avoid the crowd.

2. Wayfinding strategies – shortest path, least time, fewest turns, scenic, first noticed, different from previous.

3.4 Physical movement aspects that was analysed

Based on Figure 1, the researcher has selected six sub-components of wayfinding physical movement to be analysed. The sub-components and their aspects that were analysed are like below:

1. Direction – to determine whether participant's direction was clockwise/anti-clockwise (Asakura and Iryo 2007)
2. Distance, duration, and speed – during the whole wayfinding task.
3. Arrival time – how long does it take to reach each destination.
4. Location – Show route (edge) usage frequency based on gender and age.

3.5 ESRI Tracking Analyst and ET GeoWizards

The main intention of using ESRI ArcGIS 9.3 for this research is to look into the movement patterns of each respondent. This is done through the Tracking Analyst. Using Tracking Analyst, the point data can be manipulated to show the movement of participants throughout the time of carrying the GPS receiver. This is called spatio-temporal analysis in GIS where different time (temporal) can determine different locations for the same spatial entity. The outcome will be various spatial patterns indicating the participant's movement. Tracking Analyst was very useful in determining the direction, location and arrival time sub-components of wayfinding.

Another important role of ESRI ArcGIS 9.3 is to derive the distance used by each respondent during their wayfinding task. This is done through an add-on called ET GeoWizards which has the capability of deriving the distance of the route used by respondent. After obtaining both distance and duration for each respondent, the researcher is able to derive the speed by dividing the distance in kilometre by time.

3.6 Data Analysis

At first place, the researcher is using Odds-ratio test to seek relations between individual differences with wayfinding behaviours. Odds-ratio test is a method to compare whether the probability of a certain event is the same for two groups, in other words, how strong the difference is. For example, this method can be used to compare the probability of females and males using track surface in the KCC. An Odds-ratio greater than one implies that the behaviour is more likely to happen in the first group. An Odds-ratio less than one implies that the event is less likely to happen in the first group. The weakness of this method is that, it only can compare binary data such as male and female, old and young, and Republican or Democrat.

In order to overcome the weakness, the researcher has chosen other statistical tests such as below:

1. Chi-square test for independence as this research goal is to compare two unpaired groups and the type of data is binomial.
2. Independent-measures ANOVA as this research goal is to compare three or more unmatched group and the type of data is interval.
3. Independent-measures t Test as this research goal is to compare two unpaired groups and the type of data is interval.

4. RESULTS

4.1 Landmarks and strategies based on age

Table 1 indicates that there was a significant age difference in the usage of fewest turns as a wayfinding strategy. Old age respondents were more likely to use the first noticed strategy as a wayfinding strategy with four or 14 % of the respondents who were using it. In terms of landmarks, there were two landmarks that produced significant differences with age.

The two landmarks were the signposts with a p-value of 0.061 and vegetation types with a p-value of 0.014. The usage of signpost was closely related to middle age respondents with 41 or 82 % of the respondents using it. Old age respondents were found to utilise vegetation types with 18 % or five respondents who were using it compared to other age respondents.

Table 1. Age differences in wayfinding strategies and landmarks usage

Strategies or Landmarks	Chi-square		
	Value	df	Asymp. Sig. (2-sided)
First noticed (Old)	5.651	2	0.059
Signpost (Middle)	5.607	2	0.061
Vegetation types (Old)	8.533	2	0.014

4.2 Landmarks and strategies based on gender

Based on Table 2, it can be noticed that female respondents were more likely to use first noticed as a wayfinding strategy with an odds of 2.045. In terms of landmarks used, a landmark has proven to produce a significant difference which was signboard with a p-value of 0.054. The odds of 2.198 has shown that, females were more than twice higher than male's odds.

Table 2. Gender differences in wayfinding strategies and landmarks usage

Strategies or Landmarks	Chi-square		
	Value	df	Asymp. Sig. (2-sided)/ Odds-Ratio
First noticed (Female)	3.201	1	0.074/2.045
Signboard (Female)	3.7	1	0.054/2.198

4.3 Landmarks and strategies based on travel group

Table 3 shows that only one wayfinding strategy and three landmarks have recorded significant differences which were shortest path as the strategy and signboard, signpost, and vegetation types as the landmarks. P-value for shortest path was 0.05, signboard was 0.035, signpost was 0.008, and vegetation types was 0.062. Groups were more likely to use shortest path strategy with 17 % or one respondent who was using it. In terms of signboard and signpost usage, it is noticed that, couples were more likely to utilise it during wayfinding task with 94 % or 31 respondents who have been using signboard and 79 % or 26 respondents who have been using signpost. Vegetation types were more likely to be used by groups where 10 % or one of them who was using it.

Table 3. Travel group differences in wayfinding strategies and landmarks usage

Strategies or Landmarks	Chi-square		
	Value	df	Asymp. Sig. (2-sided)
Shortest path (Groups)	12.607	3	0.05
Signboard (Couples)	13.56	3	0.035
Signpost (Couples)	17.379	3	0.008
Vegetation types (Groups)	12.011	3	0.062

4.4 Landmarks and strategies based on familiarity with the environment

Table 4 indicates that there were significant differences produced in terms of landmarks and there were multiple. Firstly, there was a significant difference in the usage of signboard and signpost in terms of familiarity with the environment. It can be noticed that first visitors that follow interest were more likely to use signboard and signpost as landmarks with 82 % or 81 respondents who were using signboard and 79 % or 78 respondents who were using signpost.

There were also significant differences between the usage of track surface and vegetation types in terms of familiarity with the environment. It has been found that first visitors who follow landmarks were more likely to utilise track surface where 64 % or seven of them who were using it. The same respondents have also been found to utilise vegetation types with 27% or three respondents who were using it. Not any of the familiar visitors has been found to generally utilise any landmarks.

Table 4. Familiarity with the environmnet differences in wayfinding strategies and landmarks usage

Strategies or Landmarks	Chi-square		
	Value	df	Asymp. Sig. (2-sided)
Signboard (First visitors, interest)	11.725	3	0.008
Signpost (First visitors, interest)	11.235	3	0.011
Track Surface (First visitors, landmarks)	7.316	3	0.062
Vegetation types (First visitors, landmarks)	10.811	3	0.013

4.5 Direction used based on gender

There was a significant difference with a p-value of 0.066 between genders in the direction chosen during wayfinding. Based on Table 5, it can be noticed that the majority of male respondents used anti-clockwise direction with 19 respondents or 54 % using it. For female respondents, 65 % or 39 of them have chosen to use a clockwise direction during the wayfinding process. Thus, males were more likely to use an anti-clockwise direction while females were more likely to use clockwise.

Table 5. Gender differences in the direction used

Direction	Chi-square		
	Value	df	Asymp. Sig. (2-sided)
Clockwise (Female)	3.373	1	0.066
Anti-clockwise (Male)			

4.6 Distance used based on gender

There was a clear significant difference with a p-value of 0.004 between genders in distances used during wayfinding. Based on Table 6, the mean distance of male respondents was 1.40640 km which was below the general mean distance whilst for females, the mean distance was 1.71572 km which was above the general mean distance. It can be noticed that the mean distance taken by females was longer than males.

Table 6. Gender differences in the distance used

Distance (km)	Chi-square		
	Value	df	Asymp. Sig. (2-sided)
1.40640 (Male)	3.373	1	0.066
1.71572 (Female)			

4.7 Arrival Time Based on Gender

Based on Table 7, two significant differences exist in terms of gender differences in arrival time involving arrival time at Woodland Boardwalk and Woodland Bush. For Woodland Boardwalk, the p-value was 0.075 while for Woodland Bush the p-value was 0.063. For both destinations, the mean arrival time for males was around 10 minutes later than female.

Table 7. Gender differences in the arrival time

Destination	Mean Arrival Time (minutes)	T-Test		
		Value	df	Sig. (2-tailed)
Woodland Boardwalk	38.55 (Male)	1.842	33	0.075
	29.73 (Female)			
Woodland Bush	48.18 (Male)	1.926	33	0.063
	38.55 (Female)			

4.8 Arrival time based on familiarity with the environment

Based on Table 8, there was a significant difference with a p-value of 0.092 between arrival time at Woodland Bush and familiarity with the environment. First time visitors took less time in finding their way to the destination when compared to familiar respondents. First time visitors who followed landmarks tended to be the earliest respondents arriving at Woodland Bush with a mean arrival time of 33 minutes.

Table 8. Familiarity with the environment differences in the arrival time

Mean Arrival Time (minutes)	T-test		
	Value	df	Sig. (2-tailed)
64.5 (Very familiar)	2.355	33	0.092
50 (Familiar, not sure)			
40.21 (First visit, interests)			
33 (First visit, landmarks)			

4.9 Route usage based on gender

Based on Table 9, route AD or DA had the highest frequency of usage of 96 movements compared to the others. Table 9 shows that female used these routes the most with 50 respondents. Route AC or CA have the lowest usage of 28. Males used these routes the most with 28 respondents. We could also observe that all routes to and from the Koala Boardwalk such as DE or ED and CD or DC were dominated by female.

Table 9. Gender differences in route usage frequency

Route	Usage Frequency	
	Male	Female
AB or BA	30	29
AC or CA	20	8
AD or DA	46	50
BC or CB	37	45
BE or EB	47	48
CD or DC	34	42
CE or EC	25	21
DE or ED	20	21

4.10 Route usage based on age

Based on Table 10, young age respondents dominated route AD or DA with 44 respondents followed closely by middle age with 42 respondents. Although route AD and DA were dominated by young age, all routes toward and from Koala Boardwalk such as CD or DC and DE or ED were dominated by middle age.

Table 10. Age differences in route usage frequency

Route	Usage Frequency		
	Young	Middle	Old
AB or BA	24	30	5
AC or CA	8	17	3
AD or DA	44	42	10
BC or CB	34	37	11
BE or EB	39	50	6
CD or DC	22	47	7
CE or EC	23	20	3
DE or ED	15	21	5

5. DISCUSSION

One of the significant outcomes of this research was the prominence of gender differences in relation to five of the tourist behaviours. Females were more likely to use signboards as a landmark. This shows a similarity with Xia, Packer, and Dong's finding that suggests females are more eager to use a signboard (Xia, Packer, and Dong 2009). There was also a significant difference between gender and first noticed. Females were more likely to use that strategy. Again, this was the same with Xia, Packer, and Dong's finding (2009) that suggest females prefer utilising first noticed in their wayfinding task.

In terms of direction, females were more likely to use the clockwise direction while males were more likely to use the anti-clockwise. In terms of distance, males' average distance was 1.40640 km while females' average distance was 1.71572 km. This corresponds with the mainstream findings that suggest males are better than females in wayfinding, thus distance taken by males should be shorter (Chebat, Gelinas-Chebat, and Therrien 2008). However, this can be argued that tourist wayfinding can be categorised as leisure activity. Hence, the longer the distance is, the more tourists want to enjoy the tour by visiting more places, producing larger distance. This is called hedonistic values (Babin, Darden, and Griffin 1994). Hedonistic values has also influenced the time taken to arrive at Woodland Boardwalk and Woodland Bush where time taken by males were generally 10 minutes longer than females.

Another important finding was with the familiarity with the environment differences where it has shown relations with two of the tourist wayfinding behaviours. Firstly, in terms of landmarks used, first visitors who follow interests were more likely to use signboards and signposts while first visitors that used landmarks were more likely to use track surfaces and vegetation types as landmarks. The more tourists are familiar with the environment, the less chances of them using landmarks (Xia, Packer, and Dong 2009). In terms of arrival time, the more familiar the tourists are with the environment, the longer time they will take to end their wayfinding tasks. The shortest time taken was by first visitors who follow landmarks and this has proved that the utilization of landmarks has greatly assisted the wayfinding tasks (Sorrows and Hirtle 1999).

Both age and travel group differences have shown relations with two of the tourist wayfinding behaviours which were wayfinding strategies and landmarks used. Firstly, in terms of age, middle age respondents were more likely to use signposts as their landmarks. Old age respondents were more likely to use vegetation type. This is parallel with Xia, Packer, and Dong's finding (Xia, Packer, and Dong 2009). Secondly, in terms of wayfinding strategy used, old age respondents were more likely to use the first noticed strategy.

Lastly, in terms of travel group differences in relation with landmarks used, couples were more likely to use the signboard and signpost. This is similar to Xia, Packer, and Dong's finding that suggested couples were the highest travel groups who have been using signboard during wayfinding (Xia, Packer, and Dong 2009). Groups were more likely to use vegetation type. In terms of travel group differences in relation to strategy used, group respondents were more likely to use the shortest path. This was due to the fact that the only group respondent was a tourist guide and he might want to finish the task as quickly as possible due to frequency of his visits.

The researcher has also generated some outputs involving the location part of the tourist wayfinding behaviours. In terms of route usage, females have dominated routes towards and from the Koala Boardwalk while males dominated routes towards and from the Woodland Bush. In terms of age, same as female, middle age respondents have dominated routes towards and from the Koala Boardwalk. These have shown the effects that the difference of age and gender has in producing different interest in the determination of route towards any destination.

6. CONCLUDING REMARKS

Individual differences such as gender and familiarity with the environment have been proven to have impact on tourist wayfinding behaviours. Tourism is one of the most rapidly developing industries in the world. The methodology developed and the findings in this thesis can assist tourists, tourist agencies and tour operators in designing tour itinerates and packages and help tourist organisations improve facility management. This methodologies and findings can also be used to further clarify and develop the knowledge of tourist movements.

REFERENCES

Arrowsmith, C., Zanon, D., Chhetri, P. 2005. Monitoring visitor patterns of use in natural tourist destinations. In *Taking tourism to the limits: Issues, concepts and managerial perspectives*, ed. C. Ryan, Pageand, S., Aicken, M., 33-52. The Netherlands: Elsevier.

Asakura, Y., Iryo, T. 2007. Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transportation Research Part A* 4, pp. 684-690.

Babin, B. J., Darden, W. R., Griffin, M. 1994. Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of Consumer Research* 20(4), pp. 644-656.

Chebat, J-C., Gelinas-Chebat, C., Therrien, K. 2008. Gender related wayfinding time of mall shoppers. *Journal of Business Research* 61, pp. 1076-1082.

Chen, C.-H., Chang, W.-C, Chang, W.-T, Gender Differences in Relation to Wayfinding Strategies, Navigational Support Design, and Wayfinding Task Difficulty, *Journal of EnvironmentalPsychology* (2008),doi:10.1016/j.jenvp.2008.07.003.

Cherry, K. E., Park, D. C. 1993. Individual differences and contextual variables influence spatial memory in younger and older adults,. *Psychology and Aging* 8, pp. 517-526.

Golledge, R. G. 1999. Human wayfinding and cognitive maps. In *Wayfinding behavior cognitve mapping and other spatial Processes*, ed. R. G. Golledge. The Johns Hopkins University Press, Baltimore , pp. 5-45.

Gramann, K., Muller, H. J., Eick, E., Schonebeck, B. 2005. Evidence of separable spatial representations in a virtual navigation task. *Journal of Experimental Psychology: Human Perception and Performance* 31(6), pp. 1199-1223.
Hallahan, L., Bomford, J. 2005. *Phillip Island Victoria/Australia*. Scancolor (Australia) Pty. Ltd, Melbourne.

Harrell, W. A., Bowlby, J. W., Hall-Hoffarth, D. H. 2000. Directing wayfinders with maps: the effects of gender, age, route complexity, and familiarity with the environment. *Journal of Social Psychology* 140(2), pp. 169-179.

Holscher, C., Meilinger, T., Vrachliotis, G., Brosamle, M., Knauff, M. 2007. Up the down staircase: Wayfinding strategies in multi-level buildings. *Journal of Environmental Psychology* 26(4), pp. 284-299.

Iaria, G., Palermo, L., Committeri, G., Barton, J. J. S. 2009. Age differences in the formation and use of cognitive maps. *Behavioural Brain Research* 196: 187–191.

Kirasic, K. C. 2000. Age differences in adults’ spatial abilities, learning environmental layout, and wayfinding behavior. *Spatial Cognition and Computation* 2: 117-134.

Lawton, C. A. 1994. Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles* 30, pp. 765-779.

Lawton, C. A., Kallai, J. 2002. Gender differences in wayfinding strategies and anxiety about wayfinding: a cross cultural comparison. *Sex Roles* 47 (9-10), pp. 389-401.

Malinowski, J. C., Gillespie, W. T. 2001. Individual differences in performance on a large-scale, real-world wayfinding task. *Journal of Environmental Psychology* 21(1), pp. 73-82.

O' Connor, A., Zerger, A., Itami, B. 2005. Geo-temporal tracking and analysis of tourist movement. *Mathematics and Computers in Simulation* 69 (1-2), pp. 135-150.

Passini, R. 1984. Spatial representations: a wayfinding perspective. *Journal of Environmental Psychology* 4, pp. 153 164.

Reed, A. 2000. *Management of Koala Conservation Centre*. Phillip Island Nature Park, Melbourne.

Sorrows, M. E., Hirtle, S. C. 1999. *International Conference COSIT '99, August 25-29, 1999: The nature of landmarks for real and electronic spaces*. Stade, Germany: Springer-Verlag.

Walmsley, D. J., Jenkins, J. M. 1992. Tourism cognitive mapping of unfamiliar environments. *Annals of Tourism Research* 29(3), pp. 268-286.

Xia, J. 2007. Modelling the spatio-temporal movement of tourists. PhD diss., RMIT University. http://adt.lib.rmit.edu.au/adt/uploads/approved/adt VIT20080110.161021/public/01front.pdf

Xia, J., Arrowsmith, C., Jackson, M., Cartwright, W. 2008. The wayfinding process relationships between decision-making and landmark utility. *Tourism Management* 29(3), pp. 445-457.

Xia, J., Packer, D., Dong, C. 2009. *18th World IMACS / MODSIM Congress, July 13-17, 2009: Individual differences and tourist wayfinding behaviours*. Cairns.

EVALUATION OF ONLINE ITINERARY PLANNER AND INVESITAGATION OF POSSIBLE ENHANCEMENT FEATURES

H.M. Tam[1] & Pun-Cheng, L.S.C.[2]

Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University
([1]tamhoming@gmail.com & [2]lspun@polyu.edu.hk)

KEY WORDS: Online Itinerary Planner, Location-based Service, Web-based application, Geo-visualization, Tourist guide

ABSTRACT:

Current information available to the tourists visiting Hong Kong seems to be abundant and fragmented. Usually, individual tourists tend to plan their itinerary well ahead they arrive, either through tourist guide or online information sources (e.g. existing itinerary planning websites). With the enhancement of information technology, it is expected that online itinerary planning would be a complete supplementary to the hard copy travel guides and magazines in the future. However, current online itinerary planner overlooks the transportation link and optimum travel plan. For example, several tourist spots can be visited in a sequential manner so as to optimise the time spent. It is the place where this project plays its role. This project aims to develop a scheduling algorithm based on the Greedy Algorithm that helps prepare an itinerary for tourists visiting Hong Kong on an individual basis. Tourists' limited knowledge in the spatial extent of and transportation facilities of Hong Kong have always been obstacles of arranging an efficient (optimal) travelling plan. With the view to minimize the travelling time and maximize the time of sightseeing, shortest travelling time between tourist spots have been adopted as the principle in deriving the solution. The benefit brought by the presence of this system and the current availability of tourist information (both transport and tourist spot) will be assessed and evaluated.

1. INTRODUCTION

This chapter aims to let readers to understand the background, objectives and scope of the system. Such information will be important for readers' further understanding in the system architecture.

1.1 BACKGROUND

Current information available to tourists visiting Hong Kong is abundant. However, it is rather fragmented: you may find transport or tourist spot information available separately on different web sites. For instance, the Hong Kong Tourism Board has set up a website titled "Discover Hong Kong" for some time. It provides information from tourist spots, accommodations to dining and shopping. The information available on sites like "Discover Hong Kong" is indeed quite diverse in content. You may need some time to integrate the data you have received on your own. A tourist may have no idea how to travel between two tourist spots - they may finally come up with an itinerary that covers all the destinations, but such itinerary is not necessarily cost efficient.

Sequencing a time and cost efficient itinerary can be a hard question to tourists, especially at times when they don't have much knowledge about the complex transportation and the spatial proximity of their desired visiting places. They will not be able to make good use of the well developed transport facilities throughout their trips. It is the place where this project plays its role to help. It will help tourists schedule the optimal path (with sequence) of visiting their desired spots so as to achieve the highest efficiency and save their time and money. It is their limited time of stay in Hong Kong making every second become important. It is beyond question that different tourists will visit the same list of tourist spots at different sequence. Optimizing their travelling sequence and preparing transport suggestion will bring much convenience to them.

1.2 OBJECTIVES

This project aims to develop a scheduling algorithm that helps prepare an itinerary for tourists visiting Hong Kong on an individual basis. Tourists may select their own tourist spots. Besides deriving a solution which includes all the tourist spots selected by the user, the itinerary solution also includes suggested transport means (and also the time taken for the means), time spent on each tourist spot and so on. In short, the objectives of the system are:

1. To provide a tool for tourists to help them schedule their visit
2. To provide an itinerary solution
3. To provide transport information between tourist spots and the hotel
4. To provide maps showing the vicinity of the points of interest (both tourist spots and the hotel)
5. To suggest the time to be spent at a particular tourist spot

It is the ultimate objective of the system to serve as a complete substitute to all other collaborations of tourist information sources so that users can have their own itinerary. They may also print their itinerary far before they set off.

1.3 SCOPE OF STUDY

Hong Kong is a city with well developed transport network. This transport network includes a great variety of transport modes, e.g. railway, bus, mini-bus, taxi, tram etc. Given two places, you can usually name more than 1 means of transport connection available. As the aim of the study is to investigate the tourist scheduling algorithm, we would restrict the route chosen to be the shortest possible travelling time between two spots.

Another issue that the system should cater is the number of places (tourist spots and hotels). It is certainly impossible to include all the tourist spots in Hong Kong - there are too many. Besides, classifying a place as a "tourist spot" is a subjective decision. The system is a pilot study and therefore limits the number of tourist spots selected, yet, spreading throughout the territory. It is not only the number of tourist spots that matters, but also their spatial extent (in terms of travelling time among spots), especially when it involves cross-district travelling. With the reasons stated above, the system developed would be restricted to two issues:

1. Number of tourist spots, and;
2. Routes with the shortest time in between tourist spots

Other criteria (other than travelling time and cost) should also be put into consideration. The rationale of the scheduling algorithm will be discussed in later sections.

2. LITERATURE REVIEW

The context in this section is actually giving answers to a very simple yet difficult question: "What constitutes an itinerary planner?"

2.1 ITINERARY

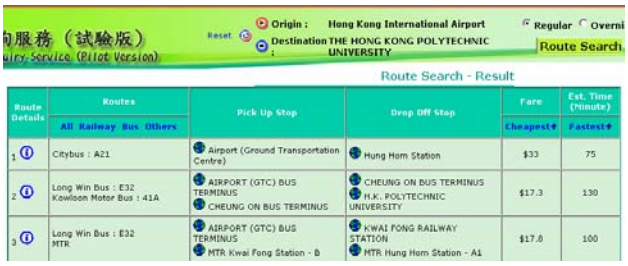
An itinerary is usually referred as tourist guide, or guide book. It is a book that aims to provide information of high practical value to tourists. These may include geographic location, transport, shopping, dinning or any other information of tourist spots limited to a particular area. An itinerary may also solely refer to the way of getting from one place to another. The term "itinerary" in this paper is confined as the one that tourists can rely on during their stay in Hong Kong.

2.1.1 Information provided: "Lonely Planet" is a renowned brand of travel guidebook in the world. In the contents of the book "Hong Kong & Macau – Pick & Mix Chapter", the key contents include eating, shopping, entertainment, sports & activities and transport. (Lonely Planet, 2008) This classification is indeed sensible and reasonable – tourist spots and activities have always been tourists' utmost concern. The system developed should serve this principal function.

It is obvious that not every single tourist will purchase a guide book like "Lonely Planet". Instead, they look for free resources available on the internet or other channels. The information availability is therefore assessed.

2.1.2 Tourist spot information: The Hong Kong Government is always keen in promoting Hong Kong to the tourists and branding Hong Kong as the "Asian's World City". The site titled "Hong Kong Fun in 18 Districts" set up by the Home Affairs Department includes a list of renowned tourist spots. Hong Kong Tourism Board (HKTB) is an organization which is responsible for promoting Hong Kong to the world. Since Hong Kong is always referred as the "Shopping Paradise", the Board has set a web site titled "Where to Shop". The site includes textual information, supplemented by photos that help tourists learn about the details of individual shopping area.

2.1.3 Transport information is crucial to a tourist. Some cities, e.g. Sydney, have detailed arrival and departure timetables for every transport. It is not the case in Hong Kong – only the frequency of transport (usually given as interval, say, 8-12 minutes) will be shown publicly (though detailed departures exist for managerial purpose). Most of the transport companies in Hong Kong have made their transport information online so that the public can get the information anytime. In early 2009, the Transport Department launched the Public Transport Enquiry Service (PTES). This service provides transport route suggestions once user specifies their origin and destination (Figure 1).


詢服務（試驗版）
uiry Service (Pilot Version)
Reset
Origin : Hong Kong International Airport
Destination : THE HONG KONG POLYTECHNIC UNIVERSITY
Regular
Overni
Route Search
Route Search - Result

Route Details	Routes All Railway Bus Others	Pick Up Stop	Drop Off Stop	Fare Cheapest	Est. Time (Minute) Fastest
1	Citybus : A21	Airport (Ground Transportation Centre)	Hung Hom Station	$33	75
2	Long Win Bus : E32 Kowloon Motor Bus : 41A	AIRPORT (GTC) BUS TERMINUS CHEUNG ON BUS TERMINUS	CHEUNG ON BUS TERMINUS H.K. POLYTECHNIC UNIVERSITY	$17.3	130
3	Long Win Bus : E32 MTR	AIRPORT (GTC) BUS TERMINUS MTR Kwai Fong Station - B	KWAI FONG RAILWAY STATION MTR Hung Hom Station - A1	$17.8	100

Figure 1 Transport route suggestions by PTES

2.1.4 Itinerary planning service provider: There are websites that provide itinerary planning service, TripIt is one of them. It processes travel confirmation emails, weather, driving directions etc into an itinerary. The objective of the site is to save the time that travellers spent on arranging their own itinerary. TripIT may sound great to cities like San Fransico where driving between spots are usual practices. Yet, the place where the pilot study take place – Hong Kong – is a small place with 7 million people. Car rental is certainly not a good option for travellers in Hong Kong - public transport is too well developed. Narrow and congested road conditions of Hong Kong also discourage tourists from renting a car.

HKTB launched a website which provides an interface for the user to plan their itinerary. User first selects the duration of stay, time of the day for arrival and departure. This itinerary is impracticable since its time interval is only up to morning, afternoon and night sessions only. It does not cater any dining needs nor shopping needs of the tourists.

2.2 SCHEDULING ALGORIHM

The scheduling of an itinerary that takes the shortest time is actually a practical case of the Travelling Salesman Problem (TSP). Given a list of places, a salesman must find a path to visit all the places for exactly one time. TSP is a well-known NP-complete problem, meaning that there is no algorithm that can solve TSP efficiently. The exact algorithm to solve TSP is to compute all the permutation and look for the one with lowest cost. It is therefore the time taken for scheduling an itinerary with n cities is actually $O(n!)$. It works fine for a small number of cities, but becomes impracticable for cases with more than 20 places.

Heuristics and approximation algorithms are therefore devised. They give good solutions quickly. With these algorithms, extremely large problem can also be solved within a reasonable amount of time. These solutions are probably 2-3% away from the optimal solution. The common approach of these algorithms is to first construct a possible path. Then, improve the solution using different improvement algorithms, like iterative improvement or randomized improvement.

3. METHODOLOGY

An online itinerary planner should provide solutions quickly and promptly since the user is too impatient to wait. Therefore, the system devised in this paper employs the approximation algorithm with improvement tactics to compute the solution in a timely manner. Assumptions are made for better modelling of the system.

3.1 ASSUMPTIONS

Below is a list of assumptions that have been considered in the initial stage of the system design:

1. User values their time in Hong Kong of the utmost importance. They want to keep their travelling time as short as possible.
2. Users will only need to specify the hotel they stayed and the tourist spots they would like to visit. System will then schedule the itinerary accordingly.
3. Users will always set off their trip from their hotel. They also end their trip there. The hotel specified by the user will always be set as the origin and the destination of the trip.
4. Users will specify the start time and end time for their visit.
5. In case of the time slot (duration between start time and end time) is too short to finish all the visiting. The system will put the rest of the spots on another day, starting from the same time stated by the user.
6. User may need to know the time they need to get from one spot to another. The suggested staying time at a particular tourist spot should also be made available.
7. User may want to know the environment of the vicinity of the tourist spots.

3.2 IMPLEMENTATION PLATFORM

To implement the system, a free and open source cross-platform web server package - XAMPP is employed. It is a solution stack of software with open source software to run web sites on servers. It is easy to install and therefore suitable to be used as the platform of the development. Although, the time taken to run the page may differ from loading a page on a WWW server, it is the features of the system to be assessed. The efficiency (or the networking issue) is left for further enhancement of the project.

3.3 DATABASE

3.3.1 Design: With respect to the assumptions listed, the database schema of the system is to be designed accordingly. Since the basic function of the system is to schedule the itinerary and display the public transport data between tourist spots, the database developed should have fields storing corresponding information. Three tables are constructed in the database, namely SPOT, JOURNEY and SUBJOURNEY. The three tables are designed with different purposes: (Table 2 & Figure 3)

Table 2 Purposes of the three tables in the database

Table	*Purposes*
SPOT	1. Keeps details of individual tourist spots and hotels (e.g. name, district) 2. Facilitates information storage and retrieval 3. Latitude and Longitude are stored to enable the use of Google Map for accurate location display
JOURNEY	1. Stores the transport route suggestions of all combinations of origin-destination 2. Time taken to travel using the respective transport route suggestion
SUBJOURNEY	1. Stores the information of individual transport route – pickup and dropoff point will also be given. 2. Details of different legs within a journey – interchange is unavoidable in some cases.

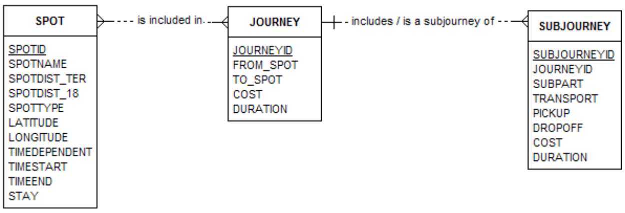
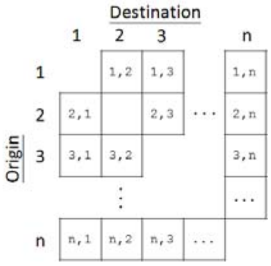


Figure 3 Entity-Relationship diagram of the system

3.4 DATA PREPARATION

3.4.1 Data Preparation: The scheduling algorithm adopted in the system uses the travelling time as the means of scheduling. Real transport data was collected through the PTES (see Section 2.1.3). Throughout the collection of data from the PTES, the fastest route is chosen since time of utmost concern to tourists.

Hotels and tourist spots can be treated as the same during the collection of transport data. They can be considered as Points Of Interest (POIs). Schematically, the cost (time taken) between the POIs data can be represented like a square matrix (Figure 4). The first cell on the 2nd row, named with (2,1), should hold the cost of getting from POI2 to POI1. If (2,1) is considered as the inbound journey, then (1,2) would be the outbound of the same POI pair. If the POIs are drawn as vertex in a graph, the two edges between each POI pair would be the cost of the inbound and outbound journey between them (Figure 5). The graph is an asymmetric graph.


Destination

Origin	1	2	3		n
1		1,2	1,3		1,n
2	2,1		2,3	···	2,n
3	3,1	3,2			3,n
		⋮			···
n	n,1	n,2	n,3	···	

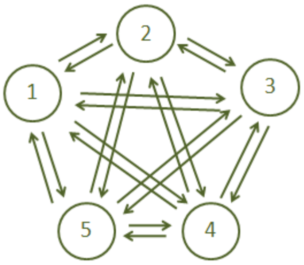


Figure 4 Schematic representation of the transport routes to be prepared (Left)

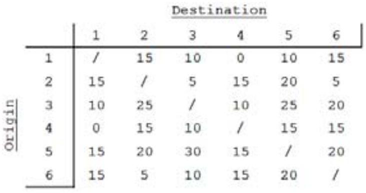
Figure 5 Asymmetric graph formed upon completion of the transport route (Right)

3.5 SCHEDULING

There are two ways to schedule the list of tourist spots selected by the user.

1. Schedule all the tourist spots with the use of greedy algorithm together with the k-opt heuristic search (General Solution)
2. Place the time constrained spots at particular position that satisfy the constraints, then schedule the rest of them (Time-constrained solution)

(1) has been devised in the prototype platform while (2) is still on theoretical side. Real data would be used for (1) and the cost matrix is given here (Table 6).


Destination

Origin	1	2	3	4	5	6
1	/	15	10	0	10	15
2	15	/	5	15	20	5
3	10	25	/	10	25	20
4	0	15	10	/	15	15
5	15	20	30	15	/	20
6	15	5	10	15	20	/

Table 6 Cost (Time) matrix of possible tourist spot pairs

3.5.1 General Solution: Consider a case where the user has selected the hotel with the SPOTID of 1, and some tourist spots with SPOTID from 2-6. With the use of the greedy algorithm (GA) alone, the solution will be like Figure 7. User shall start from POI A and end at POI A again. sequence (numbers inside circles represents tourist spots) while the figures between ten with :

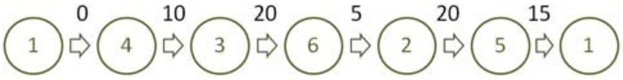


Figure 7 The scheduling solution with Greedy Algorithm

Since there is a possibility that the path generated may be the longest path (Jergen et. al, 2004), the solution needs to be passed through an improvement step. The tour improvement heuristic will be adopted as the improvement tool. The result generated with improvement tool applied will be:

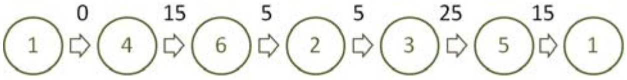


Figure 8 The scheduling solution with Greedy Algorithm with tour improvement heuristic

Attention is to be drawn on the difference in the position of the 3 spots located in the middle before and after the application of tour improvement heuristic. The difference in position of the 3 spots is marked by the rectangles in Figure 9.

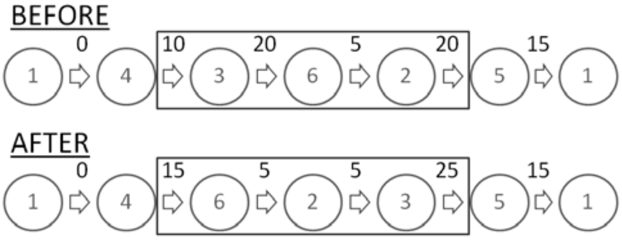


Figure 9 Difference in position before and after tour improvement heuristic. (3-Opt adopted)

The difference resulted in Figure 9 is actually a result of the K-Option improvement heuristic. At this case, the value of K is set as 3. The repositioning took place since a shorter path is available (note only the figures above the arrows in both rectangles). There should be at least 5 spots in the itinerary (4 if start and end point are the same). The first node and the last node will be fixed (because of their nature). In an arbitrary list with spots A, B, C, D, E. It would yield *3!* (a total of 6) permutations (only three spots in between can be swapped).

For a list with more than 5 nodes, the 3-Opt will be first performed on the 2nd, 3rd and 4th spot. Then the floating window will be shifted to the 3rd, 4th and 5th. It proceeds and processes until the last node of the floating window touches the *(n-1)th* node.

By illustrating how the floating window works, the difference in the sequence generated solely by greedy algorithm and greedy algorithm with 3-Opt employed is now justified.

The reason of setting the K value as 3 in the improvement heuristic is due to the spread of the tourist spots at different districts – there are about 3 tourist spots at the same district in the database. Since it is rather unlikely that the tour will be improved (in terms of reduction in total travelling time) by having tourist spots, the value 3 is justified.

3.5.2 Time-constrained Solution: The time slot fitting approach is the algorithm which first deals with time constrained tourist spot first. To make it easier for understanding, the illustration only includes ONE time constrained spots only. Lan Kwai Fong is selected as the time constrained spot so as to demonstrate this approach. It is assumed that the bars in Lan Kwai Fong only operates from 18:00 – 24:00 daily. (Figure 10)

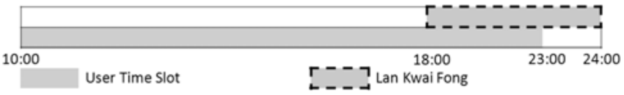


Figure 10 User defined time slot and operating hours of time constrained spot

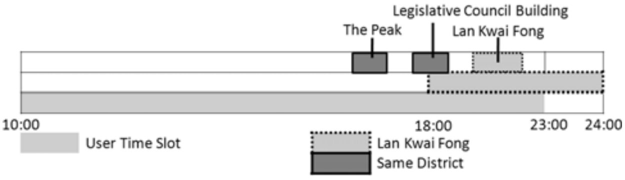


Figure 11 Time constrained spot with some same district spots inserted

Then, Lan Kwai Fong would be inserted as late as possible in the user defined time slot (Figure 11). Visit to the Lan Kwai Fong should end some time before the user defined time slot since it takes sometime for the tourist to back to the hotel. Tourist spots of the same district will then be inserted before the time that Lan Kwai Fong resides. (Figure 11)

Once the tourist spots (of the same district) have been inserted, the algorithm shall look for the possibility of making the whole trunk to be switched to an earlier time. (Figure 12) The rationale of this move is due to the consideration that there is a chance that the schedule is subject to delay (by traffic congestion). If there is a delay of 30 minutes for the day, the maximum time that the user can spend on Lan Kwai Fong would reduce by 30 minutes.

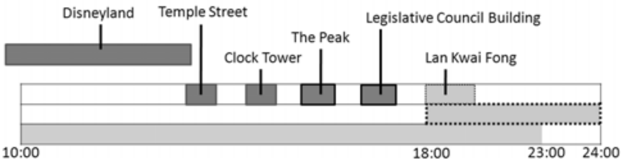


Figure 12 Time constrained spot and two other spots shifted forward

Once the shifting is completed, tourist spots of other districts will be inserted before the scheduled tourist spots. If the amount of time that takes to finish the trip is loner than the current time available. The system shall be looking for the possibility for a shift to a later time (Notice that the time visiting the time constrained spot has been shift earlier, so, there are some rooms for another shift).

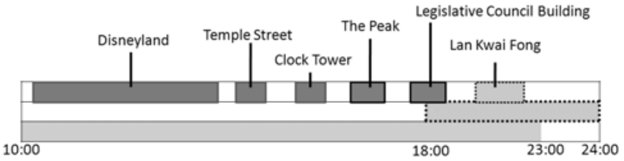


Figure 13 Forward shift of the whole route in order to fit in an extra tourist spot

Extra tourist spots will then be fitted in (Figure 13). Optimization will then take place by implementing the k-opt heuristic iterative search. The sequence before the time constrained spot is said to be fixed after the optimization. (Figure 14) Please be noted that the swapping in position between Temple Street, Disneyland and Clock Tower is fictional. It only serves the illustration of optimization (3-Opt) only.

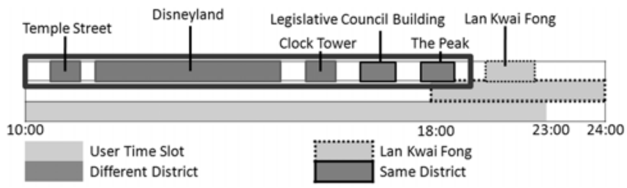


Figure 14 Optimization completed for the spots before the time constrained spot

The scheduling of the spots later than the Lan Kwai Fong will then be continued. The scheduling mechanism is much simpler – time constrained spot (Lan Kwai Fong) as the start point and the hotel as the end point.

4. SYSTEM IMPLEMENTATION

This chapter aims to provide details on the issue considered during the preparation of the user interface and other relevant issues.

4.1 SYSTEM WORKFLOW

It takes 4 steps to build the customized itinerary (Figure 15). Hotel should be chosen first since it is the first and last node of the itinerary. Secondly, the arrival and departure time would let the system know how much time the user would be make himself available each day. User should then select the tourist spots. The system would then click the “Schedule” button. The system would then process the information and generate the itinerary using the algorithm specified in the section 3.5.

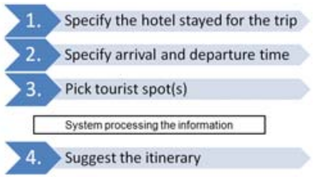

1. Specify the hotel stayed for the trip
2. Specify arrival and departure time
3. Pick tourist spot(s)
System processing the information
4. Suggest the itinerary

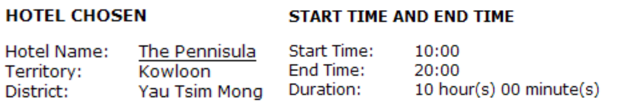
Figure 15 The 4-step system workflow of the database

4.2 USER INTERFACE & REPORT PRINTING

It is the intention of the system to be developed with a user-friendly user interface. Short text description will be given in the first page to show the procedure of the scheduling function.

The whole program will be divided into 3 steps:

1. User selects a hotel
2. User specify the duration of visits to be done each day.
3. System will print the information specified by user for checking (Figure 16)
4. User selects a number of tourist spots.



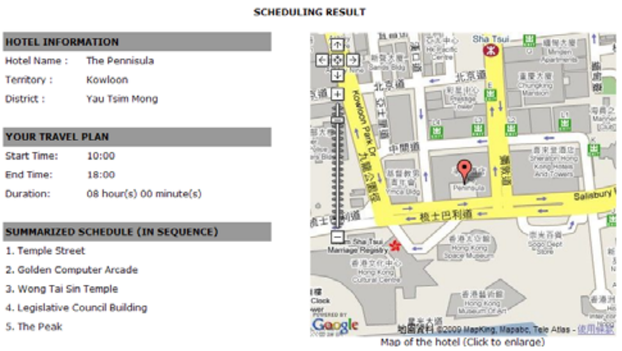
HOTEL CHOSEN

Hotel Name:	The Pennisula
Territory:	Kowloon
District:	Yau Tsim Mong

START TIME AND END TIME

Start Time:	10:00
End Time:	20:00
Duration:	10 hour(s) 00 minute(s)

Figure 16 Information shown upon the completion of step 1 and 2



SCHEDULING RESULT

HOTEL INFORMATION

Hotel Name :	The Pennisula
Territory :	Kowloon
District :	Yau Tsim Mong

YOUR TRAVEL PLAN

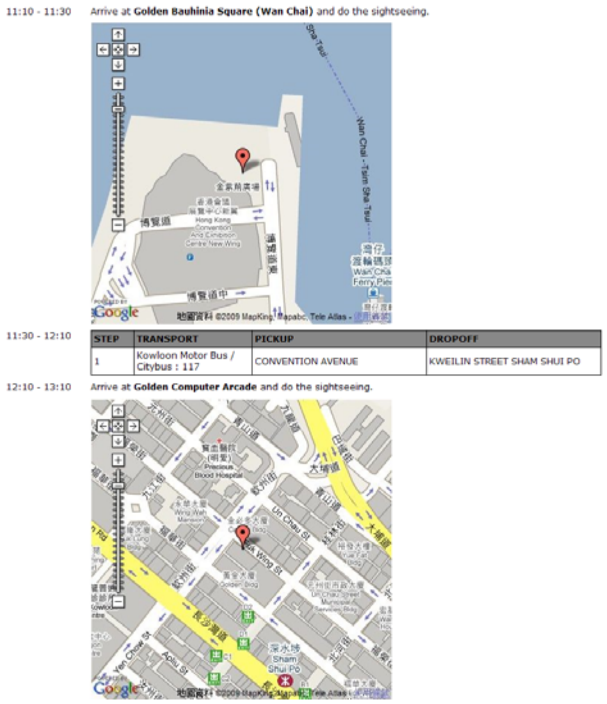
Start Time:	10:00
End Time:	18:00
Duration:	08 hour(s) 00 minute(s)

SUMMARIZED SCHEDULE (IN SEQUENCE)

1. Temple Street
2. Golden Computer Arcade
3. Wong Tai Sin Temple
4. Legislative Council Building
5. The Peak

Map of the hotel (Click to enlarge)

Figure 17 Itinerary generated (extract)



11:10 - 11:30 Arrive at **Golden Bauhinia Square (Wan Chai)** and do the sightseeing.

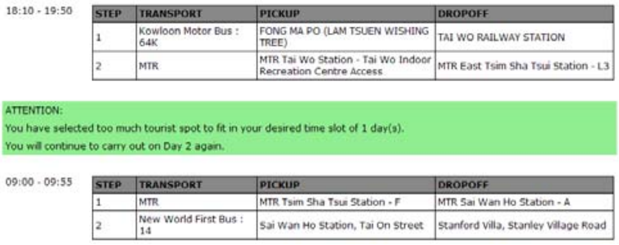
11:30 - 12:10

STEP	TRANSPORT	PICKUP	DROPOFF
1	Kowloon Motor Bus / Citybus : 117	CONVENTION AVENUE	KWEILIN STREET SHAM SHUI PO

12:10 - 13:10 Arrive at **Golden Computer Arcade** and do the sightseeing.

Figure 18 Detailed transportation means tourist spots

The user may then select the spots and let the system to do the scheduling. Figure 17 and 18 show a typical layout of the itinerary generated. Besides, the system is capable of handling exceptional case. For example, an ambitious user may select numerous tourist spots that cannot be visited within the time slot specified. It is therefore the system prompts the user of this fact and continue the journey on another day. (Figure 19)



18:10 - 19:50

STEP	TRANSPORT	PICKUP	DROPOFF
1	Kowloon Motor Bus : 64K	FONG MA PO (LAM TSUEN WISHING TREE)	TAI WO RAILWAY STATION
2	MTR	MTR Tai Wo Station - Tai Wo Indoor Recreation Centre Access	MTR East Tsim Sha Tsui Station - L3

ATTENTION:
You have selected too much tourist spot to fit in your desired time slot of 1 day(s).
You will continue to carry out on Day 2 again.

09:00 - 09:55

STEP	TRANSPORT	PICKUP	DROPOFF
1	MTR	MTR Tsim Sha Tsui Station - F	MTR Sai Wan Ho Station - A
2	New World First Bus : 14	Sai Wan Ho Station, Tai On Street	Stanford Villa, Stanley Village Road

Figure 19 Prompt of overtime visiting

5. SYSTEM EVALUATION & ENHANCEMENT

5.1 SYSTEM PERFORMANCE

The system receives users input and generates the path without any error. Yet, the execution time was not recorded. The limited size (in terms of number of tourist spots) does not put much stress in the time of computation.

Since there is only one user in the development platform, it is rather hard to perform either stress test or any similar testing. Normally, the system is able to return a result within 5 seconds at a case where all the tourist spots have been selected. Around 500 database transactions were taken place.

5.2 SYSTEM LIMITATION & POSSIBLE ENHANCEMENT

5.2.1 Programming language: PHP is easy to program and implement. It is possible for later system to consider the compilation time of different languages and choose a better and more efficient one. Compatibility between the scripting language and the server support should also be considered.

5.2.2 Development Platform and Implementation Platform: In this project, the combination of PHP and MySQL is used to implement the scheduling. It should be noted that this combination is adopted because of its convenient and readily available technical support.

5.2.3 Scheduling algorithm: The rationale of setting the K value of K-opt heuristic as 3 has been discussed earlier. This “magic number” is to be revised once necessary. It should be noted that the selection of the k value should be careful and be limited to small numbers or the number of combination will be increasing geometrically. The system currently uses the greedy algorithm together with the heuristic search to seek for possible improvement of the result. Though the problem size in this study is small enough to perform exhaustive search, it is not the intention of the system to compute the shortest solution, but an optimal solution instead. The real time computation of the total travelling time of an exhaustive search possesses too many queries to the database.

To avoid doing exhaustive search in deriving an optimal solution, an alternative is to screen out some of the “impractical cases” from the exhaustive search. For example, given 5 tourist spots, the tourist spots can be further divided district-wise. It is to filter out certain impracticable combinations so as to save the computation time.

5.2.4 Solution with time constrained spots: It is the limitation in time and difficult in terms of coding to implement the time dependent solution. The current scheduling ignores the time effect of the spots – the time effect refers to the opening hours (e.g. the Hong Kong Disneyland), operating hours (e.g. the Museums), or favouring hours (e.g. visiting the Peak at night is more preferable than afternoon time). The algorithm to solve the time constrained spots have been discussed earlier.

5.2.5 Which weighs higher: time or distance? The pilot study only consider the predetermined travelling time when performing scheduling. It is however that traffic condition differs a lot in peak and non-peak hours. The distance factor may be applied in order to find out the chances of having traffic congestion during the peak hours.

5.2.6 From the nearest district to the farthest district? The system currently prepares the initial solution using the Greedy Algorithm, which makes the itinerary resulted visiting the nearest spot first. That means the solution will end up starting from less transport time to longer transport time. It is questionable if the total time of the trip will be lower if the trip can be started another way round – the system first pick the furthest spot from the origin and do the greedy algorithm, then the improvement heuristics.

5.2.7 Mapping of results: Currently, there is only a map showing each spot. Yet, the map does not provide detailed information to enable user to get to the transport facilities that the system suggests.

5.3 SUGGESTIONS

The computation of the initial solution (which needs to be improved at later part) is derived from the greedy algorithm. Some studies have pointed out the greedy algorithm may not come up with a good solution, or even a decent solution for later improvement to take place. The choice of the algorithm was mainly due to its simplicity. Further enhancement should explore the possibility of using geographical relationship as aid to schedule. Better utilization of Goggle Map may help user learn more about the environment surrounding the area.

6. CONCLUSION

As stated in the very beginning, the initiative of the project is to devise a system that capable to schedule an itinerary that satisfies the conditions specified by the users. The scheduling algorithm and the database architecture is rather simple. It is expected that further enhancement is to be made on both the algorithm and the database architecture. The use of hardware to accommodate / suit the needs of the system is to be assessed afterwards.

REFERENCE

1. CORMEN T.H., LEISERSON C.E. & RIVEST L.R. (1989) Introduction to Algorithms. The Massachusetts Institute of Technology
2. EVANS R. J. and MINIEKA E. (1992). Optimization Algorithms for Networks and Graphs. Marcel Dekker, Inc.
3. MICHAEL S. (2002). Internet Travel Planner : How to Plan Trips and Save Money Online. Globe Pequot Press
4. PUN-CHENG Lilian, MOK E.C.M. et.al. (2007), EASYGO-A public transport query and guiding LBS. Lecture Notes in Geoinformation and Cartography, Location Based Services and TeleCartography, Springer, pp. 545- 554
5. WONG, C.S. (2004). A web-based GIS for Scheduling Travel Journey in Hong Kong. The Hong Kong Polytechnic University
6. HOME AFFAIRS DEPARTMENT, HKSAR (2008). Hong Kong Fun in 18 Districts. Hong Kong Special Administrative Region. (http://www.gohk.gov.hk/, Accessed 31st March, 2009)
7. Public Transport Enquiry Service (Pilot Version) http://ptes.td.gov.hk/ (accessed 27th Nov, 2009)THE HONG KONG TOURISM BOARD (2009) Discover Hong Kong (http://www.discoverhongkong.com, Accessed 1st February, 2009)
8. TripIt - Online travel itinerary and trip planner http://www.tripit.com/ (accessed 27th Nov, 2009)

APPENDIX

Implementation of the pilot study is available at:
http://myweb.polyu.edu.hk/~06159619d/v5.php

A SIMPLE AND EFFICIENT SQL-BASED APPROACH FOR RETRIEVAL OF GEOSPATIAL DATA IN MOBILE GIS APPLICATION

G. Y. K. Shea [a], J.N. Cao [b]

[a] Dept. of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China – geoffrey.shea@polyu.edu.hk
[b] Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China - csjcao@comp.polyu.edu.hk

Commission II

KEY WORDS: Mobile GIS, mobile computing, mobile geospatial database

ABSTRACT:

The core objective of mobile GIS application is to provide vicinity geospatial information to the moving clients based on their current position. Since most of the mobile devices are limited in resources such as processing power, storage capability in terms of memory size, communication speed. Sending lot of map contents are normally required for all those applications as well. Therefore, the success of a mobile GIS application depends mainly on the efficiency of the geospatial data delivery method used and how well the adaptation can be employed. Two customised spatial index numbering schemes have been devised. Evaluation has been conducted in the actual mobile environment to determine the relative efficiency for this two numbering schemes. The main purpose for devising a unique spatial index numbering system is to provide a faster and simpler way other than using R-Tree algorithm to fetch the regularly geo-referenced map base images. A prototype mobile application is developed based on this customised spatial index scheme with an open source spatial-enabled RDBMS engine. The work flows and algorithms in the initial formation and subsequent updating process of the local dynamic moving geospatial database will also be described in this paper.

1. GROUPING OF GEOSPATIAL DATA

Basically, the transmission of map contents to the mobile client is comprised of two kinds of information – geographic and attributes. There is a common set of geospatial data that needs to be served to the mobile clients irrespective of the type of applications This common set of geospatial data contains: (a) road features; (b) building features; (c) hydrology features such as coastline, rivers and reservoirs; and (d) major point features such as bus stops and address. Usually, we called this set of data as “map base”. Map base data bears the following characteristics:

- Delivered to the client as a single image
- Each single image is geo-referenced (i.e. position information is embedded in the image)

The map base data serves as the backdrop for displaying other geographic related data. Therefore, another set of data in association with the map base will be sending to the client in the same time. We called this set of data as “application data”.

Usually, the application data will contain more descriptive information other than the position information and the number of layers to be included can be one or more.

The format of this set of data is varying and is depending on the type of application. It can be stored in a database, in XML format, in text file format, etc. Figure 1 depicts the geographical relationship between the map base data and application data.

According to George Cantor (1845-1918), the founder of set theory, “A set is a collection of definite, distinguishable objects of perception or thought conceived as a whole”. The map base data can be regarded as a set of object collection containing geo-referenced images. The notation for this set can be expressed as follows:

$$M = \{m_1, m_2, \cdots m_n\} \qquad (1)$$

where m_i is the map base element and i = 1, 2, ... n

Similarly, the application data, set A, can be regarded as a set of data comprising of two other sets, namely position data, set P, and descriptive data, set D.

$$A = P \cup D \qquad (2)$$

Usually, both sets of data M and A will be transferred to the

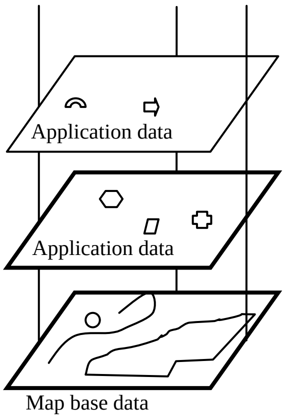


Figure 1. Geographical relationship

client as a whole while the application is executing. That means the client will receive a set of geospatial data, set C, containing both the map base and the application data.

$$C = M \cup A \qquad (3)$$

2. DEFINING A UNIQUE SPATIAL INDEX NUMBERING SYSTEM FOR MOBILE GIS

Two numbering schemes have been devised and tested on their efficiency in the actual mobile environment. The design of unique spatial index number for both schemes is based on a local plane coordinate system as detailed in the following section.

The main purpose for devising a unique spatial index numbering system is to provide a faster and simpler way other than using R-Tree algorithm to fetch the regularly geo-referenced map base images. The usual case is to determine the map base tile that contains the input point, which is a spatial search algorithm called point-in-polygon test. Therefore, we are devising an algorithm to find out the spatial index number of the map base tile that contains the input coordinate of a point.

2.1 Hong Kong 1980 Plane Coordinate System

The local plane coordinate system used in Hong Kong is based on Hong Kong 1980 Grid System (HK80GS) that devised by Survey and Mapping Office of Lands Department. The map name convention for scales 1:20000, 1:10000, and 1:5000 used in this system is shown in Figure 2.

The whole of Hong Kong is covered by a series of map with extent measures 60 Km in the x direction and 48 Km in the y direction. The origin of this grid system is located at the lower left corner with coordinate's value (800000, 800000). The x and y axis is named easting and northing respectively.

2.2 Spatial Index Scheme 1

The base image to be partitioned is created from 1:5000 maps, e.g. 13SWA. The ground coverage for this base image is 3750m wide in the x direction and 3000m high in the y direction. The corresponding image size for this base image is 4800 x 3840 pixels. Each base image is partitioned into 15 rows by 15 columns evenly to give a total of 225 small tiles.

Following this partitioning concept, the whole of Hong Kong is covered by 57600 small tiles (16 x 4 x 4 x 225) mesh of 240 rows by 240 columns.

The unique spatial index number for each small tile in this scheme is a 7- or 8-digit integer coming from 5 parts as illustrated in Figure 3.

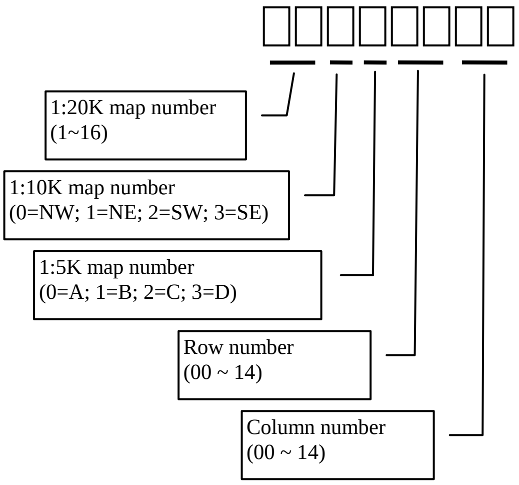
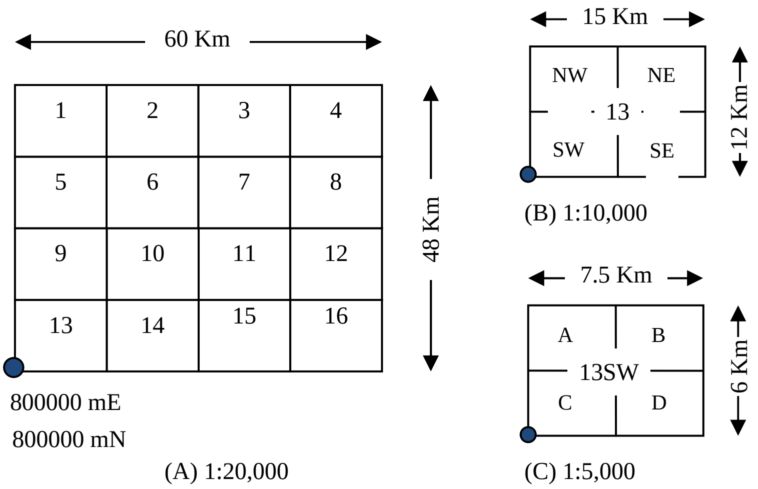


Figure 3. Scheme1 spatial index number

- The first part, P1, is a 1- or 2-digit number of the corresponding map in scale 1:20000.
- The second part, P2, is a 1-digit number of the corresponding map in scale 1:10000. NW is presented as 0, NE is presented as 1, SW is presented as 2, and SE is presented as 3.
- The third part, P3, is a 1-digit number of the corresponding map in scale 1:5000. Similar as above A is presented as 0, B is presented as 1, C is presented as 2, and D is presented as 3.
- The fourth part, P4, is a 2-digit row number.
- The fifth part, P5, is a 2-digit column number.
- Both row and column number starts at the lower left corner with value 0.



(A) 1:20,000

(B) 1:10,000

(C) 1:5,000

Figure 2. Hong Kong map naming convention

For example, all those small tiles within map 13SWD will be in the range of 13230000 and 13231414 as shown in Figure 4. Similarly, those small tiles within map 13SWA will be in the range of 13200000 and 13201414.

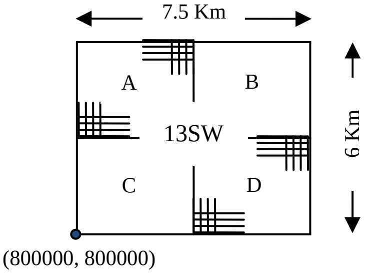


Figure 4. Examples spatial index numbers (Scheme 1)

2.2.1 Finding Scheme 1 Unique Id from Coordinates

The spatial index id is actually depending on a number of factors: (1) point coordinates, (2) relative location in map scale 1:20K, (3) relative location in map scale 1:10K, (4) relative location in map scale 1:5K, (5) relative row number to lower left corner of map scale 1:5K, and (6) relative column number to lower left corner of map scale 1:5K. When express in mathematical function would be:

$$id = f(x, y, P1, P2, P3, P4, P5) \qquad (4)$$

For example, a point with coordinates (804550, 802500) will be contained in small image tile with spatial index id equals to 13231203.

2.2.2 Finding Tile Minimum Bounding Rectangle from Scheme 1 Spatial Id

There are two ways to express a minimum bounding rectangle: (1) lower left corner coordinates and upper right corner coordinates, and (2) lower left corner coordinates, width and height of the bounding rectangle. We use the second to store the small image tile minimum bounding rectangle information. The equations to find the lower left corner coordinates are expressed below.

$$x_{tile} = x_{LL20K} + x_{LL10K} + x_{LL5K} + 250 \times RowN \qquad (5)$$

$$y_{tile} = y_{LL20K} + y_{LL10K} + y_{LL5K} + 200 \times Colur \qquad (6)$$

The width and height for each of the minimum bounding rectangle is 250m and 200m respectively.

For example, the lower left corner of the MBR for the tile with spatial id equals to 13231203 is (804500, 802400).

2.3 Spatial Index Scheme 2

In Scheme 2, the method of partitioning a large image into a number of small tiles is same as the method used in scheme1 above. That is the whole of Hong Kong will be covered by 57600 small tiles mesh of 240 rows and 240 columns after partitioning. The difference is the way to number each small tile uniquely.

The unique spatial index number for each small tile in scheme 2 is an 8-digit number coming from the row number and column number of the small tile position in this mesh. The first four digits are designated for row number and the last four digits are designated for column number. The lower left corner of the small tile will have a row number 0 and column 0. The row number is increased upwardly and the column number is increased toward right hand side. In order to avoid leading zeros, both row number and column number are shifted by a value of 8000 (refers to Figure 5).

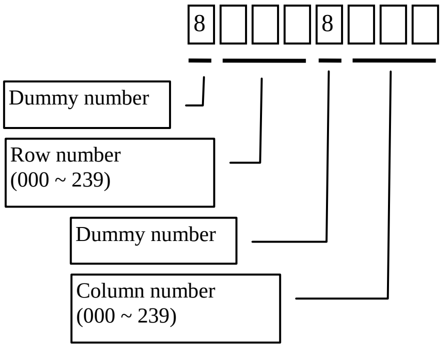


Figure 5. Scheme 2 spatial index number

Therefore, the spatial index number for the lower left corner of the small tile is 80008000, whereas, the upper right corner tile will have an index number 82398239. Some examples are shown in Figure 6.

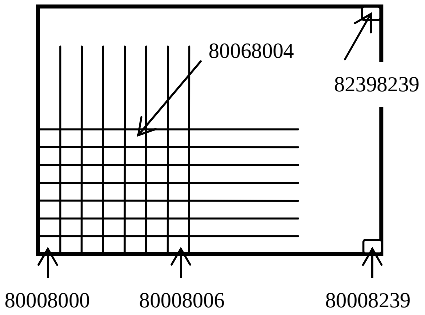


Figure 6. Examples of spatial index number

2.3.1 Finding Scheme 2 Spatial Id from Coordinates

The spatial id for scheme 2 is comparatively simpler than scheme-1 and is depending on the coordinates of the lower left corner of the small image tile which depends on which row and column band that the small tile falls into. When express in mathematical function would be:

$$RowNr = f(y, multiples\ of\ 200) \qquad (7)$$

$$ColumnNr = f(x, multiple\ of\ 250) \qquad (8)$$

Using the same example point in scheme 1, i.e. (804550, 802500), the point is contained within the small image tile with spatial index id equals to 80128018.

2.3.2 Finding Tile Minimum Bounding Rectangle from Scheme 2 Spatial Id

The equations for determining the lower left corner coordinates are expressed below:

$$x_{ttle} = 800000 + ColumnNr \times 250 \qquad (9)$$

$$y_{ttle} = 800000 + RowNr \times 200 \qquad (10)$$

The width and height for each of the minimum bounding rectangle is 250m and 200m respectively.

For example, the lower left corner of the MBR for the tile with spatial id equals to 80128018 is (804500, 802400).

2.4 Analysis of the Two Index Schemes

Theoretically, investigating the equations given in Sections 2.2.1 and 2.3.1, we found that the expense to calculate the spatial index number using scheme 2 is simpler and faster than using scheme 1. It is because there is no need to calculate the relative location in map scales 1:20K, 1:10K and 1:5K using scheme 2, which is a substantial reduction of calculation cost in mobile device.

Empirically, a test is conducted in a mobile device against two sets of small tile images database, which is created using the scheme 1 and scheme 2 algorithms. The test results revealed that the retrieval of a record from scheme 2 table is faster than from scheme 1. The difference is ranging from 50 to 100% faster.

3. AN ADAPTIVE APPROACH TO TRANSFERRING GEOSPATIAL INFORMATION

The proposed algorithm is trying to deliver geospatial data in a systemic approach after considering the client's situation and need such as speed of client (4; 50; 70 Km/h), and moving direction (e.g., North; South; East; West). The primary idea is to partition the features geographically into a number of blocks according to the client's allowable display area and send the blocks asynchronously to the client. The blocks of features received by the client will form a dynamic moving local geospatial database on the client device. Thus an AJAX-like agent is developed to implement on the client application tier and is responsible for both interacting with the client user and communicating with mobile GIS middleware on behalf of the user.

Basically this simple approach is employing to ensure a continuous graphic area comprised of at most eleven by eleven tiles of map image is provided on the client's device for viewing as if it were seamless.

3.1 Assembling of Base Block of Objects

There are two variants for assembling the base block of objects (BBO) to form part of a dynamic moving local geospatial database: (1) even number of block assembling method, and (2) odd number of block assembling method.

3.1.1 Even number of block assembling method

The dynamic moving local database is a square block of BBO producing by assembling even number of rows and the same even number of columns, e.g. 4x4, 6x6, 8x8, etc.

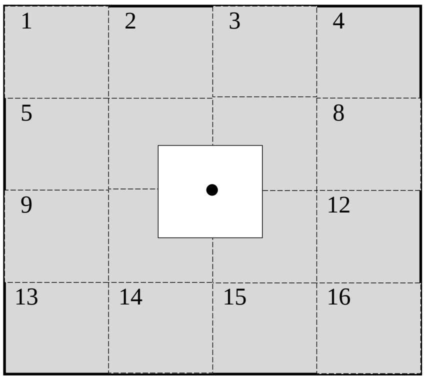


Figure 7. Assembling of 4 rows by 4 columns of BBO

Take a 4x4 assembling for illustration. In the 4x4 assembling method is a combination of 16 blocks of features. The centre of the display area (same size as BBO) is positioned at the lower right corner of block 6. In other word, the display area of the device will cover a quarter of each of the blocks 6, 7, 10 and 11 as shown in the following diagram (Figure 7).

The payload for building the initial geospatial database is the transmission of 16 blocks of BBO features. The number of blocks to be transferred in the subsequent move is depending on the updated position of the device's centre as follows:

- Position 6, 7, 10, or 11 – no block to be downloaded
- Position 2, 3, 5, 8, 9, 12, 14, or 15 – 4 blocks of features
- Position 1, 4, 13, or 16 – 7 blocks of features

3.1.2 Odd number of block assembling method

Again, this assembling method will guarantee a local dynamic moving geospatial database containing a square block of BBO features is available on the mobile client device at any time. This assembling method will guarantee the viewing area is exactly one of the constituent BBO located at the centre of the larger block. Take a 5x5 block as an example, block number 13 will be displayed on the device screen and containing the client current location as shown in Figure 8.

Again, the volume of data to be transferred is depending on two situations in order to maintain a 25-block local database.

- **Initial stage**. It is necessary to send 25 blocks of features to the client in the beginning. The transfer must start with block 13 and then follow by inner ring blocks (i.e., blocks 8, 12, 14, 18, 7, 9, 17, and 19). Finally, the remaining outer ring blocks.
- **Subsequent move**. Moving to positions 8, 12, 14, or 18 will trigger the transmission of 5 blocks of features. Whereas, moving to positions 7, 9, 17, and 19 will trigger the transmission of 9 blocks of features.

3.2 Total Number of Blocks Transferred to Form the BBO

Typically, the volume of blocks to be transferred to form the local database is depending on two situations.

- **Initial formation**. The total number of blocks to be transferred to form the BBO is n × n (where n is the width of the BBO).
- **Subsequent update of database**. The total number of blocks to be transferred to update the BBO is (i) n for movement directions in North, South, East and West; and (ii) 2 × n -1 for movement directions in NW, NE, SW and SE.

3.3 Assembling of Rectangular Buffer Strip

Rectangular buffer strip is part of the local dynamic moving database. It is the outer region of the square block of BBO. The size of this strip is varying and default to a strip of 1 block around the square block of BBO.

Typically, the total number of blocks to form the rectangular buffer strip is given by the following equation.

$$B_{RBO} = 4s(n + s) \qquad (11)$$

Where s is the strip size in block value, and n is the width of BBO in block value.

3.4 Formation of Dynamic Moving Geospatial Database

The flow of the information is carried out in two stages: (1) Initial formation of local geospatial database; and (2) Updating of local geospatial database.

1	2	3	4	5
6	7	8	9	10
11	12	●	14	15
16	17	18	19	20
21	22	23	24	25

Figure 8. A block of 5x5 BBOs

3.4.1 Initial formation of dynamic moving database

The initial formation of dynamic moving database involves the formation of BBO and RBO. The formation of BBO and RBO is the process of requesting constituent blocks from the master database, which can be resided locally on the device or remotely on a server.

The total number of blocks to be transferred is the sum of blocks for BBO and RBO as given by the following equation:

$$B_{LDB} = n^2 + 4s(n + s) = (n + 2s)^2 \qquad (12)$$

The SQL select statement to extract the relevant blocks for BBO and RBO from the master database is listed as follows:

```
SELECT PKUID_s2, TileImage
FROM basemap_image_06071011
WHERE PKUID_s2 IN ( list_of_ids )
ORDER BY PKUID_s2 ASC
;
```

42	43	44	45	46	47	48
35	36	37	38	39	40	41
28	29	30	31	32	33	34
21	22	23	24	25	26	27
14	15	16	17	18	19	20
7	8	9	10	11	12	13
0	1	2	3	4	5	6

Figure 9. Arrangement for 1-dimensional array storing IDs of BBO

23	24	25	26	27	28	29	30	31
21								22
19								20
17								18
15								16
13								14
11								12
9								10
0	1	2	3	4	5	6	7	8

Figure 10. Arrangement for 1-dimensional array storing IDs of RBO

Since we are employing the scheme 2 spatial index numbering method, there are two characteristic for this scheme: (1) same row of blocks will have the smallest ID located on the left hand side and will have the largest ID located on the right hand side; (2) the ID for a block in the higher row is larger than the ID in the lower row.

Based on the characteristics of this scheme, we can express the ID information and their spatial relationship with two 1-dimensional arrays for BBO and RBO as illustrated in Figures 9 and 10.

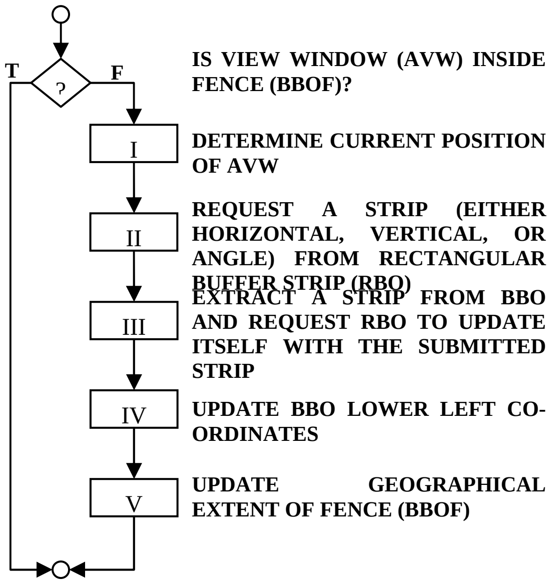

T
F
?
IS VIEW WINDOW (AVW) INSIDE FENCE (BBOF)?
I
DETERMINE CURRENT POSITION OF AVW
II
REQUEST A STRIP (EITHER HORIZONTAL, VERTICAL, OR ANGLE) FROM RECTANGULAR BUFFER STRIP (RBO)
III
EXTRACT A STRIP FROM BBO AND REQUEST RBO TO UPDATE ITSELF WITH THE SUBMITTED STRIP
IV
UPDATE BBO LOWER LEFT CO-ORDINATES
V
UPDATE GEOGRAPHICAL EXTENT OF FENCE (BBOF)

FIGURE 11. ALGORITHM FOR GENERIC UPDATING OF LOCAL DB

The ID list provided to the SQL select statement is sorted in ascending order according to the characteristics of this scheme.

3.5 Generic Updating Process on Local Dynamic Database

The local dynamic database is needed to go through an updating process when the centre of the view window is touching or exceeding the fence boundary. The generic algorithm for updating local database is carried out as shown in Figure 11. The RBO updating process as mentioned in Task III of the algorithm is a separate process that needs interaction with the master geospatial database to fetch new base blocks. The detailed updating process is varying depending on the relative position of the view window with respect to the fence boundary.

4. IMPLEMENTAION OF THE PROPOSED APPROACH ON A WINDOWS MOBILE-BASED DEVICE

The development of a Windows Mobile-based application called amGIS.Viewer.SA is employing all the concepts and algorithms described earlier. The application is using a set of DLLs customised for Windows Mobile device as illustrated in Shea and Cao (2010) as the building block for application development.

Figures 12 through 14 illustrated the execution of this application in a Dell Axim X51V pocket PC running Windows Mobile 5.0.

5. CONCLUSIONS

We proposed an algorithm to define a unique spatial index numbering system for fast retrieval of geospatial data under standard RDBMS environment on a Windows Mobile-based device without using R-Tree algorithm. The benefit of applying this unique spatial index number in fetching geospatial data over the use of R-Tree algorithm is faster retrieval of result. It is because the expensive computing cost of R-Tree is not required.

REFERENCES

Cantor, G., 1845-1918. Transfinite Numbers and Set Theory. http://www.math.utah.edu/~pa/math/sets.html (accessed 18 Aug. 2009)

Shea, G.Y.K., Cao, J.N., 2010. Use of open source programs to create a foundation for developing serious GIS application on mobile device. Proceedings of the XXIV FIG International Congress, 11-16 April 2010, Sydney, Australia.

ACKNOWLEDGEMENTS

Acknowledgement is due to The Hong Kong Polytechnic University for the financial assistance of this study. Thanks also go the Lands Department of The Government of the Hong Kong Special Administrative Region (HKSAR) for providing the 1:10000 and 1:20000 of Hong Kong digital map data in Arc/Info export format (.e00) for this project.

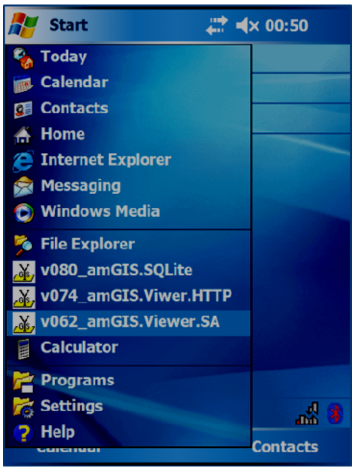
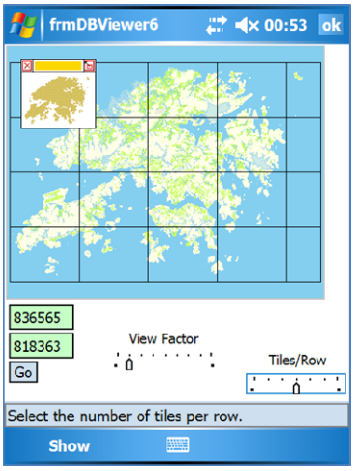
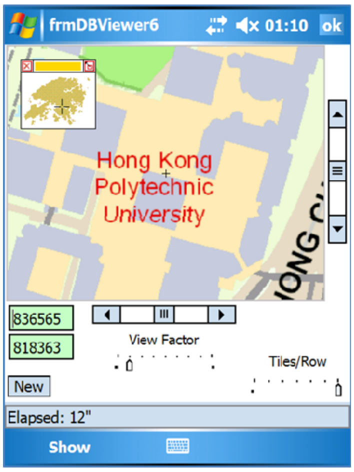


Figure 12. Start amGIS.Viewer.SA from the Start Menu



Select the number of tiles per row.

Figure 13. Select the number of tiles per row and select a point



Elapsed: 12"

Figure 14. Centre tile of the 9x9 block of image is shown

MOBILE ROUTING SERVICES FOR SMALL TOWNS USING CLOUDMADE API AND OPENSTREETMAP

Jianghua Zheng[a], Xiaoyu Chen[a], Błazej Ciepłuch[a], Adam C. Winstanley[a], Peter Mooney[a, b] and Ricky Jacob[a]

[a] Department of Computer Science, National University of Ireland Maynooth, Co. Kildare. Ireland
jianghua.zheng@nuim.ie
[b] Environmental Research Centre, Environmental Protection Agency, Richview, Clonskeagh, Dublin 14. Ireland

Commission VI, WG VI/4

KEY WORDS: routing, navigation, CloudMade API, OpenStreetMap, Location Based Services, OSM

ABSTRACT:

This research presents a practical solution for mobile routing services for small towns using open sources. Free mapping application program interfaces (API) provided by web map services, including routing services, are available to create customised map based web services combining their cartographic base data with the users own data. However, most applications focus on big cities. Location based services in small towns are generally few as many people believe there is a little demand in such areas. However, the demand of LBS applications in some small towns can be as strong as big cities, for example university towns and tourist resorts. Better location based services, especially routing services, can help strangers get familiar with the environment in a short time and lead them to places of interest. However, there are two problems to overcome for such systems. One is cost both in terms of data costs and development time. Open source data and mash-up technology could provide an answer. The other problem is the availability of suitable data of the required accuracy and detail. This is more serious as most free map services, such as Google Maps and Microsoft Bing Maps (Virtual Earth), don't provide sufficient detailed and accurate data for routing services. One feasible and economical way is to create the map ourselves and have it updated by the public. OpenStreetMap (OSM) is a free, open and fast developing map of the world. Detailed data was collected using a GPS logging device and uploaded to OpenStreetMap. The CloudMade API was used to provide multi-mode routing services together with turn-by-turn descriptions for car users, bicycle riders, and pedestrians. This solution is relatively easy and fast to deploy. Maynooth, a small university town in County Kildare Ireland, was used as a test bed. A prototype navigation system was developed for mobile users using the Windows Mobile platform. The system demonstrates that a solution to detailed navigational services for pedestrians, cyclists and drivers can be economical and feasible for small towns.

1. INTRODUCTION

Routing is one of the most important services provided by Location Based Systems (LBS). Mobile routing services encompass way-finding applications for vehicle drivers, pedestrians and cyclists, delivered using mobile terminals. Much research has been carried out on mobile routing algorithms because of the complexity of application environments and variety of user requirements (Huang *et al.*, 2007; Huang and Wu, 2008). Like POI (Point of Interests) query services, mobile routing services have become more and more popular in the real world, especially those applications for vehicle drivers, TomTom car navigation systems are typical examples. NAVITIME (Japan) (Arikawa *et al.*, 2007; Zheng *et al.*, 2006) and Nokia Maps 2.0 (Dominique, 2008) are two typical examples for pedestrians (Zheng *et al.*, 2009). Some free map platforms, such as Google Maps, Yahoo Maps and Microsoft Bing Maps, include direction modules which provide turn-by-turn routing services. Google also plans to allow Android 2.0 phones to give users real-time turn-by-turn walking directions (Adhikari, 2009). However, most current mobile routing services are provided mainly with detailed content only in relatively big cities or for major streets in rural areas. Location based services for small towns are generally ignored as many people and most companies believe there is little demand in such areas and there is no need to invest in detailed data collection for such places.

However, the demand of LBS applications in some small towns may be as strong as that in big cities. University towns and some tourist resorts are typical examples of small towns where there are many visitors potentially requiring access to LBS. They do not have much time to get familiar with the places. Better local location-based services, especially routing services, can help strangers get familiar with a strange environment in a short time. The aim of this work is to provide an effective, efficient and low cost solution to providing LBS, especially routing services, for small towns and tourist resorts. The research question contains three constraints: effectiveness (containing useful data and services); lowest cost; and efficiency (fast system development and easy maintenance). A solution using CloudMade API and OpenStreetMap (OSM) is described. The work takes Maynooth, the only University Town in Ireland, as an example and we put forward a mobile routing services prototype to demonstrate the solution.

The rest of the paper is organized in four sections. Section 2 discusses why OpenStreetMap is used as a data source. The CloudMade API is described in the following section. In section 4, we provide a detailed discussion of the development and implementation of the prototype for the LBS for small towns. Finally, the paper closes with a discussion of conclusions from the work and puts forwards some suggestions for future work. The work is part of the eCampus project, which is constructing a major testbed for StratAG, the Strategic Research Cluster in Advanced Geotechnologies (www.stratag.ie) centred at

National University of Ireland Maynooth. It focuses on constructing a campus information system including diverse location-based services.

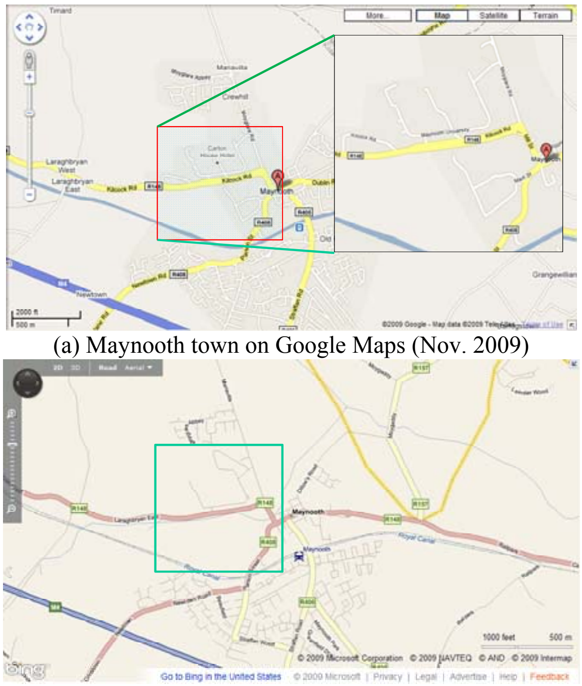
2. OPENSTREETMAP

2.1 Experimental Area and Data Collection Methods

Maynooth is a university town located in north County Kildare, Ireland. It is about 25km west of Dublin city centre. Accurate and sufficient data of the area is the basis for local mobile routing services together with POI queries. There are four major ways to obtain local spatial data and associated attribute data:

- From current data available for LBS
- From professional survey companies or agencies
- From free map providers
- Collect the data by yourself or by crowd sourcing

In most cases, the former two are not suitable, especially the second one which might cost a lot of money or have low cost performance. The third one is the most convenient. However, if we check Maynooth town, like most other small towns and tourist resorts, there is poor representation on the most popular commercial free map platforms, such as Google Maps, Yahoo Maps and Microsoft Bing Maps. Navteq and TeleAtlas, the two biggest map data companies in the world, are the map providers of those platforms and their data just focuses on the requirements of vehicle navigation. Figure 1(a) shows the map of Maynooth on Google Maps and Figure 1(b) the content on Microsoft Bing Maps.


More...
Map
Satellite
Terrain
Maynooth
Laraghbryan West
Laraghbryan East
Kilcock Rd
R148
R406
Royal Canal
2000 ft
500 m
©2009 Google - Map data ©2009 Tele Atlas
1000 feet
500 m
© 2009 Microsoft Corporation © 2009 NAVTEQ © AND © 2009 Intermap
Go to Bing in the United States
Privacy | Legal | Advertise | Help | Feedback

(a) Maynooth town on Google Maps (Nov. 2009)

(b) Maynooth town in Microsoft Bing Maps (Nov. 2009)

Figure 1 Maynooth town on web map services

The rectangles in Figure 1 show the approximate area of National University of Ireland Maynooth. Only the main streets and a few POIs of the town can be obtained from Google Maps and Microsoft Bing Maps. This poor spatial data coverage is not sufficient for establishing effective LBS routing services for this area. As purchasing commercial data is expensive, we have to collect the data ourselves. OpenStreetMap provides an outstanding example of a spatial data source for this area to which we can contribute. The next section describes the OpenStreetMap project.

2.2 OpenStreetMap

OpenStreetMap (OSM) is a free map of the entire world. It allows you to view, edit and use geographical data in a collaborative way from and for anywhere on Earth. It uses a crowd sourcing model to provide user-generated street maps. There have been various geo-wiki applications that utilise user-generated content for maps (Jacob *et al*., 2009). However, OSM is probably the most extensive and effective project currently under development (Haklay and Weber, 2008). OSM development is geographically unbalanced. In general, it is more complete in Europe. Unlike other web map services, OSM provides a set of tools to create a free editable map of the world. The maps are created using data from portable GPS devices, aerial photography, other free sources or simply from local knowledge (http://en.wikipedia.org/wiki/OpenStreetMap). It integrates some useful tools for importing, editing, exporting and generating geometry from GPS trails which encourages users to be not only users but also data generators. For this purpose, there are also some offline editing tools, such as JOSM and OSM2Go. All these tools are free and easy to use.

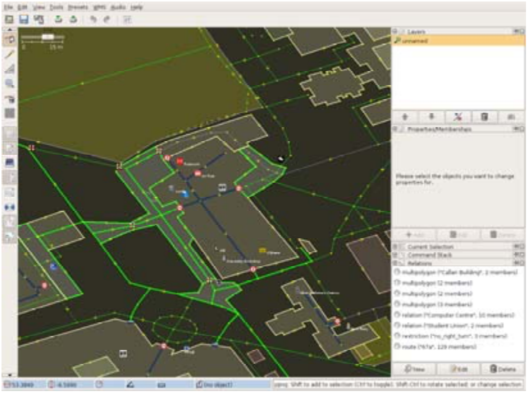


Figure 2 Interface of JOSM editor

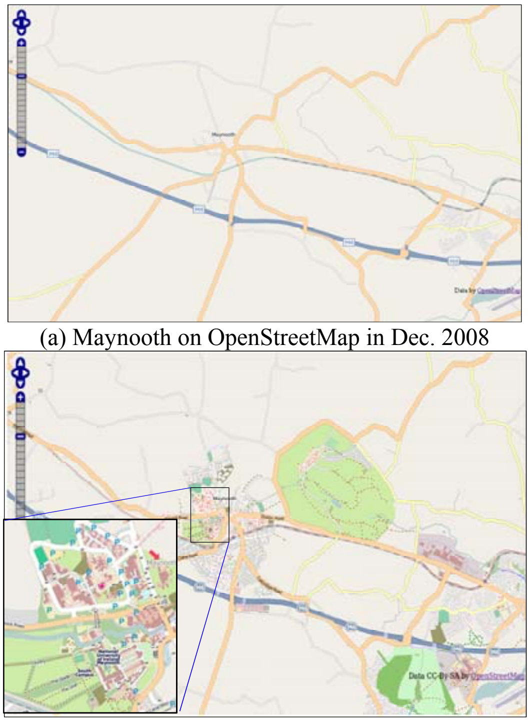
More and more organizations and individuals provide APIs using OSM data, such as for map rendering and routing such as the CloudMade Routing API. The reasons for selecting OpenStreetMap as a platform for data collection and representation can be summarised as:

- ***Totally free***. All data are generated by the public with little usage restrictions and no cost. The OSM tools and APIs are powerful but also are also free.
- ***Multiple outputs***. It is possible to render various popular formats of data of the same area from the OSM dataset giving flexibility of use.
- ***More vivid map data with various attributes***. OSM provide the public a set of powerful tools to render your own style OSM data for your personal map based applications. This attributes to its two major components, Mapnik and Osmarender.
- ***More current data***. OSM data is being updated constantly. OSM also has a mechanism for users to update local data. For example, the OSM data of Ireland is updated weekly.

OSM data can be stored and managed easily and efficiently using PostgreSQL/PostGIS, a powerful open-source spatial database management system, making it easy to create new applications.

2.3 Data Creation

The OSM data collection of Maynooth town was done by members of our research group and student and volunteer helpers, mainly using GPS logger devices, such as GlobalSat® DG-100 GPS Data Logger. Much of the roads and paths were collected by bicycle.



(a) Maynooth on OpenStreetMap in Dec. 2008

(b) Maynooth on OpenStreetMap in Oct. 2009

Figure 3. Data Creation of Maynooth town

Figure 3(a) shows there was little information of Maynooth on OSM platform in December 2008. However, now following the upload and editing of the logged data, there is abundant information of Maynooth town, including various POIs, buildings, streets, bicycle lanes, and even data inside offices of some buildings.

3. THE CLOUDMADE ROUTING API

3.1 Developing Modes for Routing

There are two typical software development strategies to construct routing procedures:

- Third party APIs
- Implementing original routing algorithms.

The first is convenient and fast for the developer to deploy applications. However, if there are special requirements, for example if people have preferences such as passing through buildings as much as possible because it is raining heavily, it becomes necessary to build more parameterised routing modules.

3.2 Why use CloudMade Routing API?

Routing algorithms are easy to realize on a well-formed link-node network (Zhan, 1997). However, the network data obtained from OSM is not suitable for direct optimal path computing. This is because the network data is generated from public submissions that may not result in a well prepared link-node network. A road will not usually be captured in segments though it might have several intersections with other roads. Rather than having to manually edit the road network, the PostgreSQL/PostGIS spatial database provides powerful and helpful standard functions to generate intersections and build the link-node network. Third-party routing APIs allow developers the possibility of quick development provided they fulfil the specific routing requirements of the application.

There are several third-party routing APIs, which could be used with OSM data, such as the CloudMade Routing API (http://cloudmade.com/, 2009) and pgRouting. Both of them are open source. The main objective of pgRouting is to provide routing functionality for PostgreSQL/PostGIS (http://pgrouting.postlbs.org/, 2009). It includes several routing algorithms such as traditional Dijkstra, A*, Shooting Star and Travelling Sales Person (TSP). Since OSM data is stored and managed in PostgreSQL/PostGIS, we can add and use pgRouting as standard functions of PostGIS. The CloudMade Routing API is currently the most used with OSM data. It provides car, foot and bicycle modes for users. It also provides turn-by-turn direction descriptions with multi-lingual templates, including Chinese. Another advantage of CloudMade is it generally serves requests more quickly than a local server does. As a whole, using CloudMade Routing API makes routing services easily and economically realized.

4. PROTOTYPE OVERVIEW

We could build the prototype with B/S or C/S architecture. The trend is B/S architecture. However, we have selected C/S architecture in our prototype for fast design and easy testing. The application is designed to run on a smart phone (HTC 3470), running Windows Mobile OS. C# is the development language for the mobile terminal. PHP is used to develop the server-side program.

4.1 System Architecture

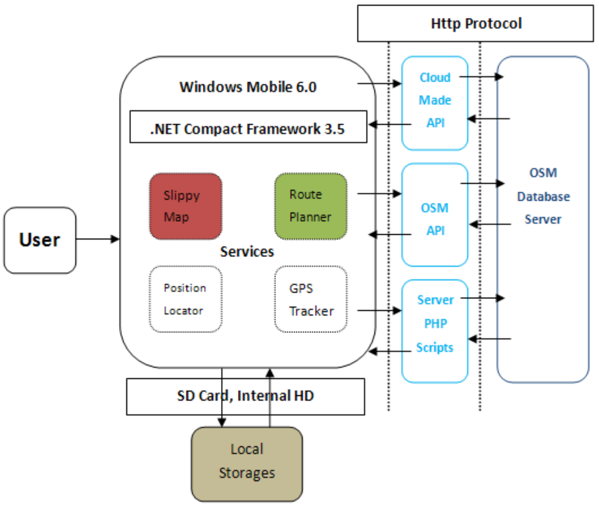


Figure 4 System Architecture of the prototype

Figure 4 shows the architecture of the prototype. The basic spatial data is on an OpenStreetMap database server and it provides map services through the OSM API. The CloudMade API provides powerful routing algorithm for both web and mobile applications based on OSM data. We provide three travel modes: car, walking and bicycle.

4.2 Slippy Map for Mobile Terminals

Fast and continuous map presentation on the mobile terminal is a basic requirement of this application. Slippy map mechanisms are widely used in map-based applications. For example, Google Maps and Microsoft Bing Maps have their own slippy map system based on commercial map databases (Haklay, 2008). There are also third party web services which provide the slippy map interface for developers, which can be used to render various map resources. Mapnik and OpenLayers are mainly used for such proposes. Unfortunately, how to implement the slippy map on a mobile device is hardly seen for developers to reference. We have implemented our own module for this.

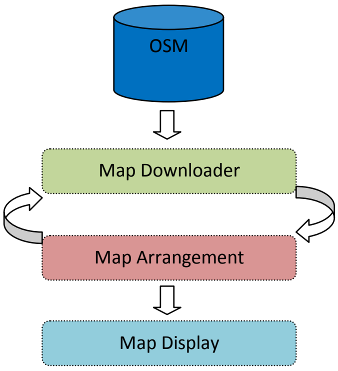


Figure 5. Main software architecture for slippy map

Figure 5 shows the software architecture for slippy map implementation. The core of the map downloader class is OSM API calls wrapped within a C# http request. OSM provides a structured URL to obtain the map tiles. The URL is a combination of zoom level, X and Y coordinates, as in the example in figure 6.

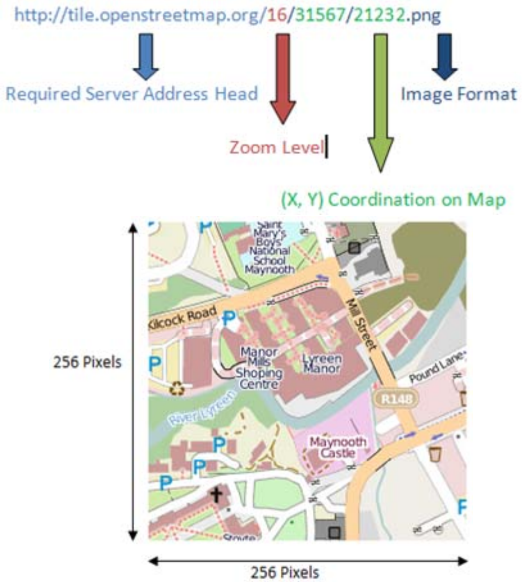
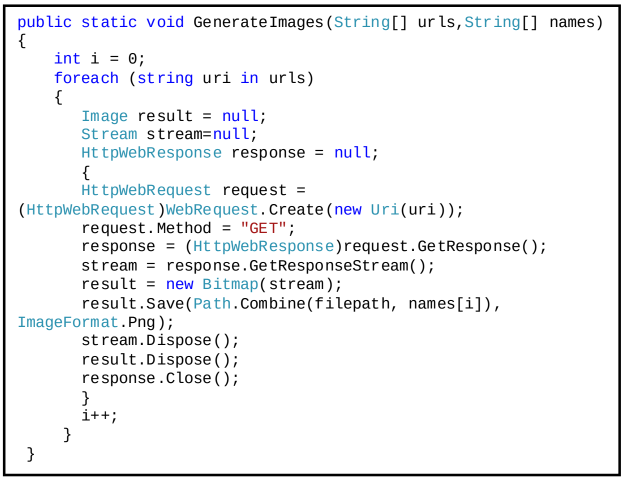

http://tile.openstreetmap.org/16/31567/21232.png
Required Server Address Head
Zoom Level
Image Format
(X, Y) Coordination on Map
256 Pixels
256 Pixels
Kilcock Road
Mill Street
Manor Mills Shoping Centre
Lyreen Manor
Pound Lane
R148
Maynooth Castle
River Lyreen
Saint Mary's Boys National School Maynooth

Figure 6. Map as rendered on mobile terminal

Figure 6 displays the map representation on a mobile terminal after the slippy map processing. C# code segment for downloading a OSM map tile is as figure 7.



```
public static void GenerateImages(String[] urls,String[] names)
{
    int i = 0;
    foreach (string uri in urls)
    {
       Image result = null;
       Stream stream=null;
       HttpWebResponse response = null;
       {
       HttpWebRequest request =
(HttpWebRequest)WebRequest.Create(new Uri(uri));
       request.Method = "GET";
       response = (HttpWebResponse)request.GetResponse();
       stream = response.GetResponseStream();
       result = new Bitmap(stream);
       result.Save(Path.Combine(filepath, names[i]),
ImageFormat.Png);
       stream.Dispose();
       result.Dispose();
       response.Close();
       }
       i++;
     }
 }
```

Figure 7. C# code segment for downloading OSM map tile

Map re-arrangement is triggered when certain events are detected, for instance a map container going out of valid region, or relocation of the map centre. Four map tiles are used to fully cover the screen of the device. If an empty space occurs during map operation, the tile positions will be rearranged accordingly. Figure 8 shows examples of scrolling resulting in the retrieval and addition of new map tiles.

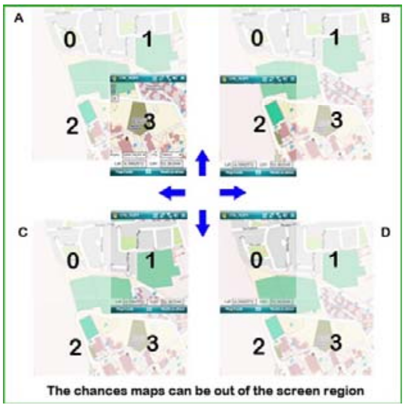
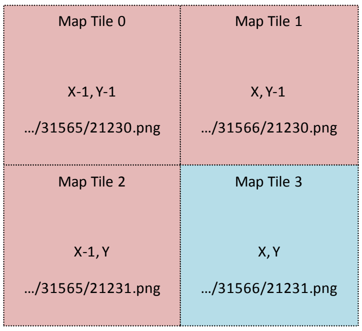


Figure 8. Four situations trigger map arrangement event



Map Tile 0 X-1, Y-1 .../31565/21230.png	Map Tile 1 X, Y-1 .../31566/21230.png
Map Tile 2 X-1, Y .../31565/21231.png	Map Tile 3 X, Y .../31566/21231.png

Figure 9. Positional relationship of tiles in OSM

Figure 9 illustrates the determination of which tiles to download during scrolling. Map tile 3 is assigned to be the reference tile on right-lower corner of the map container and from this the URL references of surrounding tiles can be calculated. OSM provides a formula to derivate X-Y coordinates of maps from latitude and longitude values. The formula can also work in reverse.

```
OSM Formula to derivate X Y Coordination
n = 2 ^ zoom
xtile = ((lon_deg + 180) / 360) * n
ytile = (1 - (log(tan(lat_rad) + sec(lat_rad)) / π)) / 2 * n
```

```
C# Implementation of OSM Tile Name Derivation Formula

public PointF WorldToTilePos(double lon, double lat, int zoom)
{
    PointF p = new Point();
    p.X = (float)((lon + 180.0) / 360.0 * (1 << zoom));
    p.Y = (float)((1.0 - Math.Log(Math.Tan(lat * Math.PI / 180.0) +
                1.0 / Math.Cos(lat * Math.PI / 180.0)) / Math.PI) /
2.0 * (1 << zoom))
    return p;
}
public PointF TileToWorldPos(double tile_x, double tile_y, int zoom)
{
    PointF p = new Point();
    double n = Math.PI - ((2.0 * Math.PI * tile_y) / Math.Pow(2.0,
zoom));
    p.X = (float)((tile_x / Math.Pow(2.0, zoom) * 360.0) - 180.0);
p.Y = (float)(180.0 / Math.PI * Math.Atan(0.5 * (Math.Exp(n) -
      Math.Exp(-n))));
    return p;
}
```

Figure 10. OSM derivation formula

The two C# functions in figure 10 can be used to construct the URL for a map tile. For example, test data (latitude= 53.381646, longitude= -6.582667, zoom= 16) here is chosen to input into C# function WorldToTilePos(). Point (31569, 21232) is returned after execution. From this the URL for this map tile is http://tile.openstreetmap.org/16/31569/21232.png . If this tile is the reference tile in slippy map system, URLs for rest of tiles are as in Figure 11.

Map Tile 0 http://tile.openstreetmap.org/16/31568/21231.png	Map Tile 1 http://tile.openstreetmap.org/16/31569/21231.png
Map Tile 2 http://tile.openstreetmap.org/16/31568/21232.png	Map Tile 3 http://tile.openstreetmap.org/16/31569/21232.png

Figure 11. URLs of four tiles in Slippy Map System

Displaying the map is the final and simplest of the three components of slippy map. The image in the control area is changed according to the required arrangement of the map tiles. After the required tiles are downloaded, an image merging function is used to join them together in memory in order to fit into map container and then render the data to the screen. The advantage of doing this is to create a clean and easy drawing environment for other map functions, like route planning and image icon display.

4.3 Routing using CloudMade API

The CloudMade Routing uses an http protocol. The structure of the URL is similar to the OSM tile querying URL, but instead of returning a PNG image, a GPX or JS format file is returned.

```
http://routes.cloudmade.com/YOUR-API-KEY-GOES-HERE/api/0.3/start_point,[transit_point1,...,transit_point N],end_point/route_type[/route_type_modifier].output_format [?lang=(en|de)][&units=(km|miles)]
```

Figure 12. CloudMade Routing planning URL structure

Figure 12 shows the CloudMade route planning URL structure. The general parameters in the structure are "start_point", "end_point", "route_type" and "output_format". Other parameters are optional.

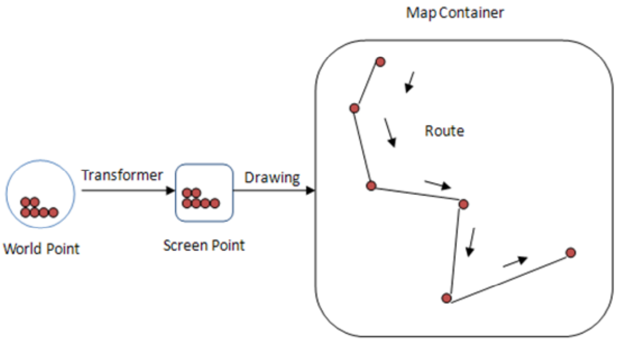
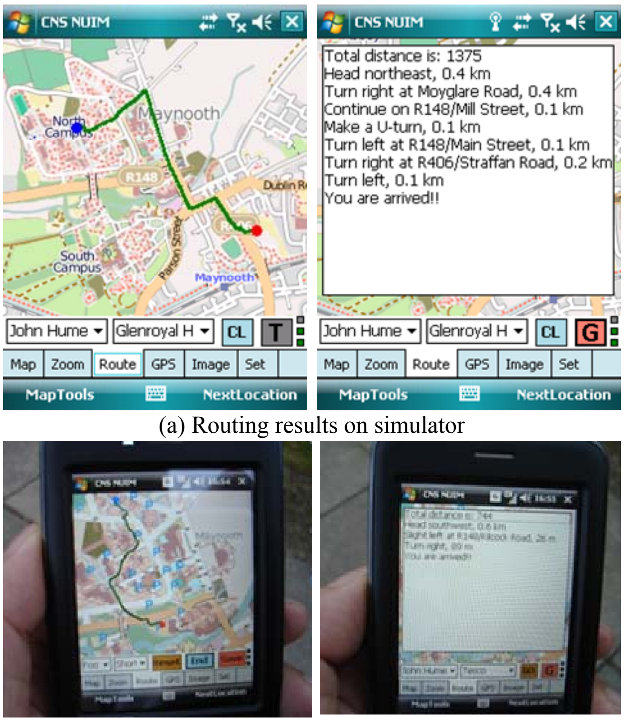


Figure 13. Drawing route on map of mobile phone

Figure 13 displays the main steps to draw a route on the map displayed on the mobile terminal.

4.4 User interfaces of the prototype

Figure 14 shows examples of the map and textual routing interfaces on both the API simulator and as actually rendered on the mobile terminal.



Total distance is: 1375
Head northeast, 0.4 km
Turn right at Moyglare Road, 0.4 km
Continue on R148/Mill Street, 0.1 km
Make a U-turn, 0.1 km
Turn left at R148/Main Street, 0.1 km
Turn right at R406/Straffan Road, 0.2 km
Turn left, 0.1 km
You are arrived!!

(a) Routing results on simulator

(b) Routing on HTC 3470 model

Figure 14. Routing interfaces

5. CONCLUSION AND FUTURE WORK

The demand of LBS applications in some small towns might be as strong as big cities; university towns and some tourist resorts as examples of such small towns where there are many tourists and people potentially requiring access to LBS. This paper describes an effective, low cost and efficient solution of mobile routing services for such small towns. The core parts of the solution are a spatial database platform (OpenStreetMap and PostgreSQL/PostGIS) and a routing module (CloudMade Routing API). These are all open source products. The work also discussed a successful method to display slippy maps on mobile terminals. By demonstrating a prototype of a mobile routing service for Maynooth town, we have showed how this could benefit more small towns.

The interfaces and response speed should be optimized in the future. During the routing process, landmarks are helpful for pedestrian users to make certain that he is walking in the right direction (Millonig and Schechtner, 2007; Hile *et al*., 2009). We plan to carry out some landmark based applications, such as using geotagged photography, as the extension of the solution described in this paper.

ACKNOWLEDGEMENTS

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

REFERENCES

Adhikari R., 2009. Android 2.0 Phones Get New Google Nav App, http://www.technewsworld.com/story/Android-20-Phones-Get-New-Google-Nav-App-68496.html (accessed 30th Oct. 2009)

Arikawa M., *et al*., 2007. NAVITIME: Supporting Pedestrian Navigation in the Real World. *Pervasive Computing*, 6(3), pp. 21-29

Bonte D., 2008. The Mobile World Congress 2008: Pedestrian Navigation at Last. 11 Feb. 2008, Barcelona, Spain. http://www.abiresearch.com/Blog/Telematics_Blog/474 (accessed 28 Oct. 2009)

Jacob R., 2009. Campus Guidance System for International Conferences Based on OpenStreetMap. In: *LNCS: Web and Wireless Geographical Information Systems,* Springer, Berlin / Heidelberg, German, Volume 5886/2009, pp. 187-198

Haklay M. and Weber P., 2008. OpenStreetMap: User-generated street maps. *Pervasive Computing,* 7(4), pp.12-18

Hile, H., *et al*., 2009. Landmark-Based Pedestrian Navigation with Enhanced Spatial Reasoning. In: *LNSC: Pervasive Computing*, Springer, Berlin/Heidelberg, Volume 5538/2009, pp. 59-76

Huang, B., Wu, Q., and Zhan, F. B., 2007. A shortest path algorithm with novel heuristics for dynamic transportation networks. *International Journal of Geographic Information Science*, 21(6), pp. 625-644.

Huang, B. and Wu, Q., 2008. Dynamic Accessibility Analysis for Location Based Service Using an Increment Parallel Algorithm. *Environment and Planning B*, 35(5), pp. 831–846.

Millonig A. and Schechtner K., 2007. Developing Landmark-Based Pedestrian-Navigation Systems. *IEEE Transactions on Intelligent Transportation Systems*, 8(1), pp. 43-49

Zhan F B., 1997. Three Fastest Shortest Path Algorithms on Real Road Networks. *Journal of Geographic Information and Decision Analysis*, 1 (1), pp. 69−82

Zheng J.H., *et al*., 2006. Study on Data Organization of Personal Navigation Services, *Computer Engineering*, 32(24), pp. 41-47

Zheng J.H., *et al*., 2009. Spatial characteristics of walking areas for pedestrian navigation. In: *Proceedings of the 2009 Third International Conference on Multimedia and Ubiquitous Engineering,* IEEE, Piscataway, NJ, USA, pp.452-458

A MULTI-MODAL ROUTE PLANNING APPROACH WITH AN IMPROVED GENETIC ALGORITHM

Haicong Yu*, Feng Lu

State Key Laboratory of Resources and Environmental Information System,
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences,
No. 11A, Datun Road, Chaoyang Dist., Beijing 100101, P. R. China - (yuhc, luf)@lreis.ac.cn

KEY WORDS: Route Planning, Multi-modal, Multi-criteria, Genetic Algorithm

ABSTRACT:

The purpose of multi-modal route planning is to provide the traveler with optimal, feasible and personalized route between origin and destination, which may involve public and private transportation modes. The strategy driven approach (i.e. routing by certain predefined transfer order) is useful but can hardly provide free combination of multiple travel modes and some feasible results may be consequently missed. A genetic algorithm (GA) is proposed in this paper to solve the multi-modal route planning problem. Variable length chromosomes with several parts (subchromosome) are utilized to represent routes in multi-modal travel environment, where each part describes a kind of transportation mode. Crossover and mutation operators are redefined in single mode; two new operators, hypercrossover and hypermutation, are defined as inter-mode operation. A multi-criteria evaluation method using a p-dimensional vector to represent multiple criteria is adopted in the fitness function for selecting the optimal solutions. The experimental results show a various mode combination, and some results conform experience well.

1. INTRODUCTION

Multi-modal transportation systems are designed to provide various travel approaches. However, problems yield accompany with the convenience of so many travel mode choices including public and private transportation. How to make a sensible decision among these travel modes or the combination of certain kinds is becoming a Gordian knot. The multi-modal route is defined as a route involves two or more different modes among which the traveler has to make a transfer (van Nes, 2002). Free combination of different travel modes is an essential characteristic. However, rare researches are concentrating on it. With the consideration of free mode combination, the time consumption of strategy drive approach is a hard burden. How to combine different modes to meet the travelers' individual needs with least computational cost is a bottleneck of optimal multi-modal route planning.

Route planning problem is used to be represented as an optimal problem which is aiming at providing feasible route under certain needs. In single criterion era (the shortest path problem), Dijkstra algorithm has been considered as a representative solution. It is an exact algorithm which always determines the exact shortest one. However, real-world optimization problems can hardly be expressed with just one criterion, one result neither. When considering multi-criteria, conflict appears often, that is, improve in one criterion could lead to deteriorate in another. The exact algorithms could not handle it well. As genetic algorithms (GA) handle a set of solutions simultaneously and have multiple various solutions during one process, it is an effective algorithm to deal with multi-criteria optimal problems.

This paper presents a multi-modal route planning approach utilizing an improved genetic algorithm to solve the multi-modal route planning problem. Variable length chromosomes with several parts (subchromosome) are utilized to represent routes in multi-modal travel environment, where each part describes a kind of transportation mode. Crossover and mutation operators are redefined in single mode; two new operators, hypercrossover and hypermutation, are defined as inter-mode operations. A p-dimensional vector representing multiple criteria with the concept of *dominate* is adopted for selecting the optimal solutions.

The remainder of this paper is organized as follows. Some related work introduced in Section 2. Section 3 describes the problem we are concentrating on. Then a multi-modal network model is introduced, and an improved GA for multi-criteria route planning approach is proposed in Section 4. The implementation with experiment and result are shown in Section 5. Finally, conclusion and future work is introduced.

2. RELATED WORK

In the past decades, a growing number of multi-modal route planning systems are available either on the internet or on desktop which appear as spatial decision support tools to provide available travel suggestions for travellers. Most of the online route planning systems (e.g. Google Map, Bing Map, *etc.*) support multiple travel modes. Nevertheless, few could provide free combination of travel modes.

Utilizing Genetic Algorithm to solve route planning has been reported for years. Gen and Cheng *et al.* (1997) proposed a priority-based encoding method to represent all possible paths in a graph. The chromosomes are of the same length, and the encoding is also complex, but their work provided a new approach to such kinds of difficult-to-solve problems. Delavar

* Corresponding author.

and Samadzadegan *et al.* (2001) proposed a genetic algorithm with a part of an arterial road regarded as a virus to select route to a given destination on an actual map under a static environment. They generated a population of viruses in addition to a population of routes. Crossover and virus theory of evolution were used to improve the rate of search, but mutation did not used. The length of the route was taken as fitness. Davies and Lingras (2003) presented a GA based strategy to find the shortest path in a dynamic network, which adapted to the changing network information by rerouting during the course of its execution. In their case the problem was no longer the shortest path, but the shortest walk. Crossover operator was used when walking condition changed to bad and spliced the new walk into the old one. All these researches concentrate on single criterion problems.

In multi-criteria research, Chakraborty (2004, 2005) proposed a GA based algorithm with a novel fitness function for simultaneous multiple routes searching for car navigation to avoid overlap. The alternate routes considered some attributes such as distance, number of turns, passing through mountain and passing by the side of river, and use penalty for fitness. In the work of Huang and Cheu, *et al.* (2004), GA introdueced for determining the weights of different criteria, which eventually achieve a series value of each criterion and sum the up as the final cost. Hochmair, H. H. (2008) utlized GA for Pareto Optimal route set searching in order to reduce the number of route selection criteria. In this work, fitness function was not included. Instead, the distinct solutions were kept. Random walking was used to generate initial population in single criterion and two parents to create a mutated offspring and replaces one parent approach is not as the traditional. However, the specific detail is not introduced. The GA based solution for multimodal shortst path problem presented by Abbaspour, R.A. and Samadzadegan, F. (2009) shown the robustness of this approach through an experiment and concluded that proposed algorithm can efficiently explore the search space to find the shortest multimodal path. Two criteria, length and waiting time, were summed in order. However, the time cost calculation is time-table based and estimated. Nevertheless, these researches are useful exploration.

For multi-criteria problem, some label setting algorithms (e.g. Martins, 1984; Corley and Moon, 1985) and label correcting algorithms (e.g. Skriver and Andersen, 2000) are reported. However, these labeling methods may cause exponential running time. As a high efficient search strategy for global optimization, genetic algorithm demonstrates favorable performance on solving the combinatorial optimization problems. GAs for dealing with multi-criteria optimal problem has been considered as an effective approach in Multi-objective Optimization. It can process a set of solutions simultaneously, and obtain various effective ones. According to these, we introduce GA in multi-modal route planning field considering multiple evaluation criteria in order to provide optimal alterable routes with free combination of travel modes.

3. PROBLEM DESCRIPTION

The single mode road network is often represented as a directed graph model $G=(V,E)$, in which $V=\{1,...,n\}$ represents a set of nodes *n*, and *E* is the set of direct arcs. Each arc is denoted by a pair (edge) $e=(i,j)$ connecting node *i* to node *j*, $\forall i,j\in V$. Weight of each edge $w(e)$ is assigned on *e* in graph *G*. A route $R(s,t)$ in the road network between two distinct nodes is defined as $R(s,t)=\{s=i_1,(i_1,i_2),i_2,...,i_{j-1},(i_{j-1},i_j),i_j=t\}$. The length of a route from *s* to *t* is calculated as the sum of the weight that assigned on edges along $R(s,t)$. The shortest path problem is to get the minimal value of $\sum w(e_{i,j}),e_{i,j}\in R(s,t)$.

In the multi-modal case, *G* can be considered as the union of multiple subgraphs, representing different travel modes respectively. The graph is not simple, that is for any two nodes n_1, $n_2\in V$ there may be more than one edge. In a three modes involved travelling network, $G=G_D\bigcup G_P\bigcup G_W$ where $G_D=(V_D,E_D)$, $G_P=(V_P,E_P)$, $G_W=(V_W,E_W)$ represent driving network, public transportation mode (i.e. bus, subway, tram etc.) and pedestrian walking. The subgraph nodes V_D,V_P,V_W are defined with $V_{DW}=V_D\bigcap V_W\neq\varnothing$ as well as $V_{PW}=V_P\bigcap V_W\neq\varnothing$, which describe the connection among different transportation modes. Similarly, E_{DW},E_{WD} define the transfer between driving mode and walking mode; E_{PW},E_{WP} between public mode and walking mode.

Associated weight of each edge $w(e)$, in multi-criteria environment is represented as a p-dimensional vector of criteria $C(e)=(C_1(e),C_2(e),...,C_P(e)),e\in R(s,t)$. The value of any criterion $k\in(1,...,p)$ for the given route $R(s,t)$ is defined as $C_k^{R(s,t)}=\sum_{e\in R(s,t)}C_k^e$. So the optimal problem can be stated as $\min C^{R(s,t)}$. Criteria such as time, transfer, fare etc. can be described together in multi-criteria problem.

4. MODELING APPROACH

4.1 Data Modeling

Before introduce the proposed GA, the data model has to be declared in advance, for the later evolution operators depend on it. The multi-modal network used in our research is organized on the concept of subgraph. The subgraph approach is a widespread method considering several existing single modes independent representation, which lays the gaps for modeling mode transfer, and further makes the route guidance inaccurate and unclear. In order to remain the independence of each network and maintain the connectivity of each other simultaneously, transfer nodes and transfer links are proposed. Note that transfer nodes are those belonging to or could be attached to both source and target travel modes. In multi-modal network each mode is represented as a horizontal layer, while transfer links or nodes connect each of them. The transport modes concerned in our research involve both public and private transportations. Driving, bus, subway and walkway networks are established separately.

Road network using navigation dataset consists of roadways of different classes, and is represented as a directed graph. In this road network, *Node* denotes the intersection, roadway start and end, entry and exits. *Roadway Section* states a directed path between two neighbor nodes. A series of *Vertex* constitute a *Roadway Section*. *Roadway* is a part of a road over which vehicles travel. One *Roadway* contains several *Roadway Sections*. The start and end of a trip as well as the transfer node could be any point of interest (POI), bus stops and subway stations. *Nodes* and *Roadway Sections* have been attached with *ID* attribute for identification, which would be used in transfer relationship representation. *Roadway Sections* also have *Fnode*, *Tnode* attributes to express the direction from Node *i* to Node *j*.

A feasible pedestrian walking network should comprise walking facilities such as crosswalk, overpass, underpass etc. In our earlier work (see Yu, H. and Lu, F. 2009) walking network is automatically built based on the road network data and the walking concerned facilities data with spatial manipulation and semantic analysis on the involved dataset. It has similar attributes as road network. Besides, the associated roadway ID has been taken as an attribute, which could make connection between these two modes. The transfer defined here is critical as they provide an access to other travel modes and connect them together.

Bus network encompasses different bus routes (such as Bus No. 839) and is built with dynamic segmentation technology. The opposite direction of the same bus route is defined as two bus lines and assigned with different *ID* value. Bus stops are assorted as physical stops and logical stops. The former denotes the actual bus stop location with ID as its only identity. Physical stops with the same stop name but different direction are identified separately. The latter denotes the logical relationship among physical stops. Through logical stops, we can identify stops with the same coordination, name, bus line assigned, and roadway ID attached. By doing this, the relationship of bus network and road network has been established. That is, these two modes could make transfer at bus stops. So does the walking network.

Subway network is represented as an undirected graph with timetable-based attributes. The node here is represented as subway station, which is the access to transfer to other modes. Each subway station has calculated the available roadway ID, walkway ID and bus stops for transfer. These are one to more relations.

In our research, POIs are taken as the origin and destination spots and the transfer nodes, which could be easily attached to the modes mentioned above. For most of the time, transfer links defined as a virtual one. In order to represent the transfer cost between modes, transfer links also assigned weight of everyone the same as those defined in single mode. This definition is useful in calculating routes in multi-modal network, such as waiting time, transfer delay and so on.

Note that the multi-modal network is independent in physical, which can be maintained and managed separately, but connected in logical, which can be easily rebuilt when single mode data updated. This approach ensures the independency and connectivity simultaneously.

4.2 The Proposed GA

We assume the reader is familiar with the simple GA. For detailed information please see Goldberg (1989). As the simple GA does not support evolution in multi-modal environment, an improved one is proposed in this paper to solve the multi-modal route planning problem. The proposed GA (Figure 1) have the similar procedure as the simple one, but differ at representation and evolution operators and using p-dimensional vector to represent environment pressure.

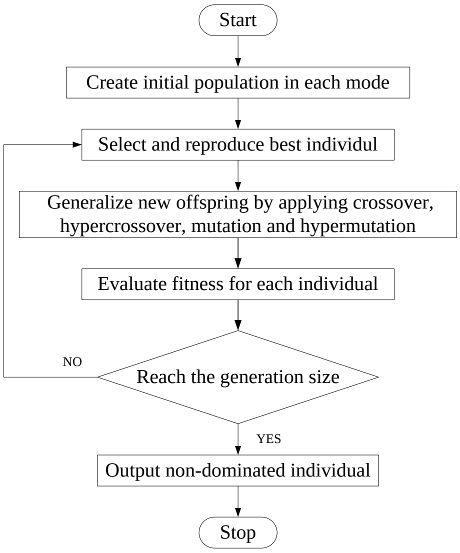

Start
Create initial population in each mode
Select and reproduce best individul
Generalize new offspring by applying crossover, hypercrossover, mutation and hypermutation
Evaluate fitness for each individual
NO
Reach the generation size
YES
Output non-dominated individual
Stop

Figure 1. Flowchart of proposed genetic algorithm

4.3 Genetic Representation

To represent multi-modal routes as genes is critical for developing a genetic algorithm. Special difficulties arise from a) a route contains multiple modes and for each mode encompasses different number of genes, b) a random sequence of edges usually does not correspond to a route (Gen and Cheng *et al*. 1997), and c) allowing evolution operators among multiple modes in one route may cause deficiency.

In this study, variable length chromosomes have been used for encoding the problem. A chromosome or an individual (route) encompasses several sequences of positive integers which represent the IDs of the representative modes and the same number of negative integers that identify the mode of the following genes. In other words, mode tag has been added in front of its genes as a negative integer (Figure 2). The precondition of this encoding approach is that the values of genes (IDs) are positive integers. The ID coded chromosome contains arc IDs in road and walking modes and node IDs in bus and subway modes. The reasons that we do not use node for all modes are as follows. First, arc feature to represent road is more reasonable in logical and so does the point feature to denote bus stops and subway stations. Second, it is easy to attach a point feature to the surrounding arc which makes modes transfer more apparent. Third, when calculating on road network, turning impedance, temporal regulation, even real-time dynamic traffic information are base on arc feature, which makes it possible to apply this approach in dynamic environment.

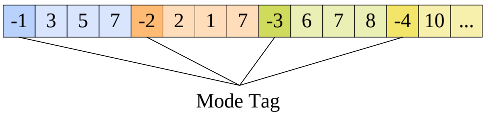


Figure 2. An example of a chromosome

4.4 Evaluation Function

An evaluation function plays the role of the environment, rating solutions in terms of their "fitness" (Michalewicz, 1996). The fitness is not a scalar value as in single criterion problem. It is represented as a p-dimensional vector, each dimension associated with one evaluation criterion. In multi-objective optimization, conflicting objectives (criteria) functions result in a set of optimal solutions, instead of one. These optimal solutions are known as Pareto-optimal solutions, that is, no reduction can be made to one of its cost components, without increasing at least one of the others. Dominate concept is using to identify the relationship of vectors. In our case, for two routes *R1* dominates *R2*, are represented as $f(R1) \preceq f(R2)$, when $C_i^{R1} \le C_i^{R2}$ for each $1 \le i \le p$ and $C_i^{R1} < C_i^{R2}$ for at least one *i*. Then, *R1* is called the non-dominated solution. Multi-objective ranking method proposed by Goldberg (1989) has been used. In this method, all non-dominated individuals are assigned as rank 1. Besides the current non-dominated ones, compare the rest individuals and assign the non-dominated as rank 2, repeat until finish all individuals. Each criterion has its own fitness function calculated independently. The rank values identify the optimal solutions.

4.5 Population Initialization

The composition of initial population is remarkably different compare to the simple GA. As we are looking forward to getting free mode combination of travel modes, the initial population has been generated in each mode respectively. Thus, the initial population encompasses four sets of sub-population in our study, and in each sub-population, random generalization is applied.

Note that in some travel mode such as subway there is not always available route from the input *Origin* to *Destination* (*O-D*). We allow these incomplete solutions but only in initial population, with the precondition that these incomplete solutions are a part or would be a component of complete ones in later evolution. One incomplete solution must generate in the possible reaching space decided by the *O-D* position and through evolution operators it could generate a complete one. This could insure the pureness and diversity of initial population.

4.6 Evolution Operators

Conventionally, evolution operations are achieved through *selection*, *crossover* and *mutation* operators. Another two operators created in this case to adjust to multi-modal environment. The *Crossover* operator and *Mutation* operator are defined in intra-modal environment, i.e., in single mode. Accordingly, *Hypercrossover* operator and *Hypermutation* operator are defined in inter-modal environment, and achieve new individuals from different travel modes.

The *Selection* operator reproduces the best individual. Those selected ones are the current optimal with the lowest rank value, i.e. the Pareto-optimal solutions, which guarantee the elite genes keeping in the next generation. On other hand, to avoid large numbers of similar solutions, individuals with the same chromosome are taken as a “single” one. If the count of “single” non-dominated does not satisfied the purpose selection size (decided by the selection probability), the next rank value has to be taken into account, until finish the target. On the country, random select the target number of individuals into the next generation.

Single point crossover strategy is being utilized in both *Crossover* operator (Figure 3) and *Hypercrossover* operator (Figure 4). For two randomly selected (according to crossover probability) individuals (parents): first, detect all the possible operational modes (with the same mode tag), and select one randomly as the current crossover mode; then, for the current operation mode utilize the corresponding single point crossover operation; if failed, return to select other modes; if all failed turn to *Hypercrossover*. The detailed procedures of single point crossover are as follows:

a) Detect all the candidate crossover genes;
b) Select a pair of candidates and make crossover;
c) Loop detection and repair.

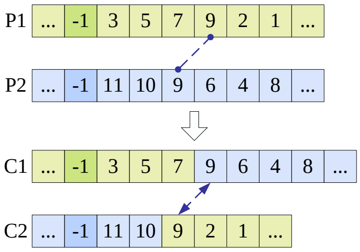

P1 ... -1 3 5 7 9 2 1 ...
P2 ... -1 11 10 9 6 4 8 ...
C1 ... -1 3 5 7 9 6 4 8 ...
C2 ... -1 11 10 9 2 1 ...

Figure 3. Crossover operator

The detected candidate crossover genes are those with the same gene value. One point crossover could generate loops or other illegal path, which have to be eliminated.

Hypercrossover operator is more complex than the above one. The combination mode set formed by the permutation of all the modes of each parent individual. Determination of the current mode combination is achieved by random selecting. Then, crossable candidates are detected and selected one pair. This function execute on the bases of the relationship among multiple modes, which established in Network Modeling Section. The detected crossable candidates are those with mode transfer relationship or within certain distance (walking) or have some connection in attribute, which could make a transfer. Besides loop path, incomplete solutions could be generated (a certain gap between two crossover genes). Thus, gene repair is used to amend this kind of problem by adding appropriate genes in certain mode, such as walking.

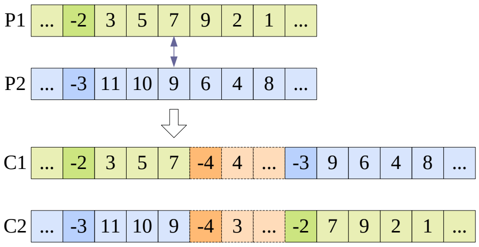

P1 ... -2 3 5 7 9 2 1 ...
P2 ... -3 11 10 9 6 4 8 ...
C1 ... -2 3 5 7 -4 4 ... -3 9 6 4 8 ...
C2 ... -3 11 10 9 -4 3 ... -2 7 9 2 1 ...

Figure 4. Hypercrossover operator

The *Mutation* operator and *Hypermutation* operator (Figure 5, 6) are similarly defined in different environment. Unlike the crossover operators, mutation operators are unary operations, that is, with one individual each time. In mode detection step, the length of the mode (the size of genes), has to be taken into account, and neglected the too short one, because those one can hardly make operation. For the random selected two mutable genes, another route (genes) is generated and replaces the original one. Make sure that the lengths between the two selected genes are long enough to generate another path. In some mode such as subway, mutability detect is recommended, for there’s not always other route existing between two stations.

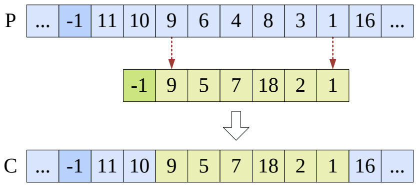


Figure 5. Mutation operator

Hypermutation operator is designed to achieve performance improvement in at least one criterion. Therefore, each mode has defined certain target mutation modes to reduce cost in some criterion. Walking mode is seldom taken as target mode except the select genes are within walking distance, and fulfils the probability. Like hypercrossover operation, gene repair method is mostly used, for there would be many gaps in inter-mode transfer or intra-mode transfer (bus transfer).

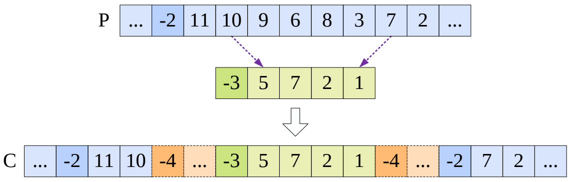


Figure 6. Hypermutation operator

5. EXPERIMENT AND RESULT

To evaluate the performance of the proposed approach, we implement the improved GA in our multi-modal, multi-criteria route planning system. Data utilized in this prototype comprise detailed road network navigation dataset with 53,997 *Roadway Sections* (arcs) and 33,402 nodes, 786 bus lines with 23,590 logical bus stops organized by dynamic segmentation technology, 7 subway routes with 113 subway stations, and walkway with 52,509 arcs and 33,225 nodes. And 839 POIs distribute in Beijing downtown. Those data stored separately as arc, node or point features. Network topology built on each single mode firstly; then, the topology of the multi-modal network built according to the relationship among various travel modes.

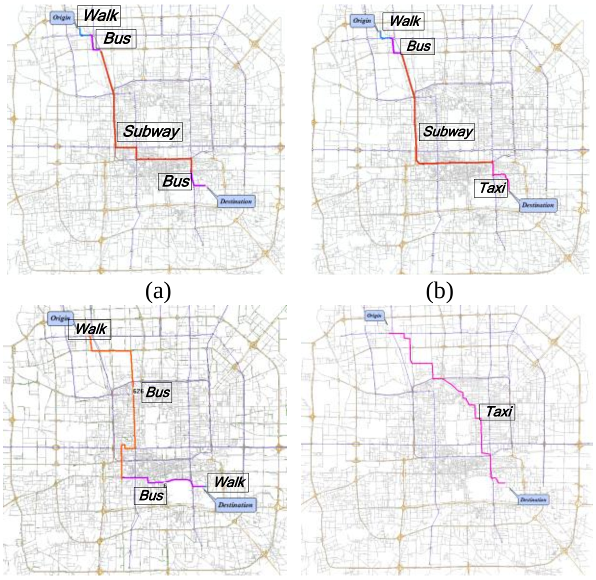
In our experiment, the initial population size is 100, with 13 taxi routes, 40 in bus and subway each, and the rest in walking mode. Table 1 lists the parameters of the proposed GA. The testing *Origin* and *Destination* (*O-D*) are located in the northwest and southeast of Beijing with Euclidean distance over 1,300m.

Parameter	Value
Initial population size	100
Selection size	10~20
Crossover probability	0.25
Hypercrossover probability	0.2
Mutation probability	0.25
Hypermutation probability	0.2
Max generation number	30

Table 1． Parameters of proposed GA

Evaluation criteria vector is calculated as the cost of each route. Criteria considered are travel time, fare and transfer numbers. The shortest path is not taken because in practical environment it could not lead to direct value representation as the other criteria. Time cost encompasses different modes. In pedestrian walk, length of connected edges and average walking speed (5 km/h) are included; in bus mode, besides length and velocity, the interval in bus route line and the entrance and exit platform time are considered; subway mode is schedule based, and the transfer time is counted; driving mode has the similar representation as walking, however, the velocity is fluctuated with time and provided by traffic velocity prediction on historical traffic reasoning with real-time traffic information. The fare structures of taxi, bus and subway systems are different. In Beijing, taxi fare depends on both travel distance and waiting time (velocity less than 10km/h); bus is distance and bus line related with 60% discount for Travel Card users; a fixed amount of 2 RMB is charged per trip for subway. Inter-mode and intra-mode transfers are counted together to represent transfer times including walking.

Figure 7 shows the routes provided for alternation: (a) a bus-subway-bus transfer route, with walking guidance; (b) the fastest route with bus-subway-taxi transfer; (c) a bus-bus transfer route; (d) a taxi route, which is faster but more expensive than (c).



(a) (b)

(c) (d)

Figure 7. Alternative results

Table 2 lists the routing results of each non-dominated individuals. In this table, the least time route using bus-subway-taxi mode combination, which conforms the experience well. In other mode combination, except the common bus-bus transfer, bus-subway-bus, walk-subway-bus and bus-subway-taxi transfer also achieved.

Time	Fare	Transfer	Modes
29.2’	2.8	3	WBSB
29.9’	2.4	2	WSB
48.7’	0.8	1	WB
28.2’	17.4	3	WBST
29.8’	32	0	T
183.1’	0	0	W
43.2’	0.8	2	WBB

Table 2. Routing Result List

6. CONCLUSION

The purpose of multi-modal route planning is to provide the traveler with optimal, feasible and personalized routes between origin and destination, which may involve public and private transportation modes. This paper proposed an improved GA for route planning in multi-modal, multi-criteria environment. Crossover and mutation operators are redefined in single mode; hypercrossover and hypermutation are defined as inter-mode operation. In order concerning various requirement, Vector based evaluation has been utilized to represent multiple criteria. Through applying multi-objective ranking method, the optimal solutions are provided.

This approach implements a free combination of travel modes with concerning various individual needs, which has little manipulation and more intelligence. An experiment has been conducted on the base of our multi-modal network. And the results show a various mode combination, which could adapt to different situations and the results conform experience well.

In the future, multi-objective optimization has to be studied in order to improve the evolution operation for non-dominated solutions. The current multi-objective ranking method is useful but consumes much running time. Other time consuming aspect should be studied further. The performance of the proposed approach in dynamic environment has not tested yet, which leaves a lot work to do.

ACKNOWLEDGEMENTS

This research is supported by the Hi-Tech Research and Development Program of China under grant No. 2007AA12Z241; Chinese NSFC program under grant No. 40871184; The Knowledge Innovation Project of CAS under grant No. KZCX2-YW-308. Their supports are gratefully acknowledged.

REFERENCE

Booth, J. and P. Sistla, et al., 2009. A data model for trip planning in multimodal transportation systems. In: *Proceedings of the 12th International Conference on EDBT: Advances in Database Technology*, Saint Petersburg, Russia, ACM, New York, pp. 994-1005.

Car, A. and H. Mehner, et al., 1999. Experimenting with hierarchical wayfinding. Technical report 011999, Department of Geomatics, University of Newcastle, Newcastle upon Tyne NE1 7RU, United Kingdom.

Chakraborty, B., 2004. GA-Based Multiple Route Selection for Car Navigation. *Applied Computing*. Springer-Verlag, Berlin, pp. 76-83.

Chakraborty, B., 2005. Simultaneous multiobjective multiple route selection using genetic algorithm for car navigation. *Pattern Recognition and Machine Intelligence*. Springer-Verlag, Berlin, pp. 696-701.

Chiu, D., Lee, O., and Leung, H., et al., 2005. A multi-modal agent based mobile route advisory system for public transport network. In: *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-38)-Track 3*, Big Island, HI, USA, IEEE Computer Society.

Corley, H. and Moon, I. 1985. Shortest paths in networks with vector weights. *Journal of Optimization Theory and Applications*, 46(1), pp. 79-86.

Davies, C. and Lingras, P., 2003. Genetic algorithms for rerouting shortest paths in dynamic and stochastic networks. *European Journal of Operational Research*, 144(1), pp. 27-38.

Delavar, M.R.; Samadzadegan, F. and Pahlavani, P., 2001. A GIS-assisted optimal urban route finding approach based on genetic algorithms. Dept. of Surveying and Geometrics, Faculty of Engineering, University of Tehran, Tehran, Iran, Commission II, Working Group II/5.

Foo, H. M. and Y. Lao, et al., 1999. A Multi-Criteria, Multi-Modal Passenger Route Advisory System. In: *Proceedings of 1999 IES-CTR International Symposium*, Singapore.

Gen, M., Cheng, R. and Wang, D., 1997. Genetic algorithms for solving shortest path problems, In: *Proceedings of 1997 IEEE International Conference on Evolutionary Computing*, pp. 401-406.

Goldberg, D. E. 1989. *Genetic algorithms in search, optimization, and machine learning*. Reading, Mass: Addison-Wesley Pub. Co.

Hochmair, H. H., 2008. Grouping of optimized pedestrian routes for multi-modal route planning: a comparison of two cities. *The European Information Society*. Springer, Berlin, pp. 339-358.

Huang, B. and Cheu, R. L. *et al.* 2004. GIS and genetic algorithms for HAZMAT route planning with security considerations. *International Journal of Geographical Information Science*, 18(8), pp. 769-787.

Jung, S. and S. Pramanik, 2002. An efficient path computation model for hierarchically structured topographical road maps. *IEEE Transactions on Knowledge and Data Engineering*. 14(5), pp. 1029-1046.

Martins, E., 1984. On a multicriteria shortest path problem. *European Journal of Operational Research*, 16(2), pp. 236-245

Michalewicz, Z., 1996. *Genetic Algorithms + Data Structures = Evolution Programs* (Third Edition). Springer-Verlag, Berlin.

Skriver A.J.V.A and Andersen K.A., 2000. A Label correcting approach for solving bicriterion shortest-path problems. *Computers & Operations Research* 27(6), pp. 507-524.

Yu, H. and Lu, F. 2009. A Multi-modal, Multi-criteria Dynamic Route Planning Approach with Accurate Walking Guidance. In: *11th International Conference on Computers in Urban Planning and Urban Management*, Hong Kong, China.

van Nes, R., 2002. Design of multimodal transport networks: a hierarchical approach. PhD dissertation, Technische Universiteit Delft, Delft, Netherlands.

The Extended Route Service
Based On Dynamic Update Frame: From Design To Deployment

Guo Shanxin[a], Meng Lingkui[a], Yu Wanli[b]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.-vincentradcliffe@yahoo.com.cn
[b] Graduate University of Chinese Academy of Sciences, Beijing, China.

Commission VI, WG VI/4

KEY WORDS: Route analysis, Dynamic Framework, GIS, Space-time Prism, Windows Socket, E-map

ABSTRACT:

This paper establishes a dynamic route service framework based on windows socket and .net remoting technology, which ensures the update of extended route service data in time. To improve the quality of urban travel navigation service, the route analysis is extended in several means, including the integration of route service with time information and improving the algorithm in different travel modes. Using the principles of space–time prism concept as a source reference, the problem for choosing the appropriate facility in restrained time cost is discussed. Potential path area model of space-time prism is simplified to find the appropriate facility with a high speed.

1. Introduction

Route service is one of the unique functions in e-map. It bases on the route navigation and supplies the function of Optimal Path Analysis and navigates for users (Goodchild, 1999). Different from other services of e-map, route service requires Timeliness and high accuracy map data (Alan Kwok, 2006), e.g., in the urban traffic field, real-time update traffic information is required. By ensuring the reliability of data, users obtain reliable results of the analysis. To serve the requirements of the Real-time update data in this service, a dynamic update framework was established.

The dynamic update frame is numerous interactive modes between clients and servers. These interactive modes ensure that the synchronization and symmetry of information between clients and servers in network environment. There are many ways to achieve synchronized information between clients and servers. Windows Socket communication frame is one of the solutions widely used (WeiMeng Lee, 2005c). Extra corresponding adjustments were done to make the transportation of geographical data more speediness and safety. . Net Remoting technology is a distributed calling technology, which makes up the limitations of DCOM network communication in the Internet and has the flexibility as Web Service (David Curran, 2002), therefore, .Net Remoting was used to complement the Windows Socket to complete the geographical data of the remote dynamic update is appropriate.

The route services are widely used in vehicle navigation, trip planning, logistics, transportation and other aspects. In this paper, two ways are used to achieve the dynamic expansion of route service, one is by the integration of route service with time information and the other one combines different travel modes in services.

2. The Dynamic Update Frame of Route Service

As mentioned in the introduction, the route service requires geographic data with high timeliness and accuracy, especially the road network data. In order to make a credible path to achieve the results of services, update information is transported to every client when the data has been changed. The transportation requires a series of Interactive modes between client and servers.

2.1 Two communication types between client and server

In the classic three-tier client-server mode, the communication between client and server becomes very critical. The client and server communication is divided into two categories, command communication and data communication. Command communication is established to complete the command transportation between client and server. Data communication enables the data versions in both client and server synchronized. Command communication has small size of data, but the structure of it is variable, while data communication has big size of data with little change in data structure, and it requires higher data security, e.g., in map service, many command communications between client and server are needed before the transportation of map data, and contents of each command communication are completely different. When sending map data to client, server needs to encrypt it to ensure the safety of the map. Then, client need to complete the decryption. No matter what maps the client requests, the communication format of them don’t have many changes.

2.2 The Communication Process between client and server

In order to achieve a good effect of command communications and data communications, Windows Socket is used to carry out the transportation of commands between the client and server. .Net Remoting technology was used for the transportation of map data. Figure 1 shows four logical layers between client and server.

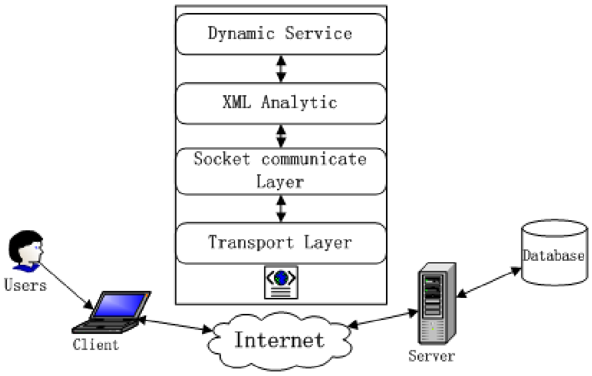


Figure 1 Client and server-side hierarchical model diagram

Both client and server include transport layer, Socket interaction layer, XML parsing layer, and dynamic services layer. Through the transport layer to carry out the transmission of map data, Socket interaction layer completes the client and server-side command interaction, XML parsing layer is an XML-based interpretation of the map data and finally maps are provided to the dynamic service application layer. Figure 2 shows the logical process when client requires map data from map server. The client publish data request to the server with Socket communications. Socket analytic server captures requests, parses requests, and transmits the request to the data server which is working for these requests. Data server receives data request, and the data is converted to XML files and through Socket server to interact with client. Data Server and the client open the corresponding data transmission channel. We transport the map data from server to client with .Net remoting technology.

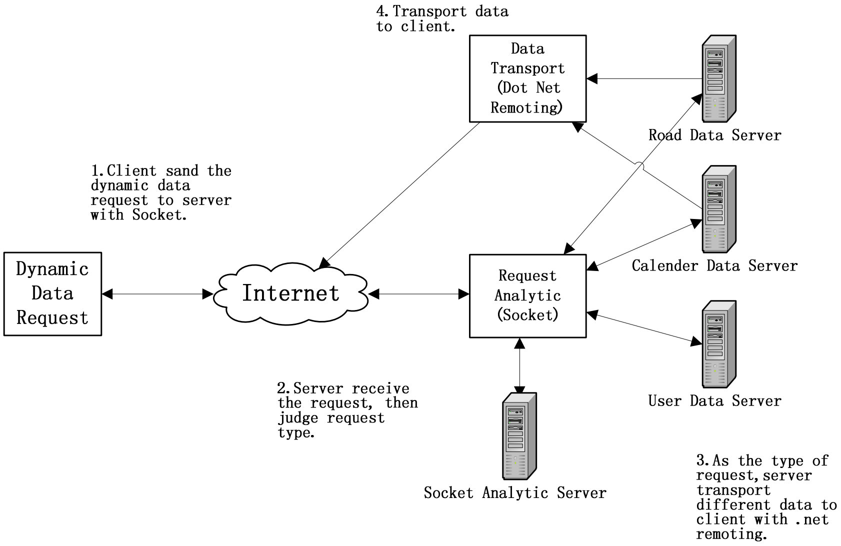

4.Transport data to client.
Data Transport (Dot Net Remoting)
Road Data Server
1.Client sand the dynamic data request to server with Socket.
Calender Data Server
Dynamic Data Request
Internet
Request Analytic (Socket)
User Data Server
2.Server receive the request, then judge request type.
Socket Analytic Server
3.As the type of request, server transport different data to client with .net remoting.

Figure 2 The logical structure diagram of client requesting for map data

In the multi-client data request process, Socket server maintains a list of users to monitor users' status, and achieves the allocation of user requests. As shown in Figure 3, when the client logs on, Socket server assigns a separate thread for each client to complete client request listening (Simon G. M. Koo, 2003). Socket server updates the user list, the user's IP address and data request, and sends request to back-end data server. Through this dynamic update framework, the synchronization of data is completed between client and server.

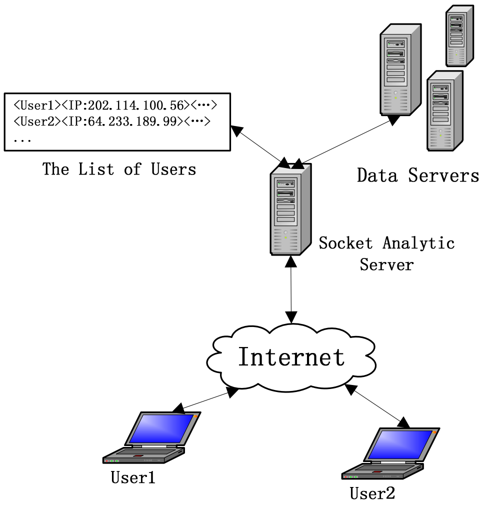


Figure 3 Multi-user data requests

3. The Extended Route Service based on time information

Adding time information into route analysis makes it closer to the real application needs. There are two aspects of this issue, integrating the result of route analysis with schedule and combine time-limited information with route service to improve the accuracy of result.

In the first aspect, using the result of route analysis improves the accuracy of schedule. To consider business hours, gate open and close time of facility with analysis, the result contains exact time information. This information can improve schedule to fit reality. In the second aspect, time-limited information is considered into analysis. The key issue is the facility choice in time-limited information.

3.1 Combine the result of route analysis with schedule

To combine the result of route analysis with schedule, two types of time was considered, the "hard" time, e.g., bank hours, open and close time of gates, and the "soft" time, including the idle time of service points and the time of road traffic jams. Hard time has direct impacts on the route analysis results, e.g., before the gate opens, the gate is a barrier in the path, so another path to get around the gate was chosen in route analysis. The bank becomes a non- reached point out of its business hours, and another solution is achieved. Soft time impacts on the time of reaching destination. No barriers or non-reached were produced by "soft" time, so it has the indirect impact on the route analysis results.

To improve the route analysis arithmetic, "hard" time information is taken into consideration. There are two solutions when "hard" time exists in a path, waiting until it opens, and choosing another way or changing the facility. Users may feel uncomfortable with the first solution, because no one likes waiting for something. The second solution might not obtain the best routes from logical view, but users feel satisfied. As shown in Figure 4 (a), a user chooses bank A as a stop point, but bank A is not open, so maybe bank B is a better choice. In Figure 4 (b), there is a gate in path A, and if it isn't 7:30am-23:00pm, the gate is a barrier, so the user will either choose to wait or path B instead. The guide line of different choices is totally time-consuming. The route analysis arithmetic makes the decision to give an optimal path. For the "soft" time, the arithmetic just needs to put this information into schedule and provide the accuracy time information in each point.

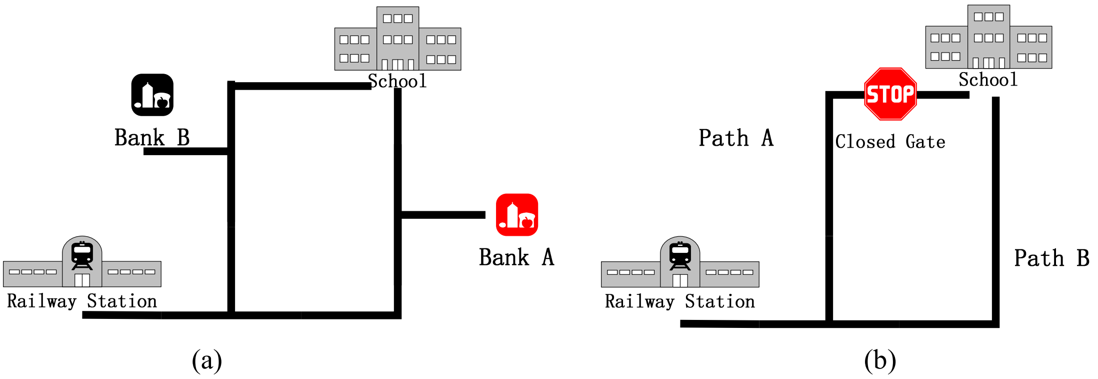


Figure 4 two solutions in route analysis with "hard" time information

3.2 Combining time-limited with route service

The space–time prism concept in Hagerstrand's time geography provides an effective model to identify the opportunities under spatio-temporal constraints (HongBo Yu, 2008; T.Neutens, 2007). Using the principles of space–time prism (Lenntorp, 1976), the problem of how to choice the appropriate facility in fixed time cost will be discussed. Figure 5 describes the space-time prism with isotropic cost road network. Here XY axis represents geographical space, and the T-axis represents time. Someone starts from T1 point and after a certain period of time reaches the end of T2. Its projection on the 2D plane is a region known as a potential path area (PPA) (Lenntorp 1976) in isotropic cost of network, the PPA is an ellipse. In 3D plane, there is a prism between the origin and ending. All possible routes were contained in this area, and the routes out of this area are forbidden in reality.

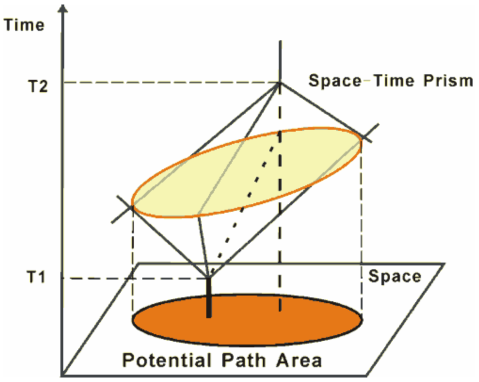


Figure 5 space–time prism models

In actual geography of the world, due to the complexity of road network data, seeking a PPA is difficult. Figure 6 shows the simplified model of PPA. In this picture, the total time between railway station and school is T, and there are three facilities around. Which one is appropriate facility to ensure the total path time less than T? If the cost of road network is isotropic, the theoretical PPA is shown by broken line as an ellipse. In the simplified model of PPA, as start point (school) and end point (train station) as the center, each with T / 2 as the radius, two service areas R and H are drawn, and this map is divided into three regions. The relationship between three facilities and two service areas is described as below. The whole region of map is U.

$$
\begin{aligned}
&A\in U-R\cup H \\
&B\in R\cup H-R\cap H \qquad (1)\\
&C\in R\cap H
\end{aligned}
$$

As the figure 6 shows, A is beyond the area of PPA, so facility A is not the appropriate one. Contrarily, C is a facility point in the PPA within the region, so facility C is the appropriate one. Region B might be in the PPA or not, so we need to screen each of these service points, retention to meet the conditions of service points. This simplified model quickly determines whether the service points meet the conditions and finds all the appropriate facilities.

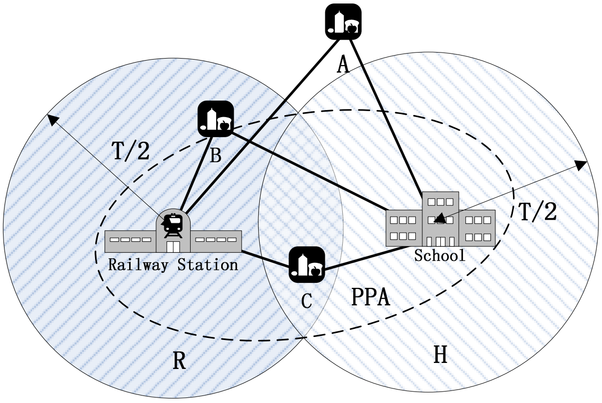


Figure 6 The simplify model of PPA

4. Path analysis model based on the integration of persons and vehicles

4.1 Vehicle Path Analysis

Numerous roads are closed to traffic on the map. In the process of vehicle navigation, vehicle barriers are dynamically loaded onto the map for analysis. Figure 7 shows the flow chart of the algorithm for vehicle navigation.

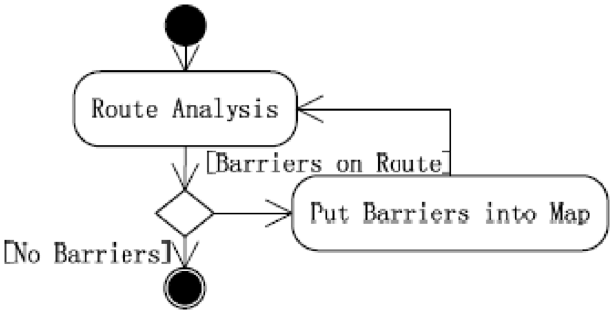


Figure 7 Vehicle Path Analysis Algorithm Process

The algorithm calculates a path after a user sets starting and ending points and checks whether any barriers exists on the path. If any barriers exist, barriers are added to the analysis. Then, the algorithm calculates a path again. Repeating the process above until there is no barrier on the path. This path is the optimal path.

4.2 Route analysis model for the integration of people and vehicles

Sometimes, the path achieved from chapter 4.1 is not the best result. In reality, the locations we get on and off are not always our starting point or destination. In order to obtain a satisfactory path analysis results in this situation, a comprehensive analysis based on the integration of people and vehicles route analysis is introduced. The Vehicle Path Analysis Algorithm showed in chapter 4.1 is improved. When the destination cannot be reached by vehicles, Path analysis model for the integration of persons and vehicles determines the appropriate off location and get the best path. Figure 8 shows the flow chart of the improved algorithm.

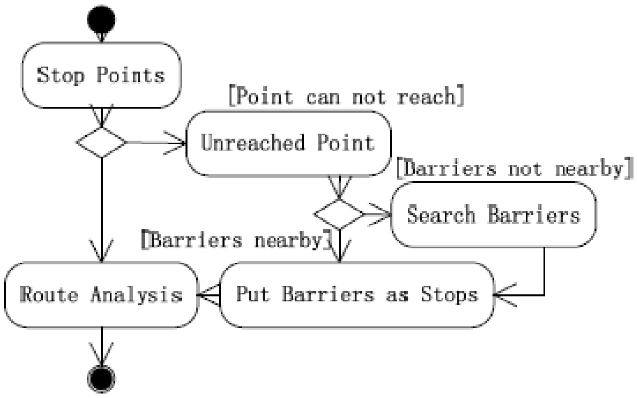


Figure 8 Algorithm Process for Path analysis model based on the integration of people and vehicles

As shown in Figure 8, whether there are any unreachable points in the target points that imported by user. If any, store these points in the queue of unreachable points. Remove the point one by one from the queue, and perform the following procedure. After being removed from the queue, the existence of barriers around the point is determined; if none, broaden the search range, and continue to search until barriers points are found and marked as on and off points. Participate in path analysis as a special target point, when the points in the unreachable points queue have performed the procedure. Then, path analysis based on the integration of people and vehicles has completed. At this point, the location of on and off will appear around the unreachable points on the path, and it is the result of the analysis for integration of persons and vehicles.

5. Experiment and Realization

Wuhan University Navigation System was designed and implemented by using ESRI ArcEngine and Wuhan University campus map data. The platform follows the dynamic-update architecture as discussed in the first section. If the traffic information changes, the server will notify each client to update data. Through a series of interactions between server and client-side, the map data was dynamically updated. The path navigation services of the platform are expanded in time and travel patterns. Users obtain precise time information of each point through the path analysis services, and the results help arrangements of their trips. Path analysis model based on the integration of persons and vehicles discussed in section 4.2 is used in the platform to provide users with more personalized navigation services.

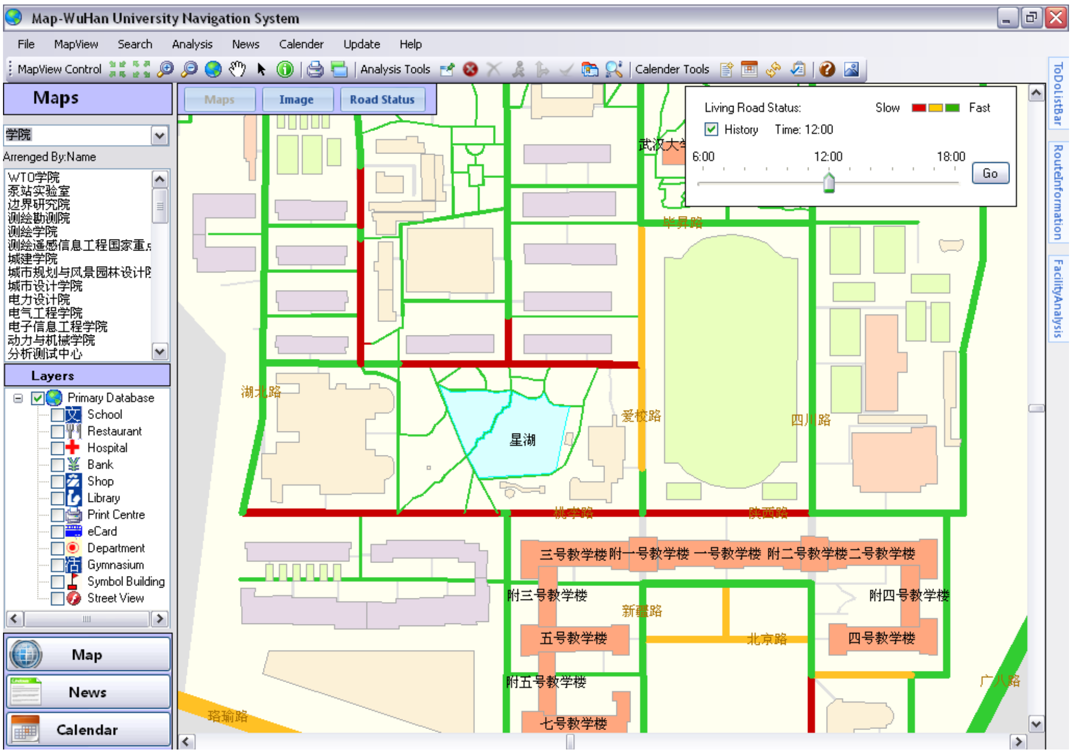
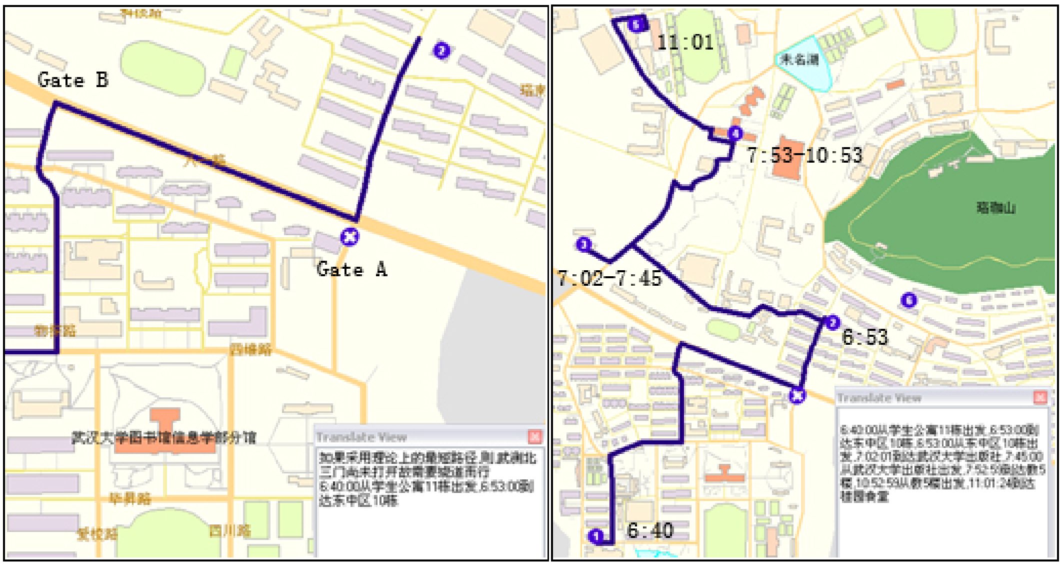


Figure 9 traffic updates under the framework of dynamic-update

In Figure 9, the client requests the server-side the latest traffic information through the dynamic update framework. Figure 10 shows the extension on time of the path analysis, Fig. 10 (a) shows how does the entrance gate switching time impact the path analysis. In this example, a user started from the starting point at 6:40 am. When he arrived in Gate A, Gate A was not open, after further analysis, Gate B was feasible. Fig. 10(b) shows that the path analysis integrates service time of each service point and users' residence time. Take point No. 3, Wuhan University Press as an example, Wuhan University Press is open from 7:30 am to 18: 00 pm. A user reached the Press at 7:02 am, because there was no second around, so the user need to wait. The process sets the user stayed in the press for 15 minutes, and the user left the press house at 7:45 am.



如果采用理论上的最短路径,则武测北三门尚未打开故需要绕道而行
6:40:00从学生公寓11栋出发,6:53:00到达东中区10栋

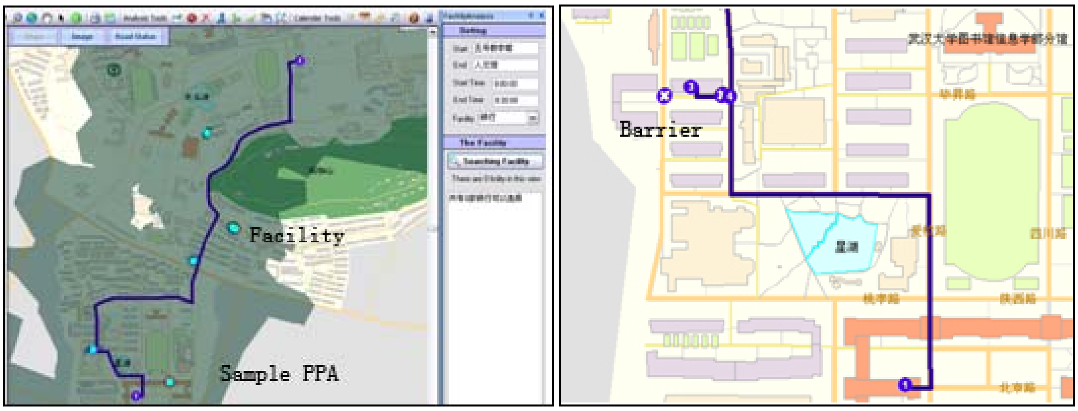
Translate View

6:40:00从学生公寓11栋出发,6:53:00到达东中区10栋,6:53:00从东中区10栋出发,7:02:01到达武汉大学出版社,7:45:00从武汉大学出版社出发,7:52:59到达教5楼,10:52:59从教5楼出发,11:01:24到达桂园食堂

(a) (b)

Figure 10 The combination of Path Analysis and time information

Figure 11 (a) shows service-points selection problem under a certain time-limited window. The gray area is simplified as the scope of PPA, and highlight points are the service points satisfying the constraints. The detail time information of each is provided. Fig. 11(b) shows the solution on the situation about the unreachable mid-points in the path analysis model based on the integration of persons and vehicles. Barriers exist on both sides of the 3rd point, which is unreachable. After the path analysis model for the integration of persons and vehicles, the 4th point was the best location as a on and off point, so the platform recommends that users get off at the 4th point, walk to the 3rd point, and get on at the 4th point, drive to next target point。



(a)

(b)

Figure 11 Facility Choice under time-limit (Left) and Route analysis model for the integration of people and vehicles (Right)

6. Conclusions

A dynamic route update framework is designed based on Socket and .Net Remoting technology after discussion of dynamic data update in route service, time information and different travel modes. Business hour of service points, open and close time information of gates are added in analysis to extend route service. Selection of proper facility points in a restrained time window is solved by simplifying potential service area in route analysis based on time-space prism concept. Route analysis model based on the integration of persons and vehicles is designed to obtain an optimal route, for which has the problem as existing some points that vehicles cannot reach between the depature point and users' destinations.

The integration of GIS with related subjects and technology may take place with the ongoing research about principles and methodology in dynamic GIS. Web 2.0, cloud computing and grid computing are strong supports of GIS. The dynamic data update method is one of the numerous means, and better results might be gained by a combination of several different ways. With further research of this update framework, perhaps it will be an effective solution of dynamic GIS. To achieve a better solution of persons' requirements in daily life, route service might be further extended in dynamic framework, and dynamic optimized routes and schedules are our ongoing research.

References

Alan Kwok, Lun Cheung., 2006. Representational Issues in Interactive Wayfinding Systems: Navigating the Auckland University Campus. *Lecture Notes in Computer Science*, Springer Berlin. 4295, pp. 90-101

David Curran, AndyOlsen et al., 2002. *Visual Basic.NET Remoting Handbook*. Wrox Press.

Goodchild, Michael., 1999. *Interoperating geographic information systems*. Kluwer Academic, Boston.

WeiMeng Lee. Practical., 2005c. *.NET 2.0 networking projects* Springer-Verlag New York Inc, New York, pp.1-65.

S. Koo, C. Rosenberg, H. Chan, and Y. Lee., 2003. Location-based E-campus Web Services: From Design to Deployment. *In Proceedings of the First IEEE International Conference on Pervasive Computing and Communication ,* PerCom.

Lenntorp, 1976, Paths in Space–Time Environments: A Time Geographic Study of Movement Possibilities of Individuals. Lund Studies in Geography B: Human Geography. *Lund Studies in Geography*, Series B, No.44

T.Neutens, F.Witlox,N. Van De Wenhe, PH.De Maeyer. 2007. Space-time opportunities for multiple agents: a constraint-based approach. *International Journal of Geographical Information Science*,21(Nos.9-10), pp.1061-1076.

Acknowledgements

The authors would like to thank Ms. Den Wenting of The School of Engineer, Cranfield University for her help in deploying Wuhan university navigation system.

ENHANCING TRAVEL TIME FORECASTING WITH TRAFFIC CONDITION DETECTION

Yingying Duan, Feng Lu, Jun Ouyang, Chuanbin Chen

State Key Laboratory of Resources and Environmental Information System,
Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences,
No. 11A, Datun Road, Chaoyang District, Beijing 100101, P. R. China - (duanyy, luf, ouyj,chencb)@lreis.ac.cn

KEY WORDS: Travel time forecasting, traffic condition detection, natural language understanding, spatial-temporal data mining

ABSTRACT:

Short-term traffic forecasting aims to provide more reliable travel information service, so as to assist people in making more reasonable travel decisions. With the increasing availability of traffic data along with the development of communication technology, both the capability and accuracy of travel time forecasting have been significantly enhanced in real-time conditions and a great number of forecasting methods have been carried out during recent years. However, they are inadequate when confronted with the real world traffic problems, since the real-time traffic condition can be affected easily and changed constantly. In our study, a hybrid forecasting approach is presented from a more practical perspective, based on a traffic condition detection method which monitors the real-time traffic condition and performs the travel time forecasting according to different traffic conditions.

In particular, we first build a traffic conditions evaluation system to detect different sorts of traffic conditions. In this study, the traffic conditions are divided into four types including light, stable, congested and abnormal traffic condition according to travel time cost. We use a clustering tool to obtain traffic flow patterns of different traffic conditions. And the process characterize the state of the system with respect to the deviation of current conditions from an expected ones based on historical data as a definition for abnormalities in the traffic stream. Then the hybrid forecasting approach, in which several methods are used to deal with different traffic conditions, is trained to judge with certain confidence which method performs the best according to the certain traffic condition with historical traffic data. Then the travel time forecasting is taken out after the detection of real-time condition by the hybrid forecasting approach with fixed historical data and received real-time traffic information. Case studies are carried out using a real-time traffic dataset in downtown Beijing.

1. INTRODUCTION

Travel information services develop quickly all around the world and make great efforts in the area of intelligent transportation systems (ITS). With the increasing public travel demands, traditional travel information services such as shortest path search on static maps are out of requirements. Since the accuracy and practicality of dynamic navigation highly depends on the short term traffic forecasting, a lot of short term travel time cost forecasting models on the roadways have been developed during recent years. The short term travel time cost forecasting research can be categorized as follows: direct and indirect model-based approaches, data-driven approaches (Van Lint et al., 2005) , and hybrid approaches.

Model-based approaches solve the travel time forecasting problem by forecasting traffic conditions from a number of time periods ahead and then subsequently deduce mean travel times from these forecasted traffic conditions. Examples include METANET (Smulders et al., 1999) , DynaSMART (Hu, 2001) , and DynaMIT (Ben-Akiva et al., 2001) . Data driven models are able to directly learn the complex traffic dynamics from the data and to forecast traffic flows and speeds. Many successful efforts have been reported including ARIMA models (Williams, 2001) ; artificial neural networks (Mark et al., 2004); fuzzy neural networks (Yin et al., 2002);and support vector regression models (Wu et al., 2004)and so on. All the data driven methods have in common that they correlate mean travel times or traffic conditions to current and past traffic data, and they do not require extensive expertise on traffic flow modelling. The easily use and the accuracy in progressing non-linear problem make the data driven models highly developed. A lot of forecasting models have been developed during recent years but none of them could consistently outperform the others. In the real-world applications, traffic forecasting conditions can be affected by a lot of factors. That leads to the research on hybrid approaches on travel time forecasting. The hybrid approaches combine with many possibly applicable candidate models for various traffic conditions (Zhu, 2009).

Studies on short-term traffic forecasting have shown that, one of the key functions of traffic management systems is to monitor traffic conditions and detect the presence of conditions that are abnormal or may not be expected (Turochy, 2006) . The traffic condition detection problem can be viewed as recognizing the non-recurrent congestion patterns from observed data series obtained from loop detectors. In recent years, computational intelligence approaches including neural-computing, evolutionary computing; wavelet analysis and fuzzy logic have been employed to solve the complex and mathematically intractable incident detection problems (Jin et al., 2006a).

However, most of them are based on single roadway pattern, geospatial neighbourhoods relationships between roadways do not involve in these pattern detection methods. And the real traffic data used in those researches are obtained from fixed sensors such as cameras and magnetic loops. These traffic data can easily be used in the freeway, but in the urban road network, the fixed sensors are most settled at the intersections in arterial roads. That causes the lack of detecting on the roadway segments and on the other roads.

Aiming at such problems, we propose a travel time cost forecasting approach based on traffic condition detection. Although traffic incidents are not predicable, we can enhance the travel time forecasting by detecting the abnormal traffic conditions and using different forecasting strategies under certain situation. Here we use two ways to get the abnormal traffic information. One is detecting the traffic condition; the other is collect traffic events described in natural language. The major contribution of this traffic condition detection method proposed is considering both spatial and temporal information in detection using the traffic velocity collected by floating vehicles since it is an appropriate measurement to indicate the congestion. When the real time traffic speeds on the roadways are received, the traffic detection module is working on detecting abnormalities in the traffic. And a traffic events collection module is also used to catch the traffic incidents described in natural language at the same time. After that, two types of neural networks are built to forecast the speeds in the future time intervals according to the traffic condition is recurrent or not. The method presented in this paper is argued to provide a practical solution for real-time public travel information service.

2. TRAFFIC CONDITION DETECTION

The real-time traffic data updated every five minutes is collected from the floating vehicles. We use a new algorithm outperformed the California algorithm consistently under various scenarios. The proposed approach includes spatial-temporal data mining and online detecting using California Algorithm. The California Algorithm is one of the earliest and most popular algorithms which based on the logical assumption that a traffic incident increases the traffic occupancy upstream of the incident and decreases the traffic occupancy downstream of the incident significantly(Jin et al., 2006b) . The incident detection process can be divided into four steps as listed below.

Step 1 Pre-processing
In this task, the raw velocity data obtained from float vehicles are transformed in the format needed for the algorithm. Common pre-processing approaches include calculating the cumulative values of time-series data and interpolating for the missing data.

Step 2: Traffic Model Generation
The second stage analyzes the traffic data on different weekdays to construct traffic models respectively. We de-noising and enhancement of the signal output obtained from the pre-processing and using statistics methods to confirm the confidence interval for each time interval on each roadway. This is an important stage because noise corruption is one of the primary reasons for poor reliability of the incident detection algorithms. Each dataset contains the velocity collected on one roadway at the same time on the same day of the week. Test the datasets follow which kind of distributions, and calculate the confidence interval by the cleaned data. The final traffic model will be generated as the confidence interval of the remaining historical data.

Step 3 Non-recurrent conditions confirmation
The detection will be performed by calculating the distance between the real-time data and the limit of the confidence interval. Once the real-time data is outlier of the confidence interval, the traffic data will be considered as a suspect abnormal condition. Then the traffic conditions on this roadway and on the adjacent roadways in the past time intervals should be taken into consideration. If the past time intervals are confirmed abnormal conditions or suspect condition, the current time interval can also be confirmed as non-recurrent condition. Or if the relational roadways such as the upstream and downstream roadways have been confirmed abnormal, the current time interval on the current roadway can be recognized as non-recurrent condition. If the traffic condition on current time interval can not be confirmed, it remains as a suspect condition.

3. TRAFFIC EVENTS COLLECTION

3.1 Traffic Events Reported in Nature Language

Neither of the magnetic loops, float vehicle or cellular phone signal analysis technologies can obtain the abrupt traffic events on spots or road cross turns. Once the abrupt traffic events happen, the traffic policemen, onlookers or people concerned will report the events and resulted influence (on the spot or monitor viewing) to the information center via cellular phones, short messages or other instant message systems. Since these traffic events are detected by human, those reported massages are most described in natural language. It has been a time consuming task to translate the messages into the valuable information that suit the applications and requires artificial work. The bottleneck focuses on understanding the natural language describing traffic and matching understood traffic information in LRS forms with the underlying road network spatial dataset, including the matching of address with the geometrical information, matching of multi-source LRS, and LRS and GIS positioning manner.

Although the natural language can not be truly understood automatically nowadays, the key words could be caught by the describing rules in some specific field. Thus the automatic understanding of traffic events reported by natural language should focus on the characteristic rules of traffic information. Those messages are always described briefly in certain forms combined with "<Address/Landmark> + {Direction} + {Offset} + <Traffic Events>".

According to the description rules, we proposed a cross-step word segmentation algorithm to process real-time traffic events represented in natural Chinese in this approach. Four libraries are built to process the word segmentation. The Address word library contains the addresses, landmarks as well as point of interests; The Direction word library is filled with the directions such as "由南向西"(from south to west) and so on; The Event word library stores the traffic events like "车行缓慢", "两车刮蹭", "追尾"; And the Ext Location word library used to save the offset value.Considering the record length distribution of the word libraries depicting real-time traffic information, this algorithm sets corresponding steps of word segmentation for address, direction and event libraries, and improves the one step running of the string pointer in classical Chinese word segmentation to flexible multiple steps running, so as to aggregate possible Chinese words efficiently. Then the addresses should be matched with the roadways and the influence of the events should be quantified for the velocity forecasting.

3.2 Traffic events processing module

Understanding the traffic events only concerns how to rapidly and accurately process the key words representing addresses, directions and events. Two sequential stages of the traffic events processing are identified: (1) natural language segmentation; (2) semantic understanding of the events. Figure 1 shows the flow of the processing.

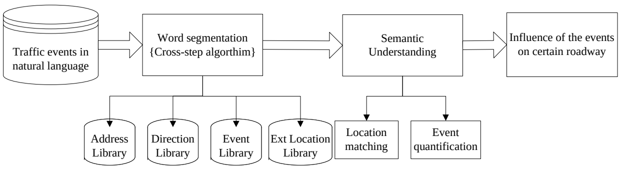


Figure 1 Technical work flow for traffic events processing

Here we use a novel MM algorithm to identify the key words. The classical maximum matching (MM thereafter) algorithm is working as follows: denote D as the word library, Max as the longest length in D , Str as the string needed segment. The MM algorithm searches in D with the substring from Str at the length of Max . If the substring matches with the word, the substring is denoted as a word and the pointer moves on the whole sentence. If no word matches the substring, the length of the substring should be reduce and search again (Feng et al., 2006).

Since the Max is longer than most of the words length, the classical MM algorithm may cause low efficiency. Aiming at this specific issue, we calculate the length in the word libraries and use the appropriate length as the initial length of the substring. Taking the address library as an example:

Record Length	Number	Ratio/%
2	29	0.71
3	797	19.48
4	1 480	36.18
5	1 218	29.77
6	427	10.44
>=7	140	3.42
total	4 091	100

Table 1 Length Distribution of Address Records in traffic events

We can see that more than 90% of the records have a length more than two characters. So here we use two as the initial length for the algorithm. A brief work flow of the improved MM algorithm is shown in figure 2:

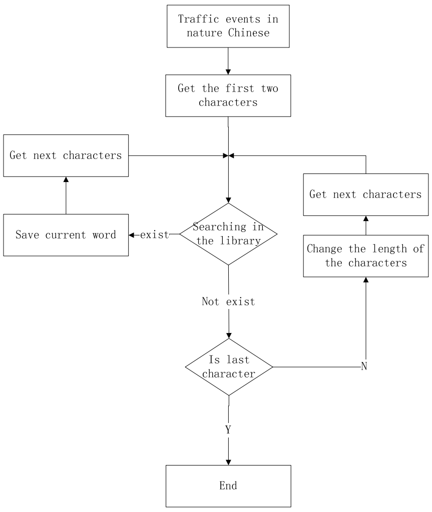


Figure 2 Work flow for word segmentation

This algorithm sets corresponding steps of word segmentation for address, direction and event libraries, and improves the one step running of the string pointer in classical Chinese word segmentation to flexible multiple steps running, so as to aggregate possible Chinese words efficiently. The proposed algorithm runs 10 times faster than an improved MM algorithm, while keeps similar accuracy and robustness.

The semantic understanding of the traffic events is about how to locate the address to the roadway and quantify the influence of the event. The traffic events reported in nature language most represented by the LRM(Linear Reference Methods). The representation can be defined by two ways shown in table 2.

Formula	Instance
Position Point + Direction＋Offset	北沙滩桥往东 100 米
Position Line + Direction + Offset	大屯路往东 100 米

Table 2 The standard LRM Formula

After the word segmentation, the key words can be used to match the roadway. The geographical positions of the addresses are used to confirm the position point/line, the direction and offset are transformed into distance in geographical coordinate system. The matching process is carried out by four steps.

1. Finding position point/line

According to different address types (Roadway, Intersection, POI, Overpass), using the geographical information such as the starting position, end position to confirm the position of the event.

2. Detecting the position type

Figure out the event is posited by point or line.

3. Matching the roadways

If the event is described by position point, find the start roadway which nearest the position and has the same direction with the offset. If the event is described by position line, the roadway matches the position line is the start roadway. Then search the roadways connected with start roadway to identify the end roadway. The searching considers both direction and offset distance which means the angle between the roadways should follows the direction and the cumulative length of the

roadways should larger than the offset. The traffic condition on the matched roadways will be effect by the traffic event.
The influence that the traffic events brings is reflected through the possible driving speed lost on underlying roadways, which is argued to be correlative with the event types and degrees.

The real-time traffic and events undoubtedly pose important influence on the turn costing. In our study, the famous HCM is utilized to model the influence of the real-time driving speed on the turn costing, with a precondition that the driving speed has a positive relation with the traffic flow.

4. FORECASTING METHOD

4.1 Travel time forecasting

A real-time traffic forecasting approach based on BP (Back Propagation) neural network is utilized in our study. If the traffic condition is been assessed as normal condition, which contain the recurrent congestion and the smooth traffic, the input will consider the spatial-temporal characteristics of travel time cost, the change pattern of city traffic is identified and adopted to obtain the rules of travel time cost forecasting using BP neural network. Otherwise, if the traffic condition is been assessed as abnormal condition which refers to non-recurrent traffic condition, the input is the speed in the past time intervals on the roadway as well as the speeds on the upstream and downstream roadways.

It is well known that the dynamic characteristics of traffic are very complicated and restricted by the statistics granularity of the traffic time series. When the granularity of the traffic time series is too small(less than 2 minutes), the randomness plays the key role which makes the time series too complex. If the granularity of the traffic time series are too large(more than 15 minutes), the real time traffic influence in the future time segment can be ignored. Based on the empirical results, we use 5~15 minutes as the granularity of the traffic time series.

4.2 Forecasting module

The BP neural networks are trained independently for each roadway. In normal traffic condition, two types of traffic data are required for the BP neural networks training when forecasting the travel time cost on roadway L_i at time $T+t$:

(1) Historical data:
The historical data includes the historical speed of the roadway L_i at time $T+t$ on the same day of the week.

(2) Real-time data:
The real-time data includes the real time collected speed of the roadway L_i before time $T+t$ which means the time intervals before T .

The real travel time cost of the roadway L_i at time $T+t$ is used as the teacher to supervise the training.
After well trained, the travel time cost on each roadway in the future time segment can be forecasted.

In non-recurrent congestion, two types of traffic data are required for the BP neural networks training when forecasting the travel time cost on roadway L_i at time $T+t$:

(1) Traffic environment data:
The real-time collected speeds at time T on the roadways which are the upstream and downstream to roadway L_i

(2) Real-time data:
The real-time data includes the real time collected speed of the roadway L_i before time $T+t$ which means the time intervals before T .

The real travel time cost of the roadway L_i at time $T+t$ is used as the teacher to supervise the training.

After well trained, the travel time cost on each roadway in the future time segment can be forecasted.

5. CASE STUDY

Traffic network of Beijing, China is selected as test bed for this study. The traffic information is gathered by the floating car data in Beijing. Floating car data (FCD) is a method to determine the traffic speed on the road network. It is based on the collection of localization data, speed and direction of travel and time information from driving vehicles.

5.1 Choice of testing datasets and data pre-processing

The raw data is the mean speed data of over 2000 major roadways collected every five minutes in Beijing from July 1st to September 30th, 2007. Since the day of the week effect the traffic, our experiments are taken separately on each day of the week.

Rough data caused by the data collecting, transmitting and matching with the road network can not be used directly since the outliers and missing data are too numerous. Hence, we provide a inverse time weighted interpolation and completion algorithm including two steps:

1). Outliers detection is based on the following two criteria:
(a) Detection of two low speed measures, below 5km/h for more than one hour;
(b) Detection of inconstant speed measures, less than half of the average speeds for less than 10 minutes.

2). Missing data completion is carried out by taking the weighted mean of the available data.

5.2 Schematic System Structure

The system framework is presented as shown in Figure 3. This figure illustrates the logical sequence of how proposed forecasting system operates.

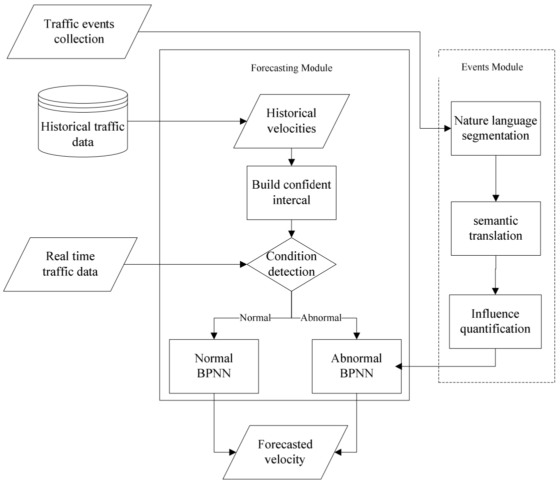


Figure 3 Schematic structure of system design

5.3 Procedure:

The template model is carried out for each roadway by statistical the historical data. Here we take the FuChengRd as an example. The velocities collected on this roadway around 8:00 AM on weekdays are shown in Figure 4.

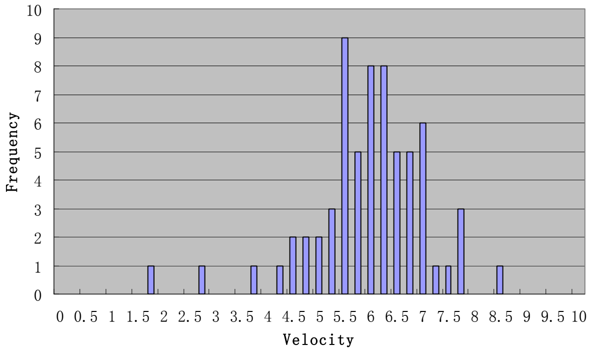


Figure 4 Velocity collected on FuChengRd

The distribution is tested as Gaussian distribution. The confident interval at 8:00 AM on weekday can be calculated according to the formula in Gaussian distribution. And the other time intervals are calculated in the same way. In Figure 5, the cyan line shows the template model on FuChengRd on weekdays.

The forecasting results on FuChengRd are shown in figure 5. It shows the velocity on this roadway from 5 AM to 8 PM on one Tuesday. The box in the figure highlight the non-recurrent congestion detected by the proposed approach and forecasted by the BP neural network for abnormal traffic condition.

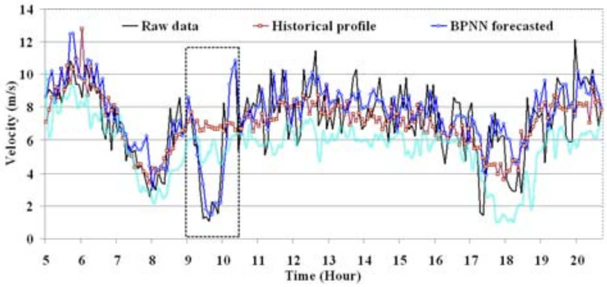
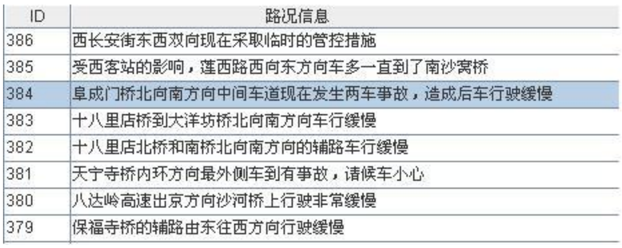


Figure 5 Velocity on roadway

The traffic events are got from the Beijing Traffic Radio in Chinese nature language at first, and then be syncopated and parsed into segmented words by semantic rules. The segmented words are translated into address, direction, event and location by the four word libraries using the improved maximum matching algorithm that presented in this paper. After that, the address is matched with map by geometrical information.

The traffic events are collected from Beijing Traffic Radio in Chinese nature language firstly (as shown in figure 4), and then be syncopated and parsed into segmented words with semantic rules. The segmented words are translated into address, direction, event and location libraries using a cross-step word segmentation algorithm. Then the resulted traffic influence is matched with road network maps for further applications
When the BP neural network based traffic forecasting server receives the real-time traffic and events information, the real-time traffic condition is detected and the forecasted driving speed in next 5~15 minutes for every roadways is obtained from the well trained network, and the network is adjusted at the same time.

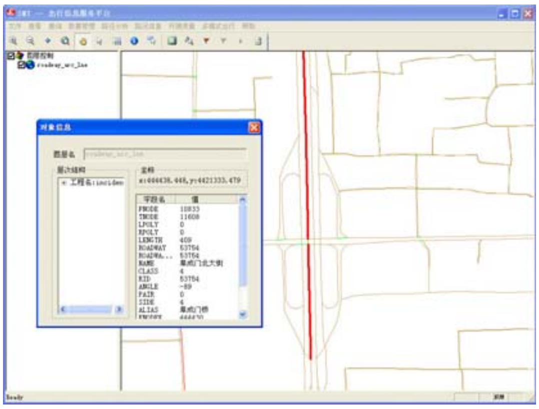
Fig.6 shows an example for natural Chinese understanding. By the use of the improved MM algorithm proposed above, the sentence “阜成门桥北向南方向中间车道现在发生两车事故造成后车行驶缓慢” is segmented into five key Chinese words, namely, “阜成门桥” as an address, “北向南” as a direction, and “事故” and “行驶缓慢” as two events.



ID	路况信息
386	西长安街东西双向现在采取临时的管控措施
385	受西客站的影响，莲西路西向东方向车多一直到了南沙窝桥
384	阜成门桥北向南方向中间车道现在发生两车事故，造成后车行驶缓慢
383	十八里店桥到大洋坊桥北向南方向车行缓慢
382	十八里店北桥和南桥北向南方向的辅路车行缓慢
381	天宁寺桥内环方向最外侧车到有事故，请候车小心
380	八达岭高速出京方向沙河桥上行驶非常缓慢
379	保福寺桥的辅路由东往西方向行驶缓慢

Figure 6 Traffic events described in natural language

Then the address is matched with the roadway “阜成门北大街” in the spatial data, and the events effect the traffic by changing the velocity of this matched roadway.Fig.7 shows the how this event effects the traffic. The roadway “阜成门北大街” is turned red since a congestion occurred here and the velocity is very slow.



x:444438.448,y:4421333.479

字段名	值
FNODE	10833
TNODE	11608
LPOLY	0
RPOLY	0
LENGTH	409
ROADWAY	53754
ROADWA...	53754
NAME	阜成门北大街
CLASS	4
RID	53754
ANGLE	-89
PAIR	0
SIDE	4
ALIAS	阜成门桥

Figure7 Influence caused by a traffic event

The forecasting is taken by different BPNN models according to the certain traffic condition. Here we use historical profile as the comparative study. Historical profile is calculating the average of historical traffic data as the forecasted traffic data. The forecasting is taken independently with the forecasting time interval as 5 minutes, 10minutes and 15minutes. The results carried out at normal conditions are shown in table 3.

Day	AVG	BP5	BP10	BP15
Mon	0.122	0.121	0.124	0.125
Tue	0.105	0.11	0.111	0.114
Wed	0.115	0.118	0.119	0.121
Thu	0.115	0.115	0.118	0.120
Fri	0.113	0.108	0.111	0.114
Sat	0.125	0.123	0.125	0.126
Sun	0.128	0.126	0.129	0.130

Table 3 Forecasting errors on normal conditions

The forecasting errors on detected abnormal condition are shown in table 4.

Method	Error
Average	1.27
BPNN in next 5 min	0.546
BPNN in next 10 min	0.689
BPNN in next 15 min	0.995

Table 4 Forecasting errors on abnormal conditions

6. CONCLUSION AND DISCUSSION

In this paper, we proposed a new method to forecasting travel time costs on roadways by identify the traffic condition in real time. The traffic conditions are detected by a spatial-temporal data mining and differentiating as recurrent and non-recurrent conditions. Different input data and neural networks are used according to the certain condition. The approach can take advantages both in historical and real-time traffic information. The changes in traffic patterns can be well taken into consideration to enhance the accuracy of the forecasting. The results show that the process can enhancing short-term travel time cost forecasting especially in the abnormal conditions.

Further exploration of this topic may include other detection methods with other distance metrics and other forecasting equations as well。Since the point here is to detect the different traffic condition and to recognize the traffic incident, the influence of the traffic incidents learned in the paper are only used to forecasting the changes of velocity in a very short time. For the future work, the estimation of the influence both in spatial and temporal should be well discussed.

ACKNOWLEDGEMENTS

This research is supported by the Hi-Tech Research and Development Program of China under grant No. 2006AA12Z209 and 2007AA12Z241; Chinese NSFC program under grant No. 40871184. Their supports are gratefully acknowledged.

REFERENCES

Ben-Akiva, M., Bierlaire, M., Burton, D., Koutsopoulos, H. and Mishalani, R., 2001. Network state estimation and prediction for real-time transportation management applications. Networks and Spatial Economics, 1(3/4): 293-318.

Feng, L., Zhu, L. and Zhang, J., 2006. A Model for Ontology driven Web Information Retrieval and Its Realization. Journal of the China Society for Science and Techenical Information, 25(003): 276-281.

Hu, T.Y., 2001. Evaluation framework for dynamic vehicle routing strategies under real-time information. Transportation Research Record: Journal of the Transportation Research Board, 1774(-1): 115-122.

Jin, Y., Dai, J. and Lu, C.T., 2006a. Spatial-Temporal Data Mining in Traffic Incident Detection, SIAM DM 2006 Workshop on Spatial Data Mining. Citeseer, Bethesda.

Jin, Y., Dai, J. and Lu, C.T., 2006b. Spatial-Temporal Data Mining in Traffic Incident Detection, SIAM Conference on Data Mining. Citeseer, Maryland, USA.

Mark, C.D., Sadek, A.W. and Rizzo, D., 2004. Predicting experienced travel time with neural networks: a PARAMICS simulation study, pp. 906-911.

Smulders, S.A., Messmer, A. and Knibbe, W., 1999. Real-time application of metanet in traffic management centres.

Turochy, R.E., 2006. Enhancing Short-Term Traffic Forecasting with Traffic Condition Information. Journal of Transportation Engineering, 132(6): 469-474.

Van Lint, J.W.C., Hoogendoorn, S.P. and van Zuylen, H.J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research Part C (13): 347-369.

Williams, B.M., 2001. Multivariate vehicular traffic flow prediction: Evaluation of ARIMAX modeling. Transportation Research Record: Journal of the Transportation Research Board, 1776(-1): 194-200.

Wu, C.H., Ho, J.M. and Lee, D.T., 2004. Travel-time prediction with support vector regression. IEEE transactions on intelligent transportation systems, 5(4): 276-281.

Yin, H., Wong, S.C., Xu, J. and Wong, C.K., 2002. Urban traffic flow prediction using a fuzzy-neural approach. Transportation Research Part C, 10(2): 85-98.

Zhu, J., 2009. A self-learning short-term traffic forecasting system through dynamic hybrid approach, 11th International Conference on Computers in Urban Planning and Urban Management, Hong Kong, China.

EVALUATION OF AUTOMATICALLY EXTRACTED LANDMARKS FOR FUTURE DRIVER ASSISTANCE SYSTEMS

Claus Brenner and Sabine Hofmann

Institute of Cartography and Geoinformatics
Leibniz Universität Hannover
Appelstrasse 9a, 30167 Hannover, Germany
{claus.brenner, sabine.hofmann}@ikg.uni-hannover.de

KEY WORDS: LIDAR, Mobile Laser scanning, Extraction, Matching, Mapping, Accuracy, Navigation

ABSTRACT:

In the future, vehicles will gather more and more spatial information about their environment, using on-board sensors such as cameras and laser scanners. Using this data, e.g. for localization, requires highly accurate maps with a higher level of detail than provided by today's maps. Producing those maps can only be realized economically if the information is obtained fully automatically. It is our goal to investigate the creation of intermediate level maps containing geo-referenced landmarks, which are suitable for the specific purpose of localization.

To evaluate this approach, we acquired a dense laser scan of a 22 km scene, using a mobile mapping system. From this scan, we automatically extracted pole-like structures, such as street and traffic lights, which form our pole database. To assess the accuracy, ground truth was obtained for a selected inner-city junction by a terrestrial survey. In order to evaluate the usefulness of this database for localization purposes, we obtained a second scan, using a robotic vehicle equipped with an automotive-grade laser scanner. We extracted poles from this scan as well and employed a local pole matching algorithm to improve the vehicle's position.

1 INTRODUCTION

1.1 Driver Assistance Systems

In recent years, research and development in the area of car driver information and safety systems has been quite active. Beside the goal to provide information to the driver, active safety systems should completely prevent accidents, instead of just reducing the consequences. Therefore, vehicles are using more and more onboard sensors such as cameras, laser scanners or radar to gather spatial information about their environment. There are numerous examples for driver assistance systems, like adaptive cruise control, collision warning/ avoidance, curve speed warning/ control, and lane departure warning. Compared to car navigation, these so-called advanced driver assistance systems (ADAS) operate at a very detailed scale, requiring large scale maps, and thus, even more detailed field surveys. The NextMAP project has investigated the future map database requirements and the possible data capturing methods (Pandazis, 2002). The key point is to align the evolution of in-vehicle sensors with advances in mapping accuracy and map contents in such a way that sensible ADAS applications can be realized (at minimum cost). For example, additional mapping of speed limit, stop, priority, right of way, give way and pedestrian crossing signs, traffic lights, as well as the capture of the number of lanes, the road and lane width, and of side obstacles such as houses, walls, trees, etc. has been proposed. However, the NextMAP project has not dealt with the commercial aspects of producing such maps with very high detail and accuracy. Some investigations with regard to this have been made in the 'Enhanced Digital Mapping' project (EDMap, 2004). ADAS applications have been classified, according to their requirements, into 'WhatRoad','WhichLane', and 'WhereInLane', the latter requiring a mapping accuracy of ±0.2 m. It was concluded that contemporary mapping techniques would be too expensive to provide such maps for the 'WhichLane' and 'WhereInLane' case.

A key observation is that these findings assume a 'traditional' map production, i.e. the map is a highly abstract representation in terms of a vector description of the geometry, with additional attributes. However, depending on the application, this is not always necessary. For example, the accurate positioning of vehicles using relative measurements to existing map features does not necessarily require a vector map in the usual sense. Rather, any map representation is acceptable as long as it serves the purpose of accurate positioning. This means that map production to derive a highly abstract representation (which usually involves manual interaction) is not required, but rather a more low-level representation is suitable as well (which can be derived fully automatically and thus, inexpensively).

Also, accuracy considerations nowadays usually assume that the position of a vehicle relative to the map is obtained using an absolute measurement for both, typically using the global positioning system (GPS) for the vehicle and geo-referenced maps (possibly obtained by GPS measurements as well). However, when the vehicle uses relative measurements to known objects in the environment (such as poles and crash barriers), a very high absolute map accuracy is not required. Thus, one is relieved from unreasonable high requirements regarding absolute map accuracy and absolute vehicle positioning, nevertheless making relative centimeter level positioning feasible.

1.2 Mapping Based on Mobile Laser Scanning

Mobile mapping systems using laser scanners combined with GPS and inertial sensors are well suited for the production of large scale maps, since they reach a relative accuracy of down to a few centimeters. With a rate of more than 100,000 points per second, they capture the environment in great detail (Kukko et al., 2007).

The problem of building a representation of the environment using laser scanners has also been investigated in robotics (although, of course, laser scanners were not the only sensors considered). Iconic representations, such as occupancy grids, have been used as well as symbolic representations consisting of line maps or landmark based maps (Burgard and Hebert, 2008). One of the major problems to solve is the simultaneous localization and mapping (SLAM), which incrementally builds a map while a robot drives (and measures) in unknown terrain (Thrun et al., 2005). Often, 2D laser scans (parallel to the ground) are used for this, but it has also been extended to the 3D case (Borrmann et al., 2008). However, it is probable that it will still take some time until such full 3D scanners are available in vehicles. What can be expected in the near future, though, are scanners which scan in a few planes such as the close-to-production IBEO Lux scanner (IBEO, 2009), which scans in four planes simultaneously.

Apparently, it would be unreasonable to provide dense georeferenced point clouds for the entire road network, as this would imply huge amounts of data to be stored and transmitted. On the other hand, as pointed out in the previous section, obtaining a highly abstracted representation usually requires manual editing, which would be too expensive. Following the work in robotics, a landmark based map would be a suitable approach. For example, to provide a map useful for accurate positioning, landmarks should fulfill the following requirements: (1) they should be unique in a certain vicinity, either by themselves or in (local) groups, (2) their position / orientation should be stable over time, (3) few of them should be needed to determine the required transformation with a certain minimum accuracy, (4) they should be reliably detectable, given the available sensors and time constraints.

In this paper we describe the use of pole-like objects as landmarks. We assess the accuracy of the extracted poles using a terrestrial survey and investigate the applicability of pole matching for the problem of vehicle localization.

2 DATA ACQUISITION

In the following sections, we describe the acquisition of the two data sets used. For our experiments, we obtained two different data sets of the same area in Hannover, Germany. The first data set is a very dense point cloud, obtained by the Streetmapper mobile mapping system and is described in section 2.1. The second data set was acquired by a vehicle, equipped with an automotive-grade scanner and is described in section 2.2.

2.1 Reference Data

A dense laser scan of a number of roads in Hannover, Germany, was acquired using the Steetmapper mobile mapping system. This system was jointly developed by 3D laser mapping Ltd., UK, and IGI mbH, Germany (Kremer and Hunter, 2007). The scan was acquired with a configuration of four scanners. Two scanners Riegl LMS-Q120 were pointing up and down at an angle of 20°, one was pointing to the right at an angle of 45°. Another Riegl LMS-Q140 was pointing to the left at an angle of 45°. The LMS-Q120 has a maximum range of 150 m and a ranging accuracy of 25 mm. All scanners were operated simultaneously at the maximum scanning angle of 80° and scanning rate of about 10,000 points/s. Positioning was accomplished using IGI's TERRAcontrol GNSS/IMU system which consists of a NovAtel GNSS receiver, IGI's IMU-IId fiber optic gyro IMU operating at 256 Hz, an odometer, and a control computer which records all data on a PC card for later post processing to recover the trajectory. The scanned area contains streets in densely built-up regions as well as highway like roads. The total length of the scanning trajectory is 21.7 kilometers, 70.7 million points were captured. On average, each road meter is covered by more than 3,200 points.

2.2 Vehicle Data

The second scan was obtained by RTS Hanna, a vehicle which was developed by the Real Time Systems Group (RTS) at the Institute for Systems Engineering at the Leibniz Universität Hannover, Germany (Figure 1). As main sensor they use two pairs of custom-built 3D laser range scanners for environmental detection, the 'RTS-ScanDriveDuo'. Each pair consists of two 2D SICK LMS 291-S05 laser scanners, which are mounted on the ScanDrive, a rotary unit. In vertical direction, the scanners cover an area of 90°, when reading remission values. Based on the center point of a single scanner, 20° of the upper half-space and 70° of the lower half-space are recorded. The angular resolution is 1° in vertical direction. The horizontal scan range of each scanner pair is 260°, with an angular resolution of approximately 2°. Taking the driving direction as 0°, the left scanner covers the area from $-210°$ to $+50°$, the right scanner covers the area from $-50°$ to $+210°$. With a rotational speed of 150°/s, it takes 2.4 s for a full rotation. Using a pair of scanners where each scanner only covers one half-space, the scanning time is reduced to 1.2 s. New scan lines are provided every 13 ms. Therefore, each 3D scan consists of 180 scan lines. The scanners have a maximum range of 30 m and a ranging accuracy of 60 mm.

For positioning, RTS-Hanna uses a Trimble AgGPS 114 without differential corrections. A speed sensor at the gear output and an angular encoder at the steering provide the odometer data.

For our experiments, all data, including GPS positions and headings, odometer data and scanned point clouds, were stored on a PC card for post processing. A single point cloud contains post processed data of one scanner rotation calculated relative to a fixed center point. Each scan consists of approximately 6,000 object points.

To compare results, this second scan was acquired along the trajectory of the first scan, excluding highway like roads because of speed limitations.

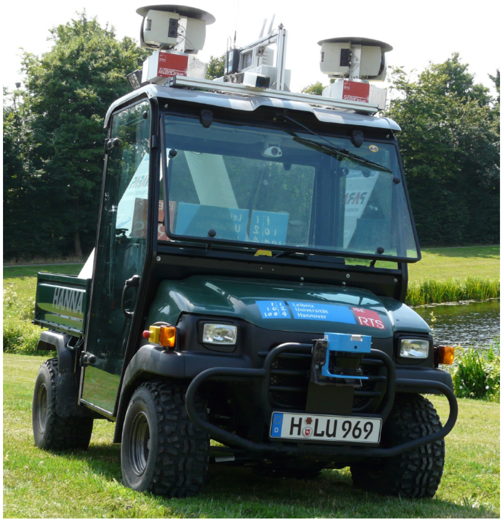


Figure 1: The RTS Hanna robotic vehicle. The point clouds were acquired using the four laser scanners mounted on the two ScanDrives on top of the vehicle. Image courtesy of RTS.

3 FEATURE EXTRACTION

We decided to use poles, such as poles of traffic signs, traffic lights, and trees, as landmarks. Poles are usually abundant in street corridors, especially in inner city scenes. It has been shown earlier by (Weiss et al., 2005), that a combination of GPS, odometry, and a (four-plane) laser scanner can be used to estimate the position of a vehicle, given a previously acquired map of poles.

The reported accuracies were in the range of a few centimeters.

In order to form matches, poles have to be extracted in both the reference scan and the vehicle scan, as described in the following two subsections.

3.1 Extraction of Poles from Streetmapper Data

The automatic extraction of simple shapes (like cylinders) from laser scanning data has already been done in other contexts, e.g. the alignment of multiple terrestrial scans in industrial environments (Rabbani et al., 2007). However, as can be seen from Figure 2, left, single poles are not hit by very many scan rays. Thus, methods which rely on the extraction of the surface, of surface normal vectors, or even of curvature, are not applicable.

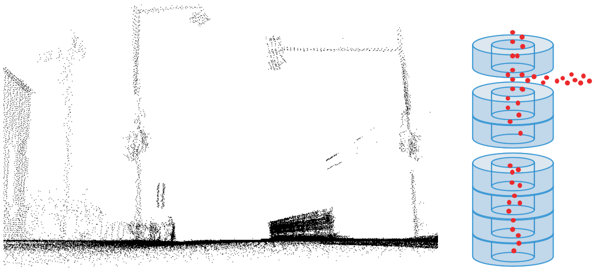


Figure 2: Left: Part of the Streetmapper scene containing a sign post, street (and traffic) light, and traffic light. Right: Illustration of the pole extraction algorithm using a stack of hollow cylinders.

Instead, we use a simple geometric model for pole extraction, assuming that poles are upright, there is a kernel region of radius r_1, where laser scan points are required to be present, and a hollow cylinder, between r_1 and r_2 ($r_1 < r_2$) where no points are allowed to be present (Figure 2, right). The structure is analyzed in stacks of hollow cylinders. A pole is confirmed when a certain minimum number of stacked cylinders is found. After a pole is identified, the points in the kernel are used for a least squares estimation of the pole center. The method also extracts some tree trunks of diameter smaller than r_1, which we do not attempt to discard.

In the entire 22 km scene, a total of 2,658 poles were found fully automatically, which is one pole every 8 meters on average. In terms of data reduction, this is one extracted pole per 27,000 original scan points. Although the current implementation is not optimized, processing time is uncritical and yields several poles per second on a standard PC. Note that this method owes its reliability to the availability of full 3D data, i.e., the processing of the 3D stack of cylinders. We classified the extracted poles manually and found 46% trees, 21% street lights, 5% traffic lights, 5% tram poles, 8% sign posts and other poles and 9% of non identifiable pole-like objects. The group of non identifiable objects contains all pole-like objects which are located along roads, but could not be identified definitely by manual inspection, e.g. since too few laser beams had hit the pole. The error rate (false positives) was about 6% on average, with higher rates of up to more than 30% in densely built-up areas. False positives are mainly due to occlusions which generate narrow bands of scan columns, similar to poles.

3.2 Extraction of Poles from RTS Hanna Data

Hanna scan data consists of scan frames, where one frame is a horizontal sweep of any of the four scanners. One orientation is given for each single frame. As the vehicle moves during the scan, scanpoints are motion compensated in such a way that the single, given orientation holds for all points of the frame.

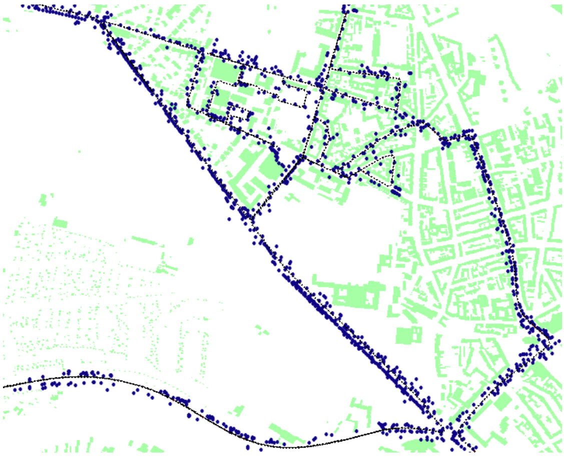


Figure 3: Extracted poles (blue dots) in a larger area with buildings from the cadastral map (green). The thin dotted line is the trajectory of the Streetmapper van.

Unfortunately, applying the same algorithm as in the Streetmapper case to extract poles does not work well. The major problem is the low point density. From the 3D point cloud of a single scan, it is almost impossible to tell if a column of stacked points is due to an actual pole or just a consequence of the low density (Figure 4). On the other hand, if one tries to combine several scans along the trajectory in order to obtain a higher density, the positioning errors will lead to several images of the very same pole, as can be seen in Figure 5. The extracted poles may be more than one meter apart, even if they were derived from successive scans.

Therefore, we extracted the poles from single scans, using the raw data rather than the 3D point cloud. Since the points are recorded in succession, column by column, a neighborhood can be defined on the row and column indices. This means that we perform our analysis on the depth image, indexed by the horizontal and vertical scan angles. One has to keep in mind, though, that this image is not an exact polar representation due to the forward movement of the vehicle during the scan. We extract poles by searching for vertical columns which have height jumps to both their left and right neighboring columns. The resulting point subsets are checked for a minimal height. Then, a principal component analysis is performed and the selected column subset is accepted based on the eigenvalues, equivalent to fitting a line and checking the residual point distances.

Figures 4 and 5 show poles which were extracted using this method. In total, 1,248 poles were detected in 1,384 scans of the first trajectory and 1,794 poles in 3,237 scans for the second trajectory. Due to the limited range and resolution of the scanners, there are around 50% (first trajectory) to 60% (second trajectory) of scans where no poles were found and around 30% of scans where only one pole was found. In both trajectories, the maximum number of detected poles in any scan was 6. Figure 6 shows the distribution of the number of scans in which n ($0 \leq n \leq 6$) poles were found.

Figure 7 compares the reference poles, extracted using the Streetmapper van, with the poles obtained from the RTS Hanna scans. The latter were extracted in each scan separately, as described above. They were then transformed to the global coordinate system using the transformation provided for each scan

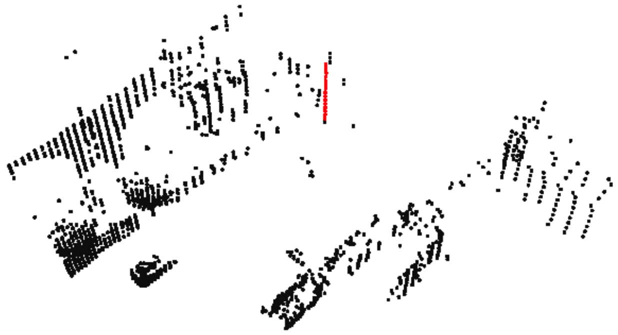


Figure 4: Single pole (red dots) extracted from RTS Hanna data (black dots). Ground points were removed for better visibility.

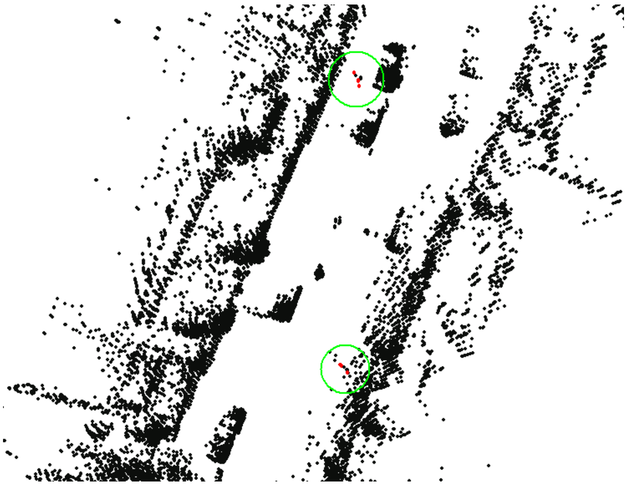


Figure 5: Poles (red dots inside green circles) extracted from RTS Hanna data (black dots). Ground points were removed for better visibility. Extracted poles appear at several locations due to the positioning error of the vehicle.

frame, which is based on GPS positioning and odometry. In Figure 7, left, the situation is especially obvious at the parking lot in the upper right, where a systematic shift is present (corresponding poles are marked by circles). Again, it can be seen that single poles appear multiple times in the RTS scans due to the low positioning accuracy. Figure 7, right, shows another scene at an intersection, where the reference poles are to the left and right of the street, whereas the poles extracted from RTS Hanna often appear inside the buildings or in the middle of the road. See Figure 12 for final results on these two scenes.

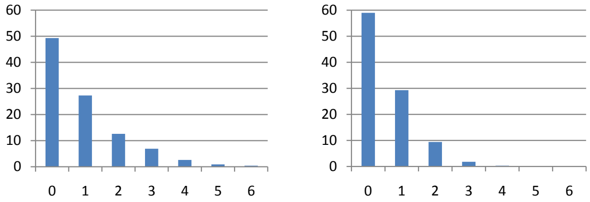


Figure 6: Histograms of the percentage of scans with n extracted poles, for the first (left) and second (right) trajectory of the RTS vehicle.

4 DATA ANALYSIS

4.1 Verification of Reference Data

Before using the reference data provided by the Streetmapper system, we have to make a statement on the quality of this data. Therefore, an analysis of the accuracy of position of the extracted landmarks is required. Ground truth was obtained along a selected inner-city junction by a terrestrial survey, using a total station to measure every single pole manually. The accuracy of position is < 10 cm for every measured pole.

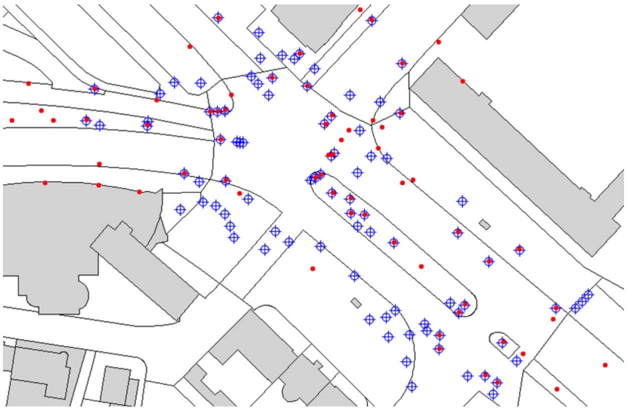


Figure 8: Accuracy of position: ground truth obtained by a terrestrial survey (blue crosshairs) compared to automatically extracted poles (red dots) of the reference map.

As ground truth and extracted reference poles are both given in DHDN 1990, 3rd meridian strip, calculating the residuals is very easy. We found 28 poles with a maximum distance of 32 cm from their corresponding pole. Using larger search radii only leads to mismatches between poles. Calculating the residuals using these points, we achieved a maximum distance between corresponding poles of 22.9 cm and a minimum distance of 1.4 cm. The root mean square error was 12.1 cm. Figure 9 shows the error distribution.

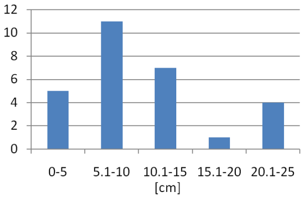


Figure 9: Histogram of pole measurement errors.

Comparing the extracted poles to the ground truth also revealed the problem of up-to-dateness. As shown in Figure 10, some of the extracted poles, marked with red crosses, have a large distance to ground truth poles, marked with black circles (the outer circle has a distance of 50 cm to the center of the pole). In this case there is no gross error in the feature extraction algorithm, but the junction was rebuilt between scanning the area and measuring poles manually.

5 EXPERIMENTS

The goal of our experiment is to find out if the position of the vehicle can be determined using a matching of poles. To this end, we implemented a local matching algorithm, as described in the following (see also Figure 11).

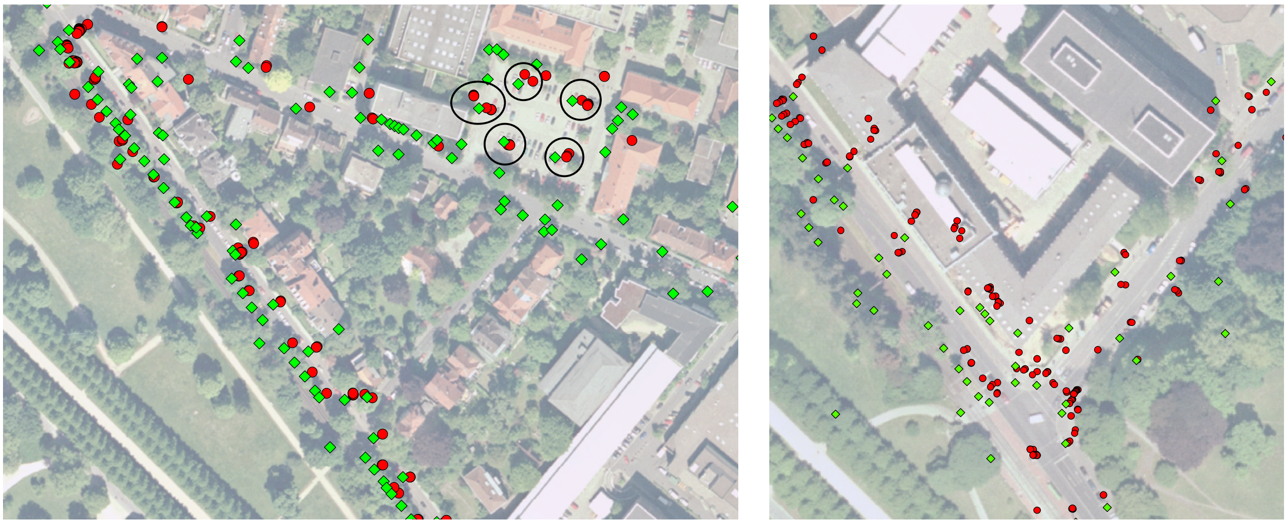


Figure 7: Comparison of reference poles (green squares) and poles extracted using the RTS Hanna vehicle (red dots).

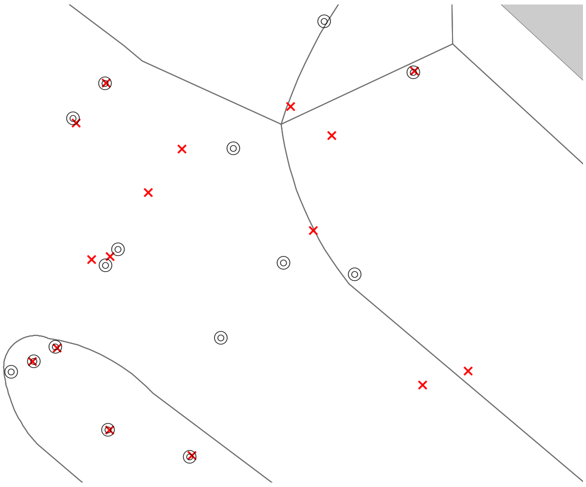


Figure 10: Detailed view of the junction: large distances between extracted poles (red crosses) and ground truth (black circles, radius: inner circle 25 cm, outer circle 50 cm) are not measurement errors but due to a rebuilt of the junction.

The vehicle maintains a current transformation, T, and a current set of active poles, P, which is a set of 2D points. As the vehicle moves forward, new scans become available and some of them lead to newly detected poles. Those poles are then added to the active set. At the same time, old poles above a given horizon distance ε_h are removed from the active set. The reason for this is that the positions in the active set can only be assumed to be approximately correct within a certain distance, as their transformation is based on the given scan frame orientation, which uses GPS and odometry.

Whenever the active set is updated, a match to the global database of poles, Q, is attempted. For each point in the active set, all neighbors in the database within a distance smaller than a search distance ε_s are retrieved. For any translation vector between a pole in the active set and a corresponding element among its neighbors, the active set is transformed and the number of 'hits' of all poles in the active set to the database are counted. A hit is assumed when the distance is below a 'matching' threshold ε_m. This is similar to random sampling consensus (RANSAC), however, due to the low number of elements we can afford checking all possible translation vectors instead of having to pick them randomly. The transformation T' with the largest consensus set is estimated using a least squares estimation of a similarity transform. It is then taken as the new local transformation T. We used ε_h=100 m for the horizon, ε_s=15 m as search radius, and ε_m=2 m as match radius in our experiments.

```
$Q$ is the reference point set
Initialize the transformation $T$ to identity
Initialize the current active set of points $P$ to be empty
**while** new points $p$ become available **do**
    $P \leftarrow P \cup \{p\}$
    Remove all points $p'$ from $P$ with dist$(p', p) > \varepsilon_h$
    For each point $p \in P$ compute
        $N_\varepsilon(p) \leftarrow \{n | \text{dist}(n, p) < \varepsilon_s\}$
    **for** all pairs $p, q$ with $p \in P$, $q \in N_\varepsilon(p)$ **do**
        Compute transformation $T'$ mapping $p$ to $q$
        Map all $p \in P$ using $p' = T'(p)$ and count all $p'$
            for which $\exists n \in N_\varepsilon(p)$ with dist$(n, p') < \varepsilon_m$
    Let $T'$ be the transformation with the highest count
    Find closest point pairs using $T'$ and determine
        $T \leftarrow$ best L.S. estimation based on those pairs
```

Figure 11: Algorithm used for local point pattern matching.

The results for the two scenes of Figure 7 can be seen in Figure 12, which shows the positions of the vehicle according to the given orientation per scan frame (red dots) and the corrected position after applying the transformation T (green squares). For example, in the left image, the trajectory starts at the parking lot, where both positions are identical in the beginning (only the red dots are visible). After a few meters, the poles are matched and the green trajectory is corrected to its true position. The corrected trajectory is also clearly better in the left and lower part, where the original trajectory obviously suffers from GPS shadowing by the trees. Similarly, in the right image, the original coordinates are far off the correct lane not only after the U-turn, but also when driving straight on (presumably also due to shadowing). As there is no ground truth for the real trajectory, e.g. obtained with a GPS/IMU system with substantially higher accuracy, our analysis does not include an evaluation for the entire trajectory. Clearly, there are also many parts in the trajectory where our lo-

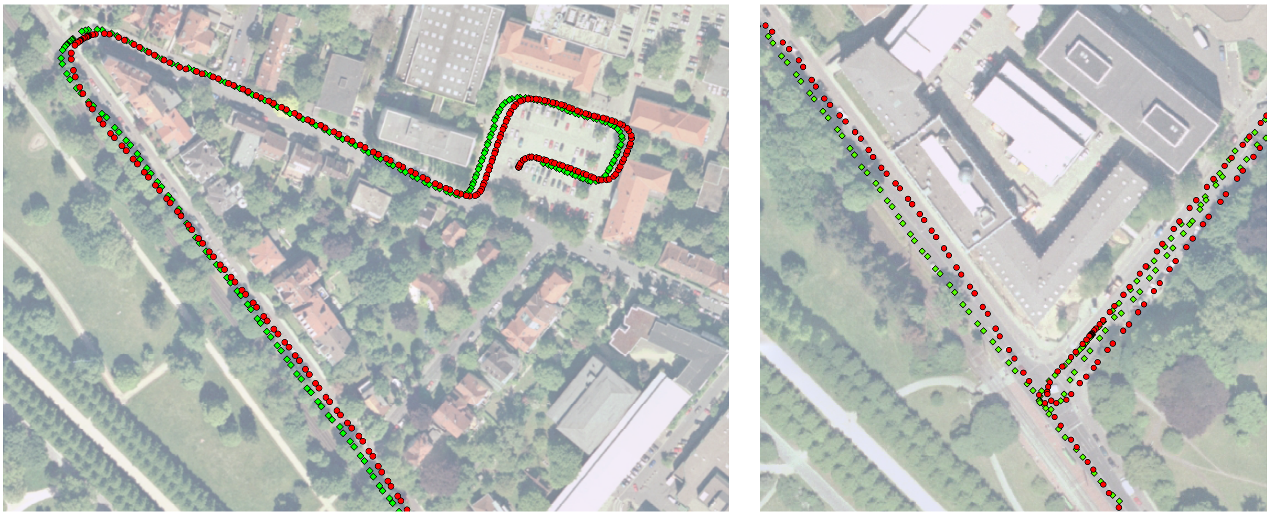


Figure 12: Comparison of original vehicle trajectories (red dots), as obtained from the RTS Hanna GPS/odometry positioning system and corrected trajectories (green squares), using pole matching.

cal matching algorithm failed so far. Sometimes, too few poles lead to failure. On the other hand, along the alleys, there are often too many poles and our local matching algorithm finds wrong correspondences, leading to a wrong transformation.

6 DISCUSSION AND OUTLOOK

In this paper, we presented an approach to use automatically extracted landmarks for the positioning of vehicles. We focused on pole-like objects and extracted them both in a high-resolution, high accuracy scan (the refereence database) as well as in low-resolution, low accuracy scans of a robotic vehicle, using two different algorithms. We assessed the accuracy of the extracted poles and made first experiments regarding the improvement of the trajectory using a local pole matching.

As we mentioned, pole matching did not succeed in all cases. Too few as well as too many poles may cause problems, and erroneous poles in the vehicle scans and in the database lead to wrong feature associations and thus, transformations. We think that a key to a successful matching is the short-term stability of the vehicle trajectory. If a high precision of the poles can be obtained locally, many false matches can be ruled out by tight distance bounds. However, even then, there is sometimes the problem of too few poles in the database and it is obvious that a key element for a reliable localization will be to include additional landmarks.

For the future, we plan to include such additional features, e.g. local planes. Also, the local matching algorithm should be extended and all measurements should be included in a Kalman filter solution to estimate the vehicle position.

ACKNOWLEDGEMENTS

We thank Matthias Hentschel from the Real Time Systems Group at the Institute for Systems Engineering of the Leibniz Universität Hannover for providing the RTS Hanna scans.

REFERENCES

Borrmann, D., Elseberg, J., Lingemann, K., Nüchter, A. and Hertzberg, J., 2008. Globally consistent 3D mapping with scan matching. Journal of Robotics and Autonomous Systems (JRAS) 56(2), pp. 130–142.

Burgard, W. and Hebert, M., 2008. Springer Handbook of Robotics. Springer, chapter World Modeling, pp. 853–869.

EDMap, 2004. Enhanced digital mapping project final report. Technical report, United States Department of Transportation, Federal Highway Administration and National Highway Traffic and Safety Administration, http://ntl.bts.gov/lib/jpodocs/repts_te/14161.htm, (accessed 30 Nov. 2009).

IBEO, 2009. Ibeo Lux laser scanner. http://www.ibeo-as.de (accessed 30 Nov. 2009).

Kremer, J. and Hunter, G., 2007. Performance of the streetmapper mobile lidar mapping system in real world projects. In: Photogrammetric Week 2007, Wichmann, pp. 215–225.

Kukko, A., Andrei, C.-O., Salminen, V.-M., Kaartinen, H., Chen, Y., Rönnholm, P., Hyyppä, H., Hyyppä, J., Chen, R., Haggrén, H., Kosonen, I. and Ĉapek, K., 2007. Road environment mapping system of the finnish geodetic institute – FGI roamer. In: IAPRS (ed.), Proc. Laser Scanning 2007 and SilviLaser 2007, Vol. 36 Part 3/W 52, pp. 241–247.

Pandazis, J.-C., 2002. NextMAP: Investigating the future of digital map databases. In: e-Safety Congress, Lyon, http://www.ertico.com/en/activities02/ completed_projects/ websites/ nextmap_website.htm (accessed 30 Nov. 2009).

Rabbani, T., Dijkman, S., van den Heuvel, F. and Vosselman, G., 2007. An integrated approach for modeling and global registration of point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 61(6), pp. 355–370.

Thrun, S., Burgard, W. and Fox, D., 2005. Probabilistic Robotics. The MIT Press, Cambridge, Mass.

Weiss, T., Kaempchen, N. and Dietmayer, K., 2005. Precise ego localization in urban areas using laserscanner and high accuracy feature maps. In: Proc. 2005 IEEE Intelligent Vehicles Symposium, Las Vegas, USA, pp. 284–289.

TEMPORALLY ADAPTIVE A* ALGORITHM ON TIME-DEPENDENT TRANSPORTATION NETWORK

N. B. Zheng [*], F. Lu

Institute of Geographic Sciences and Natural Resources Research, CAS, 11A Datun Road, Chaoyang District, Beijing, 100101, China – (zhengnb, luf)@lreis.ac.cn

KEY WORDS: Shortest Path; A* algorithm; Time-Dependent Network; FIFO condition; Traffic Information

ABSTRACT: Traditional solutions to shortest path problems on time-varying transportation networks only use traffic information at definite moment so as to ignore the fact that the travel time through a link is dependent on the time to enter it. In this paper, the travel speed instead of the travel time on each link of road networks was modelled as a time-interval dependent variable, and a FIFO-satisfied computational function of the link travel time was then deduced. At last, a temporally adaptive A* shortest path algorithm on this FIFO network was presented, where the time factor was introduced into the evaluation function, and the Euclidean distance divided by the maximum possible travel speed was used as heuristic evaluator. An experiment on the real road network shows that the proposed algorithm is capable of foreseeing and bypassing those forthcoming traffic congestions, only with a cost of about 10 percent more computational time than the traditional algorithm. Furthermore, frequent path reoptimization caused by the traditional algorithm gets avoided effectively.

1. INTRODUCTION

It is increasingly necessary for a vehicle navigation system or a map search website to calculate fastest paths by using real-time, historical, or predicted traffic information (Fawcett & Robinson, 2000; Yamane *et al.*, 2004; Ishikawa, 2005). Traditional solutions to this problem only consider traffic information at definite moment, and calculate optimal paths by either classic or heuristic shortest path algorithms including Dijkstra, A*, branch pruning, hierarchical search, etc (Lu & Guan, 2004; Fu *et al.*, 2006; Klunder *et al.*, 2006; Cho & Lan, 2009). As a result, while travelling, the planned path will have to be re-optimized frequently to respond the periodical renewal of that near-real-time traffic information (e.g., every five minutes). Obviously, this process will be considerably time-consuming, even if some practical accelerating techniques, such as dynamic window scheme (Kim & Jung, 2002) and incremental search approach (Huang & Wu, 2007), are introduced into it. Besides, this type of algorithms is incapable of computing the path and evaluating the travel time for a trip from an overall perspective. The reason is that they have overlooked the fact that the travel time through a link is dependent on the time to enter it. Therefore, the time-dependent shortest path algorithms are needed for the dynamic route planning.

The time-dependent shortest path problem (TDSPP) models the travel time through a link as an arrival-time-dependent variable, and can be resolved by some modified labelling algorithms (i.e., Dijkstra, A*, etc), only if the First-In-First-Out (FIFO) premise is satisfied (Kaufman & Smith, 1993; Sung *et al.*, 2000; Horn, 2000; Chabini & Lan, 2002; Kanoulas *et al*., 2006). In practice, the travel time through a link is usually taken as time-interval dependent in a nutshell, which ignores the fact that the travel speed may vary with time intervals during a trip along this link, and at the same time, violates the FIFO condition. In view of this, this paper defines the travel speed rather than the travel time on a link as a time-interval dependent variable, and then computes the FIFO-satisfied time-dependent link travel times, which will be used for the temporal adaption of the A* shortest path algorithm at last.

The remains of this paper are organized as follows. Section 2 defines a time-dependent network with time-interval dependent link travel speeds. Section 3 derives a computational function of FIFO-satisfied link travel time. Section 4 presents a temporally adaptive A* algorithm. Section 5 tests the proposed algorithm on a real road network. Conclusions are drawn in section 6.

2. TIME-DEPENDENT NETWORK

Suppose the time T is segmented into the following intervals: $[t_0, t_1)$, $[t_1, t_2)$, ..., $[t_k, t_{k+1})$ ($k = 0, 1, ..., m–1$, $t_k < t_{k+1}$), then the time-dependent network is defined as $G = (N, A, L, V, W)$, where $N = \{0, 1, ..., n–1\}$ denotes the node set; $A \subseteq \{(i, j)|(i, j) \in N \times N\}$ denotes the directed link set; $L = \{l_{ij}|(i, j) \in A, l_{ij} > 0\}$ denotes the link length set; $V = \{f_{ij}(t)|(i, j) \in A, t \in T\}$ denotes the set of the time-interval dependent link travel speeds, and for each $t \in [t_k, t_{k+1})$, have $f_{ij}(t) = v_{(i,j)k}$; $W = \{w_{ij}(t)|(i, j) \in A, t \in T\}$ denotes the set of the time-dependent link travel times, where $w_{ij}(t)$ denotes the travel time along link (i, j) when departing from node i at time t.

Let $T_{ij}(t)$ denote the arrival time to node j with departure time t from node i along link (i, j), have $T_{ij}(t) = t + w_{ij}(t)$. Let $p = \{i_1, i_2, ..., i_u\}$ denote a path from node i_1 to node i_u, then the path arrival time function of path p is given by the composition of the link arrival time functions along p:

$$T_p(t) = T_{i_{u-1}i_u}(T_{i_{u-2}i_{u-1}}(T_{i_{u-2}i_{u-3}} \cdots \ T_{i_1i_2}(t))) \ .$$

Let $P(o, d)$ denote the path set from origin node o to destination node d, $ET_{od}(t)$ denote the earliest arrival time while leaving from origin node o at time t to destination node d, then:

$$ET_{od}(t) = \min\{T_p(t) : p \in P(o,d)\} \ .$$

[*] Corresponding author. Email: zhengnb@lreis.ac.cn.

Obviously, the time-dependent shortest path problem is just a problem of computating the earliest arrival time $ET_{od}(t)$.

(FIFO conditon) For each link $(i, j) \in A$, if the following inequality is satisfied, we call this link FIFO-satisfied:

$$T_{ij}(s) \le T_{ij}(t) \quad \forall s \le t\,.$$

If every link of a time-dependent network is FIFO satisfied, we call this network a FIFO network. Kaufman & Smith (1993) has proved: the shortest path problem on the FIFO network can be well dealt with by the traditional labelling algorithms.

3. COMPUTATION OF LINK TRAVEL TIME

Consider a link $(i, j) \in A$ of length l (for simplicity, we drop the subscript (i, j) in this section) with travel speed v_k changing with the time intervals $[t_k, t_{k+1})$ ($k = 0, 1, ..., m–1$). If a vehicle sets off from node i at time t and reaches node j at time $T(t)$, then its travel time along this link will be $T(t) – t$. The computational procedure of $T(t)$ is as follow:

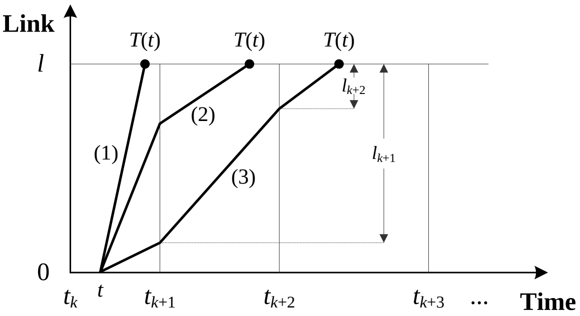


Fig. 1 A trip along a link with time-interval-varying speeds

1) if $l_k - v_k \times (t_{k+1} - t) < 0$ (See Fig. 1(1)), then
$$T(t) = t + l_k / v_k\,,$$
else
2) if $l_{k+1} - v_{k+1} \times (t_{k+2} - t_{k+1}) < 0$ (See Fig. 1(2)), then
$$T(t) = t_{k+1} + l_{k+1} / v_{k+1}\,,$$
else
3) if $l_{k+2} - v_{k+2} \times (t_{k+3} - t_{k+2}) < 0$ (See Fig. 1(3)), then
$$T(t) = t_{k+2} + l_{k+2} / v_{k+2}\,,$$
else
$\vdots$
i) if $l_{k+i-1} - v_{k+i-1} \times (t_{k+i} - t_{k+i-1}) < 0$, then
$$T(t) = t_{k+i-1} + l_{k+i-1} / v_{k+i-1} \qquad \forall i = 2,3...\,,$$
else
$\vdots$

where
$$\begin{aligned} l_k &= l \\ l_{k+1} &= l_k - v_k \times (t_{k+1} - t) \\ l_{k+2} &= l_{k+1} - v_{k+1} \times (t_{k+2} - t_{k+1}) \\ &\vdots \\ l_{k+i} &= l_{k+i-1} - v_{k+i-1} \times (t_{k+i} - t_{k+i-1}) \\ &\vdots \end{aligned}$$

In practice, if the time intervals are long enough, a travel along the link will cover no more than two time intervals, namely, $[t_k, t_{k+1})$, $[t_{k+1}, t_{k+2})$. In this case, the above computational procedure can be reduced to:

$$T(t) = \begin{cases} t + l/v_k\,, & t \in [t_k,\ t_{k+1} - l/v_k) \\ t_{k+1} + [l - v_k(t_{k+1} - t)]/v_{k+1}\,, & t \in [t_{k+1} - l/v_k,\ t_{k+1}) \end{cases}.$$

The FIFO satisfaction of $T(t)$ is illustrated by Fig. 2 and Fig. 3. Assume $T(t)$ violates the FIFO condition, that is, an overtaking happens: $T(s) > T(t)\ \forall s \le t$, two trajectories are sure to intersect at some position, e.g., point C in Fig. 2. Thus, there will have two different speeds in the same time interval, e.g., $[t_{k+2}, t_{k+3})$ in Fig. 2. In following two figures, speed value is indicated by the slope of the trajectory. Apparently, this violates the definition of the time-dependent network in Section 2. Consequently, it can be surely concluded that $T(t)$ satisfies the FIFO condition. In fact, two travel trajectories departing at different times will be parallel in the FIFO network, as shown in Fig. 3.

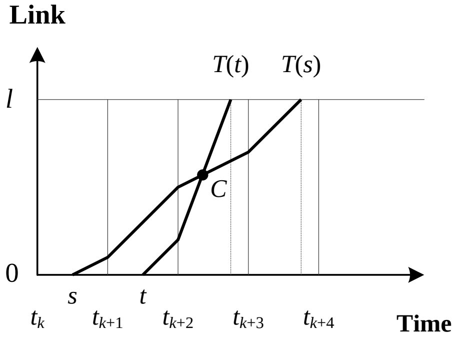


Fig. 2 FIFO condition violated

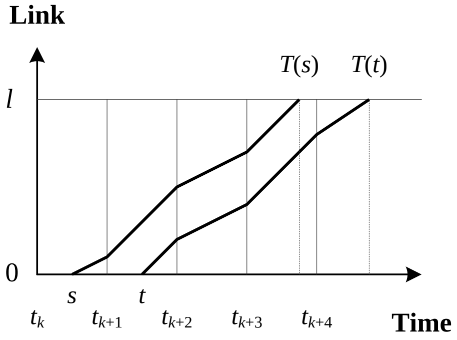


Fig. 3 FIFO condition satisfied

4. TEMPORALLY ADAPTIVE A* ALGORITHM

The A* shortest path algorithm was first proposed by Hart *et al.* (1968), and further proved applicable to road networks by Fu *et al.* (2006) and Zeng & Church (2009). The A* algorithm makes use of a heuristic evaluation function $F_i = L_i + e_{(i,d)}$ as a label for node i, where L_i is the travel time of the current evaluated path from the origin node to node i; $e_{(i,d)}$ is an estimated travel time from node i to node d. The sum of these two functions, F_i, is the weight of node i, and reflects the likelihood of node i being on the shortest path.

As for the time-dependent network, it is necessary to introduce time factor into the evaluation function: $F_i(t) = T_i(t) + e_{(i,d)}(T_i(t))$, where t denotes the departure time from the origin node; $T_i(t)$ denotes the arrival time of the current path from the origin node at the time t to the node i; $e_{(i,d)}(T_i(t))$ is an evaluated travel time from the node i at time $T_i(t)$ to the destination node d. $e_{(i,d)}(T_i(t))$ controls the accuracy and efficiency of the algorithm. Chabini & Lan (2002) has proved that if $e_{(i,d)}(T_i(t))$ is not more than the real minimum travel time from node i to the destination node d with departure time t, the temporally adaptive A* algorithm will be strictly optimal. In view of this, let

$$e_{(i,d)}(T_i(t)) = \frac{D_{(i,d)}}{V_{\max}}\ ,$$

where $D_{(i,d)}$ denotes the Euclidean distance from node i to the destination node d; $V_{\max}$ denotes the maximum possible travel speed.

Let o denotes the origin node, d denotes the destination node, t denotes the departure time, P_i denotes the straight preceding node of node i along the shortest path, Q denotes the scan eligible node set, and R denotes the permanent labelled node set (That is, the shortest path from the origin node to any node in this set has been found), then the basic steps of the temporally adaptive A* (TAA*) algorithm are as follows:

```
Step 1: Initialization: Set $T_o = t$; $F_o = T_o + e_{(o,d)}(T_o)$; $F_j = T_j = \infty\ \forall j \neq o$; $P_o = -1$; $Q = \{o\}$; $R = \varnothing$;
Step 2: Node Selection: Select and remove the node $i$ with the
        minimum label $F_i(t)$ from the scan eligible node set $Q$,
        and label it permanently:
            $i = \mathrm{argmin}_{j \in Q}\{F_j\}$; $Q = Q \backslash \{i\}$; $R = R \cup \{i\}$;
        if $i == d$, the shortest path has been found, then
            STOP;
        else
Step 3: Node Expansion: Scan each link outgoing from node $i$.
        For each link $(i, j) \in A$ & $j \notin R$, if
            Arri_time($(i, j), T_i$) + $e_{(j,d)}$(Arri_time($(i, j), T_i$)) < $F_j$,
        then
            $T_j$ = Arri_time($(i, j), T_i$); $F_j = T_j + e_{(j,d)}(T_j)$; $P_j = i$;
            insert node $j$ into $Q$: $Q = Q \cup \{j\}$;
        where Arri_time($(i, j), T_i$) is a procedure of computing
        the arrival time of link $(i, j)$ according to Section 3;
Step 4: Termination: If $Q = \varnothing$, then STOP;
        else: goto step 2.
```

Tab.1 Default speed values of different road types (Speed: km/h)

	Expressway	Arterial	Sub-arterial	Minor	Overpass	Ramp
Monday morning peak (7:00-9:00) & Friday evening peak (17:00-19:00)	35	35	35	28	35	30
Other peaks (7:00-9:00; 17:00-19:00)	45	40	45	40	30	30
Nights (21:00-24:00; 0:00-6:00)	60	60	60	50	45	45
Others (6:00-7:00; 9:00-17:00; 19:00-21:00)	55	55	55	50	55	45

Tab 2 A naive model for time-dependent turn delays (Time: min)

	Straight	Right	Left	U-turn	Others
Peaks (7:00-9:00; 17:00-19:00)	1.0	0.5	1.0	0.8	0.3
Night (21:00-24:00; 0:00-6:00)	0.5	0.0	1.0	0.5	0.5
Others (6:00-7:00; 9:00-17:00; 19:00-21:00)	0.5	0.1	1.0	0.5	1.0

5. EXPERIMENT

This paper implements and tests the proposed algorithm in the self-developed Urban Public Travel Path Service System with runtime environment: dual-core CPU 1.6GHz, RAM 1.0GB, OS Windows XP professional. The experimental data include road network data within the Beijing's Fourth Ring and carriageway-based floating car traffic data of July, August, and September, 2007. As for the algorithm, the scan eligible node set is realized by quad-heap priority queue (Lu *et al.*, 1999); the maximum possible travel speed is set to 60km/h. Besides, in order to deal with turn delays and prohibitions in the transportation network, an arc-labelling strategy is introduced (Gao & Lu, 2008). To be specific, the node and the link exchange their roles each other and the turn delays are accumulated into the path arrival time.

5.1 Data Preparation

The topology of the road network is carriageway-based, and one carriageway corresponds to one link of the time-dependent network in Section 2. After generalizing those virtual links and virtual nodes that represent a same road intersection into one topological node, the network contains 53997 links in all.

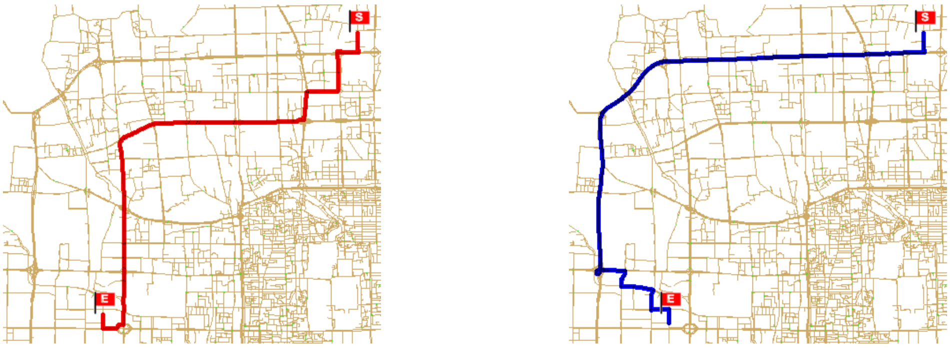
The traffic data of one day are discreted into 288 time intervals, and each interval is 5 minute duration. The traffic data items in each time interval include: carriageway ID, carriageway length, travel time, and congestion level. In order to support the TAA* algorithm, we convert these raw travel time data into the travel speed data by the division of carriageway length by travel time. Furthermore, in view of the cycling nature of the urban traffic flow, we average those speed values with the same day category and the same time interval, which results in 7 speed data files.

For the carriageways not covered by floating car traffic data, we give them default speed values according to the road type and the entrance time, as shown in Tab. 1. Moreover, we build a naive time-dependent turn delay model: *Turn* = 0.5 × *f*(*time*, *turn_type*), where 0.5 denotes the probability of waiting at intersections; *f*(*time*, *turn_type*) is evaluated as Tab. 2.

5.2 Experimental Results

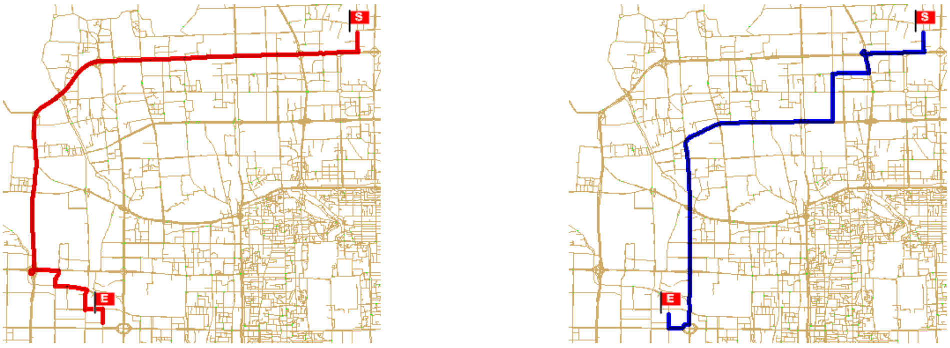
We choose the Beichendong Road as an origin road and the Cuiwei Road as an destination road, set 7:00 and 8:00 on Tuesday as two departure times, and then compute four least-time paths by the TAA* algorithm and the real-time A* (RTA*) algorithm respectively. Computational results are shown in Fig. 4 and Fig. 5. As compared with the TAA* algorithm, the only and significant distinct of the RTA* algorithm is that the RTA* algorithm merely uses the traffic data at the departure time.

As seen from the Fig. 4 and Fig. 5, while departing at 7:00, the TAA* algorithm foresees and bypasses the forthcoming traffic congestion on the Third Ring Road, but the RTA* algorithm do not. While departing at 8:00, the TAA* algorithm predicts that the current traffic congestion on the Third Ring Road will die away at the entrance time, and therefore avoids an unwanted detour caused by the RTA* algorithm.



(a) RTA* (b) TAA*

Fig. 4 Comparison of computational results of TAA* and RTA* (Departure time: 7:00)



(a) RTA* (b) TAA*

Fig. 5 Comparison of computational results of TAA* and RTA* (Departure time: 8:00)

Furthermore, we select 30 pairs of Origin-Destination (O-D) carriageways, set 7:30 on Tuesday as departure time, and then calculate the least time path between each O-D pair by the TAA* algorithm, RTA* algorithm and RTA*_M algorithm separately. The comparisons between them are shown in Fig. 6, Fig. 7 and Fig. 8 (The 30 paths are sorted by length). Thereinto, the RTA*_M algorithm works as follow: i) At the beginning of the trip, compute an optimal path by the RTA* algorithm; ii) during the trip, re-optimize the path by the RTA* algorithm as long as new traffic information arrives (This is simulated by the historical traffic data); iii) repeat step ii) until the destination is reached. So the computational time of the RTA*_M algorithm is the sum of that of each RTA* algorithm, and the path of the RTA*_M algorithm is that actual travelled path.

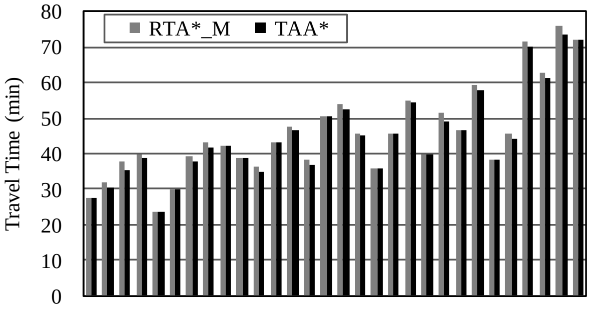


Fig. 6 Comparison of path travel times of TAA* and RTA*_M

As seen from the Fig. 6, although the low traffic information coverage and the immature of the traffic prediction model and the turn delay model make a few paths less differentiated, the path travel times of the TAA* algorithm are less than those of the RTA*_M algorithm on the whole.

It is concluded from the Fig. 7 that the TAA* algorithm will cost about 10 percent more computational time than the RTA* algorithm. The extra time is exhausted mainly by the procedure of link travel time computation. Its time complexity is O(m), where m denotes the number of the time intervals of the link.

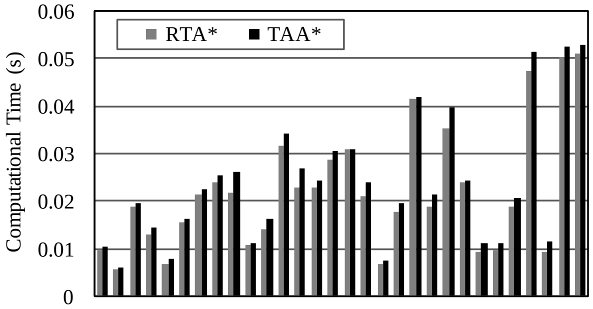


Fig. 7 Comparison of computational times of TAA* and RTA*

It can seen from the Fig. 8 that the TAA* algorithm cost much less computational time than the RTA*_M algorithm. The reason is that the TAA* algorithm only need to carry out the computational procedure once because it has taken the future traffic status into account already in advance, but comparatively, the RTA*_M algorithm must execute the computational procedure again and again because of its "short views".

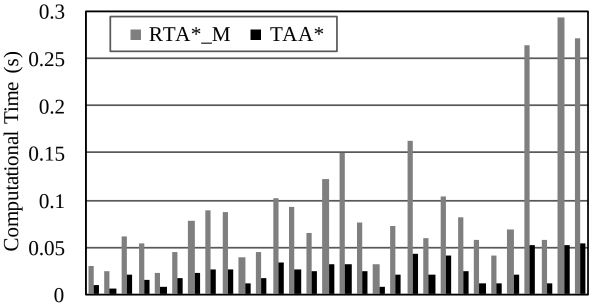


Fig. 8 Comparison of computational times of TAA* and RTA*_M

6. CONCLUSIONS

This paper suggests a novel time-dependent network model, where the travel speed instead of the travel time is regarded as changing with different time intervals, and from which a FIFO-satisfied computational function of link travel time is derived. Furthermore, this paper develops a temporally adaptive A* algorithm to calculate least-time paths on the defined time-dependent network, where the link-travel-time computational function was used to the evaluation function, and the Euclidean distance divided by the maximum possible travel speed was designed as the heuristic evaluator. An experiment on the real road network shows that the proposed algorithm is capable of foreseeing and bypassing those forthcoming traffic congestions, saving the travel time, and raising the overall efficiency of the navigation system.

The implementation of the TAA* algorithm needs the support of link travel speed data and turn delay data. This paper only takes the simple means of historical link travel speed data as future link travel speeds. Besides, turn delays between links are set empirically rather than theoretically. Therefore, we will keep on our research work on accurate and available link travel time prediction method and turn delay measure model in the near future.

Acknowledgements:
The work described in this paper is supported by the National Natural Science Foundation of China (40871184), the China Postdoctoral Science Foundation (20090450563) and the National High Technology Research and Develop Program of China (2007AA12Z241).

References:

Chabini, I., Lan, S., 2002. Adaptations of the A* algorithm for the computation of fastest paths in deterministic discrete-time dynamic networks. *IEEE Transactions on Intelligent Transportation Systems*, 3(1), pp. 60-74.

Cho, H.J., Lan, C.L., 2009. Hybrid shortest path algorithm for vehicle navigation. *The Journal of Supercomputing*, 49(2), pp. 234-247.

Fawcett, J., Robinson, P., 2000. Adaptive routing for road traffic. *IEEE Computer Graphics and Applications*, 20(3), pp. 46-53.

Fu, L., Sun, D., Rilett, L.R., 2006. Heuristic shortest path algorithms for transportation applications: state of the art. *Computers & Operation Research*, 33(11), pp: 3324-3343.

Gao, S., Lu, F., 2008. An arc-labeling shortest time path algorithm. *Geo-Information Science*, 10(5), pp. 604-610. (In Chinese)

Hart, E.P., Nilsson, N.J., Raphael, B., 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transaction of System Science and Cybernetics*, SSC-4(2), pp. 100-107.

Horn, M., 2000. Efficient modelling of travel in networks with time-varying link speeds. *Networks*, 36(2), pp. 80-90.

Huang, B., Wu, Q., Zhan, F.B., 2007. A shortest path algorithm with novel heuristics for dynamic transportation networks. *International Journal of Geographical Information Science*, 21(6), pp. 625-644.

Ishikawa, H., 2005. Development of a local storage type traffic prediction for car navigation system. In: *The 12th World Congress on Intelligent Transportation System*, San Francisco, USA.

Lu, F., Guan, Y., 2004. An optimum vehicular path solution with multi-heuristics. M. Bubak *et al.* (Eds.): ICCS 2004, *Lecture Notes in Computer Science*, 3039, pp. 964-971.

Lu, F., Lu, D., Cui, W., 1999. Improved Dijkstra algorithm based on quad-heap priority queue and inverse adjacent list. *Journal of Image and Graphics*, 4A(12), pp. 1044-1050. (In Chinese)

Kanoulas, E., Du Y., Xia, T., Zhang, D., 2006. Finding fastest paths on a road network with speed patterns. In: The 22nd International Conference on Data Engineering, Atlanta, USA.

Kaufman, D.E., Smith, R.L., 1993. Fastest paths in time-dependent networks for intelligent vehicle-highway systems application. *IVHS Journal*, 1, pp. 1-11.

Kim, J., Jung, S., 2002. A dynamic window-based approximate shortest path re-computation method for digital road map databases in mobile environments. M.H. Shafazand and A.M. Tjoa (Eds.): EurAsia-ICT 2002, *Lecture Notes in Computer Science*, 2510, pp. 711-720.

Klunder, G.A., Post, H.N., 2006. The shortest path problem on large-scale real-road networks. *Networks*, 48(4), pp. 182-194.

Sung, K., Bell, M.G.H, Seong, M., Park, S., 2000. Shortest paths in a network with time-dependent flow speeds. *European Journal of Operational Research*, 121(1), pp. 32-39.

Yamane, K., Endo, Y., Fujiwara, J., Machii, K., 2004. Statistical traffic information for navigation system. In: The *11th World Congress on Intelligent Transportation System*, Nagoya, Aichi, Japan.

Zeng, W., Church, R.L., 2009. Finding shortest paths on real road networks: the case for A*. *International Journal of Geographical Information Science*, 23(4), pp. 531-543.

WAVELET DE-NOISING OF TERRESTRIAL LASER SCANNER DATA FOR THE CHARACTERIZATION OF ROCK SURFACE ROUGHNESS

K. Khoshelham*, D. Altundag

Optical and Laser Remote Sensing, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands - (k.khoshelham@tudelft.nl)

KEY WORDS: Laser scanning, Roughness length, Measurement noise, Wavelet decomposition, Thresholding, Fractal dimension.

ABSTRACT:

The application of terrestrial laser scanning to the study of rock surface roughness faces a major challenge: the inherent range imprecision makes the extraction of roughness parameters difficult. In practice, when roughness is in millimeter scale it is often lost in the range measurement noise. The parameters extracted from the data, therefore, reflect noise rather than the actual roughness of the surface. In this paper we investigate the role of wavelet de-noising methods in the reliable characterization of roughness using laser range data. The application of several wavelet decomposition and thresholding methods are demonstrated, and the performances of these methods in estimating roughness parameters are compared. As the main measure of roughness fractal dimension is derived from 1D profiles in different directions using the roughness length method. It is shown that wavelet de-noising in general leads to an improved estimation of the fractal dimension for the roughness profiles. The choice of the decomposition method is shown to have a minor effect on the de-noising results; however, the application of hard or soft thresholding mode does have a considerable influence on the estimated roughness measures. The presented results suggest that hard thresholding yields more accurate de-noised profiles for which the estimated roughness measures are more reliable.

1. INTRODUCTION

The measurement of the surface roughness of rock masses has been traditionally based on manual measurement tools such as carpenter’s comb and compass and disc clinometers. The manual measurements are limited to small samples at accessible parts of the rock. Terrestrial laser scanning is an attractive alternative measurement technique, which offers large coverage, high resolution, and the ability to reach inaccessible high rock faces. A fundamental limitation of this technique, particularly in the characterization of rock surface roughness in millimeter scale, is the measurement noise inherent in laser scanner data. In general, error in laser scanner data may originate from three main sources: the imprecision of the scanning mechanism and the ranging technique (Dorninger et al., 2008), environmental conditions (Borah and Voelz, 2007) and the physical and geometric properties of the scanned surface itself (Soudarissanane et al., 2009). Normally, the systematic components of the error are eliminated or modeled through a proper calibration procedure (Lichti, 2007). The remaining random error is in the order of a few millimeters for a typical medium-range (1-150 m) terrestrial laser scanner, and is commonly referred to as measurement noise.

The effect of laser scanner measurement noise on roughness characterization has been pointed out in a few previous studies. Fardin et al., (2004) reported that the fractal dimension obtained from raw laser data of a rock face is larger than the expected range (according to Kulatilake and Um (1999) 1.2-1.7 for 1D profiles, and 2.2-2.7 for 2D patches). They attributed the overestimated roughness to the irregular distribution of the points in the original point cloud, and performed an interpolation of the points into a uniform distribution to reduce the fractal dimension to within the expected range. Rahman et al., (2006) suggested that the overestimation of surface roughness obtained from raw laser data is due to the fact that roughness measures reflect more noise in the data than the actual roughness of the surface. They used radial basis functions to interpolate the data into a smooth surface, which resulted in roughness measures within the expected range. Although data smoothing by interpolation has been the common approach to reduce the influence of noise on roughness characterization, it is generally not considered an adequate noise reduction method (Gonzalez and Woods, 1992). The basic assumption in data smoothing is that the measured surface is actually smooth and so by smoothing one can reduce the noise without degrading the data related to the actual surface. As this assumption is not valid when dealing with rough surfaces, the result of data smoothing is the loss of roughness information. Thus, a careful treatment of noise in laser range data is of great significance if a realistic characterization of rock surface roughness is of concern. In this paper we investigate the influence of range measurement noise on roughness characterization of rock surfaces using the roughness length method (Malinverno, 1990). We demonstrate the application of wavelet transform (Hardle et al., 1998; Strang and Nguyen, 1996) to removing noise from roughness profiles derived from laser scanner point clouds, and compare the performance of various wavelet decomposition and thresholding methods in the context of surface roughness characterization.

The paper proceeds with an overview of the laser scanning technique and the derivation of roughness profiles from laser range data in Section 2. In Section 3, the principles of wavelet-based de-noising are presented along with a description of various decomposition and thresholding methods. Section 4 reports the experimental analysis of the influence of noise on roughness characterization and the results of wavelet de-noising of roughness profiles. The paper concludes with some remarks in Section 5.

* Corresponding author.

2. ROCK SURFACE ROUGHNESS FROM LASER RANGE DATA

Laser scanning is an active measurement technique based on emitting laser beams to a surface of interest and recording the reflections. A scanning mechanism, usually a rotating mirror, deflects the emitted beam towards the surface in such a way that the entire surface is scanned at regular horizontal and vertical angular intervals. The range measurement principle in medium-range terrestrial laser scanners is most often based on the phase difference between the emitted and received waveforms. From the measured range and horizontal and vertical scan angles, 3D coordinates are computed for each point in a Cartesian coordinate system with its origin at the centre of the scanner. Today's laser scanners can measure more than a hundred thousand points per second at an angular resolution smaller than 0.01 degrees (see for instance Faro (2009)). By scanning at such high resolution from a few tens of meters distance to a rock face one can acquire a dense point cloud that represents the geometry of the scanned surface in great detail.

Before roughness information is derived from a point cloud it is convenient to rotate the point cloud such that surface roughness corresponds to variations in the direction of *Z* axis. Based on the assumption that the point cloud represents a more or less flat surface, the rotation can be computed simply by performing the principal components analysis (Jolliffe, 2002). The eigenvectors and eigenvalues of the covariance matrix of the points describe the axes of maximum and minimum variation in the point cloud, and provide a transformation of the points to these principal axes. By fitting a smooth (usually planar) surface to this rotated point cloud a representation of the roughness as the residual height of the points can be obtained.

A common method for roughness characterization, which is also adopted in this paper, is the fractal-based roughness length method (Malinverno, 1990). In this method, roughness is characterized by two measures: fractal dimension and amplitude. Both measures can be derived from a 1D profile or a 2D patch extracted from the point cloud. In either case, the roughness measures are estimated based on a power law relation between the standard deviation of the residual height of the points, *s*, and the length of a sampling window *w*:

$$s(w) = Aw^{H} \tag{1}$$

where parameters *A* and *H* are called amplitude and the Hurst exponent respectively. These parameters are estimated from the intercept and slope of a log-log plot of *s* versus *w* for several lengths of the sampling window. The main measure of roughness is the fractal dimension, which is derived from the Hurst exponent as D = 2-H for a 1D profile, and D = 3-H for a 2D patch. A large fractal dimension indicates a very rough surface with abrupt changes of the residual height whereas a small fractal dimension implies a smooth surface without much roughness. More details on the estimation of fractal dimension for 1D profiles can be found in Kulatilake and Um (1999), and for 2D patches in Fardin et al., (2004). In the rest of the paper we focus on the characterization of roughness in 1D profiles.

3. WAVELET DE-NOISING OF ROUGHNESS PROFILES

Wavelet de-noising is based on the wavelet transform (Strang and Nguyen, 1996) for decomposing a signal into several components of different scale and resolution. The basic principle is that high-frequency components are more likely to contain noise than low-frequency components that contain the general trend of the signal. The purpose of wavelet decomposition in de-noising laser range data is to remove noise only from the high frequency components so as to preserve the low frequency content of the data as much as possible. The procedure for the wavelet de-noising of a roughness profile consists of several steps as shown in Fig. 1. The first step is the decomposition, which can be done by the discrete wavelet transform or by the wavelet packet method. The actual de-noising is performed by applying a threshold to the high-frequency components. The value of the threshold depends on an estimation of the level of noise in the data and the threshold selection method. The application of the threshold can also be done in the hard as well as soft mode. The final step involves the reconstruction of the thresholded components to yield the de-noised profile. The following sections provide a more detailed description of the wavelet de-noising procedure.

3.1 Wavelet decomposition and reconstruction

The wavelet decomposition process consists of two operations: filtering and downsampling. Filtering separates the signal into components of different scale: convolution with a low-pass filter generates the low-frequency components known as approximation coefficients, and convolution with a high-pass filter results in the high frequency components known as detail coefficients. The downsampling operation reduces the resolution of the coefficients to one-half. The decomposition process may be iterated in several levels. In multi-level decomposition we distinguish between two decomposition principles. In the discrete wavelet transform (DWT), the decomposition is applied to approximation coefficients only. In the wavelet packet method (WP) both the approximations and the details are decomposed.

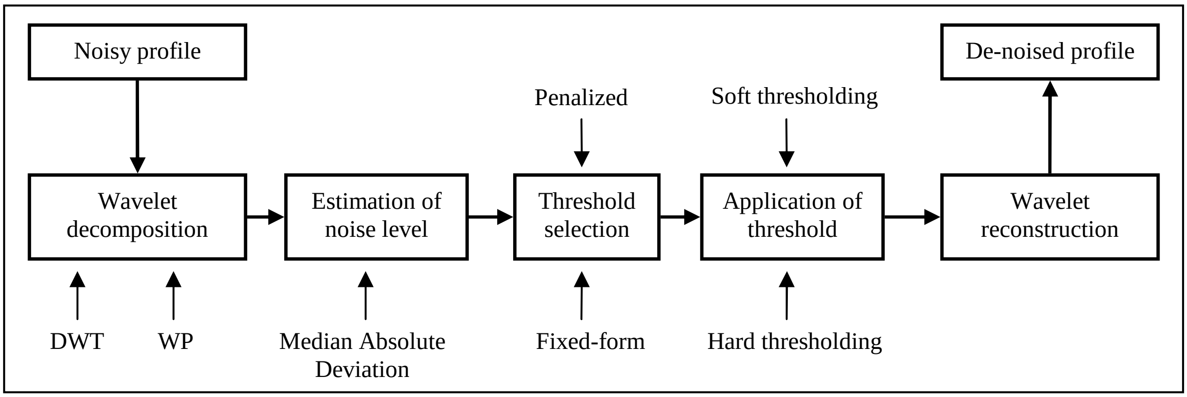


Fig. 1. Wavelet de-noising procedure.

Wavelet reconstruction is the process of recovering the original profile from its components. The reconstruction process consists of two operations: upsampling and filtering. The components are upsampled by inserting zeros between the samples and then convolved with the reconstruction filters. The approximation coefficients are convolved with a dual low-pass filter, and the detail coefficients are convolved with a dual high-pass filter. The reconstructed approximations and details are then summed up to yield the reconstructed profile. The decomposition and reconstruction filters should meet certain requirements in order to guarantee a perfect reconstruction of the data from the coefficients. A detailed description of the design of wavelet filters can be found in Strang and Nguyen (1996).

3.2 Thresholding of wavelet coefficients

De-noising by the thresholding of wavelet coefficients is based on an important property of wavelet decomposition that transforms white noise into white noise (Donoho and Johnstone, 1995). Since normally systematic errors are eliminated from the laser scanner data it is prudent to assume that the remaining error is white noise with Gaussian distribution. The thresholding is usually applied to the detail coefficients to ensure the preservation of the actual data. There are several methods for the estimation of the threshold value. In this paper, we compare two main threshold estimation methods: fixed-form thresholding and penalized thresholding.

The fixed-form thresholding method was proposed by Donoho and Johnstone (1994). For the detail coefficients of a profile obtained by the discrete wavelet transform the fixed-form threshold is estimated as:

$$t^f = \sigma_n \sqrt{2\log(d)} \qquad (2)$$

where d is the length of the detail coefficients at the first level of decomposition, and σ_n is the standard deviation of noise. For the wavelet packet decomposition of a profile the fixed-form threshold is estimated as:

$$t^f = \sigma_n \sqrt{2\log\left(d\log(d)/\log(2)\right)} \qquad (3)$$

To estimate the standard deviation of noise from the data the median absolute deviation (MAD) of the coefficients has been proposed by Donoho and Johnstone (1995):

$$\sigma_n = \frac{1}{0.6745}\text{Median}(|w_k|) \qquad (4)$$

where w_k are the detail coefficients at the first level.

The penalized thresholding method was proposed by Birge and Massart (1997). This method is based on minimizing a penalty function defined as:

$$t^* = \underset{t=1,\dots,n}{\arg\min}\left[-\sum(w_k^2,\ k<t) + 2t\sigma_n^2(\alpha + \log(\frac{n}{t}))\right] \qquad (5)$$

where α is a sparsity parameter and n is the number of detail coefficients w_k sorted in descending order. The penalized threshold for both the discrete wavelet transform and the wavelet packet is then estimated as:

$$t^p = |w_{t^*}| \qquad (6)$$

The sparsity parameter α can be tuned to obtain different threshold values. Three levels of penalized thresholding are common: penalized low (α = 1.5); penalized medium (α = 2); and penalized high (α = 5).

The application of the threshold can also be done in two modes. The standard hard thresholding criterion is defined as:

$$\hat{w}_{j,k}^h = \begin{cases} w_{j,k} & if \quad |w_{j,k}| \geq t \\ 0 & if \quad |w_{j,k}| \langle t \end{cases} \qquad (7)$$

where t is the threshold and $w_{j,k}$ are wavelet coefficients. The soft thresholding criterion is defined as:

$$\hat{w}_{j,k}^s = \begin{cases} sign(w_{j,k})\ (|w_{j,k}| - t) & if \quad |w_{j,k}| \geq t \\ 0 & if \quad |w_{j,k}| \langle t \end{cases} \qquad (8)$$

The soft thresholding criterion for wavelet de-noising was suggested by Donoho (1995). In contrast to hard thresholding, which can result in discontinuities (sharp drops) in the de-noised profile, soft thresholding yields a smooth output. Fig. 2 demonstrates the difference between the hard and soft thresholding modes. In the hard thresholding mode data beyond the threshold are preserved, but discontinuities are inevitable. Soft thresholding on the other hand shrinks the entire profile in order to prevent the occurrence of discontinuities.

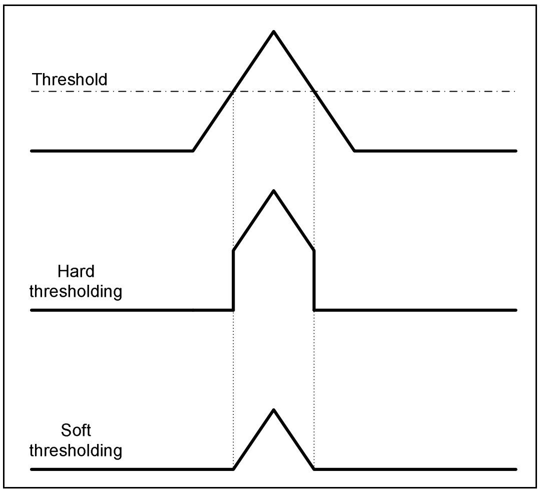


Fig. 2. The concept of hard and soft thresholding.

4. EXPERIMENTS AND RESULTS

The wavelet decomposition and thresholding methods were applied to roughness profiles extracted from a laser point cloud of a rock surface with millimeter-scale roughness. Fractal dimension was estimated for the de-noised profiles as well as the original laser profiles, and also for the manually measured roughness profiles to serve as reference. The following sections describe the experimental setup followed by the results and comparisons.

4.1 Study area

The scanned rock is situated in Tailfer, about 20 km south of the city of Namur and on the east side of the Meuse River in southern Belgium. The geological character of the scanned rock is a slightly metamorphosed limestone that is part of Lustin formation of carbonate mounts.

4.2 Data description

The rock surface was scanned with a Faro LS880 terrestrial laser scanner (Faro, 2009). The scanner was positioned at approximately 5 meters distance to the rock surface, and operated at the highest possible angular resolution, i.e. 0.009 degrees. The resulting point cloud contained about 1.2 million points on the rock surface with a point-spacing of 1 mm on average. According to the technical specifications of the laser scanner, the nominal range precision at a perpendicular incidence angle, which was roughly the case in our scan, is between 0.7 mm and 5.2 mm respectively for objects of 90% and 10% reflectivity at a distance of 10 m.

Roughness data were also collected manually along three profiles on the rock surface by using a carpenter's profile gauge with metallic rods at 1 mm intervals. These profiles were marked with white chalk and were visible in the reflectance image of the laser scanner data. Fig. 3 shows the profiles along which manual measurements were made, and their traces in the reflectance data of the point cloud.

The principal components were computed for a cutout of the point cloud that contained the profiles. The transformation parameters were then applied to rotate the point cloud into a more or less horizontal surface. Guided by the chalk traces in the reflectance image, three corresponding roughness profiles were extracted from the point cloud with samples interpolated at regular 1 mm intervals. The results of this procedure were three pairs of roughness profiles derived correspondingly from the manual and laser measurements with the same length and spatial resolution. We refer to these as the horizontal, diagonal and vertical profiles. Fig. 4 depicts the corresponding manual and laser roughness profiles in the horizontal direction.

4.3 Results

Using the roughness length method the fractal dimension was estimated for roughness profiles from both the laser scanner data and the manual measurements. The unit of profile length was chosen as 1 cm for all profiles to guarantee an appropriate density of 10 points per unit length. The power law relation was determined for each profile by calculating the standard deviation of the profile height within windows of 8 different sizes ranging from 3% to 10% of the profile length. Fig. 5 illustrates the power law relation between the window size and the standard deviation of the profile height for the laser and manual profiles in the horizontal direction. Here, the fractal dimension is estimated at 1.17 for the manual profile, and 1.96 for the laser profile. Considering the expected range of 1.2-1.7, the laser profile yields a clearly overestimated measure of roughness, while the fractal dimension of the manual profile is also slightly below the expected range.

To study the role of wavelet de-noising, different wavelet decomposition and thresholding methods were applied to the laser profiles and the estimated fractal dimensions for the de-noised profiles were compared with those of the manual profiles. For all profiles the decomposition was performed in 3 levels using a Daubechies wavelet of order 3 (db3). The standard deviation of noise was estimated at 1.8 mm, 1.3 mm, and 1.5 mm, respectively for the laser profile in the horizontal, diagonal and vertical direction. From these estimated noise levels thresholds were computed using the methods described in Section 3.2, and were applied to the detail coefficients globally at all decomposition levels. Table 1 summarizes the fractal dimensions estimated for the de-noised profiles obtained by using the discrete wavelet transform as the decomposition method. The same measures estimated for the de-noised profiles obtained by using the wavelet packets are summarized in Table 2. It can be seen that the fractal dimensions of the de-noised profiles vary across different thresholding methods; the variation is however smaller across different decomposition methods.

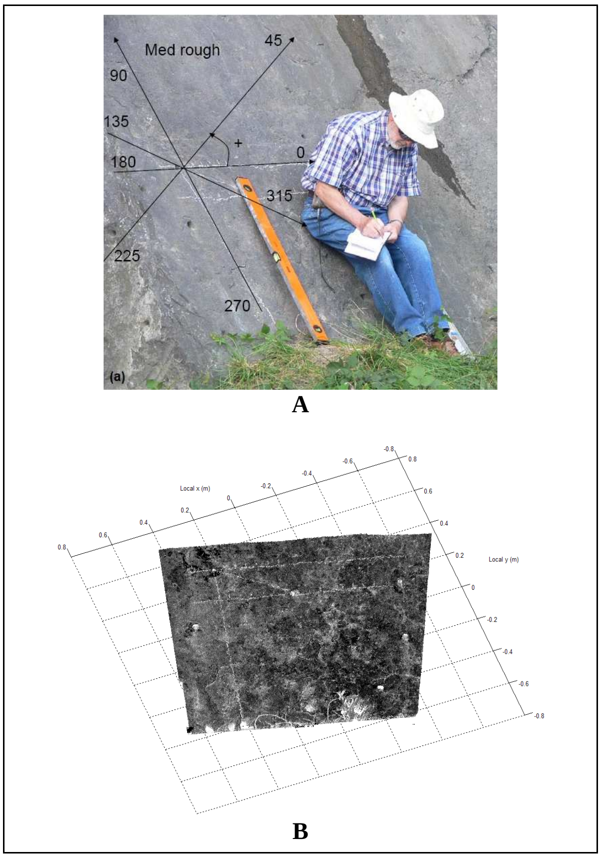


Fig. 3. A. manual measurement of roughness profiles; B. cutout of the rotated point cloud of the rock surface visualized with reflectance values.

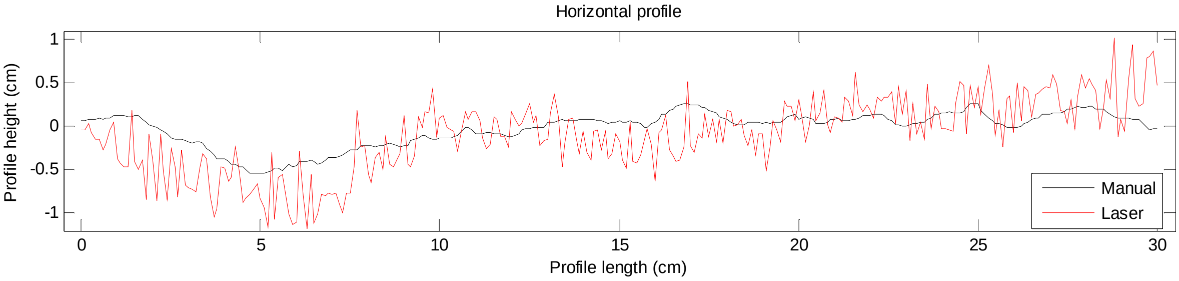


Fig. 4. Manually measured and laser scanned roughness profiles in the horizontal direction.

Fig. 6 depicts the variation in the fractal dimension of the de-noised profiles in the horizontal direction across different decomposition and thresholding methods. As it can be seen, the choice of the decomposition method has a minor impact on the fractal dimension of the de-noised profiles: the fractal dimensions pertaining to the discrete wavelet transform are only slightly larger than those of the wavelet packets. This can be verified also for the diagonal and vertical profiles from Table 1 and Table 2.

A noticeable difference in the performance of the de-noising methods can be seen in the application of hard and soft thresholding modes. Fig. 7 shows the influence of hard and soft thresholding on the fractal dimension obtained for the de-noised profiles in the horizontal direction. Soft thresholding results in too smooth de-noised profiles for which the estimated fractal dimensions are smaller than that of the manually measured profile and below the expected range. On the contrary, the de-noised profiles obtained by hard thresholding yield fractal dimensions that are within the expected range, except when penalized-high thresholding method is used. The fractal dimensions corresponding to the penalized high thresholding with both decomposition methods are in fact smaller than 1. The difference between the performances of hard and soft thresholding methods can be seen also for the diagonal and vertical profiles in Table 1 and Table2.

An examination of the results of different thresholding methods suggests that the fixed-form threshold applied in hard mode to the coefficients obtained by the wavelet packet decomposition yields fractal dimension values that are closer to those of the manually measured profiles and are also within the expected range. With the discrete wavelet transform as the decomposition method the penalized low thresholding method applied in soft mode seems to be an appropriate choice. Fig. 8 shows the result of penalized-low soft thresholding applied to the DWT coefficients of the horizontal laser profile, which compares well with the corresponding manually measured profile.

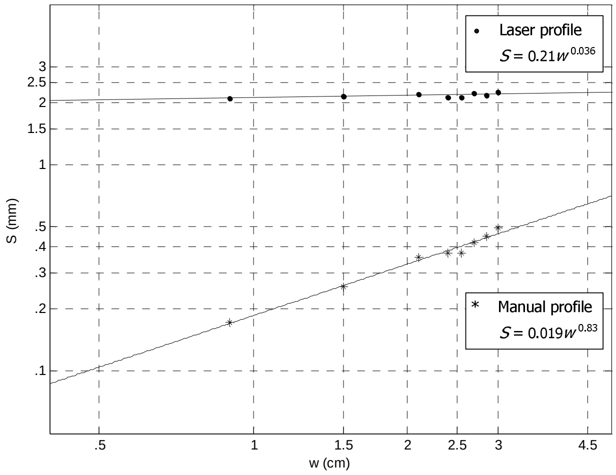


Fig. 5. The log-log plot of the standard deviation of profile height against window length for the laser and manually measured profiles in the horizontal direction.

5. CONCLUDING REMARKS

We investigated the role of wavelet de-noising of laser range data in reliable characterization of rock surface roughness. It was shown that fractal dimension values estimated for profiles derived from laser scanner data are generally larger than the expected values. The role of wavelet de-noising was investigated through the comparison of fractal dimensions estimated for the de-noised profiles with those of the corresponding manually measured profiles. The results of wavelet de-noising methods in general led to an improvement of the roughness measures estimated for the laser profiles. The fractal dimensions obtained for most of the decomposition and thresholding methods were within the expected range. The choice of the decomposition method was not found to affect the de-noising result; however, the application of hard or soft thresholding mode did have an impact on the estimated roughness measures. The presented results suggest that hard thresholding yields more accurate de-noised profiles for which the estimated roughness measures are more reliable.

In this research, the de-noising methods were applied to 1D profiles extracted from the laser scanner point cloud. Future research will focus on 2D wavelet de-noising of a range image, which is the fundamental data structure of terrestrial laser scanners. Other topics for further research include an investigation of the role of point density and profile length, and an analysis of the de-noising results using other roughness characterization methods.

				Horiz. profile	Diag. profile	Vert. profile
Original profile extracted from laser data				**1.96**	**1.89**	**1.90**
De-noised profiles Discrete Wavelet Transform	Soft Thresh.	Fixed-form		1.05	1.23	0.81
		Penalized	Low	1.23	1.38	1.19
			Med.	1.07	1.32	1.07
			High	0.94	1.19	0.62
	Hard Thresh.	Fixed-form		1.51	1.46	1.33
		Penalized	Low	1.68	1.76	1.68
			Med.	1.51	1.69	1.59
			High	0.94	1.44	0.62
Manually measured profile				**1.17**	**1.32**	**1.20**

Table 1. Fractal dimension values estimated for the de-noised profiles using discrete wavelet transform as the decomposition method.

				Horiz. profile	Diag. profile	Vert. profile
Original profile extracted from laser data				**1.96**	**1.89**	**1.90**
De-noised profiles Wavelet Packets	Soft Thresh.	Fixed-form		0.95	1.09	0.74
		Penalized	Low	1.10	1.40	1.11
			Med.	1.10	1.26	0.95
			High	0.94	1.08	0.66
	Hard Thresh.	Fixed-form		1.42	1.38	1.30
		Penalized	Low	1.52	1.79	1.66
			Med.	1.52	1.74	1.47
			High	0.94	1.19	1.18
Manually measured profile				**1.17**	**1.32**	**1.20**

Table 2. Fractal dimension values estimated for the de-noised profiles using wavelet packets as the decomposition method.

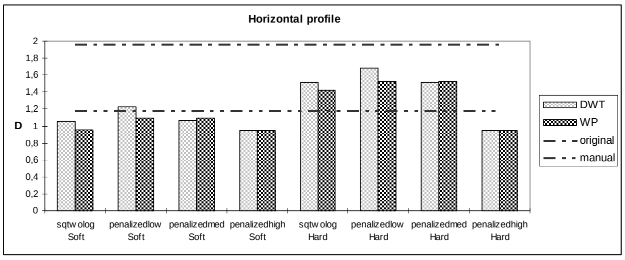


Fig. 6. Effect of decomposition method on the fractal dimension of de-noised profiles.

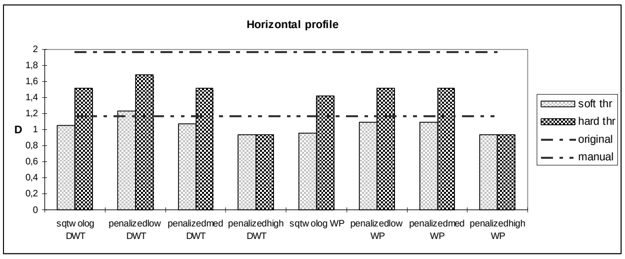


Fig. 7. Effect of hard and soft thresholding on the fractal dimension of de-noised profiles.

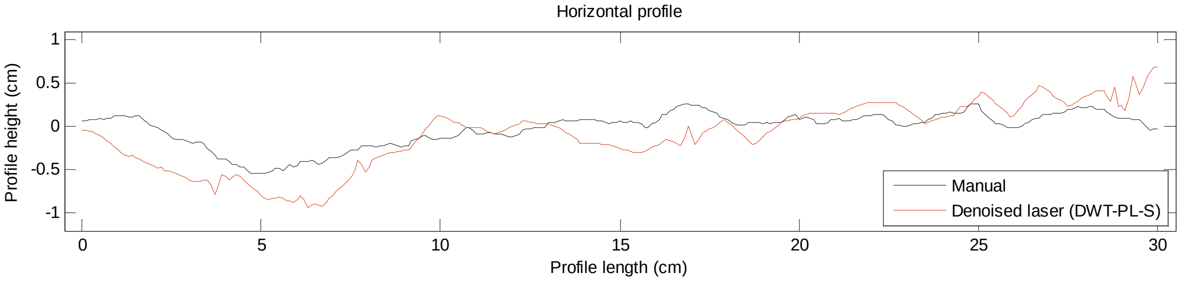


Fig. 8. De-noised laser profile obtained by penalized-low soft thresholding of the DWT coefficients compared with the corresponding manually measured profile.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Dominique Ngan-Tillard and Prof. Massimo Menenti of TU Delft for their helps and valuable comments.

REFERENCES

Birgé, L. and Massart, P., 1997. From model selection to adaptive estimation. In: D. Pollard, E. Torgersen and G.L. Yang (Editors), Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics. Springer-Verlag, New York.

Borah, D.K. and Voelz, D.G., 2007. Estimation of laser beam pointing parameters in the presence of atmospheric turbulence. Applied Optics, 46(23): 6010-6018.

Donoho, D.L., 1995. De-noising by soft-thresholding. IEEE Transactions on Information Theory, 41(3): 613-627.

Donoho, D.L. and Johnstone, I.M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3): 425-455.

Donoho, D.L. and Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association, 90(432): 1200-1224.

Dorninger, P., Nothegger, C., Pfeifer, N. and Molnár, G., 2008. On-the-job detection and correction of systematic cyclic distance measurement errors of terrestrial laser scanners. Journal of Applied Geodesy, 2(4): 191-204.

Fardin, N., Feng, Q. and Stephansson, O., 2004. Application of a new in situ 3D laser scanner to study the scale effect on the rock joint surface roughness. International Journal of Rock Mechanics and Mining Sciences, 41(2): 329-335.

Faro, 2009. Laser Scanner LS 880 Techsheet, pp. Accessed September 2009 http://faro.com/FaroIP/Files/File/Techsheets%20Download/UK_LASER_SCANNER_LS.pdf.PDF.

Gonzalez, R.C. and Woods, R.E., 1992. Digital image processing. Addison-Wesley, New York, 716 pp.

Hardle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A., 1998. Wavelets, approximation and statistical applications. Lecture Notes in Statistics, 129. Springer Verlag, 254 pp.

Jolliffe, I.T., 2002. Principal component analysis. Springer, New York.

Kulatilake, P.H.S.W. and Um, J., 1999. Requirements for accurate quantification of self-affine roughness using the roughness-length method. International Journal of Rock Mechanics and Mining Sciences, 36(1): 5-18.

Lichti, D.D., 2007. Error modelling, calibration and analysis of an AM-CW terrestrial laser scanner system. ISPRS Journal of Photogrammetry and Remote Sensing, 61(5): 307-324.

Malinverno, A., 1990. A simple method to estimate the fractal dimension of a self-affine series. Geophysical Research Letters, 17(11): 1953–1956.

Rahman, Z., Slob, S. and Hack, R., 2006. Deriving roughness characteristics of rock mass discontinuities from terrestrial laser scan data, Proceedings of 10th IAEG Congress: Engineering geology for tomorrow's cities, Nottingham, United Kingdom.

Soudarissanane, S., Lindenbergh, R., Menenti, M. and Teunissen, P., 2009. Incidence angle influence on the quality of terrestrial laser scanning points, ISPRS Workshop Laserscanning 2009, Paris.

Strang, G. and Nguyen, T., 1996. Wavelets and filter banks. Wellesley-Cambridge Press, Wellesley MA USA, 490 pp.

CCSSM: A TOOLKIT FOR INFORMATION EXTRACTION FROM REMOTELY SENSED IMAGERY

Y. Ge [a, *], C. Zhang [a,b], H. X. Bai [a]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing - (gey, zhangch, baihx)@lreis.ac.cn
[b] School of the Earth Sciences and Resources of China University of Mine and Technology, Xuzhou

KEY WORDS: Information Extraction, Spatial Information, Spectral Information, Multiple-point Simulation

ABSTRACT:

This paper presents a method named CCSSM (Classification of Combining Spectral information and Spatial information upon Multiple-point statistics) which is the derivation of two probability fields from the supervised classification for the spectral extraction and multiple-point simulation (MPS) for the spatial information, which then are fused. The performance of CCSSM for two-class classification has been discussed in our previous research works. This paper mainly introduces the software toolkit of CCSSM. A multiple-class classification using CCSSM is then given.

1. INTRODUCTION

Incorporating spatial structural information and spatial correlation information with spectral information is one way to improve the accuracy of classification of remotely sensed imagery (Ge et al., 2008b). In past decades, much effort has been directed towards developing excellent methods such as contextual classification, classification using texture structural information, and classification utilizing geostatistics. Recently, Ge et al. (2006; 2008a; 2008b; 2009) introduced multiple-point simulation (MPS) into the process of information extraction to increase the classification accuracy of remotely sensed imagery. MPS was used to characterize the structural and spatial association properties of geographical objects through a training image. MPS is one of the main applications of multiple-point statistics (Zhang et al. 2005), while SNESIM is one of the effective algorithms of MPS. The method of integrating MPS with spectra information was named as the Classification of Combining Spectral information and Spatial information upon Multiple-point simulation (CCSSM) (Ge et al., 2008c). The performance of CCSSM for the two-class classification has been substantiated by experiment and accuracy assessment using a Landsat Thematic Mapper (TM) 30 m image (Ge et al., 2008a; Ge et al., 2008b). This paper mainly introduces the software toolkit of CCSSM. A multiple-class classification using CCSSM is then given.

2. CCSSM

2.1 General idea

CCSSM is the derivation of two probability fields from the supervised classification and multiple-point simulation (MPS) and based on spectral and spatial information, which then are fused. CCSSM completely uses the training image and the template information to acquire the structural information of geographical objects and compensates for the inadequacy of using only spectral information for the information extraction. It not only takes spectral information of the remotely sensed imager into account, but also considers the spatial structure information and spatial correlation information of geographical objects. Therefore, it can effectively extract the investigated object with distinct spatial structural characteristics through a training image, and in particular can assist the user to extract objects smaller than the spatial resolution of the imagery (Ge et al., 2008a ,2008b and 2009).

In previous studies, the experiment undertaken (Ge et al., 2008a; 2008b and 2009) was only concerned with two-class classification, for instance, road and non-road. In fact, the algorithm is not limited to two-class classification. For example, assume that there are three categories with distinct structure characteristics. There are two means to simulate there classes. (1) First draws three training images which correspond to the simulations of three categories. Second, arbitrarily select a category as the first category to be simulated with the corresponding training image and all other categories will be treated as one category. Then the second category is simulated versus all other categories pooled together except the first category. Similarly, the third category can be simulated. (2) The second means is to first draw one training image with three categories and then simulate three categories simultaneously with the training image. In this paper, the second means will be implemented in the example.

2.2 CCSSM toolkit

This toolkit of CCSSM is written by c++ with GTK. This toolkit is composed of three parts for processing input data, intermediate data and output data as shown in Figure 1. The input data includes the probability field from the supervised classification which could be maximum likelihood classification (MLC) and training image. The supervised classification can be carried out in commercial software package for remotely sensed data such as PCI Geomatica, EARDAS and IDRISI. The training image can be drawn in commercial image software such as Photoshop, and then converted into a raster text file as input data.

[*] Corresponding author. gey@lreis.ac.cn.

The second part is for processing the intermediate data which includes hard data and search tree. The hard data for MPS can be obtained from the training data for MLC or the classified results from MLC. The search tree is a dynamic data structure that stores the conditional probability distribution of the training image and allows retrieval of all conditional probability distributions existing in a training image (Strebelle 2000, Ge et al., 2008b). Single normal equation simulation (SNESIM) is one of the effective algorithms of MPS. In CCSSM, SNESIM is adopted to extract the spatial structural information (Strebelle 2000). In the SNESIM algorithm, the template is an important parameter in the simulation process. It is a search window that consists of a set of ranked pixels and MPS use it to capture patterns from the training image. The SNESIM algorithm with multiple grid approach is performed to simulate the unsampled data through training image and then obtain the MPS result (Ge et al., 2009).

The third part is for fusing the probability fields from MPS and MLC. CCSSM toolkit provides three kinds of fusion methods which are the consensus-based fusion method, the evidence-based fusion method and the probability-based fusion method. In CCSSM toolkit, the formats of input data, intermediate data, and output data are the text type.

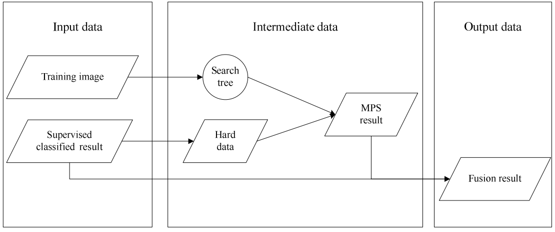


Figure 1. Components and data flow in CCSSM toolkit

3. EXAMPLE

The experimental procedure consists of four steps: MLC classification, MPS, data fusion using the consensus theory, and an accuracy assessment using error matrices. As described before, we first perform the MLC classification, then obtain MPS results with the revised SNESIM algorithm, and finally fuse the results using the consensus theory.

3.1 Data description

The example data was selected from QUICKBIRD 2.5 m imagery. The image size is 614 x 787 pixels and its resolution is 2.5 m as shown in Figure 2. From this image, it can be seen the roads and buildings have strong spatial structural features; for instance, in the study area, most roads are of a cross network structure and a certain inclination and most buildings distribute in square in the middle of the image.

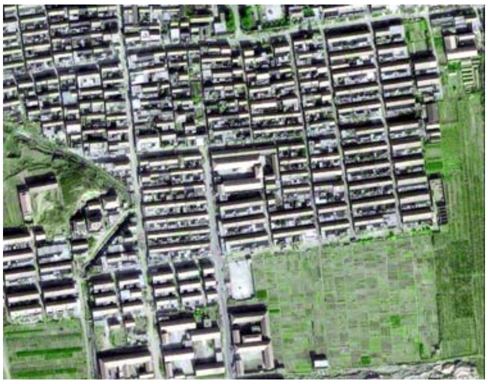


Figure 2. QUICKBIRD 2.5 m imagery

3.2 Training image

The training image was drawn by hand according to common characteristics of interested objects. In this remotely sensed imagery, the classes of buildings, gardens, roof shadow and streets have obvious structures and their spatial distributions are also homogeneous except in those areas of lawn.

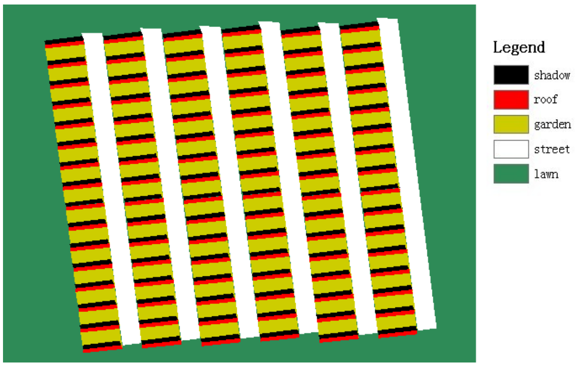


Figure 3. Training image (560x400）

3.3 MLC

From the Figure 2, the geographical objects can be roughly classified into 5 classes which are building roof, garden between the buildings, street road, lawn and roof shadow. As to investigate the performance of CCSSM method on multiple-class classification, the training data is selected in according the five classes. Then the MLC classification algorithm is performed using PCI Geomatica.

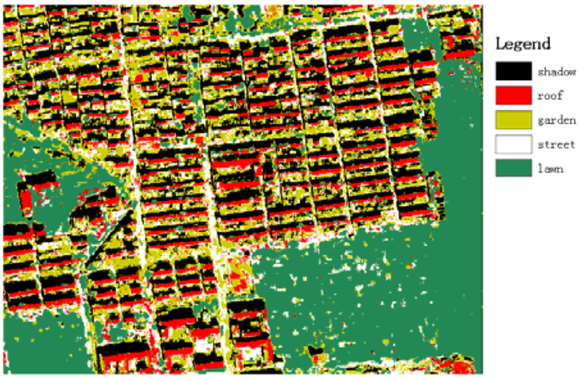
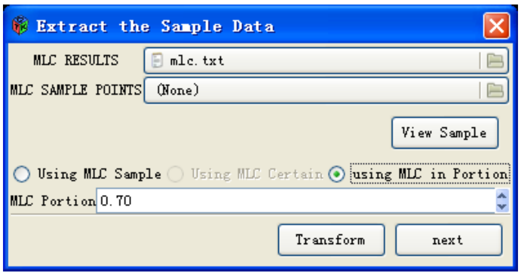


Figure 4. MLC results

3.4 MPS simulation

3.4.1 Hard data: There are two ways to obtain the hard data for MPS as shown in Figure 5. One is to use directly the sample data for MLC as hard data. The other is to select some pixels from the classified results of MLC by setting a threshold such as 0.7. The threshold can be set in terms of the resolution of remotely sensed imagery and the classification accuracy of MLC.


Extract the Sample Data

MLC RESULTS mlc.txt

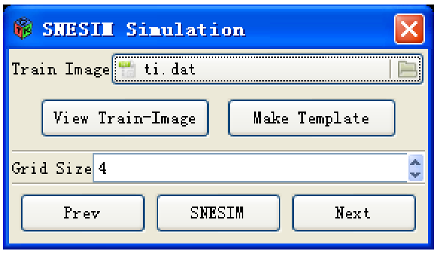
MLC SAMPLE POINTS (None)

○ Using MLC Sample ○ Using MLC Certain ● using MLC in Portion

MLC Portion 0.70

Figure 5. Interface for extracting sample data for MPS

3.4.2 SNESIM: First select the training image and press “View Train-Image” to view the training image as shown in Figure 6. The function of the button of “Make Template” is to set the template, for instance, the template can be set 5x5 or7x7 pixels. The grid size is to set the parameter of multiple-grid approach. This parameter can be set to 1， 2， 4， 8， …. Sort Interval denotes to sort the random path again after simulating a fixed number of pixels (Ge et al., 2008b). After setting all parameters, press the button of SNESIM and then get MPS result as shown Figure 7.


SNESIM Simulation

Train Image ti.dat

Grid Size 4

Figure 6. Interface for setting parameters for SNESIM

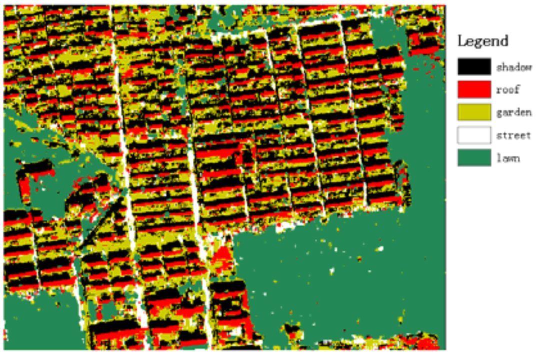
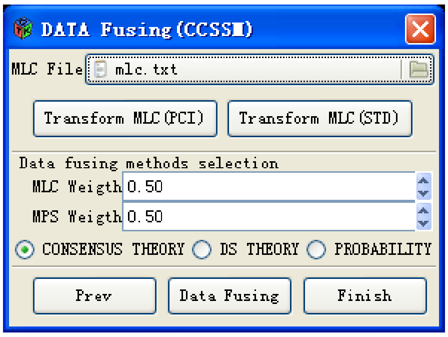


Figure 7. MPS result

3.5 Fusion

First one presses the button of “Transform” to extract the probability filed from MLC and then selects the fusion method as shown in Figure 8. In this example, the CONSENSUS-based fusion method is chosen to fuse two classification results. Finally, one presses the button of “Data Fusing” to obtain the fusion result as shown Figure 9. The result then is saved to the format of text file and can be viewed in ENVI software package.


DATA Fusing (CCSSM)

MLC File mlc.txt

Data fusing methods selection

MLC Weigth 0.50

MPS Weigth 0.50

● CONSENSUS THEORY ○ DS THEORY ○ PROBABILITY

Figure 8. Interface for setting parameters for Fusion

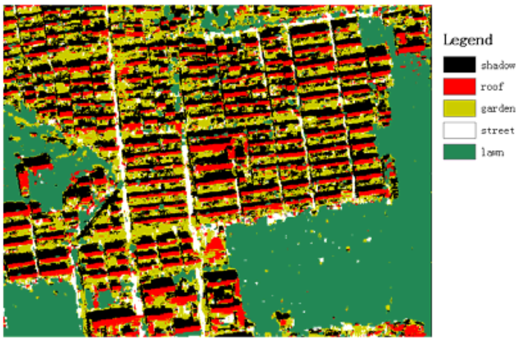


Figure 9. Fusion result

4. CONCLUSION AND FUTURE WORKS

This paper briefly introduced the general idea of CCSSM and its toolkit. A multiple-class classification is then given to demonstrate the use of CCSSM. Experiments have shown that this method can effectively extract the interested feature information with distinct structural characteristics. Furthermore, there are some aspects need to be discussed. For example, the software toolkit needs further improvement and can be downloaded for free to interested users around the world. The template is designed by a more reasonable and scientific way which considers the spatial structure and scale characteristics of the study area. The training image is a key factor in MPS and directly influences the simulation results, therefore, it is essential to design the training image with a realistic design.

REFERENCES

Ge, Y., Bai, H.X. and Le, D.Y., 2006, A classification method for remotely sensed imagery by integrating with spatial structure information. In 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. 5–7 JULY 2006, Lisboa, Portugal (Lisboa: Instituto Geográfico Português), pp. 588–594.

Ge, Y., Bai, H.X., Cheng, Q.M., 2008a, The solution of multiple-point statistics to extracting information from remotely sensed imagery. Journal of Chinese University of Geoscience, 19, pp. 421–428.

Ge, Y., Bai, H.X., Cheng, Q.M., 2008b, New classification method for remotely sensed imagery via multiple-point simulation: experiment and assessment. Journal of Applied Remote Sensing, 2, 023537.

Ge, Y., Bai, H.X., 20009, Multiple-point simulation-based method for extraction of objects with spatial structure from remotely sensed imagery. International Journal of Remote Sensing (accepted).

Strebelle, S. (2000). Sequential simulation drawing structures from training images: Unpublished doctoral dissertation, Stanford University, California.

Zhang, T., Switzer, P. and Journel, A., 2005, Merging prior structural interpretation and local data: The bayes updating of multiple-point statistics, In Proceeding of International Association for Mathematical Geology, 21–26 August 2005, Toronto, Canada (Toronto: the Geomatics Research Laboratory, York University and Wuhan: China University of Geosciences), pp. 615–620.

ACKNOWLEDGMENTS

This work was supported in part by the National High Technology Research and Development Program of China (Grant No 2006AA06Z115, 2006AA120106) and the National Natural Science Foundation of China (Grant No. 40671136, 40971222) and State Key Laboratory of Independent Innovation Team Project (Grant No. O88RA203SA).

CONSTRAINT ENERGIES FOR THE ADAPTATION OF 2D RIVER BORDERLINES TO AIRBORNE LASERSCANNING DATA USING SNAKES

J. Goepfert, F. Rottensteiner, C. Heipke[1], Y. Alakese, B. Rosenhahn[2]

[1]Institute of Photogrammetry and GeoInformation (IPI), Leibniz Universität Hannover
(goepfert, rottensteiner, heipke)@ipi.uni-hannover.de
[2]Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover
rosenhahn@tnt.uni-hannover.de, y_alakese@yahoo.de

Commission II

KEY WORDS: snakes, vector data, river, ALS, intensity, consistency

ABSTRACT:

The German Authoritative Topographic Cartographic Information System (ATKIS) stores the height and the 2D position of the objects in a dual system. The digital terrain model (DTM), often acquired by airborne laser scanning (ALS), supplies the height information in a regular grid, whereas 2D vector data are provided in the digital landscape model (DLM). However, an increasing number of applications, such as flood risk modelling, require the combined processing and visualization of these two data sets. Due to different kinds of acquisition, processing, and modelling discrepancies exist between the DTM and DLM and thus a simple integration may lead to semantically incorrect 3D objects. For example, rivers may flow uphill. In this paper we propose an algorithm for the adaptation of 2D river borderlines to ALS data by means of snakes. Besides the two basic energy terms of the snake, the internal and image energy, 3D object knowledge is introduced in the constraint energy in order to guarantee the semantic correctness of the rivers in a combined data set. The image energy is based on ALS intensity and height information and derived products. Additionally, features of rivers in the DTM, such as the flow direction or the river profile, are formulated as constraints in order to fulfil the semantic properties of rivers and stabilize the adaptation process. Furthermore, the known concept of twin snakes exploits the width of the river and also supports the procedure. Some results are given to show the applicability of the algorithm.

1. INTRODUCTION

1.1 Motivation

Many applications of geospatial information require a 3D representation and visualisation of the Earth's surface and the related topographic objects. In Germany the Authoritative Topographic Cartographic Information System (ATKIS®) is one of the main sources of such data. However, the relevant information is stored in a dual system that basically consists of the Digital Landscape Model (DLM) and the Digital Terrain Model (DTM). The objects in the DLM are modelled as 2D vector data, whereas the DTM represents the Earth's surface by terrain heights in a regular grid. Due to different methods of acquisition, processing, and modelling, discrepancies exist between the DLM and the DTM. This leads to semantically incorrect results if the two data sets are integrated without a geometrical adaptation. For instance, in a combined data set bodies of standing water having impossible height variations, streets with invalid height gradients, and rivers flowing uphill can be found. Therefore, the two data sets have to be suitably adapted for accurate combined visualization and processing. The Surveying Authority of the German Federal State of Schleswig-Holstein has conducted a state-wide Airborne Laser Scanning (ALS) flight for the derivation of the DTM. ALS delivers a 3D point cloud representing a Digital Surface Model (DSM) along with intensity values for each laser point. These intensities are related to the reflectance properties of the illuminated surfaces. Both the DSM and the intensities can be used to create additional features for an automated procedure adapting the DLM to the height data. It is the specific goal of this paper to present a new method for matching the borderlines of rivers to the ALS data. Rivers typically have an accuracy of 3-5 m in the ATKIS DLM, with local deviations that may reach 10 m. The ALS point cloud we use has an accuracy of approximately 0.3 m in the horizontal position and 0.15 m in height, so that a considerable improvement should be achievable. To increase the accuracy of river borderlines using ALS data, features extracted in the DSM could be matched with DLM objects. In this paper, a top-down method based on active contours is proposed instead. Whereas the contour is initialized by the vector data, the DSM or related products act as the image energy attracting the contour to salient features. Additionally, object knowledge about the specific appearance of rivers in a DSM is used to formulate constraint energies in order to increase the robustness of the approach.

1.2 Related Work

There has been growing interest in an integration of 2D GIS data and the related height information. A procedure for merging incorporated the 2D geometry of the objects into the DTM structure based on a triangulated irregular network (TIN) (Lenk 2001). However, no inconsistencies between the vector data and the DTM were considered. Koch (2006) improved the TIN-based integration methods by a least squares adjustment using equality and inequality constraints in order to incorporate some semantic properties of the objects. This method, besides being very sensitive to the weights of the observations, does not exploit the implicit information about the vector objects contained in the height data, e.g. structure lines at embankments related to roads or rivers. In (Goepfert & Rottensteiner, 2009) we have done so by adapting active contours to ALS data to improve the positional accuracy of road networks.

Snakes or parametric active contours are a well-known concept for combining feature extraction and geometric object representation (Kass et al., 1988; Blake & Isard, 1998). They explicitly represent a curve with respect to its arc length. In the standard formulation they cannot handle changes in the topology such as splitting and merging of entities (McInerny & Terzopoulos, 1995). This is not a problem for the adaptation of the 2D vector data to ALS features, because the initial topology is taken from the GIS data base and should be held fixed during the process. In the context of river borderlines, the concept of twin snakes introduced by Kerschner (2003) is relevant, because twin snakes connect two approximately parallel lines with a predefined distance. Earlier, Fua (1996) proposed a similar concept called ribbon snakes. Here the width of the ribbon is modelled by an extension of the internal energy and the image energy is calculated at the borderlines of the ribbon. Snakes are widely used in image and point cloud analysis as well as GIS applications. For example, Burghardt and Meier (1997) propose an active contour algorithm for feature displacement in automated map generalisation. Cohen and Cohen (1993) introduce a finite elements method for 3D deformable surface models. Borkowski (2004) shows the capabilities of snakes for break line detection in the context of surface modelling. Laptev et al. (2001) extract roads using a combined scale space and snake strategy. Several papers also deal with the modelling of rivers in height data. Brockmann and Mandlburger (2001) present a technique to extract the boundary between land and river based on bathymetric and ALS data. Fua (1998) simultaneously modelled drainage channels and the surrounding terrain from multiple images using a model-based optimisation scheme in order to express geometric, photometric and physical properties of the object of interest.

2. METHOD

2.1 General Work Flow

In this paper a new method based on an expansion of the twin snakes concept is proposed in order to adapt river borderlines from the ATKIS DLM to ALS data. The image and constraint energies are reformulated so that they represent the specific appearance of rivers in height data while the river borderlines conform to the physical properties of rivers. Whereas the 2D vector data are used for the initialization of the contour (cf. section 2.2), preprocessed ALS height and intensity data define the image energy forcing the snake to salient features (cf. section 2.3). Additionally, object knowledge is introduced in the algorithm (cf. section 2.4) in order to stabilize the optimization process and to obtain a suitable solution. For this purpose the energy functional of the snake is extended by constraint energies derived from three properties of rivers: (1) simple rivers have two approximately parallel borderlines, (2) rivers flow downhill, and (3) the terrain heights increase monotonically from the river banks to the adjacent areas.

After defining and weighting the different terms of internal, image, and constraint energies an iterative optimisation process is started. In the iteration process, the position of the snake is modified. The average change of the position of the contour points in the current iteration can be used to determine the convergence of the algorithm. Afterwards, the new position of the contour should match the corresponding features for the river borderlines in the ALS data. The novelty of the paper consists in the adaptation of the twin snake concept (Kerschner, 2003) to the requirements of our application, in a new formulation of constraints taking into account the physical properties of rivers (river flow, gradient directions at river embankments), and in a new definition of the image energy of the snakes based on the appearance of rivers in ALS data.

2.2 Snakes

In the initial idea of snakes, introduced by Kass et al. (1988), the position of contours in images is assessed by an energy functional, which consists of three terms:

$$\begin{aligned} E^*_{snake} &= \int_0^1 E_{Snake}(v(s))ds \\ &= \int_0^1 (E_{\text{int}}(v(s)) + E_{image}(v(s)) + E_{con}(v(s)))ds \end{aligned} \tag{1}$$

where *v(s) = (x(s), y(s))* represents the parametric curve with arc length s. The internal energy E_{int} determines the elasticity and rigidity of the curve considering the natural behaviour of the desired objects. The features of the object of interest should be represented in the image energy E_{image} in an optimal way. These features attract the contour to a suitable position. Furthermore, constraints can be defined in additional energy terms (E_{con}), which determine external forces. For example, the contour could be connected to fixed points using spring forces. The internal energy is motivated by the definition of curves forcing the contour to stay smooth linear objects:

$$E_{\text{int}}(v(s)) = \frac{\alpha(s)\cdot|v_s(s)|^2 + \beta(s)\cdot|v_{ss}(s)|^2}{2} \tag{2}$$

In Equ. 2, v_s and v_{ss} are the derivatives of *v* with respect to *s*, whereas *α(s)* and *β(s)* are weight functions. The first derivatives, weighted by *α*, impose a penalty to the arc length of the contour. For that reason, high values of *α* cause very straight curves. The curvature of the snake is modelled by the second order term, which is controlled by *β*. Thus, high values of *β* result in smooth contours, whereas small weights enable a zigzag like behaviour.

If dark or bright lines represent the objects of interest in the images, the image energy (cf. section 2.3) can be simply defined by the grey values. For objects represented by image edges the magnitude of the gradient image may describe this energy term. The global minimum of the energy functional (Equ. 1) determines the optimal position and shape of the snake in the image with respect to the defined energy terms. In order to simplify the optimisation process, the sum of image and constraint energies is replaced by the external energy E_{ext} and the functions *α* and *β* are set to constant parameters. After these modifications the minimisation of the functional results in the two independent Euler equations (for x and y):

$$\frac{\delta E_{ext}(v(s))}{\delta v(s)} + \alpha\cdot v_{ss}(s) + \beta\cdot v_{ssss}(s) = 0 \tag{3}$$

Because the derivatives can not be calculated analytically, they have to be approximated by finite differences. With the discrete formulation of the energy functional the weights can be varied easily for each node of the contour. Considering the substitution $f_v(v_i) = \partial E_{ext} / \partial v_i$ with v_i representing the node *i* the Euler equations can be rewritten:

$$\begin{aligned} &\alpha_i\cdot(v_i - v_{i-1}) - \alpha_{i+1}\cdot(v_{i+1} - v_i) \\ &+ \beta_{i-1}\cdot(v_{i-2} - 2v_{i-1} + v_i) - \beta_i\cdot(v_{i-1} - 2v_i + v_{i+1}) \\ &+ \beta_{i+1}\cdot(v_i - 2v_{i+1} + v_{i+2}) + (f_v(v_i)) = 0 \end{aligned} \tag{4}$$

The matrix form of Equ. 4 reads:

$$Av + f_v(v_i) = 0 \tag{5}$$

The pentadiagonal banded matrix *A* in Equ. 5 includes the weights of the internal energy. Equ. 5 is solved by setting the right side equal to the product of a step size γ and the negative time derivatives of the left side. Assuming that the derivatives of the external energy $f_v(v_i)$ are constant during a time step, an explicit Euler step regarding the image energy is obtained. The internal energy completely determined by the banded matrix results in an implicit Euler step if the time derivative is calculated at time *t* rather than *t-1*. These considerations lead to:

$$Av_t + f_v(v_{t-1}) = -\gamma(v_t - v_{t-1}) \tag{6}$$

The time derivative vanishes at equilibrium and Equ. 6 degenerates to Equ. 5. The solution is obtained by:

$$v_t = (A + \gamma \cdot I)^{-1} \cdot (\gamma \cdot v_{t-1} - f_v(v_{t-1})) \tag{7}$$

The following sections describe the derivation of the image energy and the definition of different constraint energy terms.

2.3 Image Energy

Before defining the image energy, we have to discuss what the optimal position of the river borderline in the DTM is. Fig. 1 visualises a cross section of a river bed, the points of interest (black arrows), and the derivatives of the DTM. The points of interest mark the boundary between land and water and correspond to the maximum of the second derivative. The cross section of the intensity data in the vicinity of a river shows a similar structure (Fig. 2a and b). Due to the specular reflection and a high absorption rate in the laser wavelength (1064 nm) water surfaces appear dark in the intensity image. Under certain conditions the reflected echo from water bodies is even too weak to be detected in the receiver device. Therefore, gaps in the scanning pattern can exist. However, very high intensity values can be observed in near-nadir points due to total reflection. This phenomenon does not occur in the analysed data set, but has to be considered in future studies.

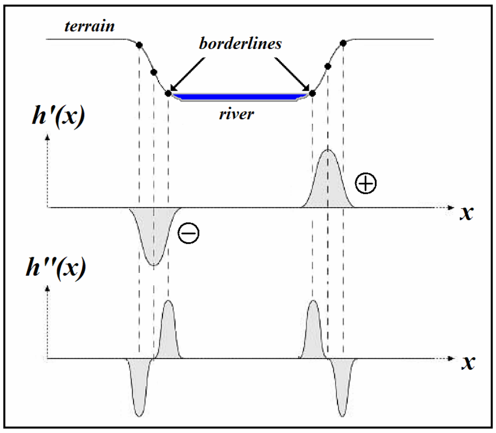


Figure 1. Cross section of rivers in the DTM

The image energy is generated in three steps. The ALS height and intensity data are first transferred into a regular grid by Kriging (Cressie, 1990). Unfiltered data are used in order to analyse the ability of snakes to bridge disturbances, such as shrubs at river banks in the DSM. The resulting height image I_{DSM} and the intensity image I_{Int} are smoothed by a 3 *x* 3 median filter to remove outliers while preserving the river edges. In a second step the two images are combined by a weighted sum:

$$I = a \cdot I_{Int} + b \cdot I_{DSM} \tag{8}$$

In Equ. 8, *I* is the combined image, and *a* and *b* are weights. Analysing the histograms of the images, these weights are set so that the two sources have a similar range of values and thus a comparable influence. The image energy is calculated by the negative magnitude of the second derivatives of the resulting image (Equ. 9). This is realised by convolving the image with the corresponding derivatives of a Gaussian *G* (Fig. 2c).

$$E_{img} = -\sqrt{\left(\frac{\delta^2 G}{\delta x^2} * I\right)^2 + \left(\frac{\delta^2 G}{\delta y^2} * I\right)^2} \tag{9}$$

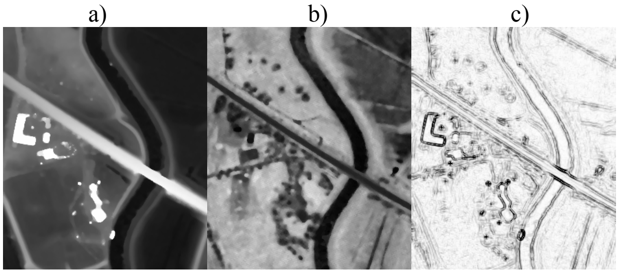


Figure 2. a) DSM b) intensity c) image energy

The boundaries between land and water appear dark in the created image energy, which forces the snake to low grey values. However, other edges in the vicinity of the desired borderline disturb the optimisation process. Therefore, some constraint energies are defined in the next section in order to increase the robustness of the algorithm.

2.4 Constraint Energy

The constraint energy E_{con} is the sum of three components:

$$E_{con} = E_{Twin} + E_{Flow} + E_{Gradient} \tag{10}$$

The energy E_{Twin} is related to the concept of twin snakes (Kerschner, 2003) and is explained in Section 4.2.1. E_{Flow} is related to the downhill flow direction and is explained in Section 4.2.2. Finally, the term $E_{Gradient}$ takes into consideration the gradient direction of the terrain in the vicinity of the river bank; it is explained in Section 4.2.3.

2.4.1 Twin snakes: Similar to the original constraint energy in Kass et al. (1988), which connects the contour to fixed points by spring forces, Kerschner (2003) developed the idea of twin snakes. In this concept two snakes are linked by a predefined distance d_0. For that purpose the energy functional of each snake is extended by an additional term E_{Twin} that is minimised if the distance of the snakes fulfils the demand:

$$E_{Twin} = \kappa_{Twin} \cdot (d(v_i) - d_0)^2 \tag{11}$$

In Equ. 11, $d(v_i)$ is the actual distance to the twin at node v_i and κ_{Twin} is the weight of this energy term.

The minimisation of the energy functional is conducted for each contour individually and alternates between the left and right snake. The corresponding other snake is fixed in the meantime. In the optimisation process the derivatives of Equ. 10 with respect to the image coordinates *x* and *y* have to be calculated (cf. E_{ext} in Equ. 3) to determine in which direction this energy term moves the contour. The twin energy acts perpendicular to the snake direction either with an attraction or repulsion force. Due to the fact that rivers vary in width a constant value for the predefined distance is not suitable for the proposed application. Assuming a relative quality that is higher than the absolute accuracy of the DLM this distance is derived from the vector data for each node individually. These values are calculated either as point to point or point to polyline distances. After each iteration the distances are updated. In order to find the inner edges of the bundle of edges along the rivers in the image energy, it is useful to decrease the value for river width from the DLM. A smaller predefined distance increases the attraction force of the two corresponding snakes. The twin energy term supports the delineation of two image edges with a predefined adjustable distance and helps to overcome local minima.

2.4.2 Flow direction: The downhill flow direction of rivers in the DTM is one of the physical properties which are integrated in the algorithm. Due to the effect of the Earth's gravity the nodes in flow direction should show a decreasing or at least a constant height. Initially, a weight $w_{\Delta hi}$ depending on the height differences between the upstream and the downstream nodes h_{i-1} and h_i of one snake is computed:

$$w_{\Delta hi} = \begin{cases} |h_{i-1} - h_i| & if \quad h_{i-1} - h_i < 0 \\ 0 & else \end{cases} \qquad (12)$$

Introducing $h(v_i)$ for the terrain height for node v_i and a weight parameter κ_{Flow}, the flow energy term is then defined by:

$$E_{Flow} = \kappa_{Flow} \cdot w_{\Delta hi} \cdot h(v_i) \qquad (13)$$

In the minimisation process the derivatives of E_{Flow} at the nodes with respect to the image coordinates x and y have to be calculated (cf. E_{ext} in Equ. 3) similar to the strategy used for the image energy. This approach attracts the nodes that do not satisfy the requirements of the flow direction, from surrounding terrain to lower river areas. The height differences in $w_{\Delta hi}$ act as additional weights. The larger the deviation from the flow constraint for two nodes the stronger is the gradient descent force in the DSM for the node in downhill direction. If some nodes of the snake have already been attracted to the river borderline this energy term has the effect to shift the nodes on higher terrain in their vicinity to the lower river area.

2.4.3 Gradient direction: The third energy term in Equ. 10 exploits the fact that rivers usually flow in valleys and thus the terrain heights increase on each side with the distance from the river. Initially, edge pixels in the DSM are calculated by using the Sobel operator with post processing steps (non-maxima-suppression and hysteresis threshold) following the strategy of Canny (1986). Afterwards, the gradient direction of the edge pixels is computed. Fig. 3b visualises an image in the vicinity of a river with grey value coded gradient directions. Due to the use of the unfiltered data vegetated areas disturb the depicted edges.

For the following considerations we assume the directions of the river borderlines from ATKIS to be correct within certain limits. At each node of the snake we define the direction of the cross-section to be perpendicular to the snake direction. We then analyse the gradient orientations within a small rectangular buffer that is aligned with the direction of the cross-section. The gradient orientations are compared to the directions that are expected based on a model of the cross-section similar to the one depicted in Fig. 1. If these directions are close to each other the edges are labelled as positive, otherwise as negative (see the signs in Fig. 1 for the first derivatives *h'(x)* for the right snake). By analysing the order of the edges in the vicinity of the river the most probable edge for the current node is chosen considering the typical cross sections of the river (cf. Fig. 3b). We search for two edge pixels with opposite directions in correct order concerning the river profile without another edge pixel in between. Due to the smoothing effect of the internal energy this constraint is already useful, if the predominant number of choices is correct. The related energy term is defined in a similar way as E_{Twin} in Equ. 14. It is minimised if the distance between the current node and the edge is zero:

$$E_{Gradient} = \kappa_{Gradient} \cdot d(v_i) \qquad (14)$$

In Equ. 14, $d(v_i)$ is the distance between the node v_i and the chosen edge and $\kappa_{Gradient}$ represents the weight parameter. The comparison of the snake direction and the gradients enables the approach to deal with poor initialisation. The influence of suitable edges depends on the size of the rectangular buffer. The gradient energy shifts the snakes to the correct gradients (first derivatives) close to the desired edges (second derivatives) and thus into the range of influence of the image energy.

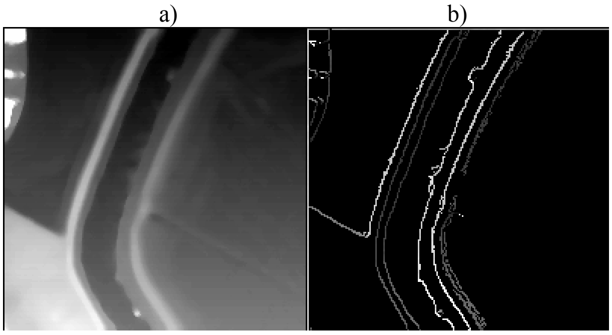


Figure 3. a) DSM b) grey value coded gradient direction

3. EXPERIMENTS

3.1 Data

We use an ALS data set captured near the village Kellinghusen by TopScan during a countywide flight campaign of Schleswig-Holstein between 2005 and 2007. Flying at an altitude of 1000 m the system ALTM 3100 from Optech was used in the first and last echo mode to provide an overall point density of 3-4 points/m^2 and an accuracy of 0.15 m (height) and 0.3 m (position). From the ALS data a 1 m grid is interpolated. The ATKIS vector data set (river borderlines only) was manually manipulated in order to create different initialisation scenarios.

3.2 Effects of the Constraint Energy Terms

The first four experiments emphasise the advantages of the constraint energies in certain conditions. For that reason a higher weight is assigned to the constraint currently considered. The example in Fig. 4a visualises the behaviour of the original snake without constraints. If the initialisation is situated close to the desired edge (upper part of the left snake), the snake moves to the river borderlines. However, with a poor initialisation either the influence range of the image energy is too small to attract the contour (lower part of the left snake) or other edges along the river disturb the final position of the snake (right snake). We want to tackle these problems by integrating the three constraint energy terms. In Fig. 4b the advantages of the twin energy are highlighted. The distance of the initial borderlines is too large, but the approximate river width is known. This information is exploited by the twin energy term. In this case the two snakes are attracted by this part of the energy functional until the range of influence of the image energy is reached. Using actual ATKIS data the assumption is made that the river width is nearly correct and can be used as the predefined distance between the two snakes. With the twin energy the two contours support each other on their way to a suitable result. Fig. 4c emphasises the effect of the integration of the flow direction in the method. If some parts of the snake have already reached the river borderline, the other nodes in the vicinity are attracted by this force from the surrounding terrain to the lower river region. Due to the high weight for this energy in Fig. 4c and the disturbance caused by shrubs, which hide the true river borderline in the DSM, the nodes in the middle of the right snake move to the centre of the river. The vegetation hampers the fulfilment of the flow direction constraint. Using the gradient direction information (Fig. 4d) helps to deal with poor initialisation. The snakes are able to jump across several wrong edges in the image energy. This additional energy extends the range of the influence of the suitable edges similar to approaches modifying the image energy based on a distance transformation (gradient vector flow by Xu & Prince, 1997).

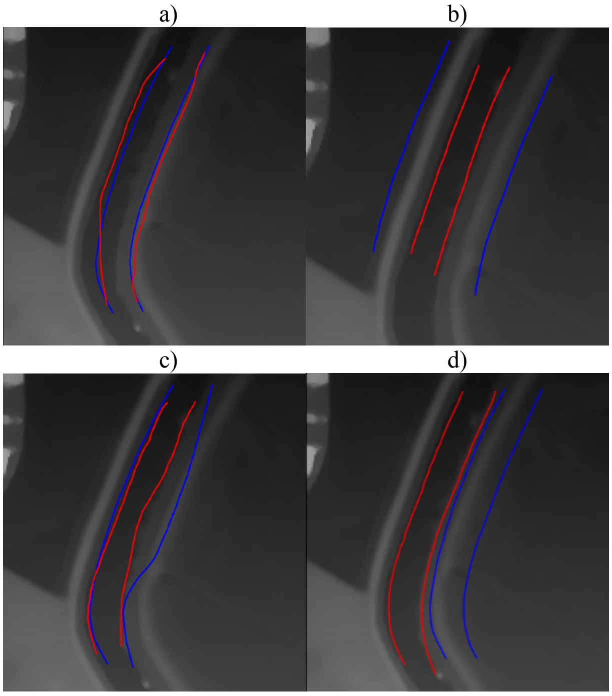


Figure 4. Initialisation scenarios for the different constraints (blue: initialisation; red: final solution): a) without constraints; b) twin snakes c) flow direction; d) gradient direction.

	α	β	κ_{Image}	κ_{Twin}	κ_{Flow}	$\kappa_{Gradient}$
Figure 5	0.16	1.2	3	0.2	10	0.05

Table 1: Weights for the energy terms used for Fig. 5 and 6.

3.3 Results and Discussion

Fig. 5 depicts the adaptation of river borderlines with different initialisations to show the robustness of the method. The empirical weights for the different energy terms of all following examples are shown in Table 1. The algorithm can adapt the 2D vector data, systematically shifted in x-direction by 10 m (Fig. 5a) and 15 m, to the ALS information. Even vegetation and the bridge in the middle of the image do not significantly influence the result. However, if the vector data are shifted by 20 m (Fig. 5b) some parts of the snakes do not reach the correct position any more. These problems (yellow arrows) occur at both ends of the snakes, in areas of strong curvature, and in the vicinity of vegetation and the bridge, where the lack of suitable image energy affects the outcome of the algorithm. For evaluation purposes the root mean square (RMS) values of the perpendicular distances of the nodes to the reference are calculated (Table 2). In case of 10 m and 15 m shift the errors are smaller than 1.5 m for the left snake and 3 m for the right snake. The poorer result of the right snake is caused by the systematic shift of the initialisation. The entire right snake has to cross the river embankments with several edges in order to move to the correct position. The constraint energy terms attenuate while approaching the desired edge and the snake is sometimes caught in a local minimum related to another edge from the river borderline. Therefore, it is assumed that the algorithm is able to solve random errors of the ATKIS data significantly better, which sometimes cross the correct position.

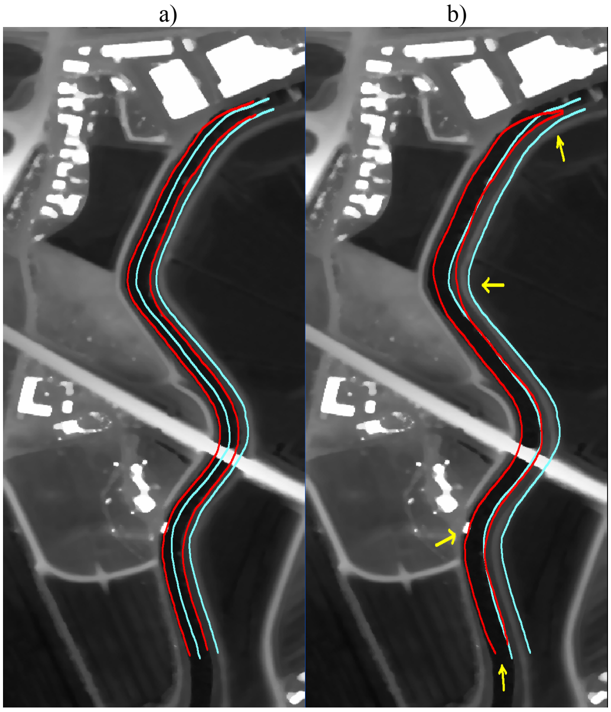


Figure 5. Adapted river with different initialisation (light blue: initialisation; red: final solution) a) x-shift: 10 m b) x-shift: 20 m

RMS of point to line distances (m)	shift: 10 m		shift: 15 m		shift: 20 m	
	left	right	left	right	left	right
Initialisation	8.90	8.95	13.34	13.42	17.77	17.88
Solution	1.15	2.48	1.23	2.67	3.40	3.31

Table 2: Evaluation of the results in Fig. 5 (left snake: 124 nodes; right snake: 120 nodes)

The method was applied to a second example without changing the parameter settings (Fig. 6 and Table 3). The DSM is strongly influenced by vegetation (Fig. 6a). Tree branches hanging across the river generate strong edges in the image energy (yellow arrows). Without adapting the parameters of the snake the constraint energy terms are not able to release the upper snake from the related minima. By using filtered ALS data the snakes move to a suitable solution comparable to the first example without changing the parameters (Fig. 6b). However, once more the systematic shift affects the snake in the shift direction (in this example the lower snake) more than the other. In summary, the accuracy achieved is not entirely satisfying yet for the snake that is initialised to be furthest away from the river. However, difficulties in fixing a suitable reference position, which is done manually, and the resolution of the used grids of 1 m have to be considered in this context.

RMS of point to line distances (m)	DSM		DTM	
	upper	lower	upper	lower
Initialisation	8.16	8.13	8.16	8.13
Solution	3.01	2.53	1,04	2.42

Table 3: Evaluation of the results in Fig. 6 (upper snake: 101 nodes; lower snake: 104 nodes)

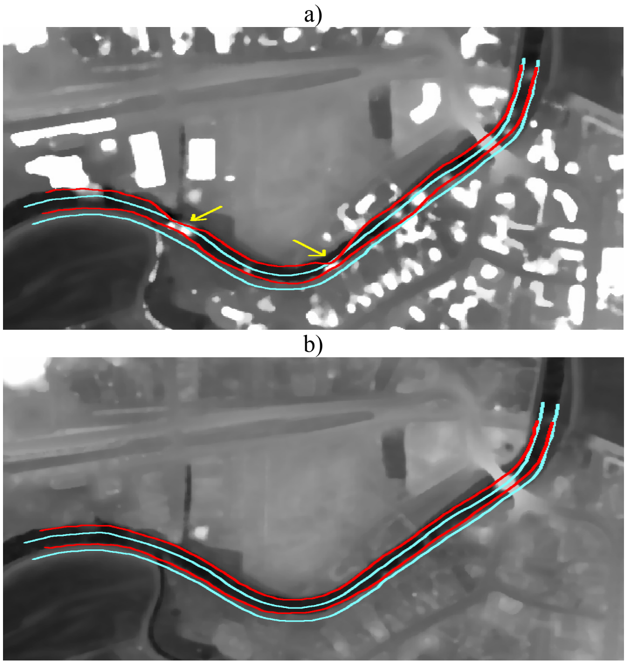


Figure 6. Adapted river with 10 m y-shift using unfiltered (a) and filtered (b) ALS data (light blue: initialisation; red: final solution)

4. CONCLUSION

This paper is focused on a method for adapting river borderlines to ALS features by means of active contour. For that purpose the snake algorithm is extended by three constraint energy terms derived from object knowledge. The connection of the two river sides in the twin energy, and the information about the flow direction as well as the gradient direction at the river banks stabilise the robustness of the basic active contour algorithm and enable the method to deal with poor initialisations. In future work a combination of the proposed constraints and the topology of larger river networks will be integrated in the concept. This algorithm is only one step to a larger framework which we develop in order to solve the inconsistencies between the DLM and height information. All objects in the vector data, which are represented by suitable features in the terrain model, should be adapted. This process provides a dense network of shift vectors which can be used in addition to prior accuracy knowledge in order to improve the consistency of the DLM and ALS data. Furthermore, some objects, such as river and road networks, can be treated together. For example, the bridge in Fig. 5 disturbs the adaptation process of rivers. Bridges are strong feature in the terrain model which can be used to connect rivers and roads in a combined optimisation.

ACKNOWLEDGEMENT

This research was supported by the surveying authorities of Lower Saxony “Landesvermessung und Geobasisinformation Niedersachsen” and Schleswig-Holstein “Landesvermessungsamt Schleswig-Holstein”. We also express our gratitude to the mentioned surveying authorities for providing the data.

REFERENCES

Blake, A. and Isard, M., 1998. Active contours. *Springer, Berlin Heidelberg New York*, 351 p.

Borkowski, A., 2004. Modellierung von Oberflächen mit Diskontinuitäten. *Habilitation*, TU Dresden, Germany, 91p.

Brockmann H., Mandlburger G., 2001. Aufbau eines Digitalen Geländemodells vom Wasserlauf der Grenzoder. *Publikationen der DGPF*, Band 10, pp. 199–208.

Burghardt, D. and Meier, S., 1997. Cartographic displacement using the snake concept. In: Förstner, Plümer (eds.), *Semantic modeling for the acquisition of topographic information from images and maps*, Basel, Birkhäuser Verlag, pp. 59-71.

Canny, J., 1986. A Computational Approach to Edge Detection. *IEEE TPAMI*-8(6): 679-698.

Cohen, L. D. and Cohen, I., 1993. Finite element methods for active contour models and balloons for 2-D and 3-D images, *IEEE TPAMI* 15(11): 1131-1147

Cressie, N. A. C., 1990. The origins of Kriging, *Mathematical Geology* 22: 239-252.

Fua, P., 1996. Model-based optimization: Accurate and consistent site modeling. *IntArchPhRS 31 B 3*:222–223.

Fua, P., 1998. Fast, accurate and consistent modelling of drainage and surrounding terrain. *Int. J. Computer Vision 26(3):*215-234.

Goepfert, J., Rottensteiner, F. 2009. Adaptation of roads to ALS data by means of network snakes: *IntArchPhRS 38-3/W8*:24-29.

Kass, M, Witkin, A, Terzopoulos, D., 1988. Snakes: active contour models. *Int. J. Computer Vision* 1(4):321-331.

Kerschner, M. 2003. Snakes für Aufgaben der digitalen Photogrammetrie und Topographie. Phd Thesis. IPF, Vienna University of Technology.

Koch, A., 2006. Semantische Integration von zweidimensionalen GIS-Daten und Digitalen Geländemodellen. PhD Thesis, University of Hannover, DGK-C 601.

Laptev, I., Mayer, H., Lindeberg, T., Eckstein, W., Steger, C. and Baumgartner, A., 2000. Automatic extraction of roads from aerial images based on scale space and snakes, *Machine Vision and Applications* 12: 23-31.

Lenk, U., 2001. 2.5D-GIS und Geobasisdaten - Integration von Höheninformation und Digitalen Situationsmodellen. PhD Thesis, University of Hannover, DGK-C 546.

McInerney, T. and Terzopoulos, D., 1995. Topologically adaptable snakes. *Proc. ICCV*, pp 840-845.

Xu, C, Prince, J.L., 1997. Gradient Vector Flow: A new external force for snakes, *Proc. IEEE-CVPR*, pp. 66-71.

TRACKING AREAL OBJECT IDENTITY IN SNAPSHOT SEQUENCES

Mingzheng Shi and Stephan Winter

Department of Geomatics
The University of Melbourne
Victoria, 3010, Australia
m.shi@pgrad.unimelb.edu.au, winter@unimelb.edu.au

Geosensor networks are deployed to detect and track dynamic geographic phenomena, or objects, over space and time. Object identity is a unique characteristic to distinguish different spatial objects. Based on tracking and analysis of the relations of object identities for example different classes of events can be derived. This paper introduces a spatiotemporal data model for the storage and maintenance of areal object lifespan in a decentralized network. Different from previous work, our algorithm uses snapshot sequences to analyze the relations of spatial objects, and it allows both abrupt and incremental changes. In contrast to previous approaches the presented approach can deal with change beyond incremental change. Our model can also become an essential component for analyzing static topological relations of spatial objects in a decentralized manner, or for detecting dynamic topological changes.

1 INTRODUCTION

Our world is dynamic, and many geographic phenomena are continuously changing in time and space by different causalities. There are two ontologies that are frequently used for analyzing change, referred to SNAP and SPAN (Grenon and Smith, 2004). The SNAP ontology has a snapshot-based view, where entities are organized as temporal sequences of snapshots. Snapshots are commonly used in geographical applications. Conventional geographical information systems (GIS), for instance, acquire spatiotemporal data by remote or in-situ observations of typically relative fine spatial, but coarse temporal granularity, and thus, represent spatial phenomena as static snapshots (Peuquet, 2001). When comparing two snapshot sequences, changes may appear to be abrupt due to the low temporal frequency of observations. Moreover, interpolation of the snapshots is not always appropriate to reconstruct the underlying dynamic processes of reality. In this case, the snapshot approach is unable to trace events that occur between snapshots.

Event-based approaches in SPAN ontology then become promising in theoretical studies (e.g., Worboys, 2005). An event-based spatiotemporal data model, for example, is proposed by (Peuquet and Duan, 1995). The model only records the timestamp when change occur, therefore it can be understood as a chronicle-based model that is dual to the snapshot-based model (Galton, 2004). In this case, changes or events can be explicitly stored in a compacted timeline, and additionally the redundancy of spatiotemporal data in the snapshot-based model can be reduced. However, one of the issues that the event-based model needs to resolve is the spatial, temporal, and thematic granularities of observations. With certain spatial and temporal resolution of observation, events may not be recorded somewhere, sometime. Even if someone may argue that continuous observation is possible with future technologies, there is still the problem of redundancy of observations. Large amount of energy can be consumed to observe an environment where no change has occurred.

Therefore, there exist situations where data from applications, e.g., snapshots, cannot satisfy theoretic analysis, and theoretic models, e.g., event-based model, are difficult to apply to real world applications. It is the emergence of new observation tools, e.g., wireless sensor networks (WSN), that bring the theory and application closer, since WSN can provide new options of spatial, temporal and thematic granularities of observations. A WSN is a network of computing devices that can collaborate via radio communications. A WSN is also a network of observation devices, since each node in the WSN is equipped with sensors that enable thematically fine-grained observation of the environments. A WSN is able to observe change in real-time, and hence, with almost any temporal granularity (but typically coarser spatial granularity).

In the context of geographic information science, snapshot-based approaches and event-based approaches have been used to model WSNs. Snapshot-based approach are commonly used in applications, (e.g., Wark et al., 2007), where sensor nodes are tasked to periodically sense and store snapshots of an environment by setting the sensor network to a certain temporal sensing resolution. Although a snapshot approach is more practical, theoretical studies, (e.g., Worboys and Duckham, 2006), are more interested in event-based model of WSNs due to its advantages of detecting salient changes or events. But, as discussed above, event-based approach is based on the underlying assumption of continuous observation, which is not practical in real world scenarios. Moreover, the redundancy of continuous observations becomes more critical in the case of WSNs, since WSNs are battery-powered and energy resources are highly constrained for sensor nodes.

In this circumstance, this paper develops a new spatiotemporal data model for WSNs that incorporates both snapshot-based and event-based approaches, such that the model can adapt to different spatial, temporal and thematic granularities of observations, can detect, store, and analyze changes or events, and can reduce the redundancy of spatiotemporal data. Since temporal granularity of observation is allowed in our model, geographic phenomena are represented as sequences of snapshots. Also, our spatiotemporal data model is an object-based model, i.e., the geographic phenomena are represented as objects or entities. We are primarily interested in areal objects or regions. Our model relies on areal object identities for spatiotemporal data representation. Thus, each areal object will be provided a unique identity, and our goal is to decentrally and dynamically trace the areal object identities in snapshot sequences over time.

This paper has three major contributions. Firstly, we introduce a new model that integrates both snapshot-based and event-based approaches for spatiotemporal data representation. Secondly, we develop an approach to decentrally store the completed lifespan of areal objects in a sensor network. Finally, a decentralized algorithm is designed to maintain the lifespan of areal objects in snapshot sequences.

2 RELATED WORK

Geographic phenomena may continuously change over time their location, size, orientation, or their character. The change investigated in this paper is the change of spatial entities. A change occurs whenever spatial entities possess different spatial attributes at different times (Galton, 2000). Grenon and Smith (2004) classify entities in the spatiotemporal world into two categories: one is continuants, and the other is occurrents. The study of this paper is on the perspective of continuants that exist at a given time at a given level of granularity and undergo different types of changes over time.

Hornsby and Egenhofer (1997) suggest a qualitative representation of change. Their notion of change is based on object identity, and a set of operations that either preserve or change identity. This paper applies their concept of identity in sensor networks for decentralized maintenance of areal object identities. Stell (2003) starts from a partially ordered time scale and dynamic objects with four fundamental operations on these objects (birth, death, merge and split), similar to Hornsby and Egenhofer (1997). He then introduces granularity of objects by two further operations, amalgamation and selection. This means his concept of granularity is based on sets. In contrast, the granularities discussed in this paper are related to observations. A discussion of spatiotemporal data model can be found in Abraham and Roddick (1999).

As discussed in Section 1, decentralized approaches in WSNs can be classified into snapshot-based approaches and event-based approaches. Lian et al. (2007) propose a gradient boundary detection approach, where snapshots of a monitored region are transmitted to a designated central computer via the *sink* of the network at certain time intervals. A snapshot at a given time t is abstracted as a contour map, which consists of gradient boundaries. Only the nodes on the gradient boundaries need to report to the sink, which constructs a contour map. A unique work for collecting snapshot data is proposed by Skraba et al. (2006). Their method is called a sweep, i.e., a wavefront that traverses the whole network and passes all nodes in the network exactly once. Sarkar et al. (2008) apply the sweep method to construct iso-contours in a continuous scalar field. Given the constructed iso-contours of a snapshot, users can acquire contours at given values or ranges. All the above approaches are limited to static snapshots. They cannot dynamically analyze snapshot sequences to derive salient changes.

Duckham et al. (2005) develops a model to track salient changes or events in the environment. They model a sensor network as a triangulation network. And the triangulation can change over time in response to the movement of spatial phenomena. Worboys and Duckham (2006) also use triangulations, and spatial objects is represented as a set of triangles. Changes of spatial objects are represented as the insertion or deletion of triangles from the set. However, their approach is limited to incremental changes, i.e., the change of a single triangle at each time step. In contrast, our approach allows both abrupt and incremental changes. Sadeq and Duckham (2008) investigate different sensor network structures for detecting topological events. The fundamental topological events include appearance, disappearance, merge and split, etc. Their approach also assume incremental changes. Their classification of events is based on the theoretical study of Jiang and Worboys (2009). Although there are a large amount of existing work on sensor network data models, this paper is the first work to investigate areal object lifespans in snapshot sequences.

3 SPATIOTEMPORAL DATA REPRESENTATION

For geographic data representation, there is a distinction between field-based and object-based models. Our spatiotemporal data model is an object-based model, and we use an object-oriented approach to implement our model. In our approach, a WSN is modeled as a set of point objects with point observations. An appropriate data structure is essential for efficient interaction or communication among point objects. The primary structure in our model is an edge. Edges are important components for spatial objects. For example, an area is embedded in a polygon that is constituted as a sequence of edges. By sensor network point observations, geographic phenomena will be abstracted as areal objects that consist of points, and these points will then be organized as sequence of edges.

Figure 1 illustrates an example about the representation of dynamic areal objects in a sensor network. Suppose Figure 1(a)-(d) are at four different time steps, e.g., t_1, t_2, t_3, and t_4. Between t_1 and t_2, two polygons expand. Between t_2 and t_3, the two polygons merge into a larger polygon. And the larger polygon expands again between t_3 and t_4. If the temporal granularity of observation is fine enough to acquire all the four snapshots, then we can infer three changes or events, i.e., expand, merge, and expand, between the four snapshots respectively. If the given temporal granularity can only acquire two snapshots at t_1 and t_4 in Figure 1(a) and (d), then only one event, i.e., merge, can be inferred between the two snapshots. These changes between snapshots may include incremental changes, but there are also non-incremental changes, for example, the change between Figure 1 (c) and (d). By allowing granularity of observations in our model, we should also allow granularity of changes.

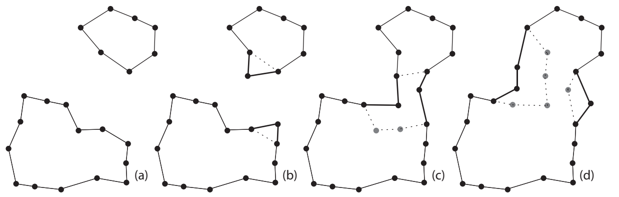


Figure 1: The evolving of spatial objects over time. (a)-(d) are at different time steps, e.g., t_1, t_2, t_3, and t_4.

There are always tradeoffs between granularity of observations and granularity of spatiotemporal data. If a finer granularity of observation is applied to a sensor network, more energy will be consumed for observation, communication, and data processing in the network, but the acquired data would have higher quality. On the other hand, coarser granularity of observation means less energy consumption and lower data quality. Our model can be applied to different granularity of observations. Note that there is limited number of point observations in a sensor network deployed area, which means spatial granularity of observation has a limit. Even the finest temporal and thematic granularity of observations are applied, inaccuracy and imprecision will be always associated with the observed spatiotemporal data. In this paper, we assume a certain granularity of observations can be decided for an application, so that the energy consumption and the accuracy of spatiotemporal data can be balanced.

This paper will focus on decentralized tracking of dynamic areal objects in a sensor network under certain temporal granularity of observations, as the example in Figure 1. Since temporal granularity of observation is taken into consideration, a sequence of snapshots will be acquired by sensor networks. Each snapshot in the sequence will contain a set of areal objects. By the evolution of areal objects over time, such as expansion, contraction, split, and merge, edges of areal objects could remain unchanged, or be inserted or deleted. The insertion and deletion of edges are important for our data model. In the rest of this section, we will define the representation of inserted and deleted edges in a sensor network.

Firstly, a sensor network can be modeled as a directed planar graph $G = (V, E)$, which is built by Delaunay triangulation (Preparata and Shamos, 1985). In the graph G, V is a set of nodes and E is a set of edges, e.g., (v, v'), which represent the direct communication links between the nodes $v \in V$ and $v' \in V$. We assume that E is symmetric, i.e., if there are $v, v' \in V$, $(v, v') \in E$, then $(v', v) \in E$. Note that the direction of a directed edge will be illustrated in a figure when it is relevant, otherwise the representation of G can be simplified as in Figure 2(a).

The set of neighbors of $v \in V$ is denoted as $neighbor(v) = \{v' : (v, v') \in E\}$. For example in Figure 2(a), $neighbor(v) = \{a, b, c, d, e\}$, and the set $neighbor(v)$ is sorted into clockwise order. Our approach is a decentralized approach in the sense that each point object only stores information about itself and its immediate neighbors. Each point object has its own datasets, and global knowledge, e.g., the information about the whole network G, is unavailable in point objects.

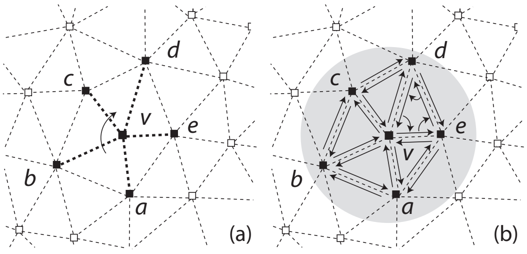


Figure 2: Sensor network structure. (a) An example of a communication graph G, where the set $neighbor(v)$ of each node v is sorted into clockwise order. (b) An areal object, illustrated as an solid circle, is represented by a sequence of boundary edges (a, b), (b, c), (c, d), (d, e), and (e, a).

A node v and its neighbors are organized into *directed cycles*, or simply *cycles*. A cycle is denoted as (v, b, a, v), where the three nodes v, a and b are distinct, and there is an edge for any two consecutive nodes in the cycle, i.e., (v, b), (b, a), $(a, v) \in E$. Examples of cycles can be found in Figure 2(b), where (v, b, a, v) and (v, c, b, v) are cycles. Since $neighbor(v)$ is in clockwise order, all the cycles are counterclockwise, and an example is shown in Figure 2(b). Based on the above definition, we can define the representation of areal objects in a sensor network:

Definition 1 *A directed edge $(v, v') \in E$ is an* object edge, *if both v and v' are located in an areal object.*

Definition 2 *A cycle (v, b, a, v) is an* object cycle, *if all the vertexes of the cycle, i.e., v, a, b, are located in an areal object.*

Definition 3 *An object edge $(v, v') \in E$ is a* non-boundary edge, *if the edge belongs to an object cycle.*

Definition 4 *An object edge $(v, v') \in E$ is a* boundary edge, *if it does not belong to any object cycles.*

Areal objects are represented as sequence of boundary edges in the network. An example is shown in Figure 2(b), in which the nodes v, a, b, c, d and e are located in an areal object. The edges (a, b), (b, c), (c, d), (d, e), (e, a) are boundary edges, since they do not belong to any object cycles.

Since areal objects are evolving over time, sequences of boundary edges would dynamically change at different snapshots, as illustrated in Figure 3. The difference of boundary edges between two consecutive snapshots is represented by the insertion and deletion of boundary edges. If a boundary edge appears between t_{i-1} and t_i, then we will insert the boundary edge into relevant datasets with a timestamp of t_i, and vise versa. Each sensor node in the network may be involved in the insertion and deletion of boundary edges in its own datasets. To represent the change of areal objects in a sensor network, we require two definitions:

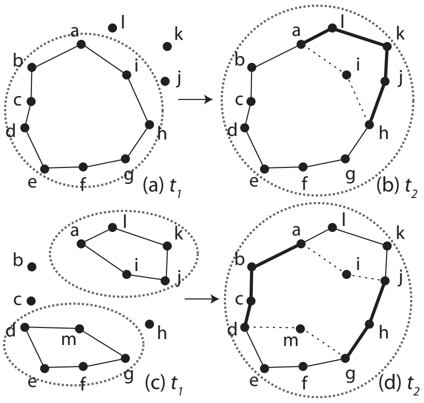


Figure 3: Examples of creation and deletion of edges. The created edges are denoted as thick continuous lines, and the deleted edges are denoted as dashed lines. Areal objects are marked by dashed-line ellipses. Note that the directions of edges are not shown in the figure.

Definition 5 *A boundary edge $(v, v') \in E$ is an* inserted edge, *at time t_i if the edge is not a boundary edge at previous time t_{i-1}, but become a boundary edge at time t_i.*

Definition 6 *A boundary edge $(v, v') \in E$ is an* deleted edge, *at time t_i if the edge is a boundary edge at previous time t_{i-1}, but is not a boundary edge at time t_i.*

The inserted and deleted edges will be decentrally stored at different sensor nodes in the network with a timestamp. Two local datasets, i.e., $insertedEdge(v, t)$ and $deletedEdge(v, t)$, are required at each node v to store inserted and deleted edges over time. For example, an area object is expanded from Figure 3(a) to (b), so that edges al, lk, kj, and jh have been inserted and edges ai and ih have been deleted. And the other edges remain unchanged. Six nodes $a, i, h, j, k,$ and l need to record the insertion and deletion of boundary edges. For example, there are $insertedEdge(a, t_2) = \{(a, l)\}$ and $deletedEdge(a, t_2) = \{(a, i)\}$ for the node a.

4 AREAL OBJECT IDENTITY

4.1 Leader and Trajectory

In our model, spatiotemporal data is decentrally stored in each node in the network. These spatiotemporal data acquired at different time steps may need to be accessed for qualitative and quantitative analysis. For example, two consecutive snapshots at time steps t_i and t_{i-1} could be used to calculate the area difference of an areal object between two time steps. This section introduces the concept of leader and trajectory, which will enable the decentralized retrieval of historical spatiotemporal data.

We define a leader of an areal object as a selected node that is currently located inside the areal object. A leader trajectory, or simply trajectory, is the path of the leader over time. A trajectory consists of a sequence of edges, called trajectory edges. Since areal objects are changing over time, the leaders and leader trajectories need to be dynamically updated. Figure 4(a)-(d) show a general idea about the update of leaders and leader trajectories. Assume that there is no areal object at initial time step t_0. Between time step t_0 and t_1, an areal object appears. A node a on the boundary edges of the areal object is selected as the leader. The leader a reports the appearance of the areal object to the sink by greedy routing (Karp and Kung, 2000). During the routing, the information of the leader a will be recorded along the path from a to the sink. We regard the path between the leader a and the sink as one part of the leader trajectory.

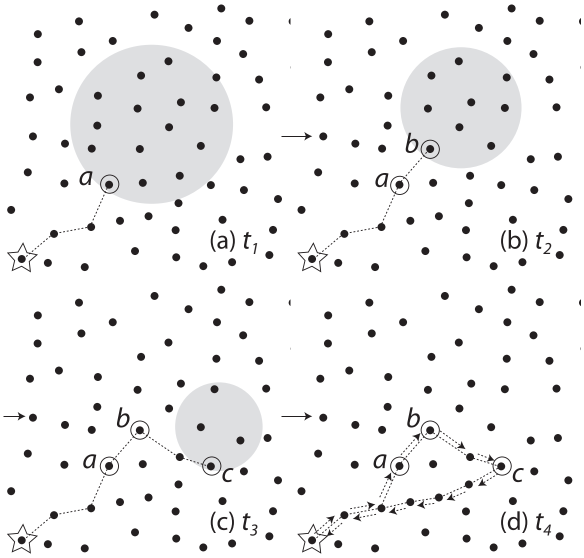


Figure 4: Leaders and leader trajectories need to be updated when areal object changes over time. Leaders are marked by circles, and dashed lines represent the leader trajectories.

As in Figure 4 (a) and (b), the areal object contracts between t_1 and t_2, and the leader a is not longer in the areal object. The node b is then selected as the new leader, and the path between the two leaders a and b is added into the leader trajectory. Similarly, at t_3 in Figure 4(c), there is an update of the trajectory between leader b and c. Finally, between t_3 and t_4, the areal object disappears. The leader c will report the disappearance of the areal object to the sink. And at the same time, the nodes along the path between c and the sink will record the trajectory. A completed trajectory that enables the access of the full history of the areal object is then formed at the sink. The completed trajectory is a closed traversable trail, as shown in Figure 4(d). The leader trajectory can be considered as a chronicle that links all the relevant snapshots of an areal object.

4.2 Trajectory Identity

To correctly access spatiotemporal data of areal objects, each leader trajectory should have an identity. Since each areal object has only one trajectory, the trajectory identity is also used as the identity of the areal object. The identity of a trajectory will be decentrally stored in each trajectory edge. For example, in Figure 5(a), node a is a leader and a trajectory identity, e.g., A, is given to each edge on the trajectory. To correctly access each snapshot of an areal object from the trajectory, each boundary edge of the areal object at each snapshot should also have an identity. And the identity of boundary edges should match the identity of the trajectory. In Figure 5(a), each boundary edge of the areal object is also provided the same identity A.

The identity A will be used to identify the trajectory over time. Thus, when the areal object changes over time, the same identity should be maintained. In Figure 5(b), the areal object has contracted at t_2, and the leader is changed from a to b. The same identity A should be updated for both new trajectory edges and new boundary edges, as shown in Figure 5(b).

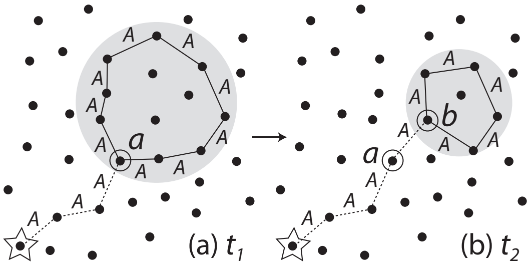


Figure 5: Trajectory Identity. Boundary edges of the areal object have the same identity as the trajectory edges.

As discussed in Section 3, each boundary edge of areal objects has a timestamp. For instance, the boundary edges in Figure 5 (a) and (b) are associated with a timestamp of t_1 and t_2 respectively. The snapshot sequences of areal objects can be easily retrieved using trajectory and its identities. For example, if a query inserted from the sink is interested in the areal objects at t_2, then the relevant data can be retrieved by a traversal from the sink to the leader b. And the leader b (with identity A) can then traverse all the boundary edges that have a timestamp of t_2 and have the identity of A. Before the trajectories can be used for accessing spatiotemporal data in a decentralized network, we firstly need to maintain the trajectory identities when areal objects change over time.

4.3 Trajectory Maintenance

There are six types of changes of areal objects that require possible updates of trajectories. These six types of changes are: appearance, disappearance, expansion, contraction, merge, and split. When an areal object appears, a new trajectory needs to be created for the areal object, and the trajectory needs a unique identity. Figure 6 has a abstract representation of the changes of trajectories. In the figure, the star represents the sink, circles or nodes represent leaders, and a continuous line represents a path between two nodes. In Figure 6(a)-(b), an areal object appears, a leader is selected, and an identity of A is given to the leader. Also the path between the leader and the sink has the same identity A. An areal object with a trajectory identity A will be simply called areal object A.

As discussed in the previous section, when an areal object expands or contracts, the trajectory of the areal object may need to be updated. In Figure 6(c), a new leader is selected for the trajectory A due to the expansion of the areal object, and there is a path between the two leaders. Between t_2 and t_3, areal object A contracts and a new areal object B appears, and their trajectories are updated correspondingly.

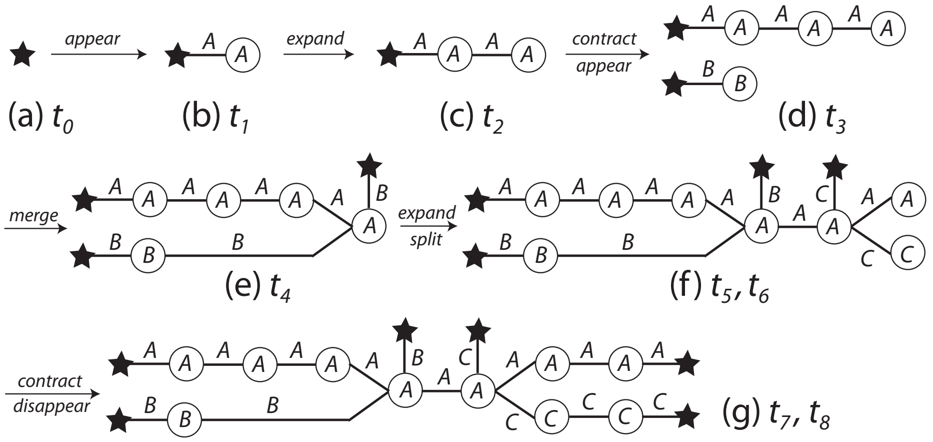


Figure 6: The trajectories need to be updated when areal objects change over time. The star represents the sink, different circles or nodes at one time step represent different leaders, and a continuous line represent a path between two nodes. There may be multiple stars at one time step, but these stars represent the same sink at a single location.

If two areal objects A and B merge, the two trajectories of the areal objects have to merge into one trajectory, as in Figure 6(e). One identity, e.g., A, can remain for the merged trajectory, and the other trajectory B is concluded to the sink. With the conclusion of trajectory B, the sink will also aware the merge of two areal objects A and B. In contrast, if an areal object A splits into two areal objects, the trajectory A has to split into two trajectories, e.g., A and C, as shown in Figure 6(f). In Figure 6(g), both areal objects A and C has disappeared, and both trajectories A and C conclude to the sink.

5 DECENTRALIZED ALGORITHMS

Since our algorithms are run decentrally in each sensor node in the network. The algorithm for updating trajectory identity will be dependant on different types of nodes in the network. We assume that each sensor node has only binary readings, i.e., 0 and 1. If a sensor node is located inside an areal object, its reading is 1, otherwise the reading would be 0. We define that a node is an *active* node if it changes readings between two consecutive time steps. We use the set $active(t)$ to represent all active nodes at the time step t.

In Figure 7 (a) and (b), for instance, an areal object expands from t_1 to t_2, and there are inserted edges (c, d), (d, e), and (e, a), deleted edges (c, v) and (v, a), and active nodes d and e. We define another type of edges:

Definition 7 *An inserted or deleted edge is a transition edge if the edge connects an active node and a non-active node.*

For example, in Figure 7(b), the inserted edge (c, d) is a transition edge that has a initial node $c \notin active(t_2)$ and a terminal node $d \in active(t_2)$. Also (e, a) is a transition edge, but the edge begins at an active node e and ends at a non-active node a.

Transition edges are always in pairs. If there is a transition edge with an initial non-active node and a terminal active node, then there always exists another transition edge with an initial active node and a terminal non-active node. A pair of transition edges is always connected by a trial. For example, (c, d) and (e, a) are connected by the trial $c \rightarrow d \rightarrow e \rightarrow a$. The set of transition edges of a node v is denoted as $transitionEdge(v, t)$, for instance, $transitionEdge(c, t_2)$ = $\{(c, d)\}$ in Figure 7(b).

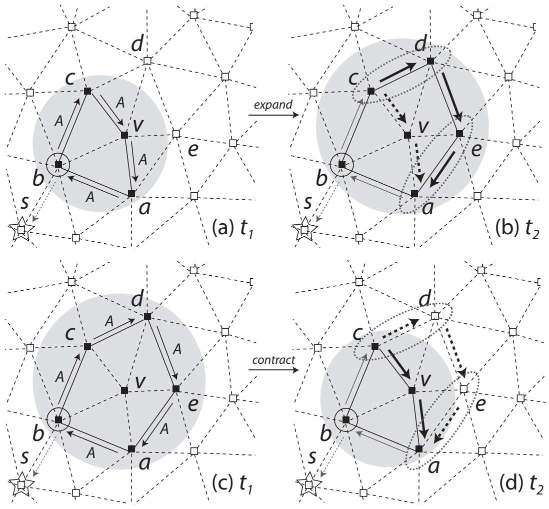


Figure 7: Expansion and contraction. Leaders are marked by circles, and transition edges are marked by dashed-line ellipses.

As discussed in Section 4.3, trajectories need to be updated differently when areal objects undergo different types of changes over time. Our decentralized algorithms should firstly distinguish different types of changes, so that trajectories can be update properly. A large number of nodes may be involved in the changes of areal objects, and thus in-network aggregations among relevant nodes are necessary. In our algorithm, graph traversals are used to organize the aggregation of decentralized spatiotemporal data. If all the relevant nodes initialize traversals in the network, there would be a large number of redundant traversals, and the efficiency of the algorithm will be significantly reduced. In this case, only a small subset of sensor nodes is eligible to initialize traversals.

In the algorithm, only the node v that has transition edges, i.e., $transitionEdge(v, t) \neq \emptyset$, will be eligible for starting traversals. And the traversals will always start at non-active nodes. For example, in Figure 7(b), $\{a, c\} \notin active(t_2)$ and $\{d, e\} \in active(t_2)$, so the traversal would start from non-active nodes a or c. Since node c is the initial node of the transition edge (c, d), node c will start a traversal to visit all inserted edges, i.e., $c \rightarrow d \rightarrow e \rightarrow a$, and reach another non-active node a. Node a will then notice that c and a have the same identity A, and the traversal has visited three inserted edges (c, d), (d, e), and (e, a). An expansion of the areal object A is detected at node a. Similarly, in Figure 7(d), the non-active node c will start a traversal to visit all deleted edges: $c \to d \to e \to a$. And node a will detect a contraction of areal object A.

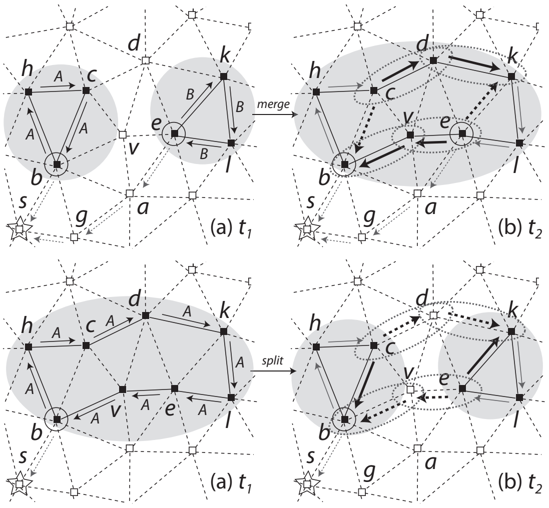


Figure 8: merge and split.

In Figure 8(b), two areal objects A and B have merged. There are two active nodes d and v and four transition edges (c, d), (d, k), (e, v), and (v, b). Both node c and e are eligible for starting traversals, since $\{c, e\} \notin active(t_2)$ and they are initial nodes of transition edges. The traversal starts from c will traverse inserted edges (c, d) and (d, k) and reach another non-active node k: $c \to d \to k$. The other traversal from e is $e \to v \to b$. Node k will notice c and k have different identities, and the traversal has visited inserted edges. Thus, a merge of two areal objects is detected at node k. Note that b can also detect the merge of two areal objects. In Figure 8(c) and (d), an areal object splits into two areal objects. Similarly, splits can also be detected by traversals.

6 CONCLUSION

This paper has introduced a new spatiotemporal data model that enables the storage of areal object lifespan in snapshot sequences. Each areal object has a unique trajectory that links all the relevant snapshots of the areal object together. The evolving of trajectories is dependent on the changes of areal objects. By the maintenance of trajectories in a decentralized network, different types of salient changes or events can also be derived by decentralized processing of snapshot sequences. An important area of future work is to use the areal object trajectories for qualitative and quantitative analysis of a static snapshot or dynamic snapshot sequences.

References

Abraham, T. and Roddick, J., 1999. Survey of spatio-temporal databases. Geoinformatica 3(1), pp. 61–99.

Duckham, M., Nittel, S. and Worboys, M., 2005. Monitoring dynamic spatial fields using responsive geosensor networks. In: Proceedings of 13th Annual ACM International Workshop on Geographic Information Systems (GIS05), pp. 51–60.

Galton, A., 2000. Qualitative Spatial Change. Oxford University Press.

Galton, A., 2004. Fields and objects in space, time, and space-time. Spatial Cognition and Computation 4(1), pp. 39–67.

Grenon, P. and Smith, B., 2004. Snap and span: Towards dynamic spatial ontology. Spatial Cognition and Computation 4(1), pp. 69–104.

Hornsby, K. and Egenhofer, M., 1997. Qualitative representation of change. In: S. Hirtle and A. Frank (eds), COSIT 1997, Volume 1329 of Lecture Notes in Computer Science, Springer, Heidelberg, pp. 15–33.

Jiang, J. and Worboys, M., 2009. Event-based topology for dynamic planar areal objects. International Journal of Geographical Information Systems 23(1), pp. 33–60.

Karp, B. and Kung, H., 2000. Gpsr: Greedy perimeter stateless routing for wireless networks. In: MobiCom 00: Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, ACM Press, New York, NY, USA, p. 243254.

Lian, J., Chen, L., Naik, K., Liu, Y. and Agnew, G., 2007. Gradient boundary detection for time series snapshot construction in sensor networks. IEEE Transactions on Parallel and Distributed Systems 18(10), pp. 1462–1475.

Peuquet, D., 2001. Making space for time: Issues in space-time data representation. Geoinformatica 5(1), pp. 11–32.

Peuquet, D. and Duan, N., 1995. An event-based spatiotemporal data model (estdm) for temporal analysis of geographical data. International Journal of Geographical Information Systems 9, pp. 7–24.

Preparata, F. and Shamos, M., 1985. Computational Geometry: An Introduction. Springer, New York.

Sadeq, M. and Duckham, M., 2008. Effect of neighborhood on in-network processing in sensor networks. In: T. C. et al. (ed.), GIScience 2008, LNCS, volume 5266, Springer, Heidelberg, pp. 133–150.

Sarkar, R., Zhu, X., Gao, J., Guibas, L. and Mitchell, J., 2008. Iso-contour queries and gradient descent with guaranteed delivery in sensor networks. In: Proceedings of the 27th Conference on Computer Communications, IEEE, pp. 960–967.

Skraba, P., Fang, Q., Nguyen, A. and Guibas, L., 2006. Sweeps over wireless sensor networks. In: IPSN 06: Proceedings of the 5th international conference on Information processing in sensor networks, p. 143151.

Stell, J., 2003. Granularity in change over time. In: M. Duckham, M. Worboys and M. Goodchild (eds), Foundations in Geographic Information Science, Taylor & Francis, London, pp. 95–115.

Wark, T., Corke, P., Sikka, P., Klingbeil, L., Guo, Y., Crossman, C., Valencia, P., Swain, D. and Bishop-Hurley, G., 2007. Transforming agriculture through pervasive wireless sensor networks. IEEE Pervasive Computing 6(2), pp. 50–57.

Worboys, M., 2005. Event-oriented approaches to geographic phenomena. International Journal of Geographical Information Systems 19(1), pp. 128.

Worboys, M. and Duckham, M., 2006. Monitoring qualitative spatiotemporal change for geosensor networks. International Journal of Geographical Information Systems 20(10), pp. 1087–1108.

THE STUDY FOR MATCHING ALGORITHMS AND MATCHING TACTICS ABOUT AREA VECTOR DATA BASED ON SPATIAL DIRECTIONAL SIMILARITY

Guo Li, Lv Zhiping, Zhang Bin, Wang Yaoge

Institution of Surveying and Mapping, Information Engineering University, Zhengzhou, China. Postcode: 450052
Email: gl_750312@163.com

KEY WORDS: matching algorithm, matching tactics, data matching, data fusion, spatial directional relation, spatial directional similarity

ABSTRACT:

The matching technique is the key to geospatial data integration and fusion. In this paper, we describe spatial directional relation through direction relational matrix, and discuss calculation methods of the spatial directional similarity. Combining with the area vector data matching algorithms, we introduce some matching tactics, which contribute to accomplish many-to-one, many-to-many matching. We describe matching process based on spatial directional similarity and matching tactics. Then we take a test, using building property area data and town planning area data at a scale of 1：500. In the test, we take the spatial directional relation matrix algorithm, and adopt two-steps matching tactics. Based on the test, we draw a conclusion: matching efficiency depends on not only algorithms, but also tactics. Proper tactics can help us accomplish more complex task.

1. INTRODUCTION

There are different expression forms for the same phenomenon. Data matching technique can recognize the same object from different expression forms. So data coherence matching technique is the key to spatial data integration and fusion. There are abundant semantic information, complex topological relations, different geometry shapes and location differences in the vector spatial data, so the automatic matching techniques is always the difficult point and hotsNW ı the corresponding study (Zhang Qiaoping, 2002).

Spatial vector data matching approaches consists of geometry matching, topological matching and semantic matching. In order to improve the matching efficiency and validity, we also have to make full use of matching tactics. We combine the spatial directional similarity algorithm and two step tactics to realize complex area objects matching.

2. SPATIAL DIRECTION CONCEPTION DESCRIPTION

Spatial direction can be described as an orientation from one spatial object pointing to another object in a direction reference system. There are two aspects, quantitative and qualitative attributes can be used to describe spatial directional relation. Quantitative directions model gives an accurate direction angle to depict the direction. Qualitative directions model have 8 kinds of models, including direction relation matrix model (Deng Min, 2006). We adopt directional relation matrix model to describe directional relation and to calculate similarity value.

2.1 Description for Spatial Directional Relation

Relation matrix model divide the space into 9 absolute direction pieces (Figure 1). The outer 8 pieces express 8 directions, and the centre piece is named of O. Generally, the reference object lies in the centre piece(Guo Qingsheng, 2004).

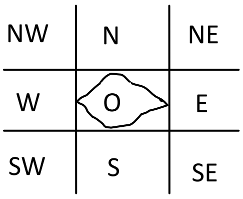


Figure 1. Spatial directions separations

2.2 Spatial Direction Relation Matrix

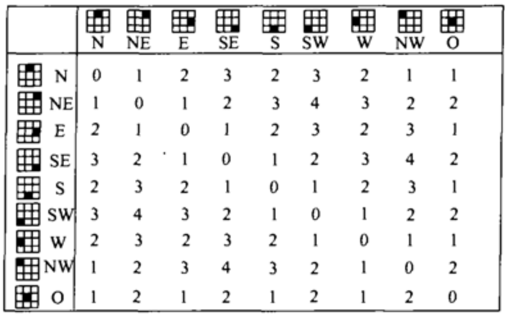
If A is a reference object, we divide the space according the minimum bounding rectangle of the reference object. A 3×3 direction relation matrix is the following(Ding Hong, 2004):

$$dir(A,B)=\begin{bmatrix} A_{NW}\cap B & A_N\cap B & A_{NE}\cap B \\ A_W\cap B & A_0\cap B & A_E\cap B \\ A_{SW}\cap B & A_S\cap B & A_{SE}\cap B \end{bmatrix} \qquad (1)$$

Where A is a reference object and B is the target object.
A_{NW}, A_N, A_{NE}, A_W, A_O, A_E, A_{SW}, A_S, A_{SE} are 9 direction pieces.
$A_{NW}\cap B$ means the part of B, which is located in the direction of piece A_{NW}.

2.3 Spatial Directional Distance

Spatial directional distance represents the difference of spatial direction pieces; and its inverse presents spatial direction similarity. As the object changes its direction, the distance changes accordingly. The distance is defined in figure 2. For example, as the direction changes in one direction piece, the distance is defined 0; as the direction changes from NE to SW or from NW to SE, the distance is defined 4. Other instance refers to figure 2 (Roop K Goyal, 2000).



	N	NE	E	SE	S	SW	W	NW	O
N	0	1	2	3	2	3	2	1	1
NE	1	0	1	2	3	4	3	2	2
E	2	1	0	1	2	3	2	3	1
SE	3	2	1	0	1	2	3	4	2
S	2	3	2	1	0	1	2	3	1
SW	3	4	3	2	1	0	1	2	2
W	2	3	2	3	2	1	0	1	1
NW	1	2	3	4	3	2	1	0	2
O	1	2	1	2	1	2	1	2	0

Figure 2. The distance of four neighbouring directions pieces

2.4 Equation of Spatial Directional Similarity Value

The spatial directional similarity value between two directional is defined as the following equation (Roop K Goyal, 2000):

$$S_i(D^0, D^1) = 1.0 - \frac{d(D^0, D^1)}{D_{\max}} \tag{2}$$

Where D_{max}= 4

$d(D^0,D^1)$ = the least direction distance from matrix D^0 to matrix D^1.

$S_i(D^0,D^1)$ = the spatial directional similarity value of D^0 to matrix D^1.

3. CALCULATION FOR SPATIAL DIRECTIONAL SIMILARITY VALUE

Equation 2 is tight, but is hard to calculate. We can realize the algorithm of spatial directional similarity calculation efficiently if we adopt raster data format. Let's take an example, the directional relation matrix between the target area object and the reference area object can be changed to calculate the matrix summation between all the raster cells of the target area object to the reference area object.

3.1 Calculation for Spatial Directional Similarity in Ideal Instance

We can get the average directional distance value by adding up each raster cell direction distance value and dividing by the sums of cells. The formula of calculating the area objects similarity value is the following (Ding Hong, 2004):

$$S(I_r, I_p) = 1 - \frac{1}{4n}\sum_{i=1}^{n} d_i \tag{3}$$

Where I_r = the reference object

I_p = the target object

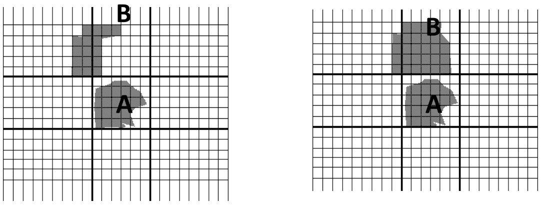
n = the raster cells numbers of the target object

d_i = the moving distance of each cell

$S(I_r,I_p)$ = the similarity value of the two object.

3.2 Calculating for Spatial Directional Similarity in Ordinary Instance

Because of different data sets, the same object may present the different shape. The above-mentioned method does not adapt to the ordinary instances. How does it to calculate the similarity between the objects with different shape? We can solve it by this way.
Let's take Figure 3 for example.



3(a) 3(b)

Figure 3. Similarity calculation for object with different shape

Firstly, we take A as the reference object. B is the object with different shape in 3(a) and 3(b).
We calculate the similarity value of B from figure 3(a) to figure 3(b). We take the raster cells of B in figure 3(a) as norm. The principle is:
① comparing to B in figure 3(b), if the cell does not exist in B of figure 3(a), but exist in B of figure 3(b), the moving distance is 0.
② comparing to B in figure 3(b), if the cell exist in B of figure 3(a), but does not exist in B of figure 3(b), the moving distance is 4.
③ the other instance, referring to figure 2.
The calculating course is as follows:
The number of raster cells in B of figure 3(a) is 16. There are 5 cells in B of figure 3(a) moving form NE to N comparing to B of figure 3(b). The spatial directional similarity value of B from figure 3(a) to figure 3(b) is *Sim*1(*a,b*).

$$Sim1(a,b) = 1 - \frac{1}{4\times 16}(1\times 5) = 0.92 \tag{4}$$

Secondly, we take the raster cells of B in figure 3(b) as norm. Then we calculate the similarity value of B form figure 3(b) to figure 3(a). The number of raster cells in B of figure 3(b) is 30. There are 5 cells in B of figure 3(b) moving from N to NE, comparing to B of figure 3(a). Furthermore, there are 13 cells exist in B of figure 3(b), but don't exist in B of figure 3(a), the moving distance is 4. The spatial directional similarity value of B from figure 3 (b) to figure 3 (a) is *Sim*2(*a ,b*)

$$\begin{aligned} &Sim2(a,b) \\ &= 1 - \frac{1}{4\times 30}(1\times 5 + 4\times 13) = 0.525 \end{aligned} \tag{5}$$

Finally, we set a threshold *K*. If the two similarity values are both bigger than *K*, we think that the object B in figure 3(a) and figure 3(b) are the same object. It is one to one match.

3.3 Matching Tactics for Complex Instances

The above-mentioned method can realize one-to-one match, but cannot solve matching problems of one-to-many, many-to-one, especially many-to-many. In order to solve complex matching question, we adopt some matching tactics. The main thought is the following:

There are two different data sets, *A* and *B*. The area objects in dataset *A* are $\{a_1,a_2,...,a_i,...,a_m\}$, and area objects in data B are $\{b_1,b_2,...,b_j,...,b_n\}$. The objects numbers of the two datasets may be not equal, that is to say, m may be not equal to n. The aim for match is to find out homonym entities which exist in different data sets.

3.3.1 One-to-one Matching Tactic

There are entities:

$a_i \in A, i=1,2,...m$;

$b_j \in B, j=1,2,...,n$;

A and *B* are two data sets.

If: $Sim(a_i,b_j) \geq$K, and $Sim(b_j,a_i) \geq$k;

K is the threshold, and $0<K<1$.

We can draw a conclusion: the object b_j in data set B is the homonym to a_i; meanwhile a_i in data set A is the homonym to b_j. So, a_i and b_j are homonym entities, and the matching relation is one-to-one matching. The chart is shown in the Figure 4.

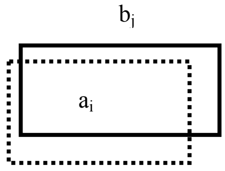


Figure 4. One-to-one matching

3.3.2 Many-to-one Matching Tactic

There are entities $a_i,a_t,b_j,$:

$a_i \in A$, $i=1,2,...m$;

$a_t \in A$; $t=1,2,...m$;

$i \neq t$;

$b_j \in B, j=1,2,...,n$;

A and *B* are two data sets.

If: $Sim(a_i,b_j) \geq$K and $Sim(a_t,b_j) \geq$K;

K is the threshold, and $0<K<1$.

We can draw a conclusion: the entity a_i and a_t are match to the entity b_j. So, the aggregation of { a_i , a_t }are match to entity b_j . The match relation is many-to-one. The chart is shown in the Figure 5.

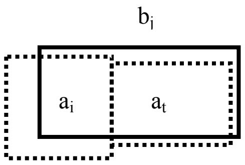


Figure 5. Many-to-one matching

3.3.3 Many-to-many Matching Tactic

There are entities a_i ,a_t ,b_j , b_k ,b_r;

$a_i \in A$, $i=1,2,...m$;

$a_t \in A$, $t=1,2,...m$;

$i \neq t$;

$b_j \in B, j =1,2,...,n$;

$b_k \in B, k =1,2,...,n$;

$b_r \in B, r =1,2,...,n;$

$j \neq k \neq r$;

A and *B* are two data sets.

If: $Sim(a_i ,b_j) \geq$K, *and* $Sim(a_i ,b_k) \geq$K,

and $Sim(a_t ,b_k) \geq$K, *and* $Sim(a_t ,b_r) \geq$K;

K is the threshold, and 0<K<1.

We can draw a conclusion:

a_i is match to $\{b_j , b_k \}$,and a_t is match to $\{ b_k ,b_r \}$.

So, the aggregation of $\{a_i ,a_t \}$are match to $\{b_j , b_k ,b_r \}$. The match relation is many-to-many. The chart is shown in the Figure 6.

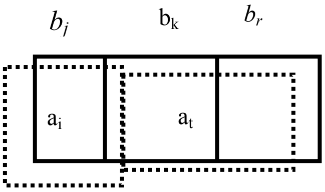


Figure 6. Many-to-many matching

4. DATA MATCHING PROCESS DESCRIPTION

If there are different data sets, we can adopt above algorithm and tactics to realize homonym entities matching. The matching process is shown in the Figure 7:

4.1 Vector Data Rasterization

In order to make use of above algorithm to realize data matching, we must change the data format from vector to raster.

4.2 Ascertain the Reference Object

In certain area, we should choose an appropriate object as reference object.

4.3 Similarity Value Calculation

We adopt above similarity value algorithm, calculate the similarity values of objects which are around the reference object.

4.4 Preliminary matching

According to experiment, we fix a range for threshold value. If the similarity values of some objects are within the scope of threshold, we can get a preliminary conclusion: these objects are probably homonym entities.

4.5 Matching Tactics

Based on the matching algorithms, the tactics can accomplish many-to-one, many-to-many matching.

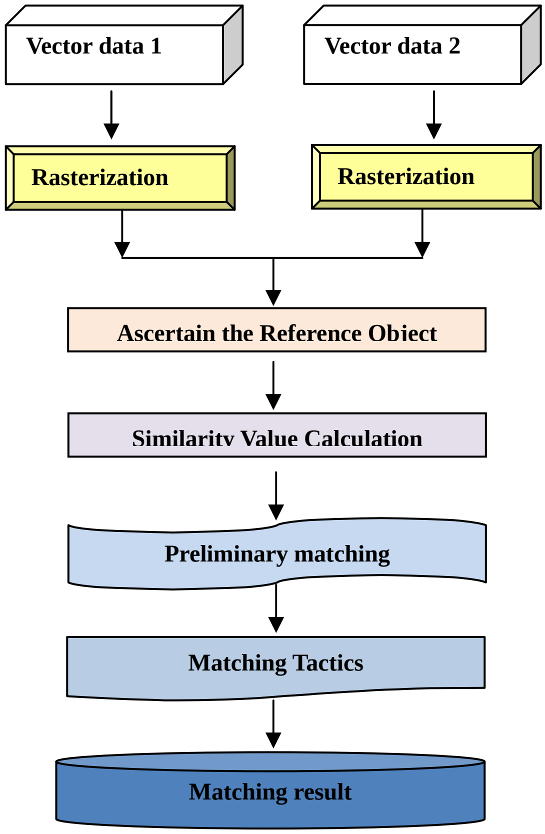


Figure 7. Matching Process

5. CONCLUSION

In the end, we take test and analysis for area vector objects data matching with building property area data and town planning area data at the scale of 1:500. The building property data is shown in the figure 8 and town planning data is shown in figure 9.

We take the spatial directional relation matrix algorithm, and adopt two-steps matching tactics. Firstly, we set up a big threshold to accomplish one-to-one matching. Secondly, we reduce the threshold to match the rest objects, realizing many-to-one, many-to-many matching.

Based on the test, we get some experimental data and make two tables. From table 1, we can see, when the thresholds are big, from 0.6 to 0.9, the accuracy of one-to-one match is from 53% to 85%. When the threshold is 0.8, the accuracy is the highest.

From table 2, we also can see that when we reduce the threshold, the accuracy of many-to-one, many-to-many matching are increased. When the threshold is 0.3, the accuracy of many-to-one, many-to-many matching is highest.

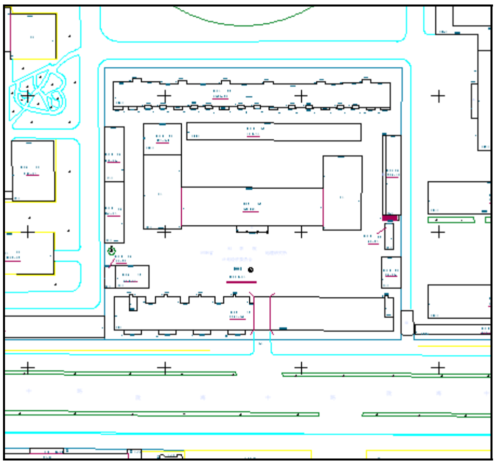


Figure 8. Building property Data

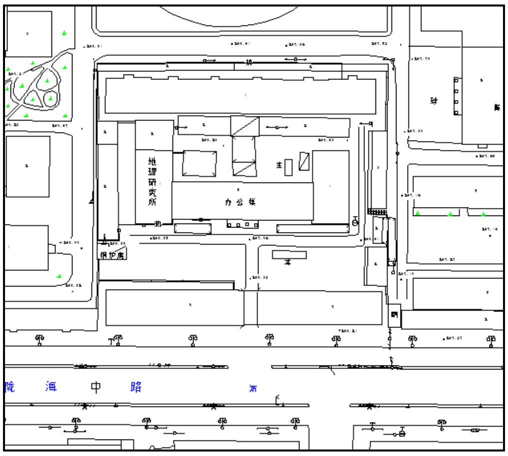


Figure 9. Town planning data

Threshold	To be matching number	Correct matching number	Accuracy (%)
0.6	72	38	53
0.7	72	58	81
0.8	72	61	85
0.9	72	51	71

Table 1 Big threshold match (one-to-one match)

Threshold	To be matching number	Correct matching number	Accuracy (%)
0.3	14	11	78
0.4	14	9	64
0.5	14	7	50
0.6	14	3	21

Table 2 Small Threshold Matching (many-to-many match)

Based on the test figures, we can draw a conclusion. The matching efficiency is not only lies on algorithm, but also depends on tactics. Proper tactics can help us accomplish more complex task.

6. REFERENCES

Deng Min, 2006， Liu Wenbao, Li Junjie, Sun Dian . Computational Model of Spatial Direction Relations in Vector GIS. *Journal of Remote Sensing*, 10(6):821-828

Ding Hong, 2004. *A Study on Spatial Similarity Theory and Calculation Model.* Doctoral Thesis, Wuhan University.

Guo Qingsheng, 2004, Ding Hong. Similarity for Spatial Directions Between Areal Objects in Raster Data. *Editorial Board of Geomatics and Information Science of Wuhan University*, 29(5):447-450

Goyal P K，2000．*Similarity Assessment for Cardinal Directions Between Extended Spatial Objects*：[Ph.D.Thesis]．Orono：The University of Maine.

Roop K Goyal, 2000. *Similarity assessment for Cardinal Directions between Extended Spatial Objects*: [Ph.D.Thesis]. The University of maine.http://www.spatial. maine.edu/ Publications /phd_thesis/ Goyal2000.pdf

Zhang Qiaoping, 2002. *The Research on Areal Feature Matching among the Conflation of Urban Geographic Databases.* Doctoral Thesis, Wuhan University.

7. ACKNOWLEDGEMENTS

The paper presented here is supported by *National High Technology Research and Development Program (863) Project*（*2009AA12Z305*）.

The author likes to thank the anonymous reviewers for their valuable comments.

LINEAR FEATURE ALIGNMENT BASED ON VECTOR POTENTIAL FIELD

David N. Siriba and Monika Sester
Institute for Cartography and Geoinformatics (IKG), Leibniz Universität Hannover
Appelstraße 9a, 30167 Hannover, Germany
E-mail: (david.siriba, monika.sester)@ikg.uni-hannover.de

KEY WORDS: Feature Alignment, Snakes, Vector Potential Field, Positional Accuracy Improvement

ABSTRACT:

An approach to align a linear feature in one dataset with a corresponding feature in another dataset that is considered more accurate is presented. The approach is based on the active contours (snake) concept, but implements the external force as a vector potential field in which case the source of the force is in vector form; further the snake feature is implemented as a non-closed snake. This is different from the conventional implementation of the snake, where the source of the external force is an image and the force is implemented as a gradient flow and usually as a closed snake.

In this approach two conditions: the length and alignment conditions have to be satisfied to obtain a good alignment. Whereas the length condition ensures that the length of the snake feature is nearly equal that of the reference feature, the alignment condition requires that the snake and the reference feature are properly aligned. The length condition is achieved by fixing the end points of the snake feature to those of the reference feature. The alignment condition is achieved by segmenting the reference feature so that there is uniform external force from all parts of the feature. One assumption in this approach is that the snake and the reference feature are matched prior to alignment. An outstanding challenge therefore is to find out how to consider the effects of non-corresponding but neighbouring reference features on a snake feature in circumstances where prior matching has not been undertaken.

1. INTRODUCTION

Feature alignment is one of the most critical steps in geometric data integration. It entails a geometric adaptation of features in two different datasets to obtain a better geometric correspondence between them, in which one of them is considered to be of a better geometric accuracy. Feature alignment is generally achieved through geometric transformations and depending on the application, different techniques can be used in feature alignment.

In a data integration problem, feature alignment is usually carried out prior to, or after feature matching. In the former case, feature alignment would involve a global transformation, while the latter would involve a local non-rigid transformation in addition to a global transformation. When the scale (resolution) and the accuracy of the datasets involved is almost similar, the expected discrepancies in position after a global transformation are not so large, otherwise the discrepancies could be quite significant.

The most common method used for feature alignment is the Iterative Closest Point (ICP) algorithm introduced by Besl and McKay (1992). ICP is a rigid transformation based on the 4 parameter Helmert transformation for a 2D case, although any transformation can be used and requires a good initial alignment. ICP takes the closest point, the problem however is that ICP will eventually match a point that is closest among all the candidate points even though the distance is very far. In this case, ICP can be implemented by specifying a certain threshold distance. An approach that does not require a specification of the threshold distance would be required in particular where the relative accuracies of the datasets involved in the alignment is not known. Active contours or snakes as commonly known have been used for feature alignment without the need for a threshold distance. Although other parameters such as the segmentation distance of reference feature, which affects the final alignment, have to be specified.

An active contour or snake is a force minimizing curve that is influenced by internal force coming from the curve itself and external forces computed from the image data (Kass et al., 1988). A curve is referred to as an active contour by the fact that the curve segments are adjusted iteratively by moving the control points that connect the vertices to lock on the edges of an image. Snakes may be understood as a special case of a more general technique of matching a deformable model to an image by means of energy minimization.

A number of modifications to the original concept of the active contours have been suggested and implemented. In most of the modifications and implementations however, the snakes form closed loops in the image, although this is not necessarily true for all the snakes. Based on the original concept, most implementations convert the reference feature to an image even when it is in vector format, in which case the external force is then modelled as a gradient flow.

In this paper, an approach that implements a non-closed snake and the external force as a vector field is presented. The development of this approach was motivated by a problem in data integration, in particular feature alignment of two vector datasets, i.e., the snake and the reference feature are both in vector format. Furthermore, the accuracy of the snake feature is unknown (Siriba, 2009). The approach wants to avoid converting the reference feature to an image.

The next section of this paper presents an overview of active contours. This is followed by a discussion of vector based snakes that form the basis of the approach presented in this paper. The experimental results of the implementations of the individual and combined forces is presented, which is then followed by the evaluation of the quality of the alignment using relative curve length and the Hausdorff distance. Conclusions and suggestions for further work come at the end of the paper.

2. RELATED WORK

The original definition of active contours or snakes is that they are curves defined within an image domain that can move under the influence of internal forces coming from within the curve itself and external forces computed from the image data. In most image processing applications, snakes are used particularly to locate object boundaries, model shapes, segmentation of images, motion tracking (Xu and Prince, 1998). In another novel application, snakes have been used in the smoothing of line objects within the framework of linear feature generalization (Burghardt, 2005) and for displacement (Burghardt and Meier, 1997).

The snake as a spline is represented as a parametric curve as:

$$V(s) = (x(s), y(s)) \tag{1}$$

Where (s) is proportional to the curve length, x and y are the curve coordinates.

The snakes's force function is composed of internal and external force components and is given as:

$$E_{snake} = \int_0^1 E_{snake}(V(s))ds = \int_0^1 (E_{\text{int}ernal}(V(s)) + E_{external}(V(s)))ds \tag{2}$$

In discrete form (2) could be represented as:

$$E_{snake} = \sum_{i=1}^{n} (E_{\text{int}ernal}(i) + E_{external}(i)) \tag{3}$$

Where *i* denotes a vertex point on the snake and *n* is the total number of vertices in the snake.

The internal spline force is composed of the first and the second order terms, which are the first and second derivatives of the curve. The first (elasticity) term keeps the snake from stretching or contracting along its length, while the second (bending) term keeps the snake from bending too much. The internal force can be represented as:

$$E_{Internal} = \alpha(s)\left|v_s(s)\right|^2 + \beta(s)\left|v_{ss}(s)\right|^2 / 2 \tag{4}$$

Where v_s and v_{ss} are the first and the second derivatives, $\alpha(s)$ and $\beta(s)$ are functions of the arc length along the snake. These are similar to *R* and *K* variables in (6). In discrete form (4) becomes:

$$E_{Internal}(i) = \alpha_i\,|v_i - v_{i-1}|^2 + \beta_i\,|v_{i-1} - 2v_i + v_{i+1}|^2 /2 \tag{5}$$

The external force provided by the image is a weighted combination of force functional which attracts the snake to lines, edges and terminations. These energies are functions of the image intensity, image gradients and curvature of level line respectively. A discussion on the implementation and numerical methods are available in (Kass et al., 1988).

In the various implementations of the snakes, there are notable differences in some basic aspects. For instance, the snakes form closed loops in the image, although not necessarily true for all the snakes. This means that open snakes are also possible as implemented in (Burghardt and Meier, 1997; Burghardt, 2005). Moreover, for closed snakes, depending on the problem and the initialization, the snake can be made to grow or shrink, although the latter case is common.

The active contours as initially conceived require edges in an image onto which they are accurately localized (Kass et al., 1988). The edges provide the external force to pull the snake. Although the external force is modeled based on the image intensity or the magnitude of the image gradient, the external force can be modeled as a gravitational field (Honea et al., 2002), in which case the snake is attracted to edges in the image even if they are some distance away.

Active contours are further differentiated as either parametric or geometric. While the original model proposed by (Kass et al., 1988) constitutes a parametric implementation, a geometric implementation is proposed by (Caselles et al., 1993). In both models, most implementations define single closed object boundaries. In a situation where multiple objects have to be handled, the maintenance of the topology is important. A method called network snakes that incorporates a complete topological and shape control was introduced by Butenuth (2008).

Another aspect of active contours which differ in concept and implementation is the manner in which the force is minimized. Some of the methods used to minimize the force of the active contour include finite differences as used in the Eulerian equations (Kass et al., 1988), dynamic programming and greedy algorithm (Lam and Yan, 1994).

These differences in the various aspects of active contours provide an opportunity for various combinations for different applications. While the approach presented in this paper generally uses the original concept, it differs with most implementations in that the source for the external force does not have to be converted to an image as used for instance in (Song et al., 2006, Burghardt, 2005) for feature displacement and line smoothing respectively. As a consequence of this, the external force is modelled as a vector potential field instead of image gradient. Although this approach has considered the snake paradigm proposed by Honea et al. (2002), it differs from the concept by considering the source of the external force as a vector instead of an image. Although Bader (2001) modelled active contours as force vectors rather than as an image, it is informed by the traditional concept of snakes. Again, in this implementation, the snake is assumed to be an open snake as opposed to closed snakes which are very common, particularly in image processing. In general this approach uses geometric implementation that uses finite elements method to minimize the forces. The next section presents the theoretical concepts upon which this approach is based.

3. SNAKES BASED ON VECTOR POTENTIAL FIELD

The problem in feature alignment is that two linear features deemed to be similar but are not in perfect correspondence after initial alignment are required to be brought to a perfect correspondence. One of the features is considered fixed (the reference) while the other one, in this case referred to as the snake is iteratively moved until it is aligned with the reference feature.

To implement the alignment within the vector domain, reference is made to the paradigm proposed by Honea (2002), in which the total force ($\overline{F}$) acting on a point (vertex) on the snake is given as:

$$\overline{F_i} = w_e \overline{E_i} + w_1 \overline{R_i} + w_2 d_k \overline{K}_i \qquad (6)$$

Where w are the weights, $\overline{E}_i$ is the external force, while $\overline{R}_i$ and $\overline{K}_i$ are the elastic and the bending components of the internal force respectively acting on a point on the snake.

3.1 External Force

Let $X = \{\bar{x}_i\}, i = 1, 2...n$ denote the vertices of the snake and $S = \{\bar{s}_i\}, i = 1, 2...m$ denote the vertices of the reference linear feature as illustrated in figure 1.

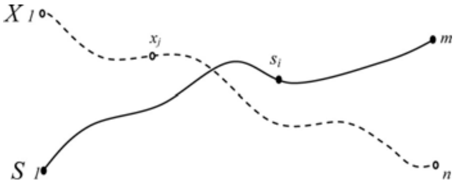


Figure 1: Snake and reference linear features

The net external force from the vertices of the reference feature on apoint (vertex) on the snake is then represented as:

$$\overline{E_i} = \sum_{j-1}^{m} E_{ij} \frac{\overline{s_i} - \overline{x_j}}{\left|\overline{s_i} - \overline{x_j}\right|} \qquad (7)$$

In which E_{ij} is the gravitational field potential of the vertex, s_i in the reference feature on the snake vertex, x_j and is given as:

$$\overline{E_{ij}} = \alpha \frac{e(\overline{s_j})}{d_{ij}^2} \qquad (8)$$

Where d_{ij} is the Euclidean (or other metric) distance from $\bar{x}_i$ to $\bar{s}_i$; e and α are constants and are taken to be unit.

3.2 Internal Force – Elastic Component

The internal force of the snake is composed of two parts like in the traditional case. If c_i denotes the snake point at x_i and let c_{ix} and c_{iy} be the x and y-coordinates of c_i, then the elastic force acting on a snake point from all other snake points is expressed as:

$$\overline{R_i} = (\sum\nolimits_{j \neq i} \frac{1}{c_{ix} - c_{jx}}, \frac{1}{c_{iy} - c_{jy}}) \qquad (9)$$

This force is responsible for maintaining the topology of the snake feature.

3.3 Internal Force – Bending Component

Let c_{i-1}, c_i and c_{i+1} be three consecutive snake points as illustrated in figure 2 and K be the unit vector normal to (c_{i-1}, c_{i+1}), with its magnitude proportional to the distance d_k.

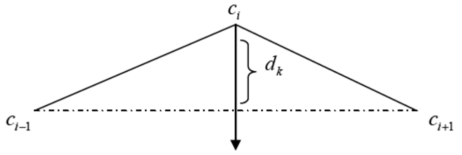


Figure 2: Internal Force (Bending Component)

3.4 Combining the Forces

Equation (6) shows the net force acting on a snake point. While the main issue in the equation is how to determine the weights, two conditions must however be fulfilled:

i) The length of the final snake should be the same or nearly the same as that of the reference snake (length condition)

ii) A perfect or near perfect alignment of the snake with the reference feature should be achieved (alignment condition).

Therefore the forces can be combined using any weights as long as these conditions are achieved. The next section presents an analysis of the effects of these forces based on an experimental snake and reference feature.

4 EXPERIMENTAL RESULTS

4.1 External Force

The external force is considered the most important force in the system because it is the one that will eventually influence the snake to move to the required position. Ideally, the snake should eventually be aligned with the feature that provides the external force. However, depending on the number of iterations (dynamism) and conditions specified for the system, the snake could behave in a number of ways, for instance, if the number of iterations is too small, the snake may not move to its final alignment with the reference feature, otherwise it will move beyond the required alignment position.

A linear feature consists of vertices that are normally not uniformly spaced, in which case the feature is considered to be unsegmented. If additional vertices are introduced at equal intervals the feature could be considered to be segmented. This operation is necessary in the implementation of the external force to ensure uniform pulling force from the reference feature. This is similar to setting the resolution of the image in the conventional snake as small as possible to ensure uniform external force.

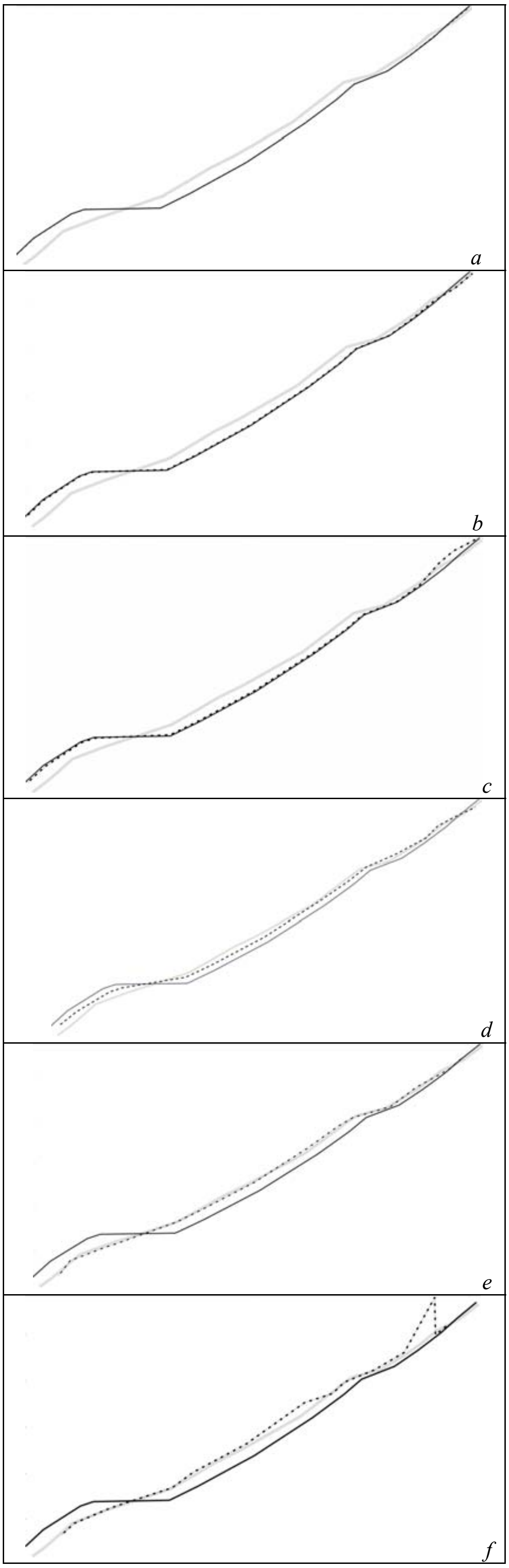


Figure 3: External Force after 2, 5 and 10 iterations for the unsegmented (*a,b,c*) and segmented (*d,e,f*) reference feature respectively

Figure 3 illustrates the aligned position of the snake (in dashed black line), the reference feature (bold gray line) and the initial snake position (black line) after applying only the external force. In the figure, a, b and c shows the result after applying the external force at 2, 5 and 10 iterations for unsegmented reference feature, while d, e and f are the results for a segmented reference feature for the same number of iterations. The length of the reference feature is approximately 470m and the segmentation distance is 1.0m.

It is noted that, for the unsegmented reference feature a lot more iterations are required to bring the features into alignment, however for segmented reference feature, alignment is achieved quite fast. The downside of segmenting the reference feature is that the snake shrinks as the number of iterations increases and unwanted kinks are introduced after certain number of iterations, as a result of possible infinitesimal distance between points on the snake and the reference feature.

The length of the example reference feature is 471.161m, while the original length of the snake feature is 478.528m. Generally, the snake feature shrinks with successive iterations. This is because the snake will tend to be pulled towards the centroid of the reference feature. The rate of shrinkage will depend on the relative forces of the snake and the corresponding reference feature. In particular, the force from the corresponding reference feature depends on the number of points (vertices) on the reference feature.

4.2 Internal Force – Elastic and Bending Components

Figure 4 illustrates the elasticity effect on the example snake, with bold gray line representing the initial snake position and the dotted black line representing the final snake position after 10 iterations. As the number of iterations increases, the snake reduces to a stretched and smoother line, which is almost a straight line. In other words, some vertices in the feature can be eliminated without compromising the structure of the snake feature.

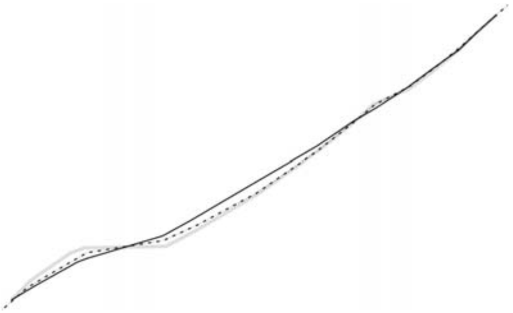


Figure 4: Initial snake position (in bold gray), final position after applying the elastic force (black line) and the bending force (dotted black line)

The bending force minimizes the curvature of the snake by pulling a vertex towards the line between its two neighbouring vertices. This force ensures that the successive structure of the snake points is maintained even when the external force may tend to distort that structure. If implemented separately, this force will result in a snake which has maintained its length since the end points do not experience any bending. In figure 4, the effect of the bending internal force on the example snake after 10 iterations is illustrated, with the dotted black line. The net effect of the bending force is the retaining of the initial character of the snake, and the effect of the force is equivalent to line simplification.

Since this force does not consider the reference feature and the snake points change their positions, the net effect is similar to a simplification of a linear feature. The extent of simplification depends on the number of iterations specified during execution. Overall, the internal forces ensure that the topologic structure of the snake is maintained.

4.3 Combined Forces

During the implementation, the best approach is to combine the individual force components and then compute the snake displacements iteratively. This is as opposed to iteratively computing the individual components separately and then combining them.

Combining the forces results in the net force acting on a snake point, this could be acceleration or displacements with instantaneous velocity depending on the nature of dynamism defined in the system. Dynamism is defined through the parameters – i.e., the weights and the number of iterations. Figure 5 (a) illustrates the snake position (in dotted black line) after 10 iterations, the reference feature (in bold gray) and the initial snake (in black) after combining the forces with equal weights.

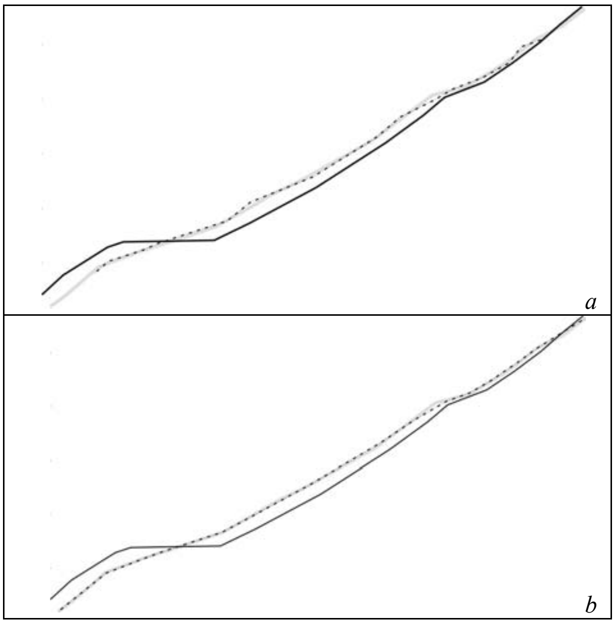


Figure 5: Aligned snake feature without conditions (a) and with conditions (b)

It can be observed in figure 5 (a) that even after combining the forces, the final snake has contracted. This is because equal weights were assumed for the forces. So a strategy to determine the weights is required so that the two conditions stated in section 3.3 are simultaneously achieved.

(i) Length condition

During the application of the forces, the elastic and bending force ensures that the length of the snake feature does not change significantly, while the external energy is one that affects the length of the snake as it evolves. Therefore the length condition can only be fulfilled by considering mainly the external energy. If left without including some constraint, the final length of the snake feature under the influence of the reference feature will increase or decrease, for example as illustrated in figure 5 (a). This is because the end points of the snake feature are not only affected by the terminal points of the reference feature but also by the intermediate points (vertices) of the reference feature. The effect of the intermediate points on the terminal points of the snake can be cancelled by fixing the coordinates of the terminal points of the snakes to the terminal points of the reference feature.

(ii) Alignment condition

Once the terminal points are fixed, the alignment condition has to be fulfilled, which requires the vertices of the snake to be aligned with the reference feature. In this case the relative importance of the forces should be considered. Since the ultimate objective is to achieve a near perfect alignment, the external force carries more weight compared to the elastic and bending forces. The elastic force is partly considered in the length condition, however, the intermediate snake vertices need to be spaced appropriately with regard to the initial snake configuration. In this case a way to determine the weights has to be established as described in the next paragraph.

(iii) Weights

To fulfill the alignment condition, an appropriate strategy for determining the relative weights is required. Some of the strategies that can be considered include the following: (i) assigning equal weights or (ii) assigning relative weights.

Assignment of equal weights for the forces has been adopted in this implementation and figure 5 (b) illustrates the finally aligned snake.

One of the implied conditions is to maintain the structure of the snake as much as possible. This condition therefore does not allow for the segmentation of the snake, because additional vertices would be introduced. Since no additional vertices would be introduced, there won't be a perfect alignment of the snake with the reference feature; at least some sections will be misaligned.

4.4 Evaluation of the Quality of the Alignment

The quality of the alignment can be evaluated in terms of the ratios of the curve lengths and the maximum displacements between the snake and the reference feature. For positional misalignment, the classical Hausdorff distance (Hangouet, 1995) can be used as a quality measure. Whereas the ratio of the curve lengths is a good parameter for quality evaluation, the Hausdorff distance is a better indicator for the alignment quality because it indicates by how much the positional error of the snake feature has been improved, in case the objective of the alignment was to enhance the positional accuracy.

Table 1 shows for the values for the curve length ratios and the Hausdorff distances after the alignment for various cases. A value of 1.0 for the ratio of the curve lengths of the snake and reference feature would indicate a perfect alignment. Unless there is a perfect alignment, this value would usually vary and the value of the variation from 1.0 is proportional to the quality of the alignment. The initial snake to reference feature length ratio and the Hausdorff distance were 1.0156 and 13.77m respectively.

Calculating the Hausdorff distance for the final case when the conditions are fulfilled is straight forward; however for the other cases this is not trivial. This is because the length of the snake feature changes significantly, further, the distance evaluated for the unsegemented cases may be misleading, because the number of vertices in the reference feature are few and not uniformly distributed.

Case	*Snake to Reference Length Ratio*	*Hausdorff Distance (m)*
Figure 3 (a)	1.0136	19.46
Figure 3 (b)	1.0125	19.35
Figure 3 (c)	1.0119	19.17
Figure 3 (d)	0.9641	7.56
Figure 3 (e)	0.9132	2.75
Figure 3 (f)	1.0021	30.91
Figure 4 – Elastic force	1.0197	13.21
Figure 4 – Bending force	0.9921	12.04
Figure 5 (a)	0.8210	7.34
Figure 5 (b)	0.9955	1.28

Table 1: Evaluation of Feature Alignment

In the experiment the accuracy of the reference feature is considered to be higher than that of the snake feature. A better evaluation of the quality of the alignment would be achieved if the relative accuracies of the features is known before the alignment process.

5 CONCLUSIONS AND FURTHER WORK

A segmented reference feature provides a fast alignment, but after a certain number of iterations, the snake starts to shrink and even unwanted kinks are introduced. This is overcome by fulfilling the length and alignment conditions as described here. Adaptive determination of the number of iterations is part of further being considered.

In general, the approach presented here could be useful for feature alignment applications that do not necessarily require the snake features to be closed and where the accuracy of the snake feature is so low that using buffers as the basis for alignment may not be sufficient. From the example illustrated here the positional accuracy of the snake feature was improved from 13.77m to 1.28m. Although this is a general indicator for the quality of the alignment, considerations for the relative accuracies of the features will constitute a better quality assessment. A further investigation will involve establishing whether the distance used to segment the reference feature affects the overall result, particularly the Hausdorff distance.

Although the assignment of equal weights to the forces seemed sufficient, it would be interesting to establish a way to determine the relative weights for the forces as possible further work.

The approach discussed herein was based on the assumption that the features to be aligned are already matched to their corresponding reference features. This assumption keeps the snake feature from the effects of the external forces of other reference features. Further work will also investigate, if different matching candidates can be evaluated in an integrated way using their attractive forces. In this way, reference features which are closer will have a higher influence than features that are further away. This would allow to start the adaptation without prior matching.

REFERENCES

Bader, M. 2001. Energy Minimization Methods for Feature Displacement in Map Generalization. PhD thesis, University of Zurich, Switzerland.

Besl, P. and McKay, N., 1992. A Method for Registration of 3-D Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence (Special issue on interpretation of 3-D scenes – part II) 14(2), pp. 239 – 256.

Burghardt, D. and Meier, S., 1997. Cartographic Displacement Using the Snakes Concept. In W. Förstner and L. Plümer (Eds), Semantic Modeling for the Acquistion of Topographic Information from Images and Maps. Birkhäuser-Verlag: Basel.

Burghardt, D. 2005. Controlled Line Smoothing by Snakes. GeoInformatica 9(3), pp. 237 – 252.

Butenuth, M., 2008. Topology-Preserving Network Snakes. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B3a. Beijing, pp 229 – 234.

Caselles, V., Catte, F, Coll, T. and Dibos, F., 1993. A geometric model for active contours. In: Numerical Mathematics Vol. 66, pp. 1 -31.

Hangouet, J.F., 1995. Computation of the Hausdorff distance between plane vector polylines, Twelfth International Symposium on Computer-Assisted Cartography, Charlotte, North Carolina.

Honea, D., Synder, W. and Bilbro, G., 2002. Active Contours Using a Potential Field. 16th International Conference on Pattern Recognition (ICPR'02) - Volume 2, pp. 757 – 760.

Kass, M., Witkin, A. and Terzopoulos, D., 1988. Snakes: Active Contour Models. International Journal of Computer Vision 1(4), pp. 321 – 331.

Lam, K.-M. and Yan, H., 1994. Fast Greedy Algorithm for Active Contours. Electronics Letters. Vol. 30(1), pp. 21 – 23.

Siriba, D. 2009. Positional Accuracy Assessment of a Cadastral Dataset based on the Knowledge of the Process Steps used, Proceedings of 12th AGILE Conference on GIScience. Hannover, Germany.

Song, W., Haithcoat, L. and Keller, J.M., 2006. A Snake-Based Approach for TIGER Road Data Conflation. Cartography and Geoinformation Science, Vol. 33, No. 4, pp. 287 – 298.

Xu, C., and Prince, J., 1998. Snakes, Shapes and Gradient Vector Flow. IEEE Transactions on Image Processing Vol. 7 No. 3., pp. 359 – 369.

ACKNOWLEDGEMENT

The comments made by our colleague at the institute, Stefan Werder towards improving this paper are greatly acknowledged.

RESEARCH ON LANDSLIDE PREDICTION MODEL BASED ON SUPPORT VECTOR MODEL

Xiaowen Zhao [a,]*, Min Ji [a], Xianguo Cui [a]

[a] Geomatics College, Shandong University of Science and Technology, 579 Qianwangang Road, Economic & Technical Development Zone, Qingdao, China,266510,wen1987zi@163.com, jimin@sdust.edu.cn, cxg824@163.com

KEY WORDS: mine landslide, SVM, prediction model, GIS, LIBSVM, cross validation, grid search

ABSTRACT:

The Landslide，which is caused by mining activities, has become an important factor which constrains the sustainable development of mining area. Thus it becomes very important to predict the landslide in order to reduce and even to avoid the loss in hazards. The paper is to address the landslide prediction problem in the environment of GIS by establishing the landslide prediction model based on SVM (support vector machine). Through differentiating the stability, it achieves the prediction of the landslide hazard. In the process of modelling, the impact factors of the landslide are analyzed with the spatial analysis function of GIS. Since the model parameters are determined by cross validation and grid search, and the sample data are trained by LIBSVM, traditional support vector machine will be optimized, and its stability and accuracy will be greatly increased. This gives a strong support to the avoidance and reduction of the hazard in mining area.

1. INTRODUCTION

The increasing demand for coal of the industrial society has led to more and more serious coal mining. The Mining-induced landslide hazard has seriously influenced the sustainable development of mining areas. So it is important for us to select a suitable method to predict mine slope stability. However, there are many influencing factors for landslide and the effects of the same factors in various areas are different. The mathematical relationship between the factors which impact landslide and the landslide stability prediction is hard to obtain. Therefore, it is a comparatively accurate method to get a statistical analysis model with the historical data. SVM can get solution by solving a convex quadratic programming question. The solution is global optimal solution, and its ratio is high. The Prediction with SVM will use structural risk minimization principle instead of the empirical risk minimization principle, maximize the generalization ability of learning machine, make sure that the independent test set which was gotten from a limited sample of training set remains a small error, and get a non-linear mathematical relationship with a higher dimension at the same time. In this paper, we identified the complex relationship between influencing factors and stability prediction of landslide by training samples with SVM and predicted results of unknown data with the relationship. It was proved that mine landslide prediction based on SVM got satisfactory result, and it was a prediction model with a high accuracy and stability.

2. SVM (SUPPORT VECTOR MACHINE)

SVM was a new general-purpose machine learning method based on statistical learning theory, and it was built under the theory framework and general approach of machine learning with limited samples. Its basic thought was to transform the input space to a high-dimension one by using the non-linear transformation defined by inner product function and to find a non-linear relationship between the input variables and the output ones in the high-dimension. SVM had a better generalization than neural network which used empirical risk minimization principle. (Hsu, 2009)

2.1 The generalized optimal separating hyper plane

SVM developed from the optimal separating hyper plane in a linear condition. To make it clear, we started from the two-dimension situations.

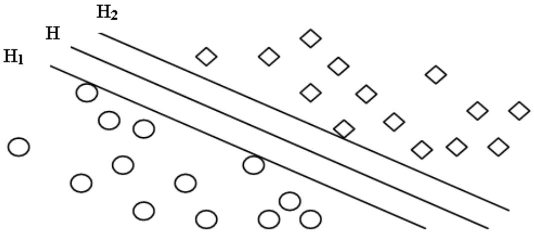


Figure 1. Separating hyper plane

As Figure 1 showed, set circular and diamond graphics as two kinds of samples, the straight line H as category line, H_1 and H_2 as straight lines which lines the samples nearest to category line were on and paralleled to the category line, the distance between the two lines called class interval.(Li X.Z.,2009; Zhao H.B., 2008). The optimal separating line asked for correct separation of the two kinds of samples and the largest class interval. The general equation could be formula (1):

$$x \cdot \omega + b = 0 \qquad (1)$$

x - the samples data
w, b – parameter for the line

* Corresponding author. E-mail address: wen1987zi@163.com;

Formula (1) could be normalized. We needed to make linear separable sample set (2) satisfy formula (3):

$$(x_i, y_i)\,(\mathrm{i}=1,\cdots,\mathrm{n}, \mathrm{x}\in \mathrm{R}^{\mathrm{d}}, \mathrm{y}\in\{+1,-1\}) \qquad (2)$$

$$y_i[(\omega\cdot x_i+b)]-1\geq 0 (i=1,2,\cdots,n) \qquad (3)$$

x_i, y_i - the sample point data

At this time, the classification interval was $\frac{2}{\|\omega\|}$. The maximum of classification interval equalled the minimum $\|\omega\|^2$. So the hyper plane which satisfied the formula (3) and made the $\|\omega\|^2$ get the minimum was the optimal separating hyper plane. The sample points on two straight lines (H_1 and H_2) were support vectors. (Jiang Q. W., 2005)

We used Lagrange method to transform the hyper plane question to a dual one, subjected to:

$$\sum_{i=1}^{n} y_i\alpha_i = 0 (\alpha_i\geq 0, i=1,2\cdots,n) \qquad (4)$$

α_i - Lagrange multiplier parameter

And Looked for the maximum value with the parameter α_i from the function $Q(\alpha)$, its expression was:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i\cdot x_j) \qquad (5)$$

Parameters α_i were Lagrange multipliers corresponding to every parameter, so function $Q(\alpha)$ was a question which asked for the best solution in the quadratic function under inequality constraint, there was a unique solution. There were few non-zero α_i, so the samples which were corresponding to these α_i were support vectors. The optimal separating function $f(x)$ could be obtained from the questions above:

$$f(x)=sign\{(\omega\cdot x)-b\}=sign\{\sum_{i=1}^{n}\alpha_i y_i (x_i\cdot x)-b^*\} \qquad (6)$$

The summation in formula(6)just contained the support vectors in fact, and b^* were classification threshold which could be obtained by any of the support vectors (vectors which satisfies the equation in formula(3)) or the mid-value of any two support vectors from the two kinds of samples.(Dong J. X., 2003)

2.2 Non-linear SVM

In non-linear condition, added a relaxation $\xi_i \geq 0$ to formula (3), then it changed to (7):

$$y_i[(\omega\cdot x_i+b)]-1+\xi_i\geq 0 (i=1,2,\cdots,n) \qquad (7)$$

The goal changed to be the minimum value of formula (8):

$$(\omega,\xi)=\frac{1}{2}\|\omega\|^2+C(\sum_{i=1}^{n}\xi_i) \qquad (8)$$

That was equally to the generalized optimal separating hyper plane considering of the least wrongly classified sample and the maximum separating interval. C > 0 was a constant and the penalty parameter of the error term. The dual problem of optimal separating hyper plane in non-linear situation was almost the same as the linear ones. The multipliers α_i objected to:

$$0\leq\alpha_i\leq C(i=1,2,\cdots,n) \qquad (9)$$

The method which SVM used to construct separating decision function in non-linear condition contained two steps. First step was translating training data from raw mode to high dimension space by non-linear transformation of special kernel function. Second step was looking for optimal separating hyper plane in feature space. The hyper plane was corresponding to non-linear separating surface in the raw mode. So, there was only one more mapping link in non-linear condition than the linear ones when using SVM. We supposed the non-linear mapping to be $x\rightarrow\varphi(x)$, the function $Q(\alpha)$ changed to:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i\cdot x_j) \qquad (10)$$

$K(x_i\cdot x_j)=\varphi(x)\cdot\varphi(x)$ - (kernel parameter)

The separating decision function in non-linear SVM changed to:

$$f(x)=sign\{(\omega\cdot x)-b\}=sign\{\sum_{i=1}^{n}\alpha_i y_i K(x_i\cdot x)-b^*\} \qquad (11)$$

The kernel functions $K(x_i\cdot x_j)$ in formula (11) were in accordance with the mercer condition, and in corresponding to the inner product in the transformation space. (Gallus D.)
In the choice of kernel function, there were three options:

(1)Multinomial kernel: $K(x \cdot x_i) = [(x_i \cdot x_j) + 1]^d$ (12)

(2) RBF: $K(x \cdot x_i) = \exp\{-\frac{|x_i - x_j|^2}{\sigma^2}\}$ (13)

(3) Sigmoid function: $K(x \cdot x_i) = \tanh[v(x_i \cdot x_j) + a]$ (14)

3. THE APPLICATION OF SVM IN PREDICTION FOR LANDSLIDE IN MINING AREA

The prediction of the landslide by using the SUV contained the prediction on time-series and prediction on space. We used prediction on space which depended on identification of the stability of side slope. We chose on proper influencing factors for stability and picked up these factors to construct distinguishing model. We used training of samples data in the mining area to get certain discriminant function of side slope stability and then used the function to get decision outcomes of unknown sample data and the stability outcome of the sample points.

3.1 Determination of influencing factors for stability of landslide in mining area

Because of the special environment in mining area, influencing factors were more complicated than normal areas. In normal areas, the factors such as altitude, slope, aspect, vegetation coverage, litho logical character, geologic structures and so on appreciably affected the occurrence of landslide. From the survey on the study of mining area we found that the most appreciably influencing factors contained litho logical character, geologic structures, thickness of overlying strata, precipitation, precipitation intensity, slope, aspect, slope mining conditions. The data of Litho logical character, geologic structures, slope, aspect could be achieved from remote sensing image interpretation, thickness of overlying strata from field reconnaissance, precipitation and precipitation intensity from updating information on the web, slope mining conditions could be grated after field Reconnaissance.

3.2 Acquisition of sample points for prediction of landslide

In order to get enough sample data with an acceptable quality, we needed to choose suitable sample points in mining area, and get the influencing factors on each sample point. We found that it will not be enough if we only used the landslide points in mining area to be the historical data for training, so we needed to pick up certain numbers points besides centre points.

We set certain principle when picked up points besides centre points in order to make sure the samples were well distributed so that they would not affect the accuracy and stability. When the landslide area were less than 4 pixel areas, we picked up 1 point; when 4-5 pixel areas, we picked up 4 points; when 5-9 pixel areas, we picked up 5 points; according to landslide area, we may choose 1, 4, 5, 9, 16, 25 points on one landslide. We added sample points on the basis of primary points, for the primary points were certain, so we could raise the number of training samples and ensure the accuracy and stability of training model. (Ma Z. J., 2003)

3.3 Landslide database

From the analysis of database we knew that landslide database needed to contain infrastructure data layer and landslide disaster layer. The infrastructure layer involved spatial data and properties data. Non-numeric data in property layer needed to be transformed to numeric ones which were easy to use. The transformation was with certain principles. The fields in the infrastructure data list were built according to the rules in related stipulate. The graphic layer contained thematic map of each influencing factor, topographic map and so on. Landslide disaster data was the most important part in the database. There were both property and spatial data of landslide point and area. We used GIS technology to establish property database and spatial database because there were both kinds of data. The structured chart was showed in Figure.2.

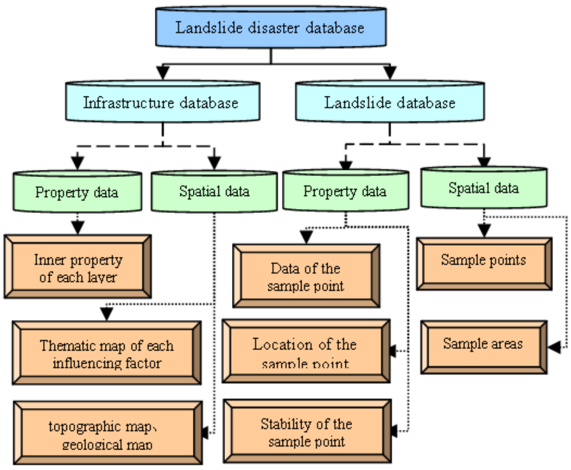

Landslide disaster database
Infrastructure database
Landslide database
Property data
Spatial data
Property data
Spatial data
Inner property of each layer
Data of the sample point
Sample points
Thematic map of each influencing factor
Location of the sample point
Sample areas
topographic map、geological map
Stability of the sample point

Figure 2. Database structure

3.4 Selection of parameters in landslide prediction model

We used training samples and testing samples to do the experiment. When we trained the model, fitting accuracy of testing samples was different with different input parameters in prediction model. In order to find the best fitting function, we did lots of parameter tests. The best fitting function required suitable parameters, C , σ .

We used grid search method to choose parameters and estimate the generalization of each set of parameters. Grid search method gave different numeric values to M penalty parameters for C and N kernel parameters for σ , constructed M*N combinations for different SVM models. We estimated the generalization of parameters to choose grid points of (C, σ) with the highest generalization.

The method to determine parameter based on grid search involved training of SVM and comparison of prediction accuracy. To the same grid point of (C, σ) , different training method would get discriminant function with different prediction accuracy. To n samples, picked up n-1 samples to train prediction model and got expected value of error rate. Then we used the value to determine capability of the grid point. We provided a possible interval of C (or σ) with the grid space. Then, all grid points of (C, σ) were tried to see which one gave the highest cross validation accuracy. We picked up the best parameters to train the whole training set and generate the final model.

The implementation was as follows: we chose a range for grid points of (C,σ), for example, $\sigma = 2^{10} \sim 2^{-15}$, step-size in research was -1 and $C = 2^{-10} \sim 2^{15}$, step-size was 1, then two-dimension grids of C and σ were structured on the coordinate system. We got the accuracy of each grid point, drew contour line of the accuracy to determine the best grid point. If these points couldn't get the requested accuracy, we reduced the step-size to do ransack.

The advantage of grid research was searching 2 parameters at the time and we could finally get the best grid point to make the accuracy get the highest level. In computational process, grid points could decouple between each other in order to do parallel computation and get a high efficiency. Parallel research cut down the time for searching the best parameter value, but when the samples were large in numbers, repeating the process one by one would also cost lots of time. Then we used cross validation via to do approximate evaluation. That was to divide samples into K sets, pick up K-1 set to be training set and get a decision function, then use the function to predict the testing set. This would reduce the frequency of training and ensure the accuracy for prediction, and it was called K-fold cross-validation. (Wang X. L., Li Z. B., 2005)

3.5 Application of prediction model

The Process of prediction in mining area by using the SVM contains acquisition of training samples, selection of SVM kernel function and training software, generation of SVM prediction model and decision outcomes of unknown data. The process was showed in the Figure.3.

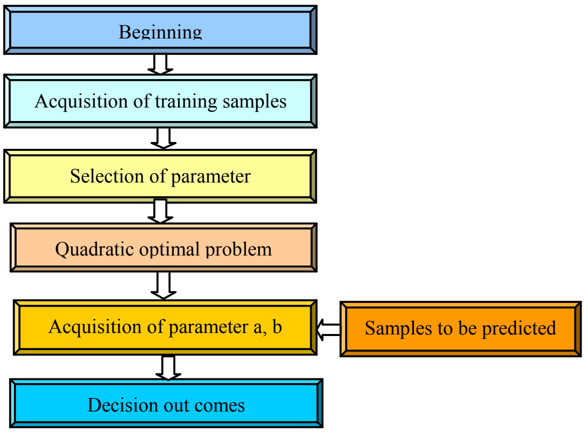
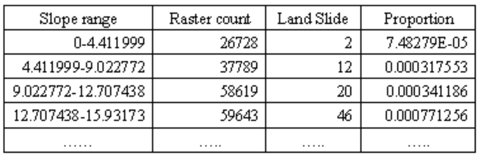

Beginning
Acquisition of training samples
Selection of parameter
Quadratic optimal problem
Acquisition of parameter a, b
Samples to be predicted
Decision out comes

Figure 3. predicting process

3.5.1 Collection and pre-processing of materials: Before the application of SVM in landslide, we gathered materials which were to use for research. After the determination of influencing factors, we did pre-processing to the samples at first. According to assessment objective and the characteristics of the regional environment, we determined the grading standard for each factor. In order to make it easy for us to express prediction outcome in the classification diagram in GIS software on the basis of statistic analysis with the regional survey data, we found out the highest and lowest value to determine ranges for each factor data. Then we divided influencing factor value into grades according to divided principle and translated the influencing factor into quantitative value taken in accordance with actual situation.



Slope range	Raster count	Land Slide	Proportion
0-4.411999	26728	2	7.48279E-05
4.411999-9.022772	37789	12	0.000317553
9.022772-12.707438	58619	20	0.000341186
12.707438-15.93173	59643	46	0.000771256
......	.....	.....	.....

Figure 4. slope range

3.5.2 Grid unit division: Whether the division was applicable would affect reasonableness of the assessment outcomes and influence complexity of parameter acquisition in prediction process. When the unit ware was defined, we could take every evaluation unit as an independent individual. Usually, two ways were prepared for the division, regular division and irregular division. We used regular division to divide the study area into N grid units. Took each unit to be a point and extracted data from the thematic map using weighted average method.

3.5.3 Selection of kernel function: The study showed that types of kernel had little to do with the capability of prediction model, the important ones were kernel parameters and error penalty parameter. We chose RBF function as the kernel function in designing prediction model. Because there were only two parameters when using RBF function, this would keep the stability of function. (Dong H., 2007)

3.5.4 Selection of the training software: We verified from lots of literatures that using LIBSVM to train sample data would get better result than others. LIBSVM was a library for support vector machines (SVM). Its goal was to make users use SVM as a tool easily. We prepared data in special format for LIBSVM and referred to the file "*heart scale*" which is bundled in official LIBSVM source archive. The input file format is as follows.

[Label1] [index1]: [value1] [index2]: [value2]...
[Label2] [index1]: [value1] [index2]: [value2]...
(http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

3.5.5 Result analysis: From above tasks, we got training samples and testing samples, the factors charts of study field. Then we used training samples to do the test. We inputted different parameter in training process and did comparison on the fitting degree by testing samples. We constructed prediction models with the best parameters obtained by the above comparison, then used the model to train the testing samples and picked up support vectors in the samples set to get the discriminant function. We used LIBSVM to train samples, and then got the prediction model to estimate the unknown grid units divided according to the study field. The outcome was expressed in GIS software by showing in the map. In order to verify the method, we used another method to do the test on the accuracy with the help of the data that we've got. The similar ways contained logistic regression model, stability factor prediction and so on. We put the landslide layer in to do a test and found that SVM prediction model got a higher accuracy. (Wang H. W., 2007)

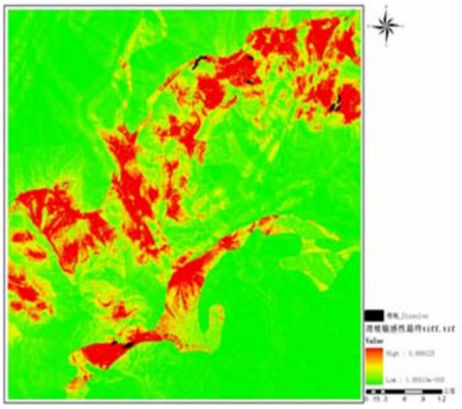
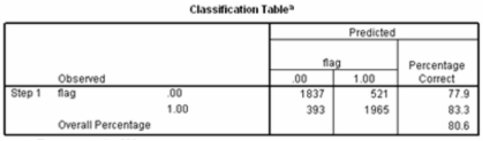


Figure 5._result chart



Classification Table^a

			Predicted		
			flag		Percentage Correct
Observed			.00	1.00	
Step 1	flag	.00	1837	521	77.9
		1.00	393	1965	83.3
	Overall Percentage				80.6

Figure 6._accuracy list

4. CONCLUSION

The Prediction of landslide was related to several influencing factors and there were interactions in factors. It was hard to use traditional mathematical analyzing method to get a certain linear relationship between prediction outcome and influencing factors. SVM avoided these questions by getting non-linear relationship using historical data. SVM got result from solution of convex quadratic programming problem without numerous samples, and the solution was global optimal solution with a high accuracy. Small sample size will take a great advantage in mining area where sample data was hard to get. Using function which was obtained with sample data would make each possible influencing factor be considered. From principal component analysis we could determine the primary factors and it also took the interactions in factors into account. The prediction model obtained by training known landslide point data and stable point data was possessed of a high capability and accuracy. Display in GIS software was useful for expressing the outcome visually intuitive and further analysis to outcome.

The Application of landslide prediction based on SVM in mining area is always in exploration, and there are still many things which need to be researched further. For example, the initial selection of kernel parameter and penal parameter need to be directed much better in order to save the time on finding the best parameters.

REFERENCES

Dong H., 2007, Prediction of landslide displacement based on Takens theory and SVM, China Journal of Highway and Transport, PP,13-18

Dong J. X., Krzyzak A., 2003, A Fast Parallel Optimization for Training Support Vector Machine, Canada

Gallus D., Abecker A., Classification of Landslide Susceptibility in the Development of Early Warning Systems

Hsu C.W., Chang C.C., 2009, Department of Computer Science, A Practical Guide to Support Vector Classification, Taiwan

Jiang Q. W., 2005, ArcGIS-based assessment of regional landslide hazard, China

Li X. Z., Zhao Z. Y., 2009, Basis and application prospects of Support Vector Machine, SCIENCE&TECHNOLOGY INFORMATION, China, PP 431,461

Ma Z. J., 2003, Prediction of landslide based on Support Vector Machine theory, Journal of Zhejiang University (Science Edition), China, PP, 592-596

Wang H. W., 2007 No.4, Predictive modelling on multivariate linear regression, Journal of Beijing University of Aeronautics and Astronautics, China

Wang W., Zhang N., 2008, General approach of solving Convex constrained quadratic programming problem, Journal of Hainan Normal University（Natural Science）, China

Wang X. L., Li Z. B., 2005, Grid-based search of the Support Vector Machine Kernel Parameters, PERIODICAL OF OCEAN UNIVERSITY OF CHINA, China

Wu J., 2006, *Localized Support Vector Machines For Classification*

Zhao H. B., 2008, *Analysis of Support Vector Machine in Rock and Soil Mechanics and Engineering*, PP, 19-31

LIBSVM -- A Library for Support Vector Machines，Chang Chih-Chung and Lin Chih-Jen,

http://www.csie.ntu.edu.tw/~cjlin/libsvm/

APPENDIX

Supported By Chinese National Programs for High Technology Research and Develop-ment(2009AA12Z147);
Supported By the Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping(2009B14);
Supported By Key Laboratory of Advanced Engineering Surveying of SBSM(TJES0805);

PRELIMINIARY INVESTIGATION OF WEB GIS TRUST: THE EXAMPLE OF THE "WIYBY" WEBSITE

A. Skarlatidou [a,], M. Haklay [a], T. Cheng [a]

[a,] Dept. of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT UK – a.skarlatidou @ ucl.ac.uk

KEY WORDS: Web-Mapping, Usability, Trust, Trust-Cues, Human-Computer Interaction, Guidelines

ABSTRACT:

Public access to environmental information is now a common requirement by national, international and European Union legislation. It is widely recognized that web-based GIS can enhance access to environmental information and can support public participation in environmental decision-making. Yet when these systems are used by non-experts might be challenging because of the GIS complexity. Considerations about data accuracy and errors during the analysis further increase the elements of risk, complexity and uncertainty, which are preconditions of trust. Many lay users are partially aware of the technicalities related to spatial data handling. Thus, the issue of trust in such systems, and how user's trust is built is an important consideration. Online trust has been repeatedly identified as a major concept for online information systems and its value recognised, especially in the context of e-commerce, as it influences the intentions to engage, the use and acceptance of online systems and the overall User Experience. However, there is very limited, if at all, knowledge about how trust is constructed in web-mapping systems. To improve knowledge in this domain, this paper describes the concept of online trust and its characteristics and models developed in different fields. The UK Environment Agency 'What's In You Back Yard' (WIYBY) website is examined using techniques derived from the Human-Computer Interaction field. A Heuristic Evaluation and a Cognitive Walkthrough were undertaken by three evaluators, to identify what influences trust and how perceived trustworthiness can be enhanced through interface design. Trust cues suggested in the literature were also considered for their applicability and relevance in web-mapping. Based on the findings a set of guidelines is presented which covers the dimensions of graphic, content, structure, map functionality and trust-cue design.

1. INTRODUCTION

Existing Web GIS applications instruct, advise users and provide information and analysis, which according to Fogg (2003) are amongst these situations where computers' credibility matters. Furthermore, Web GIS incorporate the element of risk especially when they are used in domains such as environmental decision-making, a popular domain especially for web-based Public Participation GIS (PPGIS). Although uncertainty is an inherent characteristic of geographical data, the uncertainty in Web GIS is further increased because of the complexity of these interfaces and the fact that mostly used by non-experts (Unwin, 2005; Haklay and Zafiri, 2008). For all these reasons, trust in the context of Web GIS establishes an important area of research which, no one to our knowledge yet considered.

There is not a commonly agreed definition for trust. Trust was examined in different disciplines, with each approaching trust from a different perspective and the result was *"a confusing potpourri of definitions"* (Shapiro, 1987, p.625). Another aspect of trust, which challenges scientists to agree on a common definition (Wang and Emurian, 2005) is that trust encapsulates different meanings (Williamson, 1993), as for example, credibility, reliability, honesty and confidence. Despite the lack of a commonly agreed definition, trust researchers agree on specific trust components which can help conceptualize trust.

Online trust can be defined as a trustor's willingness to depend or rely on a trustee, which can be an online system or online information (Chopra and Wallace, 2003). In the context of online trust the following components are of particular importance: a specific context; the preconditions of dependence, uncertainty and risk; the trustor's confidence that trust will be upheld and the willingness to act on that confidence; the factors, which influence trustor's trust perceptions (e.g. propensity to trust); the dimensions of trust (cognitive and affective trust); and the trustee attributes (Chopra and Wallace, 2003).

Online trust is a well-researched area from a Human-Computer Interaction (HCI) perspective and particularly for the e-commerce domain (Riegelsberger et al., 2005). Existing studies suggest that people's trust perceptions about electronic online environments, influence the intentions to engage, the use and acceptance of these systems, enhance cooperative behaviours and influence the User Experience (UXP) (Shneiderman, 2000; Egger, 2001; Fogg, 2003). Several studies focus on the online trustee attributes as these influence the people's perceptions about the trustworthiness of online systems and thus it is suggested that a trust oriented interface design which emphasizes on the improvement of these attributes can subsequently enhance the perceived trustworthiness.

As it is unknown what influences public trust in Web GIS, the wider research framework that this study follows is based on an investigation of different interfaces with non-expert users using HCI methodology in order to understand how trust perceptions are formed. In particular the aim of this study is to identify the interface elements and functionality attributes which influence the trustworthiness of Web GIS and to subsequently build a set of preliminary trust-based guidelines which can eventually improve the trustworthiness of these systems.

2. BACKGROUND

Chopra and Wallace (2003) think that for a system to be trustworthy it should be competent, predictable, and ethical and should have positive intentions. Grabner-Krauter et al. (2006) suggest two trust dimensions, a soft dimension which encapsulates attributes such as benevolence, honesty, integrity and credibility and a hard dimension which is referred to the system's functionality and encapsulates attributes such as reliability, correctness, and availability and so on. McKnight et al. (2002) believes that the trustworthiness of a system is influenced by the system's reputation (competence, benevolence and integrity) and the system's quality (functionality and aesthetics). In a similar perspective, Fogg (2003), who uses the term credibility instead of trust, believes that the credibility of a system is influenced by its reputation, integrity and expertise. Finally, according to Corritore et al. (2003) the ease of use (usability) affects the perceived credibility and risk about the system, which in turn influences its trustworthiness.

It can be concluded that two categories of attributes are of particular importance in the discussions about the online trustee attributes. The first category involves perceptual attributes which refer mainly to the source's reputation (e.g. is the source perceived as honest and reliable?). The second category involves attributes which refer to the system's functionality and overall quality. In this category aesthetics, professional look and feel and usability are of critical importance.

Shneiderman (2000) highlights that a good design emphasizing on clear commitments and usability can improve trustworthiness. Wang and Emurian (2005) amongst others suggest the use of pastel and cool tones and colour combinations and the use of high quality photographs. The design quality is also mentioned in the Cheskin Research Report (1999) but also in several other studies (Nikander and Karvonen, 2000). Karvonen (2000) links aesthetics to trustworthiness and in particular focuses on how the beauty of simplicity (clean and clear design) influences usability and affects online trust (affect-based trust). Fogg (2003) also provides a detailed list of elements which increase the perceived trustworthiness of a system (e.g. external non-broken links).

Trust-inducing features are also widely recognized as interface elements which can further improve trustworthiness. For example, several studies emphasize on elements such as seals of approval (Cheskin Research Report ,1999), branding and logos (Cheskin Research Report,1999; Shneiderman,2000), feedback mechanisms (Ba and Pavlou, 2002), external links, citations and contact details (Fogg, 2003), photographs, videos and chats (Wang and Emurian, 2005), which influence positively the formation of trust perceptions.

The majority, if not all, of the previously cited studies focus on e-commerce environments while Web GIS have their own special characteristics. There are several studies in the GIS literature which investigated user aspects of mainly offline environments or stand-alone web-based applications. These studies investigate amongst others, geovisualisation barriers (e.g. Ishikawa et al., 2005), different map functionality interaction options and cognitive aspects of users (e.g. Hornbaek et al., 2002; Fabrikant, 2001; Harrower and Sheesley, 2005). Issues such as the usability of Web GIS applications has only recently considered with Skarlatidou and Haklay (2006) who published the first, to our knowledge, study investigating how public web-mapping sites are used by novices.

Concerning the usability element of Web GIS, existing studies highlight that the end users and especially non-experts have significant problems while interacting with these systems (Skarlatidou and Haklay, 2006). Nivala et al. (2008) performed a usability evaluation of public web-mapping sites (Google Maps, MSN Maps & Directions, MapQuest and Multimap) where they found 403 usability problems and they provide the first list of usability guidelines for the design of similar applications. It is not surprising that some GIS research studies emphasize on the importance of User-Centered Design (UCD). For example, Kramers (2008) described the benefits of a UCD approach in order to overcome difficulties that the non-expert face when tools are purely based on technology-driven designs. The significance of a UCD approach is also acknowledged by Van Elzakker (2005) in his study about maps' usability.

Although usability is linked to trustworthiness, it was never considered as for its relevance to trustworthiness and thus it is not clear what usability problems influence trust. Usability is yet only one of the attributes which influence trustworthiness. Other elements that should be considered in the context of Web GIS, involve amongst others the content of these websites, if the information provided satisfy the user needs, and the aesthetics and functionality of the GIS element. Also, trust cues should be explored separately in order to understand how trust can be induced in this context according to the trust-based needs and expectations of the end users.

3. METHODOLOGY

The Web GIS selected to be evaluated as for its trustworthiness is the "What's In Your Back Yard" (WIYBY) website provided by the Environment Agency. The WIYBY website provides environmental information to the UK public (e.g. about air pollution, water quality, and risk of flooding, waste sites) and it was selected for different reasons. First, as it is anticipated by national, European and international legislation, the public should have access to environmental information and several studies suggest that GIS can be used to enhance public access and participation, exactly because it changes the ways that people can interact and communicate with maps, and can support visual thinking in the decision-making process (MacEachren, 1994; MacEachren and Kraak, 1997; Sieber, 2006; Dunn, 2007). In this context the WIYBY website serves this purpose, however before the people rely on the system and make a decision based on the information that it provides (e.g. where to buy a house based on flood occurrences), they should trust it.

Another reason for selecting the WIYBY website is that the elements of uncertainty and risk, which are necessary preconditions for examining trust, are existent. Environmental problems involve conflicting views and ethical considerations and are usually ill-defined, which increases the element of risk. Also, as Haklay (2002) suggests, accuracy and uncertainty are internal to environmental information and errors through its analysis are always existent. The fact that non-experts users have a limited knowledge about spatial data handling, GIS operations and expertise in this domain creates additional trust concerns. Finally, a previous usability study of the WIYBY website revealed that end users had significant interaction problems (Alsop, 2008), which further increases the complexity, while it is acknowledged that in complex situations people develop mental shortcuts one of which is trust (Grabner-Krauter and Kaluscha, 2003).

In order to identify what elements of the system influence its trustworthiness, and how the quality and the usability are linked to trustworthiness, the method of Heuristic Evaluation (HE) was firstly applied. HE Evaluation is a popular and informal inspection method, where the evaluators judge the system based on a list of usability principles (Nielsen, 1994). A list of the heuristics from Xerox Corporation was used for the evaluation of the overall User Interface. Especially for the Web GIS element the GIS heuristics which were developed by Nivala et al. (2008), were used. The evaluators were asked for each problem identified, to document whether they think that influences trust and to also provide a severity rating. The severity rating scale used in this study was from 1 (minor problem) to 5 (critical problem).

One limitation of the HE is that focuses mainly on popular usability problems and it does not take into consideration the cognitive and affective processing of the end user. To overcome this problem the method of Cognitive Walkthrough (CW) was further implemented. Nielsen (1994) explains, that CW is a method that simulates the users' problem solving practices, thus it was expected that the method of CW would help to capture more trust related problems. The evaluators were provided with two persona-based scenarios which reflected the user needs' and expectations. The first persona reflected the needs of a scientist with extensive experience in both environmental data and Web GIS applications. The second persona, involved a novice user with increased Internet suspicion.

During the CW, the evaluators were provided with a list of questions to consider for each task, which amongst others included questions, such as: What is the effect that the user will try to produce? Are there any elements which might decrease user's trust perceptions? As it was expected, the CW allowed the evaluators to add into their observations more elements and concerns (including cognitive and emotional aspects of the interaction) that a potential user has while interacting with the WIYBY interface, capturing in that way problems that are not provided by usability heuristics.

Moreover, a list of trust inducing features and elements suggested in the literature (e.g. branding, testimonials/stories, pictures, videos, chats, blogs, external links, contact details) was given to the evaluators, in order to judge their applicability in the Web GIS context. In addition to that the evaluators were asked to document additional interface elements, which they thought that can further induce trust.

A critical concern with the implementation of both methods, is the number of evaluators that inspect the user interface. It is generally recommended that three to five evaluators can identify the majority of the user interface and thus the subjectivity can be eliminated. For the purposes of this study three evaluators were recruited. All the evaluators were GIS experts, which was essential in order to ensure that the GIS element was examined thoroughly. The first evaluator had used in the past the method of Heuristic Evaluation (HE), the second evaluator was experienced in both methods (HE and CW), while the third evaluator had never performed neither a HE nor a CW.

4. RESULTS

Tables 1 and 2 summarize the number of problems found by each evaluator during the HE and the CW, respectively. The trust related problems, which considered most critical (the Severity Rating given was either 4 or 5), are listed separately. Also note that the problems associated with the GIS element, are listed separately from the general User Interface (UI) problems.

Method: CW	Problem Type	No. of Problems	No. of Trust problems	SR*=5,4 (&T*)
First Evaluator	GIS	22	16	15 (12)
	UI	19	10	9 (7)
	Total	**41**	**26**	**24 (19)**
Second Evaluator	GIS	13	11	11 (10)
	UI	14	9	10 (8)
	Total	**27**	**20**	**21 (18)**
Third Evaluator	GIS	3	3	3 (3)
	UI	6	4	4 (2)
	Total	**9**	**7**	**7 (5)**

Table 1. Heuristic Evaluation- Problems Found
SR* =Severity Rating, T* =Trust

Method: HE	Problem Type	No. of Problems	No. of Trust Problems	SR*=5, 4 (&T*)
First Evaluator	GIS	31	15	12 (9)
	User	15	10	6 (6)
	Total	**46**	**25**	**18 (15)**
Second Evaluator	GIS	18	11	14 (10)
	UI	19	9	13 (8)
	Total	**37**	**20**	**27 (18)**
Third Evaluator	GIS	5	3	5 (3)
	UI	2	1	1 (1)
	Total	**7**	**4**	**6 (4)**

Table 2. Cognitive Walkthrough – Problems Found
SR* =Severity Rating, T* =Trust

All evaluators consistently considered the majority of the identified problems as trust related. Although the third evaluator who was a GIS expert but with no significant experience in using methods such as the HE and CW found less problems, the pattern between total problems found and trust related problems is the same. For example, with the method of HE and for the first evaluator, 54.3% of the total problems found were considered as trust-related. The percentages for the second and third evaluator are 54% and 57%, respectively. Thus, more than half of the problems identified by each evaluator were considered as trust related.

The majority of the trust related problems were considered as being critical (with a severity rating of either 5 or 4). A 60% of the general trust problems were considered to be critical by the first evaluator with the method of HE and 73% with the method of CW. For the second evaluator a 90% of the trust problems were found to be also critical with the method of HE and 90% with the method of CW. Finally, the third evaluator considered as critical all the trust related problems (100%) that found with the method of HE and a 71.5% with the method of CW.

The HE resulted in the identification of more general problems compared with the problems found with the CW, although the CW resulted in the identification of more trust-related problems. This was an expected result as the method of CW supports the consideration of the cognitive and affective needs of a potential end user. It should be mentioned that although some of the problems found were common between the two methods and between the evaluators, the method of CW identified more identical trust-related problems.

The specific trust related problems found during the preliminary expert evaluations provided the basis for the establishment of a list of trust-based guidelines in the context of Web GIS and which are discussed in the next section.

5. DISCUSSION OF GUIDELINES

The majority of the trust related problems found, are similar to those described in the e-commerce literature. Wang's & Emurian's (2005) trust model was modified to effectively group the problems found and introduce a preliminary list of trust-based guidelines in the Web GIS context (Table 3 – Appendix A). The guidelines are grouped in five dimensions and for each dimension the User Interface and GIS guidelines are listed separately, except from the last dimension which is concerned with the trust cue design.

The Graphic Design dimension is concerned with the quality of graphics and other interface elements that are used in the Web GIS context. For example, it should not be ignored that the GIS component increases the complexity of these applications as non-experts require additional time to familiarize themselves with it. Thus it is believed that other interface elements should match popular visualisations (e.g. menus, visited links) so that the users can concentrate on the GIS component.

Concerning the Graphic Design dimension of the GIS element, the evaluators documented that in cases where information on map or legend was not communicated easily (e.g. because of the colour combinations and overlapping symbols), this could potentially reduce trust. The map size was also considered important for the formation of trust perceptions, as a small map size reduces the amount of information on screen and this might give the impression that the operator is trying to "hide" something from the users.

Several of the trust-related problems found were associated with the Structure Design dimension of the website. An efficient structure brings transparency, thus it is necessary to efficiently group information and also provide the users with well organized menus. In this perspective a menu item for the GIS component is essential.

For the Content Design it is critical amongst others that the vocabulary is simple, information is updated and the expectations and needs of both novices and experts are met. For example, information as for how the maps were constructed might not be important for novices, but taking into consideration the user's progress from a novice to an expert level when the application is used constantly, this information might be essential in the future. In the same perspective, instructions and tutorials about the GIS tasks should be provided for the novices. Generalization and scales used are important considerations for trust formations, but these features should be further explored using HCI techniques which involve real users.

The GIS functionality Design should focus on consistency and on users' expectations. If a feature is not functioning in the way that it is expected to be, the system will be considered as being unpredictable and the evaluators thought that predictability in this case is strongly linked to trustworthiness. It should be also mentioned that for example the GIS component of the WIYBY website is only working with Internet Explorer (IE) while there is no direct error response when a user attempts to access the website using a different web browser. In such cases it is very likely that the user assumes that the GIS component is not working at all and thus the whole website loses credibility.

The fifth dimension is concerned with the trust inducing features. In general, the evaluators thought that aesthetics, usability, professionalism and other elements such as the existence of external links are important attributes which can increase trustworthiness in the Web GIS context. Features such as videos, chats, blogs are probably not directly relevant to the GIS context, although further research is required in order to investigate the users' trust expectations and needs in this context. The evaluators suggested that features such as data copyrights and logos of the data providers could eventually help increase trust. The evaluators' suggestions as for the trust inducing features are summarized in Table 4.

Trust – Cue Design
1. The logo of the site operator or provider should be clearly visible from all pages and of high quality. 2. Copyright and data issues (e.g. data provider) about maps should be immediately visible. 3. In case of external links, the website operators should check regularly each link provided. Messages such as "We are not responsible for the content of the websites" can decrease trust. 4. Professionalism and Aesthetics are significant in trust improvement. 5. The layout and functionality of both the User Interface and the Map element should be of high quality. 6. Vocabulary should be simple. 7. Contact details should be easy to find. 8. When additional services, such as the "Sign up for floodline warnings" are used, it is essential to clarify how user data is used and that it is not passed to third parties.

Table 4. Trust-based Guidelines for Trust-Cue Design

6. CONCLUSIONS

Based on simple inspections methods and the example of the WIYBY website this study provides a preliminary set of trust-based guidelines that can be applied in the wider context of Web GIS in order to improve trustworthiness. However, as web- based GIS applications are used in different contexts, and trust perceptions vary according to context, it is necessary to investigate these elements separately. Simple, time efficient and easy to apply methods such as the HE and CW can guide this process.

The majority of the trust-related problems identified by the evaluators, match the problems that are described in the trust-based literature and which refer to mainly e-commerce websites. However, for Web GIS, it is essential to run additional HCI-based experiments which involve real users and who might have additional problems. Although the evaluators considered separately the trust-inducing features, experiments with non-experts can help to identify the users' trust expectations and thus recommend additional features which can be incorporated into this context in order to increase trust.

Finally, it should not be ignored that not only the functional attributes influence trust, but also elements such as the trustor's propensity to trust and the source's (or the website's provider) reputation and credibility. Therefore, in order to understand trust in depth these different elements should be combined and only experiments with real users can reveal how these elements interact with each other, for the formation of the overall trust perceptions in the Web GIS context.

7. REFERENCES

Alsop, R., 2008. "What's in your backyard?" A usability study by persona. *MSc GIS: University College London*

Ba, S. and Pavlou, P. A., 2002. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behaviour, *MIS Quarterly*, 26(3), pp. 243-268.

Cheskin Research and Studio Archetype/Sapient, 1999. *eCommerce trust study* http://www.cheskin.com/cms/files/i/articles//17__report-eComm%20Trust1999.pdf (accessed 13 Sep. 2009)

Chopra, K. and Wallace, W.A., 2003. Trust in Electronic Environments. In: *Proceedings of the 36th Annual Hawaii international Conference on System Sciences, Hawaii* January 06 - 09, pp. 331-340.

Corritore, C. L., Kracher, B. and Wiedenbeck, S., 2003. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, 58(6), pp. 737-758.

Egger, F. N., 2001. Affective design of e-commerce user interface: How to maximize perceived trustworthiness. In: *Proceedings of the International Conference on Affective Human Factors Design,* Singapore June 27-29, pp. 317-324 .

Fabrikant, S.I., 2001. Evaluating the Usability of the Scale metaphor for Querying Semantic Information Spaces. In: Montello, D.R (ed.) *Spatial Information Theory: Foundation of Geographic Information Science*, Berlin: Springer-Verlag, pp. 156-171.

Fogg, B. J., 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann Publishers, Elsevier, San Francisco, pp. 1-283

Grabner- Kräuter, S. and Kaluscha, E. A., 2003. Empirical research in on-line trust: a review and critical assessment. *International Journal of Human-Computer-Studies,* 58(6), pp. 783-812.

Grabner-Kräuter, S., Kaluscha, E. A., and Fladnitzer, M., 2006. Perspectives of online trust and similar constructs: a conceptual clarification. In: *Proceedings of the 8th international Conference on Electronic Commerce: the New E-Commerce: innovations For Conquering Current Barriers, Obstacles and Limitations To Conducting Successful Business on the internet*, New Brunswick, Canada, pp. 235-243.

Haklay, M., 2002. Public Environmental Information Systems: Challenges and Perspectives. Ph. D. : University College London

Haklay, M. and Zafiri, A., 2008. Usability Engineering for GIS: Learning From A Screenshot. *The Cartographic Journal*, 45(2), pp. 87-97.

Harrower, M., and B. Sheesley, 2005. Designing better map interfaces: A framework for panning and zooming. *Transactions in GIS*, 9(2), pp. 77-89.

Hornbaek, K., Bederson, B. & Plaisant, C., 2002. Navigation Patterns and Usability of Zoomable User Interfaces with and without an Overview, *ACM Transactions on Computer-Human Interaction*, 9(4),pp. 362- 389.

Ishikawa, T., Barnston, A.G., Kastens, K.A., Louchouarn, P. & Ropelewski, C.F., 2005. Climate Forecast Maps as a Communication Decision-Support Tool: An Empirical Test with Prospective Policy Makers. *Cartography and Geographic Information Science*, 32(1), pp. 3-16.

Karvonen, K., 2000. The beauty of simplicity. In: *Proceedings of the ACM Conference on Universal Usability.* Arlington, Virginia, pp. 85-90.

Kramers, R. E., 2008. Interaction with Maps on the Internet- A User Centred Design Approach for the Atlas of Canada. *The Cartographic Journal*, 45(2), pp. 98-107.

McKnight, D. H., Choudhury, V. & Kacmar, C., 2002. The impact of initial trust on intentions to transact with a website: a trust building model. *Journal of Strategic Information Systems*, 11(3-4), pp. 297-323

Nielsen, J., 1994. Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), *Usability Inspection Methods*. John Wiley & Sons, New York, pp 25-62

Nikander, P. and Karvonen, K., 2000. Users and Trust in Cyberspace. In: *Proceedings of Cambridge 2000 Workshop on Security Protocols*, Cambridge, pp.24-35.

Nivala, A.M., Brewster, S. & Sarjakoski, L.T, 2008. Usability Evaluation on Web Mapping Sites. *The Cartographic Journal*, 45(2), pp. 129-138.

Riegelsberger, J., Sasse, A. and McCarthy, J., 2005. The Mechanics of Trust: A Framework for Research and Design. *International Journal of Human-Computer Studies*, 62(3), pp. 381-422

Shneiderman, B., 2000. Designing trust into online experiences. *Communication of the ACM,* 43(12), pp. 57-59

Shapiro, S. P., 1987. The social control of impersonal trust. *American Journal of Sociology*, 93(3), pp. 623-658

Skarlatidou, A. & Haklay, M., 2006. Public Web-Mapping: Preliminary Usability Evaluation. In: Proceedings of *GIS Research UK*, Nottingham April 5-7

Unwin, D., 2005. Fiddling on a different planet. *Geoforum,* 36(6), pp. 681-684

Van Elzakker, C.P.J.M., 2005. From Map Use Research to Usability Research in Geo-information Processing. In: *Proceedings of the 22nd International Cartographic Conference*, Spain 9-16 July

Wang, Y. D. and Emurian, H. H., 2005. An overview of online trust: Concepts, elements, and implications. *Computers In Human Behaviour*, 21(11), pp. 105-125

Williamson, O. E., 1993. Calculativeness, trust, and economic organization. *Journal of Law and Economics,* 24(1), pp. 453 – 486

Xerox Corporation, Usability Techniques: Heuristic Evaluation – A System Checklist [Online] http://www.stcsig.org/usability/topics/articles/he-checklist.html (accessed 25 Nov. 2009)

8. ACKNOWLEDGEMENTS

This project is funded by the Engineering and Physical Sciences Research Council (EPSRC) and Arup and it is part of an Engineering Doctorate at University College London.

9. APPENDIX A

	Graphic Design	Structure Design	Content Design	Functionality Design
USER INTERFACE	1. The menu should match popular menu visualisations. 2. The Graphical User Interface elements should offer affordance and should be designed according to Internet standards (e.g. visualisation of links should match expected colour codes and visualisation patterns). 3. The Visualisation of page elements or features should be throughout the website. 4. Use high quality graphics to reveal professionalism.	1. Fix broken links or "Not Found" pages. 3. Group menu in a logical manner. 4. Provide links and hyperlinks to increase the accessibility of information from different pages. 5. Textual information on different pages should be grouped effectively and should be relevant to the context. 6. Titles, headings and subheadings should me meaningful. 6. Provide an index. 7. Provide a menu item for the GIS element.	1. Vocabulary and Terminology should be easy to understand. 2. Information should be recently updated. 3. Advice and error messages should be easily communicated. 4. The website must support both experts and novices. (e.g. in case an expert user expects additional information, provide external links)	-
GIS ELEMENT	1. Colour combinations should be effective (consider colour deficiency). 2. Map results should be communicated clearly and efficiently (not overlapping symbols, different colours or shapes and transparency levels to communicate information). 3. Map size should not be too small. 4. Selected objects should be easy to identify. 5. Base maps should be of high quality and relevant to the context of the application. 6. Scales should be chosen so that each provides high quality and useful maps. 7. Legend should be of high quality and easily communicated.	1. Legend should not block the map. 2. Search box should be immediately visible.	1. Information about map features and results should be easily accessible (ideally next to the map). 2. Generalization should not be such that leads to perceptions for limited accuracy or make maps difficult to read. 3. Scales should be selected, so that they support and are meaningful to the tasks. 4. Map information should support both experts' and novices' needs and expectations. 5. Information as for how the maps were constructed should be provided. 6. Provide Help & Documentation/instructions/or tutorials about maps' tasks.	1. Ensure browser compatibility. 2. Map functionality should be consistent and unique (do not design more than one function for the same task). 3. Map functionality should be consistent at all scales. 4. An undo or cancel feature should be provided. 5. The search box should handle gazetteer.

Table 3. Trust-based Guidelines -Graphic, Structure, Content and GIS functionality Design dimensions

THE DESIGN AND IMPLEMENTATION OF A WEB SERVICES-BASED APPLICATION FRAMEWORK FOR SEA SURFACE TEMPERATURE INFORMATION

HE Ya-wen[a,b,c], SU Fen-zhen[a], DU Yun-yan[a], Xiao Ru-lin[a,c], Sun Xiaodan[d]

[a.]Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
[b.]Yantai Institute of Coastal Zone Research for Sustainable Development, CAS, Yantai 264003, China;
[c.] Graduate School of Chinese Academy of Sciences, Beijing, 100049,China;
[d.] Shandong University of Science and Technology, Qingdao 266510, China

heyw@lreis.ac.cn

KEY WORDS: Web Services, Sea Surface Temperature Information, GIS, Framework

ABSTRACT:

Based on the study of the sea surface temperature data and application, this research put forward a Web Service-based application framework for sea surface temperature information, which can solve the problems in heterogeneity, distribution, and efficiency triggered by networking. An application prototype was successfully designed based on the framework, namely: The Application Service Platform of Sea Surface Temperature Information in the South China Sea. It can integrate heterogeneous sea surface temperature data services and application services, and provide users with transparent, "one-stop" Web applications on sea surface temperature. Users can access the platform to search and fetch valuable information and value-added applications. On the platform, all of the heterogeneous and distributed sea surface temperature information is encrypted, decrypted, monitored, and hence interchangeable according to international standards. The results confirm the feasibility of the application of Web Services to sea surface temperature information integration, and this study can also be referenced by other marine information.

1. INTRODUCTION

With the developing of network, techniques, especially the grid and the Web service, and with the increasing of multi-source sea surface temperature data, which come from remote sensing or investigation, Implementation the Web-based analysis of long time series sea surface temperature data, has become an urgent demand of ocean information services. Because of the dynamic and multiple characters of ocean data, such as Argo data, sea surface temperature field data, Ocean Current data field data and so on, the traditional method of data integration and visualization cannot meet the people's needs of accessing spatial information in anywhere at any time and getting the changing information of sea elements dynamically. And thus, the real-time analysis and evaluation of sea phenomenon quantitatively with high accuracy cannot be executed.

The sea surface temperature field data has been applied in a more widespread domain, and the researchers have proposed many application models of sea surface temperature field data, the traditional applications of these models have some drawbacks, the limit of accessible users and centralization of most of all resources on the local machine. Due to the limit of resources and performance in traditional applications, it is very difficult to realize reuse and sharing of these models. This research proposed the sea surface temperature field data application service framework based on the Web Services technology, which can realize the interoperate of data and models in the distributed environment flexibly. The services included in this framework can be accessed directly by URL. Due to the loose coupling of the Web Services, they can also be integrated in other applications. Study on the demand of Web-based visualization of sea surface temperature field data, the dynamic visualization method and the long-distance localization inquiry method of spatial and temporal process of the sea surface temperature field based on the framework were proposed in this paper.

2. THE WEB SERVICES-BASED APPLICATION FRAMEWORK OF SEA SURFACE TEMPERATURE FIELD INFORMATION

2.1 Web Services architecture

A Web service (also Web Services) is defined by the W3C as "a software system designed to support interoperable machine-to-machine interaction over a network". Web services are frequently just Web application programming interfaces (API) that can be accessed over a network, such as the Internet, and executed on a remote system hosting the requested services.
The system architecture of Web Services was shown in Figure 1.

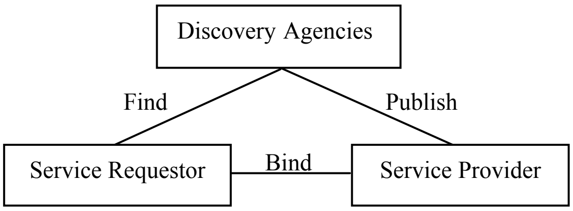


Figure 1. Architecture of Web Services

Software agents in the basic architecture can take on one or all of the following roles:

1. Service requester -- requests the execution of a Web service.
2. Service provider -- processes a Web service request.
3. Discovery agency -- agency through which a Web service description is published and made discoverable.

A software agent in the Web services architecture can act in one or multiple roles, acting as requester or provider only, both requester and provider, or as requester, provider, and discovery agency. A service is invoked after the description is found, since the service description is required to establish a binding.

The figure above illustrates the basic Web services architecture, in which a service requestor and service provider interact, based on the service's description information published by the provider and discovered by the requester through some form of discovery agency. Service requesters and providers interact by exchanging messages, which may be aggregated to form MEPs.

2.2. The application framework for sea surface field information

In the multi-domain sea Surface temperature field data research, many application models have been designed, however, these application models are heterogeneous caused by many of reasons (in detail). Moreover, in the distributed environment, it is very difficult to achieve the transparency application. The researchers always choose to develop independent applications, or to integrate and share heterogeneous and distributional applications through manual methods. The Web service can solve above difficult problems, in the distributed environment, the Web Services-based applications can be reused and used transparently.

The Web Services-based application framework for sea surface field information can integrate the distributed and heterogeneous data and applications. Sea surface temperature application models are encapsulated into web services based on the application framework for sea surface field information. Web Services integrate Web-based applications using the XML, SOAP, WSDL and UDDI open standards over an Internet protocol backbone. XML is used to tag the data, SOAP is used to transfer the data, WSDL is used for describing the services available and UDDI is used for listing what services are available.

Figure 2 is the Web Services-based application framework for sea surface field information. Firstly, the sea surface temperature data is encapsulated as Web service, the uniform interface of data access is the foundation of application services. Secondly, the sea surface temperature field application models are encapsulated as Web Services. Finally, these data services and application model services were registered using the UDDI interface. The programming interface for UDDI consists of four parts: an inquiry (search) API, a publishing API, a delete API and an update API.

3. THE VISUALIZATION OF SEA SURFACE TEMPERATURE FIELD DATA

3.1. The Physical Framework of prototype application

The experimental environment consists of three web services nodes, a web services resource management center and an application service platform of sea surface temperature information in the South China Sea. Node 1(provide web services of sea surface temperature data), node 2(provide web services of sea surface temperature data) , node 3(provide web services of application models) and the application service platform of sea surface temperature information are arranged in the institute of geographical sciences and natural resources research Chinese academy of science. The web services resource management center is arranged in the Northeastern University. In node 1, node 2 and node 3, all the web services are deployed on the IIS 6.0. In addition, ArcGIS Server should be installed in all nodes. Furthermore, the ArcGIS Server ADF should also be installed in the node that has the application services.

Figure 3 is the physical framework for application services of sea surface temperature information. The framework contains three nodes, a web services resource management center and an application service platform of sea surface temperature information. The three nodes are host of the sea surface temperature information, they provide the sea surface temperature data services and the application models services, and all the web services can be registered in the web services resource management center. Finally, the application service platform of sea surface temperature information in the South China Sea provides unification application.

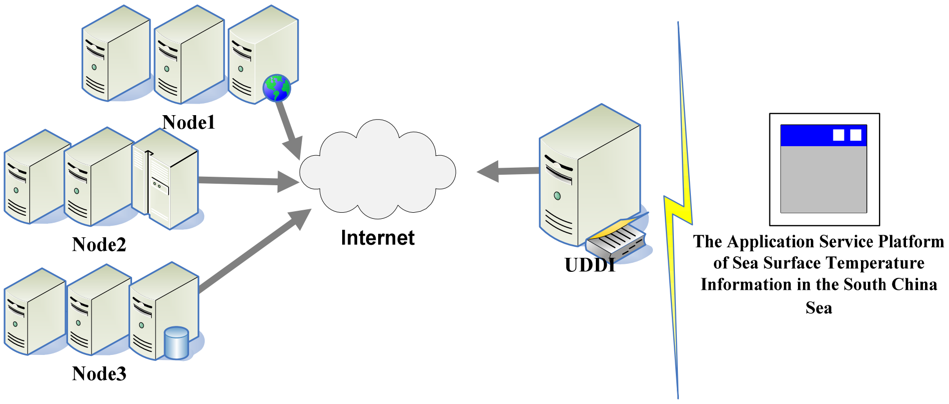

Node1
Node2
Node3
Internet
UDDI
The Application Service Platform of Sea Surface Temperature Information in the South China Sea

Figure 2. the Physical Framework for Application services of Sea Surface Temperature Data

3.2. The realization of data services and model services

The South China Sea is selected as the focus area (105°30' ~122°15' E, 3°~26°30' N), the sea surface temperature data of HDF format is selected as object of study, the data preprocessing flow is shown in Figure 4. The sea surface temperature data distributes separately in two data nodes. All the data are preprocessed according to the data preprocessing flow. The WSDL document segment of sea surface temperature data services as following:

```
<service name="SeaTemperatureDataServer1">
<soap:address location="http:// */arcgis/services/SeaTemperatureDataServer1/MapServer" />
</service>
```

```
<service name="SeaTemperatureDataServer2">
<soap:address location="http:// */arcgis/services/SeaTemperatureDataServer3/MapServer" />
</service>
```

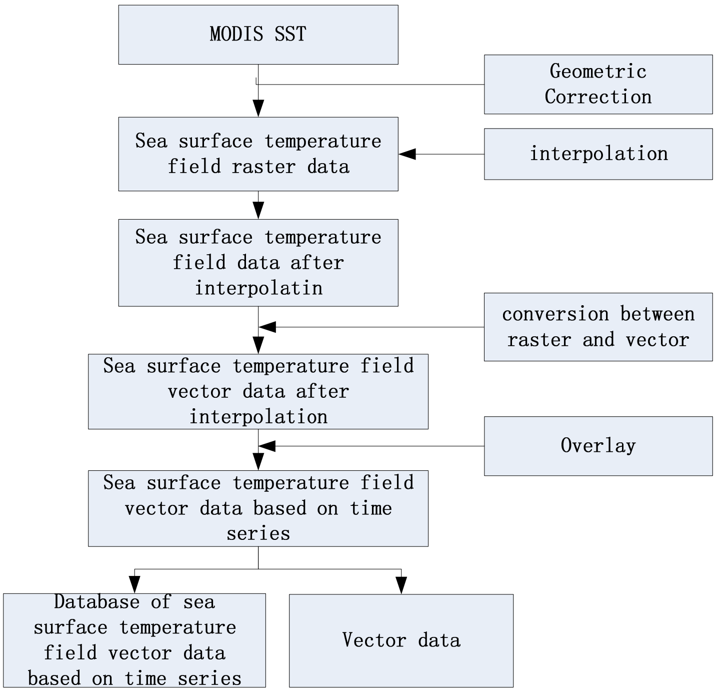

MODIS SST
Geometric Correction
Sea surface temperature field raster data
interpolation
Sea surface temperature field data after interpolatin
conversion between raster and vector
Sea surface temperature field vector data after interpolation
Overlay
Sea surface temperature field vector data based on time series
Database of sea surface temperature field vector data based on time series
Vector data

Figure 3. The pre-procession and organization flow of sea surface temperature field data

The point process network visualization method aims at the spot objects in spatial. In network environment, the marine environment information dynamic process of a spot object in marine can be expressed by curve with time as abscissa and marine environment information as ordinate.

The surface process network visualization method takes the marine environment element field as the research object. The tow above-mentioned sea surface temperature field data visualization models are encapsulated as service. The WSDL of this Web Service is:

```
<service name="SeaTemperatureService">
<soap:address location="http://*/ SeaTemperature/ SeaTemperatureService.asmx " />
</service>
```

The sea surface temperature field data services and model services can be registered in the web services resource management center, which is built based on UDDI(Universal Discovery Description and Integration). The resource management center is the yellow pages of Web services. As with traditional yellow pages, you can search for a provider that offers the services you need, read about the service offered and contact someone for more information. You can, of course, offer a Web service without registering it in resource management center, just as you can open a business in your basement and rely on word-of-mouth advertising but if you want to reach a significant market, you need UDDI so your customers can find you. The "one stop" marine environment information service can be provided based on the resource management center. The registration information of sea surface temperature information services, which were registered in the resource management center, as follow

Name	Type	Location	Provider	Publish time	Valid time
SeaTemperatureDataServer1	Map Service	Node 1	Heyw	2009-2-12	Working day
SeaTemperatureDataServer2	Map Service	Node 2	Heyw	2009-2-20	Working day
SeaTemperatureService	Web Service	Node 3	Heyw	2009-2-18	Working day

Tablel 1. the Registration Information of Sea Surface Temperature Information Services

3.3. The Application examples

In the prototype application, the integration of data and application is realized with dynamic loading the Web services. The user may search the data services and model services through the name of service provider, the service name as well as the data spatial scope, and then load them in the platform.

With the above integrated the sea surface temperature data service, the user may produce the curves of sea surface temperature data and the profile curves of sea surface temperature data and the dynamic visualization of sea surface temperature data, these functions can be realized through the SeaTemperatureService.

Figure 4 is the 2D visualization of sea surface temperature field data (provieded by IGSNRR, CAS, in South China Sea (105°30′–122°15′E, 3°–26°30′N)) is carried out by the chrome/gray method. Figure 5 is the profile curve of sea surface temperature data, in the curve system the vertical axis is the valve of sea surface temperature, the horizontal is position and the different colours represent time. Figure 6 is the dynamic visualization of sea surface temperature data, each figure in this system represents the value variety at one time, so we can set an interval for all the figures to display in order of time.

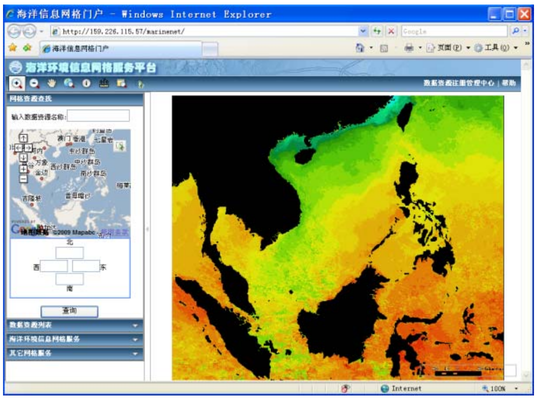


Figure 4. the Integration of the Sea Surface Temperature

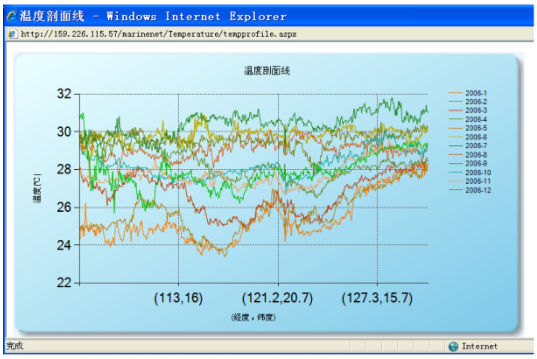


Figure 5. the Profile Curve of Sea Surface Temperature Data

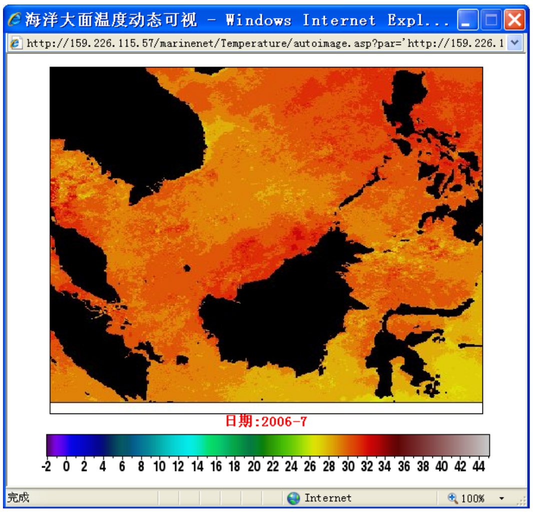

海洋大面温度动态可视 - Windows Internet Expl...
http://159.226.115.57/marinenet/Temperature/autoimage.asp?par='http://159.226.1
日期:2006-7
-2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
完成
Internet
100%

Figure6. the Dynamic Visualization of Sea Surface Temperature Data

4. CONCLUSION

Although the gradual and further application of the sea surface temperature data, but the study about the integration and sharing of the sea surface temperature data just begun, and many problems need to be solved and consummated immediately. Based on the researches of the current application status and characteristic of sea surface temperature information, this article put forward a Web Service-based application framework for sea surface temperature information. Web service was used to solve the problems in heterogeneity, distribution, and efficiency triggered by networking, and to realize the sea surface temperature data share in different application fields. A prototype application was successfully designed based on the framework, namely: The Application Service Platform of Sea Surface Temperature Information in the South China Sea. It can integrate heterogeneous sea surface temperature data services and application models services, and provide users with transparent, "one-stop" web applications on sea surface temperature field information. Users can access the platform to search and fetch valuable information and value-added applications. Through the platform, all of the heterogeneous and distributed sea surface temperature information can be encrypted, decrypted, monitored, and then interchanged according to international standards. This research demonstrates the typical application of grid Web Services technology. The results confirm that Application Service Platform of Sea Surface Temperature Information is feasible in data integration and sharing. Furthermore, this study can also be referenced by other marine information.

References

[1] Su Fenzhen, Zhou Chenghu, and Yang Xiaomei, "The Fundamental Study on Key Technologies for marine geographic information system", *Acta Oceanologica Sinica*, Beijing, 2004, 26(6):22-28.

[2] Lu Feng, Zhou Daliang, "Probe into Network Oriented Large-scale GIS Platform Structure", *Journal of Remote Sensing*, Beijing, 2002. 36-44.

[3] Su Fenzhen, Zhou Chenghu, "Definition and structure of marine geographic information system", *Acta Oceanologica Sinica*, Beijing, 2004,11(26), 26-33.

[4] Whright D J, "Coastal mapping and charting", *Geospatial Solutions,* 2004, 14(9):20-21.

[5] Li Deren,Wang Yangdong,Zhu Qing,et al, "Data Model and visualization of 3D City Landscape Based on Intergrated Database", *Geo-Spatial Information Science*, Beijing, 1999, 2(2):21-25.

[6] Gong Jianhua, Lin Hui, Xiao Lebin, and Xie Chuanjie, "Perspective on Geo-Visualization", *Journal of Remote Sensing*, Beijing, 1999, 3(3):236-243.

[7] HU H B, LIU Q Y, LIN X P, "The South Pacific subtropical mode water in the Tasmansea", *Journal of Ocean University of China*, Qingdao, 2007,6 (2) : 107-116.

[8] Li Deren, Zhu Xinyan, and Gong Jianya, "From Digital Map to Spatial Information Multi-grid——A Thought of

Spatial Information Multi-grid Theory", *Journal of Remote Sensing*, Beijing, 2003,28(6):642-650.

[9] Li Deren, Huang Junhua, and Sao Zhenfeng, "Design and Implementation of Service-Oriented Spatial Information Sharing Framework for Digital City", *Journal of Remote Sensing*, Beijing, 2008,33(9):881-885.

[10] I Forster and C Kesselmanl, *The Grid: Blue print for a new Computing Infrastructures*, CA:Mogan Kaufmann Publishers, San Francisco, 1998.

A STUDY OF SPATIAL DATA SHARING SYSTEM WITH WEB SERVICES

Fan Li [a, *], Xu Zhang [a], Xian Jiang [a], Yong Shan [b]

[a] Institute of Forest Resource Information Technique, CAF, Beijing, -（lifan,zhangxu,trista）@caf.ac.cn
[b] Daniel B. Warnell School of Forestry & Natural Resources, Georgia State University,Georgia-shanyongsy@hotmail.com

KEY WORDS: Web Services, Spatial data, OGSA.

ABSTRACT:

Web Services is service-oriented architecture advanced in recent years, with its definitive agreement and complete platform and language independence and a high degree of loosely coupled, gradually become an important direction of application integration. According to design ideas of spatial data sharing system, through exploring the Web Services technology, the design was proposed, based the Web Services technology node in the system. Based the description of the overall framework of the spatial data sharing system, the approach of the spatial data sharing service node was elaborated in detail.

1. INTRODUCTION

Recently, people have developed a lot management information systems for the demands of the forestry information building .The next building step is to share these data and information and accelerate the forestry information system more interconnected ,large-scale and integrated .At the same time, all the data are distributed in different department which form the information silos.We should integrate the data which scattered in various information systems and spatial data node ,by analyzing the integration of forestry to the needs of existing information systems, integrating the technology can be used to establish, supporting for heterogeneous databases and building operating system distributed spatial data sharing platform.

Spatial Data is the important basic data resources of multi-disciplinary innovation, eco-environmental monitoring and national sustainable development research. Multi-disciplinary innovation include Contemporary international earth science, environmental science, ecology, meteorology, oceanography, land science, natural resources, science, natural disasters, agriculture, forestry, grassland science and so on.

In order to extensively dispense Spatial data resource and satisfy the demands of science research , share spatial data is urgently. There’s some problem:

1. Institute of Forest Resource Information Technique is receiving different kind data from 1990 which capacitance has increased to 30TB （TeraByte） , even now the capacitance is increased by more than 10GB per day.
2. The current spatial data even be processed and compressed which also size in few hundred MB . To obtain these data , network should as the transmit tool. And in the transmission process ,network should offer the the breakpoint Resume function while the network experiencing intermittent.
3. Our cooperant department Remote Sensing Ground Station of Chinese Academy of Sciences also has large amount data and developed spatial data share system,we should use the Web Services technology to integrate the different nodes spatial data and different platforms .Two units of spatial data sharing model of the design shown in Figure 1 Followed:

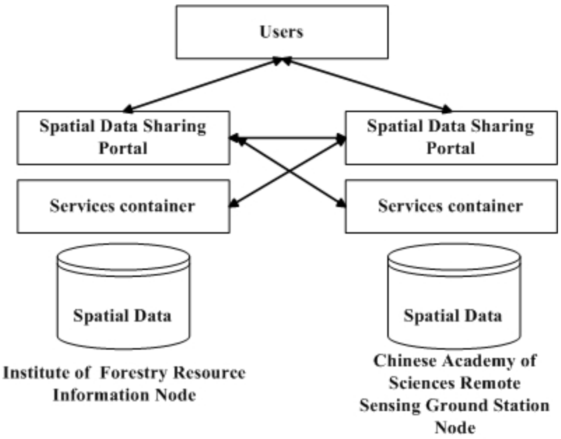

Users
Spatial Data Sharing Portal
Spatial Data Sharing Portal
Services container
Services container
Spatial Data
Spatial Data
Institute of Forestry Resource Information Node
Chinese Academy of Sciences Remote Sensing Ground Station Node

Figure 1 Spatial data sharing model

2. WEB SERVICES TECHNOLOGY

Web Services is a revolutionary distributed computing. It uses XML-based message processing as a basic data communication, to eliminate the use of different component models, operating systems and programming languages that exist between different systems, so that heterogeneous systems can be used as part of the computing network to run concurrently. Developers can use to create distributed applications such as the use of components, by creating Web Services from a variety of sources, combined with the application. Because Web Services are built on the basis of some common protocols, such as HTTP (Hypertext Transfer Protocol), SOAP (Simple Object Access Protocol), XML, WSDL (Web Services Description Language), UDDI (Universal Description, Discovery, and Integration, etc..

Web Services as a new method for the function and application integration technology, solve the original integration of technology in the Internet telecommunications issues. Web Services based on XML service description documents, service requests and feedback the results on the Internet can be passed through the HTTP protocol, it is easy to be accessed and return results. Web Services is a dynamic integration program, all services can be dynamically through UDDI standard was found，bound and use, easy to adapt to changes in the system, improve system flexibility and scalability. The basic model of compositive system using Web services is followed.

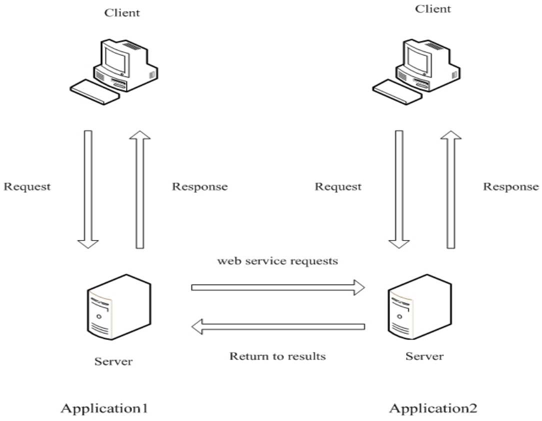


Figure 2 The basic model of compositive system using Web service

3. DATA SHARING PROGRAM

Data sharing management is the general name for establishing, clarrifying, controling, accessing and maintaining of metadata sets. A metadata database is combined by various single metadate sets. The metadata management system is used to manage and control the metadata databse by a centralized or distributed way. Users are allowed to share the metadate. This system can provide demanding metadate for data administrators, database administrators, system analysts, programmers and ultimate users. The database interoperating technology is adopted by this system, the metadate could be extracted from the database at anytime.

This system supports diversification of database, differentiation of operation platform, non-unity of data type and storage methods. The main problems that should be resolved is that: isomerism elemination, standards based, diffuse coupling and data security. And Web Services based data sharing is the best way to resolve those problems.

4. DESIGN OF DISTRIBUTED SPATIAL DATA SHARING PLATFORM

4.1 Functional Design of Distributed Spatial Data Sharing Platform

Distributed spatial data sharing platform is not only the important part of data sharing research and construction, but also the main interface through which to provide convenient and high-efficient services to end users. Considering spatial data’s properties in terms of multiple sources as well as ensuring that this service sharing platform can become a main platform providing authoritative, consistent, fleet services, it is necessary to conduct a great deal of research and technology improvement in the fields of meta-data, data management model, users management strategy and data index, and to buildup a uniform platform that can provide data transition service and distributed information from spatial ’s portals to all departments, parties and the public. Meanwhile, it is also necessary to leverage data portal technology to setup uniform data accessing interfaces so that to create uniform interfaces to all types of data for purpose of user management, authentication, and security control intensively. Currently the spatial data sharing platform consists of several levels, such as portal level, service sharing level, core service level, resource management level and network platform level, and functions such as meta-data management, data publishing, search and exploring, data download and user management, etc. The multi-node spatial data portal’s functions are displayed as below.

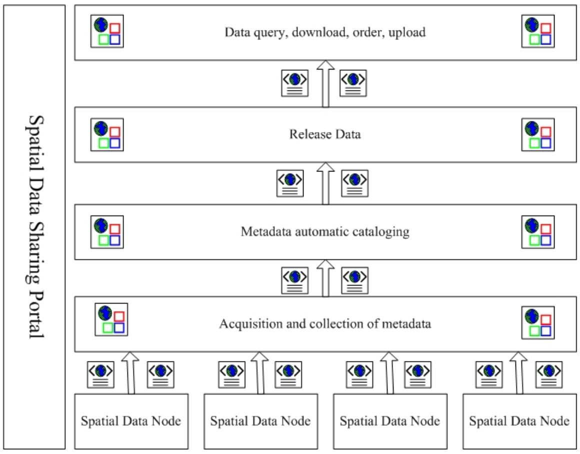

Spatial Data Sharing Portal
Data query, download, order, upload
Release Data
Metadata automatic cataloging
Acquisition and collection of metadata
Spatial Data Node
Spatial Data Node
Spatial Data Node
Spatial Data Node

Figure 3 Multi-node spatial data portal function

1. Data searching/exploring and booking system: Be able to provide service abilities such as sharing spatial data set, searching/exploring and booking download text data set, searching/exploring of spatial data based on several methods.

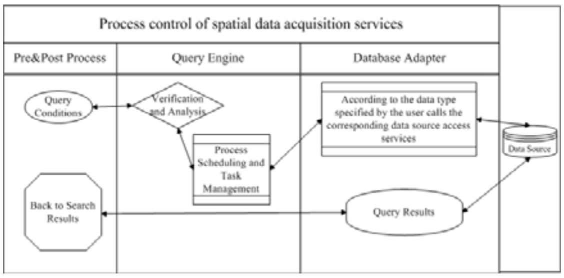

Process control of spatial data acquisition services
Pre&Post Process
Query Engine
Database Adapter
Query Conditions
Verification and Analysis
Process Scheduling and Task Management
According to the data type specified by the user calls the corresponding data source access services
Data Source
Back to Search Results
Query Results

Fig 4 The services process for obtain data

(1) Metadata-based data set query: Provide metadata query service to sharing spatial data set, display to users by means of fast-graphic, fast-graphic-description, etc, and provide multi-query ways based on time, satellite type, latitude and longitude coordinate.

(2) Data set download: Provide booking function to sharing spatial data set. Be able to provide download service via FTP of HTTP according to a user ’ s own requirement.

2. Data publishing sub-system: Can manage and publish information of data sharing websites, including homepage customization, publishing content management, news information, latest updated information and images.

3. User management sub-system: Support functions like multi-type user ID authentication, authentication management, log tracking, etc. First, to manage administrator, data exchanging people and normal users by means of certification, password; Second, to grant different access to users according to different user type; Third, access information can be statistically collected and inquired online no matter to administrator, data exchanging people or normal users.

4. Metadata catalog service sub-system: Provide metadata catalog services to implement metadata’s collection, database creation, setup/upload/maintain catalogs. Coordinate current metadata for spatial data set, add related data service, download information, delete expired data, etc, provide spatial data set’s catalog services.

5. Administration functions implemented for the platform:

(1) Save metadata into database: The program can analyze metadata’s file name that to be uploaded automatically, and then judge whether the going-to-upload data exists already or not. If the data doesn’t exist, it will be written into the database’s specific field. Otherwise, wrong operation information will be shown to administrator and requires the administrator to re-input again. The administrator input all the contents to fields manually, and the program will check with the record exists or not. If the record doesn’t exist, it will be written into database. Otherwise, administration domain error will be displayed, re-input will be required.

(2) Upload fast-view: The selected fast-view will be upload to dedicated position on server, and related information will be written into specific field in database according to the file name parsing result.

(3) Data export: Data can be exported to Excel sheet according to input data period.

(4) Data import: A given Excel sheet can be parsed according to specific fields, and such data can be written into database accordingly.

(5) Files in server can be written into database automatically: By scanning a dedicated catalog’s file storage structure, data can be written into related tables in database.

(6) Check booking information: After log in, the administrator can check all users’ booking information.

The distributed spatial data spatial ’s well running not only increases the sharing services’ efficiency and affinity, but also provides strong technical support to future construction of information sharing platform and data integration and sharing in a standardization way.

4.2 Composition of Spatial Data Sharing Services

Currently the spatial data sharing services constructed include data service (data processing, data mining, data provision), statistics analysis service (statistical sheets, graphic analysis processing), spatial analysis calculation service (buffer, graphics overlying, key factor’s filter, topology analysis, etc), spatial information publishing service (supporting roaming, zoom out, zoom in, selection, etc), interface processing service (support text, statistical graphic, table, video, vector graphic presentation, etc), project data interface service (implement Returning farmland to forest business data extraction, conversion, and loading). These web services can be distributed among different servers.

4.3 Register Spatial Data Sharing Services

The register process management of spatial data sharing service includes register information syntax checking, web service ID creation, web service register. Detail web service information consists of two XML documents, which are detail document of web service and interface document (WSDL document). The register information required during registering process include service name, selection of service catalog, saving path of the service, service description. The system records a registered user according to his login information automatically. Web service process is depicts as below in Figure 5.

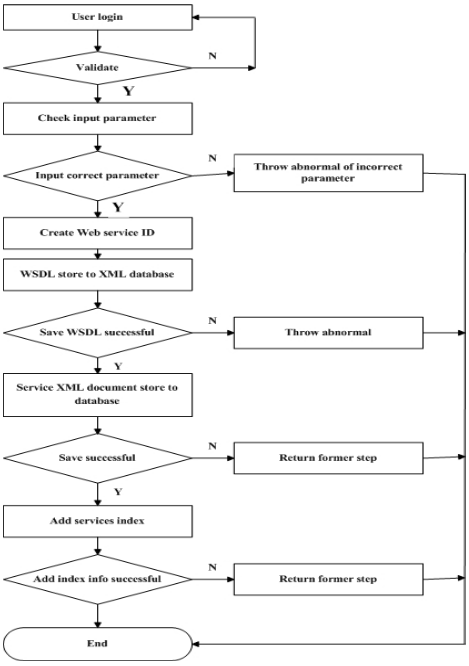


Figure.5 The flow of the web services registering

4.4 Invoke Mechanism of Spatial Data Sharing Services

During an invoke process of sharing services, spatial data management platform will invoke web services access interface (service container's client interface) first, then communicate from service container's client to service container's publishing interface via related protocol, the service contain plays a role of invoking services. The invoke process is virtualized to invoke services directly by a series of protocols, and a series of details during the process are hidden. There are many services registered in spatial data management platform, so the services can be looked up according to service name, service catalog, and finally check the details.

4.5 Spatial Data Sharing Services' Combination Model

SDS, the spatial data sharing services combination model, is a three-mode group (D, S, R, X), as

1. D = (d1 , d2 , , dn) is the set of spatial metadata;
2. S = (s1 , s2 , , sm) is the set of data services;
3. R: the set of services' logical relationship, $\sum_{1}^{n} s_i \bigcup \sum_{1}^{m} s_j \bigcup \cdots\cdots \bigcup \sum_{1}^{t} s_k$, where I, j, k are arbitrary elements belong to service set.
4. X=(x1 , x2 , , xt) is the set for use of transferring XML messages;

Combination relationship is constructed by the procedure from service query to data achievement.
X（Query message）——>SRS（service conformation）——>X（transfer message）——>D（invoke data）——>X（transfer message）——>S（return to data service）——>X（return to message）

4.6 Construction of Spatial Data Sharing Services Nodes

It is required that all types of data service integrate to work coordinately and combine as thick-size format. Meanwhile, the automation of spatial data service is required to describe the integration and coordination of services, as well as to implement the internal process for spatial data processing nodes.

The spatial data service nodes publish and register the services to dedicated service resource management center, and update the services periodically. After receiving users' invoke queries, the spatial data service nodes achieve data from remote data server or local ones according to users' requirements, and trigger services to monitor the running status. The spatial data services nodes divide tasks from up level into sub-tasks, and distribute such sub-tasks to PCs (or high-capacity computers) in data pool, execute, callback, and integrate the execution results, finally, the spatial data service nodes return the results to up level. As a separate node, the data service node will provide a simple interface to outside, and the user can access this node via IE explore.

5. CONCLUSION AND FUTURE PROSPECT

Web Services technology brings new solution for spatial geography information sharing, inter-operation and integration. Future spatial information sharing will be presented as service format, including spatial data query service, spatial data processing service, etc, and all these services can be integrated into a dedicated a new service to present.
Web Services technology constructs a distributed sharing system model for spatial data in the scope of Internet, it helps to increase the scalability and inter-operation abilities and implements a multiple levels system structure. Meanwhile, it leverages group environment to process TB level spatial data, supports spatial data distribution calculation, spatial information distribution sharing in the network environment.

6. REFERENCES:

Zhiwei XU, Wei LI, Research of Vega network's System Structure. Computer Research and Development, 2002，39(8)：923～929

The Globus Project. http://www.globus.org

UDDI.org. http://www.uddi.org

Zhiwei XU, Baiming FENG, Wei LI. Grid computation technology[M] Beijing, PHEI，2004。25-58

Ian Forster，Carl Kesselman. The Grid 2[M] Beijing，PHEI，2005。124-220

Jingbo XIA, Ying LIU, Shengrong WANG. Grid Mechanism and Development[M] Xi’an, Xi’an Electric Science Technology University Publishing, ，2006.. 30-86

SUN．Sun One white paper．http://www.sun.com.cn，2002.1

W3C.Web Service Related Standards. http://www.w3c.org,2003-06．

ZDNetChina.http://www.cn-java.com/target/news.php news_id=1648，2002-05．

Juanle WANG, Yunqiang ZHU, Chuanjie XIE. Design and Development of the Earth System Science Data Sharing Network Platform[J]. Cutting edge of Geography，2006， 54～59．

Dingcheng HUANG．General Architecture of Science Data Sharing Project[J]．Chinese Basic Science，2003， 63～68．

Huqing LIANG，Ronghua MA．Metadata Design and Research of General Geography Information System in Provinces [J]． Journal of Remote Sensing，2002，272～278．

Jinhua WU．Discussion of Geography Spatial Metadata[J]．Journal of Xi&apos;an University Engineering Science and Technology ，2002， 59～61．

Zhaoning WEN, Hong ZHANG. Web-Service-Based Distributed Spatial Data Sharing Model [J]. Computer Engineering，2005，31(6)，25～26,62.

Lin YANG, Xiaoping DU, Shunping ZHOU. Data Exchange Strategy of Distributed-Geological-Map-Based Database [J]. Journal of China University of Geosciences，2006，31(5)，659～662.

Caifu HE. Application of Earth Observing System (EOS) and Spatial Information[J]. Jiang Xi Weather Technology and Science，2002，26（Supplement），182～185.

Yanfang CHENG, Qingping GUO. Application Research of Distributed Calculation Technology and Coordinated Design System [J]. Journal of Wuhan University of Technology（Jiaotong Kexue yu Gongcheng Ban），2007，31(2)，296～299.

LI Fan, ZHANG Xu, LIU Yan. “Research and development of Web Servicess design standard in forestry science data sharing system”, Journal of Northwest Sci-Tech University of Agriculture and Forestry, 2007，35（extra edition）12～16.

LI Fan, ZHANG Xu, CHEN Yan. “Research and Implementation of Web Servicess management and mechanics in Digital Forestry platform”, Forestry Science, 2006，42(extra edition）,11～114.

7. ACKNOWLEDGEMENTS

Many thanks to Zengyuan LI(Vice Superintendent of Institute of Forest Resource Information Technique, CAF)’s great instruction and guidance to this project. This paper was supported in part by Science Foundation of The Chinese Academy of Forestry, under contract number IFRIT200801.

AN AUTOMATED INTERNET GEOINFORMATION SERVICE FOR INTEGRATING ONLINE GEOINFORMATION SERVICES AND GENERATING QUASI-REALISTIC SPATIAL POPULATION GIS MAPS

S. Shi and N. Walford

Centre for Earth and Environmental Sciences Research, School of Geography, Geology and the Environment, Kingston University London, Penrhyn Road, Kingston upon Thames KT1 2EE, UK – (s.shi, nwalford)@kingston.ac.uk

KEY WORDS: Automatic spatial data handling, internet-based Geoinformation Service, quasi-realistic GIS models, census, spatial population distribution, dasymetric maps, OGC, geoprocessing

ABSTRACT:

This paper presents the design and realisation of a low-cost automated internet geoinformation service, named Kingston Automated Geoinformation Service (KAGIS), that provides the United Kingdom's academic community with fully processed quasi-realistic GIS maps of spatial population distribution, known as dasymetric maps. The key innovative aspect of the service is that it undertakes spatio-temporal interpolation for census area statistics and enable users to overcome the difficulties associated with inconsistent boundaries used by different census events. It is designed and programmed to carry out live, on-demand online integration of Edinburgh University's geospatial data service (EDINA) and Manchester University's census data service (MIMAS), to automatically geoprocess and deliver digital dasymetric maps for any of 2001, 1991 and 1981 census years and any area of interest at all census area levels in the United Kingdom as per user's request. The system is capable of automatically handling large GIS datasets in accordance with users issuing requests. The system unleashes a large amount of geospatial and census data held in these formerly isolated digital repositories and carries out all necessary data processing to generate the requested spatial information products in a novel workflow. The online data transmission, retrieval, integration, processing and generation of quasi-realistic spatial population (dasymetric) GIS maps at the English county level typically take 2 minutes.

KAGIS enables census data special interest group in the UK's academic community to access and utilise this newly available geoprocessing service for a wide range of applications, such as analysing population changes over the time and undertaking fine resolution detailed spatial population estimation for planning at various scales and areas of interest. Online interactive selection of datasets and the entire procedure of online heterogeneous data integration, areal interpolation and data processing for dasymetric maps with ancillary landuse maps derived from remotely sensed imagery are automated to offer end users opportunities to simply follow a few instructive checkbox and button-clicking steps to obtain resultant digital dasymetric maps.

The design of KAGIS adopts the service oriented design architecture to couple the latest web portal technology and ArcGIS Server technology to realise the new secure internet geoinformation service with a set of specially designed and coded geoprocessing programs performing coherent chained actions in a novel workflow and it is programmed as a novel stateful application to be able to provide the runtime online data integration, processing and resultant GIS map delivering service to concurrent internet users.

Novel aspects of KAGIS include its operation as a new fully integrated service system that acts as an automated online integrator, data processor, e-content generator and secure service deliverer. The system innovatively carries out live, flawless, on-demand combination of formerly isolated geospatial and census data-only repositories and services according to user requests. The novel design of coherently chained modules and workflow enables fully automated processes for real-time digital data acquisition, real-time data processing for on-demand GIS model generation and service delivery.

The whole and parts of the secure online service technology of KAGIS are generic and universally applicable in any countries in the world. With the process of globalisation and development, all countries have been and are undergoing significant development and change. The technology and know-how experience of KAGIS can assist any country in the world to develop its secure internet GIS services, integrating geographical information for specific decision making systems as well as providing low-cost and efficient ways to process and utilise existing geographical data for a multitude of purposes such as development planning.

1. INTRODUCTION

Edina is an internet service provider of digitised British census boundary datasets and Mimas is an online service provider of population statistics datasets of the United Kingdom (Figure 1).

To analyse change is one of the most important functions of a census (Rees, 1998) and another aspect of important use of census is to focus on the up-to-date information as representing the contemporary demographic and socio-economic condition of the population.

However, existing dataset services and datasets themselves do not offer users any easy and friendly ways to process datasets to summarise and compare geographical and demographic change. Datasets are disparate, census boundaries between separate census years changed and there is also inconsistency in the aggregate statistical counts produced (Martin, 1998a&b; OFFICE FOR NATIONAL STATISTICS, 2001; Mackaness and Towers, 2002; Rees and Martin, 2002; Martin, 2003; Norman, Rees and Boyle, 2003; Walford and Shi, 2009).

This presents a fundamental problem and challenge when analysis of socio-economic change is to be undertaken due to incomparability of the datasets.

Furthermore, aggregate statistics are summarised for each census areas. Landuse information is not taken into account and population is assumed to be evenly distributed throughout the areas. Historically, ward is the smallest spatial resolution area for census data. Thus, conventional GIS map depiction of spatial population distribution is far from the reality.

To produce GIS datasets in order to compare for analysing change over time, GIS datasets need to be processed to conform to the same set of census boundaries. To achieve the result of more realistic spatial population distribution, ancillary landuse information needs to be taken into account. Necessary data processing consists of areal interpolation and dasymetric mapping interpolation (Eicher and Bewer, 2001). Figure 2 synoptically illustrates the effect of areal interpolation and dasymetric mapping techniques by using simplified hypothetical zones of separate census years, as described in Walford and SHI (2009).

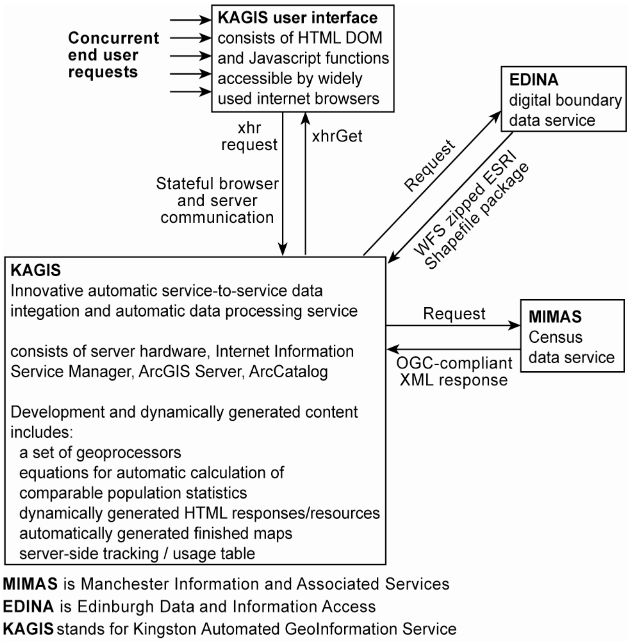

Concurrent end user requests
KAGIS user interface
consists of HTML DOM and Javascript functions accessible by widely used internet browsers
xhr request
xhrGet
Stateful browser and server communication
EDINA
digital boundary data service
Request
WFS zipped ESRI Shapefile package
KAGIS
Innovative automatic service-to-service data integration and automatic data processing service
consists of server hardware, Internet Information Service Manager, ArcGIS Server, ArcCatalog
Development and dynamically generated content includes:
a set of geoprocessors
equations for automatic calculation of comparable population statistics
dynamically generated HTML responses/resources
automatically generated finished maps
server-side tracking / usage table
Request
OGC-compliant XML response
MIMAS
Census data service
MIMAS is Manchester Information and Associated Services
EDINA is Edinburgh Data and Information Access
KAGIS stands for Kingston Automated GeoInformation Service

Figure 1: Collaborative relationship of internet services and data flow chart for creating the automatic geoprocessing service

Areal interpolation and dasymetric mapping techniques have been well researched in previous studies and there is a growing interest in achieving comparability over space and time and applying these methods (Flowerdew and Green, 1992 and 1994; Goodchild, Anselin and Deichmann, 1993; Wilson and Rees, 1998; Eicher and Bewer, 2001; Dorling, Martin and Mitchell, 2003; Dorling and Rees, 2003; Geddes, Gimona and Elston, 2003; Norman, Rees and Boyle, 2003; Reibel and Bufalino, 2005; Gregory and Ell, 2006; Hess, 2007; Longford, 2006 and 2007; Shroeder, 2007; Wilson and Rees, 1998).

This paper presents an innovative automated internet service solution to make available comparable quasi-realistic spatial population distribution GIS datasets of any area of interest based on the census boundaries of any of 2001, 1991 and 1981 census years in the United Kingdom to advance analysis of demographic and socio-economic change over time.

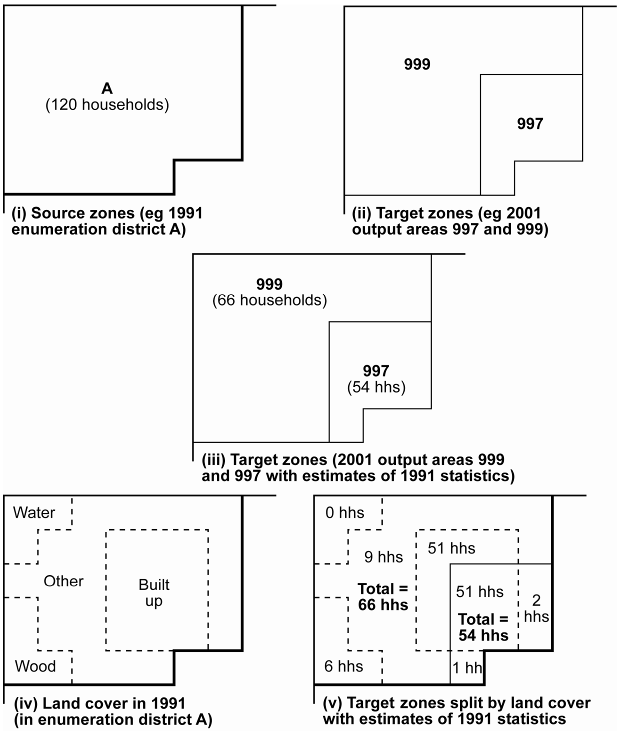

A
(120 households)
(i) Source zones (eg 1991 enumeration district A)
999
997
(ii) Target zones (eg 2001 output areas 997 and 999)
999
(66 households)
997
(54 hhs)
(iii) Target zones (2001 output areas 999 and 997 with estimates of 1991 statistics)
Water
Other
Built up
Wood
(iv) Land cover in 1991 (in enumeration district A)
0 hhs
9 hhs
51 hhs
Total = 66 hhs
51 hhs
2 hhs
Total = 54 hhs
6 hhs
1 hh
(v) Target zones split by land cover with estimates of 1991 statistics

Figure 2: Synoptic illustration of the effect of areal interpolation and dasymetric mapping techniques by using simplified hypothetical zones of separate census years (hhs denotes households. Scenario: source year = 1991 & target year = 2001 effect of areal interpolation (iii) is the result of transferring population statistical data in (i) into 2001 census boundaries (ii); effect of dasymetric mapping (v) is the result of reallocating population statistics in (iii) by incorporating landuse information in (iv))

2. AIM

The aim of the Kingston Auto Geoinformation Service (KAGIS) project was set to develop an innovative automated internet-based geoinformation service to implement automatic spatial data handling mechanism, procedures and modular programs to integrate heterogeneous datasets, automate necessary data processing to allow end users to simply follow a few instructive checkbox and button-clicking steps to obtain digital quasi-realistic GIS models of spatial population distribution of any area of interest based on a set of zone boundaries of a selected census year in the United Kingdom.

In order to compare GIS maps of spatial population distribution across time, these GIS maps need to be produced by using the same set of geographical boundaries. To satisfy a diverse range of users, the service system needs to be designed to allow users to choose any of existing census boundaries. Automatic spatio-temporal interpolation with areal interpolation needs to be carried out based on user's decision on the selection of a specific boundary dataset pertaining to a particular census year as target year's geographical boundaries. Any of other year's census data can be set as source year's population statistical data. The spatio-temporal interpolation (areal interpolation) processes the data and allocates population statistical data into the target year's geographical boundaries.

By utilising landuse GIS map derived from remotely sensed imagery, dasymetric mapping interpolation processes the resultant target GIS map to result in quasi-realistic spatial population GIS dataset.

3. SYSTEM ARCHITECTURE, DESIGN AND OPERATION

In the KAGIS project, the system design adopts service oriented architecture and couples the latest web portal technology and ArcGIS Server technology. ArcGIS Server and ArcCatalog are installed on an internet server which runs Windows Server Operating System with Internet Information Services Manager (IIS) (ArcGIS Resource Centre).

On the server, a set of specially designed and coded geoprocessing modules written in Python scripting language are published by ArcCatalog, managed by ArcGIS Server Manager and made available as Representational State Transfer services (REST services) by ArcGIS Server. Python geoprocessing tools/scripts are designed and programmed for integrating web services and ESRI tools at the back-end of KAGIS. Server-to-server communication and data exchange are achieved through services' compliance with OGC standards (Gardels, 1997; Doyle and Reed, 2001; OGC, 2004).

3.1 Applying unified ontological coding approach for both spatial and statistical datasets

Central to the successful design, development and implementation of automatic spatial data handling mechanism, procedures and modular programs employed in the KAGIS system is a unified ontological coding system for spatial and statistical datasets. The standardised codes are used in the design and implementation of both front-end user interface and server-side lookup tables. The logical design of the coding system and the logical design of scripts of KAGIS allowed full automation. In addition, a linking table between legacy population statistical variables and new comparable statistical variables have been developed and used as a code look up table for practical solution of correct data extraction.

Within the Edina's dataset, each census zone is coded with a unique identifier. This identifier possesses a regular pattern and presents an ontological hierarchical relationship among census zones enabling service-to-service data retrieving specific to demand. Mimas uses this coding system together with a standardised population statistical variable coding system newly developed at Kingston for service call to search and produce responses of OGC-compliant XML data streams (OGC, 2004). Automatic service calls are made by dynamically composing service request strings with variable values utilising these coding systems.

3.2 Automatic session and information management for stateful personalised service

The application is designed to offer a personalised service in order to meet envisaged concurrent individual demand which varies from session to session and enables session and state of use tracking. The approach of creating and using a session identifier and implementing session management plays an important role in realising the concurrent personalised service of the application. A session identifier is used to create a temporary folder on the server specific to each use session, which stores HTML DOM objects and intermediary and final products of map data files, automatically generated at various stages of that particular session. Each use session consists of a number of stages/steps with a user making choices and selections to query into the databases, and commanding geoprocessors to perform (Figure 3).

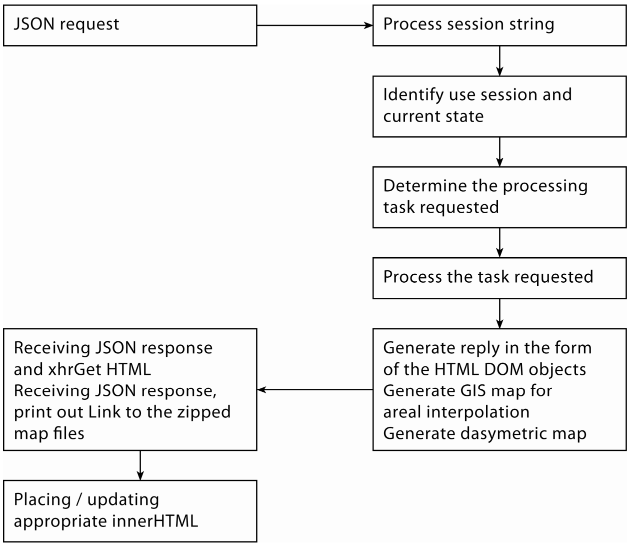

JSON request
Process session string
Identify use session and current state
Determine the processing task requested
Process the task requested
Generate reply in the form of the HTML DOM objects Generate GIS map for areal interpolation Generate dasymetric map
Receiving JSON response and xhrGet HTML Receiving JSON response, print out Link to the zipped map files
Placing / updating appropriate innerHTML

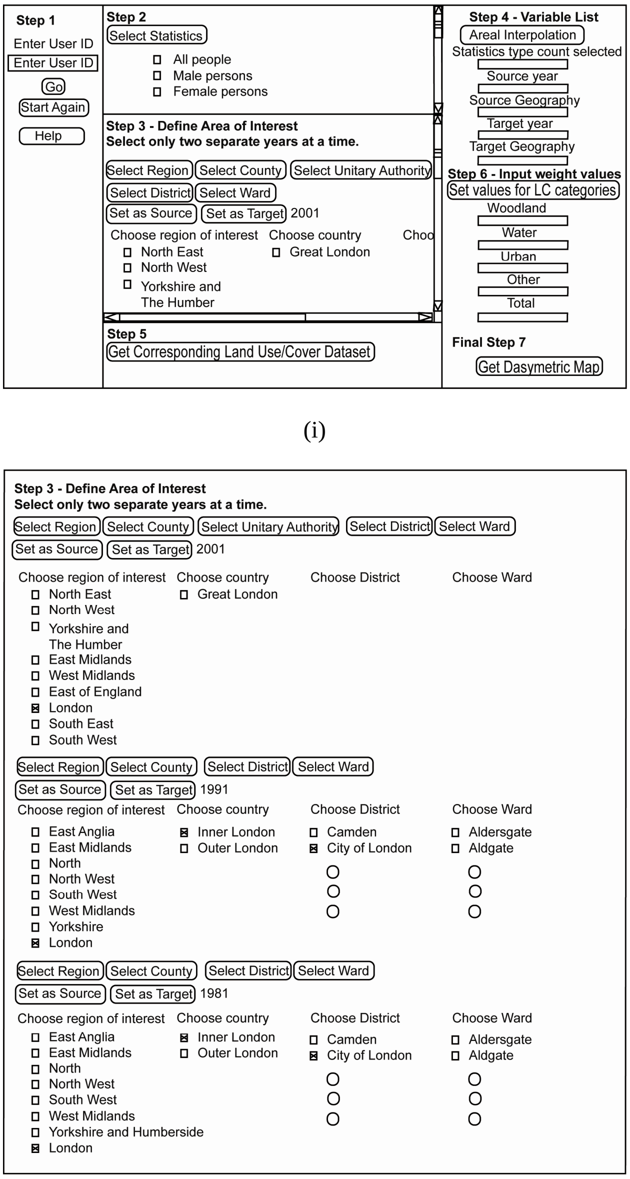
Figure 3: Application flow chart for stateful automatic generation of responses in the form of HTML DOM objects

The system is a secure service system, prior approval is required and approved registered user will be given a user identifier. Only with issued user identifer, user can trigger off a service-side geoprocessor to create a personalised session folder to store all dynamically generated intermediatory files and geoprocessed results.

3.3 Design of front-end user interface, accessing and consuming geoprocessing services

The front-end user interface is designed in the form of HTML webpage allowing user access with commonly available internet browsers over the internet. HTML Document Object Model (HTML DOM) and Javascript scripting was made to produce user interface and interactive feedback messaging. Javascript functions contained in the user interface HTML document creates instances of geoprocessors and consumes REST services and monitor the progress and status of REST services through ESRI's Javascript API. Ajax data exchange employs ESRI's Javascript API and dojo toolkit functions in client-side browsers to interact with REST services of ArcGIS Server technology. Dojo functions are embedded in the Javascript to undertake JavaScript Object Notation (JSON) Ajax calls to ArcGIS Server's REST services to invoke geoprocessors and monitor their progress (Figure 3) (ArcGIS Resource Centre; dojotoolkit). The webpage is partitioned into a number of interactive functional sections as HTML divs (Figure 4) and the programming of Javascript functions enables information interchange between sections and validates user interaction and input to ensure the necessary procedure is followed.

The system presents users with historical British census databases (statistical and spatial) reflecting the development of such data resources over recent enumerations. It allows users to make selections of geographical areas of interest at various levels (region, county, unitary authority, district, ward and output areas) and key population statistical accounts dynamically (Figure 4).



(i)

(ii)

Figure 4. Schematic illustration of (i) KAGIS user interface and (ii) sub-window of dynamically generated listings of geographical area names

3.3.1 Dynamic re-presenting historical heritage of census data resources and user monitoring of selection

The service offers automatic listing of geographical names at the different levels in HTML markup language with unique identifier codes to provide users with a means of interactively making choices for querying into and searching geographical databases (Figure 4). These listings for 2001, 1991 and 1981 censuses respectively are presented as a stack in sections as HTML divs to allow interaction with back-end databases for areas associated with 2001, 1991 and 1981 censuses. Users' querying into these databases and selection of census areas are processed in parallel and the service system automatically generates HTML DOM objects and presents these as innerHTML.

Applying the mechanism described above, the content of large databases is unleashed and presented to users for dynamic interrogation into databases. Once users have selected areas of interest at a particular level, the required statistics, the source and target zones and years are specified as variables to be fed into a server-side geoprocessor. These are dynamically presented in the variable sub-window (Step 4 in Figure 4) for user monitoring of selection of population statistics variable, source geography and target geography, ready to be submitted to server. These variables/parameters are passed onto server-side geoprocessors to search and extract requested data from distant servers, which allows the user requested data to be obtained in preparation for geoprocessing.

3.4 Automatic spatial data extraction and downloading as per dynamic user request

The Edina's geospatial dataset service produces server-to-server responses that are OGC compliant Web Feature Service (WFS) responses. WFS supports either GML or zipped ESRI Shape File package, as returns for service requests. The dataset service used by KAGIS is known as UKBORDERS, sponsored by the Economic and Social Research Council (ESRC) and Joint Information Systems Committee (JISC), U.K..

On demand calling and extracting of necessary map datasets are achieved by means of a specially designed Python module callable and executable by a server-side geoprocessing script. It feeds into the distant server with search patterns of census zone codes obtained from the front-end user selection for fetching source and target GIS datasets.

The geoprocessing script programmatically takes the responses of zipped files, saves these into the folder specific to the session and automatically unzips them by importing and executing a Python unzip module to perform to make available datasets for further processing.

In the event of user's multiple selections of wards, the geoprocessing script automatically detects the number of datasets to be fetched and the process is iterated until completion. Thereafter, the shape files are merged automatically into unified source and target maps ready for further processing.

3.5 Design for runtime generation of comparable population statistics

Three arrays of newly researched and developed equations for calculating comparable population statistics for 1981, 1991 and 2001 census datasets from data extracted at runtime from Mimas are designed and contained in a Python module to be dynamically called and executed.

3.6 Automatic extracting statistical data and calculating comparable population statistics as per dynamic user request

Mimas broadcasts server-to-server responses by using Geolinked DATA Access Service specification of Open GIS consortium Inc. (2004) and its service responses are in the form of OGC-compliant XML streams. The specific dataset service used is known as Casweb as a census programme sponsored by the Economic and Social Research Council (ESRC) and Joint Information Systems Committee (JISC), U.K.. An *ad hoc* Python module has been developed and implemented in the KAGIS system to make service call to the Mimas service with values of census zone identifiers and population statistical variables requested by user, read into memory the XML stream and is callable and executable in the geoprocessing script. A high performance Python XML parsing module (lxml) is employed and imported to perform extraction of data contained in XML into a Python list and extracted data are automatically inserted into an automatically created column in the database associated with the source ESRI map shape file.

Within the geoprocessing script, automatic decision making is devised and implemented to check against lists of statistical variables which need to calculated for comparable population statistics. If calculation is required for a specific statistical variable contained in the list, an *ad hoc* calculation function is invoked to call and execute the corresponding equation contained in the arrays of equations to produce values of comparable statistics for the variable prior to automatic creation of the column and automatic insertion of data values.

As a number of statistical variables are involved in calculation, the geoprocessing script is devised to be capable of automatically deciding on the number of iteration of multiple service calls and data extraction to be made and carry out calculation of comparable statistical values.

3.7 Automatic spatio-temporal interpolation with areal interpolation

A specially designed geoprocessing script for areal interpolation transfers population statistical data from the source map to the target map and automatically generates a new resultant map.

Once a user has selected the source zones and required statistics associated with a particular census year and the target zones for a certain year (e.g. total male population, source: in 2001 output areas; and target 1991 enumeration districts), upon clicking on the button (Step 4 in Figure 4), user triggers off server-side geoprocessor to perform tasks described in 3.3, 3.4, 3.5 and 3.6 and the task of areal interpolation.

3.8 Preparing landuse map with the same geographical coverage

Large landuse GIS maps derived from remotely sensed imagery are stored in a dedicated folder of the KAGIS server. Maps are vectorised polygons of land cover categories derived from the Ordnance Survey (OS) Strategi® dataset. A dedicated geoprocessing service is published for consuming in the front-end HTML to be activated by clicking the button (Step 5 in Figure 4). The geoprocessor extracts from large landuse datasets for the UK on KAGIS and produces the landuse map with the coverage of the map created by the above areal interpolation.

3.9 Empowering census analysts with decision making in assigning weights to landuse categories

Users are presented with opportunities to make decision on weights to be used and click a button to confirm (Step 6 in Figure 4). The weights assigned overwrite the default preset values.

3.10 Automatic generation of quasi-realistic spatial population distribution count models (digital dasymetric maps)

Once the areal interpolated map is produced, landuse map with the same geographical coverage is prepared and user's decision on weights is applied on the landuse categories, dasymetric mapping geoprocessing module can be triggered off to perform the task upon user's click on an *ad hoc* button (Step 7 in Figure 4). The Javascript function associated with the button submits the service request and commands the server-side dasymetric mapping geoprocessor to produce quasi-realistic spatial population distribution count model incorporating landuse information.

The algorithm used in the dasymetric mapping geoprocessing script has been described in Mennis and Hultgren (2006) and it produces dasymetric GIS dataset as actual spatial population distribution count model, by applying the principle illustrated in Figure 2.

3.11 Instant delivery of spatial statistical models processed with automatic spatial data handling

Immediately upon completion of dasymetric mapping geoprocessing, Python zipping function imported into the geoprocessing script zips the session folder and creates a zipped package file on the server. The front-end Javascript function detects the completion of the server-side geoprocessing task and prints out the internet address link to the automatically generated zip file as innerHTML. All user needs to do is to click on the link, download the zip file and save it locally for examination and analysis.

4 CONCLUSION AND IMPLICATION

KAGIS innovates the ways of digital data requisition, data processing, digital GIS model generation and service delivery as a fully integrated and automatic service and represents a landmark in the development of census-related services to the U.K Higher Education sector.

The newly available KAGIS service will enable inter-censal comparison of spatial population models and analysis of demographic and socio-economic change to be undertaken at any of existing spatial scales and of any area of interest in the United Kingdom. The system allows users to obtain digital quasi-realistic spatial population distribution models at any existing spatial scales and of any area of interest by utilising any of 1981, 1991 and 2001 census boundaries.

The service system is designed to be extensible to offer fully processed digital dasymetric models for the forthcoming 2001 census (Martin, 2000) to offer the most up-to-date digital quasi-realistic spatial population distribution models for a multitude of purposes such as quasi-realistic service and development planning. KAGIS will enable development of advanced research and teaching in a wide range of areas such as quantitative demography and population geography, regional science, spatial economics and quantitative social policy.

With the process of globalisation and development, all countries have been and are undergoing significant development and change. The technology and know-how experience of KAGIS can assist any country in the world to develop its secure internet GIS services, integrating geographical information for specific decision making systems as well as providing low-cost and efficient ways to process and utilise existing geographical data for a multitude of purposes such as development planning.

ACKNOWLEDGEMENTS

The research conducted in this paper forms a project titled *Web-based Spatio-temporal Interpolator for Census Area Statistics* which has received financial support from ESRC (RES-348-25-0012). The work is based on data provided with the support of ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-LINE Consortium. The work is based on 2001 Census Output Area Boundaries, Crown copyright 2003. Crown copyright material is reproduced with the permission of the Controller of HMSO. The work is based on an Ordnance Survey/EDINA supplied service, © Crown Copyright/database right 20(01).

REFERENCES

ARCGIS RESOURCE CENTERS, JAVASCRIPT APIs. http://resources.esri.com/help/9.3/arcgisserver/apis/javascript/arcgis/help/jshelp_start.htm (accessed 1st August 2008)

Dojotoolkit. http://www.dojotoolkit.org/ (accessed 1st August 2008)

DORLING, D., REES, P., 2003, A nation still dividing: the British Census and social polarisation 1971-2001, *Environment and Planning A*, **35:** 1287-1313.

DORLING, D., MARTIN. D., MITCHELL, R., 2003, *Linking censuses through time*, project website, http://census.ac.uk/cdu/software/lct/ (accessed 30 January 2006).

DOYLE, A. REED, C., (Eds.), 2001, OGC, *Introduction to OGC Web Services, OGC interoperability white paper*. Available online at: www.opengis.org (accessed 26 June 2006).

EICHER, C.L., BEWER, C. 2001 Dasymetric mapping and areal interpolation: implementation and evaluation, *Cartography and Geographic Information Science*, **28**: 125-138.

FLOWERDEW, R., GREEN, M., 1992, Developments in areal interpolation methods and GIS, *Annals of Regional Science*, **26**: 67-78.

FLOWERDEW, R., GREEN, M., 1994, Areal interpolation and types of data. In FOTHERINGHAM, S., ROGERSON, P. (eds.) *Spatial Analysis and GIS.* London: Taylor & Francis, pp. 121–145.

GARDELS, K., 1997, *The Open GIS Approach to Distributed Geodata and Geoprocessing*, Open Geospatial Consortium. http://www.opengis.org (accessed 1st August 2008)

GEDDES, A., GIMONA, A., ELSTON, D.A., 2003, Estimating local variations in land use statistics, *International Journal of Geographical Information Science*, **17**: 299-319.

GOODCHILD, M.F., ANSELIN, L., DEICHMANN, U., 1993, A framework for areal interpolation of socioeconomic data, *Environment and Planning A*, **25**: 383-397.

GREGORY, I.N., ELL, P.S., 2006, Error-sensitive historical GIS: Identifying areal interpolation errors in time-series, *International Journal of Geographical Information Science*, **20**: 135-152.

HESS, D.B., 2007, Transformation of spatial data to a new zone system: a survey of US metropolitan planning organizations, *Environment and Planning B – Planning and Design*, **34**: 483:500.

LANGFORD, M., 2006, Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers Environment and Urban Systems*, **30**: 161-180.

LANGFORD, M., 2007, Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps, *Computers Environment and Urban Systems*, **31**: 19-32.

LONGLEY, P.A., GOODCHILD, M.F., MAGUIRE, D.J., RHIND, D.W., 2001, *Geographical Information Systems and Science*. (Chichester, UK; Wiley).

MACKANESS, W., TOWERS, A., 2002, Handling and accessing census boundary data. In REES, P., MARTIN, D., WILLIAMSON, P. (Eds.) *The Census Data System*. (Chichester, UK; Wiley). pp. 85-95.

MARTIN, D., 1998a, Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science*, **12**: pp. 673-685.

MARTIN, D., 1998b, Census geography 2001: designed by and for GIS? In CARVER, S. J. (Ed.) *Innovations in GIS 5*. (London; Taylor and Francis), pp. 198-209.

MARTIN, D., 2000, Towards the geographies of the 2001 UK Census of Population, *Transactions of the Institute of British Geographers*, **25**: 321-332.

MARTIN, D., 2003, Extending the automated zoning procedure to reconcile incompatible zoning systems, *International Journal of Geographical Information Science*, **17**: 181-196.

MENNIS, J., HULTGREN, T., 2006, Intelligent dasymetric mapping and application to areal interpolation, *Cartography and Geographic Information Science*, **33**: 179-194.

NORMAN, P., REES, P., BOYLE, P. 2003, Achieving data compatibility over space and time: creating consistent geographical zones, *International Journal of Population Geography*, **9**: 365-386.

OFFICE FOR NATIONAL STATISTICS, 2001, *Administrative area boundary changes in England and Wales between the 1991 Census and the 2001 Census.* Available online at: http://www.statistics.gov.uk/geography/downloads/ONSG_user_guide_2001.pdf (accessed 25 June 2006).

OPEN GIS CONSORTIUM INC., 2004. *Geolinked Data Access Service (GDAS)*. Open GIS Discussion Paper.

REES, P., MARTIN, D. 2002, The debate about census geography. In REES, P., MARTIN, D., WILLIAMSON, P., (Eds.), 2002, *The Census Data System*. (Chichester, UK; Wiley). Pp. 27-36.

REIBEL, M., BUFALINO, M.E., 2005, Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems, *Environment and Planning A*, **37**: 127-139.

SHROEDER, J.P., 2007, Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data, *Geographical Analysis*, **39**: 311-335.

W3C, 1999. HTML 4.01 Specification. W3C Recommendation. http://www.w3.org/TR/html401/ (accessed 1st August 2008)

WALFORD, N., & SHI, S., 2009: Web-based Spatio-temporal Interpolator for Census Area Statistics. 2011 Census research: new data, linkage and outputs. The Royal Statistical Society, London. 13 May 2009.

WILSON, T., REES, P. 1998. *Lookup tables to link 1991 population statistics to 1998 local government areas.* Working Paper 98/5 School of Geography, University of Leeds, wpaper/wp98-5.pdf.

A DYNAMICALLY LOAD AND UNLOAD ENABLED AGGREGATION MODEL BASED ON THE WPS

JianBo Zhang [a, b], JiPing Liu [a], Bei Wang [c]

[a] Research Center of Government Geographic Information System, Chinese Academy of Surveying and Mapping, 100039,Beijing, China
[b] School of Resource and Environmental Science, Wuhan University, 430079,Wuhan,China
[c] Institute of Geographic Sciences and Natural Resources Research, 100101,Beijing,China

KEY WORDS: SDI, Aggregation model, logic object layer, Geoprocessing chains

ABSTRACT:

The processing of web-based spatial data is an important issue in the field of Spatial Data Infrastructures (SDI). After analyzing the architecture of the fundamental conception and application supported by the OGC Web Processing Service specification, a common loading and unloading enabled aggregation model is designed, this provides the ability to conflate of geoprocessing service chains dynamically. Common processes to be pushed in the model, the relationship and the intercommunion among geoprocessing chain have to be offered in term of uml sequence. The experiments represent lightweight spatial analysis solutions could be construct flexibly.

1. INTRODUCTION

The rapid development of Networking and distributed computing technologies evolve geographic information system into the common geospatial services. Web services can be simply defined as the information related software entities running in network environment which provides special user to meet the needs of specific information or processing capabilities. With the characteristic of self-contained, and self-describing, geospatial services do not depend on the context or state of other services (Anders Friis-Christensen, 2006). The collection of spatial services has created a technology evolution that moves from standalone GIS applications towards a more loosely coupled and distributed model based on interoperable GI services (Díaz.L, 2007, Granell.C, 2007), Furthermore, most users of traditional GIS systems use only a small percentage of their systems' functionalities; the services model provides users with just the services and data they need, without having to install, learn, or pay for any unused functionalities, that is the motivation of the SDI which concentrate on the interoperability between spatial resource and the applications among the spatial domain.

Recently, traditional spatial data infrastructures had combined various kinds of geospatial services mainly include WMS (Web Mapping Services), WCS (Web Coverage Services), and WFS (Web Feature Services). (OGC,2004a; OGC, 2004b; OGC, 2005b; OGC, 2005c; OGC, 2005d) we can get the meta spatial information from SDI, also view and mapping the data online, the SDI not only provides the functionality on downloading the data but also provides an open spatial data access interface and so on. Thus currently SDI's main focus lies on distributed data storage in the form of spatial services, the retrieval through catalogues, and the visualization in form of web map services. Nevertheless, going on with the geoprocessing service applications in-depth, the functionality of SDI cannot meet the processing and modelling requirements of special users (Caldeweyher D, 2008). At the same time, geospatial services distinguish form business services, While traditional business workflows are oriented towards document processing, task management and control-flow, spatial processing workflows typically are data- and/or compute-intensive, dataflow-oriented, and often involve data transformations, analysis, and simulations. In order to complete a spatial analysis function normally, the task needs to call multiple spatial information services, access to many different types of spatial data through multiple steps, sometimes even repeat the steps. Thus, the need for adaptable interfaces and tools for accessing scientific data and executing complex analyses on the retrieved data has risen in a variety of disciplines (e.g., geology, biology, ecology). The mechanism for such assembly of services is often referred to as service chaining (Alameh N, 2003), the process of combining or pipelining results from several complementary services to create customized applications. GIS services have specific middleware requirements that current Web service technologies can only partially meet.

In this paper, firstly, we present a architecture of aggregation model describes composing the geoprocessing services for disaster assessment. In the model, we present the spatial resource such as shp file, raster, OGC services to be the objects we call layers based logic object conception. The model provides uniform access to the vast amount of spatial data and highly heterogeneous services based on open standards and Internet in the SDI environment. the model that We propose with a plug-in method, on describing the dynamics of their work processes to build geoprocessing chains to demonstrate how within our framework, services can be composed into scientific workflows and executed to perform scientific tasks.

The paper proceeds as follows: Section 2 discusses SDI and geoprocessing services. The design of the aggregation model contributes to services chains that enable to load and unload dynamically. The implementation presented in Section 5 by illustrating a specific application scenario. The paper ends with the main conclusions and an outlook for planned work.

2. SDI AND GEOPROCESSING SERVICES

2.1 Geospatial service

Geospatial Services encapsulates the spatial data access, retrieval, processing and analysis functions and be depicted through a simple and convenient way (xml) for users to provide a unified interface-oriented SOA (service-oriented architecture). spatial information services with the pattern range of the discovery, binding, execution gradually replace the tight coupling pattern between components. Its data and functional characteristics of loosely coupled build the right conditions for spatial information access and the distributed processing. In order to achieve full interoperability among spatial data, functions and software, OGC, and W3C work together to develop uniform standards. The most widespread standards adopted in web oriented spatial Information Technology are including WMS, WFS, WCS, WPS, GML(Geography Markup Language) (Anders Friis-Christensen, 2008).

The Web Map Service (WMS) is responsible for generating dynamically maps from geospatial data, custom map for the "Space Geographic Information Drawn into a suitable screen display digital image files. it can display and integrate various layers of geographic datasets onto the same map. A WMS provides a standardized access to maps rendered in a format such as PNG, GIF or JPEG by using the operations getMap and getFeatureInfo.

The Web Feature Services (WFS) provide access to the distribution of vector based data and support the function of insert, update, and delete, search and discovery geographical elements. The WFS returns a Geography Markup Language (GML) document, which is an XML grammar to express geographical features.

The Web Coverage Services (WCS) provide the location information or attribute contains in a raster layers, rather than accessing a static map. According to HTTP protocol, client requests to send the appropriate data, including images, multi-spectral images and other scientific data, examples of different data formats supported by a WCS are: DTED, GeoTIFF, or NITF.

WPS defines a standardized interface that facilitates the publishing of geospatial processes, and the discovery of and binding to those processes by clients describes process operation that returns a description of a process including inputs and outputs and the execute operation that performs the calculations and returns the result.。

2.2 SDI orients geoprocessing service

Above the standardized base for GI service we present take the SDI share the spatial data、model、and the applications for user one step further facilitate access to distributed, heterogeneous geospatial data through a set of policies, common rules, and standards that together help improve interoperability (Efrat Jaeger, 2005) Currently, SDI open standards predominantly support the download, retrieval and visualization of spatial data, The following diagram briefly illustrates the architecture of traditional SDI.

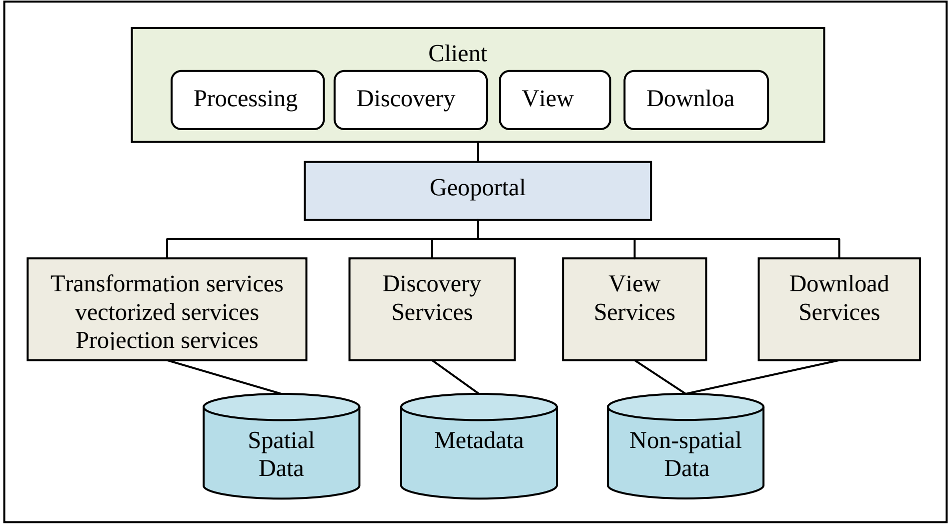


Figure 1 the traditional SDI architecture

From the figure we can see the traditional architecture of the SDI is divided into three layers. The data tier contains metadata directory libraries, geospatial data stored in the space-based database, and the professional sectors based on the topic of application. Service tier mainly provides discovery services, basic spatial data processing services, browsing services, and spatial data download service. In which discovery services called the space directory service can be extracted from the spatial metadata database from the semantic level. The basic spatial data processing services provide projection transformation, coordinate transformation and basic spatial processing such as map algebra. View services provide WMS mapping services, and WFS services enable to extract the features. Download services provide all levels of users enable to download online data; users do not need CD-ROM, and mobile storage devices to copy the spatial data they required.

However, basic data retrieval and visualization services, far from meeting the needs of professionals or specific user. Special users often want to analysis, processing, modelling spatial data in special areas, and convenient to resolve a complex geo-related issues, that is the in-depth applications about geoprocessing services. The next step of SDI not only need to access to the data in a unified standard interface, at the same time need to address the specialized data analysis to extract information, even in the service chains. However, the standard geoprocessing services traditional SDI used are static or independent. Lonely geoprocessing service or the simple aggregation of multiple services cannot take on the complex geographical analysis tasks. Therefore, the dynamic assembly of geoprocessing services is significance for SDI.

3. AGGREGATION PROCESSING MODEL

Aggregation processing model aims to address the complex spatial analysis tasks among the spatial domain. Modelling of the geoprocessing services can be deemed to assembly the service according to sequence about the tasks. We call the sequence service chain which is defined as a sequence of services where, for each adjacent pair of services, occurrence of the first action is necessary for the occurrence of the second action.(Di, 2004a) When services are chained, they are combined in a dependent series to achieve larger tasks.

Here we introduce the logic geo-object layer concepts benefit to construct the service chains. At same time a three-tier aggregation model architecture is presented (Figure 2).

3.1 Logic Geo-object layer concepts benefit to chains

Here, we first discuss the issue of granularity about spatial objects. Spatial objects granularity describes the thickness level of spatial objects at different scales and at different spatial coordinate system. In this article we discuss the granularity of the space object is the atomic-level involved in spatial data, space services, and space model of objects, reflects the independence degree of the spatial objects (Di, 2004a), presents a conception of a geo-object consider a granule of geoinformation, which consists of data itself, a set of attributes (metadata), and associations with a set of methods (transformation and creation methods) that can operate on it. At the same time describes the all geoinformation and knowledge products are derived from archived geo-objects based on the principle of object-oriented.

However, all the geoinformation defined as spatial objects can be derived from the spatial data is not enough to reality. For example, we can abstract a user's request to be a spatial object; although the request can be broken down into many independent geo-objects, however, the irrelevance between the spatial objects the need to binding by the rules-driven library.

Here, we define logic geo-object layer concepts base on the geo-object. all the space features, spatial data, space services such as WMS,WFS,WCS,WPS involved in the spatial processing will be as the independent atomic-level logic object layers comparison with the actual spatial entities stored in the logic container. At the same time the model extracts the logic object layer will involve in the geoprocessing from the user’s request as well as aggregation services to complete the operation of the semantic rules.

In order to build a dynamic aggregation model, there is no inheritance and derived relations among the atomic logical object layer, in other words, atomic object are stored in parallel, there is no hierarchy. However, there is a hierarchy between the spatial objects derived from the logical object layer and the atomic logical object layer. This service model can give a high degree of reconstruction between spatial objects. Atoms logic object layers can be seen as the smallest particle size of spatial objects. Due to atomic logical object layer has its own properties and methods that can be seen as plug with the ability of dynamically load or unload, and be generated according to user request. Thus, the aggregation model enables great flexible and reusable. Base on the management on logic object layers, we design the aggregation model. Figure 2 presents a simplified view of the model architecture. In which a range of resource tier prepare spatial data sets and spatial services and processing tier processes the resources and client applications are composing a variety of standards GI services and the show the results.

3.2 Resource tier

The resource tier consists of distributed data set and services container with different types of data and storage systems. at the top of the resource tier is the logic Object container， groups data and services instances in basic functional categories(see Figure 2). Our system model has focused mainly on processing services, but our processing engine also accesses other resource types, just like topics application data，in order to facilitate access to both geospatial data processing and to discovery, viewing and download of data.

Spatial and non-spatial data container handles data access according to the available database or file system where the spatial data type range of shp file 、raster file 、GML and Poi or statistics data sets. Data services container contains most of the spatial information services in the OGC standards like WMS、WFS、WCS. Professional users not only can access symbolic maps from the WMS services directly, but also can extract interested spatial elements from the WFS, view its spatial properties and other properties. Geoprocessing services can perform complex computations on geospatial data. Which range of coordinate transformation, Rasterize service, vectorize service and so on. In our model processing service design based on wrapping. we have identified atomic spatial data formats as well as its spatial reference information, adopt a uniform coordinate transformation to the aggregation spatial information services, if necessary。

The logic Object container manages all kinds of spatial data, spatial-based services; provide logic level data protection for space aggregation engine as the core resource container of spatial system model. For the spatial processing engine is concerned, it does not need to know the path of physical storage address for the current spatial object, even do not need to know the type of the spatial object, only need to know the uniquely identify of the logic object layer. As a logical layer of spatial objects has its independent methods and properties can be used as plug-ins。

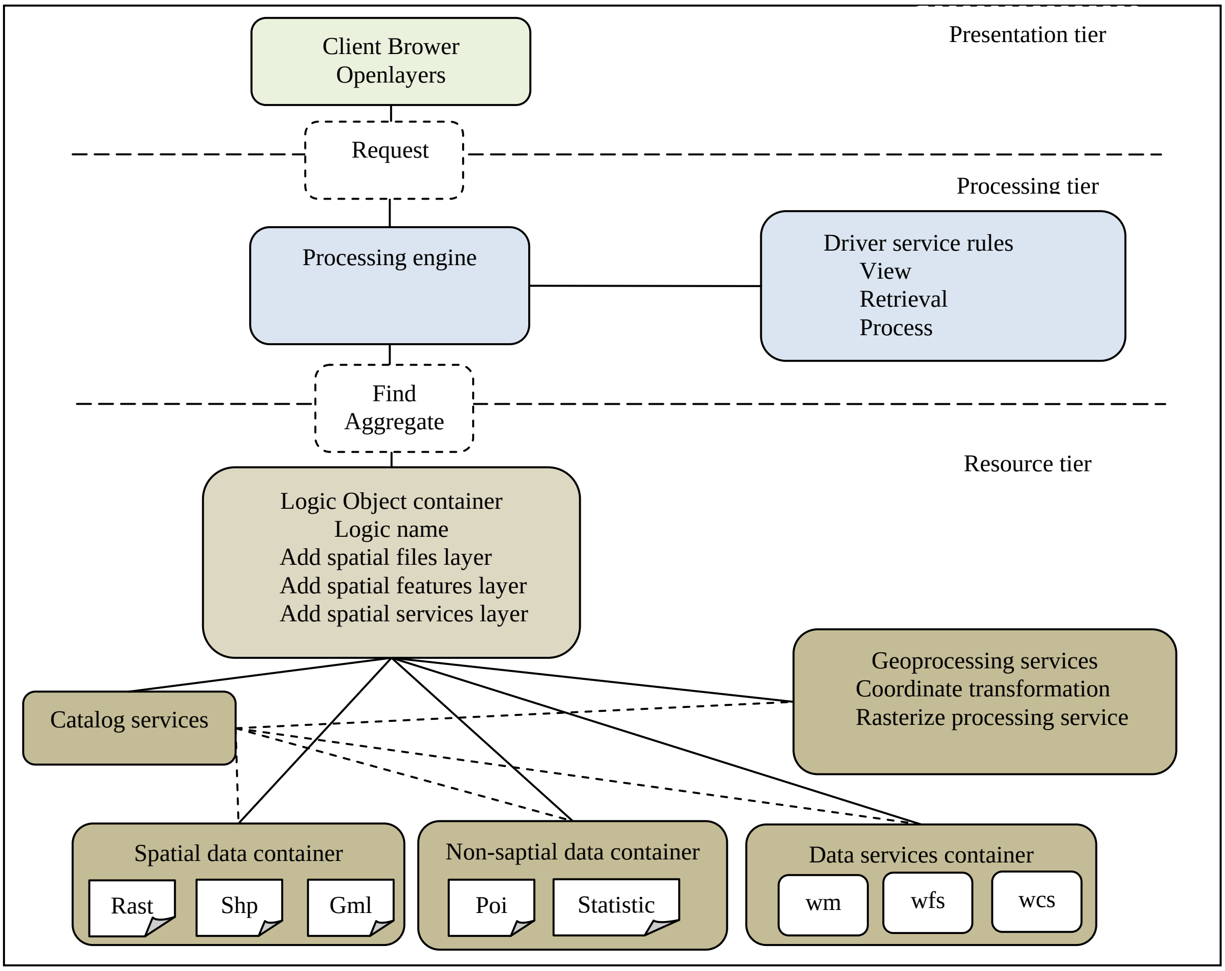


Figure 2. Aggregation model architecture

3.3 Processing tier

The Processing tier contains the business logic of the aggregate model, will be the core module of the whole model. We can take the processing engine as the brain in the whole model; in fact it is a set of GI libraries, takes analysis and decomposition of user-space process the request, while driving rules and to find the logical object layer, finally accomplishes spatial processing tasks. Space rules memory a lot of space predicate, such as browsing, searching, extraction, overlay, intersection, and buffer analysis and so on. It inducts a logical object-oriented layer, as well as behavioral constraints, which is a key node in aggregate geospatial services.

Essentially rules are many rule operators to provide semantic support and interface support; it will automatically decompose user input into a series of logical object layer and the suitable interactive computing actions for the user selection. At the same time the corresponding WPS semantic services. While receiving a series of logical object layer and the corresponding algorithms, spatial Processing engine began to query logic object layer in logic Object Layer Manager.

If it exists, the logical object layer will be instantiated according to the object-oriented thinking, then the model start to find atomic-level spatial objects from the directory services, and load the space resources. If the logic of the object does not exist, model structure it by the logic of space object metadata content which the rule presented. Thus, the task of geoprocessing service aggregation finished due to the specific rules lead the logical object layer, while processing engine accept the result produced by logical object layer with processing spatial data structure, and in the form of GML passed to the client to carry out the display. Refer to this, spatial aggregation model to complete its life cycle.

3.4 Presentation tier

As the entrance of the aggregation model provides the portal for users to access the data and services provided by our web application. Also accept user-input requests of completing specific spatial information processing and submit it to processing tier. Our client application provides one-stop discovery, access, aggregation and spatial processing results displaying of spatial information services use pure JavaScript and Ajax(Asynchronous JavaScript and XML) method Is a typical thin-client model， For data visualization, we have used the Openlayers API ,一种开源的客户端组件 an open-source client-side components for building the user-interface part of the mapping in our application .

4. IMPLEMENTATION

Based on the aggregation model architecture is depicted on Figure 2, we launch an implementation in order to gain the information of typhoon called Wipha had destroyed the GDP of the china in the neighborhood sea. This is a case of disaster assessment for making decision as shown in the figure 3.

Gong with the model needs to access varied data resource via WMS, WFS, and WCS services, besides a catalog service and four different geoprocessing services need to be involved. For the disaster assessment application we have access to the following thematic data, and services which are potentially important for disaster statistics:

- dist_county400, which is a region of the entire county in china. We use it to show the district of typhoon called Wipha traversing.
- typhoon_2007, which is a POI reference data set for the path that the typhoon called Wipha through over.
- Image _GDP_2007, which provides the properties of the GDP of china in 2007
- Image _POP_2007, which provides the properties of the population of china in 2007
- Place names used for locating specific geographic area based on geographic name input.
- The catalog service provides metadata of the various data services.
- The coordinate transformation service is responsible for transforming coordinates into requested
- The spatial intersection service is responsible for gain the region of intersection area.
- The spatial overlay service is responsible for gain the distinct code of current area.
- The GDP and population statistics service is responsible for calculating the areas affected by typhoon

Firstly, geoportal application obtains a user's request through the client application, and sent the user's request to the space rules, which is equivalent semantic decomposition unit. The model query space resources directory According to the spatial logic objects layer and spatial rules. The test under the rule generated six kinds of logical object layer, namely the background WMS layer, the user professional data we called POI published in WFS layer, distinct feature organized in WFS layers, spatial intersection WPS layer and the spatial overlay WPS layer, and the statistic WPS layer. Here, we descript rules as the assessment disaster model; it can accept four parameters, namely the region, distinct code, feature typed KML, and the poi feature. From the WFS service we can gains the distinct code and the costumed data with location information we called poi. From the spatial intersection WPS and the spatial overlay WPS we can handle the region of POI across and return the KML for showing mapping. Here we use code to extract the administrative divisions created SLD logic object layer, as WMS layer added to the client, if the pilot through the figure. The disaster statistics service is invoked after selecting all parameters necessary and a request for data is made. If necessary, coordinates are transformed into projected coordinates and then the assessed area statistics are returned and visualized as a table in the client.

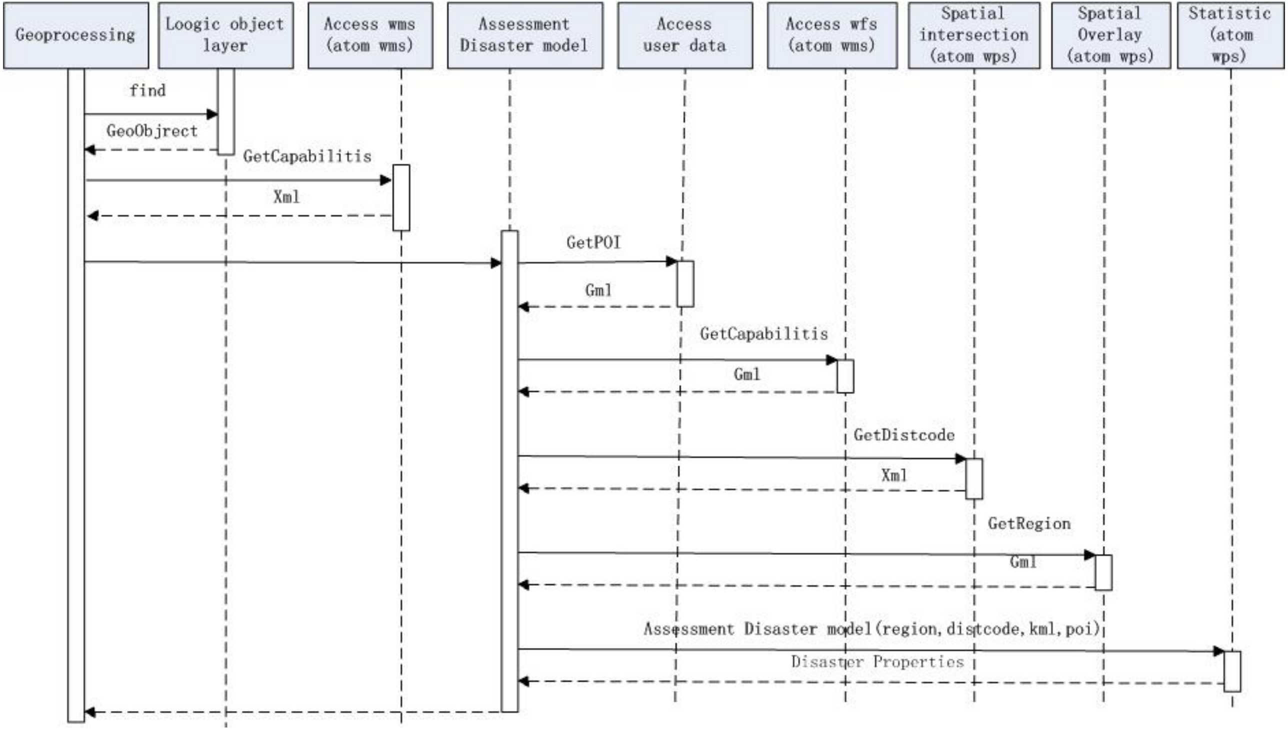
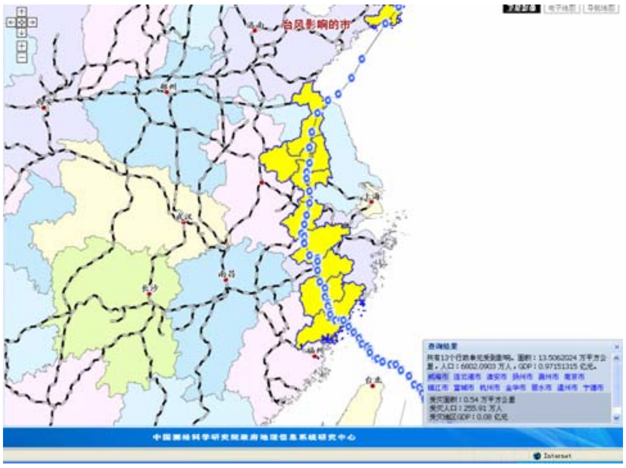


Figure 3.UML sequence of the geoprocessing flows



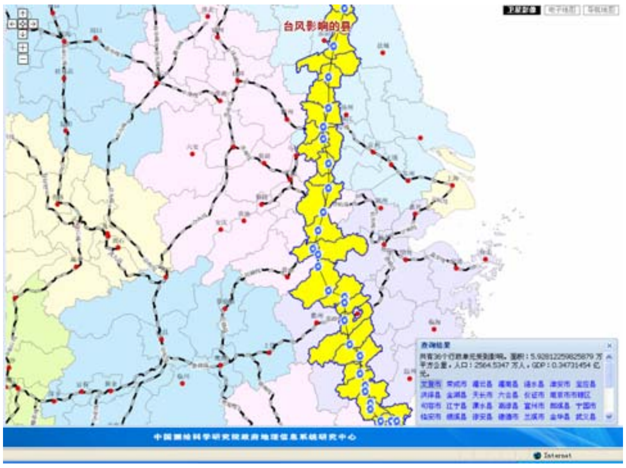
查询结果

共有13个行政单元受到影响。面积：13.5062024 万平方公里，人口：6802.0903 万人，GDP：0.97151315 亿元。

[illegible] 连云港市 淮安市 扬州市 湖州市 南京市 镇江市 宣城市 杭州市 金华市 丽水市 温州市 宁德市

受灾面积：0.54 万平方公里
受灾人口：255.91 万人
受灾地区GDP：0.08 亿元

Figure 4. The GDP losses by typhoon



查询结果

共有36个行政单元受到影响。面积：5.9281225982587 9 万平方公里，人口：2564.5347 万人，GDP：0.34731454 亿元。

[illegible] 天长市 六合县 仪征市 南京市市辖区 江宁县 宣州市 宁国市 临安市 兰溪市 金华县 武义县 [illegible]

Figure 5. The typhoon effect on the lives of people

Figure 4 and figure 5 show us the GDP losses and effect on the lives of people by typhoon via executing the geoprocessing chains that we defined in the model.

5. CONCLUSIONS

Spatial information services technology provide a good solution for spatial information interoperability and knowledge discovery, spatial-based facilities use these standardized spatial information services access by browse-getting of spatial data services forward to spatial information processing and modelling phase. In this paper, we leads the geo-object concept of m.cloud, proposed concept of plug-in logic object layer, enable spatial services dynamic into an aggregate spatial information service chain, and illustrates the flexibility of the solution through a simple Disaster Assessment experiment.

The spatial information service chain built can be applied in simple spatial analysis and disaster information extraction, did not concern other expertise area, not enough to support the complex spatial information processing model. Meanwhile, the intermediate processing results of spatial information service chain will be re-instantiated and loaded as logic object layer, will once again participate as an input parameter when the spatial information service chain operations the difficulties, the future development will focus on solving the deeper modelling applications in the Other areas of expertise of spatial information service chain, and researches of building complex spatial information service chains.

REFERENCES

Alameh N (2003). Chaining Geographic Information Web Services. IEEE Internet Computing 7(5):22-29

Caldeweyher D, Zhang J, Pham B (2006). OpenCIS-Open Source GIS-based web community information system. International Journal of Geographical Information Science 20: 885-898

Díaz.L, Costa.S, Granell.C, Gould.M(2007). Migrating geoprocessing routines to web services for water resource management applications. In Proceedings of 10th AGILE Conference on Geographic Information Science (AGILE 2007), Aalborg (Denmark)

Granell.C, Díaz.L, Gould.M(2007). Managing Earth Observation data with distributed geoprocessing services. In Proceedings of the 6 International Geoscience and Remote Sensing Symposium (IGARSS 2007), Barcelona (Spain)

Díaz,L., Granell,C., Gould,M.(2008). Case Study: Geospatial Processing Services for Web-based Hydrological Applications. In Geospatial Services and Applications for the Internet Era. Springer, New York.

Anders Friis-Christensen, Lars Bernard, Ioannis Kanellopoulos, Javier Nogueras-Iso, Stephen Peedell, Sven Schade, Cathal Thorne.Building Service Oriented Applications on top of a Spatial Data Infrastructure – A Forest Fire Assessment Example .the 9th AGILE Conference on Geographic Information Science, Visegrád, Hungary, 2006

Efrat Jaeger, Ilkay Altintas, Jianting Zhang, Bertram Ludäscher, Deana Pennington, William Michener. A Scientific Workflow Approach to Distributed Geospatial Data Processing using Web Services. Proceedings of the 17th international conference on Scientific and statistical database management 2005, Santa Barbara, CA June 27 - 29, 2005

Di, 2004a. GeoBrain-A Web Services based Geospatial Knowledge Building System. In Proceedings of NASA Earth Science Technology Conference (ESTO 2004). Palo Alto, CA. June 22-24, 2004. CD-ROM, 8p.

OGC, 2004a: Open Geospatial Consortium Inc. Catalogue Services Specification, v2.0. OGC 04-021r2.

OGC, 2004b: Open Geospatial Consortium Inc. Web Map Service specification 1.3. OGC 04-024

OGC, 2005a: Open Geospatial Consortium Inc. Catalog Services Specification 2.0-ISO19115/ISO19119 Application Profile for CSW 2.0. OGC 04-038r2

OGC, 2005b: Open Geospatial Consortium Inc. Web Feature Service Implementation Specification1.1. OGC 04-094

OGC, 2005c: Open Geospatial Consortium Inc. Web Processing Service. Discussion paper. OGC 05-007r4

OGC, 2005d: Open Geospatial Consortium Inc. OGC Web Services Common Specification. OGC 05-008

AN INDEXING METHOD FOR SUPPORTING SPATIAL QUERIES IN STRUCTURED PEER-TO-PEER SYSTEMS

Lingkui Meng [a], Wenjun Xie [a, *], Dan Liu [a, b]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] Department of Information Technology, Huazhong Normal University, Wuhan 430079, China
(lkmeng, xiewenjun)@whu.edu.cn, rock_fall@tom.com

Commission VI, WG VI/4

KEY WORDS: Peer-to-Peer, Spatial Queries, Overlap Minimization, Spatial Index, Tree Structure, Distributed Optimization

ABSTRACT:

To provide the efficient supporting spatial data queries in peer-to-peer systems has recently received much attention. Most proposed methods tried to use hop count to represent the transmission delay, and the total message count to estimate the cost of query processing. For the ignorance of the differences between DHT lookups and spatial queries, and distinction between physical networks and overlay networks, the efficiency and cost of their query processing can't be indicated properly. In addition, their experimental results are achieved by using point data sets, while the fact that the overlap of spatial objects usually exists in real applications is not considered, and it may cause multi path query processing and then results in plenty of peers visiting and routing messages. In this paper, we propose an indexing method which efficiently supports spatial queries in structured peer-to-peer systems. It adopts an overlap minimization algorithm which takes the query rate of data into account to reasonably reduce the holistic cost of queries. We also introduce a dynamically adaptive distributed optimization scheme that dynamically adapting to the time-varying overlay architecture and data usage concerns. Theoretical analysis and simulation results both indicate that our method is efficient and effective.

1. INTRODUCTION

With the rapid growth and increased importance of distributed spatial data, there is an increasing need for massive spatial data sharing in large-scale distributed systems. While the P2P (Peer-to-Peer) systems have become a powerful means for data sharing in Internet community. For their potential uses in spatial data sharing, to provide the efficient supporting spatial data queries in peer-to-peer environments has recently received much attention.

There are two types of P2P overlay networks: unstructured ones in which requests are broadcasted or routed through flooding or random walks, such as Gnutella (Ripeanu, 2001), KaZaA (Leibowitz, 2003), eDonkey (Tutschku, 2004), Freenet (Clarke, 2000), and structured ones based on DHTs (Distributed Hash Tables) in which requests are routed using routing tables, such as Chord (Stoica, 2003), CAN (Ratnasamy, 2001), Tapestry (Zhao, 2004), Pastry (Rowstron, 2001). Since they use flooding or random walks based methods for processing queries request, which results in a large number of messages and a traffic overhead, so lead to the poor efficiency of queries and the bad scalability for the systems, and these restricted their rapid development. On the other hand, it is very easy to process a query by using the assigned key in structured P2P networks, and their scalability is very good. So the structured P2P systems are more appropriate for handling data sharing. But for the reason of using DHTs, which destroy the semantics of the data objects, only exact key match queries can be supported efficiently, and it isn't an easy task to support spatial queries in structured P2P systems.

The traditional indexing methods for spatial databases can be briefly classified into two approaches: spatial sorting-based, such as Hilbert space filling curve (Bially, 1969), Z-ordering curve (Orenstein, 1986), and spatial contains relationship-based, such as R-tree (Guttman, 1984), R+-tree (Sellis, 1987), R*-tree (Beckmann, 1990). Similarly, the methods that support spatial data queries in structured P2P systems also have two kinds of implements: the one that maps multidimensional spatial data into one-dimensional by using order-preserving hash function, and the other one that distribute tree data structure in P2P environment. The main problem of former approach is the spatial relationships between spatial objects often may be destroyed, so leads to the inefficiency of the queries. This is because there are no any functions can always preserve the spatial properties. As for the latter method, a critical performance issue is the tree structure has to be queried in a top-down manner from the root node, so the communication bottlenecks are more likely to happen on the peers that take charge of the tree nodes at higher level, especially for root node, and it is also a single point of failure. For their good efficiency in centralized environment and the bad performance of the former approach, using hierarchical tree structure is a better choice. VBI-Tree (Jagadish, 2006) solved the above problem in latter approach by introduce a new routing table design using sideway index links, and DPTree (Li, 2006) handled it by propose tree branch oriented distribution.

* Corresponding author.

Although some proposed methods have achieved good results. There still exist some issues that are not considered carefully. 1) The differences between DHT lookups and spatial queries. In a DHT lookup, the query request is forwarded along a single path, while it may be routed over multiple paths for spatial query. The multiple peers visiting sometimes can't be avoided. However, we can decrease the number of multiple peers visiting. 2) The distinction between physical networks and overlay networks. So the hop count can not really reflect the transmission delay.

To address the above two issues, we propose a suite of efficient solutions, which can be used to support any kind of hierarchical tree architecture overlay. Our paper makes the following two major contributions. 1) An innovative definition of overlap is first introduced. Here we take the following properties of the systems into consideration: the non-uniform and time-varying properties of spatial data distribution and their popularity, and peer interests also are different and time-varying. Then we present an overlap minimization algorithm to minimize the number of peers need to visit for process a query. 2) We propose an efficient distributed optimization algorithm to guarantee each peer has neighbors that are physically close to it in the underlying network, and it can continuously and efficiently optimize the overlay structure under dynamic network conditions.

The rest of the paper is organized as follows. Section 2 surveys previous work, focusing on spatial data queries in structured P2P systems. In section 3 we propose the definition of overlap and the overlap minimization algorithm. The distributed optimization algorithm is explained in section 4. The experimental results of the above design, using many metrics, such as routing hop count, number of messages, distribution percentage of delay time of peers, are presented in section 5. Finally, we discuss the conclusions and future work in section 6.

2. RELATED WORK

There has been a substantial amount of research on spatial data queries in P2P systems. As mentioned, most current proposed methods generally can be briefly classified into two categories: one is the reduction of multidimensional spatial data to one dimension, then the current DHTs can be directly used to support spatial queries, and the other is distribution of spatial contains relationship-based indexing structure.

The former category includes SCRAP (Ganesan, 2004), MAAN (Cai, 2004), PRoBe (Sahin, 2005). SCRAP uses a two-step solution to partition the data space. In the first step, it map multi-dimensional spatial data down to one-dimensional by using a space-filling curve, then the one-dimensional data could be range partitioned across the dynamic available peers. MAAN supports multi-attribute range queries through multiple single-attribute resolution by using locality preserved hashing to map a range of data space to Chord, and the efficiency may be very poor. PRoBe uses a multi-dimensional logical space and maps data items onto the space based on their attribute values, and the space is divided into hyper-rectangles, with each maintained by a peer within the system.

In the latter category, VBI-Tree is an abstract data structure build based on a virtual binary balanced tree structure. It was inspired by BATON (Jagadish, 2005) structure where each peer corresponds an internal node and a leaf node. DPTree uses a tree branch oriented distribution method to distribute the tree structure among peers in a way preserving the good properties of balanced tree structures yet avoiding single points of failure and performance bottlenecks.

It is known that tree structures are very difficult to distribute in P2P systems, because searching the tree by following paths induces an uneven load on tree nodes at higher level. The above two methods solve the problem by introduce some novel designs. However, the efforts only aim to support zero dimensional data queries in multi-dimensional data space, and none of them consider the situations of multi-dimensional data in multidimensional data space. More specifically, the current methods can not efficiently handle the queries about line data or polygon data, since they omitted the overlap between adjacent tree nodes, which results in the multiple peers visiting in distributed systems. In addition, we should also try to keep each peer has neighbors that are physically close to it in the underlying network.

3. OVERLAP MINIMIZATION

For the existence of overlap in tree nodes, there will be multiple paths need to be followed even for point data queries, which results in a large number of messages and plenty of peers visiting. We therefore first give our definition of overlap in P2P systems, and then present an overlap minimization algorithm.

3.1 Definition of Overlap

In the X-tree method (Berchtold, 1996), the following two definitions of overlap are given.

If a tree node contains *n* hyper rectangles {R$_1$, …, R$_n$}, the overlap can formally be defined as

$$Overlap = \frac{\left\| \bigcup_{i,j\in\{i\ldots n\},i\neq j} (R_i \cap R_j) \right\|}{\left\| \bigcup_{i\in\{1\ldots n\}} R_i \right\|} \tag{1}$$

where $\|A\|$ denotes the volume covered by A.

For the distribution of spatial data is nonuniform, the modified definition took this into account. That is

$$WeightedOverlap = \frac{\left| \{ p \mid p \in \bigcup_{i,j\in\{i\ldots n\},i\neq j} (R_i \cap R_j) \} \right|}{\left| \{ p \mid p \in \bigcup_{i\in\{1\ldots n\}} R_i \} \right|} \tag{2}$$

where $|A|$ denotes the number of data elements contained in A.

Obviously, the popularity between data elements is different. A more accurate definition of overlap needs to take the query rate of data elements into account. So we propose the following new definition.

$$NewOverlap = \frac{q_p \times \left|\{p \mid p \in \bigcup_{i,j\in\{i\ldots n\},i\neq j}(R_i \cap R_j)\}\right|}{q_{all} \times \left|\{p \mid p \in \bigcup_{i\in\{1\ldots n\}} R_i\}\right|} \quad (3)$$

where p is a data element
q_p denotes the query rate of *p*, which is the number of received queries about *p* during the history period.
q_{all} denotes the query rate of all data elements, which is the number of received queries about all data elements stored in the local peer during the history period.

As data element popularity and peer interest is time-varying, we use exponential moving average technique to calculate the value of query rate, which give more weight to the observations in most recent periods without discarding other values, rather than directly using the observation values during the entire history period. So the new formula for calculating query rate is as follows.

$$\overline{q}_{curr} = w \times \overline{q}_{prev} + (1-w) \times q_{curr} \quad (4)$$

where $\overline{q}_{curr}$ is current valid value for query rate
$\overline{q}_{prev}$ is the previous valid one
q_{curr} is the current observed one
$w \in [0,1]$ is a constant value which represents a weight factor value for new observation.

3.2 Overlap Minimization Algorithm

If an overlap will appear on one peer, we should choose to allow its existence or adjust the position of some influenced peers to minimize the overlap. The latter action will be taken only when the benefits brought about by the existence of overlap is less than its cost. If the overlap occurred in a peer, and it will affect *n* data elements {d$_1$, d$_2$, …, d$_n$}, with the fractions of area of the overlap to them are {f$_1$, f2, …, f$_n$}, and the average cost to process query about these data elements are {c$_1$, c$_2$, …, c$_n$}. We also assume it will cost some extra system resources SR to allow the existence of the overlap. If the benefits of the overlap less than its cost, then we keep the overlap, or adjust some peers. That is

$$\sum_{i=1}^{n} f_i \times q_i \times c_i < SR \quad (5)$$

where q$_i$ is the query rate of d$_i$.

4. DISTRIBUTED OPTIMIZATION ALGORITHM

We define the delay time of a peer *p* as ***D(p)***, which is the sum of latencies to all its neighbours. Its definition can formally be defined as

$$D(p) = \sum_{n\in neighbors(p)}^{|neighbors(p)|} L(p,n) \quad (6)$$

where *p* is a peer
neighbours (p) is the set of *p*’s neighbours
n is one of its neighbours
L(p, n) is the latency of peer *p* to *n*.

It is easy for us to know that the optimum situation is to keep the minimum value of total delay time $D_{sum}(p)$ of all peers *S* involved in the system.

$$D_{sum}(p) = \sum_{p\in S}^{|S|} D(p) \quad (7)$$

where *S* is the set of all peers involved in the system

Since in the P2P systems it is impractical to calculate the above value, we propose a distributed optimization algorithm that is an iterative algorithm that each peer executes periodically and uses integer linear optimization to minimize the sum of the delay time of a peer and one of its neighbours.

The algorithm is executed periodically on two adjacent peers in the system, which we call them *seeds*. Firstly, they mutually exchange the routing table of their neighbours, and then each seed peer measures the latencies to the neighbours of the other seed peer. Finally, we can determine whether or not to swap their neighbours based on the measured values. The function that we want to minimize during the iteration of the algorithm is as follows.

$$\sum_{p\in Seeds}^{|Seeds|} D(p) = \sum_{p\in Seeds}^{|Seeds|} \sum_{n\in neighbours(p)}^{|neighbours(p)|} L(p,n) \quad (8)$$

where *Seeds* are two adjacent peers in the system

It can be proved that global convergence can be achieved but proofs are omitted due to space limitations.

5. EXPERIMENTAL RESULTS

For the evaluation of the above two algorithms we did compare the experimental results between our optimized method and the original one in a tree structure overlay. Since the structure here didn’t consider the properties of the physical network, we use an Internet node latency measurement results (Wong, 2005) from Meridian project in Cornell University. As there is only point data simulator component in the design, we also use a spatial data generator (University of Piraeus, 2006) to create polygon data, which are Zipf distribution.

We test the network with different number of nodes N from 100 to 2500. For each test, 50 point queries are executed, and then the average value is taken. In Figure 1 we present the results of number of messages to locate data to process point queries for point data and polygon data, and the number can be used as a metric of scalability of the system. As we observe that our algorithm could reduce the query processing cost for polygon data, but nearly the same as original algorithm for point data.

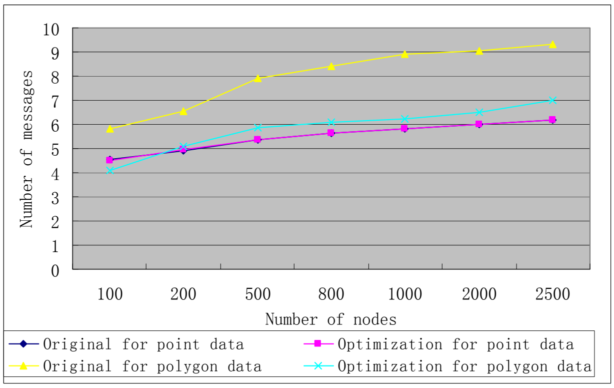


Figure 1. Number of message comparison

In Figure 2 we demonstrate the number of routing hops comparison, which can reflect the efficiency of a system. Similar as the above results, the figure indicated that our algorithm could significantly improve the efficiency for polygon data, but not for point data

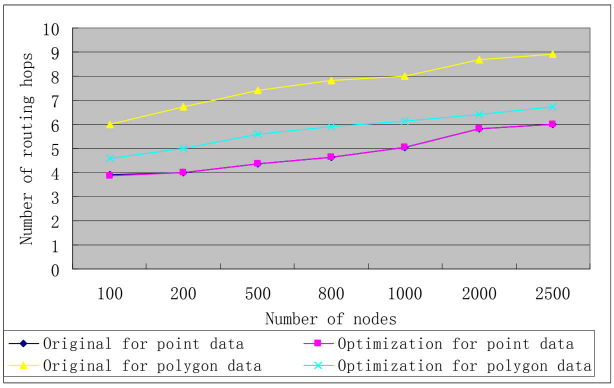


Figure 2. Number of routing hops comparison

Figure 3 shows the results of distributed optimization algorithm for keep adjacent peers in physical network as neighbors in overlay network, which prove that the method has good locality properties for tree-based overlay networks.

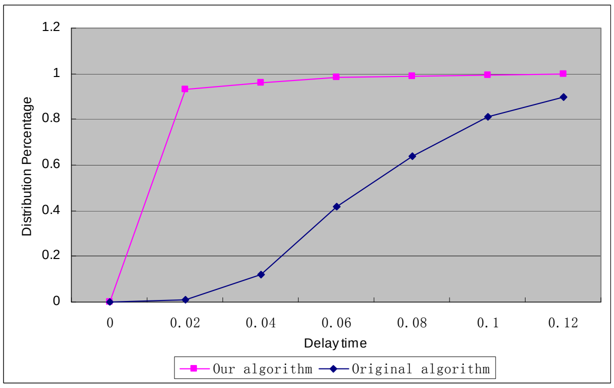


Figure 3. Distribution percentage of delay time of peers

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed our indexing method for supporting spatial queries in P2P systems, which allows us to reduce the cost of routing messages and improve the efficiency of routing hops. Additionally, by recognizing that hop count can not reflect the actual time required for processing queries, we also focus on the reduction of the delay time of each hop besides the hop count, and hence decrease the total time. The method is based on two newly proposed algorithms: overlap minimization algorithm and distributed optimization algorithm.

For the future, we will augment our method to include other efficient query algorithms, such as range query and kNN query. Upon completion of this work, we also plan to run comprehensive performance evaluation on our method.

ACKNOWLEDGEMENTS

The research is supported by Natural Science Foundation of China (40771177).

REFERENCES

Ripeanu, M. 2001. Peer-to-Peer Architecture Case Study: Gnutella Network. 1st *International Conference on Peer-to-Peer Computing (P2P 2001)*, Linköpings universitet, Sweden, pp. 99-100.

Leibowitz, N., Ripeanu, M., Wierzbicki, A. 2003. Deconstructing the Kazaa network. 3rd *IEEE Workshop on Internet Applications (WIAPP 2003)*, San Jose, California, USA, pp. 112-120.

Tutschku, K. 2004. A Measurement-based Traffic Profile of the eDonkey Filesharing Service. 5th *Passive and Active Network Measurement Workshop (PAM 2004)*, Antibes Juan-les-Pins, France, pp. 12-21.

Clarke, I., Sandberg, O., Wiley B., et al. 2000. Freenet: A Distributed Anonymous Information Storage and Retrieval System. *Workshop on Design Issues in Anonymity and Unobservability 2000*, Berkeley, California, USA, pp. 46-66

Stoica, I., Morris, R., et al. 2003. Chord: a scalable peer-to-peer lookup protocol for Internet applications. *IEEE/ACM Transactions on Networking*, 11(1), pp. 17–32.

Ratnasamy, S., Francis, P., et al. 2001. A Scalable Content-Addressable Network, *ACM SIGCOMM 2001*, San Diego, California, USA, pp. 161-172.

Zhao, B. Y., Huang, L., Stribling, J., et al. 2004. Tapestry: A Resilient Global-Scale Overlay for Service Deployment, *IEEE Journal on Selected Areas in Communications*, 22(1), pp. 41-53.

Rowstron, A. I. T., and Druschel, P. 2001. Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems. *18th IFIP/ACM International Conference on Distributed Systems Platforms*. Heidelberg, Germany, pp. 329-350.

Bially, T. 1969. Space-Filling Curves: Their Generation and Their Application to Bandwidth Reduction. *IEEE Transactions on Information Theory*, 15(6), pp. 658-664

Orenstein, J. 1986. Spatial Query Processing in an Object-Oriented Database System, ACM SIGMOD International Conference on Management of Data, Washington, D. C., USA, pp. 326-336,

Guttman, A. 1984. R-trees: A dynamic index structure for spatial searching. *1984 ACM SIGMOD Internation Conference on Management of Data*, Boston, Massachusetts, USA, pp. 47-57.

Sellis, T., Roussopoulos, N., Faloutsos, C. 1987. The R+-Tree: A Dynamic Index for Multidimensional Objects. *13th International Conference on Very Large Data Bases (VLDB 1987)*, Brighton, England, pp. 507-518

Beckmann, N., Kriegel, H. P., et al. 1990. The R*-tree: An efficient and robust access method for points and rectangles. *ACM SIGMOD 1990*, Atlantic City, New Jersey, USA, pp. 322-331.

Jagadish, H. V., B. C. Ooi, et al. 2006. VBI-Tree: A Peer-to-Peer Framework for Supporting Multi-Dimensional Indexing Schemes. *22nd International Conference on Data Engineering (ICDE 2006)*. Atlanta, Georgia, USA, pp. 34.

Li, M., W.-c. Lee, et al. 2006. DPTree: A Balanced Tree Based Indexing Framework for Peer-to-Peer Systems. *14th IEEE International Conference on Network Protocols (ICNP 2006)*. Santa Barbara, California, USA, pp. 12-21.

Ganesan, P., Yang, B., et al. 2004. One Torus to Rule them All: Multidimensional Queries in P2P Systems. *7th International Workshop on the Web and Databases (WebDB 2004)*, Paris, France, pp. 19-24.

Cai, M., Frank, M., Chen, J., et al. 2004. MAAN: A Multi-Attribute Addressable Network for Grid Information Services. *Journal of Grid Computing*, 2(1), pp. 3-14.

Sahin, O. D., Antony, S., Agrawal, D., et al. 2005. PRoBe: Multi-Dimensional Range Queries in P2P Networks. *6th International Conference on Web Information Systems Engineering (WISE 2005)*, New York City, New York, USA, pp. 332-346.

Jagadish, H. V., Ooi, B. C., Vu, Q. H. 2005. BATON: A Balanced Tree Structure for Peer-to-Peer Networks. *31st International Conference on Very Large Data Bases (VLDB 2005)*. Trondheim, Norway, pp. 661-672.

Berchtold, S., Keim, D. A., Kriegel, H. 1996. The X-tree: An Index Structure for High-Dimensional Data. *22nd International Conference on Very Large Data Bases (VLDB 1996)*. Mumbai, India, pp. 28-39.

Wong, B., Slivkins, A., Sirer, E. G. 2005. Meridian: A Lightweight Network Location Service without Virtual Coordinates. *ACM SIGCOMM 2005*, Philadelphia, Pennsylvania, USA, pp. 85-96.

Meridian Project, http://www.cs.cornell.edu/People/egs/merid-ian/

University of Piraeus. 2006. http://www.rtreeportal.org/

Cooperative Information Augmentation in a Geosensor Network

Malte Jan Schulze, Claus Brenner, Monika Sester
Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Germany
Appelstraße 9a, 30167 Hannover
{maltejan.schulze, claus.brenner, monika.sester}@ikg.uni-hannover.de

Keywords: geosensor network, simulation environment, cooperation, data acquisition, distributed processing, mobile, mapping quality, meteorology

This paper presents a concept for the collaborative distributed acquisition and refinement of geo-related information. The underlying idea is to start with a massive amount of moving sensors which can observe and measure a spatial phenomenon with an unknown, possibly low accuracy. Linking these measurements with a limited number of measuring units with higher order accuracy leads to an information and quality augmentation in the mass sensor data. This is achieved by distributed information integration and processing in a local communication range.

The approach will be demonstrated with the example where cars measure rainfall indirectly by the wiper frequencies. The a priori unknown relationship between wiper frequency and rainfall is incrementally determined and refined in the sensor network. For this, neighboring information of both stationary rain gauges of higher accuracy and neighboring cars with their associated measurement accuracy are integrated. In this way, the quality of the measurement units can be enhanced.

In the paper the concept for the approach is presented, together with first experiments in a simulation environment. Each sensor is described as an individual agent with certain processing and communication possibilities. The movement of cars is based on given traffic models. Experiments with respect to the dependency of car density, station density and achievable accuracies are presented. Finally, extensions of this approach to other applications are outlined.

1. INTRODUCTION

Geosensor networks are composed of a possibly large number of individual sensors with measuring, positioning and communication capabilities. Through local cooperation of neighboring sensors the whole network is able to perform actions that go beyond an individual sensor's capabilities and achieve a common global goal. In this way the geosensor network is able to acquire information about the environment in an unprecedented detail.

Geosensor networks mark a paradigm shift in measuring systems in two ways: from centralized to decentralized data acquisition, and from a separation of measurement and processing to integrated acquisition and analysis.

The advantages of geosensor networks lie in their scalability and also in their fault tolerance, as the role of individual sensors is not crucial - due to the high redundancy. These properties lead to a large number of applications of geosensor networks e.g. in environmental monitoring or in military.

From a computational and geoinformatics point of view, the challenge is to devise algorithms that are able to work locally and still achieve a common global solution. There are many spatial algorithms that operate in a centralized manner, presuming access to all the information; however, in the case where a local processing unit only has a limited view of the surrounding information, existing algorithms have to be adapted or new ones have to be devised to achieve a decentralized processing.

1.1 Prerequisites of our approach

Sensors can have different capabilities. In our approach, we start with the assumption that the cooperation of a large number of sensors of similar, but limited, quality and a few sensors with higher quality can lead to an enrichment of the poor quality measurement of the limited sensors. The measurements are integrated and accumulated in a Kalman Filter and thus – over time – lead to a higher accuracy of the sensed information.

1.2 Problem statement

Rainfall is the most important information source for hydrological planning and water resources management. Especially the modelling of high dynamic processes like floods and erosion rely on high resolution rainfall information. For this measurement, non-recording stationary gauges exist, which measure with a daily observation interval. These instruments are typically available in a high density (e.g. in Germany 1 station per 90 km^2). The density of recording rain stations is still inadequate (e.g. in Germany 1 station per 1800 km^2).

The idea of our approach is to densify the number of stations using unconventional sensors, which are massively available and can measure rainfall (at least approximately), namely cars: when it rains, car drivers start their wipers in order to clean the windshields. Thus, starting the wipers is an indication for liquid on the windshield; the frequency of the wiper is related to the amount of rainfall. The exact relation between wiper frequency and rainfall is unknown, however, it can be calibrated on-the-fly using measurements from the environment: on the one hand, if a car passes by a recording rain station; on the other hand, if a car passes by another car, which has been calibrated at a rainfall station recently. Thus, by locally exchanging and accumulating the measurements, the quality of the a priori unknown information, namely the amount of rainfall, can incrementally be determined and refined.

1.3 Approach

We simulate traffic and rainfall using a real road network. Traffic is simulated by generating random routes on the road network; the rainfall is simulated by generating a raincloud. Cars move in this environment and measure rainfall with their wipers. The initial coarse rainfall measurement quality of each car is iteratively improved through local cooperation of moving cars and rainfall stations.

1.4 Overview of the paper

After a description of related work, we will introduce our approach to the above described problem in section 3. We describe our simulation environment and the implementation of the Kalman filter. In section 4, examples are shown which verify the results. Section 5 gives a brief summary and an outlook on future work.

2. RELATED WORK

A general overview of wireless sensor networks is given in (Akyildiz et al., 2002). Geosensor networks for the observation and monitoring of environmental phenomena are a recent trend in GIScience. Traditional geodetic networks consist of a fixed set of dedicated sensors with a given configuration and measurement regime. The processing of the data is usually done in a centralized fashion. The advent of geosensor networks brings about the chance to move from a centralized approach to an approach using distributed sensors with computation and communication capabilities (Stefanidis & Nittel 2004).

The advantages as opposed to a centralized system are its scalability, and its high spatial and temporal resolution. In order to fully exploit a geosensor network in the way described, methods for local information aggregation have to be devised. Such methods have to take the neighbourhood and the communication range of the individual sensors into account. There are many application areas for geosensor networks, e.g. environmental observations (Duckham & Reitsma, 2009), surveillance, traffic monitoring and new multimodal traffic (Raubal et al., 2007).

Decentralized algorithms for geosensor networks have been investigated by several researchers and for different applications. Laube et al. (2008) describe an algorithm to detect a moving point pattern, namely a so-called flock pattern. A flock is described as a group of objects that moves in a certain distance over a certain time. In a similar spirit, Laube & Duckham (2009) present a method for the detection of clusters in a decentralized way. Depending on the communication range, clusters of a certain size (radius) can be detected.

Walkowski (2008) presents an approach for the optimal arrangement of geosensor nodes in order to correctly describe an underlying temporally varying phenomenon, like a toxic cloud. He assumes to have sensors that are able to move; however, the determination of the locations of lacking information has to be determined in a centralized fashion. Zou & Chakrabarty (2004) describe an approach to optimally cover an area with a given set of sensors. Sester (2009) presents an approach for cooperative detection of a boundary of a spatial phenomenon using a mobile geosensor network.

For traffic simulation there are programs that simulate not only the movements of the traffic objects on the infrastructure, but also the behaviour and the decisions of the users. For a consistent modelling of these aspects agent based approaches are used, where each traffic participant is modelled individually (Raney & Nagel, 2006).

In terms of fusing measurements in an optimal way, Kalman filtering is a widely employed technique, which is described in standard textbooks (Brown & Hwang, 1997, Simon, 2006).

The principle applicability and suitability of our approach has been investigated earlier by Haberlandt & Sester (2009). There, the main focus was to explore the quality of the interpolation taking different traffic densities and given wiper-rainfall-relationships into account.

3. APPROACH

3.1 Basic concept of simulation environment

The main objective in our work is to describe the quality of rain measurement using cars as rain gauges. In opposition to rain measurement stations that can record the rainfall data directly by using dedicated rainfall sensors, the cars in our approach do not have such sensors. We consider the wiper frequencies of a car as correlated to the rainfall intensity. When the intensity is high, one would switch the wiper frequency of the car to a high value in order to have a better visibility. When there is no rainfall at all, the wipers of the car would not be used.

The cars are considered as sensor nodes that can measure their position (for example via GPS) and their wiper frequency. In addition, they can perform calculations based on the locally collected data and share them with other cars using a wireless communication device (see Fig. 1).

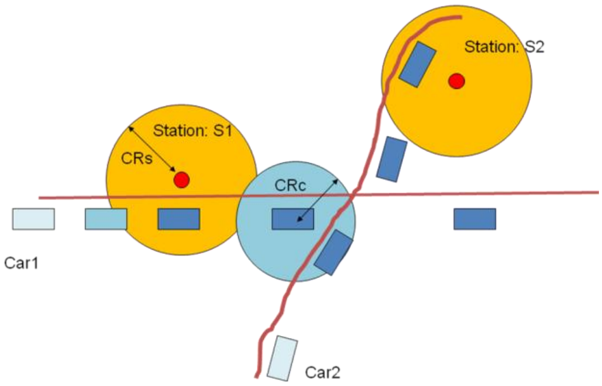


Fig. 1: Communication between cars and stations, with communication ranges CRc and CRs, respectively.

In order to determine the intensity of the rainfall from the wiper frequency information, we need a functional relationship between the wiper frequency and the rainfall intensity, otherwise the collected wiper frequency data of a car leads to a very uncertain estimation for the rainfall intensity. To simulate this case, we give cars without any information about the functional relationship a high standard deviation.

To provide high quality rain measurement data, a few weather stations, that can measure the rainfall intensity with a very high certainty, are distributed across our road network. The cars can use those high quality data, to improve their own certainty about the rainfall measurement.

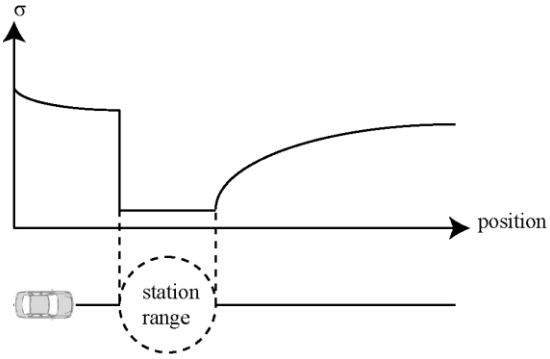


Fig. 2: Improvement of the certainty of a car by communication with a weather station.

As shown in Fig. 2, the standard deviation decreases rapidly, when the car enters the communication range of a weather station, leading to a high certainty of rainfall measurements from the car. When the car leaves the communication range, the certainty gently decreases until it reaches the original level. While decreasing, the car can still share its information with other cars that are not in the range of a weather station, helping to improve their level of certainty.

3.2 Implementation of simulation environment

Car movement
The simulation environment describes an agent based system, where each car is considered as an agent that follows a certain trajectory through a road network. We determine the movement of the cars by randomly selecting start- and endpoint of each trajectory. The movement through the road network is calculated using the A*- algorithm to determine the shortest path. The visited nodes of the road network are saved together with a timestamp. The simulation itself is based on a central start- and end time with constant time steps of 10s. For each step, the position of all cars is calculated by using a linear interpolation between two nodes.

Rainfall simulation
The rainfall intensity in our simulation environment is modelled by a mixed Gaussian with randomly distributed centers. The calculated field is normalized. The calculation of the Gaussian is based on (1). The result for the simulated raincloud is shown in Fig. 3. For this simulation, the rainfall intensity is considered to be stationary.

$$cloud(x,y) = \frac{1}{2\pi\sigma} \cdot e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} \qquad (1)$$

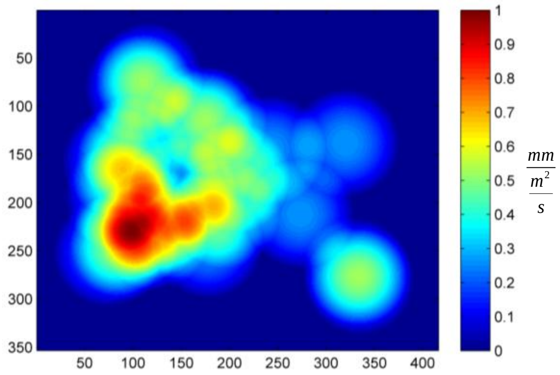


Fig. 3: Simulated distribution of rainfall intensity.

Observation of rainfall and communication strategy
For each car, a Kalman filter is implemented to describe the system state **x** and its quality $\mathbf{\Sigma}_{xx,k}^{+}$ (2).

$$\mathbf{x}_k^{+} = \begin{bmatrix} \dot{x}_k^{+} \\ \ddot{x}_k^{+} \end{bmatrix} \rightarrow \mathbf{\Sigma}_{xx,k}^{+} = \begin{bmatrix} \sigma_{\dot{x},k}^{+\;2} & 0 \\ 0 & \sigma_{\ddot{x},k}^{+\;2} \end{bmatrix}, \mathbf{\Sigma}_{ww} = \begin{bmatrix} \sigma_{w\dot{x}}^2 & 0 \\ 0 & \sigma_{w\ddot{x}}^2 \end{bmatrix}$$
$$\mathbf{x}_{k+1}^{-} = \dot{x}_k^{+} + \vartriangle t_k^{k+1} \cdot \ddot{x}_k^{+} \quad \rightarrow \quad \mathbf{\Phi_k^{k+1}} = \begin{bmatrix} 1 & \vartriangle t_k^{k+1} \\ 0 & 1 \end{bmatrix} \qquad (2)$$

The system state consists of two variables $\dot{x}$ and $\ddot{x}$. The rainfall intensity is described by $\dot{x}$, which can be considered as the rainfall speed, having the unit $mm/m^2/s$. It can be determined from the wiper frequencies of a car and is directly observed by a weather station. As the cars move underneath the stationary rainclouds, a second parameter $\ddot{x}$ is estimated, which describes the change of the rainfall intensity, having the unit $mm/m^2/s^2$. The certainty of the system state is described by the covariance matrix $\mathbf{\Sigma}_{xx,k}^{+}$. The covariance increases with the time passed, as the system noise $\mathbf{\Sigma}_{ww}$ accumulates. To make a statement about the quality of the rainfall measurement, we focus on the standard deviation $\sigma_{\dot{x},k}^{+}$ of the rainfall intensity. To predict the system state in the next epoch k+1, the transition matrix $\mathbf{\Phi_k^{k+1}}$ is used. This is a standard transition matrix usually employed for the estimation of object positions using the assumption of constant speed. To update the system state with observations, three different cases of communication are taken into account:

1. The car is located outside the communication range of other cars and stations. In this case, there is no data exchange. The car determines the rainfall intensity l_{k+1}^{own} with a high standard deviation $\sigma_{l,k+1}^{own}$ due to the uncertainty of the wiper-rainfall relationship. Only one observation is used to update the system state.

2. The car is located inside the communication range of a weather station. The weather station determines the rainfall intensity and transmits the data to the car. Once the data exchange is done, the car uses the observation $l_{k+1}^{station1}$ and its small standard deviation $\sigma_{l,k+1}^{station1}$ to update its own system state. The weather station does not update its measurements with the car measurements, because the weather station is measuring with highest accuracy and the improvement by the cars is not significant. The small standard deviation helps to improve the certainty of the system state (as shown in Fig. 2). If the car is in communication range of two or more stations, the observations are put together in a vector and their standard deviations are used to build a covariance matrix for the observation vector.

3. The car is located outside the communication range of a weather station, but inside the communication range of another car. It receives the rainfall intensity and its standard deviation from the system state of the other car and uses it as an

observation together with its own observation, if its system state is more uncertain than the system state of the other car. The rainfall intensities can be considered as equal, as the communication range of a car is very small. If more cars with a smaller standard deviation are in communication range, all observations are put together in an observation vector $\mathbf{l}_{k+1}$ and its covariance matrix $\mathbf{\Sigma}_{ll,k+1}$.

$$\mathbf{l}_{k+1} = \begin{bmatrix} l_{k+1}^{own} \\ l_{k+1}^{car,1} \\ \vdots \\ l_{k+1}^{car,n} \end{bmatrix} \rightarrow \mathbf{\Sigma}_{ll,k+1} = \begin{bmatrix} \sigma^2_{l^{own}_{k+1}} & & & \mathbf{0} \\ & \sigma^2_{l^{car,1}_{k+1}} & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma^2_{l^{car,n}_{k+1}} \end{bmatrix} \qquad (3)$$

Mapping of the rainfall

In order to map the rainfall data, the area of the road network is converted from vector to raster data. Each cell from the road network is a possible candidate to receive information about the rainfall once a car passes by. We consider two factors that will have influence on the quality of the mapped data. The quality of the information in a cell is decreasing with the elapsing time, but it will increase with the number of cars that pass this cell. In order to model this fact, a second Kalman filter for each cell that can be passed by a car is implemented. Its system state is described as follows:

$$\begin{aligned} &\dot{x}_k^+ \rightarrow \mathbf{\Sigma}^+_{xx,k} = \sigma^{+\;2}_{\dot{x},k},\ \mathbf{\Sigma}_{ww} = \sigma^2_{w\dot{x}} \\ &\mathbf{\Phi}_{\mathbf{k}}^{\mathbf{k+1}} = 1 \end{aligned} \qquad (4)$$

As in our case the simulated raincloud is static, we do not need the parameter $\ddot{x}_k^+$, which was implemented in the Kalman filter for the cars (2). The decay in quality is modelled with the system noise $\mathbf{\Sigma}_{ww}$, which is added to the system state at every time step as a part of the prediction.

Once a car passes by, the system state of the cell is updated, using the system state of the car about the rainfall intensity and its standard deviation as an observation.

After the simulation run, we are able to make a statement about the quality of the rainfall mapping by looking at the following statistics:

- The difference between the mapped data and the simulated values.
- The standard deviation of each cell.
- The number of times a cell has been visited.
- The coverage of the area.

They will be presented in the following chapter, where we discuss the first experiments that we have done in the presented simulation environment.

4. EXPERIMENTS

We took road data as well as the locations of the weather stations from a study area of approx. 3300 km^2 in the Bode river basin located in the Harz Mountains in Northern Germany (Haberlandt & Sester, 2009). Our results are based on a given car density and station distribution. Some parameters are chosen identical for every run of the simulation: The communication range for a car-car system is set to 200 m and for a station-car system to 2000 m. The simulation time is 1.5 h for each run and the cars are driving with an average speed of 70 km/h. The size for each cell is set to 200 m. The relation between the system noise and the measurement uncertainty, which controls the abatement of the car’s certainty, is identical for each run.

Simulation run with 50 cars

As a result of this run, we reached a standard deviation based on differences between the mapped rain values and the given ones of 6 %, which is acceptable. In total, 25 % of all reachable cells were mapped during the simulation. As some of them were visited twice or more often, an average visiting rate of 0.69 for each of them was reached.

An example for the improvement of the system state of a single car is given exemplarily in Fig. 4. It shows that the certainty of the system state of a car improves rapidly when it communicates with a weather station. After the communication range is left, it decreases slightly until it reaches the initial value again. Similarly, the communication with a car leads to an improvement of the quality, although it is not as high as in comparison with the weather station. An interesting fact is shown in the third break of the curve. The system state can improve even more, when two cars communicate several times in a row.

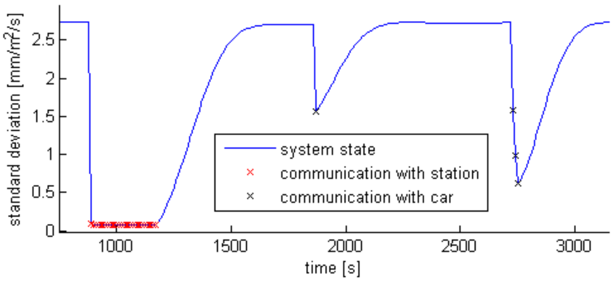


Fig. 4: Improvement of the system state by communication with other participants.

The quality of the mapped data is shown in Fig. 5. It gives an overview over the simulation area, the distribution of the weather stations and shows the standard deviation of each mapped cell.

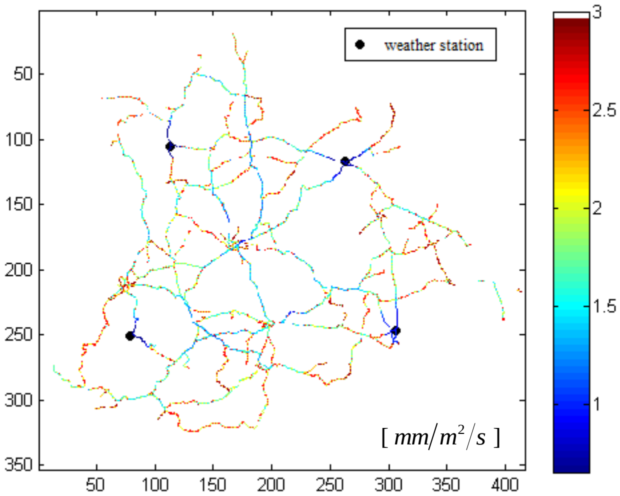


Fig. 5: Standard deviation of each reached cell with a distribution of four stations and 50 cars.

It confirms the statement of Fig. 4, as it shows dark blue areas around the station, which stands for a low standard deviation. The standard deviation on roads, that are chosen more often, seems to be on a lower level than on other roads that fork from them. This effect can be explained by the number of visits, as shown in Fig. 6.

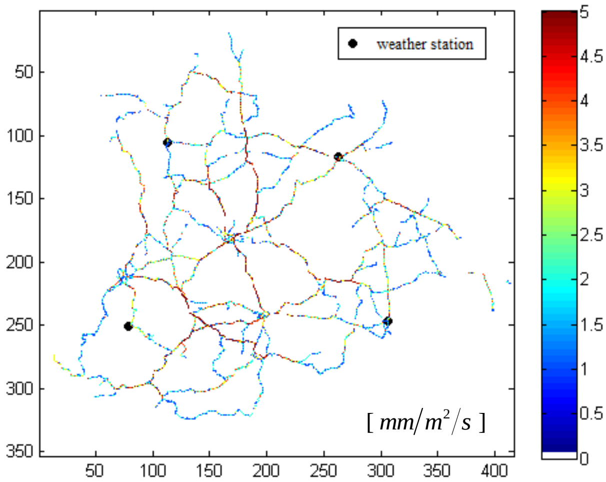


Fig. 6: Number of times a cell has been visited using 50 cars.

It shows that these roads are more often visited, than the other ones. In fact, the correlation between visiting time and variance of a cell is calculated to -0.73, which means, that the quality of mapping is not only affected by the weather station information, but also by the number of visits.

The following example shows the mapping quality results with the original station distribution. The original station distribution leads to a better mapping in the area where they are placed, although some of them are never reached by a car. It confirms the dependency of the mapping quality on the number of visiting times, because the standard deviation between Fig. 5 and Fig. 7 is nearly identical for roads, which have been chosen more often, and therefore nearly independent from the station distribution.

In order to improve the mapping quality, we did another simulation run with 100 cars. The results of this run are presented in the next section.

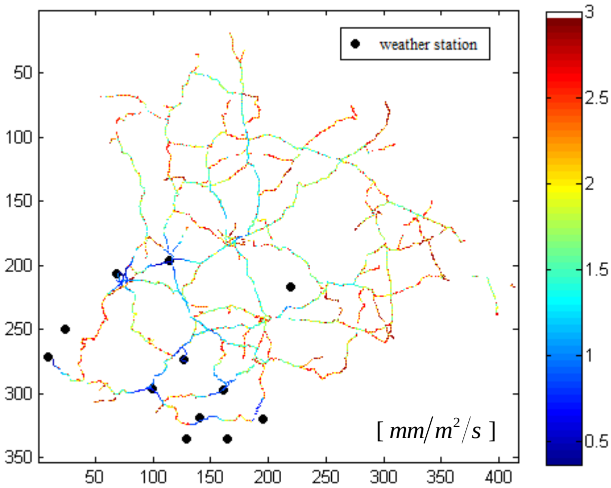


Fig. 7: Standard deviation of each visited cell, using the original distribution of stations.

Simulation run with 100 cars

As a result of this test run, we reached a standard deviation based on differences to the original rain values of about 7 %, which is the same order as the simulation above has shown. The coverage of the area is slightly higher with about 35% of all reachable cells. On average, each cell was visited 1.4 times. The standard deviation of each mapped cell is shown in Fig. 8.

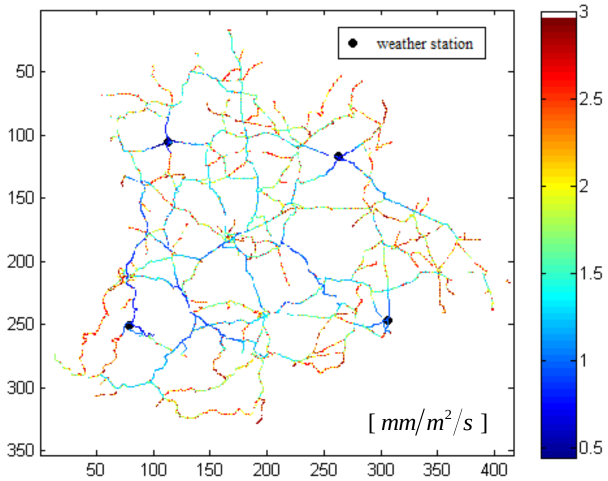


Fig. 8: Standard deviation of each visited cell with a ficticious distribution of four stations and 100 cars.

The main roads of a low standard deviation are much the same as in the tests runs that are described before, but they reached a higher level of system certainty, which can be even at the same level as the area, that is covered by the rain stations. According to the results already mentioned, this indicates that a small number of roads are chosen more often than others, these are the main roads in the network which connect the towns. This leads to the conclusion that weather stations to improve the system state of a car are much more needed at roads that are not so highly frequented, as the main roads. As the chance is high that a car, which receives information from a weather station on a low frequented road, will continue its journey on a main route is much higher than the other way around, the whole area will be mapped with a higher quality.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an approach to use a sensor network in order to predict rainfall intensities over a large area. Our sensor network is made of two different sensor types – highly accurate, but stationary, rain stations, and moving cars, which measure the rainfall only indirectly (and inaccurately) via their wiper frequencies. Although we concentrate on the rainfall application here, the basic principle can be easily adapted to other scenarios which involve moving low-budget sensors which improve their accuracy by communication with other (possibly more accurate) sensors.

In order to evaluate our approach, we used a real street network and real weather station locations. We then simulated rainfall intensity using a mixture of Gaussians as well as the positions of cars over time. From this, we derived results regarding the standard deviation of the estimated rainfall intensity, which is considered to be a measure of the system's certainty about the estimated state.

There are a number of improvements possible, which we will consider in future work. First, we assumed some constants in our simulation, especially the system and measurement noise in the Kalman filters. These constants should be verified using real data. Second, we used a rather simple model for the relationship between the wiper frequency and the rainfall intensity. However, ideally, this relationship should be more complicated and the filter should include calibration parameters, such as an offset and bias. Finally, the assumption of static rainfall could be replaced by a moving rain field and simulated traffic could be replaced by real (measured) traffic frequencies and speeds.

6. REFERENCES

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y. & Cayirci, E., 2002. Wireless sensor net-works: a survey. Comput. Netw. 38(4), 393–422.

Brown, R. G., Hwang, P. Y. C., 1997. Introduction to random signals and applied Kalman Filtering. John Wiley & Sons.

Duckham, M. & Reitsma, F., 2009. Decentralized environmental simulation and feedback in robust geosensor networks. Computers, Environment and Urban Systems, vol. 33(4), 25-268.

Haberlandt, U., 2007. Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. J. of Hydrol., 332, 144-157, 2007.

Haberlandt, U. & Sester, M., 2009. Areal rainfall estimation using moving cars as rain gauges – a modelling study. Hydrology and Earth System Sciences Discussions, vol. 6, no. 4, p. 4737-4772, 2009.

Laube, P. & Duckham, M., 2009. Decentralized spatial data mining in distributed systems. in H. Miller & J. Han, eds, ‘Geographic Knowledge Discovery, Second Edition’, CRC, Boca Raton, FL, pp. 211–220.

Laube, P., Duckham, M. & Wolle, T., 2008. Decentralized movement pattern detection amongst mobile geosensor nodes. in T. Cova, K. Beard, M. Goodchild & A. Frank, eds, ‘Lecture Notes in Computer Science 5266’, Springer, Berlin, pp. 211–220.

Raney, B. & Nagel, K., 2006. An improved framework for large-scale multi-agent simulations of travel behaviour, in: Rietveld, P., Jourquin, B., Westin, K. (Editors) Towards better per-forming European Transportation Systems, p. 42, London: Routledge.

Raubal, M., S. Winter, S.Teβmann, Ch. Gaisbauer, 2007. Time geography for ad-hoc shared-ride trip planning in mobile geosensor networks, ISPRS Journal of Photogrammetry and Remote Sensing Volume 62, Issue 5, October 2007, Pages 366-381.

Sester, M., 2009. Cooperative Boundary Detection in a Geosensor Network using a SOM, Proceedings of the International Cartographic Conference, Chile, 2009.

Simon, D., 2006. Optimal state estimation. John Wiley & Sons.

Stefanidis, A. & Nittel, S., eds., 2004. Geosensor Networks, CRC Press.

Walkowski, A. C., 2008. Model based optimization of mobile geosensor networks, in L. Bernard, A. Friis-Christensen & H. Pundt, eds, ‘AGILE Conf.’, Lecture Notes in Geoinformation and Cartography, Springer, pp. 51–66.

Zou, Y. & Chakrabarty, K., 2004. Sensor deployment and target localization in distributed sensor networks. ACM Trans. Embed. Comput. Syst. 3(1), 61–91.

SHARING LANDSCAPE INFORMATION THROUGH AN ONLINE GEOGRAPHICAL VISUALISATION PORTAL

C. J. Pettit, M. Imhof, M. Cox, H. Lewis, W. Harvey, J-P Aurambout

Future Farming Systems Research Division, Department of Primary Industries, Victoria, Australia -
christopher.pettit@dpi.vic.gov.au

Commission II, WG II/6

KEY WORDS: natural resource management, visualisation, online repositories

ABSTRACT:

The access to, and visualisation of, landscape information through online websites can effectively support natural resource management and decision making. Online repositories of information are a useful resource for community members and researchers to enhance their understanding of agricultural and natural landscapes past, present and future. In this paper we report on the development of an online geographical visualisation resource in Australia that provides access to scientific outputs created through a number of visualisation techniques.

The Victorian Resources Online website (http://www.dpi.vic.gov.au/vro) is a collection of more than 7,500 pages of natural resource information and maps. A Geographical Visualisation Portal resides within the site that includes: (i) links to contemporary media, (ii) access to technical reports and publications, (iii) video clips depicting visualisation techniques applied for understanding real and fictitious geographies, (iv) downloadable interactive content such as KMZ files, (v) a virtual soil profile used as an educational aid to increase understanding of the complex dimensions and properties of soils and (vi) a 3D object library comprising trees, shrubs, animals, built structures and rural features. The online visualisation portal provides an alternative metaphorical interface for users to access content to better understand Victorian landscapes.

1. INTRODUCTION

1.1 Research Aim

With a number of critical issues facing society such as climate change, food supply, deforestation, biodiversity loss and water shortages, landscape information and communication tools are becoming increasingly important. Geographical visualisation provides a powerful communication vehicle for communicating past, present and future landscape change scenarios.

In this paper we introduce a novel comprehensive online Geographical Visualisation Portal which aims to: (i) improve the communication of natural resource information to end users through the use of static and interactive visualisation products, and (ii) provide a resource to the broader geographical visualisation research and practitioner communities to support the development of their own visualisation products and services

1.2 Geographical Visualisation

Geographical visualisation also known as GeoViz, draws upon many disciplines including cartography, scientific visualisation, and GIScience to provide theory, methods and tools for the visual exploration, analysis, synthesis and presentation of data that contains geographic information (MacEachren and Kraak, 2001). GeoViz software packages broadly include: standard geographical information systems (GIS); digital globe packages such as Google Earth and Microsoft Virtual Earth; multi-media and 3D animation software such as Flash and 3D Studio Max; Virtual World platforms such as SecondLife and Cybertown; and Computer Gaming Engines such as Unity and Unreal. GeoViz software packages can be used to create two-dimensional, three-dimensional and four-dimensional (temporal) visualisation products to more fully engage or immerse end users in an exploratory information experience. GeoViz products can be developed as stand-alone products or on-line static or interactive semi and fully immersive environments. In this paper we focus on GeoViz as it is applied for sharing and exploring natural resource information.

1.3 Brief Review of Natural Resource Management Websites

There are numerous websites, portals and repositories containing natural resource information. A web search using the Google search engine (8th Nov 2009) resulted in 44,200,000 web hits on the search term ‘natural resource management + website’. This clearly indicates that significant natural resource management information resources exist online. These sites contain both public and private information repositories and some comprise interactive spatial mapping tools to access and share information. Others contain only textual and report information. A limited but growing number of these online sites comprise information that can be accessed and explored through the use of interactive, three-dimensional geographical visualisation interfaces.

The Development Resource Management Portal for US AID (http://www.rmportal.net/) (accessed 8th Nov 2009) provides an example of a website which comprises both public access and a dedicated members section. The Natural Resource Management Shared Land Information Portal (http://spatial.agric.wa.gov.au/slip/index.asp) (accessed 8th Nov 2009) provides an example where natural resource management information can be accessed via a java enabled interactive spatial viewer. There exist a number of Wikipedia-based natural resource management portals, for example the Environment Portal (http://en.wikipedia.org/wiki/Portal:Environment) (accessed 8th Nov 2009). None of these portals or websites enable the end user to access and explore the information using novel GeoViz techniques.

Yet, there are a growing number of online sites where natural resource information can be accessed and shared via a range new engaging GeoViz interfaces. For example the Climate Change in Google Earth site (http://www.google.com/landing/cop15/) (accessed 30th Nov 2009) includes five narrative virtual tours where visitors can explore a number of issues such as the impact of deforestation and subsequent increased carbon dioxide emissions, and increases in water stress brought about by changes in rainfall patterns (Figure 1). Also available from this site are downloadable Intergovernmental Panel for Climate Change (IPCC) climate predictions for low, medium and high emissions scenarios. Through these you can explore predicted changes in global decadal annual mean temperature and rainfall from 2000 to 2090.

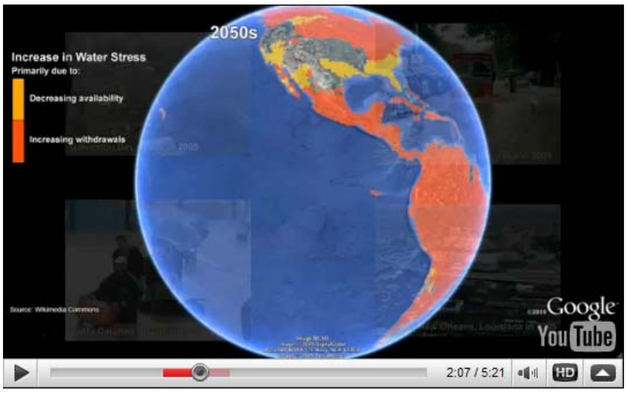


Figure 1. Google Earth used to visualise projected increase in water stress for 2050

The Mannahatta project (http://themannahattaproject.org/) (accessed 30th Nov 2009) provides an example of visualising New York pre-settlement ecology in 1609. This project uses geographical information system (GIS), Google Maps (see Figure 2) and 3D animation software to create a number of online interactive exploratory tools for visitors to explore both past and present New York.

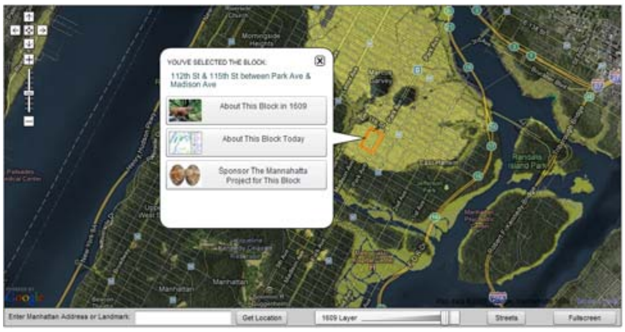

YOU'VE SELECTED THE BLOCK:
112th St & 115th St between Park Ave & Madison Ave
About This Block in 1609
About This Block Today
Sponsor The Mannahatta Project for This Block
Enter Manhattan Address or Landmark:
Get Location
1609 Layer
Streets
Fullscreen

Figure 2. Mannahatta project Google Maps exploration window. Visitors can use the slider bar to see probable landscape changes from 1609 to present day New York

National Geographic has established BlogWILD (http://blogs.nationalgeographic.com/blogs/blogwild/) (accessed 30th November 2009) for sharing information about the planet. A posting from November 3rd reports on the use of the SecondLife immersive 3D virtual world that has been used to host New Media Consortium's 'Symposium for the Future' (see Figure 3). National Geographic used this virtual world symposium to give a presentation about the planet and how new media approaches, such as virtual worlds, can be used to assist people to better understand natural systems. Virtual World platforms such as SecondLife provide a way to present landscape information through what Cartwright (1999) refers to as the 'gaming metaphor'.

Figure 3. National Geographic communicating natural resource information via a SecondLife Virtual World online symposium

2. VICTORIAN RESOURCES ONLINE

2.1 Overview of VRO

Since 1997, the Victorian Resources Online (VRO) website has been a key means for the dissemination of natural resources information via the Internet in Victoria, Australia. The website currently consists of around 7,500 web pages as well as 1,900 maps and 1000 downloadable documents and reports. Information is provided at a range of scales—from statewide and regional overviews to more detailed catchment and sub-catchment levels. At all spatial scales, information is organised around the key 'knowledge domains' of climate, landform, land use, soil, water, biodiversity and land and water management. In 2009 the website has attracted over 1,000 unique users per day. User profiling shows a wide range of users accessing and using material—including students and teachers, researchers, consultants, librarians, advisers and farm extension staff. Information presentation on the website is continually being enhanced, more recently with incorporation of visualisations to support the more standard web content, i.e. text and graphics.

2.2 Geographical Visualisation Portal

The VRO Geographical Visualisation Portal was established in 2005. Its aim was to provide alternative ways to access and interact with natural resource data, information and knowledge. The portal was established to: (i) complement and supplement existing online VRO content, and (ii) provide an online portal where users can access a number of geographical visualisation resources. This research endeavours to support a number of end users ranging from citizens, catchment managers and policy makers to geographical visualisation researchers and practitioners. The following sections of this paper will introduce the various components that comprise the Geographical Visualisation portal.

2.2.1 WWW Online Resources

This section of the portal provides a gateway to a number of related international geographical visualisation online resources assembled by Cartwright, (2005). The resource provides hyperlinks to a number of selected applications using contemporary media to visualise geography. The resources are categorised under a number of headings including (i) soil and landform, ii) life sciences, (iii) map and image collections, (iv) downloadable data storages, (v) information services with maps, (vi) online map-generation services, (vii) web atlases, (viii) hybrid products, (viii) 3D products, (ix) innovation and geographical visualisation tools, (x) developing areas of interest and (xi) references. The information is arranged hierarchically with a series of image thumbnails and hyperlinks providing links to external resources.

2.2.2 Publication Repository

This section of the portal provides a PDF version of a number of visualisation publications including: technical reports, conference papers and factsheets. There are links to peer papers published by the ISPRS working group II/6 ‘Geographical Visualisation and Virtual Reality’. Links are also available for each of these presentations that have been published as YouTube movies. This document repository by no means provides a comprehensive library of papers, reports and factsheets on the topic of geographical visualisation. Rather the repository provides a number of downloadable documents that are accessible to range of audiences from the public and policy makers to geographical visualisation researchers and practitioners. These reports and papers are scientific outputs produced by the Victorian Department of Primary Industries Geographical Visualisation Team and collaborators.

2.2.3 Video Clips

With the continual advances in Information Communication and Technology (ICT) there is a growing number of software tools available to create geographical visualisation products. These include open source software such as Virtual Reality Mark Up Language (VRML) and X3D, photorealistic visualisation packages such as Visual Nature Studio (VNS), GIS tools such as ESRI’s ArcScene, and Digital Globe products such as NASA World Wind, Biopshere, Google Earth and Microsoft’s Virtual Earth. Each of these products has strengths and weakness. For a comparative review of digital globe products see Aurambout et al. (2008).

VRML has been used to create a 3D interactive virtual Knowledge Arcade for natural resource management in Victoria. This virtual arcade includes 19 virtual shop fronts for a number of key agencies including catchment management authorities, water authorities, state government departments and universities (Pettit et al. 2008). A virtual tour of the Knowledge Arcade has been created as a video clip file embedded within the Geographical Visualisation Portal (see Figure 4).

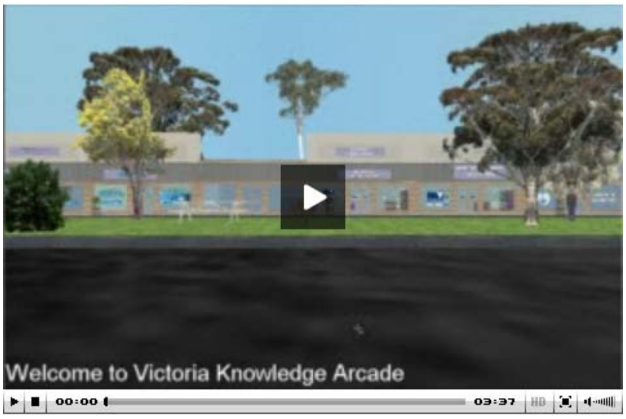


Figure 4. Natural Resource Management Virtual Knowledge Arcade video clip

There are a number of photorealistic visualisation software packages which can be used to created GeoViz products depicting past, present and future landscapes. For example Appleton et al. (2002) has produced photorealistic landscape visualisations to communicate climate change futures in agricultural areas in the United Kingdom. Figure 5 provides an example of the photorealistic landscape visualisations created for illustrating the concept of tree fencing a farm paddock in rural Victoria. This video clip shows the creation of a tree fence using fallen branches and trunks to encourage native vegetation regrowth by protecting the area from sheep grazing.

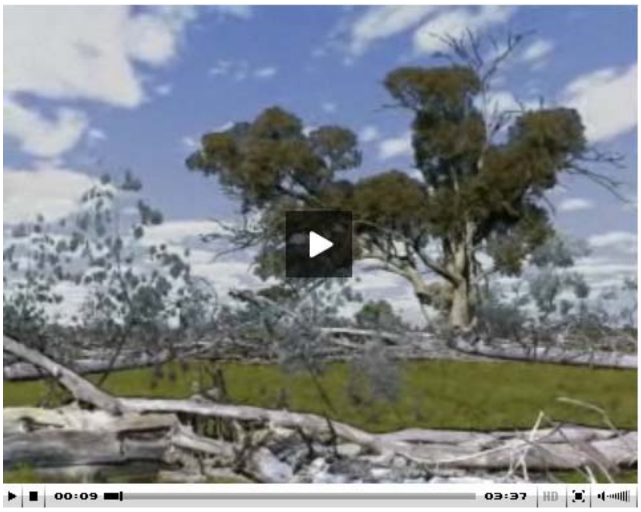


Figure 5. Photorealistic visualisation of the concept of tree fencing a farm paddock created using Visual Nature Studio

ArcScene is a GIS-based visualisation extension used to create 3D scenes and fly-through movies. The Mt Elephant video clip was created using this software (Figure 6). The basic requirements for creating the 3D movie included a digital elevation model (DEM) or LIDAR (light detection and ranging) data and a geo-referenced airphoto raster. Base height values and extrusion settings are derived from the DEM. The airphoto imagery is draped across the DEM and the render effects function enables a hypothetical illumination of the surface by creating hillshade. The ArcScene Fly tool facilitates a realistic fly-over effect to investigate the scene. A simple recording is achieved using controls that resemble a VCR (Video Cassette Recorder) on the Animation Controls dialog box that creates a movie file.

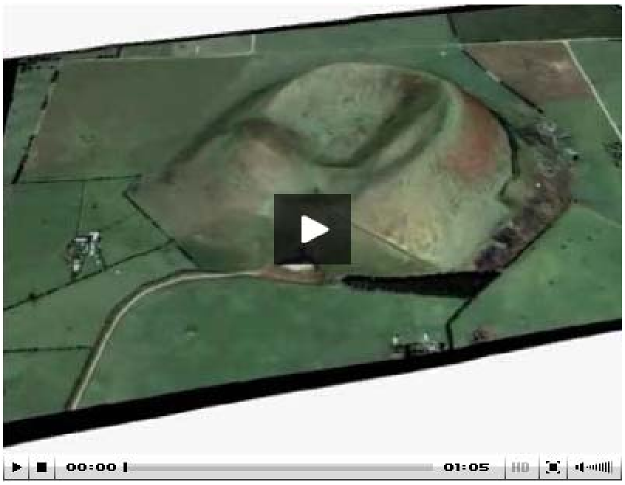


Figure 6. 3D Fly-though video clip of the Mt Elephant Geological Significant Site

The information provided on the ‘Sites of Geomorphological and Geological Significance’ section of the VRO website has been derived from a number of limited distribution publications and developed in association with retired geological specialists. The VRO website provides maps showing locations of many hundreds of these sites, as well as associated text and images. Video clips are now being routinely used to provide enhanced web content and animations. Content includes landscape fly-overs that have been developed to provide a ‘virtual tour’ of these sites (e.g. ‘virtual tour’ of Mt Elephant example shown in Figure 6), including historical footage and audio-visual recordings of retired experts describing landscapes in the field.

2.2.4 Interactive Content

A section of the portal contains downloadable interactive content. The content has been developed predominantly as OGC compliant Extensible Keyhole Markup Language (KML) files (KMZs). Visitors who have a KML compliant digital globe application can download these files and explore and interact with the digital content provided. Fishtrack is one such project where a downloadable KMZ file has been created (Figure 7). The Fishtrack project visualised the movement of a single black bream fish through the Gippsland Lakes for 1 year (2005–2006). The input data has been captured through monitoring the movement of the fish as tracked via proximity to known sensors within the lakes. Such visual information can provide insights into the migratory patterns and breeding behaviours of fish over time.

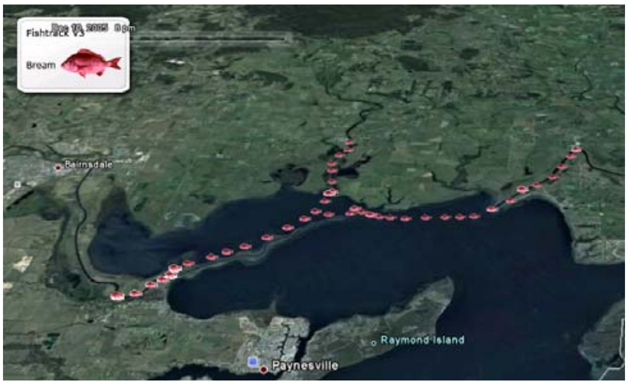


Figure 7. Fishtrack example of temporal mapping of a single black bream fish through the Gippsland Lakes

2.2.5 Virtual Soil Profile

The ‘Virtual Soil Profile’ (VSP) is an educational aid to increase understanding of the complex dimensions and properties of soils (Imhof et al. 2008). It is being developed as a visualisation tool to enhance soils education and training through improved awareness and understanding of soils and their function (see Figure 8). The initial objective is to display aspects of soil morphology and biology and their relationship to soil aggregation and management as well as depicting key processes (e.g. carbon cycle). The VSP was initially developed to support soils education and training as a stand-alone application but is being progressively incorporated into the VRO website for access by a broad range of users. Zooming functionality allows viewing at increasingly finer scales—from the pedon and its component soil horizons, to peds, through to macro- and micro-aggregates. Soil biology can also be viewed at increasing levels of detail—ranging from soil litter organisms (e.g. collembola and mites), earthworms, organisms associated with soil minerals and the living plant (e.g. protozoa), as well as bacteria and fungi.

The VSP is made up of 4 orientated digital images: north, east, west and floor images. These are used to construct the 3D VSP. It is developed and run through Coppercube 3D flash engine. The VSP is set in a panoramic scene. The panorama camera view is as if you were inside the soil pit itself. This additionally allows the panorama profiles to be used as navigation using the mouse or keyboard through the VSP features by applying hotspots to the profile textures. The 3D Coppercube engine also includes an action scripting reference. This reference allows the development of events to interact with the VSP. We used this to develop navigation menus based on soil structure and elements to move your way through and interact with the VSP by using the Coppercube actionscript switchToScene reference. The hotspots on the VSP are one of two navigation methods. By using these hotspots (links), a user can move their way between scenes and zoom in from VSP scale and macro scale down to 50 micrometres to view scanning electron microscope (SEM) imagery.

At different (zoom levels) scales, 2D and 3D animations are made available to the user to further immerse the user in the structure and processes that make up soil. The second method of navigation is through the frozen tree menu.

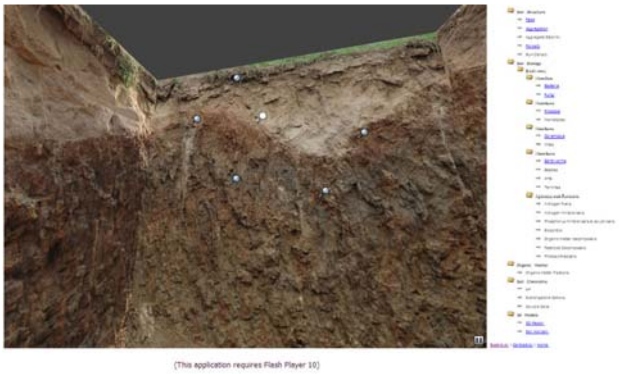


Figure 8. 3D VSP panorama with hotspots and tree menu

The 2D and 3D animations are developed as a means of depicting processes such as decomposition that happen in soil as opposed to the VSP that show features such as nematodes that are part of the makeup of soil. The VSP will be run online or alternatively off a DVD in flash and html format as an interactive application to present the features and processes that soil consists of.

2.2.6 3D Object Library

In working towards a 3D Spatial Data infrastructure the VRO Geographical Visualisation Portal contains a selection of 3D objects that can be downloaded by end users to create their own custom landscape visualisation products. The 3D object library contains over 75 individual files that represent Australian vegetation, animals, built infrastructure and rural landscapes (see Figure 9). Objects are provided in a number of file formats including GIF, WRL, FLT, 3DS and SKB. Each object can be downloaded as a ZIP file containing multiple file formats so that the objects can be used in a number of visualisation software packages (Pettit et al. 2009). The current object library is similar to the Google 3D warehouse site, http://sketchup.google.com/3dwarehouse, however focus is placed on supporting multiple file formats and objects that represent Australian landscapes.

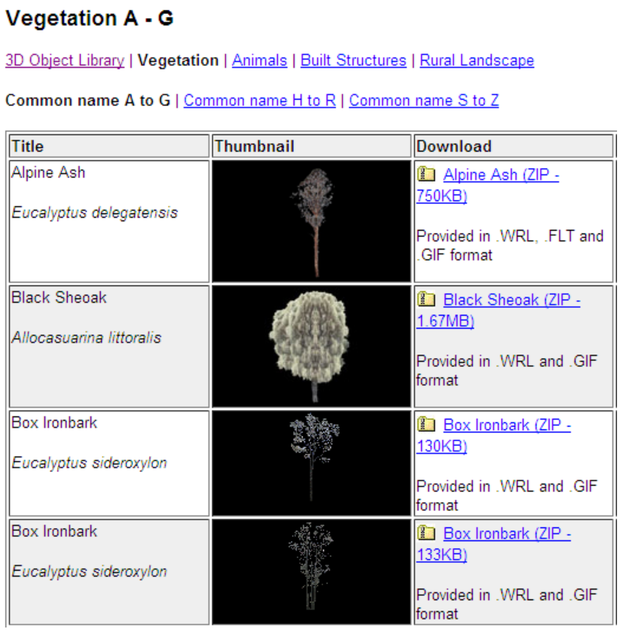
The current object library exists as a series of cascading style sheets (CSS) web pages under four broad categories. Ongoing research is focused on developing and testing a prototype geodatabase structure for accommodating a much large volume of objects and also to enable increased upload and search and discovery functionality (Bishop et al. 2009).

3. EVALUATION

Google Analytics reporting is integrated within the portal to collect user data and also provide a tool to analysis and report on usage and user trends. Analytical reports for the portal shows a consistent trend against data collected for the whole VRO web site. Usage of the portal can be seen to increase during the academic year with a peek at the end of term, usage then drops during academic holidays.

Consistently the most popular areas within the portal (26th November 2009) are the 3D Object Library and the geographical visualisation video clips section, with 43% and 23% of total traffic over the past 12 months respectively.

During the past 12 months data has been collected and a geographical profile of users has been calculated. The reports show (26th November 2009) that 48% of our users are from Australia, followed by USA (27%) and UK (4%). The Australian data can be further analysed to show that 61% of users are Victorian, 16% from NSW, 7% from Queensland and the remaining 16% are from the remaining states and territories. Interestingly, these metrics indicate that the resource has international exposure. Further research will investigate end user feedback both qualitative and quantitative, on the value of the Geographical Visualisation Portal as both an educational and decision-making resource.



Vegetation A - G

3D Object Library | **Vegetation** | Animals | Built Structures | Rural Landscape

Common name A to G | Common name H to R | Common name S to Z

Title	Thumbnail	Download
Alpine Ash *Eucalyptus delegatensis*		Alpine Ash (ZIP - 750KB) Provided in .WRL, .FLT and .GIF format
Black Sheoak *Allocasuarina littoralis*		Black Sheoak (ZIP - 1.67MB) Provided in .WRL and .GIF format
Box Ironbark *Eucalyptus sideroxylon*		Box Ironbark (ZIP - 130KB) Provided in .WRL and .GIF format
Box Ironbark *Eucalyptus sideroxylon*		Box Ironbark (ZIP - 133KB) Provided in .WRL and .GIF format

Figure 9. 3D object library downloadable vegetation trees and shrubs

4. LIMITATIONS

Victorian Government Departmental websites need: “to give citizens ready access to government websites, without discrimination.” (http://www.egov.vic.gov.au/victorian-government-resources/standards-victoria/accessibility-standard.html) (accessed 26th November 2009). Content published in the Geographical Visualisation Portal had to comply with these accessibility guidelines that state how information should be provided. In brief, consideration needs to be made to people with disabilities, people using older technology and people with poor telecommunications infrastructure, often in regional and remote areas.

Users with a disability may need access to information published within the portal, therefore the different ways they interact with the web page needs to be accounted for and solutions implemented so they not face discrimination.

To ensure content is available to people using older technologies and with poor telecommunications infrastructure we restricted the number of file formats supported so users only need to install one application to see all the objects we provide. Flash video capability has been utilised to provide quick access to a preview within the internet browser of some objects before the user needs to download it. Interactive content has been created in KMZ file format that is an Open Geospatial

Consortium (OGC) compliant format which can viewed via a number of GIS and digital globe platforms.

The government website guidelines provide challenges in being able to serve-up visualisation outputs which require plug-ins and other web components to run. Also, larger more complex visualisation outputs need to be made available as smaller files as well as optional larger downloads or higher definition files to accommodate end users with poor telecommunication infrastructure. Therefore, for example the Virtual Knowledge Arcade discussed in Section 2.2.3 that was created in VRML cannot be hosted via VRO. This is because even though VRML is a web compliant format it requires a third party plug-in to enter the virtual world. Also the Virtual Knowledge Arcade is over 600 MGBs and exceeds the permissible download standards. The video clip functionality provides a current work around for such web hosting limitations.

5. FUTURE WORK

There are a number of ongoing developments within the Geographical Visualisation Portal. These include the completion of the virtual soil profile and continual uploading of new objects into the 3D object library. Also a Virtual DemoDairy section is currently being developed. This site will enable visitors to undertake virtual tours of a dairy in south-west Victoria under current and future forecast climate change scenarios.

The next important phase of this research is to obtain metrics from end users of the value of the Geographical Visualisation Portal as both an educational and decision support resource.

6. CONCLUSIONS

In this paper we have presented the Victoria Resources Online Geographical Visualisation Portal. The site provides a number of resources to support the communication and sharing of natural resource information. The site incorporates both static and dynamic geographical visualisation products and includes a 3D object library resource to support end users in creating their own virtual landscapes.

One of the challenges facing online resources such as the VRO Geographical Visualisation portal is being able to enable a truly collaborative Web 2.0 interactive information-sharing environment. This is a current challenge to many government organisations in being able to provide online user-centred designed experiences.

REFERENCES

Appleton, K., Lovett, A., Sunnenberg, G., Dockerty, T., 2002. Rural landscape visualisation from GIS databases: a comparison of approaches, options, and problems. *Computers in Environment and Urban Systems*, 26 (2-3), pp. 201–211.

Aurambout, J-P., Pettit, C.J., 2008. Digital globes: gates to the digital earth. In: *Digital Earth Summit on Geoinformatics 2008: Tools for Global Change Research*, (Eds. Ehlers, M. Behncke, K. Gerstengarbe, F-W, Hillen, F Koppers, L Stroink, L Wachter), J Wichmann, Heidelberg pp. 233-238.

Bishop, I., Chan, P., Chan, T., Lau, A., Pettit, C., Stock, C. and Syed, D., 2009. Object libraries: the next step. In: *Spatial Data Infrastructure, Spatial Sciences Conference09*, Adelaide, 30 Sept-3 Oct.

Cartwright, W., 1999. Extending the map metaphor using web delivered multimedia. *International Journal of Geographical Information Science*, 13 (4), pp. 335-353.

Cartwright, W., 2005. "Online Resources for Geographic Visualisation" Melbourne, Victoria. http://www.dpi.vic.gov.au/dpi/vro/vrosite.nsf/pages/geovis_online_tools (accessed 26th February 2010).

Imhof, M., Mele, P., Lewis, H., MacEwan, R., Pettit, C., Bougoure, D. and Johnston, T., 2008. Virtual soil profile – an interactive tool to enhance soils education. In: *Australia New Zealand Soils Conference: SOIL – the living skin of Planet Earth*, Palmerston North, New Zealand, 1-5 Dec.

MacEachren, A.M. and Kraak, M.J., 2001. Research challenges in geovisualization. *Cartography and Geographic Information Science, Special Issue on Geovisualization*, 28(1), pp. 3-12.

Pettit, C.J. and Wu, Y., 2008. A virtual knowledge world for natural resource management. In: *Landscape analysis and visualisation: spatial models for natural resource management and planning*, (Eds. Pettit, C., Cartwright, W., Bishop, I., Lowell, K., Pullar, D. and Duncan, D.), Springer, Berlin, pp. 533-550.

Pettit, C.J., Sheth, F., Harvey, W. and Cox, M., (2009) Building a 3D object library for visualising landscape futures. In: *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation,* Cairns, Queensland, 13-17 July, pp. 2244-2250.

AN ONLINE VISUALIZATION AND DATA ANALYSIS SYSTEM FOR SOCIAL AND ECONOMIC DATA BASED ON FLASH TECHNOLOGY

Jinqu Zhang[a,b], Yunqiang Zhu[a], Yaping Yang[a], Jiulin Sun[a]

[a] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
State Key Lab of Resources and Environmental Information System, Beijing, 100101, China
[b] Spatial Information Research Centre, South China Normal University, Guangzhou, 510631, China

KEY WORDS: Visualization, Data analysis, Social and economic data, FlashGIS

ABSTRACT:

Social and economic data are almost paid greatest attention by the country leaders and used to sense the situation of a country. The manner of visualization and data analysis to social and economic data will greatly affects the knowledge detection and information acquisition, so designing a good analysis system would be very necessary. This paper tries to use flash technology to design an online visualization and data analysis system to the social and economic data. By comparisons with the traditional web GIS systems, the system designed by the paper has great priorities in following three aspects: (1) Accelerating the online data analysis speed by changing the common work flow. (2) Enhancing the system interactivity by integrating and associating the spatial map, attribute data and statistics chart. (3) Creating a distribute map for any selected element and generating a time series animation dynamically. The result shows that using flash technology can achieve difficult functions that the traditional GIS software can’t realize.

1. INTRODUCTION

Socio-economic data are the most important data for analyzing the development situation of a country. The manner of visualization and data analysis to social and economic data will greatly affects the knowledge detection and information acquisition, so designing a good analysis system would be very necessary. Usually, geographic information systems (GIS), as a good spatial-temporal information processing tool, is widely used to develop various application and decision support systems in planning and development, environmental impact assessment, and real-time analysis of spatially distributed data (Maguire et al.,1991). GIS technology has also been applied to the economic development process at local, regional, and state levels of government, as an interactive-visualization and decision-support tool dealing with the socio-economic data (Drummond, 1993). Itzhak Benenson et al. (1998) offer a simple methodology for the analysis of socio-economic networks and emphasize the role of GIS as a visualization tool.

Although GIS is widely used to visualize and analyze the socio-economic data, these applications are mostly running at local machine. It is obvious that delivering, handling and publishing geo-referenced information on the web are attracting increasing numbers of researchers and application developers. Although ArcExplore, ArcView Internet Map Server, Geomedia,etc. are well-known, web-enabled, GIS applications that permit users to access, retrieve, display and analyze GIS data over the web (Green, 1997; Plewe, 1997;Strand, 1997), these web-enabled software has big limitation in response time for visualizing the socio-economic data. The classic web GIS model includes a client program (a Web browser), which makes a request to a server program, and the server processes that request and returns the information to the client. This process model makes every operation require the response processing done at a heavy server, which is not a good model for visualizing the socio-economic data for its low efficiency. Currently, most leading GIS vendors such as ESRI, Intergraph, Autodesk, MapInfo use heavy GIS map servers and specific applications on the Internet server to provide web mapping services. But there are still some limitations except for the heavy server and time response.

This paper is trying to use flash technology to design an online visualization and data analysis system to the social and economic data with light serer but rich applications in the client. The purpose of the system is to solve the following three problems: (1) To accelerate the online data analysis speed. Usually, the analysis should first submit an analysis task to the server and then get a result response at the client, and this process is very slow, especially referring to the spatial maps data. In the system designed in this paper changes the common work flow and accelerate the analysis speed by using flash technology. (2) To enhance the system interactivity by integrating and associating the spatial map, attribute data and statistics chart. On the one hand, users can understand and recognize the problem with multidimensional views at the same time. On the other hand, by associating technology, the interactivity is greatly enhanced. When the users click the administrative area of interested, the attribute data and statistics chart will accordingly change. (3) To create a distribute map for any selected element and generate time series animations dynamically. This would be very useful to detect the element changes from time series and spatial distribution.

2. WEB-BASED FLASHGIS

2.1 Flash technology

Flash is a multimedia platform used to develop web animations with its files in the SWF format that has been the practically standard of web animations for its highest popularity and currently developed and distributed by Adobe Systems since 2005(Wikipedia, 2009). Flash can manipulate vector and raster graphics, and supports bidirectional streaming of audio and video. Previously, Flash is commonly used to create animation, advertisements, and various web page Flash components, to integrate video into web pages. Recently, with the release of its scripting language ActionScript 3.0 and Flash Player 9 in 2006, Flash was used to develop Rich Internet applications (RIA) which was introduced in March 2002 by vendors like Macromedia who were addressing limitations at the time in the "richness of the application interfaces, media and content, and the overall sophistication of the solutions" by introducing proprietary extensions (Jeremy Allaire, 2002). Now, Flash is playing a significant role in the development of RIA.

2.2 Web-based FlashGIS systems

For the excellent features in the performance of the network animation and its ability to manipulate the vector and raster graphics, flash has long been used to express geographic information, and the GIS systems developed based on flash technology can be called a FlashGIS system. Robin Hilliard (2004) tries to handle seriously detailed maps in Flash by reading and processing the Mapinfo mif/mid files. ABC news online from Australia used a simple flash-based GIS system to show the federal elections automatically online in 2004. Zong-zhi Li et al. (2004) discussed the development of webGIS based on flash. Some other articles also have a try to use the flash to show the electronic map. Although there are some flash applications in the expression of geographic information, most of them are simple and limited functionality. Until recently, with the release of ActionScript 3.0 by Adobe, its use was greatly expanded. For example, ESRI provides ArcGIS API for Flex which allows the creation of RIA on top of ArcGIS Server based on the free Adobe Flex framework. Google, Yahoo and Mirosoft and other large companies have released relative maps APIs for use with the Adobe Flex framework. By using of these various APIs, lots of web GIS systems developed by ActionScript 3.0 have been applied in various fields. However, most of these applications are not true vector-based GIS and still use a heavy server with request/response mode. Dang Van Tuyen et al.(2008) designed a flash-based tool and applied it in the management of the country’s precious forest resources, which has the similar architecture with our FlashGIS system. In this paper, we will mainly demonstrate the application of FlashGIS in socio-economic data, especially in the statistical functionalities and construct an online visualization and analysis system.

2.3 Architecture of FlashGIS

The following chart is our overall architecture of FlashGIS. Firstly, prepare the administrative units SWF files totally 385 maps, including one boundary map of China, one provincial administrative region map of China, 34 municipal administrative region maps of 34 provinces or municipalities of China and 349 county administrative region maps of 349 municipal cities of China. Each SWF file expresses one map that is converted from ESRI shp file and also contains the attribute data within the ESRI dbf file. The SWF files are named by their maps administrative codes, so that when one clicks on China map, the corresponding province map containing cities under that click point will be automatically associated, and when one continues to click on the province map the corresponding city map containing counties will be associated. By doing this, the system realizes three-level associations that are country to province, province to city and city to county. On the other hand, when one clicks on an administrative unit, the administrative code will be retrieved from SWF file content, and then a SQL query will be constructed and executed to acquire this administrative unit socio-economic data from the database. Because most of the socio-economic data are acquired according the administrative unit, taking province and county for example, and every administrative unit has a unique administrative code, so each record of socio-economic data could be associated with an administrative unit shape. By setting different colours to different administrative unit shapes according to the value of socio-economic record data, a ranges thematic map will be created. As far as the socio-economic data is concerned, it can also be separately queried and analyzed by using various statistical charts such as pie chart, bar chart, line chart and so on. Finally, the statistical charts and the thematic maps could be combined together to express much more information just in only one map. For the thematic maps, a timeline is designed and an animation standing for the distribution of one element of socio-economic data of China province would be generated automatically by setting a start time and an end time, which animation map is unique that traditional web-GIS such as ArcIMS and MapXtreme cannot realize.

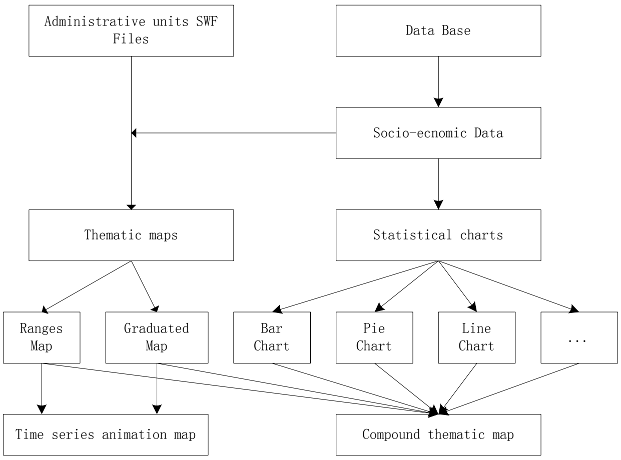


Fig 1. Flow chart of statistical flashGIS to the socio-economic data

3. SOCIO-ECONOMIC DATA VISUALIZATION AND ONLINE ANALYSIS

3.1 Socio-economic database construction

Almost every socio-economic data is recorded by the administrative unit and has an administrative code as a unique identification number, so it is possible to design a universal table structure to organize and administrate every table of socio-economic data. In our database construction, three kinds of table are designed and they are separately index table, content table and structure table.

There is only one index table of the whole socio-economic database and it is used to record the metadata information of every content table, that is to say, when you add a table to the database, you must first add a record to the index table to memorize the overall information of the new table. Table 1 shows the structure of index table. By using the index table, a content table can be easily queried according to the data content key words, cover area or statistical year et al. For a content table, it refers to one of the tables in a statistical yearbook and stores the table values. Corresponding to every content table, there is an associated structure table that stores the fields’ information in a content table. It describes the field name, field meaning, field unit and some other field information. By creating these three kinds of tables in the socio-economic database, an interactive query interface can be easily designed to acquire any table that the user wants to. Once a content table is selected, the corresponding metadata and table structure information will be automatically retrieved.

Until now, we have collected more than fifty years' data of every province in China, including more than 100 elements values of each province from 1950 to 2007 and 33 county-level data on key economic indicators of 3400 counties in China.

Table 1. The structure of index table

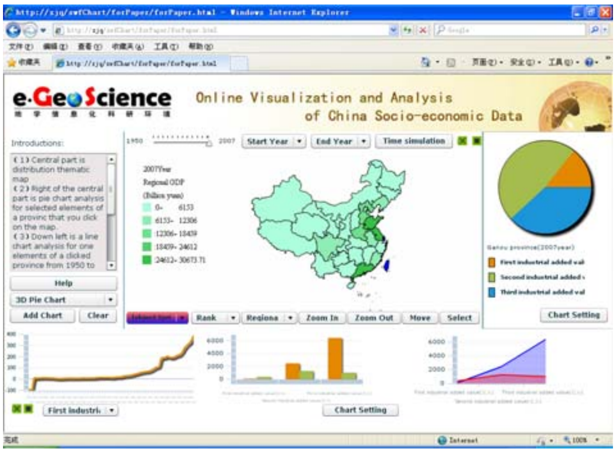
Field Name	**Field Type**	**Field meaning**
ID	Integer	The table number of a content table
Table_Name	String	The physical name of a content table
Stat_level	String	Expressing the administrative statistical unit, e.g. province, county.
Table_content	String	Expressing the key words of a content table.
Content_Type	String	Expressing the data acquisition mode, e.g. statistical data according to the administrative unit or the sampling spot check data
Cover_Area	String	Expressing the spatial range of a content table; record the covering provinces, cities and counties.
Data_Source	String	Expressing the data source of a content table.
Stat_year	String	Expressing the statistical year of a content table
Table_Structure	String	A table name recording the structure of a content table.
xMin	Number	The leftmost coordinate of the cover area of a content table
xMax	Number	The rightmost coordinate of the cover area of a content table
yMin	Number	The southmost coordinate of the cover area of a content table
yMax	Number	The northmost coordinate of the cover area of a content table
Ch_show	String	The external logical Chinese name of a content table.
En_show	String	The external logical English name of a content table.

3.2 SWF files preparing

Because the SWF file is the web animation file format with vector and raster data supported, so it is very easy to visit and browse the content of SWF directly on the internet without any configuration, which is different from the traditional GIS data, such as ESRI shape file, only be visited on the local machine. The predominant characteristics of SWF file promote us to convert the GIS information stored in the ESRI shape file to the SWF file, and then the SWF file is used for the spatialization and visualization of the attribute data queried from the socio-economic database. The shapes in a SWF file are assigned different colours according the selected field value in its corresponding record data. The shapes in a SWF file can be generated in two ways: (1) the one is to generate SWF files in advance; the other one is to store the ESRI shape files coordinates data in a database and generate SWF shapes at running time according to the user's request. In order to accelerate the internet visit speed, the first way is adopted and a series SWF files, containing a series of administrative units spatial shape, are created in advance. These SWF files are converted from ESRI shape files with attribute data and coordinates included by using a C++ programme written by ourselves. In a SWF file, each shape corresponds with a feature in the ESRI shape file and the attribute data are stored as an array with shape number associated.

3.3 Socio-economic data visualization and analysis

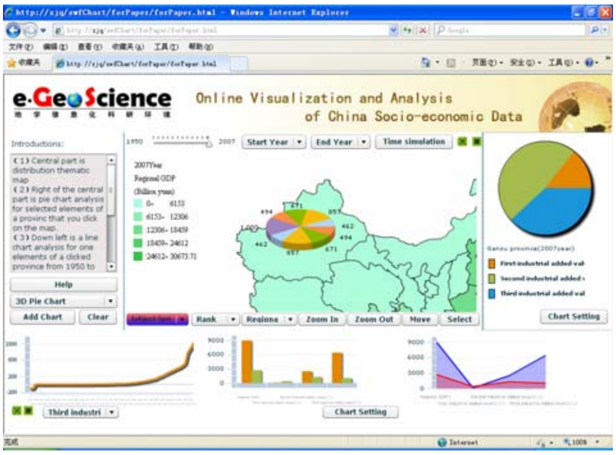
The online socio-economic data visualization and analysis system are mainly divided into three parts: (1) Spatial distribution thematic map analysis; (2) Statistical chart analysis and (3) Association analysis of the multi-level administrative units.



(1) Central part is distribution thematic map
(2) Right of the central part is pie chart analysis for selected elements of a provinc that you click on the map.
(3) Down left is a line chart analysis for one elements of a clicked province from 1950 to

Fig. 1. Index page of online visualization and analysis system

Fig.1 shows the index page of our online visualization and analysis system of china socio-economic data. At the centre of the page is a vector China map which is loaded from a SWF file. When the China SWF file loaded finished, a completed event will be activated and call for a function to run. The function will then link to the database by using a HTTPService component in flex and execute a SQL query to retrieve the regional gross domestic product (GDP) data of every province in China in the year of 2007. On the top of the map is a timeline and a time simulation button, users can click on the timeline to select one year between 1950 and 2007 and also can click the time simulation button to generate a thematic dynamic changing animation map. The most significant characteristic of this animation map is that it is online, real-time, fast speed. The regional GDP is the initial field and user can select any field he needs to generate a thematic map from more than 100 fields of the table. In addition to these, the functions of traditional GIS are provided, taking zoom in, zoom out, move and select for example. Here, the select button has special use. When the user choose select button and click on the map, the province under the click point will be automatically recognized and the statistical charts around the map, including pie chart, line chart, column chart and area chart, are also simultaneously changed. There are setting options for every chart to be set by the user. The statistical chart can still be put on the thematic map to form a multi-factor composite thematic map (Fig. 2). The detailed functions can be tested and visited on the web address of http://210.77.94.214/chart/index.html.



Introductions:

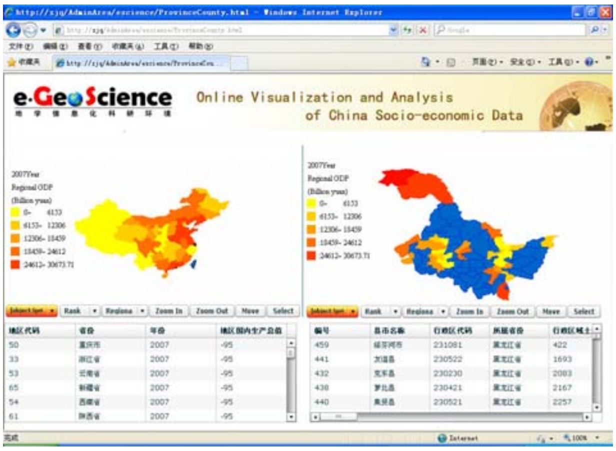
(1) Central part is distribution thematic map

(2) Right of the central part is pie chart analysis for selected elements of a provinc that you click on the map.

(3) Down left is a line chart analysis for one elements of a clicked province from 1950 to

Fig. 2. Add chart to the thematic map

Fig.3 shows the association analysis of multi-level administrative units. The main purpose of this function is to provide not only the micro spatial distribution of some element but also to detect the local distribution at the same time. In Fig. 3, the left is a whole China map, any element field can be selected to generate a thematic map of its distribution in China. When you click on the map, for example, the click point is within the range of Heilongjiang province, the Heilongjiang province map containing its counties will be automatically loaded on the right. In the bottom part is two grids showing attribute data that be selected and related to the map in the top.



地区代码	省份	年份	地区国内生产总值
50	重庆市	2007	-95
33	浙江省	2007	-95
53	云南省	2007	-95
65	新疆省	2007	-95
54	西藏省	2007	-95
61	陕西省	2007	-95

编号	县市名称	行政区代码	所属省份	行政区域土
459	绥芬河市	231081	黑龙江省	422
441	友谊县	230522	黑龙江省	1693
432	克东县	230230	黑龙江省	2083
438	萝北县	230421	黑龙江省	2167
440	集贤县	230521	黑龙江省	2257

Fig 3. Association analysis of multi-level administrative units

4. RESULTS AND DISCUSSION

By using the flash-based technology, we construct a vector-based flashgis system and apply it to develop an online visualization and analysis system for the socio-economic data of China. Originally, the socio-economic data is only attribute data with each record bounded with an administrative unit and they are incomprehensible. In order to acquire information much easier and detect more useful knowledge, we convert the China province level, municipal level and county lever GIS graphic data to the SWF files and combine them with the socio-economic data use an index field as association so that we can give a visualization analysis to the socio-economic data. Our application shows that it's a better approach to analysis the attribute data than the traditional GIS system manifested at least in three aspects: (1) web-based and cross-platform; (2) small file and fast speed; (3) rich rendering and strong expression.

4.1 Web-based and cross-platform

So far, there are many web animations created by Adobe flash software and they are all SWF file format, which make the SWF file the fact standard of web animation. Comparing with the traditional GIS data files, such as ESRI shp and Mapinfo tab files, we convert them to the SWF files and they are naturally web-based. On the other hand, SWF files are cross-platform without any additional configuration, not only on the Windows platform, but also on the UNIX platform and even the mobile phone can support the SWF files too. These characteristics make the SWF file have a high priority in developing web-based and mobile GIS system.

4.2 Small file and fast speed

The second priority of SWF files is that these SWF files are very small but with a high performance. Table 2 shows the comparison between SWF file and shp, tab files, from which we can find that the SWF file size is almost one tenth of tab file and one ninth of shp file. This is due to the high compression ratio of SWF file format and guarantees the whole file transfer to some extent. The SWF file also has the characteristics of streaming media that let it has a fast transfer on the internet.

4.3 Rich rendering and strong expression

SWF is commonly used to express web animation and good at color rendering, graphic expression and multimedia integration, and these characteristics are just applied to express the spatial information. With the release of action script 3.0, lots of open source statistical charts are developed and can be easily integrated with SWF map files. Fig. 2 just shows an easy example of integration of SWF map and statistical chart.

5. CONCLUSION

As the development and progress of Adobe flash technology, flash has demonstrated high performance and super priority in publish web GIS maps and we just take a try to apply it in the visualization analysis of socio-economic data, which shows a very successful application. However, there are still many problems to be resolved in the wide range of application. Our application is just a lightweight application, and how to support the huge amount of data still need further study. The symbolization method and construction of symbol library are still in study. Also the varieties of geographical spatial analysis, including 2D and 3D analysis, are essential to the success application of flashgis.

6. ACKNOWLEDGEMENT

This work was supported by China Postdoctoral Science Foundation (Grant No. 20090460505) and project of Earth System Science Data Sharing Network belong to the national science and technology basic condition platform construction (Grant No. 2005DKA32300).

References:

Australian Broadcasting Corporation, 2004, Federal Election. http://www.abc.net.au/elections/federal/2004/electionmaps/ (accessed 17 Nov. 2009)

Dang Van Tuyen, et ak,, 2009, Developing a Flash-based tool for GIS appilciations on the web, http://www.geoviet.vn. (accessed 17 Nov. 2009)

ESRI inc., The ArcGIS API for Flex, http://resources.esri.com/arcgisserver/apis/flex/ (accessed 17 Nov. 2009)

Green, D.R., 1997. Cartography and the Internet. *The Cartographical Journal* 34 (1), pp. 23-27.

Itzhak Benenson, Michael Sofer and Izhak Schnell, 1998, Analysis of economic networks Geographical information systems as a visualization tool, *Applied Geography*, 18(2) pp. 117–135.

Jeremy Allaire, 2002, Macromedia Flash MX - A next-generation rich client, Macromedia White Paper.

Maguire, D. J., Goodchild, M. F. and Rhind, D. W., 1991, *Geographic Information Systems: Principles and Applications*. Longman, London.

Plewe, B., 1997. GIS-Online: Information Retrieval, Mapping and the Internet. OnWord Press, Santa Fe, New Mexico pp. 311.

Robin Hilliard, 2004, Handling seriously detailed maps in Flash, http://data.aad.gov.au/aadc/portal/index.cfm?file_id=1511 (accessed 17 Nov. 2009)

Strand, E.J., 1997. Java creates new channels for GIS information. GIS World, 5, pp. 28-29.

Wikipedia, Introduction to Adobe Flash http://en.wikipedia.org/wiki/Adobe_Flash (accessed 17 Nov. 2009)

William J. Drummond, 1993, GIS as a visualization tool for economic development, *Computers, Environment and Urban Systems*, 17(6), pp. 469-479.

VISUALIZING CLIMATE CHANGE IMPACT WITH UBIQUITOUS SPATIAL TECHNOLOGIES

R. M. Bennett [a, *], C. Pettit [b], J. P. Aurambout [b], F. Sheth [b], H. Senot [a], L. Soste [b], V Sposito.[b].

[a] Department of Geomatics, The University of Melbourne, Parkville, 3010, VIC, Australia - (rohanb, herve)@unimelb.edu.au
[b] Department of Primary Industries, Victorian Government, DPI Parkville Centre, 3052, VIC, Australia - (christoper.pettit, jeanphilippe.aurambout, falak.sheth, leon.soste)@dpi.vic.gov.au

Commission II, WG 6

KEY WORDS: Climate Change, Visualization, Ubiquitous Spatial Technologies, Google Earth

ABSTRACT:

This paper further articulates the role of ubiquitous spatial technologies (e.g. Google Earth) as tools for analyzing, visualizing, and developing policy responses to predicted climate change impacts. Specifically, the efficiency and effectiveness of using the tools in the production of visualizations for the local level is studied. A brief background to climate change response reveals limited data and visualizations at the local level: ubiquitous spatial technologies can potentially fill the void. Case study data including temperature, rainfall and land suitability information from southwest Victoria (Australia) are used to test the hypothesis. The research team produced thirty short visualizations using minimal time, resources and a moderate skill base. The effectiveness of the visualizations was tested on a diverse group of stakeholders. It was found that the visuals provided contextual information and understandings of overarching climate change trends, however, integration with other datasets and higher levels of detail are required if the platform is to be used as a stand alone policy development tool. Moreover, the need to further develop design guidelines to guard against, or at least inform users about visual sensationalism is required.

1. INTRODUCTION

Over the last decade climate change science increasingly pervaded mainstream media and political discourse: debates and strategies relating to climate change permeated all levels of society. Mitigation and adaptation strategies were evident in the actions of governments, businesses and individuals: carbon footprints were assessed, emissions trading and reduction schemes developed, and the potential impacts of sea-level rises were analyzed.

The success of these strategies is a product of the data and models used for their justification. Increasingly, there is a need for these models to integrate data from a range of sources: the complex nature of climate change response requires this multi-disciplinary approach (Bell et al, 2003). Maps and graphic visualizations are a powerful tool for enabling integrated analysis: spatial coordinates can unite disparate datasets and represent them on a single platform. The ability of computers to perform this task has long been recognized (DiBase et al, 1992; Max et al, 1993). Animated weather maps provide prime examples (Gardner, 1985). However, until recently, the production and use of these maps belonged to specialized scientific communities: they remained out of the reach to local decision makers and citizens.

Ubiquitous mapping tools such as Google Earth radically democratized spatial analysis and visualization. These tools provide great utility in the realm of climate change response: amateur users from a range of disciplinary backgrounds can easily engage with climate change models and visualizations. This utility has received much attention in recent years; however, literature describing the development process is limited. Moreover, the limitations and risks associated with democratized visualization demand further research.

To this end, this paper aims to further articulate the role of Google Earth as a tool for analyzing, visualizing, and developing integrated responses to potential climate changes. Specifically, the efficiency and effectiveness of using the tool in the production of visualizations for the local level is studied. Case study data from the southwest region of Victoria (Australia) is utilized. First, a brief background to climate change response and visualization is provided. This leads to a discussion of the study’s methodology: the selected region, characteristics modelled, scenario development process, visualization design and testing procedure are articulated. Results are then discussed using imagery and preliminary user feedback. The paper concludes with a discussion of the utility of using Google Earth for localized climate change analysis, visualization and integrated policy development.

2. BACKGROUND

2.1 Contemporary responses to climate change

Contemporary responses to climate change occur at a range of scales: global, regional, national and local responses are evident. At the global level, the United Nations (UN) drives the most recognizable responses. In 1992, subsequent to the Earth Summit in Rio de Janeiro, the United Nations Framework Convention on Climate Change (UNFCCC or FCCC) was conceived. The international environmental treaty led to the creation of the Kyoto protocol, a tool for reducing the production of greenhouse gases in industrial countries. Additionally, the Special Report on Emissions Scenarios

[*] Corresponding author.

(SRES) prepared by the UN's Intergovernmental Panel on Climate Change (IPCC) in 2001 used future emission scenarios to describe potential changes to the climate. Forty scenarios divided into four families (A1, A2, B1, B2) were compiled, each based on different economic, social and environmental assumptions. The scenarios are intended to assist with climate change assessment, mitigation and adaptation strategies.

Regional and national responses are most evident through the European Union's (EU) European Climate Change Programme (ECCP) and the accompanying European Union Greenhouse Gas Emission Trading Scheme (EU ETS). Australia and New Zealand are in the process of implementing similar schemes and more recently the United States has begun development of a cap-and-trade system. Where national consensus is delayed, state based approaches emerge: Illinois (Emissions Reduction Market System) and New York State (Regional Greenhouse Gas Initiatives) provide examples in the United States, whilst New South Wales (NSW Greenhouse Gas Abatement Scheme) provides an example in Australia. These tools are largely directed at mitigation rather than adaptation.

Local level responses have been impeded previously by limited awareness, lack of specialized knowledge and minimal information at the local or landscape scale (Dockerty et al, 2005). While some visualization and analysis tools were evident during the 1990s and early 2000s (Gordin et al, 1994; Wilby et al, 2002), the pervasiveness of new ubiquitous spatial technologies and emergence of scenario building techniques (Dockerty et al. 2006; Carter, 2001) have enabled more local community engagement with respect to climate change analysis and response. These local responses are a focus of this research.

In addition to becoming more localized, contemporary climate change responses also exhibit 'integrated' natures. The complexity and scope of climate change requires such an approach: datasets, models, scientific communities, policy-making groups, and the public are incorporated into the decision-making process. Climate change literature confirms this diverse group of stakeholders (Sheppard, 2005; Gordin et al, 1994), whilst Bell et al (2003) articulate the overarching benefits and difficulties of these integrated approaches. Dockerty et al (2005) and Sheppard (2005) both highlight the need for design guidelines and caution when developing climate visualization tools for diverse audiences. For example, over-emphasis of visuals might lead to inappropriate policy responses. Integrated responses are also a focus of this research.

2.2 Modern tools for visualizing climate change

The power of computers to enable visualizations of climate systems has long been recognized (Gardner, 1985; Max et al, 1993). DiBase et al (1992) explain how animated visualizations combined with static maps, graphs, diagrams, images, and sound improve scientific expression. More recently interactive visualizations emerged enabling a range of users to undertake personal explorations of environments and scenarios. Gorden et al (1994), Wilby et al (2002), and Stock et al (2007) illustrate the advances in these tools over the last two decades: realism and available level-of-details have dramatically increased. These characteristics are worth exploring individually: they impact greatly on the design of climate visualizations.

Whilst static photo-realistic visualizations have been available for at least the last decade (c.f. Sheppard, 2005; Bishop and Miller, 2007), interactive visualizations exhibiting photo-realism emerged more recently through advances made in gaming engines. This interactively, potentially in real-time, is being rapidly translated to the scientific visualization community (Buhmann et al, 2005). In relation to climate change, studies are being undertaken to determine the utility of photo-realistic visualizations for climate change decision-making. Bishop and Miller (2007) demonstrate the utility in relation to determining the visual attractiveness of wind farms. Dockerty et al (2006) illustrate the potential in relation to changes to rural and agricultural landscapes brought about by climate change. Sheppard (2005) provides many more examples, however, like Dockerty (2005; 2006), he concedes the potential exists for sensationalism and audience manipulation through these visualizations. As such, guidelines and rules for ethical design are proposed (c.f. Sheppard, 2005).

In contrast to photo-realistic visualization, more abstract visualizations are still highly relevant. Such visualizations are borne out the weather mapping tradition where points, lines and polygons are used to represent environmental phenomena not visible to the human eye. The abstracted visualizations of Wong et al (2002) provide examples: superfluous data is purposely smoothed out of the visualization to enable better human conceptualization. Abstracted visualizations are particularly relevant to climate change where illustrative tools are required at regional, state, national and global levels. In this way, Google Earth is uniquely placed: it is a highly ubiquitous and interactive platform enabling visualization at multiple levels of detail and scale. This utility will only increase: more high-resolution imagery will be added to the platform and more users will emerge. Whilst the ubiquitous nature of Google Earth makes it an extremely useful tool for presenting climate visualization, the mass amateurism of web mapping does present some problems. Guidelines for articulating the authenticity of data and guarding against sensationalism are only now emerging (Sheppard and Cizek, 2009). At any rate, in a controlled environment, Google Earth appears to hold great potential for climate change visualization and presentation. This potential requires further exploration.

3. RESEARCH DESIGN

3.1 Overview

The specific aim of this research was to further articulate the role of Google Earth as a tool for analyzing, visualizing, and developing integrated responses to potential climate changes at the local level. To this end, a number of interactive climate change visualizations were developed using Google Earth for a case study area (southwest Victoria). The utility of the platform was tested quantitatively in terms of the time, cost and skill-base required to produce the visualizations. Additionally, qualitative feedback from a diverse set of end users was also captured. In this way, the project used a mixed methodology: qualitative and quantitative research outputs were combined.

3.2 Case study area and scenario design

The Victorian Climate Change Adaptation Program (VCCAP), a Victorian government initiative, has investigated climate change impacts and adaptation within Victoria since 2007. It aims to ensure that the Victorian farming industry is equipped with knowledge of climate change science, potential adaptation strategies, and tools for maximising economical, social and environmental outcomes. As part of this project, The Department of Primary Industry developed a pilot research program (DPI VCCAP) focusing on the southwest region of Victoria (Figure 1). This region was specifically chosen for the wide variety of agricultural commodities grown and for the high level of community engagement in regard to climate change.

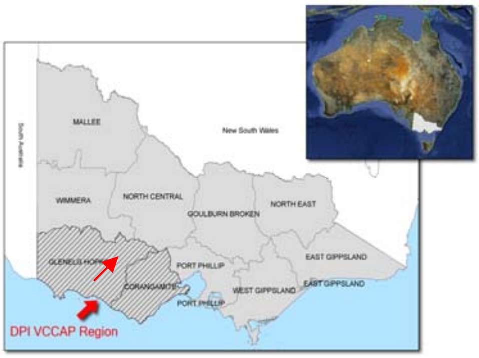


Figure 1. The DPI VCCAP study area

DPI VCCAP was guided by four key questions:(1) what are the impacts of climate change on agriculture, (2) what climate change adaptation options are available, (3) what are appropriate government policies responses and (4) how can the information be most effectively communicated (DPI, 2009). It aimed to answer these questions through multiple themes including: farming systems, scenario development, impact modelling and land suitability analysis, an e-resource centre and visualisation, communications and utilisation, and institutional adaptation and policy research.

The visualisation products developed through this project linked a number of these themes. They used data produced by the impact modelling and land suitability analysis modules, and were made available internally via the e-resource centre and externally via the Victoria Resources Online VCCAP website (http://www.dpi.vic.gov.au/DPI/Vro/vrosite.nsf/pages/climate_vccap). More specifically, they were used to inform the scenario development and analysis workshops.

The scenarios development process is now briefly discussed. Scenarios were developed around drivers of change for agriculture over which local primary producers have little or no control. These included projections of climate change and non-climate drivers. The following SRES/IPCC climate projections were used: A1FI (high growth, high carbon), A2 (divided world), and B1 (green energy). Localized climate models from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) were also used. The non-climate drivers included the global economy, trade barriers, consumer preferences, declining terms of trade, energy requirements, government policy including carbon pollution reduction schemes, dramatic change such as war or disease, and developments in science and technology. Additionally, relevant issues from regional stakeholders such as competition for land and attitudes of the urban community towards farming were included. Plausible options for how these drivers might unfold to 2050 were then built into the three scenarios. The scenarios were then analysed by a technical working group of 20 experienced stakeholders from within the region. They utilized their specialist knowledge and local experience to integrate the formal analyses with their understandings of business and community operations to provide a holistic assessment of the likely impacts and adaptive responses to climate change. These outputs were synthesised and communicated through workshops to key regional agricultural industries, agencies and policy groups.

3.3 Data acquisition

The data used in the visualizations was developed in the land suitability research theme of DPI VCCAP. The potential implications of climate change on the agriculture and forestry industry in southwest Victoria were investigated. Sposito et al (2008) modelled how projected climate changes could impact the capacity of southwest Victoria to produce a range of agricultural commodities and forestry products. The analysis used a GIS based multi criteria evaluation method to assess regional agricultural land use suitability. The model used a combination of biophysical data (soil, climate and landscape parameters) and expert judgment. The method produced GIS data layers (ESRI shapefiles) of land use suitability across the 3 climate change emission scenarios for 8 commodities: perennial ryegrass, phalaris, lucerne, barley, oats, winter wheat, blue gum and radiata pine (Figure 2). Additionally, average annual temperature and annual cumulative rainfall datasets were acquired from CSIRO.

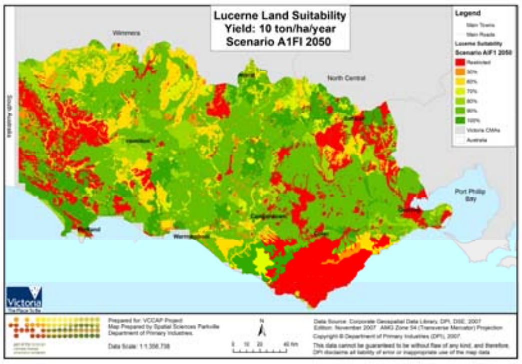


Figure 2. Example land-use suitability map (Sposito et al, 2008)

The multi-temporal datasets (2000 and 2050 epochs) produced were particularly difficult to communicate through standard paper reports and static maps: 3D visualisation techniques were therefore utilized.

3.4 Rainfall and temperature visualization development

First, data preparation was undertaken. This involved using ArcGIS to convert the temperature and rainfall features to raster for the 2000 and 2050 datasets. Second, a number of ArcGIS/Python scripts were developed to automate the frame production process. A script enabling the generation of 100 intermediate layers between the years 2000 to 2050 was developed. The visualization was intended to run for approximately 5 seconds: 100 was an appropriate number of layers. Linear interpolation was used: whilst less realistic than using individual datasets for each year, the smoothing better revealed the overarching trend. The intermediate layers produced by the script were converted to raster images. A KML file was then built: it located the raster images in space (extent) and time (span). A year counter, legend and view angle were included in the KML script. Finally, the visualization was composed in Google Earth. The KML file was opened and a tour was recorded. The collection of files was saved as a KMZ file to enable all elements to be contained in a single file, without external references.

3.5 Land-use suitability visualization development

First, the land-use data was prepared in ArcGIS. Feature classes were dissolved using a new column and a reclassification was undertaken. These were converted to raster and generalized by ‘border cleaning’: ascending order was used to privilege smaller areas. This generalization was then repeated. The resulting dataset was clipped to the relevant extent (as determined by the case study area) and was used to place symbols that represent areas effectively.

Second, the land-suitability data was prepared in ArcGIS. This was performed for each commodity variety (barley, oats, winter wheat, bluegum, lucerne, rye, pine, phalaris). The feature layers were converted to raster for the years 2000 and 2050. Reclassification occurred using percentages for non-negative values (10→100%; 9→90% etc.). Negative values were eliminated (→ NoData). Then datasets were clipped to the relevant extent. A feature point layer was created and roughly 15 symbols placed within the extent. Placement was determined by viewing the land-use layer and determining the areas aesthetically requiring symbols, and also by land-suitability values: symbol density was higher in high suitability regions.

Third, the symbols were prepared using Google Earth. Symbols for each crop were selected from various online libraries. Selection was based on semantics (how well the symbol illustrated the primary product e.g. milk bottles for cow pastures), 3D (for a more dynamic and appealing rendering), and performance (a low number of polygons was sought). The symbols were then scaled so that they were visible and appropriately proportioned compared to other symbols.

Fourth, a number of ArcGIS/Python scripts were developed to automate the layer production process. A script enabling the generation of 25 intermediate layers between the years 2000 to 2050 was developed. Linear interpolation of the land-suitability layers was used. Again, whilst less realistic than using individual datasets for each year, the smoothing better revealed the overarching trend. The intermediate layers produced by the script were converted to raster images. A KML file was then built. Again, it located the raster images in space (extent) and time (span). For each intermediate layer, the latitude and longitudes from the ‘symbol location’ layer were extracted along with the land-suitability value for that pixel(s). This information was used to place the symbol in the KML file, with height dimensions scaled in proportion to the land suitability value at the location. A year counter, legend and view angle were also included in the KML script.

Finally, the visualizations were composed using Google Earth. The KML files were loaded and a tour conducted. The complete set of files was then saved as a KMZ file without external references.

3.6 Testing the process and outputs

In order to test the efficiency of the process, indicators including total production hours, total costs ($AU), and required skills base were assessed. These were compared qualitatively against indicators for more traditional methods of production. Additionally, the effectiveness of the visualizations was tested using participants at a VCCAP workshop on July 22-23, 2009 at Warrnambool (Victoria, Australia). The quantitative outputs from the tests are not included here: these results are not the focus of this paper. Moreover, space does not permit their discussion. Instead, qualitative feedback from the session is used to inform the results.

4. RESULTS

4.1 The visualization products

In total, 30 individual animation sequences were produced: 8 commodities by 3 IPCC scenarios (A1FI, A2, B1); temperature by 3 IPCC scenarios; and rainfall by 3 IPCC scenarios. Each animation consists of ‘x’ raster data layers (between 2000 and 2050), a title, temporal labels, commodity symbols and a legend. Google Earth provides the remaining mapping infrastructure: orientation, scale, border, and underlying imagery source information. The animations run for approximately 5 seconds each. In addition to viewing the frames in sequence, the platform enables users to explore individual frames from each animation from multiple perspectives, scales and locations (Figures 3, 4, 5 and 6).

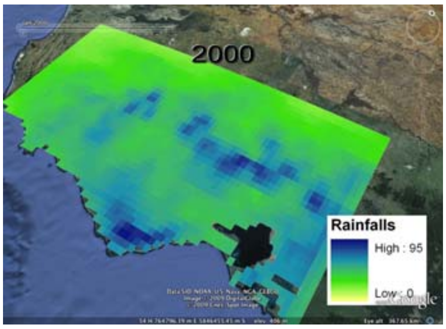


Figure 3. Rainfall animation: shades of blue and green are used to indicate rainfall amounts

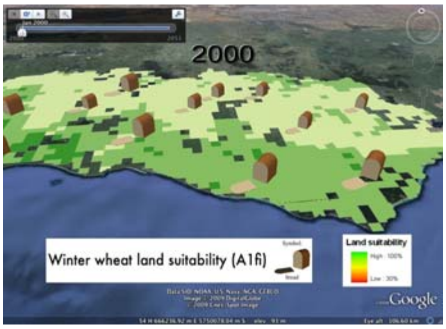


Figure 4. Winter wheat land suitability: shrinking/growing loaves of bread convey further meaning

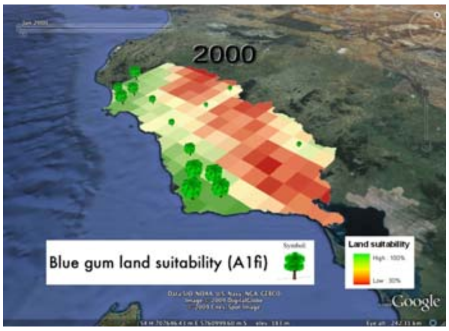


Figure 5. Blue gum land suitability for the A1FI scenario: growing symbols indicate increasing suitability

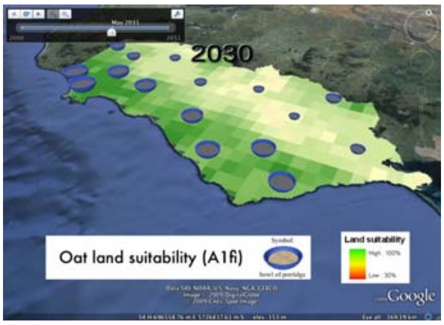


Figure 6. Oat land suitability animation: users can interact from various spatial and temporal perspectives e.g. 2030

4.2 Viewing and using the visualizations

The workshop provided access to an extremely diverse range of stakeholders including: farmers (dairy, sheep, cropping), environmental managers, social scientists, local community, emergency service workers, education workers, local government, state government, and planners. The group were exposed to the A1FI scenarios and allowed some guided interaction. Space limitations do not permit all qualitative comments to be reproduced here, however, Table 1 summarizes the overarching themes that emerged.

Theme	Description
1. An *overview* tool	The utility of the platform to provide an overarching context of potential changes occurring was recognized.
2. A *complimentary* tool	Alone, the visualizations were not enough to base decisions upon; however, they complemented tables and more specific data relating to the case study area.
3. A *collaboration* tool	The visuals provided a common language for the diverse range of stakeholders. The accessible visuals sparked discussions.
4. *Additional* data required	To be used for decision-making, more datasets would be required. Examples: social data (e.g. population, stress), environmental data (e.g. planning, sea-level rise, pests/diseases, fire/floods), economic (monthly household budgets, prices).
5. Higher *levels of detail* required	While visualizations could aid decision making at regional levels, higher resolution data would be required for the farm level.

Table 1. Themes from qualitative feedback

5. DISCUSSION

5.1 Ubiquitous spatial technologies: efficient and effective visualization platforms

Google Earth, a ubiquitous spatial technology, was found to be an efficient platform for developing climate change visualizations: good quality visualizations could be produced at low cost and within short timeframes. Coupled with ArcGIS, the tool enables fast production and accessible viewing of 3D visualizations. These characteristics have been lacking in other visualization platforms where specialized spatial knowledge was required to create and often interact with animations. However, while the overarching process can be seen as a success, a number of issues are worth further discussion.

5.2 Data and imagery: complementary in decision-making

Whilst images were found to be a useful tool for understanding overarching changes, some decision makers still desired more specific data in the form of tables or graphs. It is unclear whether this perceived limitation was a result of the user's limited exposure (or trust) to spatial technologies or whether the grid cells were too large and the legends unclear. At any rate, whilst this version of the visualizations did not provide data and graphs, Google Earth *can* be used to link text, data and graphs to geographic features. For example, upon clicking an individual land-suitability grid-cell a set of tables or graphs relating to the pixel could be displayed. More research is required to determine if visualization platforms can be used as the sole tool for landscape decision-making. It appears likely that both data and imagery will continue to complement one another in the medium term.

5.3 Abstraction vs. photo-realism: the debate continues

The grid-cell sizes and symbol scales used in the visualization pushed the visualization away from photo-realism towards a more abstracted environmental depiction. This was a conscious decision by designers: technological limitations in displaying vector graphics coupled with the low resolution of the raster datasets available guided the decision. Moreover, 'land-use suitability' cannot be perceived by the human-eye: some form of abstraction was therefore necessary. However, unconstrained resources would enable datasets at the parcel or paddock level to be produced. Additionally, technological advancements will improve the platform's ability to visualize large numbers of complex vector models simultaneously. Regardless, the 'abstraction' vs. 'reality' debate will continue to be an important design decision for any visualization project: the ability to produce photo-realistic products will not remove the issue.

5.4 Active vs. passive interaction: both are beneficial

The tool was found to promote engagement between users and the datasets. It is unclear whether tables of data would elicit a similar response from a diverse range of users, however, numerous respondent comments outlining the power of pictures and visuals suggest not. The demonstration was primarily moderator driver: respondents were able to dictate what was shown, however, they did not interact with the technology directly. Further testing of the individual interactions between the users, platform and data appears necessary. At any rate, the passive approach was found to maintain group focus and promote collaborative analysis. However, there appears to be great potential for more interactive approaches: enabling individual group members to move through the environment and express their ideas with points, lines, polygons or fuzzy zones could greatly enhance the collaborative utility of these visualizations. A parallel can be drawn with hands-on participatory tools such as touch tables or smart-boards and their ability to enable collaborative debate.

5.5 Utility in policy development: further work required

This paper focused on assessing the efficiency of using Google Earth for developing climate change visualizations: the potential of the visualizations to inform policy development and decision making was less thoroughly explored. Preliminary feedback suggests the tool has some utility in describing overarching trends. However, more detailed datasets, higher-resolution imagery, and integration with other forms of information such a tabular data and graphs would greatly enhance the application of ubiquitous spatial technologies as a participatory decision-making tool to inform planning and policy-making. The integration of additional datasets, functionality and trialling with stakeholders in an interactive session would be required to further test this hypothesis.

6. CONCLUSION

The utility of ubiquitous spatial technologies such as Google Earth to build community engagement and inform decision-making in relation to climate change holds great potential. While imagery detail and handling of complex vector graphics provide current challenges, these will be overcome in the near future. Longer-term challenges include the need to further develop and test design guidelines to guard against, or at least inform users about visual sensationalism. With the recent advent of Web 2.0 and collaborative visualization platforms there exists the research challenge to harness the enthusiasm of naïve cartographic users and visualisation producers. Whether this is through technology or educational means needs determination. At any rate, as the quality of freely available visualization products increases, ubiquitous visualization tools will play and important communicative and collaborative role in climate change policy responses.

7. REFERENCES

Bell, M.L., Hobbs, B.F., Ellis, H., 2003. The use of multi-criteria decision-making methods in the integrated assessment of climate change: implications for IA practitioners, *Socio-Economic Planning Sciences*. Volume 37, Issue 4, December, Pages 289-316.

Bishop, I.D., and Miller, D.R., 2007. Visual assessment of off-shore wind turbines: The influence of distance, contrast, movement and social variables, *Renewable Energy*. Volume 32, Issue 5, April 2007, Pages 814-831.

Buhmann, E., Paar, P., Bishop, I., Lange, E., (Eds.) 2005. *Trends in Real-Time Landscape Vizualization and Participation*. Herbert Wichmann Verlag, Heidelberg.

Carter, T.R., La Rovere, E.L., Hones, R.N., Leemans, R., Mearns, L.O., Nakicenovic, N., Pittock, B., Semonov, S.M., Skea, J., 2001. Developing and applying scenarios. In: McCarthy, J., Canziani, O.F., Leary, N., Dokken, D.J., White, K.S., (Eds). *Climate Change 2001: Impacts, Adaptations and Vulnerability*. Cambridge University Press, Cambridge/New York.

DiBiase, D., MacEachren, A.M., Krygier, J.B., and Reeves, C., 1993. Animation and the Role of Map Design in Scientific Visualization, *Cartography and Geographic Information Science*. Volume 19, Number 4, October, pp. 201-214(14).

Dockerty, T., Lovett, A., Appleton, K., Bone, A., and Sunnenberg, G., 2006. Developing scenarios and visualisations to illustrate potential policy and climatic influences on future agricultural landscapes, *Agriculture, Ecosystems & Environment*. Volume 114, Issue 1, May 2006, Pages 103-120.

Dockerty, T., Lovett, A., Sunnenberg, G., Appleton, K., and Parry, M., 2005. Visualising the potential impacts of climate change on rural landscapes, *Computers, Environment and Urban Systems*. Volume 29, Issue 3, May 2005, Pages 297-320.

DPI, 2009. Victorian Climate Change Adaptation Program, Victorian Resources Online – Statewide. http://www.dpi.vic.gov.au/dpi/vro/vrosite.nsf/pages/climate_vccap (accessed 25 November 2009)

Gardner, G.Y, 1985. Visual Simulation of Clouds, *ACM SIGGRAPH Computer Graphics*. vol. 19, no. 3, 259-268.

Gordin, D.N., Polman, J.L., and Pea, R.D., 1994. The Climate Visualizer: Sense-making through scientific visualization, *Journal of Science Education and Technology*. Volume 3, Number 4 / December, 1994, Springer.

Lange, E., Bishop, I.D., 2005. Visualization in landscape and environmental planning: technology and applications. Taylor and Francis, United States.

Max, N., Crawfis, R., and Williams, D., 1993. Visualization for Climate Modeling, *IEEE Computer Graphics and Applications*, vol. 13, no. 4, pp. 34-40, July/Aug.

Nicholson-Cole, S.A, 2005. Representing climate change futures: a critique on the use of images for visual communication, *Computers, Environment and Urban Systems*. Volume 29, Issue 3, May, Pages 255-273.

Scheraga, J.D., Ebi, K.L., Furlow, J., and Moreno, A.R., 2003. From science to policy: developing responses to climate change, In McMichael, A.J., *Climate change and human health: risks and responses*. World Health Organization.

Sheppard, S.R.J., 2005. Landscape visualisation and climate change: the potential for influencing perceptions and behaviour, *Environmental Science & Policy*. Volume 8, Issue 6, December, Pages 637-654.

Sheppard, S.R.J., Cizek, P., 2009. The ethics of Google Earth: crossing thresholds from spatial data to landscape visualisation. *J. Environ. Manage*. 90, 2102–2117.

Sposito, V.A., Pelizaro, C., Benke, K, Anwar, M., Rees, D., Elsley, M., O'Leary, G., Wyatt, R. and Cullen, B., 2008. *Climate change impacts on agriculture and forestry systems in South West Victoria, Australia*. Department of Primary Industries, Future Farming Systems Research Division. DPI Parkville Centre.

Stock, C., Bishop, I.D., and Green, R., 2007. Exploring landscape changes using an envisioning system in rural community workshops, *Landscape and Urban Planning*. Volume 79, Issues 3-4, 2 March, Pages 229-239.

Wilby, R.L., Dawson, C., and Barrow, E.M., 2002. SDSM—a decision support tool for the assessment of regional climate change impacts, *Environmental modelling and software*. Volume 17, Issue 2, 2002, Pages 145-157.

Wong, P.C, Foote, H., Leung, R., Jurrus, E., Adams, D., and Thomas, J., 2002. Vector Fields Simplification - A Case Study of Visualizing Climate Modeling and Simulation Data Sets, *The 11th Ann. IEEE Visualization Conference*. Salt Lake City, UT, October 08-13.

8. ACKNOWLEDGEMENTS

The authors would like to acknowledge our colleagues at the Department of Primary Industries (Victorian Government) and Department of Geomatics (The University of Melbourne), particularly Prof Ian Bishop, for assisting with this project.

VISUALIZING FUTURE BIOLINKS USING A TOUCH TABLE – NEW DIMENSIONS IN PLANNING

C. Bhandari [a], *, S. C. Sharma [b], I. D. Bishop [a], C. Pettit [b]

[a] Cooperative Research Centre for Spatial Information, University of Melbourne, 723 Swanston Street, Parkville Vic 3052 – (bhandari, i.bishop)@unimelb.edu.au
[b] Dept. of Primary Industry,Parkville Centre,32 Lincoln Square North, Carlton Vic 3053 – (subhash.sharma, christopher.pettit)@dpi.vic.gov.au

Commission II / II/6

KEY WORDS: Visualization, Touch table, biolinks, participatory-decision making

ABSTRACT:

Visualization is an emerging collaboration tool to support landscape planning. Integration of collaborative visualization systems and GIS, using a touch screen table has the potential to become an essential part of land-use and environmental planning and so to facilitate better decision making.

The paper focuses on a method to interactively plan for future forest regeneration areas (*biolinks*) with multiple stakeholders and conflicting objectives. This project automated three-dimensional representations of landscapes comprising key Ecological Vegetation Classes (EVC) that can be used to generate and visualize future biolink scenarios interactively. The approach involved: creation of a rule based database to generate landscape objects and EVCs; new GIS based software to facilitate freehand drawing over regional based maps and the use of touch table technology to enable participants to collaboratively design biolink scenarios.

1. INTRODUCTION

Land clearing for agriculture, or other purposes, leaves remnants of vegetation which may not be sufficient for the survival of native flora or fauna. Provision of ecological connectivity in the landscape to allow for species re-colonizations and migration has been widely called for (Brereton et al., 1995; Hilty et al., 2006; Pham and Wacher, 2004; Soule et al., 2004). Provision of areas for renewed linkage, commonly called biolinks in Australia (e.g. Mansergh et al., 2008), should be carefully planned to provide maximum ecological benefit with constraints of cost, land ownership, land suitability and other factors.

Multi-factorial and multi-criteria landscape planning with spatial information systems has a long history beginning with McHarg (1969) using manual overlays. This work provided the impetus for the computer mapping systems (such as SYMAP) and analysis tools (MAP) which evolved into geographic information systems (GIS). As ecological connectivity also has a large impact on the visual landscape, visualization tools offer an important additional component to inform the debate about future landscapes. Visualization tools, particularly when dynamically linked to GIS and realistic image libraries offer a powerful medium to assist the community, scientists, planners and policy-makers to more actively participate in the planning and redesign of new landscapes. When combined with appropriate ecological data such systems provide the facility for identification of areas for effective use of planting and regeneration to create cost effective biolinks. These digitally created landscapes that display physical properties of the real world provide a medium of visualizing the outcomes of different management approaches aimed at building in biodiversity conservation (Mansergh et al., 2008).

In this project, integration of a touch table, a visualization tool-Spatial Information Exploration Visualization Environment (SIEVE) (Stock et al., 2007) and GIS –based tools, allowed the effective use of spatial information in participatory decision-making of land use plans for future biolinks. The touch table uses a touch enabled screen, which allows multiple participants to view the screen and interact with the data using hand movements (gestures). The information can then be visualized as a pseudo-realistic three dimensional landscape using SIEVE. The essential GIS data includes EVCs (Ecological Vegetation Classes), DEM (Digital Elevation Model) and aerial photography. In our case study, these integrated technology systems have been used to communicate and exchange knowledge amongst scientists and stakeholders while viewing the existing physiographic conditions of the study area in the greater Grampian region, Southern Victoria, Australia.

2. BACKGROUND

Touch table systems assist in managing and viewing large amounts of data, primarily focussed on GIS, in a manner conducive to discussion and decision making (Arciniegas and Jensen, 2009). A common problem when dealing with large volumes of data is seeing the overall picture. When using high resolution GIS data like satellite images/ aerial photographs, DEM (Digital Elevation Model) etc, conventional displays are too small and limit the number of individuals that can access the data and engage in the discussion. Also, the problems stemming from conventional presentations are magnified since they limit the role of most participants to passive listener. (http://www.army-technology.com/contractors/data_management/touchtable/) (accessed 05 Nov 2009).

Touch table systems not only increases group comprehension and help decision making through better visualization but also helps with data management and improved participation. Interacting with touch tables is similar to working with a paper map (2D). A computer interprets the location of the hand movements on the touch surface. A single touch can be used to query source data, turn data/group layers on or off, zoom in and zoon out, set data classification levels and lock screen display. Anyone in the group can access or manipulate data simply with a touch on the touch surface of the screen as no one person controls the mouse.

A number of different types of touch table are available in the market, each one of them having specific characteristics. The choice in this case was made considering a number of factors and parameters such as product's features, specifications, applications, price and availability in Australia. Each touch table under a different manufacturer brand had a distinguishing property, such as; Object Recognition, Touchshare GIS software, Gesture-Control and Multi-Touch Surface Computing, Multitouch, Multiuser (WHO IS WHO). The brands and their key features are summarised in Table 1. Our choice was the 52" multi sensor touch screen table from Touch Screen Solutions, affordable, efficient and user friendly. The Touch Screen table integrates well with our GIS interface (ESRI's ArcMap) and the BioZone Constructor extension made for assisting Biolink planning and design. The touch tables's screen can be integrated with multi touch sensor technology using its "Infrared Multi-point Touch Screen Application Programming Interface (API)".

Product	**Diamond Touch**	**Microsoft Surface**	**TT45 & TT84**	**Touch Screen Solutions**	**Gesturetek**
Multitouch Multiuser	YES YES	YES YES	YES YES	YES YES (using API)	YES YES
Projection Type	Front Projection	Rear Projection (Cameras and infrared spectrum)	LCD	LCD	LCD
Customisable/API	YES	YES (SDK)	YES (Touchshare GIS software)	YES	YES
SPECIFICATIONS	32" & 42"& larger screens made to order	30" Display	45" & 84" respectively	32", 42", 50",52" larger screens made to order	50" Display, different sizes screens made to order
APPLICATIONS as Advertised	DiamondTouch plug-in tool for ESRI's ArcGIS.	Stores, Restaurants, Hotels and GIS, object recognition using barcode	GIS Specific-sits on top of ESRI's ArcGIS or Google Earth	Commercial like a PC or Laptop	Everywhere
PRICE	$9500 (USD) for single unit & $12500 (USD) with SDK	$12500 (USD) for single unit & $15000 (USD) with SDK	TT45 $59000 (USD) TT84 $1,79000(USD)	$8000-$13000 (AUD)	$69542 (AUD)
AVAILABLITY (in Australia)	YES	NO	YES	YES	YES

Table 1. A comparison of the key features, prices and availability of touch tables in the market (Information collected in May 2009)

3. CASE STUDY

The BioZone Constructer (figure 1) software assists users to plan future biolinks. BioZone Constructor is a standalone application built within ESRI's ArcGIS on VB.NET which enables users to build biolinks through the ArcMap interface (2 dimensional) and view the resultant landscape in SIEVE Viewer (3 dimensional).

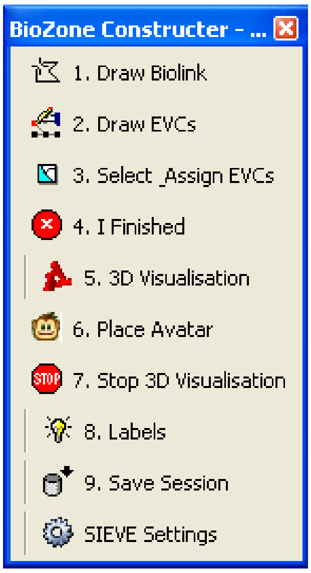
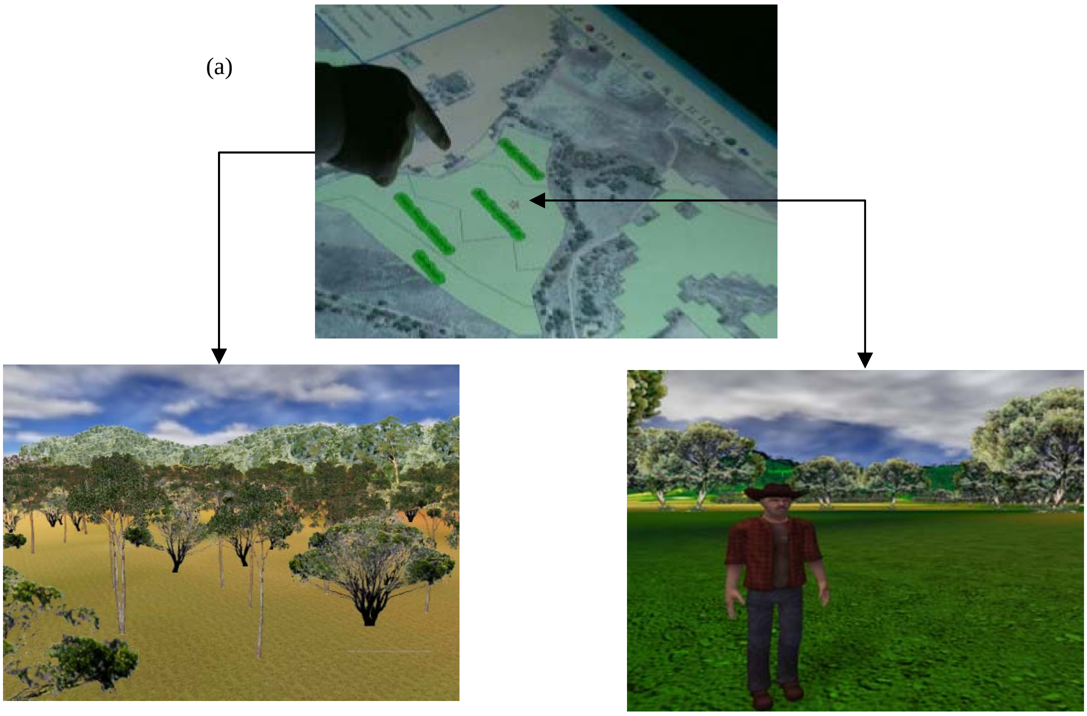


Figure 1. Screenshot of BioZone Constructor

BioZone Constructor contains tools to support participatory decision-making processes. The users choose these GIS based tools one by one to help them to construct the BioZones, divide them into Ecological Vegetation Classes (EVC) and label them with their respective names in the ArcMap interface. After finishing this task they will activate the live link with SIEVE Viewer where they will see all constructed BioZones within the existing 3D landscape filled with the correct vegetation types, heights and densities for the chosen EVCs. At the completion of the participatory planning and design process end users can save the BioZones into the Geodatabase.

The first tool in the BioZone Constructor is called the 'Draw Biolink' which enables the user to draw a biolink polygon using fingers, pen or stylus on the surface of the touch enabled screen. Every polygon made by the user is given a name of choice by the user and this shapefile automatically generates its own attribute table including area, perimeter and space for the EVC name of the polygon. Each user can make multiple polygons with different names.

Once the biolink has been drawn the user can subdivide their polygon into numerous sub-polygons using the second tool called the 'Draw EVCs'. These sub-polygons can now be assigned to one of the ten dominant EVCs in the Grampian area. In addition, the options of 'Plantation', 'Windbreak' and 'No EVC' were added as other landscape elements. Each EVC's flora type was limited to trees and shrubs only. Once the users complete drawing and assigning names to their polygons they can finish it by using the 'I finish' tool.



(b) (c)

Figure 2. Screenshots from Biolink planning process (a) ArcMap interface on Touch screen, (b) showing Birds eye view of the a proposed Biolink, and (c) Avatar position in SIEVE within Biolink

The next step in this process is to activate the 3D visualization interface - SIEVE, which allows users to explore existing spatial data and hypothetical future scenarios (newly constructed biolinks) in a real-time 3D environment. SIEVE incorporates a multi-user environment that allows users from different locations to gather in the virtual landscapes for exploring and decision-making purposes (figure 2). This functionality is frequently implemented in game engine software, on which SIEVE is built, and users are represented in the virtual world as so-called avatars. Using avatars in a collaborative environment has several advantages. It makes it easy to recognize other people and gives a feeling of physical presence. During collaborative discussions it may be important to know what features other people are looking at, or even where they are located. This can be especially useful if the meeting is run with participants involved from different locations either via a network setup or possibly with some participants involved through the use of augmented reality sets situated in the actual study area and their avatars reflect their true physical position (Chen et al., 2008; Stock et al., 2007). Remote users can communicate via a chat box and our avatars have been programmed to make gestures appropriate to the message being typed (e.g. nodding for 'yes').

The sixth tool of BioZone Constructor, 'Place avatar', helps to the user to place an avatar which means that with just a click in ArcMap interface(2D), one will start viewing an area of choice in SIEVE. "Stop 3D visualization of biolinks" allows users to disconnect from the SIEVE environment while the "Labels" tool allows users to view the names of the EVCs. Users can always switch the names on or off for their convenience. "Save session" writes output of the work session in a new Geodatabase automatically with the respective date of the work session. Users can start the SIEVE viewer on the 'local machine' or on the 'network machine' by entering the IP address of the machine into the "SIEVE Settings" window's form. Hence users can view the 3D visualization from a different location while the main Touch table interface remains in the original location.

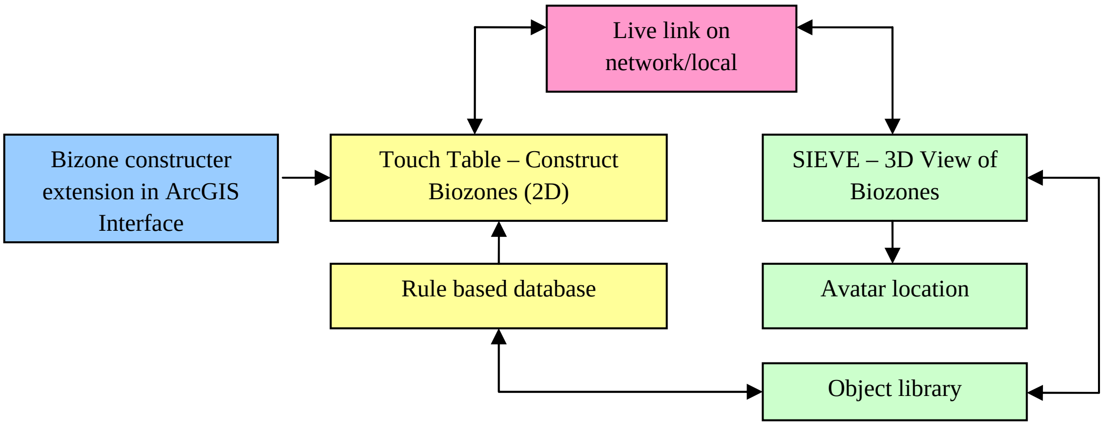

Live link on network/local
Bizone constructer extension in ArcGIS Interface
Touch Table – Construct Biozones (2D)
SIEVE – 3D View of Biozones
Rule based database
Avatar location
Object library

Figure 3. The workflow for the BioZone Constructor system

The workflow in Figure 3 illustrates how the BioZone Constructor functions for the scientists, planners and stake holder in a collaborative environment. The main component is the touch enabled screen which gives a chance to the participants to conduct a participatory and collaborative discussion whilst designing a Biolink on the touch screen. Any relevant background data (soils, land ownership, slope etc…) can be made available through the GIS view on the touch table. This gives users the opportunity to consider these factors in their biolink delineation. Using the GIS functions they can also measure distances or even run suitability analysis. The areas of drawn biolinks are available through the attribute tables.

The BioZone Constructor tools described above give users freedom to prepare multiple scenarios and to recorrect or reassign the EVC classes during the discussion. Once they have agreed on one element they might move further. The other important aspect of the system is the back end database i.e rule based database created in Microsoft Access. This database contains all attribution information of each type of object and defines some rules to project them into the 3D SIEVE environment.

The Live Link functionality between the GIS and SIEVE platform plays a significant role to connect the touch screen with a local machine or a network machine. All participants can view the designed biolink in the virtual environment containing all respective EVC species. This gives the resemblance of the future biolink in the real world. At this stage in the project, the vegetation objects are stored in SIEVE directories locally but in the future the objects will be retrieved from a server based Object Library.

Once the transformation of the biolinks is complete the users have the option to move around in the virtual environment and view the location of their Avatar on the Touch screen. This particular approach gives freedom to the participants to view different locations holding various perceptions instantaneously while also knowing where they are relative to key GIS themes. Exploration of the virtual environment can be undertaken from an above ground perspective and also from an on the ground perspective.

4. CONCLUSION

With the help of BioZone Constructor improved biodiversity and habitat plans can be created between conservation areas that were once fragmented leading to the likely loss of habitat. The current prototype system can provide:

1. A more informed way of understanding environmental processes.

2. An interactive participatory decision making approach to involve multi-stakeholders in landscape planning and design process.

3. A fast and engaging way to plan and design future biolink scenarios both in two dimensions and three-dimensions.

We note that the current prototype system does not support multi-touch functionality. However, this would be possible to implement through the multi-user touch technology using the application programming interface or migrating the current system onto a touch table platform which already supports multi-touch functionality such as the Microsoft Surface.

5. FUTURE RESEARCH

What we have reported in this paper is the development of a prototype touch table participatory planning tool. The next steps of this research will include advances in both technology and application development.

Specifically the next steps in the research and development include:

1. A more robust and comprehensive object library database platform: Both the Victorian Department of Primary Industries (DPI) and Victorian Department of Sustainability Environment (DSE) are working on vegetation and/or infrastructure libraries with appropriate management tools (Pettit et al. 2009). We anticipate building these into the product as it develops. The object library would include a wide range of vegetation types and vernacular infrastructure elements representing local landscapes.

2. More knowledge driven case studies in the landscape: a market driven project that allows scientists, stakeholders and landholders to calculate the steps and costs to improve an ecosystem. This could revolve around a number of critical natural resource management issues such as climate change, weeds, bushfires, endangered flora and fauna and their habitat and soil salinity to name a few.

6. REFERENCES:

Arciniegas, G.A. and Janssen, R., 2009. *Using a touch table to support participatory land use planning*, 18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009. http://mssanz.org.au/modsimon09, pp. 2206-2212 (accessed 3 Nov. 2009)

Chen, T., Stock, C., Bishop, I. and Pettit, C., 2008. Automated Generation of Enhanced Virtual Environments for Collaborative Decision Making Via a Live Link to GIS, in *Landscape Analysis and Visualisation: Spatial Models for Natural Resource Management and Planning*, (Eds. Pettit, C., Cartwright, W., Bishop, I., Lowell, K., Pullar, D. and Duncan, D.), Springer, Berlin, pp 571-590.

Mansergh, I., Lau, A. and Anderson, R., 2008. Geographic Landscape Visualization in Planning Adaptation to Climate Change in Victoria, Australia, in *Landscape Analysis and Visualization: Spatial Models for Natural Resource Management and Planning*, Lecture Notes in Geoinformation and Cartography.(Eds. Pettit, C., Cartwright, W., Bishop, I., Lowell, K., Pullar, D. and Duncan, D.), Springer, Berlin, pp. 469-487.

O'Connor, A., Stock, C. and Bishop, I., 2005. *SIEVE: An Online Collaborative Environment for Visualising Environmental Model*, MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005 Outputs". http://www.mssanz.org.au/modsim05/papers/oconnor.pdf, pp. 3078-3084 (accessed 2 Nov 2009)

Pettit, C.J., Sheth, F., Harvey, W. and Cox, M., 2009. *Building a 3D Object Library for Visualising Landscape Futures*. In proceedings of 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Queensland, 13-17 July, pp. 2244-2250.

Stock, C., Bishop, I.D., O'Connor, A.N., Chen, T., Pettit, C.J. and Aurambout, J-P., 2008. *SIEVE: Collaborative Decision-making in an Immersive Online Environment*, Cartography and Geographic Information Science, 35(2): 133-144.

Stock, C., Pettit, C., Bishop, I.D. and O'Connor, A.N., 2005. *Collaborative Decision-Making in an Immersive Environment Built on Online Spatial Data Integrating Environmental Process Model,* MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005 http://www.mssanz.org.au/modsim05/papers/stock.pdf, pp. 3092-3098 (accessed 03 Nov 2009)

7. ACKNOWLEDGEMENTS:

Investment from the Victorian Department of Primary Industries and the Victorian Department of Sustainability and Environment (DSE) has made this research project possible. The CRC-Spatial Information is acknowledged for the use of the SIEVE technology platform. Special thanks to Alex Lau, Jean-Philippe Aurambout and Falak Sheth for their review and testing of the touch table technology platform.

3D DATA VISUALISATION WITHIN SPATIAL DECISION SUPPORT SYSTEMS BY USING ARCGIS ENGINE

R. Laudien*, A. Christmann, S. Brocks

Dept. of Geography (GIS & RS), University of Cologne, Albertus-Magnus-Platz, Cologne- rlaudien@uni-koeln.de, christmann@www-dev.de, mail@sebastian-brocks.de

Commission II, WG 7

KEY WORDS: Software engineering, 3D modelling, Spatial Decision Support, ArcGIS Engine

ABSTRACT:

Based on the current state of science and technology, computer based Spatial Decision Support Systems (SDSSs) can be used for modern management and planning purposes. To fulfil the requirements of being advanced comprehensive systems, such modern SDSSs provide access to individual data and computer models. They combine tools and analysis functions of Geographical Information Systems (GIS), Decision Support Systems (DSS), Remote Sensing (RS), and numerical, expert or statistical models. In addition to that, such complex systems include expert knowledge and enable the user to interact with the SDSSs during run-time. To generate such user specific results, SDSSs need to access several different datasets with different formats. According to a given adapted logical decision tree, these systems are programmed in a modular and question-specific way.

This contribution shows a design and development of visualising three dimensional data within a decision support framework by using ArcGIS Engine and Java. Besides the description of the methodological approach, an example demonstrates the functionalities of the developed 3D-ArcGIS-Scene-Panel.

INTRODUCTION

Spatial data, mathematical models and expert knowledge are currently more and more incorporated in decision making processes (Bareth, 2009). Systems which compute these procedures need to access several different datasets with different formats. Modern Spatial Decision Support Systems (SDSSs) combine such data within complex process-based toolboxes (Leung, 1997; Wright, 1993; Sharma *et al.* 2006). According to a given logical decision tree, these systems are programmed question-specific (Hartkamp *et al.* 1999; Seppelt, 2003). Thus, detailed information about the requests, knowledge and personal needs of the potential users are essential (Keenan, 2006), and the developed systems need to be adapted based on the GIS- (Geographical Information System), RS- (Remote Sensing), and model-knowledge of the decision makers (Laudien and Bareth, 2007).

To meet the requirements of being comprehensive decision support tools, the individually generated components of a SDSS need to be connected (Bareth and Yu, 2007; Taylor *et al.* 1999; Rizzoli *et al.* 1997). This matter of fact is realised by coupling these modules with individually developed interfaces (Schneider 2003; Shaffer *et al.* 2001; Hartkamp *et al.* 1999, Seppelt, 2003). The interfaces access the system modules, e.g. data, models, GIS-visualizations or RS-analyses. By following this approach, models are independent from the SDSS body and vice versa. Only the results of each subsystem are exchanged (Longley *et al.* 1999). Thus, it is possible to develop different SDSS components at the same time independently from each other.

3D GIS are GIS systems providing data structures and operations for the presentation, management and processing of points, lines, surfaces and volumes in three dimensional spaces. (Breunig, 2001).

An increased interest in modelling 3D data like digital elevation models or city models in conjunction with new methodological approaches implementing three dimensional models and visualizations provide a high degree of relevance regarding planning functions (Henning, 2008). Therefore, this paper focus on the design and development of visualising three dimensional data within a decision support framework by using ArcGIS Engine and Java. The programming approach described in this paper was developed for the interdisciplinary research project IMPETUS (*An integrated approach to the efficient management of scarce water resources in West Africa*). One major task of the third project phase of IMPETUS (2006-2009) was the development, implementation and application of GIS-based SDSSs to support water resource management in selected river catchments of Benin and Morocco. The programmed systems were implemented in a Java/XML based framework (Enders *et al*. 2007; Enders and Diekkrüger,2009). Further information concerning IMPETUS is documented by (Speth *et al*. 2005).

SPATIAL DECISION SUPPORT SYSTEMS

As soon as spatial data is embedded in a Decision Support System, GIS functionalities get an important role. These functions support the user making spatially diverse decisions. In this context, the term Spatial DSS was established in the mid 1980s/1990s (Armstrong *et al.* 1986; Armstrong *et al.* 1990; Densham 1991; Goodchild *et al.* 1993; Wilson 1994; Crossland *et al.* 1995). Mennecke (1997) sees a SDSS as an easy-to-use subset of a GIS, which incorporates facilities for manipulating and analysing spatial data. In addition to that, Malczewski

(2006) addresses that a SDSS also includes functionalities of multicriteria decision analysis. Furthermore, SDSSs provide the opportunity of integrating, visualizing and evaluating various analytical models, and therefore can be used to develop management strategies (Muller, 1993; Keenan, 1996; Leung, 1997; Malczewski, 1999; Manoli *et al.* 2001; Yeh 1999). The structure of such systems is introduced considering latest technology developments and can be considered as the geo-data infrastructure for decision support related to specific spatial based questions and problems (Bareth, 2009). The SDSSs are connected to extensive geo-databases which include (i) all data for the system modelling, (ii) the models itself in a Model base Management System (MBMS), (iii) the interfaces between data, models and user knowledge, as well as (iv) the functions for data generation and data mining (Leung, 1997)

DESIGN AND DEVELOPMENT

In the interdisciplinary research project IMPETUS (Speth *et al.* 2005), the object oriented programming language Java was used to develop platform independent SDSSs. Java allows the modelling of numerous scientific questions. It reduces the complexity of a program source code in terms of abstraction, encapsulation, and defined interfaces. Object oriented programming donates an appropriate representation of the relationships between classes and objects, and increases the developer efficiency regarding the reuse of software code (Krüger, 2006). Within this programming environment, the source code is translated into byte code and then executed in a special environment, the so called Java Runtime Environment (JRE). The major part of the JRE is the Java Virtual Machine (Java-VM) which interprets and executes the byte code. Because Java-VMs exists for multiple different Operating Systems, the major advantage of programming in Java consists in the fact that all developments can run on different computers and different operating systems. Therefore, software which is developed with Java is nearly platform independent (Herter and Koos, 2006).

To fulfil the given requests of guaranteeing GIS- and RS-functionalities within the SDSSs, the ESRI developer library ArcGIS Engine 9.2 was used. With ArcGIS Engine, the software developer gets the opportunity to implement spatial analysis functions. The full version of ArcGIS Engine, which is accessible by C++, VB.NET, C#, or Java, comes with several different extensions which can be integrated, based on the users' needs. To execute the SDSSs, the JRE and the ArcGIS Engine Runtime Environment need to be installed on the computer. The IMPETUS SDSSs are developed using the programming environment Eclipse SDK which contains the Eclipse platform (Eclipse 3.2, (http://www.eclipse.org/), tools for Java programming and the environment to develop Eclipse plug-ins. In addition to that, the Subclipse plug-in is used to store the source code in a subversion repository (SVN). Such a repository allows several software developers to work on the same project independently without causing backup or versioning errors. Besides the standard Java GUI components, specific GIS- and RS-components are used for the programming. These include functionality available through the ArcGIS Engine library.

The (geo-) data (e.g. rasters, vectors, and alphanumerical data) of the single SDSSs are stored in ESRI 9.2 file-based geo-databases.

For the 3-D visualisation special framework processors and functionalities were developed. The following selection of code excerpts illustrate configuration of the *Multi Document Desktop* as well as some of the dependent processors within the client configuration. All objects instantiated during run time are stored in a global parameter map, thus instantiated objects and data residing in memory are accessible by object names.

The following code excerpt from the configuration xml demonstrates how feature classes are read from a file based geo database into memory to be further processed. The processor is capable to load several feature classes at once.

```
<processor
class="de.isdss.ext.arcgis.processors.ArcGISLoadFeature
ClassesProcessor">

<param key="readFile">${_CACHEDIR_}/farmadam/BeL4.gdb
</param>
<param key="outWorkspace">geodb.workspace</param>
<param key="featureclasses">14</param>
<!-- General Purpose Feature classes -->
<param key="inFeatureclass.1">main_cities</param>
<param key="outFeatureclass.1">fc.cty</param>
<param key="inFeatureclass.2">streets</param>
<param key="outFeatureclass.2">fc.str</param>
<param key="inFeatureclass.3">regions</param>
<param key="outFeatureclass.3">fc.reg</param>
<param key="inFeatureclass.4">departments</param>
<param key="outFeatureclass.4">fc.dep</param>
<param key="inFeatureclass.5">gemeinden</param>
<param key="outFeatureclass.5">featureclass.com</param>
<param key="inFeatureclass.6">AEZ</param>
<param key="outFeatureclass.6">fc.aez</param>
<param key="inFeatureclass.7">regions_Line</param>
<param key="outFeatureclass.7">fc.reg.line</param>
<param key="inFeatureclass.8">departments_Line</param>
<param key="outFeatureclass.8">fc.dep.line</param>
<param key="inFeatureclass.9">gemeinden_Line</param>
<param key="outFeatureclass.9">fc.com.line</param>
<param key="inFeatureclass.10">AEZ_Line</param>
<param key="outFeatureclass.10">fc.aez.line</param>
<!-- Point Feature classes to be cloned later -->
<param key="inFeatureclass.11">gem_for</param>
<param key="outFeatureclass.11">fc.com.ori</param>
<param key="inFeatureclass.12">dep_for</param>
<param key="outFeatureclass.12">fcdep.ori</param>
<param key="inFeatureclass.13">reg_for</param>
<param key="outFeatureclass.13">fc.reg.ori</param>
<param key="inFeatureclass.14">aez_for</param>
<param key="outFeatureclass.14">fc.aez.ori</param>
</processor>

<!-- Fields for Land Use -->
<processor
class="de.isdss.ext.processors.NewArrayProcessor">
<param key="name">array.attributes.landuse</param>
<param key="element.0">crops</param>
<param key="element.1">forest</param>
<param key="element.2">fallow</param>
<param key="element.3">pasture</param>
<param key="element.4">savanna</param>
<param key="element.5">other</param>
<param key="elements">6</param>
</processor>
```

To visualize the model data results, the values must be read from the data source. In this case a MS Excel sheet file containing research data and model calculation rules, that are processed each time this processor, is called providing a set of results in memory during run-time.

```
<processor
class="de.isdss.ext.processors.GetXlsArrayProcessor">
<param key="file">
${_CACHEDIR_}/farmadam/BeL4_data.xls
</param>
<param key="sheet">FARMADAM</param>
<param key="rows">10</param>
<param key="cols">
E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD
,AE,AF,AG,AH,AI,AJ,AK,AL,AM,AN,AO,AP,AQ,AR,AS,AT,AU,AV,
AW,AX,AY,AZ,BA,BB,BC,BD,BE,BF,BG,BH,BI,BJ,BK,BL,BM,BN
</param>
<param key="name">array.com.ids</param>
```

```
</processor>
```

After loading and calculating model data and feature data into memory, those two data sources must be merged together to create a visualization to be used later. For this purpose processors were developed to manage and manipulate resources during execution of the application.

```
<processor
class="de.isdss.ext.arcgis.processors.ArcGISAddFeatureL
ayerAttributesProcessor">
<param key="featureclass">layer.communes</param>
<param key="order">rows.com</param>
<param key="ids">ids.com</param>
<param key="type">lu.types</param>
<param key="attribute">lu.attr</param>
<param key="data">lu.communes</param>
</processor>
```

To prepare visualizations to be used in ArcGIS Scene or Map components, feature classes must be converted into feature layers. A feature layer is a relational, georeferenced model containing rules on how this data is to be visualized. In this case a point feature class is demonstrated having its result shown as extruded symbols on the scene it will be used in.

```
<!-- Communes -->
<!-- Land Use -->
<!-- LUC Crops -->
<processor
class="de.isdss.ext.arcgis.processors.ArcGISGet3DUnique
PointLayerProcessor">
<param key="featureclass">fc.com.point</param>
<param key="symbol.angle">0</param>
<param key="symbol.width">3.0</param>
<param key="symbol.depth">1.0</param>
<param key="symbol.size">1</param>
<param key="symbol.offset.x">-3750.0</param>
<param key="symbol.offset.y">-3750.0</param>
<param key="symbol.offset.z">0</param>
<param key="symbol.shape">character</param>
<param key="symbol.font">
ESRI Enviro Hazard Analysis
</param>
<param key="symbol.index">92</param>
<param key="color.red">255</param>
<param key="color.green">255</param>
<param key="color.blue">0</param>
<param key="scale.min">2000000</param>
<param key="scale.max">500000</param>
<param key="name">layer.com.crops</param>
<param key="by.field">NAME</param>
<param key="with.field">crops</param>
<param key="apply.on">depth</param>
<param key="value.scale">20.0</param>
</processor>
```

The so prepared set of feature layers can be incorporated into the *Multi Document Desktop* component. Since all data in memory is relational, several different perspectives can be created by projecting the loaded feature layers based on sets of fields. Each of those projections can be presented in a different way. In this case a 3-D Scene as well as a pie chart diagram is rendered using the same type of data source.

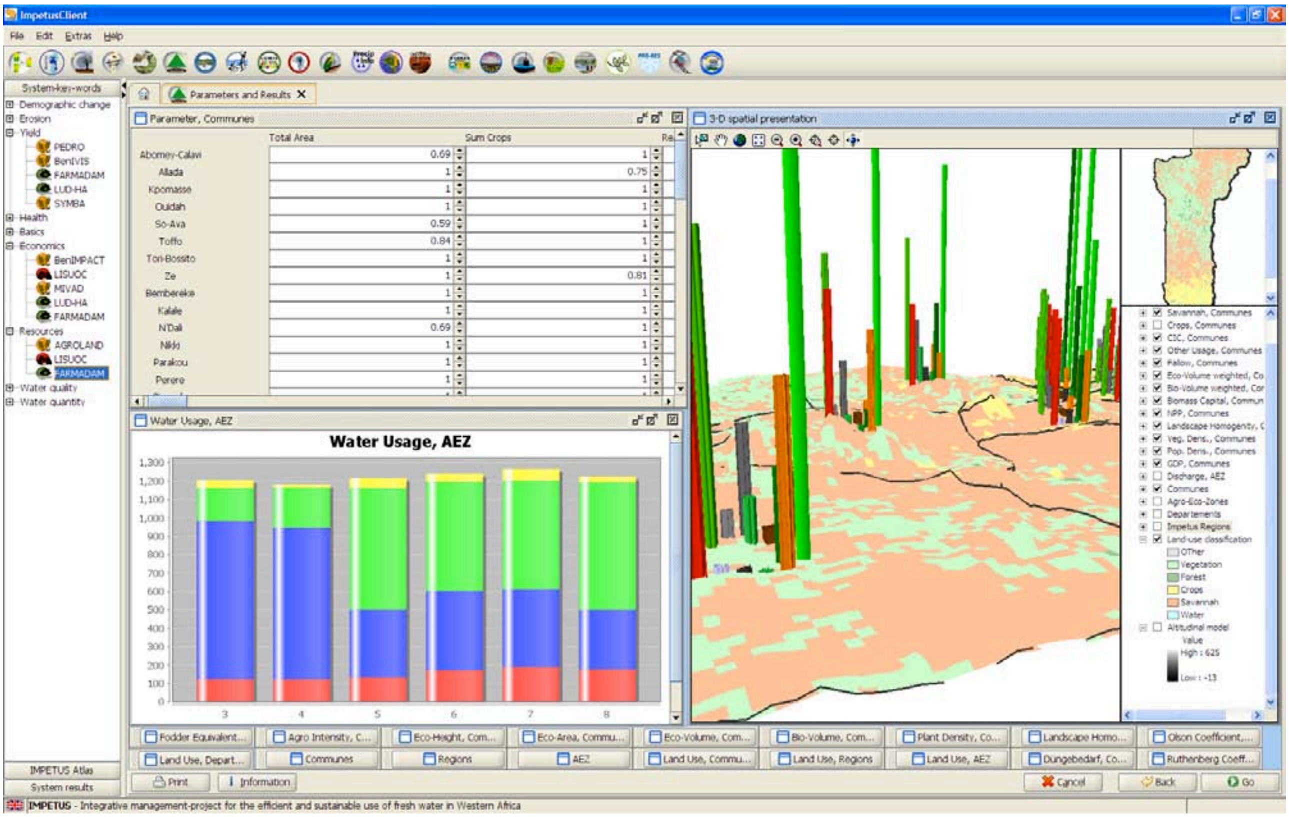
```
<function id="multipanel"
class="de.isdss.ext.arcgis.functionalities.ArcGISMultiP
anelFunctionality">

<param key="panels">5</param>

<param key="panel.1.title">scene.title</param>
<param key="panel.1.type">scene</param>
<param key="panel.1.zfactor">25.0</param>
<param key="panel.1.baseheight">layer.dgm</param>
<param key="panel.1.layers">5</param>
<param key="panel.1.layer.1">layer.cities</param>
<param key="panel.1.layer.2">layer.communes</param>
<param key="panel.1.layer.3">layer.regions</param>
<param key="panel.1.layer.4">layer.streets</param>
<param key="panel.1.layer.5">layer.luc</param>
<!-- ... -->
<param key="panel.4.title">com.luc</param>
<param key="panel.4.type">piechart</param>
<param key="panel.4.featureclass">com.luc</param>
<param key="panel.4.title">label.luc</param>
<param key="panel.4.fields">luc.fields</param>
<param key="panel.4.legendfield">NAME</param>
<param key="panel.4.order">columns</param>
<!-- ... -->
</function>
```

3-D DATA VISUALISATION WITHIN THE MULTI DOCUMENT DESKTOP ENVIRONMENT

The *Scientific Model Integration pipeLine Engine* (Enders and Diekkrüger, 2009) was used as the software framework for IMPETUS. *SMILE* is a pipeline based framework architecture providing processor interfaces written in Java and configured using a XML client configuration. Hence, complex applications could be formed by interconnecting processors in a linear way that were configured using processor specific parameters within the XML client configuration. To communicate between each other, processors share one common parameter map in memory which is organized as a hash, associating parameter names to values. As the pipeline is processed, values are stored and read from the parameter map. The parameter map is a collection of any Java based object.

Since numerous field research results and simulation models of IMPETUS were provided as spreadsheet files, the processors for reading, writing and calculating values had to be implemented as *SMILE* processors. Jakarta's POI-HSSF API was used to meet the requirements of accessing MS Excel spreadsheets. In addition to that, other processors had to be created to visualize data on screen (e.g. JFreeChart API, ESRI ArcGIS Engine components like the SceneBean).

To provide a wide variety of data representation and to show input and output at the same time, a *Multi Document Desktop* was implemented, giving the user the opportunity to choose between several visuals giving several perspectives to selected and projected data simultaneously (Fig. 1).



Parameter, Communes

	Total Area	Sum Crops
Abomey-Calavi	0.69	1
Allada	1	0.75
Kpomasse	1	1
Ouidah	1	1
So-Ava	0.59	1
Toffo	0.84	1
Tori-Bossito	1	1
Ze	1	0.81
Bembereke	1	1
Kalale	1	1
N'Dali	0.69	1
Nikki	1	1
Parakou	1	1
Perere	1	1

- ☑ Savannah, Communes
- ☐ Crops, Communes
- ☑ CIC, Communes
- ☑ Other Usage, Communes
- ☑ Fallow, Communes
- ☑ Eco-Volume weighted, Co
- ☑ Bio-Volume weighted, Cor
- ☑ Biomass Capital, Commun
- ☑ NPP, Communes
- ☑ Landscape Homogenity, C
- ☑ Veg. Dens., Communes
- ☑ Pop. Dens., Communes
- ☑ GDP, Communes
- ☐ Discharge, AEZ
- ☑ Communes
- ☐ Agro-Eco-Zones
- ☐ Departements
- ☐ Impetus Regions
- ☑ Land-use classification
  - OTher
  - Vegetation
  - Forest
  - Crops
  - Savannah
  - Water
- ☐ Altitudinal model
  - Value
  - High : 625
  - Low : -13

Figure 1. Screenshot of the 3-D Multi Document Desktop.

By using the *Multi Document Desktop* within the Client each component is rendered as a single window. Windows can be opened or closed, changed in size and moved freely on the screen. All windows that are currently closed are shown as buttons on the bottom of the *Multi Document Desktop*. Depending on the configuration each window renders a visualization of specific data (simulation results, research data) as either tables or charts. Additionally, a simulation parameter input panel (Fig. 1 upper left) allows the user to change simulation parameters at runtime as well as the visualisation output in 2-D or 3-D. Within maps and scenes rectified simulation results and parameters can be visualized.

Within the 3-D spatial presentation the user is able to navigate freely through space and select specific features. The land use classification, provided as raster data is projected into the third dimension by using a DEM. The resulting TIN could be overlaid with other vector data from the geo-database (e. g. major settlements, or borders of administrative zones).

Each zone of interest contains several sets of rendered bars, which are visualizing the simulation results. This way the user can compare different areas in regard to specific simulation results. Each value is visualized using an extruded coloured symbol.

On the right side of the 3-D spatial presentation, a configuration panel is placed in which each layer can be turned on and off.

By clicking the Go button on the bottom right of the application, the currently set simulation parameters are used to recalculate new simulation results based on the current state of the simulation and the components are refreshed. To realize these kinds of iterative simulations that use accumulating decision effects for different simulation cycles, the last processor in the 3.D visualisation processing pipeline points back to the first one forming a simulation loop using the pipeline based framework (Fig. 2).

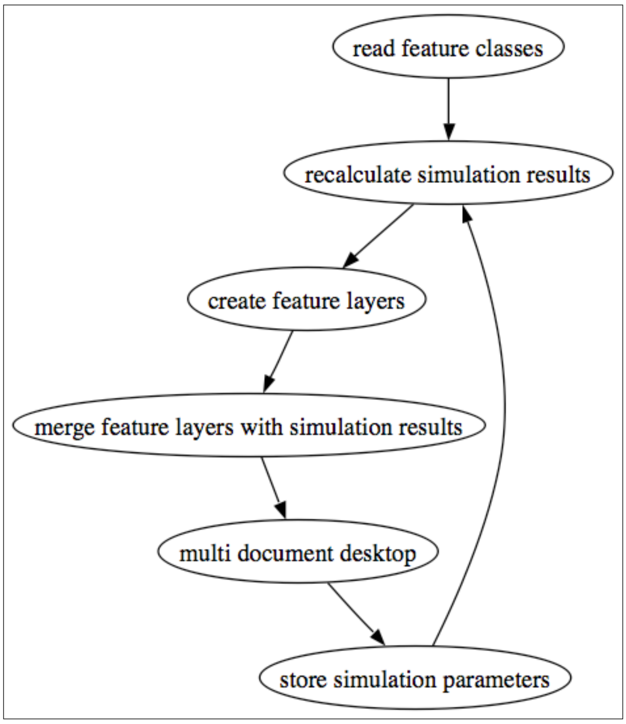

read feature classes
recalculate simulation results
create feature layers
merge feature layers with simulation results
multi document desktop
store simulation parameters

Figure 2. Farmadam pipeline structure.

Each visualized location is composed of a set of extruded symbols visualizing the simulation results and used parameters. For each commune a different set of symbols is used to allow visualization of decision effects in a three dimensional scene. Figure 3 illustrates effects on eco and bio volume of crops, forest, grassland, savannah, available water resources, farming, fertilizer requirements, urbanization, and population density as well as economic and additional parameters as an example. By using these sets of 3D symbols, different areas of interest become comparable in regard to each other showing effects of decisions. Continued execution of the simulation model visualizes long time decision effects, allowing the user to identify decisions producing unwanted results as well as the possibility to test possible countermeasures.

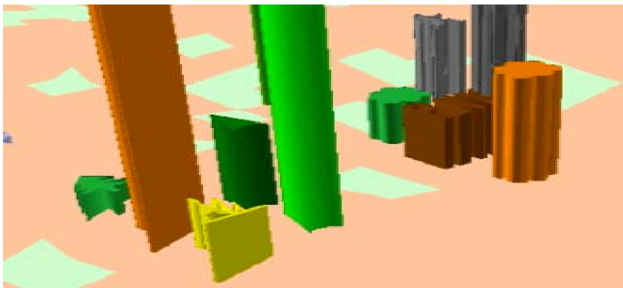


Figure 3. Extruded symbols.

Other components are used to present decision effects and parameters in more detail like different types of charts and diagrams as well as tables, allowing import and export of all relevant parameters and results. This way the current state of the simulation can be stored and imported as initial simulation state for subsequent alternative simulations.

CONCLUSION

The usage of spreadsheet files to describe simulation models and field research data reduces the necessary programming effort, especially with regard to the problem specific point of view. However, this comes at the cost of performance, since loading, recalculating and writing spreadsheet files is a time consuming process. Another drawback are the limitations of spreadsheet files, for example limiting the number of records depending on the used file format as well as the capabilities of the HSSF spreadsheet API for Java, which is used in this system. On the other hand, the usage of spreadsheet files reduced the required development time since model changes and data incorporation can be done by researchers directly without requiring special programming knowledge. By developing single generic processors with a high variability by using parameter based configuration within the XML-client configuration, development was highly flexible and changing requirements during the phase of development (often a time consuming task) could be matched very quickly. Three dimensional representations of spatial topology in conjunction with spatially represented data provide a user friendly, easy to navigate interface having an improved visibility compared to common approaches using two dimensional representations. Additionally, three dimensional visualization techniques simplify understanding of complex situations from the users point of view, especially when working with several different types of data sets in conjunction to examined locations encompassing multiple fields of interest simultaneously. Since 3D capabilities in GIS systems are increasing recently, 3D SDSS applications will become more common in the near future. Currently 3D capabilities of GIS and therefore SDSS applications are not yet fully supported by available frameworks. Hence, more research is required to match upcoming requirements to fully use the potential of modern information technology.

ACKNOWLEDGEMENTS

This study was part of the interdisciplinary scientific project IMPETUS, which was supported by the Federal German Ministry of Education and Research (BMBF) under grant No. 01 LW 06001A and 01 LW 06001B as well as by the Ministry of Innovation, Science, Research and Technology of the federal state of Northrhine-Westfalia under grant No. 313-21200200.

REFERENCES

Armstrong, A. P. and Densham, P. J., 1990. *Database organization strategies for spatial decision support systems*. International Journal of Geographical Information Systems, 4 (1), pp. 3-20.

Armstrong, M.P., Densham, P.J. and Rushton, G., 1986. Architecture for a microcomputer Based Spatial Decision Support System. In: *Proceedings of the 2nd International Symposium of Spatial Data Handling*. Williamsville: International Geographical Union), pp. 120-131.

Bareth, G. and Yu, Z., 2007. Interfacing GIS with a process based agro-ecosystem model - case study North China Plain. In: X. Tang, Y.Liu, J Zhang, and W. Kainz, eds. *Advances in spatio-temporal analysis*. London: Taylor & Francis.

Bareth, G., 2009 (in print). GIS- and RS-based Spatial Decision Support: Structure of a Spatial Environmental Information System (SEIS), *International Journal of Digital Earth*.

Breunig, M. 2001: *On the Way to Component-Based 3D/4D Geoinformation Systems*. Springer, Berlin ISBN 978-3540678069

Crossland, M.D., Wynne, B.E. and Perkins, W.C., 1995. Spatial Decision Support Systems: An Overview of technology and a test of efficacy. *Decision Support Systems*, 14 (3), pp. 219-235.

Densham, P.J., 1991. Spatial Decision Support Systems. *In* D.J. Maguire, M.F. Goodchild and D.W. Rhind, eds. *Geographical Information Systems.- Vol. 1: Principles.* Harlow: Longman Scientific & Technical, pp. 403-412.

Enders, A. and Diekkrüger, B. (2009): Development of a Spatial Decision Support Framework for IMPETUS project in West Africa.- In: Ioannis N. Athanasiadis . Pericles A. Mitkas , Andrea E. Rizzoli . Jorge Marx Gómez (Eds.): Information Technologies in Environmental Engineering, Proceedings of the 4th International ICSC Symposium, Thessaloniki, Greece, May 28-29, 2009, pp. 132-148.

Enders, A., Laudien, R. and Hoffmann, R., 2007. Spatial Decision Support Systems. In Interaktives Management-Projekt für einen Effizienten und tragfähigen Umgang mit Süßwasser in Westafrika: Fallstudien für ausgewählte Flusseinzugsgebiete in unterschiedlichen Klimazonen, Siebter Zwischenbericht, pp. 7-21 [online], Available from: http://www.impetus.uni-koeln.de/content/download/ZB2006/IMPETUS_Zwischenbericht_2006.pdf, [accessed 25 February 2009

ESRI, 2008. *ArcGIS 9.2 Desktop Help, Geo-databases and ArcSDE - Types of geo-databases* [online]. Available online at: http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Types_of_geodatabases [accessed December 2009]

Goodchild, M.F. and Densham, P., 1993. *Initiative 6: Spatial decision support systems (1990-1992)* (Santa Barbara, CA: National Center for Geographic Information and Analysis).

Hartkamp A .D., White J. W. and Hoogenboom, G., 1999. Interfacing GIS with agronomic modeling: A Review. *Agronomy Journal*, 91, pp. 761-772.

Hartkamp A .D., White J. W. and Hoogenboom, G., 1999. Interfacing GIS with agronomic modeling: A Review. *Agronomy Journal*, 91, pp. 761-772.

Henning S.D. 2008: Prozessierung von Laserscanndaten zur Erstellung eines 3D-Stadtmodells: CampusGIS-3D pp. 1. Diplomarbeit an der Universität zu Köln, Geographie (GIS / RS).

Herter, M., Koos, B. 2006: *Java und GIS. Programmierung – Beispiele – Lösungen*, Wichmann, ISBN 978-3-87907-379-5, 318p.

Keenan, P., 1996. Using a GIS as a DSS Generator. In: J. Darzentas, J. S. Darzentas, and T. Spyrou, eds. *Perspectives on DSS*. University of the Aegean, Greece, pp. 33-40.

Keenan, P.B., 2006. Spatial Decision Support Systems: A coming of age. *Control and Cybernetics*, 35(1), pp. 9-27.

Krüger, G., 2006. *Handbuch der Java-Programmierung.- 4. Auflage*. Munich: Addison-Wesley.

Laudien, R. and Bareth, G., 2007. Entwicklung und Programmierung von räumlichen Entscheidungsunterstützungssystemen mit ArcGIS Engine und Java. In: *GIS-Geoinformationssysteme*, 4, pp. 16-21.

Leung, Y., 1997. *Intelligent Spatial Decision Support Systems*. Berlin: Springer.

Longley P.A, Goodchild D.J. Maguire and D.W. Rhind, 1999. Data quality – introduction. In: P.A. Longley M.F. Goodchild D.J. Maguire and D.W. Rhind, eds. *GIS, Vol. 1*. New York: Wiley, pp. 175-176.
Malczewski, J., 1999. *GIS and multicriteria decision analysis*. New York: Wiley.

Malczewski, J., 2006. GIS-based multicriteria decision analysis: a survey of literature. I*nternational Journal of Geographical Information Science*, 20 (7), pp. 703-726.

Manoli, E., et al., 2001. Water demand and supply analysis using a spatial decision support system. *Global NEST: The International Journal*, 3 (3), pp. 199-209.

Mennecke, B.E., 1997. Understanding the Role of Geographic Information Technologies in Business: Applications and Research Directions. *Journal of Geographic Information and Decision Analyis*, 1, pp. 44-68.

Muller, J.-C., 1993. Latest developments in GIS/LIS. *International Journal of Geographical Information Systems*, 7 (4), pp. 293-303.

Rizzoli, A.E. and Young, W.J., 1997. Delivering environmental decision support systems: software tools and techniques. *Environmental Modelling & Software*, 12 (2-3), pp. 237-249.

Schneider K., 2003. Assimilating remote sensing data into a land surface process model. International Journal of Remote Sensing, 24, pp. 2959-2980.

Seppelt K., 2003. *Computer-based environmental management. Weinheim*: Wiley-VCH.

Shaffer M.J., Ma L. and Hansen, S., 2001. Introduction to simulation of carbon and nitrogen dynamics in soils. In: M.J. Shaffer., L. Ma and S. Hansen M.J. Shaffer., L. Ma and S. Hansen, eds. *Modeling carbon and nitrogen dynamics for soil management*, Boca Raton: Lewis Publishers, pp. 1-10.

Sharma T., Carmichael J. and Klinkenberg, B., 2006. Integrated modeling for exploring sustainable agriculture futures. *Futures*, 38 (1), pp. 93-113.

Speth, P., et al., 2005. IMPETUS-West Africa- An integrated approach to the efficient management of scarce water resources in West Africa – Case studies for selected river catchments in different climate zones. In: *DLR– Projektträger im DLR*, eds. *GLOWA – German Programme on Global Change in the Hydrological Cycle, Status Report 2005*, pp. 86-94.

Taylor, K., Walker, G. and Abel, D., 1999. A framework for model integration in spatial decision support systems. *International Journal of Geographical Information Science,* 13 (6), pp. 533-555.

Wilson, R.D., 1994. GIS & Decision Support Systems. *Journal of Systems Management*, 45 (11), pp. 36-40.

Wright J. R. and Buehler K. A., 1993. Probabilistic Inferencing and Spatial Decision Support Systems. In: J. R. Wright, L. L. Wiggins, R. K. Jain and T. J. Kim, eds. *Expert Systems in Environmental Planning*. Berlin: Springer, pp. 119-144.

Yeh, A., 1999. Urban planning and GIS. *In:* P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, eds. *Geographic information system.- Vol.2*. New York: John Willey and Sons. August 2001

SELECTING OFFSHORE RENEWABLE ENERGY FUTURES FOR VICTORIA

M. A. Boelen [a,], I. Bishop [b], C. Pettit [c]

[a] Dept. of Geomatics, University of Melbourne, Parkville, marika-boelen@hotmail.com
[b] Dept. of Geomatics, University of Melbourne, Parkville, i.bishop@unimelb.edu.au
[c] Spatial Information Sciences, Dept. of Primary Industries, Lincoln Square North, Carlton,
Christopher.Pettit@dpi.vic.gov.au

Commission II / II/6

KEY WORDS: Renewable Energy, Offshore, Visualisation

ABSTRACT:

Australia's population is continually growing, making land more valuable and adding to energy demand. As the coast of Victoria, Australia has regular high winds, the development of offshore renewable energy is an excellent alternative to conventional energy sources. This provides an opportunity to meet growing energy needs while caring for the environment; and supporting regional communities. There are currently no offshore energy projects in Victoria. This paper investigates demand, supply, feasibility and planning of the wind and wave power options. Analytical (GIS) and visual aids (*Google Earth*) are used to illustrate these options and so to assist the community in making an informed decision for the renewable energy approaches suitable in Victoria.

1. INTRODUCTION

1.1 The Off-Shore Option

Australia's energy demand increased by 2.1% each year – and Victoria's by 1.6% each year from 1960 to 2007 (Sustainability Victoria, 2008). The Australian Government is

> "… committed to ensuring 20 per cent of Australia's electricity supply comes from renewable energy by 2020." (Department of Climate Change, 2008)

Options for renewable energy development in Australia are predominantly wind, solar and geothermal. Discussion has previously focussed almost entirely on on-shore development potentials. However, on-shore wind farms can be controversial and as land becomes more valuable the advantages of off-shore development become more apparent.

Figure 1 shows that Victoria has an extensive coastline that may provide considerable opportunities for both wind and wave generated energy. Harries et al (2006) say the potential of offshore renewable energy resources (RER) development "…is related to the distribution of the winds, and the strongest occur between latitudes 40° and 60°…". The Victorian coast is predominantly between 38° and 39° south and so is well situated to maximise the use of wind related RER.

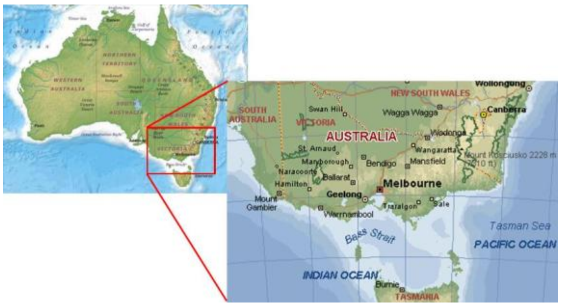


Figure 1. Location of this research (Source: World of Maps, 2009)

This paper reviews the offshore RER potential in Victoria including demand, supply, feasibility and planning.

1.2 Global Offshore Energy

1.2.1 Wave Power: Wave power is a much more recent power generation technology than wind power technology, with energy captured by turbines that are either fixed to shore, fixed to the sea floor or float on the water's surface (Sustainability Victoria, 2009; Harries, 2006). They are approximately three times more efficient than coal power stations, and have minimal visual and noise impacts. The first commercial wave farm was installed in Portugal in early 2008 and was closely followed by similar projects in Spain, the U.S.A. and U.K.

According to research analyst Gouri Nambudripad (Cleantech Group, 2008) with an investment of 500 billion British pounds, 2,000 terra Watt hours (tWh) of electricity could be produced each year from wave power.

However, the technology is still developing and is largely experimental. Developed countries have the capacity to trial wave power installations and may lead the way for longer-term adoption in the developing world.

1.2.2 Wind Power: The first modern commercial wind farm was installed in Denmark in 1991. Since then, wind turbines have become more powerful and economical, with offshore installation becoming increasingly popular.

The United Kingdom (with 590 mega watts (MW)) is the world leader in terms of installed offshore wind power; closely followed by Denmark (409MW) and the Netherlands (246MW).

Countries like China and India have also turned to offshore wind power due to their "large coastlines and vast oceanic areas, which provide excellent conditions for offshore wind power development" (Yu'an, 2009). For such developing countries, offshore power is an excellent solution as no land is required for power generating facilities; and can instead be used for housing and public services.

1.3 Economic Factors

Offshore winds are more uniform in strength and consistency compared with the onshore environment, meaning that more electricity is generated and there is less wear on electricity generating components through varying turbine speeds (Musial and Butterfield, 2004).

The costs associated with the installation and operation of offshore renewable energy can be seen in *Figure 2.* The graphs suggest that wave power is more economical than coal; while offshore wind power is cheaper than coal to install but more expensive to operate.

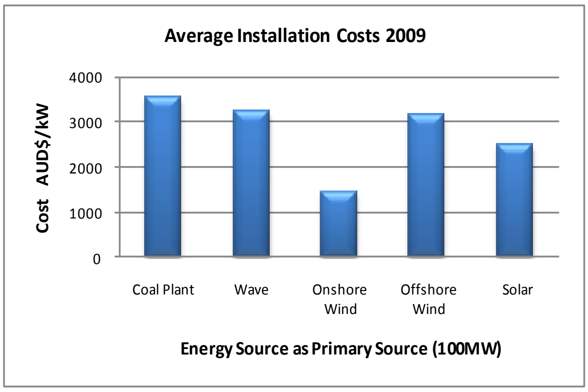


Figure 2a. Average installation costs of different energy sources in Australian Dollars (Source: Synder and Kaiser, 2009; Vining and Muetze, 2007)

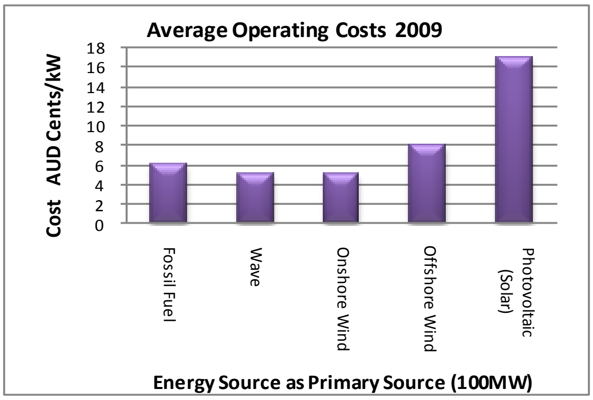


Figure 2b. Average operating costs of different energy sources in Australian cents (Source: Synder and Kaiser, 2009; Vining and Muetze, 2007)

However, there are scaling issues which have to date militated against significant investment in wave power. Large scale development is currently not feasible due to the power generating capacity of wave buoys – the largest being 150 kilo watts (kW).

The main advantage of offshore wind turbines is that visual and noise impacts are minimised. Wind turbines can be more powerful, while land can remain available for housing or other needs. Noise reduction technologies also don’t have to be used in the offshore environment, thus reducing the cost of individual turbines. As the depth of water increases, so does the cost of the structures. However, as land based wind technologies became more widely accepted, production and installation costs reduced significantly (Musial and Butterfield, 2004). So as costs reduce deeper water offshore wind technologies should become increasingly viable.

2. METHOD

2.1 Energy Demand and Generation

In the 2008/2009 year, Victorian’s consumed nearly 63 million mega watt hours (MWh) of electricity (ABARE, 2009) meaning over 7 giga watts (GW) of energy had to be generated. In 2019/2020 – the year in which the Australian Government wishes to have 20 per cent of Australia’s electricity supply to come from RER – Victoria is predicted to consume 78 million MWh of electricity (ABARE, 2009). So to meet Australia’s renewable energy target, approximately 1.8 GW of renewable energy would need to be generated. This could be achieved through:

- Approximately 720 5MW wind turbines (running at 50 per cent of their maximum capacity all year round) OR
- Approximately 14,900 0.15MW wave buoys (running at 80 per cent of their maximum capacity all year round)

Seasonal wind speeds don’t vary much around the study area. For example in the 90 mile beach area, the weakest wind speeds occur in July (13km/h) and the fastest in November (18.5km/h); with an annual average wind speed of 15.6 km/h (BoM, 2010). But as energy demand varies greatly throughout the day, more power generating devices may be needed than the amount suggested here. This energy generation scenario for 2020 is planned and visualised in this paper to enable an informed debate on the renewable energy approaches suitable for Victoria.

2.2 Spatial Analysis

Key considerations in planning for new energy infrastructure include:

- Economic Issues – cost and efficiency in relation to; water depth, location of existing infrastructure, wind speed and wave power (Harries et al, 2006; Musial and Butterfield, 2004; Sustainability Victoria, 2009)
- Environmental Issues – positioning restrictions due to the location of endangered animals and marine national parks (ABCSE, 2004; Thorpe, 1999)
- Social Issues – concerns in regards to aesthetics, noise and loss of recreation areas (Sustainability Victoria, 2009; Thorpe, 1999)

These factors are all weighted equally, in this initial assessment, and can be mapped and combined using a geographic information system (GIS) in order to determine an optimum solution for offshore RER. The maps created can be used to compare suitable locations and the impacts of the different factors. An interactive decision making system (such as a web-mapping tool like Geoscience Australia’s MapConnect; www.ga.gov.au/mapconnect) would also improve the decision making process, allowing layers to be turned on and off for comparison, although this is beyond the scope of this study.

ArcGIS (ESRI, 2009) was used for the analysis but some data the data acquired was of less than ideal resolution. For example, the bathymetry data acquired had a 250 metre grid spacing, which may have smoothed some ocean features making them unidentifiable.

2.2.1 Economic Issues: To incorporate the economic issues into the planning process bathymetry, shipping lane, petroleum platform, wind speed and wave power data was collected.

The bathymetry data determines the feasibility of offshore RER and water depths were zoned as follows using Musial and Butterfield's (2004) research:

- Most Suitable (0-30 metres deep)
- Possibly Suitable (30-50 metres deep)
- Future Suitability (50-200 metres deep)
- Not Suitable (greater than 200 metres deep)

Based on Jeng (2007) a 1 km exclusion buffer was created around major shipping lanes. Petroleum platforms can provide the infrastructure needed to transport the power created from wind and wave farms to the shore (Jeng, 2007). Many of these are due for decommissioning in Bass Strait, and a 2 km inclusion buffer was created around these as possibly suitable (considering also the dependence on corresponding depth information).

For wind turbines to be viable wind speeds must be over 5 m/s at 80 metres height and wave power required a level of sea wave energy over 20 KW/m^2. From the available wind and wave power maps (DEWHA, 2007) all the areas in this study were well above the minimum requirements and so were not included in the *ArcMap* layers.

2.2.2 Environmental Issues: The environmental factors stated above were incorporated into the planning process by collecting marine national park and endangered animal location information. These areas were given a 1 km buffer and labelled as "Not Suitable", as wind and wave farms cannot be placed in areas of environmental importance (ABCSE, 2004).

2.2.3 Social Issues: The main social issues involved with renewable energy implementation as identified by Thorpe (1999) are aesthetics, noise and loss of recreation areas.

Denmark requires that large-scale wind farms be at least 8 km from shore (Ladenburg and Dubgaard, 2007). This ensures that the visual impacts of the turbines are minimised. However, this requirement may not be practical in other countries due to the underwater topography and the current technology of offshore wind structures.

Nevertheless, an 8 km exclusion buffer was placed along Victoria's coast. The aesthetics of wave power buoys would not be an issue as they sit only 30 metres above the water level – not visible from 8km.

2.3 Visualisation

2.3.1 Building the Wind Farm Model: The wind farm model was constructed from simple shapes using *Google Sketchup*. A single wind turbine was downloaded from the *3D Google Sketchup Warehouse*. Several wind turbines were created by copy and pasting the original turbine. To work out the placement of 100 wind turbines, a 10x10 cell grid, with 600 metre spacing's was used. Each wind turbine was placed at the intersection of the grid lines (*Figure 3*), then the grid lines were deleted to prepare the model for insertion into *Google Earth*.

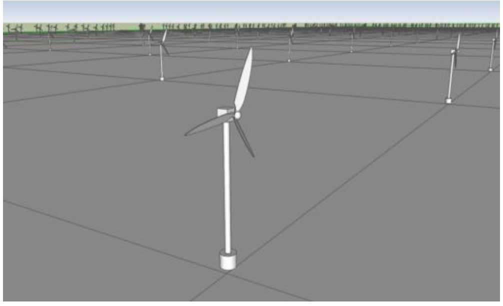


Figure 3. Place the wind turbines at the intersection of each grid line

2.3.2 Building the Wave Farm Model: The wave buoys in the wave farm model were constructed from simple shapes in *Google Sketchup* using a dimensioned buoy found at OPT (2009). The same method used above was used for positioning the wave power buoys.

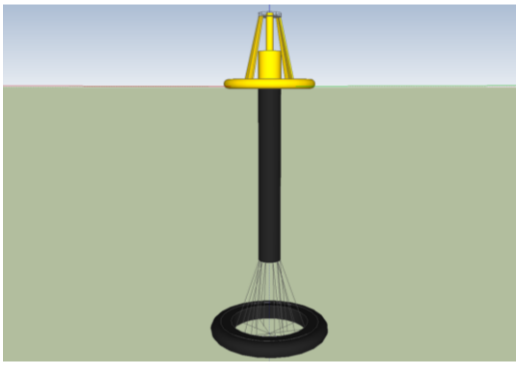


Figure 4. Wave power buoy made in *Google Sketchup*

2.3.3 Importing into Google Earth: Once the wind turbine and wave buoy models were constructed, they were positioned in *Google Earth,* based on the suitable areas defined using the GIS analysis, to create representations of the new seascapes.

The first step to create the wind and wave farm models in *Google Earth* was to import an image of the map created in the planning process by *"add → image overlay"*.

Importing the models into *Google Earth* involved adding the model as a DAE file. This was done by saving the *Google Sketchup* model as a DAE file, then opening it in *Google Earth* using *"add → model"*.

Finally, a cargo ship approximately 180 metres long 30 metres wide and 18 metres high was placed in the visualised environment to give the viewer a sense of scale (*Figure 5*).

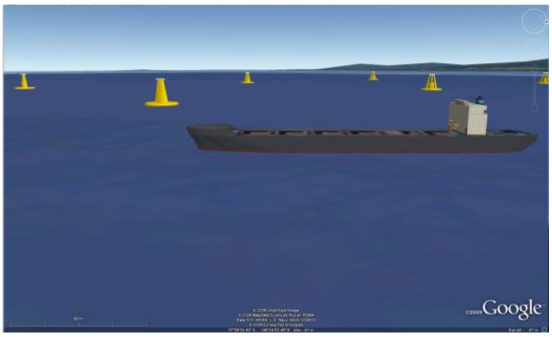


Figure 5a. 180x30x18 metre cargo ship – shows scale of wave buoy

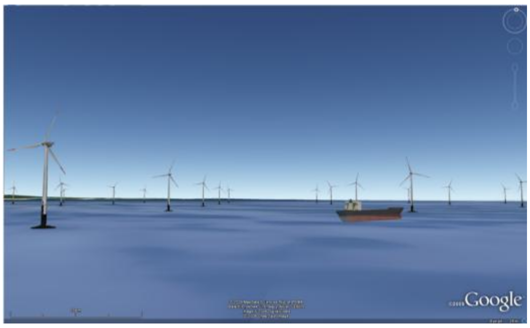


Figure 5b. 180x30x18 metre cargo ship – shows scale of wind turbines

3. RESULTS AND DISCUSSION

3.1 Spatial Analysis

As a result of the application of the data and procedures detailed in Section 2.2, *Figure 6* was produced. It is evident that the 90 Mile Beach area (a coastal region on Victoria's eastern coast) is the most suitable location, with shallow water (water less that 30 metres deep) and petroleum platforms in the area. Further offshore, the water around King Island and Flinders Island is also shown as suitable. However, the threatened fauna and marine national park data obtained does not cover these areas. This would reduce their suitability.

Therefore the area in eastern Victoria along the 90 mile beach is the focus for the visualisation component of this study.

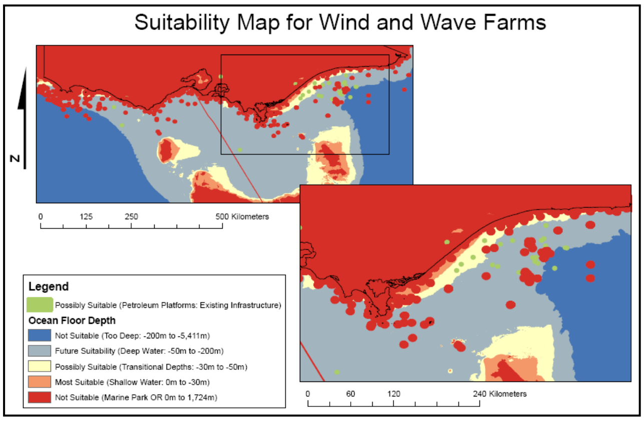

Suitability Map for Wind and Wave Farms
N
0
125
250
500 Kilometers
Legend
Possibly Suitable (Petroleum Platforms: Existing Infrastructure)
Ocean Floor Depth
Not Suitable (Too Deep: -200m to -5,411m)
Future Suitability (Deep Water: -50m to -200m)
Possibly Suitable (Transitional Depths: -30m to -50m)
Most Suitable (Shallow Water: 0m to -30m)
Not Suitable (Marine Park OR 0m to 1,724m)
0
60
120
240 Kilometers

Figure 6. Overall suitability map for the location of wind and wave farms off Victoria's coastline with 90 Mile Beach zoom

The main focus of this study was to illustrate how the transition to RER would affect the Victoria's visual environment with an overview visualisation. Thus the spatial analysis in this study was simplified. A more complete analysis would include a view-shed analysis and would use weighted factor combinations.

3.2 Visualisation

3.2.1 Different Sized Farms: *Google Earth* was the visualisation tool used in this project, primarily due to its familiarity and connectivity. The models created were easily positioned in *Google Earth* and once complete, were effective as a basic visualisation tool. However realism was difficult to achieve as there is a need for elevation and distance to see the extent of the RER impacts. Other software packages could be used to also achieve more realistic ground level visualisation.

Farms of 100, 200 and 500 wind turbines (at 100 metres tall from the ocean surface to the hub) and 100 wave buoys (at 30 metres tall from the ocean to the highest point) were visualised from the beach at approximately ground level, with an example shown in *Figure 7*.

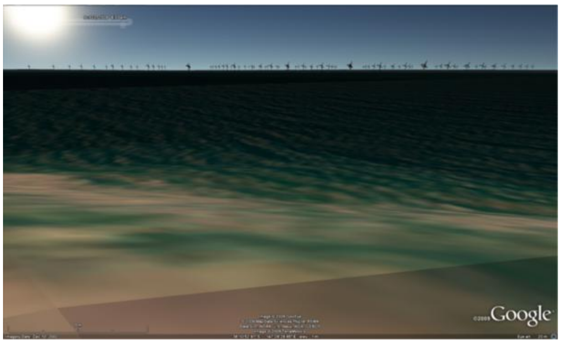


Figure 7. 100 wind turbines off the 90 mile beach coast, viewed at 8km from the beach at approximately ground level at sunrise

It is apparent from this visualisation that the individual wind turbines are very small and would have very limited individual impact on the aesthetics or recreational values of the coast. However, when 100 are seen together, particularly in conditions of high visual contrast, the effect is quite noticeable. To what extent this might have significant impact on scenic values is not known in the Australian context. European research has shown that as the number of turbines in a wind farm increases, so does the visual impact (Ladenburg and Dubgaard, 2007). However, further evaluation is necessary, especially with reference to context and comparison with the impact of alternative energy options. At 8 km noise will not be a factor.

As the wave buoys only sit 30 metres above the water, and are no closer than 4 kilometres from shore, they do not create any significant visual impacts. At this distance from shore, noise will also not be a factor and recreational activities will not be impacted.

Although wave power is commercially viable, it is not yet suitable for such large-scale projects. With current technology there would not be sufficient space to accommodate the number of wave buoys needed. For every wind turbine installed, 20 wave buoys would be needed to get the equivalent amount of power. However, smaller projects using wave energy to power remote coastal communities appear to be clearly viable.

3.2.2 Wind Power Contribution to 20% RER by 2020:

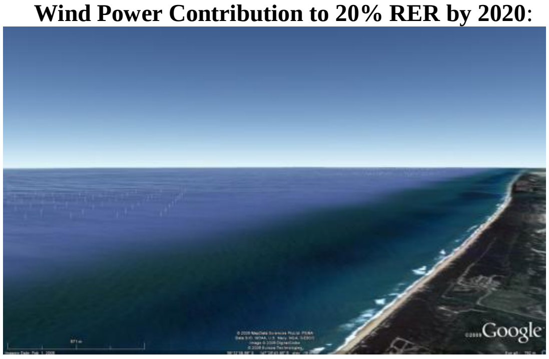


Figure 8. 20% RER by 2020 – 720 wind turbines in 3 wind farms

Figure 8 shows three wind farms – with 100, 280 and 400 wind turbines – spread across the extent of the 90 mile beach area. From this viewpoint and at this resolution the turbines are barely noticeable. However this is not a representative view point and conclusions about impact cannot be drawn. What is clear is that there will be few points along this coast which are not within view of a large number of turbines. Coastal activities – swimming, surfing, fishing etc – are not likely to be affected, but the public response to the intrusion (even in the knowledge that this could power over 2 million homes) is problematic.

3.2.3 Local Impacts of Offshore RER: Each of the wind and wave farms were positioned in the areas available derived from the spatial analysis. The most suitable area was confined to the 90 mile beach region. To help answer the question posed above about the impact of such pervasive infrastructure, further visualisations were developed from the beach; 200 metres inland from the beach (the approximate location of the 90 mile beach ocean road) and from urban centres such as Lakes Entrance; Golden Beach; Paradise Beach; and Woodside Beach.

The wind turbines were most noticeable from the beach. As the wind farms are over 8 kilometres from shore, their visibility varies depending on the weather conditions. Bishop and Miller, (2007) found significant differences in impact levels according to haze levels and their effect on the contrast between the turbines and their background. As technologies advance, the turbines could be placed in deeper water further offshore and have increasingly less visual impacts. Even with the cost reducing, putting turbines further from shore remains more expensive. The point at which society would find a balance between these costs and the amenity benefit is unknown. More detailed visualisation and survey research would be needed to address these questions.

A limited evaluation was undertaken in this research. Looking at the visualisations from the 90 mile beach ocean road, there were limited views of the wind farms. The road winds along the coast, going both behind and in front of the sand dunes. There is a lot of trees and scrub on the sand dunes, blocking the view of the ocean the majority of the time.

The visualisations and *Google Street View* also confirmed there was limited visibility of the wind farms from urban centres. *Google Street View* was used as a preliminary ground-truthing tool that confirmed the view of the wind turbines would predominantly be blocked by man-made structures, topography or vegetation (*Figure 9*). This being said, *Google Street View* is a static medium and therefore if this project were to go ahead, thorough ground-truthing would have to be performed by site visits.

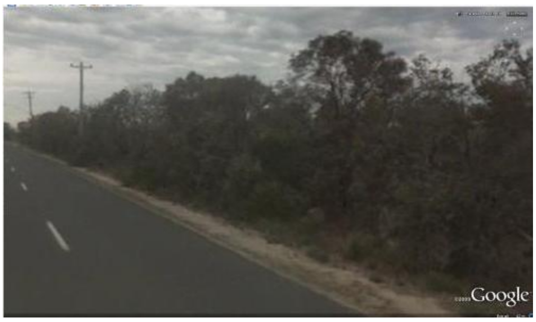


Figure 9. A typical view of the beach from the 90 mile beach ocean road (Source: *Google Street View*)

3.2.4 *Google Earth* as a Visualisation Tool: As an evaluation tool for the general public, *Google Earth* and *Google Sketchup* are effective, inter-operable and accessible programs.

When looked at from the beach, the wind turbines and water look quite realistic since there is nothing in the scene to take away from the overall effect: such as topography, trees, houses and animals. Various lighting conditions were also looked at in *Google Earth* and *Google Sketchup* to enhance the visualisation.

These effects act to enhance the overall visualisation, but in order to evoke a more valid emotional response from those evaluating the visualisations, specialised modelling programs with more advanced rendering capabilities are necessary. *Google Earth* provides a capacity for movement of the camera but not movement of the turbine blades. This movement can also have a significant effect on people's affective responses (Bishop and Miller, 2007). Game engines are another software option for providing interactive options in conjunction with dynamic elements in the landscape. Whatever product is being used the ideal is for a user to navigate through the virtual environment at their own pace and leisure.

There were also issues will the relative size of modelled objects in *Google Earth*. Although the objects were of the correct scale, they appeared to be smaller when imported to *Google Earth* than in comparable simulations in the literature. This may have been due to the angular field of view. In *Google Earth* the field of view is 60° and cannot be adjusted; whereas to give perceptual sizes similar to the typical human eye, the field of view size must be around 45°. This difference acts to make the turbines look smaller than in reality. With no field-of-view adjustment available, the only way to correct for perceived size is to somewhat increase the scale of the modelled object before importation into *Google Earth*. However, this approach can alter the visibility relationship between, for example, turbines and sand dunes, trees or houses. Hence that too can be misleading. Getting scale and visibility both correct is very important to visualisation products for public consumption.

3.3 Project Extension

Further exploration into visual impacts should be done through a community evaluation phase. This could involve RER planning workshops inspecting and evaluating the on-site visualisations and commenting on:

- the amenity of the farms – to help planners understand the implications of renewable energy sites from the ground level
- the configurations of the wind farms – 100, 200 or 500 turbines in one farm
- different renewable energy scenarios – 20, 50 or 80 per cent of energy covered by RER by 2020

- the alternatives to renewable energy – visually, would they prefer a coal fired power station or a wind turbine
- the broader implications of more renewable energy.

4. CONCLUSION

Planning for RER can begin as a straightforward process. Selecting the correct layers and using a GIS can account for economic, environmental and social issues within one program. The offshore RER can then be visualised using a simple program like *Google Earth* to get initial insights into the visual effects of the development. More rigorous evaluation of public responses would require the use of software providing for greater user control over the visualisation and dynamic objects.

Although Victoria has a large per capita energy consumption, it was found that it was feasible to provide 20 per cent of total energy production with offshore renewable energy sources. Currently wind power has the capacity to supply a greater load than wave power, although wave power would be generated a higher proportion of the time. Public acceptance may be enhanced by initially supplying smaller communities with RER, then building a base to link to the State (and now national) electricity grid.

However, a large portion of Victoria's coastline would be required to develop such resources and a significant amount of capital investment would be needed for implementation. Existing energy companies still see potential to expand the on-shore renewable capacity and are therefore not currently making any active plans for more expensive offshore installation. Moving such infrastructure offshore would therefore require financial or legislative intervention. At this stage there does not seem to be a willingness to move in this direction. Two factors might change this: (i) improved wave power technology making it a viable large scale alternative, or (ii) rapidly increasing density of on-shore wind farms to the point at which there is public pressure for an off-shore energy mix.

The consistently strong winds over Victoria's oceans, international success, climate change pressure and technological advances seem to indicate that at some future date we will see offshore RER in Victoria.

5. REFERENCES

Australian Bureau of Agriculture and Resource Economics (ABARE), 2009. "Energy in Australia", http://www.abare.gov.au/publications_html/energy/energy_09/auEnergy09.pdf (accessed 4 May 2009)

Australian Business Council of Sustainable Energy (ABCSE), 2004. Dispelling the myths about wind. In: *BioGeneration Magazine*. pp. 8-9.

Bureau of Meteorology (BoM), 2010, "Weather Observations Victoria",http://www.bom.gov.au/climate/dwo/IDCJDW0300.shtml (accessed 17 Feb. 2010)

Bishop, I., and Miller, D., 2007. Visual assessment of off-shore wind turbines: The influence of distance, contrast, movement and social variables. *Renewable Energy*, 32, pp.814-831.

Cleantech Group, 2008. "Ocean Power Technologies deploys Spanish wave unit", http://cleantech.com/news/3558/ocean-power-technologies-deploys-spanish-tidal-unit (accessed 20 Oct. 2009)

Cleantech Group, 2008. "UK holds half of Europe's wave energy potential", http://cleantech.com/news/3879/uk-holds-half-europes-wave-energy-potential (accessed 19 Oct. 2009)

Department of Climate Change, 2008. "Australian Government's Renewable Energy Target", http://www.climatechange.gov.au/renewabletarget/index.html (accessed 1 May 2009)

Department of the Environment, Water, Heritage and the Arts (DEWHA), 2007, 2008. "Renewable Energy Atlas of Australia", http://www.environment.gov.au/renewable/atlas (accessed 2 Aug. 2009)

ESRI, 2009. "ESRI Products", www.esri.com (accessed 10 Oct 2009)

Harries, D., McHenry, M., Jennings, P., and Thomas, C., 2006. Hydro, tidal and wave energy in Australia. *International Journal of Environmental Studies*, 63(6), pp. 803-814.

Jeng, D., 2007. Potential of Offshore Wind Energy in Australia. In: Offshore Technology Conference. 2*007 Offshore Technology Conference*. Houston, Texas, U.S.A, 30 April-3 May. U.S.A.

Ladenburg, J., and Dubgaard, A., 2007. Willingness to pay for reduced visual disamenities from offshore wind farms in Denmark. *Energy Policy*, vol. 35, pp. 4059-4071.

Musial, W. and Butterfield, S., 2004. "Future for Offshore Wind Energy in the United States", http://www.osti.gov/bridge (accessed 1 May 2009)

Ocean Power Technologies (OPT), 2009. "Technology", http://www.oceanpowertechnologies.com/tech.htm (accessed 10 Aug 2009)

Sustainability Victoria, 2008 & 2009. "Renewable Energy Resources", http://www.sustainability.vic.gov.au/www/html/2109-renewable-energy-resources.asp (accessed 4 May 2009)

Synder, B., and Kaiser, M., 2009. Ecological and economic cost-benefit analysis of offshore wind energy. *Renewable Energy*, 34, pp. 1567-1578.

Thorpe, T., 1999. "A Brief Review of Wave Energy", https://staff.lauder.ac.uk/ICT/Library.nsf/0/5AC80A020BA2C07F80256D820039EEA6/$FILE/A+brief+review+of+wave+energy+A+report+produced+for+the+DTI.pdf (accessed 27 March 2009)

Vining, J., and Muetze, A., 2009. Economic factors and incentives for ocean wave energy conversion. *IEE Transactions on Industry Applications*, 45(2), pp. 547-554.

World of Maps, 2009. "online Maps of Australia", http://www.worldofmaps.net/oceania/australia_maps.htm (accessed 27 May 2009)

Yu'an, Z., 2009. "China commits to offshore renewable energy", http://www.chinadaily.com.cn/bizchina/2009-06/24/content_8316184.htm (accessed 19 Oct. 2009)

6. ACKNOWLEDGMENTS

Geoscience Australia and the Department of Sustainability and Environment, Victoria provided the data used in this project.

SPATIAL RELATIONS AND INFERENCES FOR CONTEXT AWARE VISUALIZATION

O. Akcay and O. Altan

ITU, Faculty of Civil Engineering, Department of Geomatic Engineering, 34469 Maslak Istanbul, Turkey - akcayoz@itu.edu.tr, oaltan@itu.edu.tr

Commission II, WG II/5

KEY WORDS: Ontology and Concepts in GIS, Location-Based Services, Web and Wireless Applications in GIS

ABSTRACT:

Data submission and refreshment, e.g. sensors, continuously feed a distributed information system. The system has to interpret incoming data in order to get valuable information which comes from different sources. Mobile users need more adaptive visualized spatial data according to location, device, personal profile, intention etc. In a semantic approach, an ontological model provides an intelligent system which is capable of extracting implicit information from explicit one. In this paper, some spatial concepts and properties have been defined for mobile devices and their users so as to produce context-aware visualization. The aim of these mobile contextual ontologies is to obtain a semantic model which inferences some parameters for an appropriate visualization.

1. INTRODUCTION

Advanced wireless technologies enable smart applications on Location Based Services (LBS). These smart applications are capable of doing some inferences in order to obtain a visualization which is relevant to the user situation. A Geographical Information System (GIS) continuously receives various kind of data and the system always updates itself. Especially in systems like LBS, there are a lot of instant changes that affect decision process. To evaluate effects of instant changes accurately, knowledge-base systems should be established. Establishing an appropriate knowledge-base system requires some steps: Determining context, ontological approach on the context and composing knowledge-base.

According to the context, a context aware system can be provided. A context aware system is a system uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task (Dey and Abowd, 2000). Ontological approach of the context should be implemented so as to provide a system which can be interpretable by computers. Ontological concepts and relations are coded with ontological languages. A consistent knowledge-base provides a system which is sensible and capable of showing reaction to the instant changes.

In this paper some spatial concepts and their relation have been investigated in order to obtain relevant visualisation on small mobile devices. Small mobile devices are not able to handling huge amount of data. A simple computing environment structure is necessary to process spatial data on tiny mobile devices. Next section presents some related works. Section 3 then explains some concept, their relations and their some simple inferences. Concluding remarks are discussed in last section.

2. RELATED WORK

Scientifically, LBS is not only a spatial topic. Computer Sciences also contribute to LBS, because they are pervasive computing environment. Pervasive computing environments gracefully integrated networked computing device - with people and their ambient environments. A room, for example, might be saturated with a lot of devices that provide information to people without needing their active attention (Zhu et al., 2005). Establishing a pervasive computing application has become one of the major tasks in computer sciences. In particular, pervasive computing demands applications that are capable of operating in highly dynamic environments and of placing fewer demands on user attention. In order to meet these requirements, pervasive computing applications will need to be sensitive to context.

Depending on the advancement of pervasive computing, geoinformation researches have tended to focus on context-aware and semantic modelling. Hong et al., 2005 proposed an adaptive location data management strategy in order to support adaptivity and scalability of the location based system using a variety of context which can be accessed in the ubiquitous computing environment. Hong et al., 2005, however, ignored the need for a semantic projection of the context model in order to obtain a ubiquitous computing system.

Kim et al., 2005 presented the architecture of tour information services based on semantic web technologies. The aim of the service is to provide the exact tour information and interoperability between the server systems. Nevertheless, the definition of the ontologies of the tour information explains only a small part of the location based services. It does not provide a whole semantic system design. The tour ontology should be integrated to user, device and spatial ontology and relations with each other. Another ontology-based approach is personalized situation-aware mobile service supply (Weissenberg et al., 2006). The research, however, does not include the visualisation styles of any spatial entity that obtained at the context-aware service.

Chen et al., 2004 designed a context-aware architecture so as to create intelligent spaces. They established a broker federation formed by multiple brokers. Gu et al., 2004 also presented a detailed context-aware architecture for smart home applications.

Christopoulou et al., 2004 defined an ontology-based context model. Lutz and Klien, 2006 explained an ontology-based Geographical Information retrieval contributes to solving existing problems of semantic heterogeneity and hides most of the complexity of the required procedure from the requester. In this paper, as a different approach, modeling spatial visualisation has been considered for any scale of a smart application of LBS.

Ontological knowledge-based systems are composed with ontology languages. Ontology Web Language (OWL) has been accepted as a standard language of ontology (McGuinness and van Harmelen 2004). The Semantic Web Rule Language (SWRL) is another specification to extract implicit information from explicit ones. SWRL concludes acquired knowledge with a rule based XML syntax language. Therefore it can be perceived a different kind of OWL-DL specification. In any case, SWRL is based on a combination of the OWL-DL and OLW-Lite sublanguages of OWL with the Unary/Binary Datalog RuleML sublanguages of the Rule Markup Language (RML) (Horrocks et al., 2004).

3. SPATIAL RELATIONS AND INFERENCES

Basic spatial objects are polygons, polylines and points. Ontological perception of spatial objects should base on these basic shapes. In two dimensional representation, buildings, regions, territories, areas can be shown with polygons. Roads, rivers, borders can be drawn with lines. Any other object can be represented with point such as a bus station, a person and a vehicle. Figure 1 depicts a simple map with basic spatial objects.

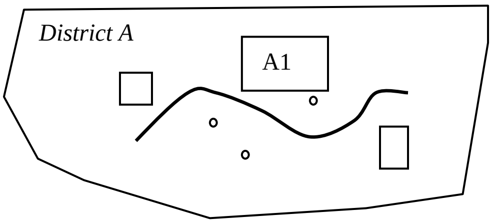


Figure 1. A basic map.

A1 is an instance of Building Concept. A1 building has to be in a district of the city. Building A1, therefore, is subsumed by a region. Let us assume that District A subsumes Building A1. A district is subsumed by a city. A city is subsumed by a country. A country is subsumed by a continent. A continent is subsumed by the world. Consequently everything which has dimensions is a part of the world. This ontological concept can be extended to three dimension.

LBS is a smart environment which can be established in different scales. It can be valid for either a home or a city. Smart environments are so complex for wide areas. It needs well organised ontological relations. Otherwise computer cannot handle huge amount of interactions occurred at the same time.

Let us assume that person A stays in the conference room. He or she wants to go to a theatre close to him. Spatial concepts connected each other with relation of isPartOf. Person A isPartOf Conference Room. Conference Room isPartOf Building A. Building A isPartOf District A. District A isPartOf City A. Theatre A isPartOf District A. Theatre B isPartOf District A and Theatre C isPartOf District C (Figure 2). We are able to inference that Theather C and Person A are not at the same district.

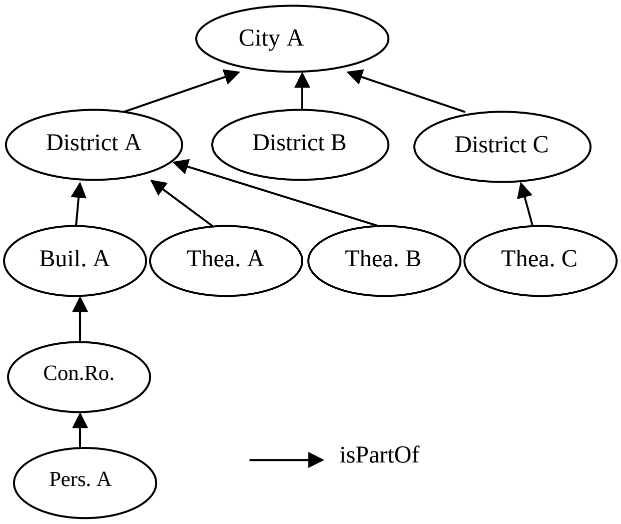


Figure 2. Ontological relations.

District A is the lowest common level of the Person A, Theatre A and B. Theatre C isPartOf District C. Theatre A and Theatre B are the closest theatres to Person A. Now one question remain that which theatre is the closest to Person A in District A. To answer this question by ontologically, smaller entities than a district should be created. partsIn this method, areas (polygons) should be divided small parts to get realistic results to the queries (Figure 3).

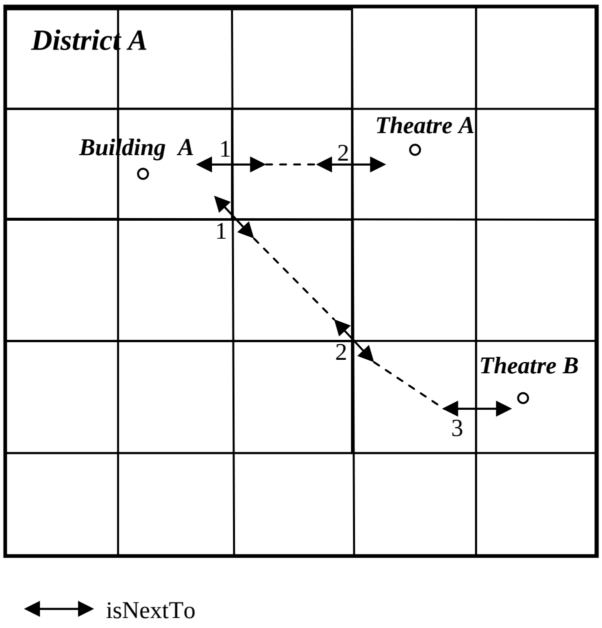


Figure 3. Retrieving the neighbours of Building A by using isNextTo in District A.

isNextTo relation defines neighbours of small District parts. Building A has two isNextTo relation to Theatre A whereas there is three relation to Theatre B. The fewer relation numbers show the closest places. Consequently Theatre A therefore can be obtained as the closest to Building A.

4. CONCLUSION

Ontological approach including concepts, relations and instances is essential for smart environments like LBS. This paper presents a simple spatial ontology in order to provide a location sensitive context aware computing system. System enables a solution for finding closest place at the smart environments. The solution differs from distance based solutions because of its applicable data structure to an ontological context aware system.

This ontological structure should be extended to handle more LBS components such as navigation and meeting queries. Ontological inferences provide implicit context so as to obtain additional information about the situation.

Statistical analyses of the ontological concepts and their relations have not been completed yet. After the analysis, some obvious benefits of the knowledge base also will be shown later. This paper also ignores roads to determine closest place. To obtain more accurate result road entity should be also added to knowledge base.

REFERENCES

Chen, H., Finin, T. and Joshi, A., 2004. An ontology for context aware pervasive computing environments. Knowledge Engineering Review 18, pp. 197–207.

Christopoulou, E., Goumopoulos, C., Zaharakis, I. and Kameas, A., 2004. An ontology-based conceptual model for composing context-aware applications. Workshop on advanced context modeling, reasoning and management, 6th International Conference on ubiquitous computing, Nottingham, England.

Dey, A. and Abowd, G., 2000. Towards a better understanding of context and context-awareness. CHI’2000 Workshop on the What, Who, Where and How of Context Awareness, The Hague, The Netherlands.

Gu, T., Wang, X., Pung, H. and Zhang, D., 2004. An ontology based context model in intelligent environments. Proceeding of Communication Network and Distributed Systems Modeling and Simulation Conference, San Diego, California, USA.

Hong, D., Kim, D., Yun, J. and Han, K., 2005. An adaptive location data management strategy for context-awareness in the ubiquitous computing environment. Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Spatial Analysis, Data Mining and Data Fusion, Beijing, China pp. 57–62.

Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B. and Dean, M., 2004. SWRL: A semantic web rule language combining OWL and RuleML, http://www.w3org/Submission/SWRL, May 2004.

Kim, J., Kim, J., Hwang, H. and Kim, C., 2005. Location sensitive tour guide services using the semantic web. Lecture Notes in Computer Sciences 3682, pp. 908–914.

Lutz, M. and Klien, E., 2006. Ontology-based retrieval of geographic information. International Journal of Geographical Information Science 20, pp. 233–260.

McGuinness, D.L. and van Harmelen, F., 2004. OWL web ontology language overview, W3C Recommendation, http://www.w3.org/TR/owl-features/, 2004.

Weissenberg, N., Gartmann, R. and Voisard, A., 2006. An ontology-based approach to personalized situation-aware mobile service. GeoInformatica 10, pp. 55–90.

Zhu, F., Mutka, M. and Ni, L., 2005. Service discovery in pervasive computing environments. Pervasive Computing 4, pp. 81– 90.

DESIGN AND DEVELOPMENT OF FIELD SYNCHRONOUS DATA COLLECTING SYSTEM OF MINING AREA SURFACE DEFORMATION INFORMATION

Yong Sun [a,*], Min Ji[a], Tao Jiang[a], Xiaojing Yao[a]

[a] Geomatics College of Shandong University of science and technology, 579 Qianwangang Road Economic & Technical Development Zone, Qingdao, China, 266510-sunyong3s@gmail.com, jimin@sdust.edu.cn, tjiang@126.com, lilyyxj@sina.com

KEY WORDS: Embedded GIS, Map Service, Surface Deformation, DGPS, Data Collecting, System Design

ABSTRACT:

In order to grasp the changing situation of ground subsidence and surface collapse caused by underground mining in mining area, and raise the efficiency of field data collecting, this paper's attention is paid to develop Field Collecting System Based on Wireless Communications to acquire spatial data and attribute data. The mobile devices use Web Services to synchronous data with the server. The system can get the required map from server by Map Service, it acquires high precision coordinate by making use of the GPS Differential Technology. The attribute data can be updated to the server by Web Services. The Field Collection System of Surface Deformation Information in Yanzhou Mining Area is developed on the basis of analyzing how the underground mining work affects the ground surface. The System will provide high precision, the lasted data for researching and monitoring the surface subsidence or collapse. It integrates field ascertainment and survey. That simplifies the work of data collecting and entering into the database, and meets the requirement of real-time data.

1 INTRODUCTION

The ground subsidence and surface collapse is common in the mining area. The major cause for surface deformation is a wide range of mined-out space underground caused by coal mining (LIU Guang, 2008). While the surface subsidence, landslides and other geological disasters have a serious impact on mine production activities and people's life, it is very important to know well the changing situation of ground surface for analyzing and predicting the change trends of ground subsidence, helping for adjusting the way of exploitation and planning of land use in the mining area, reducing the hazards of surface deformation. So it is necessary to design a field data collecting system which applies to mining area data acquisition. The system can acquire the position coordinate and the property of the surface, and it will upload the collected data to the server. Then the server will use this data to analyse the surface change trends.

2 SYSTEM DESIGN

2.1 Requirements Analysis

The Field Synchronous Data Collecting System of Mining Area Surface Deformation Information used to collect the planimetric position, elevation, the potion of mined out space, the area, the exploitation date and the picture of mining area where the subsidence has occurred. And these collected data can be uploaded to the server for analysing. There are some fixed monitoring points in the subsidence area. These points are used to analyze the position change, especially the elevation change. We can use the RTK to survey these points for high-precision coordinate. There is a CORS station in the Yanzhou coal mining area, so the GPS receiver can get differential data from the CORS by GPRS.

2.2 System Architecture

On the server, the map spatial data is managed by the ArcSDE. The layers which need to be updated are published as Map Service. And the attribute data is stored in the Oracle. The Web Services are published for access the server by the mobile device. Figure 1 shows the architecture of Field Synchronous Data Collecting System of Mining Area Surface Deformation Information.

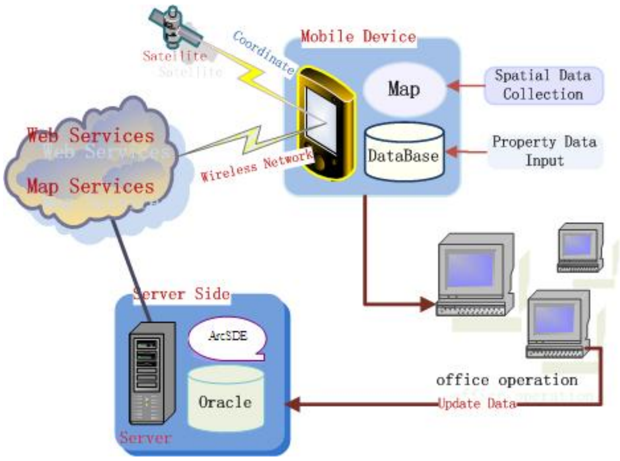


Figure 1. The Architecture of Field Synchronous Data Collecting System.

On the mobile device, the field data collecting system will get the map of the area where need to collect data from the server by the wireless network. Then the collected spatial data will be store in the map. The attribute data are stored in the embedded database-the SQL CE.

[*] Corresponding author. E-mail address: sunyong3s@gmail.com; Tel 15863012147

When the work of field data acquisition is finished, the spatial data can be updated to the map layers in the ArcSDE using the published Map Service; and the attribute data can be updated to the Oracle by the Web Services which are published on the server. The second way for uploading the data is to use the C/S mode on PC when processing data in the office. The system will avoid the operator of importing and exporting data, and it is not necessary to convert the data formats. That will raise the efficiency of field data collecting and simplify the data manage.

2.3 System Functional Design

The system is designed for acquiring data in the field. The data collected in the field are managed by a project. Each project contains a map, a database and a coordinate system file etc. One project stores the data of one region or a day's work. The project can be created as it is required.

The system functions include getting coordinate data form the GPS receiver by the port or the Bluetooth, coordinate transformation, storing the spatial data in the map, project creating and managing, attribute data entry and save, getting map from the Map Services, updating the field data by wireless network, embedded database management, GPS control, taking pictures, map operate, such as map zoom, map pan, map query and so on.

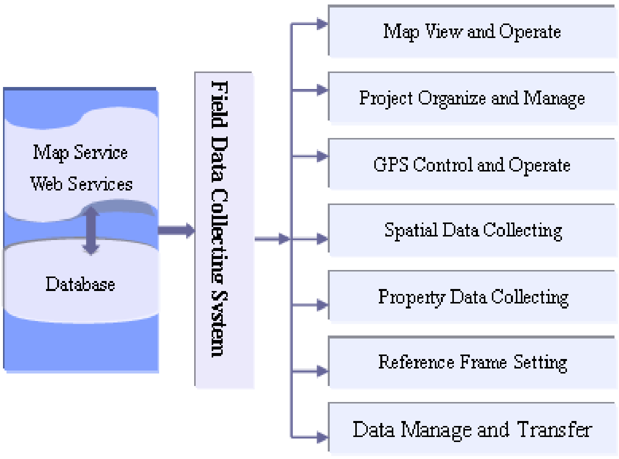


Figure 2. The System Functions Modules

A project is used to manage the collected data. The system can create, open and manage the projects. The reference frame setting module can create or edit a coordinate system file which contains the ellipsoidal parameter, the projection parameter and the parameter for coordinate conversion. A project links to a coordinate system file; the project will use the coordinate system file to convert the spatial data from the WGS 84 geographic coordinate system to the coordinate system as specified in the file when working in field. In the data manage and transfer module can manage the data and transfer the data to server.

2.4 System Project Composition

The system uses projects to manage the data collected at different time. When collecting data in the field, first of all it is to create a project to store the data. After the GPS receiver works smoothly, the system will get coordinates form the receiver. In the project， the spatial data stored in the map, and the property stored in the database. Figure 3 shows the project composition of the system.

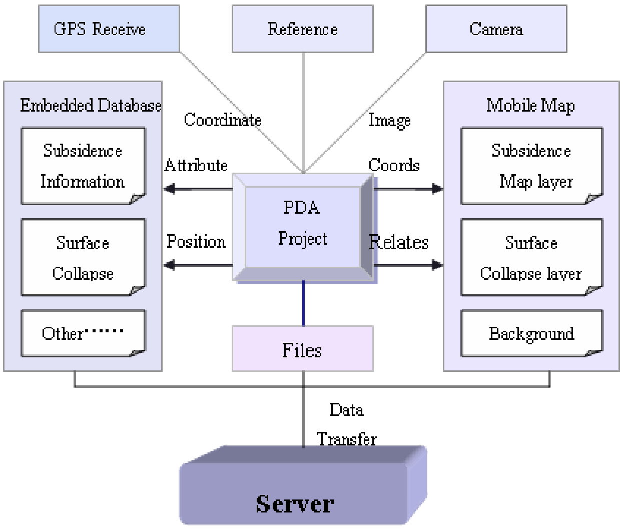


Figure 3. Project composition of the system.

There are several layers in the map. Before collecting spatial data, setting a layer which will be edited, then the coordinates will be saved in the layer. After finish a feature, can entering its property. The attribute data will be saved in the embedded database. When finished the work of data acquisition, the new data can be updated to the server by wireless network or by office operation.

2.5 The Data Organization

The system's data includes map data and attribute data. The mobile map can be acquired from the Map Service on the server or the Map Cache stored in the mobile device.

The Map Cache is created from Map Service by ESRI ArcTools or by ArcGIS Mobile from the wireless network. There are several layers in the map; the layer's type is point, line or polygon. When collecting the coordinate in the field, the coordinate data will be shaped as point, polyline or polygon and stored in one layer. The map can contain image layer. Only the map layer in the Map Service form the ArcSDE can be edited in the ArcGIS Mobile. Another kind of map used in the system is called Base Map. The Base Map created from ESRI map document (MXD file). The layer in the Base Map can not be edited. Both the Map Cache files and the Base Map files have been compressed a great deal; they are much smaller than the layer in the data source. The Map Cache can be get from the Map Service by wireless network while the Base Map can't.

The attribute data is stored in the SQL CE. It's very easy to manage the property by the SQL CE. It is not possible to work with the Oracle directly in the mobile device. So the system calls the Web Service method to query and update data with the server.

2.6 The Hardware Environment of System

The mobile device is Zhonghaida GPS PDA. The Novatel GPS OEM device is built-in in this PDA. And the GPS Receiver can make use of differential data which received from SBAS or the base station to improve the positioning precision. And the PDA includes Bluetooth, so the system can use the built-in Bluetooth to connect to an exterior GPS receiver when need more high precision positioning data. The operating system of the PDA is Windows CE 5.0. That OS is convenience for developing and deploying software. The camera in the PDA can used to take photos for recording the spot in the fields. And this PDA can use the GPRS to connect to the server for requesting or updating data.

3 CRITICAL TECHNOLOGY

3.1 Embedded GIS

Embedded GIS integrate the GIS functions with the mobile technology; it is the expansion of GIS technology from the office into the field (ZHANG Shi-huang, 2001). A mobile GIS enables field-based personnel to capture, store, update, manipulate, analyze, and display geographic information. Mobile GIS integrates one or more of the following technologies: mobile devices, positioning system (GPS, GLONASS etc), and wireless communications for Internet GIS access.

A mobile GIS based on wireless communications can connect to a server through the use of wireless network and Internet. The mobile device sends request and the server return the results what the client needs. For the professional, the mobile GIS can be used to collect field data, and then transfer the data to a server by wireless network. The collected data is stored in GIS data format, and can be updated to the server timely. That will ensure the data can be updated promptly and the real-time of GIS. For the general user, the mobile GIS can get the latest local map from the server and query the place where is interesting. Combined with GPS, the mobile GIS can get the user coordinates and find a path showing how to go to the destinations.

3.2 Web Map Services

A Web Map Service (WMS) is a standard protocol for serving georeferenced map images over the Internet that are generated by a map server using data from a GIS database. A WMS request defines the geographic layer(s) and area of interest to be processed. The response to the request is one or more geo-registered map images that can be displayed in a browser application. The interface also supports the ability to specify whether the returned images should be transparent so that layers from multiple servers can be combined or not (Jeff de la Beaujardiere, 2004). WMS is a widely supported format for maps and GIS data accessed via the Internet and loaded into GIS software, on the client side.

The WMS can provide spatial data for the use form the Internet/Intranet, and can realize the map expression, map query and map positioning (Feng Jin-jun, 2006). The WMS promotes the spatial data switching and resource sharing. ArcGIS Server can provide map services capability which not only support mapping and map viewing, but can also support modelling and geoprocessing, mobile GIS services, and open publishing as OGC WMS and KML. When publishing a map service, it can be set to support mobile GIS services. So the field synchronous data collecting system can access the spatial data which have been published by the map service.

3.3 Global Positioning System

Global positioning system (GPS) provides reliable positioning, navigation, and timing services to worldwide users on a continuous basis in all weather, day and night, anywhere on or near the Earth (Jules G., 2002). GPS has become a widely used aid to navigation worldwide, and a useful tool for map-making, land surveying, commerce, scientific uses, tracking and surveillance.

Space-based augmentation system (SBAS) works on principles similar to DGPS. Correction signals are sent from a network of ground stations to a master ground station that transforms the signals into a grid of correction signals. The grid is sent to one or more geostationary satellites that orbit 36000km above the equator, and is then broadcast to Earth. The SBAS includes the WAAS of the United States, the EGNOS of Europe, and the MSAS of Japan (LIU Wen-tao, 2008). The MSAS signal cover Asia-Pacific Region, and most parts of china can receive MSAS signal. When the positioning precision demand is not high, we can make the GPS receiver get differential data from the MSAS.

The principle of difference GPS is one receiver (Base Station) set at the coordinated point, this receiver will calculate the differential data and send them to other receiver (Roving Station) real time (Kato, 2000). When the roving station is receiving the GPS data, it is also receiving the differential data to improve the positioning precision.

There are two kinds of DGPS: Real Time Differential (RTD) and Real Time Kinematic (RTK). And the positioning precision of RTK is much higher. So when surveying the monitoring point, we can choose the RTK GPS to acquire the coordinates.

3.4 Coordinate Conversion

The reference system of coordinate data received from GPS is WGS 84, but in china the usual coordinate systems are Beijing 54 and Xian 80. So we must convert the coordinate data from the WGS 84 to Beijing 54 or Xian 80. The WGS 84 coordinate system is one of the geographic coordinate systems; and the Beijing 54 or Xian 80 coordinate system is one of the planar coordinate systems.

There are many methods to complete the conversion. It is usually need two steps for the conversion. The following introduces the conversion between WGS 84 and the planar coordinate systems. The first step is convert the rectangular space coordinates of WGS 84 ellipsoid to the rectangular space coordinates of other ellipsoid. The Burse model is a useful method. It is need 7 parameters for the conversion (Zhang Fengju, 1999).

$$\begin{aligned} X &= (1+k)X_{84} + \varepsilon_z Y_{84} - \varepsilon_Y Z_{84} + \Delta X \\ Y &= (1+k)Y_{84} - \varepsilon_z X_{84} + \varepsilon_x Z_{84} + \Delta Y \\ Z &= (1+k)Z_{84} + \varepsilon_Y X_{84} - \varepsilon_x Y_{84} + \Delta Z \end{aligned} \tag{1}$$

The second step is Gauss projection, this step convert the coordinate form geographic coordinate system to planar coordinate system (Zhang Fengju, 1999).

$$\begin{aligned} x &= X + \frac{l^2}{2} N \sin B \cos B \\ &+ \frac{l^4}{24} N \sin B \cos^3 B (5 - t^2 + 9\eta^2 + 4\eta^4) \\ &+ \frac{l^6}{720} N \sin B \cos^5 B (61 - 58t^2 + t^4) \\ y &= lN \cos B + \frac{l^3}{6} N \cos^3 B (1 - t^2 + \eta^2) \\ &+ \frac{l^5}{120} N \cos^5 B (5 - 18t^2 + t^4 + 14\eta^2 - 58\eta^2 t^2 \end{aligned} \qquad (2)$$

3.5 ArcGIS Mobile

ArcGIS Mobile SDK is provided by ESRI for developing mobile GIS application; it is belong to the ArcGIS Server. The ArcGIS Mobile lets the mobile device like PDA accesses the mobile GIS services published by the server. The map data from the ArcSDE Geodatabase can be edited in the ArcGIS Mobile online or offline. If the data are edited offline, the result will be stored in the map cache, and can be updated to the server when the wireless network is available.

ArcGIS Mobile helps organizations deliver GIS capabilities and data from centralized servers to a range of mobile devices. We can use ArcGIS Mobile to deploy intuitive and productive mobile GIS applications to increase the accuracy and improve the currency of GIS data across your organization. It's easy to use ArcGIS Mobile applications enable field staffs who do not necessarily have any GIS experience to do Mapping, Spatial query, Sketching, GPS integration, GIS editing, Wireless data access to ArcGIS Server Web services (ESRI, 2007a). With the help of ArcGIS Mobile, the staff don't have to go back to update the data that he collected in field to the geodatabase; he can update the data by wireless network.

3.6 Web Services

Web services are self-described software entities which can be advertised, located, and used across the Internet using a set of standards such as SOAP, WSDL, and UDDI. Web services encapsulate application functionality and information resources, and make them available through programmatic interfaces, as opposed to the interfaces typically provided by traditional Web applications which are intended for manual interactions. Web Services connect computers and devices with each other using the Internet to exchange data and combine data in new ways (Sheila A, 2001). Web Services can be defined as software objects that can be assembled over the Internet using standard protocols to perform functions or execute business processes.

With the help of Web Services, the mobile device could do some complex operations that he can't complete itself. It just calls the functions published by the web services to perform tasks to get and update data. It needn't to implement the functions by the software in PDA.

4 SYSTEM IMPLEMENT

4.1 System Software Environment

Figure 4 shows the soft environment of Field Synchronous Data Collecting System of Mining Area Surface Deformation Information. On the server side, the database Oracle manages all the data, and the spatial DB Engine is used to operate the spatial data. The ArcGIS Server publishes map services which support mobile GIS access. On the mobile side, the system gets map data from the map service through the use of ArcGIS Mobile, and the coordinate data collected in field stored in the map layers. The attribute data stored in embedded database. For the PDA's functions are limited, it can't operate the data directly in the Oracle. So the system will call the web services methods published on the server to update data.

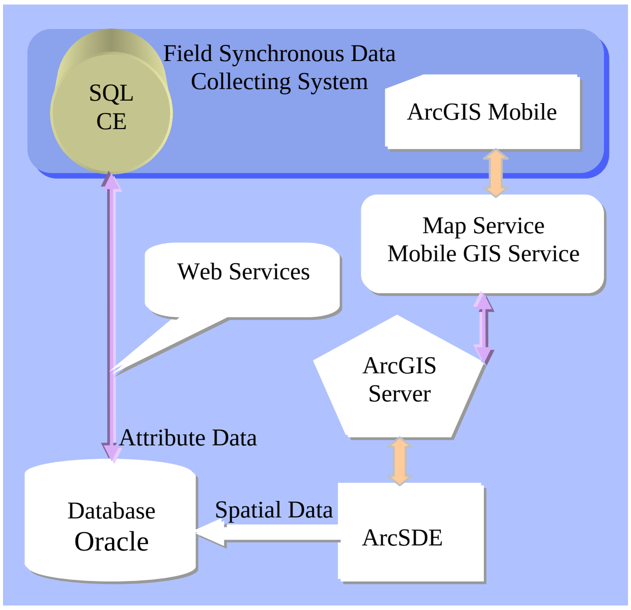
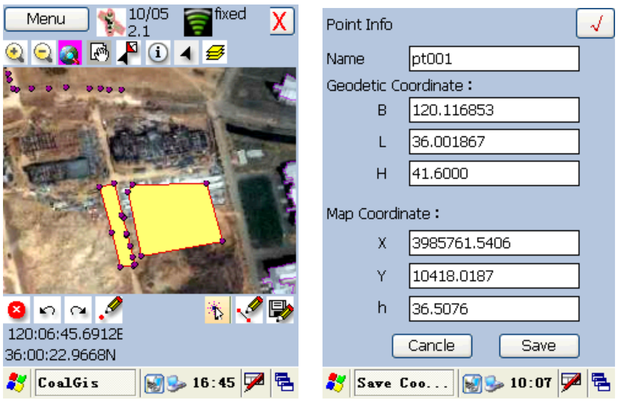


Figure 4. The software environment of System

4.2 Result

The field synchronous data collecting system of mining area surface deformation information based on wireless network supports Windows CE 5.0 and above. It is easy to use the system to collect data of Mining Area Surface Deformation and to update the data instantly to server by map service and web services from the wireless network. The system can get high accuracy positioning data by differential technology. And the system will convert the coordinate data of WGS 84 coordinate system to the coordinate system as specified in the project. Figure 5 shows two interfaces of the system.



120:06:45.6912E
36:00:22.9668N

Point Info

Name pt001

Geodetic Coordinate :

B	120.116853
L	36.001867
H	41.6000

Map Coordinate :

X	3985761.5406
Y	10418.0187
h	36.5076

Figure 5. System map interface and the interface of collecting coordinate data.

5 CONCLUSIONS

The article discusses how to use web services and mobile GIS services to update data to the server by wireless network. With the SBAS and RTK technology can improve the positioning accuracy. Using the mobile GIS technique in field data acquisition will improve the efficiency of data updating. Developing a mobile GIS system for synchronous collecting the data of mining area surface deformation information can provide the latest data for analysing the changing situation of ground subsidence.

Reference

ESRI, 2007, ArcGIS Mobile Overview, Califonia, America. http://www.esri.com/software/arcgis/arcgismobile/index.html, (Accessed 12 NOV. 2009)

Feng Jin-jun, Wang Ying, 2006. Research on WebGIS based on Web services. *Journal of North China Electric Power University*, 33(2), pp.101-104.

Jules G. McNeff, 2002. The Global Positioning System. *IEEE Transactions on Microwave Theory and Techniques*, 50(3), pp.645-652.

Jeff de la Beaujardiere, 2004. OGC Web Map Service Interface. *Open GIS Consortium Inc*, OGC 03-109r1, pp.34-36.

Kato T., Y. Terada, M. Kinoshita, H. Kakimoto, H. Isshiki, M. Matsuishi A. Yokoyama, and T. Tanno, 2000. Real-time observation tsunami by RTKGPS, *Earth Planets Space*, 52, pp.841–845.

LIU Wen-tao, XING Lu-lu, LIANG Hong, WANG Wen-hui, 2008. Bit Synchronization Design and Performance Simulation of Satellite Based Augmentation System (SBAS) Receiver, *Telecommunication Engineering*, 48(7), pp.54-56.

LIU Guang, GUO Hua-dong, RAMON Hanssen, 2008. The Application of InSar Technology to Mining Area Subsidence Monitoring, *Remote Sensing For Land & Resources*, 76(2), pp.51-52.

Sheila A, Mcllraith, Tran Cao Son, and Honglei Zeng, 2001. Semantic Web Services. *IEEE Intelligent Systems (Special Issue on Semantic Web)*, 16(2), pp.46-53.

ZHANG Shi-huang, FANG Yu, 2001. Development and Significance of Micro-Embedded GIS Software. *Journal of Image and Graphics*, 6(9), pp.901

Zhang Fengju, Zhang HuaHai, Zhao ChangSheng, Meng Lumin, Lu Xiushan, 1999. *Control Surveying*, China Coal Industry Publishing House, Beijing, pp.235-251.

APPENDIX

Supported By Chinese National Programs for High Technology Research and Development (2009AA12Z147);
Supported By Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping (2009B14);
Supported By Key Laboratory of Advanced Engineering Surveying of SBSM (TJES0805)

COASTLINE CHANGE MEASUREMENT AND GENERATING RISK MAP FOR THE COAST USING GEOGRAPHIC INFORMATION SYSTEM

D. K. Raju, K. Santosh, J. Chandrasekar & Teh Tiong-Sa

Physical Oceanography Research Laboratory,
Tropical Marine Science Institute, National University of Singapore.
12A, Kent Ridge Road
Singapore 119223

KEY WORDS: Coastline, Chart datum, Geographic Information System, Risk map

ABSTRACT:

Information on the pattern, rhythm and long-term trend of shoreline dynamics is vital to managing the coast, especially in mapping risk areas and in establishing the width of construction setback lines. Unfortunately, such data are often missing or where present of doubtful quality. In this paper, we share our experience in mapping coastline change in Singapore, using geographic information system (GIS). Most important before embarking on measuring coastline change is defining the coastline and deciding on a working definition where this line can be established on the ground, on maps or other sources of data. Complications arise where the coastline is extracted from different data sources to detect changes. In Singapore, coastline is defined is 2.515 m chart datum and this line which theoretically separates land from sea is shown on cadastral maps.

The East Coast Park of Singapore is selected to illustrate the different techniques in measuring coastline change using GIS. Techniques include setting up a series of profile lines monitored and analyzed in GIS over a decade to detect changes, and using GIS to generate the 2.515 m chart datum line from a dense network of elevation points collected on the beach. Maps showing various rates of shoreline erosion are then derived and finally a Risk map for the coast generated.

1. INTRODUCTION

Information on coastline change from seasonal to longer term trend constitute an essential and vital input in any coastal management plan, so that areas of potential loss to erosion can be identified and appropriate land use planning adopted. Unfortunately, such information is lacking and, where present, often of doubtful reliability. In Singapore, additional problems are encountered: the coastline is often <40 years old, created after land reclamation and still undergoing changes towards equilibrium. Along many coastal sectors the retreat has been planned and many beaches have been nourished to slow down shoreline retreat and maintain popular beaches. Information on beach management and the amount of sand used for nourishment is usually not available. Hence, interpreting coastline change under such circumstances is difficult.

A proper understanding of coastal processes in general and the local history of land reclamation is required to interpret the changing coastline. A retreating coastline may not be of concern as it may be planned. On the other hand, a stable coastline maintained by nourishment should be of concern. Different sectors of the ECP show diverse behavior, with some stable while others receding or advancing. This spatial pattern of different rates of erosion and accretion provides the basic information for micro management of the coast. Emplacement of canal structures and additional breakwaters has also changed the sedimentation pattern. All these have made interpreting coastline change even more difficult. Despite this, careful analysis of the data would still yield useful information on erosion hazards and risks for development. Shoreline management is not an exact science and may entail a series of responses as unanticipated problems emerge from time to time. When accurate information on shoreline dynamics is lacking, it may be wiser to adopt a precautionary principle in management. Sometimes, facilities are constructed in a high risk area for temporary use and this causes misunderstanding when not properly explained. Different types of land use strategies require different types on erosion hazards. Data on seasonal coastline change is used to establish the width required to create a buffer zone where the shore can retreat and recover unimpeded. Longer term change will aid in land use planning by establishing the highly dynamic and sensitive areas that should not be developed.

2. MEASURING COASTLINE CHANGES: CONSIDERATIONS

Important questions that need to be addressed in coastline change studies include why coastline change is measured, what is being measured and how the measurement is carried out (Teh et al, 2005). It is important to understand the reasons for measuring coastline change and the accuracy of measurement so that the results derived can be properly applied for planning purposes. Short-term changes have less value for planning compared to long-term changes. On the other hand the latter may not capture the seasonal changes provided by the former which may be critical. If the reasons for measurements are understood, this will help towards a suitable definition of the coastline and the choice of techniques to be employed. Past studies have used different datum for measuring coastline change. The popular datum or feature used includes a legally defined coastline which varies from country to country, MHWS, vegetation line, seaward foot of coastal dunes and coastal scarp. Different values are obtained with the use of different datum or coastal features. Longer term coastline changes usually use the oldest reliable topographic map as a baseline whereas short term changes employ field techniques to detect small changes in coastline position.

2.1 Short-Medium Term Measurements:

The usual technique is to conduct repeat surveys along fixed profile lines from a temporary bench mark (TBM) down to low water. Spacing of profiles and intervals between successive profiling is influenced by objectives, coastal landscapes and man power. Usually four surveys are conducted each year to capture monsoonal influence and from 1 to 5 profiles are set up for each sediment compartment to capture the spatial variation. Profiles may also be set at fixed intervals along the coast. Monitoring over several years will provide useful data on medium term coastline change.

2.2 Longer-Term Measurements:

The usual method of detecting longer term coastline change is by overlaying a series of topographic maps of the same scale from different years. The displaced coastline is then measured and the rate of change calculated. The coastline on the topographic map is represented by the blue line. This method only allows large scale changes to be detected, as a 0.2 mm width line would represent 10 m and 2 m on the ground on a 1:50,000 map and 1:10,000 map respectively. When vertical aerial photographs became easily available, this new source of information was used to supplement topographic maps. The problems that came with this new approach was trying to convert the aerial photographs, often showing tilt and of different scale from the topographic maps, to the scale of the topographic map. A more serious problem was in deciding what the blue line on the topographic map represents on the aerial photographs.

3. EAST COAST PARK, SINGAPORE

The East Coast Park (ECP) is reclaimed from the sea in several phases by Housing and Development Board of Singapore to create land for the East Coast Parkway, public housing and recreational space (Figure 1). (Wong, 1973) suggested a sequential development of beaches along the ECP to equilibrium, forming five beach types according to their developmental level. At the time of his study, he classified the bays from HL2 (Headland2) to HL4 (Headland 4) as new beaches and developing beaches.

3.1 The Study Area - Headland 2 to Headland 4

A short section of the ECP, from HL 2(Headland 2) to HL 4(Headland 4), is selected as the study area as shown in Figure 1(Ikonous Satellite image, 2000). This is in front of McDonald and represents one of the more popular coasts along the ECP. The coastline is badly eroding and shoreline retreat and loss of facilities have resulted in various responses to hold the coastline. This coastal sector was designed to consist of three inverse J-shaped bays developed between four emplaced headlands. The net littoral transport after land reclamation is westwards as suggested by the pattern of erosion and accretion adjacent to headlands and canal structures. However, there is drift reversal at times. The study area in this paper covers only a short sector of the ECP monitored for coastline change.

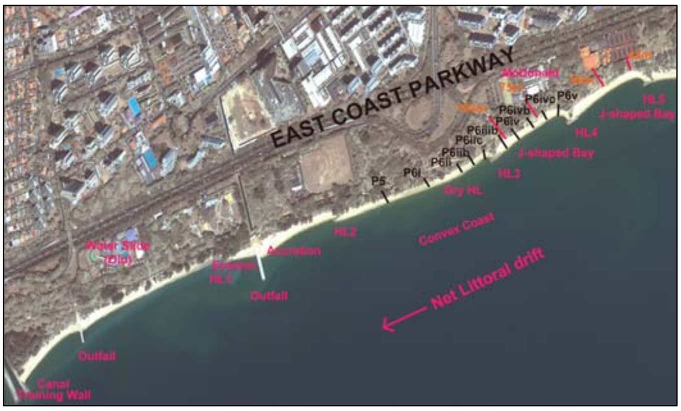


Figure 1. The East Coast Park study area

4. SHORT MEDIUM TERM COASTLINE CHANGE

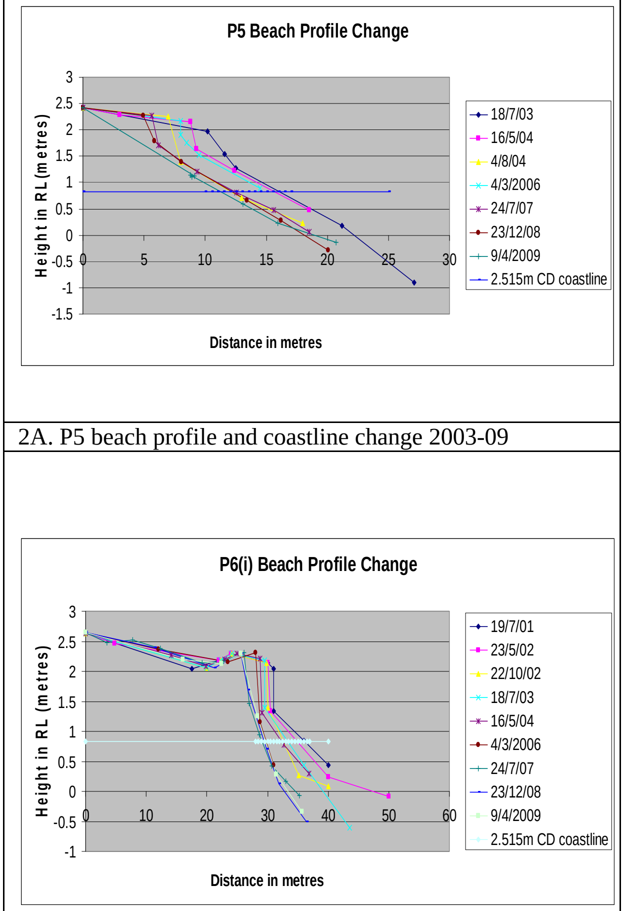
The profiles are grouped under compartment 1, consisting of P5 to P6iic and compartment 2, consisting of Piiib to P6v. The profiles are analyzed for rate of coastline change individually. The profile monitored and analyzed in the study area consists of 10 profiles (Table 1).

Profile	*Location*	*Start-end monitoring*	*No. month*	*Description*
P5	HL2-3	18/7/03-9/4/09	69	East of headland 2
P6i	HL2-3	19/7/01-9/4/09	93	Across dry headland, scarped
P6ii	HL2-3	24/12/01-9/4/09	88	Badly eroding sector, erosion mitigated, scarped
P6iib	HL2-3	13/8/04-9/4/09	56	West HL3, erosion mitigated, scarped
P6iic	HL2-3	10/9/04-9/4/09	55	West HL3, erosion mitigated, scarped
P6iiib	HL3-4	10/9/04-9/4/09	55	East HL3, nourished, scarped
P6iv	HL3-4	24/12/01-9/4/09	88	East HL3, nourished, scarped
P6ivb	HL3-4	13/8/04-9/4/09	56	East HL3, nourished, scarped
P6ivc	HL3-4	13/8/04-9/4/09	56	West HL4, nourished, bermed
P6v	HL3-4	24/12/01-9/4/09	88	West HL4, nourished, bermed

Table 1.Profiles monitored from HL2 to HL4

The field technique uses repeat surveys along fixed lines, with painted ground markers and land marks (Teh, 2000). Each profile begins with a temporary bench mark (TBM) in which the reduced level has been tied to a nearby precise level bench mark (PLBM). Beach profiling is carried out mainly by using a digital automatic level from TBM to Low water mark during low tide (Singapore Tide Tables, 2009). The collected data are processed in profile information system developed in GIS software and the coastline change measured graphically and the rate of change expressed in meter per month and meter per year.This monitored coast is highly dynamic showing changes in profile form and coastline position as a result of beach engineering and monsoonal influence. Between mid 2004 and the end of 2008 the coastland was displaced landwards.

The datum used for defining coastline in this paper is the 2.515 m chart datum or 0.863 m RL for the study area (Singapore Tide Tables, 2009). This datum is shown on cadastral maps to separate land from sea in Singapore. Figure 2 displays the graphs for selected profiles generated by the system. The system also allows the user to choose the different time period and calculate rate of recession of coast for that particular period.



2A. P5 beach profile and coastline change 2003-09

2B. P6i beach profile and coastline change 2001-09

Figure 2.Selected Profiles of Compartment 1

The profile information system manipulates all the input data for a profile and summarizes the results as a report for the individual profile as shown Table 2.

Date profiled	Coastline distance from TBM in metre	Coastline adv/retreat in metre from 24/12/01	Total months from oldest profile	Rate of change in metre per mth/ metre per yr from 24/12/01
24/12/01	12	-	-	-
24/5/02	19.1	+7.1	5	+1.42/+17.04
22/10/02	12.9	+0.9	10	+0.09/+1.08
18/7/03	17.5	+5.5	19	+0.289/+3.474
16/5/04	21.5	+9.5	29	+0.327/+3.931
28/8/07	14.3	+2.3	68	+0.034/+0.406
23/12/08	10.1	-1.9	84	-0.023/-0.271
9/4/09	11.8	-0.2	88	-0.0023/-0.0273

Table 2. P6iv change in coastline position 2001-2009

The system also generates the report for the complete set of profiles as shown in Table 3.

Profile	*Start-end monitoring*	*Months*	*Coastline change (m)*	*Rate of change (+/- m/yr)*
C1-P5	18/7/03-9/4/09	69	-4.5	-0.783
C1-P6i	19/7/01-9/4/09	93	-6.5	-0.839
C1-P6ii	24/12/01-9/4/09	88	-5.3	-0.723
C1-P6iib	13/8/04-9/4/09	56	6.1	1.307
C1-P6iic	10/9/04-9/4/09	55	4.5	0.982
C2-P6iiib	10/9/04-9/4/09	55	-6.3	-1.375
C2-P6iv	24/12/01-9/4/09	88	-0.2	-0.027
C2-P6ivb	13/8/04-9/4/09	56	0.7	0.15
C2-P6ivc	13/8/04-9/4/09	56	4.6	0.986
C2-P6v	24/12/01-9/4/09	88	-0.8	-0.109

Table 3. Coastline change over monitored period

4.1 Interpreting the Profile Data Report:

The medium term coastline change data, summarized in Table 3, showed that the rate of retreat since 2001/2003 for the coast from HL2 to the Dry HL (P5 and P6i) was greater. For the coast from Dry HL to HL3(P6ii), the medium term change recorded a retreat rate in front of the Dry HL since 2001 comparable to the coast to the west. However, towards the east near HL3 (P6iib and P6iic) the coastline has propagated since 2004 because of nourishment and from supply of alongshore sands. Along the inverse J shaped bay from HL3 to HL4, the coastline change is varied since 2004.Profiles P6iiib, P6iv and P6v has shown retreat trend while nourishment in had resulted in profiles P6ivb and P6ivc showing coastline advance. Hence GIS allows the coastal planners and managers to observe the beach profiles more accurately and take important decisions like when the beach need to nourished more appropriately.

5. LONGER -TERM COASTLINE CHANGES

The longer term coastline change of the study area is established by comparing the 1972 2.515 m CD coastline shown on cadastral maps with that of the 2007, 2.515 m CD coastline (0.863 m RL) generated in ArcGIS from a dense network of elevation points collected on the beach using Total Station. The elevation points from low water mark to 5 m were collected along the whole ECP in 2007 was used for this study.

Triangulated Irregular Network (TIN) was created from elevation data collected by field surveys using ArcGIS 3D analyst (Ormsby et al, 2009). Further the TIN (Figure 3a) was converted to Digital Elevation Model (DEM) raster layer using ArcGIS (ESRI Link,2009) at a spatial resolution of 0.25 meter as shown in the Figure 3b.

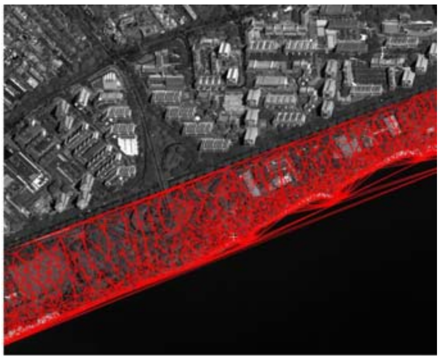


Figure 3a TIN

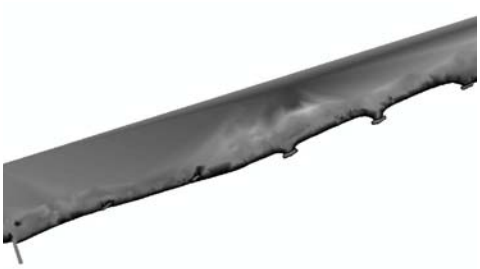


Figure 3b DEM

Figure 3.Generation of DEM

DEM can be validated spatially by comparing it with the spot heights (Herano, 2003). Spot heights provided by the Singapore Government survey agency were used for compassion in our study. The spot heights were imported as thematic layer in ArcGIS and overlaid over the DEM and compared spatially (Santosh et al, 2009).The resultant error value revealed a good correlation between the verification points and the DEM. The vertical accuracy was below 0.1m in the areas where field survey work was carried out. This DEM raster was analyzed for the elevation value 0.863m by executing spatial query using raster calculator tool in ArcGIS spatial analyst which resulted in the generation a coastline with the 0.863 m elevation.

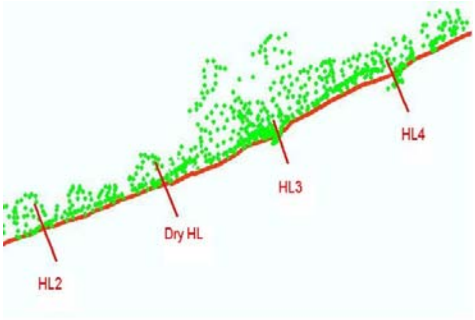


Figure 4a Coastline 2007

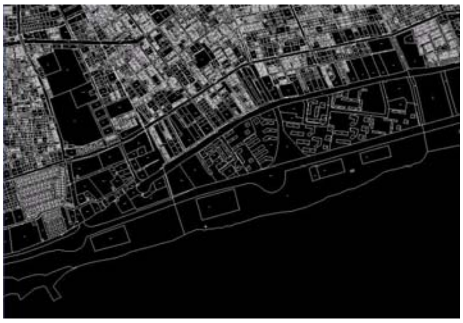


Figure 4b Coastline 1972

Figure 4. Coastline of Study area.

1972, 2.515 m CD (0.863 m RL) coastline shown on cadastral maps (Figure 4b) was obtained from Singapore Land Authority Department, Singapore. World view -1(WV-1) satellite image tasked in July 2008 for the study area was georefernced in ArcGIS and used as the backdrop layer for the comparison. The 1972 coastline and 2007 coastline were imported in ArcGIS as a thematic layer and were overlaid on the WV-1 satellite image for comparison. The average retreat was calculated spatially by identifying the area of polygon formed between the headlands and dividing it with the corresponding distance between the two headlands. The maximum retreat was calculated spatially by measuring the distance between the lines using tools in Arc Map module of Arc GIS software. Comparison of different year's topographic maps of same scale in GIS software is more accurate and reliable source for coastline change measurement and coastal land use management.

5.1 Results of the Longer-Term Coastline Change*:*

Comparing the 2.515 m CD coastline of 1972 and 2007 and examining the location of the headlands and dry headland strongly suggests that the whole study area from HL2 to HL4 has retreated since 1972, but the retreat has been varied (Figure 5). The 300 m coast between HL2 and the Dry HL recorded the least retreat, with an average of 6.55 m (-0.187 m/yr) and a maximum of 12.37 m (-0.354 m/yr). The coast from the Dry HL to HL3 shows an increasing landward displacement of the coastline eastwards. The average retreat was 29.82 m (-0.852 m/yr) and the maximum retreat 46.52 m (-1.33 m/yr). The greatest retreat took place between HL3 and HL4, where the 1972 coastline was evenly displaced landwards to the headland-breakwater built on land. Average retreat was 45.35 m (-1.29 m/yr) and the maximum retreat was 49.52 m (-1.41 m/yr).

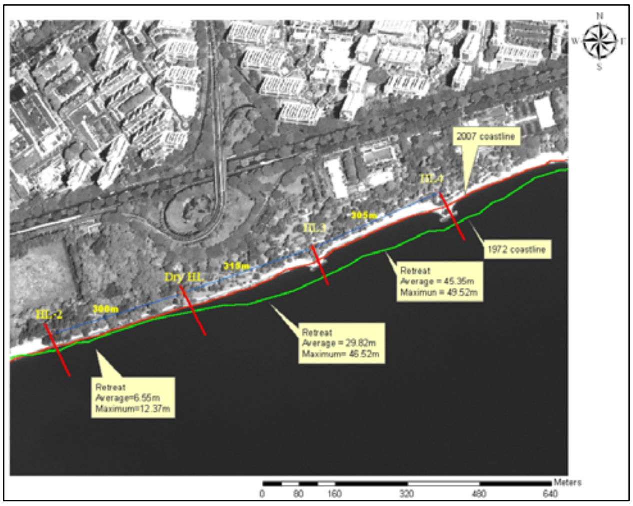


Figure 5. Retreating coastline of ECP 1972-2007 (35 years)

Long term coastline change records a retreating. The loss of beach lamps, tracks, drains and trees to erosion indicate that the erosion was unanticipated. On the whole, the coastline is badly eroding and shoreline retreat and loss of facilities have resulted in various responses to hold the coastline. Intervention using beach nourishment, relocation of tracks and trying to hold the coastline using structures are the usual responses for coasts facing critical erosion.

6. A COASTAL EROSION HAZARD AND RISK MAP OF ECP

There are different ways of classifying erosion. An obvious method is to group them under different rates of erosion (e.g. high, moderate, low) providing such data is available and reliable. In Samoa, as part of a study to identify the various coastal hazard zones and formulate policies for sustainable coastal management, coastal erosion hazard zones for 100 year time frame were mapped, based on erosion rates calculated using GIS (Taulealo and Bismark-Crawley,2002). Erosion is widespread along the whole ECP, especially within the study area. The beaches will continue to erode and intervention measures will have to be carried out to maintain the beaches and hold the coastline position. Intervention measures will become increasingly costly and will have to be carried out at shorter intervals. Other potential option to be considered include the enhancement and preservation of natural protection (e.g, replanting of mangroves and sea grass), use of softer options such as artificial nourishment and raising the height of ground of buildings and dewatering (Wong, 2003). A rising sea will lead to accelerated erosion. Managing the ECP will become increasingly more challenging and will need all the help possible. It is with this in mind that an initial attempt is made here to prepare a coastal erosion hazard and risk map for the study area, which can later be extended to cover the whole island. The coastal erosion hazard and risk map generated by using GIS shows various categories of erosion, together with the future coastline position in 15, 30, 50 and 100 years (Figure 6).

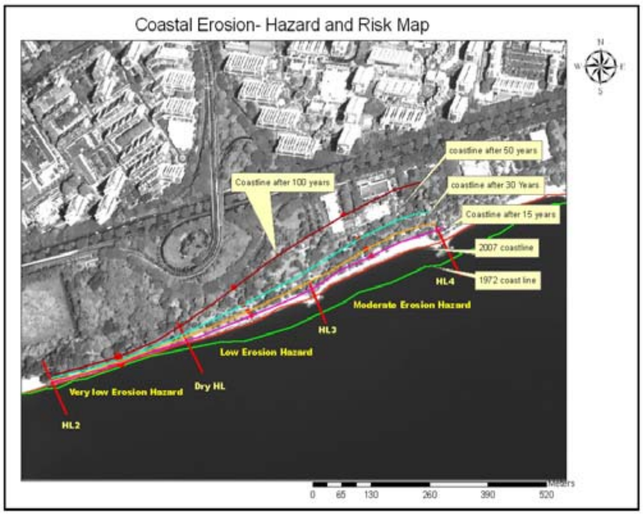


Figure 6. A coastal erosion hazard and risk map

Erosion hazards for the current study have been tentatively grouped under 5 classes based on the rate of retreat. These are <0.5 m/yr (very low), 0.5-1.0 m/yr (low), 1.0-1.5 m/yr (moderate), 1.0-1.5m/yr (high) and >1.5m/yr (very high). Three classes of erosion hazard are identified for the study area based on the long term rate of coastline change. This set of data is favoured over the shorter term change because of recent human intervention. The coast from HL2 to Dry HL is classed as very low, the coast from Dry HL to HL3 as low and the coast from HL2 to HL as moderate. The future projection of coastline was made by assuming the rate of retreat as linear. The projected future coastline position for different years allows coastal planners to assess risk for developments at different distances from the sea. The map serves as a guide and should be reviewed from time to time. Obviously the closer to the beach the greater the future risk. This coastal erosion hazard map acts as a guideline for setting up the setback lines by the coastal land use planners and other coastal land stake holders. The use of coastal construction setback lines is an important tool in coastal management and many countries have used this concept in various forms to regulate development and prevent future losses of property, structures and life caused by shoreline erosion.

7. CONCLUSION

Reliable information on short and long term trends in shoreline dynamics form the basis for identifying coastal sectors of varying sensitivities to development and aid in proper landuse planning of the coastal zone. Unfortunately, such information is lacking and should be addressed immediately. It is important to have information on the rate of natural retreat so that future shoreline positions for different years can be predicted using GIS. The general absence of data on coastline change and on sand volume used for beach nourishment make interpreting coastline change data very difficult. An ideal state would be to have such data on various coastal sectors around Singapore where human intervention had been minimal. The erosion hazard and risk map produced here using GIS represents a first approximation and will be refined as more data becomes available from the field monitoring.

Through this study a maiden attempt was made to demonstrate successfully the use of GIS system to integrate and analyse field data sets to develop erosion hazard and risk map using the short and long term coastline change data. The information provided by an erosion hazard and risk map helps coastal managers to assess degrees of risk when a particular coastal site is developed. High risk areas should be left undeveloped as a buffer zone. Since risk is related to rate of erosion which may change abruptly, an erosion hazard and risk map should be reviewed on a regular basis and changes made in accordance with changing erosion rate and pattern. The resultant map also allows the stake holders and scientist for decision making on adaptation measures e.g. beach nourishment.

8. REFERENCES

References from Journals:

Wong, P.P., 1973. Beach Formation Between Breakwaters, Southeast Coast, Singapore. *Journal of Tropical Geography*, Vol 37, pg 68-73.

Wong, P.P., 2003.Where have all the beaches gone? Coastal erosion in the tropics. Singapore .*Journal of Tropical Geography*. 24(1). 111-132.

Herano, A., 2003. Mapping of ASTER Stereo data: *DEM validation and accuracy assessment,* ISPRS *Journal of Photogrammetry and remote sensing*, 57

References from Books:

Bird, E.C.F., 2000. *Coastal Geomorphology:An introduction*. John Wiley & Sons,Ltd.Chichester.

Ormsby, T., Napoleon, E., R. Burke, R., Groessl, C. and Feaster, L., (2009). *Getting to Know ArcGIS Desktop*, ESRI Press

References from Other Literature:

Singapore Tide tables, 2009. Hydrographic department, *Maritime and Port Authority of Singapore*.

Santosh, K., Raju, D.K., Chandrasekhar, J. and Teh, T.S., 2009. *Field-based data collection techniques and remote sensing for developing a high resolution digital elevation model for coastal studies*. Proc. of the *5th Int. Conf. on Asian and Pacific Coasts,* Vol. II, pp. 304-312.

Teh, T. S., Yap, H. B. and Cheng, K.H., 2005. Sense and No Sense in Measuring Coastline Change. Paper presented in *International Conference on Southeast Asia: Issues, Problems and Prospects* 12-13 September 2005.

Teh, T.S., 2000.Beach profiling monitoring in Gurney Drive, Penang. In Teh, T.S. (Ed), Islands of Malaysia: Issues and challenges. University of Malaya, Malaysia. Pp.59-79.

References from websites:

Coastal Zone, Management, Massachusetts Office, 2009. *http://www.mass.gov/czm/hazards/shoreline_change/shorelinechange.htm*

Taulealo, T. I and Bismark Crawly, T. 2002.*Planning for Coastal Hazards in Samoa, http://www.mnre.gov.ws/documents/forum/2002/9-Tuuu-Bismarck.pdf*

ESRI Link, 2009
http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=TIN%20to%20Raster%20(3D%20Analyst)

GIMS-TECNOLOGY FOR THE ENVIRONMENTAL DIAGNOSTICS

F. A. Mkrtchyan *, V. F. Krapivin

Dept. of Informatics, V.A. Kotelnikov's Institute of Radioengineering and Electronics, RAS,
Vvedensky Sq. 1, Fryazino, Moscow region, 141190 Russia- ferd47@mail.ru

Commission VI, WG VI/4

KEY WORDS: Expert System, GIMS-Technology, Environment, Mathematical Modelling, Database, Remote Sensing

ABSTRACT:

Expert system for the operative environment diagnostics (ESOED) that is here proposed realizes GIMS-technology (GIS+Model) combining the methodic and algorithms of mathematical modeling with the land and remote observations of the environment. Links between experiments, algorithms, and models of environmental processes and subsystems are developed to realize effective procedure for the operative control and diagnostics of the environment. The ESOED functions include:

-acquisition and accumulation of data by means of in-situ and remote methods and their analysis with the subsequent subject processing;

-systematic observation and evaluation of the environment;

-evaluation and synthesis of knowledge concerning the atmosphere, soil-plant cover, and water medium change;

-predetermination of the forecasting diagnostics of the environment change under anthropogenic forcing;

-analysis of the tendencies in the environmental processes when the anthropogenic scenarios are realized;

-identification of causes of ecological disturbances and danger warning.

The objective of this Report is threefold: 1) To present a working methodology for the combined use of modeling technology and microwave remote sensing measurements in the assessment of environmental processes and biospheric subsystems dynamics. 2) To illustrate this methodology with computer calculations of global change dynamics for the various scenarios.

1. INTRODUCTION

There are many parameters describing the environmental conditions on the Earth. Among them are soil moisture and moisture related parameters such as the depth of a shallow water table and contours of wetlands and marshy areas. The knowledge about these parameters and conditions is very important for agricultural needs, water management and land reclamation, for measuring and forecasting trends in regional to global hydrological regimes and for obtaining reliable information about the water conservation estimates (Alexandrov, Oikawa, 2002; Chuklantsev et al., 2003; Ferrazzoli et al., 1996; Kondratyev et al., 2002; Kondratyev et al., 2004).

In principle, the required information may be obtained by using on-site measurements and remote sensing and by getting access to a prior knowledge-based data in the GIS databases. But the problem which arises here consists of solving the following:

- what kind of instruments are to be used for conducting the so-called ground-truth and remote sensing measurements;
- what is the cost to be paid for the on-site and remote sensing information;
- what kind of balance is to be taken under consideration between the information content of on-site and remote sensing and the cost of these types of observations;
- what kind of mathematical models may be used both for the interpolation of data and the extrapolation of them in terms of time and space with the goals to reduce the frequency and thus the cost of the observations and to increase the reliability of forecasting the environmental behavior of the observed items.

These and other problems are solved by using a monitoring system based on combining the functions of environmental data acquisition, control of the data archives, data analysis and forecasting the characteristics of the most important processes in the environment. In other words, this unification forms the new information technology called the GIMS-technology. The term "GeoInformational Monitoring System (GIMS)" is used for the description of the formula: *GIMS = GIS + Model*. There are two views of the GIMS. In the first view the term "GIMS" is synonymous with "GIS". In the second view the definition of GIMS expands on the GIS. In keeping with the second view the main units of the GIMS are considered below.

The basic component of the GIMS is considered as a natural subsystem interacting through biospheric, climatic and socio-economic connections with the global *Nature-Society* system (NSS). A model is created describing this interaction and the functioning of various levels of the space-time hierarchy of the whole combination of processes in the subsystem. The model encompasses characteristic features for typical elements of the natural and anthropogenic processes and the model development is based on the existing information base. The model structure is oriented to the adaptive regime of its use.

The combination of the environmental information acquisition system, the model of the functioning of the typical geoecosystem, the computer cartography system and the means of artificial intelligence will result in creation of the geoinformation monitoring system of a typical natural element capable of solving the following tasks:

- evaluation of global change effects on the environment of the typical element of the NSS;
- evaluation of the role of environmental change occurring in the typical element of climatic and biospheric changes on the Earth and in its territories;

* Corresponding author. This is useful to know for communication with the appropriate person in cases with more than one author.

- evaluation of the environmental state of the atmosphere, hydrosphere and soil-plant formations;
- formation and renewal of information structures on ecological, climatic, demographic and economic parameters;
- operative cartography of the situation of the landscape;
- forecasting the ecological consequences of the realization of anthropogenic scenarios;
- typifying land covers, natural phenomena, populated landscapes, surface contaminations of landscapes, hydrological systems and forests;
- evaluation of population security.

2. STRUCTURE AND FUNCTIONS OF THE ESOED

Construction of the ESOED is connected with consideration of the components of the biosphere, climate and social medium characterized by the given level of spatial hierarchy. It is based on the use of GIMS - technology.

Subsystem for Planning and Analysis of the Data Acquisition Systems. This subsystem solves the task of experimental planning by analysis of the structure of the environmental data acquisition system, making use of data from satellites, flying laboratories and movable and stationary ground observation means. The laboratories are equipped with the necessary software and hardware tools to allow determination of the degree of environmental contamination, of the ecological situation, mapping of the characteristic geological formations, detection of soil subsurface centres of ecological injury, performing the all-weather land-cover typification and detection of permafrost disturbances, oil spills, forest states and pollution of bodies of water.

Subsystem for Initial Data Processing and Data Acquisition. Methods and algorithms for synchronous analysis of aero-space information and ground measurements are realized using space-time interpolation methods. Retrieval of the data and their reduction to the common time scale is performed. Model parameters are determined. Thematic classification of the data is carried out and space-time combination is performed of images in the optical, IR and microwave ranges and of trace measurements obtained from devices of various types.

Subsystem for Computer Mapping. Algorithms are realized for creation of computer maps with characteristic markings for evaluating the ecological situation. Multilevel scaling and fragmentation of the territory is envisaged. The overlaying of output maps with the information needed by the user is provided through the user interface.

Subsystem for Evaluation of the State of the Atmosphere. Models of atmospheric pollution spread due to evaporation and burning of oil products, natural gas and other outputs of industrial enterprises are suggested. The problem of evaluation of the atmosphere dust content is solved. The gas and aerosol composition of the near-earth atmospheric layer are provided and forecasting maps of their distribution over the earth's surface are created.

Subsystem for Evaluation of the State of the Soil-Plant Cover. This subsystem solves the following tasks:

The paper must be compiled in one column for the Title and Abstract and in two columns for all subsequent text. All text should be single-spaced unless otherwise stated herein. Left and right justified typing is preferred.

- typifying of the floristic background taking into account the microrelief, soil type and its salinity, humidification and degree of soil brine mineralization;
- revealing of micro- and macrorelief peculiarities and subsurface anomalies;
- determination of the structural topology of the land cover;
- indication of forests, swamps, agricultural crops and pastures.

Subsystem for Evaluation of the State of the Water Medium. A complex simulation model of the territory is developed taking into account seasonal changes of surface and river runoff, the influence of snow cover and permafrost and the regime of precipitation and evapotranspiration. A model is constructed of water quality dynamics for the hydrologic network of the territory.

Subsystem for Risk Evaluation of the Ecological Safety and the Health of the Population. Algorithms are developed for evaluation of the damage to nature, economic stability and population health depending on changes in the environment connected with natural trends of meteorological, biogeochemical, biogeocenotic, micro- biologic, radiologic and other natural processes as well as the enhancement of environmental stress of anthropogenic origin.

Subsystem for Identification of Causes of Ecological and Sanitary Disturbances. The task of revealing the sources of environmental pollution is solved. This subsystem determines the source coordinates, the magnitude and the possible time of nonplanned introductions of contaminate substances. The dynamic characteristics of the pollution sources are given. A priori unknown pollution sources are revealed and the directions of possible transborder transfer of pollutants are determined.

Subsystem for Intelligent Support. Software-mathematical algorithms are realized for providing the user with intelligent support in performing the complex analysis of objective information formed in the framework of the simulation experiment. The necessary information for the objective dialogue with the global model is provided in a convenient form for the user. The introduction of data processing corrections is also provided. The knowledge base of anthropogenic, demographic and socio-economic processes on the territory is formed.

The ESOED functions include:

- Acquisition and accumulation of data by means of in-situ and remote methods and their analysis with the subsequent subject processing.
- Systematic observation and evaluation of the environment.
- Evaluation and synthesis of knowledge concerning the atmosphere, soil-plant cover, and water medium change.
- Predetermination of the forecasting diagnostics of the environment change under anthropogenic forcing.
- Analysis of the tendencies in the environmental processes when the anthropogenic scenarios are realized.
- Identification of causes of ecological disturbances and danger warning.

The user, following the hierarchy of the ESOED menu, can realize the following operations:

to ask for data on any identifier (array) and to correct any of its fragments;

to ask for estimates of all or part of the parameters of simulation units and to correct them;

to select the sets of parameters and identifiers for a more prompt access to them;

to synthesize a symbolic schematic map of the distribution of the estimates of the environmental characteristics;

to predict the state of the environment down to a given depth or till accomplishing the a-priori formulated criterion of assessment of the state of the water environment.

Schematically it is(as illustrated by figure 1). The user, through interface, sends a permission at each step of the command dialogue, which are assessed in the unit of the query analysis, and from its response, the controlling unit realizes a chain of needed actions of the system. Via return channels of query, the resulting prediction is arranged in the needed form, which can change in each cycle of service. The final result is presented in the form of the protocol with enumerated characteristics of the water environment by objects and territories as well as in the form of schematic maps or digital information combined with the map.

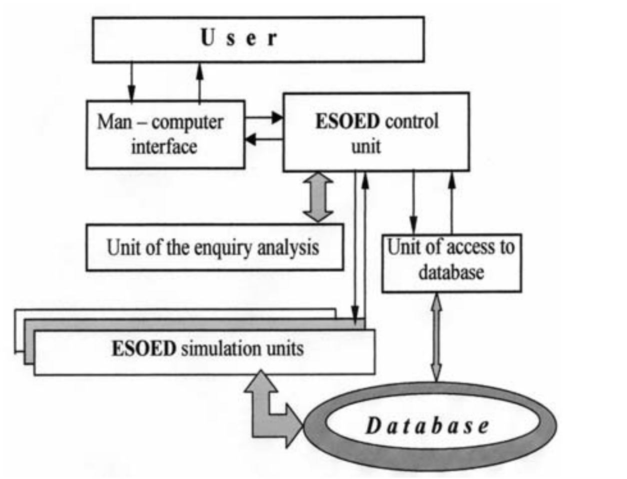


Figure 1. ESOED information units.

3. SEARCH AND DETECTION OF NATURAL DISASTERS

The ESOED allows to adjust its functions for the operative control of the environmental parameters in the regions where it is possible the natural disaster arising. It is realized by means of global model that parametrizes the dynamics of biospheric charecterisatics. Presence of the global database containing various information characteristics allows to consider and to evaluate the consequences of possible realization of the different scenarios of the NSS subsystems development. Traditional approaches to syntheses of the global models are founded on consideration of the collections of the balance equations, in which environmental parameters $\{x_i\}$ fall into the form of functions, arguments, factors and conditions of the transition between parametric descriptions of the environmental processes. As well as the other approaches are using based on the evolutionary and neuron-network algorithms. The organization of the NSS global model functioning can be presented in the manner of conceptual scheme(as illustrated by figure 2). The realization of this scheme is performed by the introduction of the geographical cell $\{\varphi_i\ ,\lambda_j\}$ with discretization steps $\Delta\varphi_i$ and $\Delta\lambda_j$ for the land surface and World Ocean by the latitude and longitude, respectively. So, all processes and NSS elements are considered as uniform and are parametrized by point models within the pixel $\Omega_{ij} =\{(\varphi,\lambda)\colon \varphi_i\leq\varphi\leq\varphi_i+\Delta\varphi_i,\ \lambda_j\leq\lambda\leq\lambda_j+\Delta\lambda_j\}$. The choice of the the pixels size is defined by set of the conditions, definied by spatial resolution of the given satellite measurements and by presence of necessary global database. In the case of water surface, the water body of pixel Ω_{ij} is divided by depth z on the layers, i.e. three-dimensional volumes $\Omega_{ijk}=\{(\varphi,\lambda,z)\colon (\varphi,\lambda)\in\Omega_{ij},\ z_k\leq z\leq z_k+\Delta z_k\}$ are formed. All elements of Ω_{ijk} are considered as uniform. Finally, atmosphere above the pixel Ω_{ij} are descretized by the height in accordance with the atmospheric pressure levels, or on typical layers by height.

It is clear that creation of global model is possible only with attraction of the knowledges and data on given multidisciplinary level. Among ensemble of the global models the most making is a model, described in (Kondratyev et al., 2002). In (Kondratyev et al., 2002; Kondratyev et al., 2004) adaptive procedure for global model fitting in the geoinformation monitoring system is offered.

Approach of the moment of the natural catastrophe arising is characterized by hit of the vector $\{x_i\}$ in a certain claster of multidimensional space X_c . In other words, going from purely verbal discourses to quantitative determination of this process, we shall enter the generalised feature $I(t)$ of the natural catastrophe and shall identify it with graduated scale Ξ. Satisfactory model that transforms the verbal portrait of a natural disaster into notions and indicator subject to a formalized description and transformation is described in (Kondratyev et al., 2002). An introduction of the characteristic I enables one to propose the following scheme of monitoring and predicting natural catastrophes. Three are three levels in the system: recorder, decision maker, and searcher, whose units have the following function:

-regular control of the environmental elements to accumulate data about their state in the regime permetted by the applied technical means;

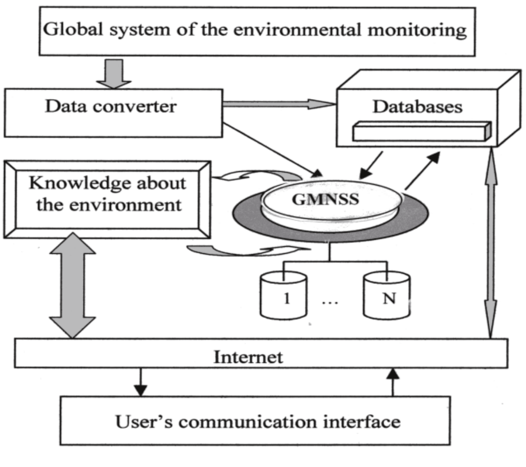
-recording suspecious elements of the environment for which the value of the indicator I corresponds to the interval of a natural anomaly danger of a given type;

-formation of the dynamic series $\{I(t)\}$ for a suspecious element to make a statistical decision about its noise or signal character and in the latter case the rest of the suspecious element by criteria of the next level of accuracy (getting of the $\{x_i\}$ vector into the cluster, etc);

-making the final decision about the approaching moment of a natural catastrophe occurence with the transmission of information to the respective environmental control services;

-iterative procedure to locate an anomaly.

Efficiency of such procedure depends on the parameters of the measuring technical facilities and algorithms for the data processing. The important role here plays the environmental model, used parallel with formation and statistical test of the row $\{\ I\ (t)\}$ and adapted to mode of the monitoring.


Global system of the environmental monitoring
Data converter
Databases
Knowledge about the environment
GMNSS
1
...
N
Internet
User's communication interface

Fig

Figure 2. Conceptual block-diagram of geoinformation monitoring and use of the global model of *Nature-Society* system (GMNSS).

As seen from the introduced criterion of an approaching natural catastrophe, the form and behaviour of $I(t)$ are special for each type of the processes in the environment. One of the complicated problems consists in determination of these forms and their classification. For instance, such frequent dangerous natural events as landslips and mudflows have characteristic features, such as preliminary changing relief and landscape, which are successfully recorded from satellites in the optical range, and together with data of aerial photography and surface measurements of relief slopes, expositio of slopes and the state of the hydro-system make it possible to predict them several days beforehand. However, restricted capabilities of the optical range under conditions of clouds or vegetation cover should be broadened by introducing the systems of remote sensing in the microwave region of the electromagnetic spectrum. Then, in addition to the indicators of landslips and mudflows, one can add such information parameters as soil moisture and biomass, since a soil moisture increase leads to landslips, and an enhancement of biomass testifies to the growth of the restraining role of vegetation cover with respect to the dislocation of mountain rocks. Particularly this it is important when supervision is snowed-stone or simply snow avalanches. Making the catalogue of such indicators for all possible natural catastrophes and contributing their in knowledgebase of monitoring system is a necessary stage of increasing of its efficiency.

4. CONCLUSION

Making the catalogue of above indicators for all possible natural catastrophes and contributing their in knowledgebase of monitoring system is a necessary stage of increasing of its efficiency.

REFERENCES

Alexandrov, G., Oikawa, T.,2002. TsuBiMo: a biosphere model of the CO_2 – fertilization effect // Climate Res., 19: 265-270.

Chuklantsev, A.A., Shutko, A.M., and Golovachev S.P., 2003.*Attenuation of electromagnetic waves by vegetation canopies in the 100-1000 MHz frequency band*. ISTC/IRE Technical Report, #2059-1, 59 pp.

Ferrazzoli, P. and Guerriero, L.,1996. Passive microwave remote sensing of forests: a modelinvestigation // *IEEE Trans. on Geosci. And Remote Sensing,* **34**(2): 433-443.

Kondratyev, K.YA. and Krapivin, V.F.(2002). Monitoring and prediction of natural disasters. *NUOVO CIMENTO*, **27**(6), 657-671.

Kondratyev, K.YA., Krapivin, V.F., Savinykh, V.P., and Varotsos C.A.(2004). *Global Ecodynamics: A Multidimensional Analysis.* Springer/PRAXIS, Chichester U.K., 658 pp.

A DECISION SUPPORT FRAMEWORK FOR THE RISK ASSESSMENT OF COASTAL EROSION IN THE YANGTZE DELTA

Li Xing[a,*], Zhou Yunxuan[a], Shen Fang[a], Kuang Runyuan[a], Wu Wen[a], Zheng Zongsheng[b]

[a]State Key Laboratory of Estuarine and Coastal Research, East China Normal University, 3663 Zhongshan North Road, Shanghai 200062 - li_xing99@yahoo.com.cn, (zhouyx, fshen)@sklec.ecnu.edu.cn, rykuang@163.com, wen1722003@tom.com
[b]College of Information, Shanghai Ocean University, 999 Huchenghuan Road, Shanghai 201306, China - zszheng@shou.edu.cn

KEY WORDS: Coastal erosion, Risk assessment, GIS, Decision support framework, the Yangtze Delta

ABSTRACT:

Coastal erosion is an issue of widespread concern. As one of the most important economic regions in China, the coast of the Yangtze Delta has been showing a trend towards erosion with global warming and the increasing human activities in the catchment and its estuary. Currently, most published studies about coastal erosion in the area focused on the causes and types of erosion. This paper presents a Decision Support Framework (DSF) for the risk assessment of coastal erosion in consideration of the potential management problems and challenges for economic development in the coastal zone of the Yangtze Delta. The framework consists of four major components: integrated database, GIS-based risk assessment models, scenarios generator and visualization toolkit. Especially, we developed a GIS-based risk assessment model for the muddy coasts of Yangtze Delta. An Analytic Hierarchy Process (AHP) method, which is instrumental in combining computer intelligence and experts' knowledge, was used to weight the variables of the model. The assessment results show the validity of the approach. Accordingly, the DSF will make the specialized data and information more accessible to managers, and has an extensive capability to facilitate communication and synergetic work between humans and computers. In this way, it is expected to make manager make more scientific decision.

1 INTRODUCTION

Coastal erosion is an issue of widespread concern. It is estimated that at least 70% of the sandy beaches in the world is retreating at a rate of 0.5-1.0 m/year (Bird, 1985). Moreover, intensified human activities and accelerated sea level rise will aggravate the coastal erosion in the future century (Zhang et al., 2004). The coasts, especially of the deltas and megadeltas, which are recognized as highly susceptible to human and natural impacts, will be exposed to increasing risks (Nicholls et al., 2007). In China, coastal erosion is also ubiquitous in most coastal area (Sheng and Zhu, 2002; Xia et al., 1993). Zuo et al. (2009) reported that most of the coasts between the mouth of the Luan River and the city of Qinhuangdao (northeast Bohai Sea, China) have been suffering from extensive erosion with an average rate of 3.7 m/year during 1986-2000. The abandoned Huanghe delta has been subjected to dramatic erosion since 1855 when the Huanghe River shifted its channel (Zhang et al., 1998). As far as the Yangtze Delta is concerned, the huge amount of sediment load from the Yangtze River contributed to the successive accretion of the Delta in the past. However, contemporarily global warming and the construction of large hydraulic engineering works in the catchment and its estuary significantly decelerate the trend, and the delta degradation has been observed since 2003 (Yang et al., 2007). Considering the combined influence of the planning dams and the South-to-North Water Diversion Project, the anticipated coastal erosion problem will be even worse.

Since the Yangtze Delta is one of the most important economic regions in China, a long-term coastal spatial planning will be an inevitable issue for the local governments. While, most coasts in the Yangtze Delta are typically muddy that are characterized by extremely complex land-sea interaction. In view of the complexity and uncertainty of human and natural dimension under the circumstances of global change, the need for integrating human and computer intelligence necessitates a Decision Support Framework (DSF) for the risk assessment of coastal erosion. Although decision support has been involved in Integrated Coastal Zone Management (ICZM) early in 1990s, it has a very limited application in coastal management (Van Kouwen et al., 2008). And most of the existing decision support related to coastal erosion was potentially underlain by the complete physical, socioeconomic datasets and the deterministic solutions. In fact, in many developing countries it is usually possible that some key parameters affecting model output are unavailable (Szlafsztein and Sterr, 2007), and the relevant research is still inadequate. In the case of Yangtze Delta, most current published studies about coastal erosion were focused on the causes and types of erosion (Cai et al., 2009; Chen et al., 1988; Gao and Wang, 2008; Hori et al., 2002; Milliman et al., 1985; Yang et al., 2001, 2006). This paper presents a decision support framework (DSF) for the risk assessment of coastal erosion in consideration of the potential management problems and challenges for economic development in the coastal zone of the Yangtze Delta.

Our primary goal is to develop a framework to provide the theoretical and methodological basis for the future practical application, further to bridge the gap between scientists and managers to provide the intelligence basis for the scientific coastal management.

2 WHY NOT DECISION SUPPORT SYSTEM (DSS)?

The idea of DSS originated in the mid-1960s (Power, 2008), and has been applied extensively in many fields. In comparison, its application for coastal management is very limited (Wiggins, 2004). The causes have been discussed by several literatures (*cf.* Jones et al., 2002; Uran and Janssen, 2003; Van Kouwen et al., 2008; Westmacott, 2001). We avoid the term DSS because of several reasons as follows:

- Nowadays, the development of computer technology has made that few computer systems of any significance haven't some functions of supporting decision making in some forms to some extent (Alter, 2004). So DSS maybe a still popular but obvious antiquated label.

[*] Corresponding author.

- The technological and administrative background is still lack for the Yangtze Delta to develop DSS for coastal management (Lau, 2005; Shi et al., 2001). China is a developing country, and the Yangtze Delta is one of its economic centers. The local authorities mainly draw their attention on the economic development. Scientific coastal management implies to give up some economic considerations to some extent. And the specific technological conditions need to be improved further.

- Coastal erosion is a complex natural phenomenon, especially to muddy coast. Although the information technology is quite advanced, decisions still have to mainly depend on experienced specialists and managers. In the first instance, we need to improve the computer system by incorporating more sophisticated models. On the other hand, it is a gradual process for human beings to understand nature. Although a great number of researches have been conducted in the Yangtze Delta over the past decades, our knowledge is very limited as to the complexity of coastal erosion. So a tool or platform is needed to facilitate the understanding the natural and human-induced coastal behaviors. DSS, as a technical artifact, can't explicitly express these meanings.

Instead, we use "decision support framework" which has three characteristics: (1) to provide reference data to promote the effectiveness of decision making; (2) to provide analysis models to promote the efficiency of decision making; (3) to bridge the gap between scientists and policy-makers by providing the scenario and visualization tools, and eventually facilitate "coevolution" of human and computer.

3 DECISION SUPPORT FRAMEWORK CHALLENGES IN THE YANGTZE DELTA

The Yangtze Delta is formed mainly by riverine sediments from the Yangtze River that depends on the river discharge and ocean processes (Chen et al., 1959) (Figure 1). The coasts are mostly muddy, along which the muddy tidal flat are extensively distributed with varying width. The coastal development is influenced by many factors, including the history of coastal evolution, ocean dynamic, terrigenous materials, human activities, socio-economic conditions, etc., so the coastal environment is very complex.

The Yangtze Delta is a tide-dominated coastal environment (Saito et al., 2001). The mean annual water discharge of the Yangtze River is 9.24×10^{11} m^3 that transports 4.80×10^8 tons of sediment to the sea (Yang et al., 2006). The water discharge has an evident yearly and seasonal variation (Chen et al., 2001). The annual tidal prism into the Yangtze estuary has a total volume of 8.40×10^{12} m^3, which is an order of magnitude higher than the water discharge (Chen et al., 1988). Along the coasts, the semidiurnal tide is predominant with different tidal ranges. The average tidal range is 1.5-1.7 m around the abandoned Yellow River mouth and along the southern part of the North Jiangsu coast (Zhou et al., 1994). And the mean tidal range is 2.7 m near the Yangtze River mouth, and the maximum tidal range approaches 4.6 m (Shen et al., 1988).

The Yangtze Delta is one of the most important industrial and agricultural areas in China where economic development is the first priority of local governments. With 2% of the area and about 10% of the population of China, the region contributed to 26% of national GDP in the first half of 2008 (http://www.ocn.com.cn/reports/2006079changsanjiao.htm, in Chinese). Along with the progress of urbanization and industrialization, more tidal flats will be reclaimed in Yangtze Delta; as a result, vulnerable coast environments will be further disturbed. Meanwhile, the delta is one of the most serious subsidence areas in China (Han et al., 2008). During 2001-2006, the mean subsidence rate is 12.7 mm/year in Shanghai (Gong, 2008). Moreover, the fluvial sediment to the coastal ocean will continually reduce as the large hydraulic engineering works are brought on line in its catchment. Additionally, the sea level is rising at a significant rate under the influence of global warming (Shi et al., 2001). It is estimated that the relative sea level rise from 1990 to 2050 will be 50 cm in Shanghai Municipality, 20-30 cm in the coastal plain of North Jiangsu and the northern bank of Hangzhou Bay, and the 100-year storm surge level will increase by 38-44 cm when the sea level reaches 50 cm (Shi et al., 2000). These will undoubtedly increase the risk of coastal erosion in Yangtze Delta.

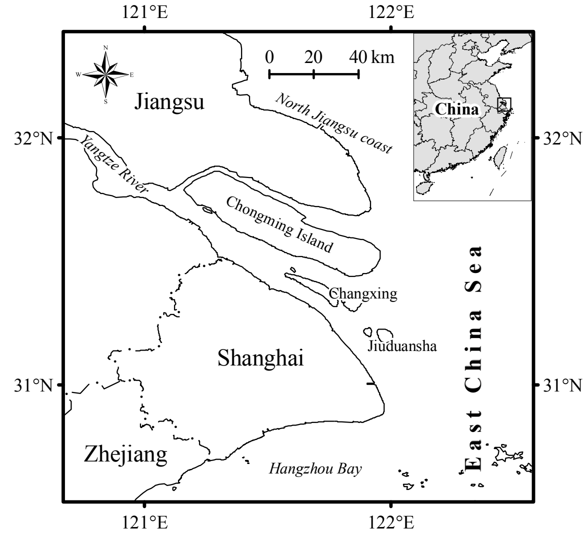


Figure 1: Location of the Yangtze Delta

4 THE DECISION SUPPORT FRAMEWORK

The DSF consists of four major components: integrated database, GIS-based risk assessment models, scenarios generator and visualization toolkit. At first, it is necessary for us to provide definitions of three concepts in order to the further description of the DSF and discussion.

a) *Visualization*: a set of computer-generated static or animated, 2D or 3D character, graphics and image to depict the properties and process of physical objects in some certain aspects, with the intention of visually communicating some information to specified audiences.

b) *Scenario*: an expression of a subset of the complete parameters that can completely depict a complex system. Scenarios are neither predictions nor forecasts, but rather attempt to picture the situation under specified conditions, by this way, to intuitively grasp the essence of events or phenomena.

c) *Decision support framework*: a conceptual structure comprising personnel and computer programs, in which manager or decision-maker and computer programs interact to make decisions and "coevolve".

The three concepts are defined mainly based on the consideration of highly complex, dynamic environment, where scientists fail to provide managers or decision-makers with effective information what they need most (Tribbia and Moser, 2008). Therefore, visualization and scenario act as tools to help those without technical expertise understand the complex physical objects and update their own knowledge. The framework contributes to combine the strengths of both human intelligence and computer program, and promote the evolution of either, and by which the framework will improve with time. In following subsections we will discuss in detail four components of DSF.

The integrated database provides a means of storage, management and retrieval all relevant data, which support the risk assessment of coastal erosion in the Yangtze Delta. It contains direct indicators data (can be directly input to model as a parameter), input data (for calculations of direct indicators) and auxiliary data (to aid visualization and scenario). These data are represented respectively in raster and vector format, and the involved data types mainly have, administrative boundary, remote sensing image, geomorphology and geology, elevation, bathymetry, shoreline, sediment transport, riverine discharge, tidal regime, wave climate, sea level rise, subsidence, extreme historical events, Land Cover/Land Use (LCLU), infrastructure, ecological hotspot, population density, etc.

GIS-based Risk Assessment Model is the core component of the DSF. It mainly is used to quantify the risk level of coastal erosion for the muddy coasts of the Yangtze Delta, and further serve as a general framework for the case of muddy coasts where possible. It will work mainly by the following steps: (i) to define homogeneous units; (ii) to identify the vulnerability indicators and to evaluate vulnerability; (iii) to identify the impact indicators and to evaluate hazard; (iv) to evaluate risk. All the calculations will be done in GIS environment where ones are allowed to take advantage of its enhanced spatial analysis capabilities.

The scenario generator is used to generate the impact of coastal erosion under varying conditions which maybe existent, hypothetic, simulated or predicted. In chief, it assists users to understand the coastal environment, define the vulnerability and impact indicators, and determine weights. Especially, it can be instrumental to incorporate human knowledge and insight into decision making processes. Before a given scenario is generated, several other steps are needed. At first, we need to plan the scenarios, such as worst case, best case or current case, etc., in terms of the ideas in our mind. Then we input the required data or parameters into relevant models to generate scenarios. Finally, we test different scenarios, and from which some knowledge is extracted. Again, some prediction models about shoreline change and sea level rise will be involved in the procedure. The visualization toolkit in the next subsection will support the entire procedure.

Visualization toolkit will provide some basic visual forms. Specially, there are some aspects of functions that are worthwhile to emphasize in the toolkit. Firstly, we endeavor to organically integrate specialized data and information on coastal erosion into GIS. In fact, visualization technique and specialized computational model have been developing in separation. And recent research indicates that GIS has been one of the most recognizable tools for managers (Tribbia and Moser, 2008). Hence, the integrated representation is feasible and necessary. It will provide insight into the hidden pattern and trend for end users. The second is to visualize the evolution process and trend of coastal elements by predictive simulation models. The data about coastal elements, such as shoreline, tidal vegetation, reclamation, nearshore sediment movement, etc., were captured at discrete time intervals without intermediate details. The function allows us to understand their complete evolution process to some extent. And the third is to link coastal landscape and dynamic elements with Digital Elevation Model (DEM) by extensive texture and 3D model. Some improved interpolation algorithms, like fractal Brownian motion (fBm), wavelet, etc., are used to enhance the accuracy and visual effects of DEM. In addition, it is important to display the uncertainty of data and results, for instance, historical shorelines and predicted shorelines.

5 RISK ASSESSMENT OF COASTAL EROSION IN THE YANGTZE DELTA

In this section, we will give the details of risk assessment for our study area that based on the DSF, especially GIS-based risk assessment model aforementioned. The study area comprises the whole coasts of Shanghai and the partial coasts of Taicang, Haimen and Qidong in Jiangsu province.

Risk assessment has varied definitions and is widely used in various fields (Del Río and Gracia, 2009; Eurosion, 2004; UN/ISDR, 2004). Usually, a whole risk assessment encompasses vulnerability assessment and hazard assessment (UN/ISDR, 2004). The former, which is often linked to the physical dimensions to determine the area, probability and/or intensity of occurrence of a hazard under specified conditions; the latter determines the damage potential of a hazard, it is often related to the socio-economic and human dimensions. Respectively, they depend on a set of vulnerability indicators and impact indicators. The resulting risk index (RI), as a single risk measurement, can be derived by combining the two aspects.

According to the steps described above, some primary indicators, which can represent the essential physical and socio-economic characteristics of the specified coastal system, need to be first selected to segment the shorelines into the relatively homogeneous units. These primary indicators include *administrative boundary*, *population density*, *geological types*, and *coastal characteristics*. The full segmentation procedure requires an intersection operation of four different segmentation results related to these four indicators, and finally products 40 segments for the whole shorelines of the study area.

The selections of vulnerability and impact indicators were followed by some premises, including representativeness, independence, availability and easy to use. Meanwhile, according the specialists' suggestions and some published literatures (Feng and Xia, 2003; Li and Yang, 2001; Wang et al., 1999), the 10 vulnerability indicators were identified as:

- coastal elevation
- coastal slope
- average annual deposit volume
- shoreline change rates
- tidal range
- significant wave height
- relative sea level rise
- intertidal width
- intertidal vegetation type
- intertidal vegetation zone width;

and the 3 impact indicators were identified as:

- population density
- land use type
- ecological hotspots.

Then both vulnerability and impact indicators were quantified or classified as numerable variables. Based on field surveys, the most recent seawall was interpreted from the Landsat TM images of April 25, 2008, and a polyline feature class, as assessment baseline, was produced in ArcGIS. Most of variables were calculated in a buffer zone of certain width, and these values were corresponding to each homogeneous unit. The variable of "coastal elevation" represents the percentage of the total area with elevation less than 2.4m above the mean sea level in a 5km wide buffer zone landwards from the assessment baseline. The "coastal slope" was calculated in a 2km wide buffer zone that 1km landwards and seawards, respectively, from the assessment baseline. The widths of buffer zones for "average annual deposit volume", "significant wave height" and "relative sea level rise" are, respectively, 2km, 5km and 5km seawards from the assessment baseline. "Intertidal width" refers to average intertidal width from assessment baseline to 2m isobaths. And the main land use type, as the variable "land use type", was identified from Landsat TM images in 2km buffer zone landwards from the assessment baseline. Vegetation data, which were interpreted from TM images based on field surveys, represents the vegetation outside the seawalls.

Subaqueous topographic data and elevation data were obtained from digitized nautical charts and 1:50,000 scale topographic map, respectively. Landsat TM images of 1990 to 2008 were used to extract shorelines. Global mean sea level data with 1/3 of a degree resolution provided by AVISO is available at `http://www.aviso.oceanobs.com/en/news/ocean-indicators/mean-sea-level/`. Significant wave height was simulated by SWAN wave model based on 11-year (1995-2005) monthly climatological data set from NOAA/NESDIS/National Climatic Data Center website (`http://www.ncdc.noaa.gov/oa/rsad/seawinds.html`). Population density data were obtained from the Gridded Population of the World version 3 (GPWv3) data set with 2.5 arc-minutes resolution (CIESIN/FAO/CIAT, 2005).

Then, depending on the nature of each of these variables, they were assigned ranks ranging from 1 to 5, with 1 representing minimum vulnerability/hazard and 5 representing maximum vulnerability/hazard.

Before evaluating the vulnerability and hazard, these variables must be weighted based on their relative importance in determining the coastal erosion vulnerability and hazard. The aims have two aspects: to avoid the underestimation or overestimation of the contribution of any variable; and to incorporate the specialists' knowledge into the DSF. To assign the objective and reliable weights for the variables, an Analytic Hierarchy Process (AHP), has been extensively used in almost all the applications related to Multiple Criteria Decision Making (MCDM) in the last 20 years (Ho, 2008), was employed for the variables weightings. According to the fundamental scale of absolute numbers (Saaty, 2008), we derived the priority scales for each variable based on the judgments of experts, and to construct the pairwise comparison matrix. For the 10 vulnerability indicators, the derived weights were, respectively, 0.182, 0.182, 0.264, 0.104, 0.023, 0.023, 0.016, 0.119, 0.035 and 0.051 with a consistency ratio of 0.037; and for the 3 impact indicators, they were, respectively, 0.751, 0.178 and 0.070 with a consistency ratio of 0.025.

The Vulnerability Index (VI) was built by calculating the weighted sum of the 10 variables (Eq. 1):

$$\mathrm{VI}_j = \sum_{i=1}^{n} w_i f_{ij}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, m \qquad (1)$$

where, VI_j is the VI for the jth homogeneous unit; w_i is the weight for the ith vulnerability indicator; f_{ij} is the scale for the ith vulnerability indicator in the jth homogeneous unit; m and n are the number of the homogeneous units and the vulnerability indicators, respectively. The calculation of Hazard Index (HI) was analogous to Eq. 1. And furthermore, the Risk Index (RI) was obtained by a simple weighted average of VI and HI. The weights for VI and HI are, respectively, 0.75 and 0.25. Finally, the RI was normalized as a percentage of 0 to 100 (Figure 2).

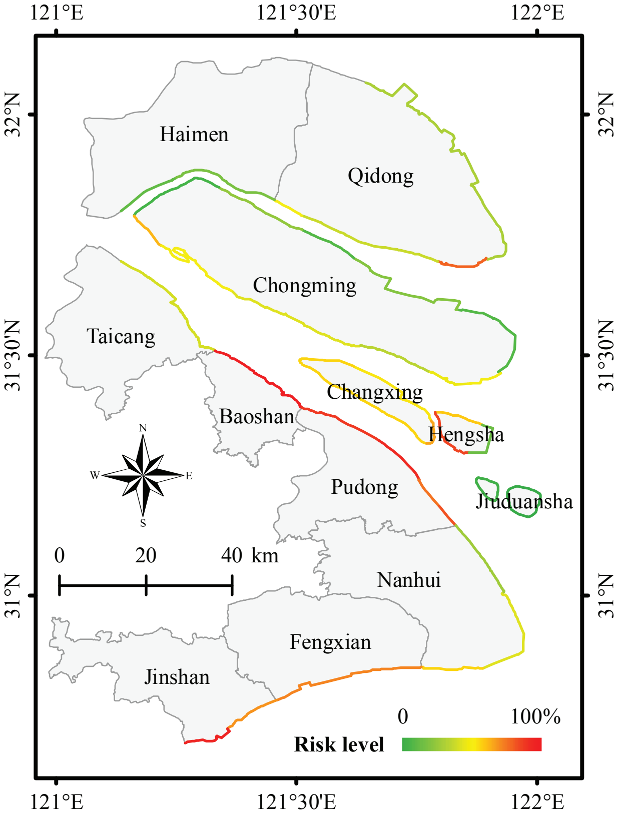


Figure 2: Distribution of risk level for the study area

The results show that the mean of RIs is 45.5%. The values over 50% of RI appear in 17 shoreline segments, amounts to over 35% of the total length of all shorelines in study area. The shoreline segments with RIs being over 60% are mainly distributed in the four districts of Baoshan, Pudong, Fengxian and Jinshan. Moderate values of RI appear in the southern part of Qidong and Chongming Island, most parts of Changxing Island, and the southern part of Nanhui district. And the low values appear in Chongming Dongtan, Jiuduansha, and the North Branch.

6 DISCUSSION AND CONCLUSIONS

As Tribbia and Moser (2008) reminded us, there exists a disconnection between science and practice in coastal management. Currently, it is scientists and researchers in the institute who are engaged in the study of coastal erosion, while the managers who need to prepare the impact of coastal erosion. In order to assure a scientific decision making, managers should adequately understand the information about coastal erosion. There is urgent need for a bridge between scientists and managers. The proposed DSF partially meets the requirement. And it has some distinctive characteristics. First of all, with GIS and the linear referencing database system, the framework will be scalable that allows the incorporation of a broad range of data and models when they are available. This is especial suitable to the developing coastal area.

Meanwhile, the scenario generator and visualization toolkit make it possible that the communication and synergetic work between humans and computers. But to implement the framework some challenges still exist.

For the Yangtze Delta, data remains one of the biggest challenges. Although a great deal of research has been focused on estuarine and coastal environment for the past decades, the data have been sporadically holding by individuals and communities. A well-organized data collection and sharing mechanism is strongly suggested. And the technique of capturing data is also problematic in highly dynamic environments (Wright, 2009). Researchers must rely on the appropriate sampling strategy to guarantee that the inherent characteristic of dynamic elements can be faithfully preserved. Alternatively, some parameters can be retrieved from remote sensing observations.

Additionally, some models are unavoidably involved in the DSF presented in this paper. Especially, the results from deterministic models should be carefully considered because, as claimed by Purvis et al. (2008), this kind of models mask the uncertainty of the complex systems and the expert knowledge is excluded from them. Dawson et al. (2009) also contested that probabilistic methods should be adopted to predict coastal erosion. Indeed for a given model, the validity of the result depends on many factors (Szlafsztein and Sterr, 2007). It is the reason why the scenario generator and the visualization toolkit are involved in the DSF in this paper. The evolution and involvement of human knowledge are indispensable in the decision making processes of complex systems.

Under the influence of global warming, there may be more urgent need to carry out the risk assessment of coastal erosion for the Yangtze Delta than other coastal areas in China. This paper presented a decision support framework (DSF) for the risk assessment of coastal erosion in Yangtze Delta. It is worthwhile to mention that we developed a GIS-based risk assessment model for the muddy coasts of Yangtze Delta. And the AHP method, which is instrumental in combining computer intelligence and experts' knowledge, was used to weight variables. The assessment results show the validity of the approach. But as described above, there are still some challenges for the case of Yangtze Delta due to the well-known reasons for developing areas. So there is still some way to go before the proposed DSF can be put into practice. We hope that our proposal can help local authorities understand the potential of coastal erosion risk and take precautions in advance. Finally, the proposed DSF, especially including the vulnerability indicators and the impact indicators, is defined for the Yangtze Delta in this paper. But, in view of its flexibility, it also available for a wide range of muddy coastal area.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Programme Strategic Scientific Alliances between China and the Netherlands (2008D FB90240), the State Oceanic Administration of China (200705020), the state key laboratory special fund from the Ministry of Science and Technology of China, the Doctoral Fund of Shanghai Ocean University. We also sincerely thank Dr. Ge Jianzhong for his help in simulating the significant wave height. A special thank is due to the anonymous reviewers for their valuable comments and suggestions for the improvement of the manuscript.

References

Alter, S., 2004. A work system view of DSS in its fourth decade. Decision Support Systems 38(3), pp. 319–327.

Bird, E. C. F., 1985. Coastline changes. Wiley & Sons, New York.

Cai, F., Su, X., Liu, J., Li, B. and Lei, G., 2009. Coastal erosion in China under the condition of global climate change and measures for its prevention. Progress in Natural Science 19(4), pp. 415–426.

Chen, J. Y., Shen, H. T. and Yun, C. X., 1988. Process of dynamics and geomorphology of the Changjiang Estuary (in Chinese). Shanghai Science and Technology Press, Shanghai.

Chen, J. Y., Yu, Z. Y. and Yun, C. X., 1959. Development of the geomorphology of the Changjiang delta. Acta Ecologica Sinica (in Chinese) 25(3), pp. 201–220.

Chen, X. Q., Zong, Y. Q., Zhang, E. F., Xu, E. G. and Li, S. J., 2001. Human impacts on the Changjiang (Yangtze) River basin, China, with special reference to the impacts on the dry season water discharges into the sea. Geomorphology 41(2-3), pp. 111–123.

CIESIN/FAO/CIAT, 2005. Gridded Population of the World: Future Estimates (GPWFE)). Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw.

Dawson, R., Dickson, M., Nicholls, R., Hall, J., Walkden, M., Stansby, P., Mokrech, M., Richards, J., Zhou, J., Milligan, J., Jordan, A., Pearson, S., Rees, J., Bates, P., Koukoulas, S. and Watkinson, A., 2009. Integrated analysis of risks of coastal flooding and cliff erosion under scenarios of long term change. Climatic Change 95(1), pp. 249–288.

Del Río, L. and Gracia, F. J., 2009. Erosion risk assessment of active coastal cliffs in temperate environments. Geomorphology 112(1-2), pp. 82–95.

Eurosion, 2004. Living with coastal erosion in europe: Sediment and space for sustainability. guidance document for quick hazard assessment of coastal erosion and associated flooding. Technical report, Service contract B4-3301/2001/329175/MAR/B3 “Coastal erosion - Evaluation of the need for action” Directorate General Environment, European Commission.

Feng, A. P. and Xia, D. X., 2003. Grading of coastal erosion disaster situation. Coastal Engineering (in Chinese) 22(2), pp. 60–66.

Gao, S. and Wang, Y. P., 2008. Changes in material fluxes from the Changjiang river and their implications on the adjoining continental shelf ecosystem. Continental Shelf Research 28(12), pp. 1490–1500.

Gong, S. L., 2008. The study on acting factors and systemic control of land subsidence in Shanghai (in Chinese). Doctoral, East China Normal University.

Han, Z. S., Jayakumar, R., Liu, K., Wang, H. and Chai, R., 2008. Review on transboundary aquifers in People's Republic of China with case study of Heilongjiang-Amur River Basin. Environmental Geology 54(7), pp. 1411–1422.

Ho, W., 2008. Integrated analytic hierarchy process and its applications - a literature review. European Journal of Operational Research 186(1), pp. 211–228.

Hori, K., Saito, Y., Zhao, Q. H. and Wang, P. X., 2002. Architecture and evolution of the tide-dominated Changjiang (Yangtze) River delta, China. Sedimentary Geology 146(3-4), pp. 249–264.

Jones, P. D., Tyler, A. O. and Wither, A. W., 2002. Decision-support systems: Do they have a future in estuarine management? Estuarine Coastal and Shelf Science 55(6), pp. 993–1008.

Lau, M., 2005. Integrated coastal zone management in the People's Republic of China–An assessment of structural impacts on decision-making processes. Ocean & Coastal Management 48(2), pp. 115–159.

Li, H. P. and Yang, G. S., 2001. Determination of classification and risk in the coastal zone of Yangtze Delta and North Jiangsu. Journal of Natural Disasters (in Chinese) 10(4), pp. 20–25.

Milliman, J. D., Huang-ting, S., Zuo-sheng, Y. and H. Mead, R., 1985. Transport and deposition of river sediment in the Changjiang estuary and adjacent continental shelf. Continental Shelf Research 4(1-2), pp. 37–45.

Nicholls, R., Wong, P., Burkett, V., Codignotto, J., Hay, J., McLean, R., Ragoonaden, S. and Woodroffe, C., 2007. Coastal systems and low-lying areas. In: Climate change 2007: impacts, adaptation and vulnerability, Cambridge University Press, Cambridge, UK, pp. 315–356.

Power, D. J., 2008. Decision support systems: A historical overview. In: F. Burstein and C. W. Holsapple (eds), Handbook on Decision Support Systems 1, Springer, pp. 121–140.

Purvis, M. J., Bates, P. D. and Hayes, C. M., 2008. A probabilistic methodology to estimate future coastal flood risk due to sea level rise. Coastal Engineering 55(12), pp. 1062–1073.

Saaty, T. L., 2008. Decision making with the analytic hierarchy process. International Journal of Services Sciences 1(1), pp. 83–98.

Saito, Y., Yang, Z. S. and Hori, K., 2001. The Huanghe (Yellow River) and Changjiang (Yangtze River) deltas: a review on their characteristics, evolution and sediment discharge during the Holocene. Geomorphology 41(2-3), pp. 219–231.

Shen, H., Guo, C., Zhu, H., Xu, H., Yun, C. and Chen, B., 1988. A discussion on the change and origin of turbidity maximum in the Changjiang Estuary. In: J. Chen, H. Shen and C. Yu (eds), Process of Dynamics and Geomorphology of the Changjiang Estuary (in Chinese), Shanghai Scientific and Technical Publishers, Shanghai, pp. 216–228.

Sheng, J. F. and Zhu, D. K., 2002. Discussion about coastline erosion and management. Marine Science Bulletin (in Chinese) 21(4), pp. 50–57.

Shi, C., Hutchinson, S. M., Yu, L. and Xu, S., 2001. Towards a sustainable coast: an integrated coastal zone management framework for shanghai, People's Republic of China. Ocean & Coastal Management 44(5-6), pp. 411–427.

Shi, Y. F., Zhu, J. W., Xie, Z. R., Ji, Z. X., Jiang, Z. X. and Yang, G. S., 2000. Prediction and prevention of the impacts of sea level rise on the Yangtze River Delta and its adjacent areas. Science in China Series D-Earth Sciences 43(4), pp. 412–422.

Szlafsztein, C. and Sterr, H., 2007. A gis-based vulnerability assessment of coastal natural hazards, state of para, brazil. Journal of Coastal Conservation 11(1), pp. 53–66.

Tribbia, J. and Moser, S. C., 2008. More than information: what coastal managers need to plan for climate change. Environmental Science & Policy 11(4), pp. 315–328.

UN/ISDR, 2004. Living with risk: A global review of disaster reduction initiatives. Technical report, United Nations Inter-Agency Secretariat of the International Strategy for Disaster Reduction, Geneva, Switzerland.

Uran, O. and Janssen, R., 2003. Why are spatial decision support systems not used? some experiences from the netherlands. Computers, Environment and Urban Systems 27(5), pp. 511–26.

Van Kouwen, F., Dieperink, C., Schot, P. and Wassen, M., 2008. Applicability of decision support systems for integrated coastal zone management. Coastal Management 36(1), pp. 19–34.

Wang, W. H., Wu, S. Y. and Chen, X. Y., 1999. Research on the assessment method of the coastal erosion disaster. Journal of Natural Disasters (in Chinese) 8(1), pp. 71–77.

Westmacott, S., 2001. Developing decision support systems for integrated coastal management in the tropics: Is the ICM decision-making environment too complex for the development of a useable and useful DSS? Journal of Environmental Management 62(1), pp. 55–74.

Wiggins, S., 2004. Coastal decision support systems in the UK. CoastNet Bulletin 8(3), pp. 18.

Wright, D. J., 2009. Spatial data infrastructures for coastal environments. In: X. Yang (ed.), Remote Sensing and Geospatial Technologies for Coastal Ecosystem Assessment and Management, Springer, pp. 91–112.

Xia, D. X., Wang, W. H., Wu, G. Q., Cui, J. R. and Li, F. L., 1993. Coastal erosion in China. Acta Geographica Sinica (in Chinese) 48(5), pp. 468–476.

Yang, S. L., Ding, P. X. and Chen, S. L., 2001. Changes in progradation rate of the tidal flats at the mouth of the Changjiang (Yangtze) River, China. Geomorphology 38(1-2), pp. 167–180.

Yang, S. L., Zhang, J. and Xu, X. J., 2007. Influence of the Three Gorges Dam on downstream delivery of sediment and its environmental implications, Yangtze River. Geophysical Research Letters 34(10), pp. L10401.

Yang, Z., Wang, H., Saito, Y., Milliman, J. D., Xu, K., Qiao, S. and Shi, G., 2006. Dam impacts on the Changjiang (Yangtze) River sediment discharge to the sea: The past 55 years and after the Three Gorges Dam. Water Resources Research 42(4), pp. W04407.

Zhang, K. Q., Douglas, B. C. and Leatherman, S. P., 2004. Global warming and coastal erosion. Climatic Change 64(1), pp. 41–58.

Zhang, Y., Swift, D. J. P., Yu, Z. Y. and Jin, L., 1998. Modeling of coastal profile evolution on the abandoned delta of the huanghe river. Marine Geology 145(1-2), pp. 133–148.

Zhou, D., Liang, Y. B. and Zeng, C. k., 1994. Oceanology of China seas. Vol. Volume II, Kluwer Academic Publishers, Dordrecht; Boston.

Zuo, X., Aiping, F., Ping, Y. and Dongxing, X., 2009. Coastal erosion induced by human activities: A northwest bohai sea case study. Journal of Coastal Research 25(3), pp. 723–733.

GIS-BASED MULTICRITERIA LAND SUITABILITY EVALUATION USING ORDERED WEIGHT AVERAGING WITH FUZZY QUANTIFIER: A CASE STUDY IN SHAVUR PLAIN,IRAN

M. Mokarram [a,]* , F. Aminzadeh [b]

[a] Dept. of Remote Sensing and GIS, Shahid Chamran University,Ahwaz, Iran (m.mokarram.313, kazemrangzan) @gmail.com

[b] Dept. of Computer Engineering, Shahid Chamran University,Ahwaz, Iran, ln.aminzadeh@gmail.com

Member of young researcher club of Islamic Azad university of Safashahr, Iran

Key words: GIS; Fuzzy quantifiers; Ordered weighted averaging; Multicriteria evaluation; Land-use suitability analysis

ABSTRACT:

Cell-based Multicriteria Evaluation (MCE) methods are used to analyse the land suitability evaluation. Land evaluation is carried out to estimate the suitability of land for a specific use such as arable farming or irrigated agriculture. land suitability evaluation is a prerequisite for land-use planning and development (Sys 1985; Van Ranst and others 1996). It provides information on the constraints and opportunities for the use of the land and therefore guides decisions on optimal utilization of land resources (FAO1983). The aim in integrating Multicriteria Decision Analysis (MCDA) with Geographical Information Systems (GIS) is to provide more flexible and more accurate decisions to the decision makers in order to evaluate the effective factors. Furthermore, By changing the parameters in this type of method, a wide range of decision strategies or scenarios can be generated in some procedures. The goal of this research is to take the advantage of incorporation of fuzzy (linguistic) quantifiers into GIS-based land suitability analysis by ordered weighted averaging (OWA). OWA is a multicriteria evaluation procedure (or combination operator). The nature of the OWA procedure depends on some parameters, which can be specified by means of fuzzy (linguistic) quantifiers. The quantifier-guided OWA procedure is illustrated using land-use suitability analysis in Shavur plain,Iran.

1. INTRODUCTION

Land-use suitability mapping and analysis is one of the most useful applications of GIS for spatial planning and management (Collins et al., 2001; Malczewski, 2004). Land-use suitability analysis is a multicriteria evaluation,which aims at identifying the most appropriate spatial pattern for future land uses according to specify requirements, preferences, or predictors of some activity (Hopkins, 1977; Collins et al., 2001). Geographic information systems (GIS) serve the multicriteria evaluation function of suitability assessment well, providing the attribute values for each location and both the arithmetic and logical operators for combining attributes (Jiang and Eastman 2000). Furthermore multicriteria evaluation may be used to develop and evaluate alternative plans that may facilitate compromise among interested parties (Malczewski, 1996). In general, the GIS-based land suitability analysis assumes that a given study area is subdivided into a set of basic unit of observations such as polygons or rasters. Then, the land-use suitability problem involves evaluation and classification of the areal units according to their suitability for a particular activity. Over the last 10 years or so, land-use suitability problems have increasingly been conceptualized in terms of the GIS-based multicriteria evaluation procedures (e.g. Banai, 1993; Jankowski and Richard, 1994; Joerin, 1995 ; Barredo, 1996; Antonie et al., 1997; Lin et al., 1997; Beedasy and Whyatt, 1999; Malczewski, 1999; Barredo et al., 2000; Mohamed et al., 2000; Bojorquez-Tapia et al., 2001; Dai et al., 2001; Joerin et al., 2001). There are two fundamental classes of multicriteria evaluation methods in GIS: the Boolean overlay operations (noncompensatory combination rules) and the weighted linear combination (WLC) methods (compensatory combination rules). They have been the most often used approaches for land-use suitability analysis (Heywood et al., 1995; Jankowski, 1995; Barredo, 1996; Beedasy and Whyatt, 1999; Malczewski, 2004). These approaches can be generalized within the framework of the ordered weighted averaging (OWA) (Asproth et al., 1999; Jiang and Eastman, 2000; Makropoulos et al., 2003; Malczewski et al., 2003; Malczewski and Rinner, 2005; Malczewski .,2006). OWA is a family of multicriteria combination procedures (Yager, 1988). Conventional OWA can utilizes the qualitative statements in the form of fuzzy quantifiers(Yager, 1988, 1996). The main goal of this paper is to produce the land suitability maps according to OWA operators for GIS-based multicriteria evaluation procedures.

2. METHODS

OWA is a multicriteria evaluation procedure (or combination operator). The nature of the OWA procedure depends on some parameters, which can be specified by means of fuzzy (linguistic) quantifiers. The GIS-based multicriteria evaluation procedures involve a set of geographically defined alternatives (e.g. parcels of land) and a set of evaluation criteria represented as map layers. The problem is to combine the criterion maps according to the attribute values and decision maker's preferences using a combination rule. each alternative (i = 1, 2, . . . ,m) is represented as a cell (raster) or a polygon and is described by a set of standardized criterion values: $a_{ij} \in [0, 1]$ for j = 1, 2, . . . ,n. A multicriteria evaluation problem involves also preferences which are typically specified as the criterion weights, $w_j \in [0, 1]$ for j = 1, 2, . . ., n, and $\sum_{j=1}^{n} w_j = 1$. Given the input data (a set of criterion map layers and criterion weights), the OWA combination operator associates with the i-th location (e.g., raster or point) a set of order weights v = v_1, v_2, . . . , v_n such that v_j ϵ[0, 1], j=1,2,..,n, $\sum_{j=1}^{n} v_j = 1$, and is defined as follows (see Yager, 1988; Malczewski et al., 2003):

$$OWA_i = \sum_{j=1}^{n} \left(\frac{u_j v_j}{\sum_{j=1}^{n} u_j v_j} \right) z_{ij} , \qquad (1)$$

where $z_{i1} \geq z_{i2} \geq . . . \geq z_{in}$ is the sequence obtained by reordering the attribute values a_{i1}, a_{i2}, . . ., a_{in}, and u_j is the criterion weight reordered according to the attribute value, z_{ij}. It is important to point to the difference between the two types of weights (the criterion weights and the order weights). The criterion weights are assigned to evaluation criteria to indicate their relative importance. All locations on the j-th criterion map are assigned the same weight of w_j. The order weights are associated with the criterion values on the location-by-location basis. They are assigned to the i-th location's attribute value in decreasing order without considering from which criterion map the value comes.
With different sets of order weights, one can generate a wide range of OWA operators including the most often used GIS-base map combination procedures: the weighted linear combination (WLC) and Boolean overlay operations, such as intersection (AND) and union (OR) (Yager, 1988; Malczewski et al., 2003). The AND and OR operators represent the extreme cases of OWA and they correspond to the MIN and MAX operators, respectively. The order weights associated with the MIN operator are: v_n = 1, and v_j = 0 for all other weights. Given the order weights, $OWA_i(MIN) = MIN_j(a_{i1}, a_{i2}, . . ., a_{in})$. The following weights are associated with the MAX operator: v_1 = 1, and v_j = 0 for all other weights, and consequently $OWA_{i(MAX)} = MAX_j(a_{i1}, a_{i2}, . . ., a_{in})$. Assigning equal order weights (that is, v_j = 1/n for j = 1, 2, . . . , n) results in the conventional WLC, which is situated at the mid-point on the continuum ranging from the MIN to MAX operators (Malczewski, 2006).
Given a set of criterion maps and a fuzzy linguistic quantifier Q, one can perform a procedure for combining the criteria based on a statement regarding the relationship between the evaluation criteria.
Based on the type of linguistically quantified
statements one can distinguish between: the absolute linguistic quantifiers and the relative (or proportional) linguistic quantifiers (Zadeh, 1983). There is no empirical evidence to show which of the two classes of linguistic quantifiers is more suitable for multicriteria evaluation. Here we will focus on a class of the proportional quantifiers known as the regular increasing monotone (RIM) quantifiers (Yager, 1996). To identify the quantifier we employ one of the simplest and the most often used methods for defining a parameterized subset on the unit interval (Yager, 1996). Specifically,

$$Q(p) = p^{\alpha} , \alpha>0 \qquad (2)$$

Q(p) is represented as a fuzzy set in interval [0, 1]. Table 1 shows a selection of the RIM quantifiers and their characteristics. By hanging the parameter, α, one can generate different types of quantifiers and associated operators between the two extreme cases of the all and at least one quantifiers. For a = 1, Q(p) is proportional to α and therefore it is referred to as the identity quantifier. As α tends to zero, the quantifier Q(p) approaches its extreme case of at least one, which corresponds to the MAX
operator. As α tends to infinity, the quantifier Q(p) approaches its extreme case of all, which corresponds to the MIN operator.
The order weights according to RIM quantifier is defined as follows:

$$v_j = \left(\sum_{k=1}^{j} u_k \right)^{\alpha} - \left(\sum_{k=1}^{j-1} u_k \right)^{\alpha} \qquad (3)$$

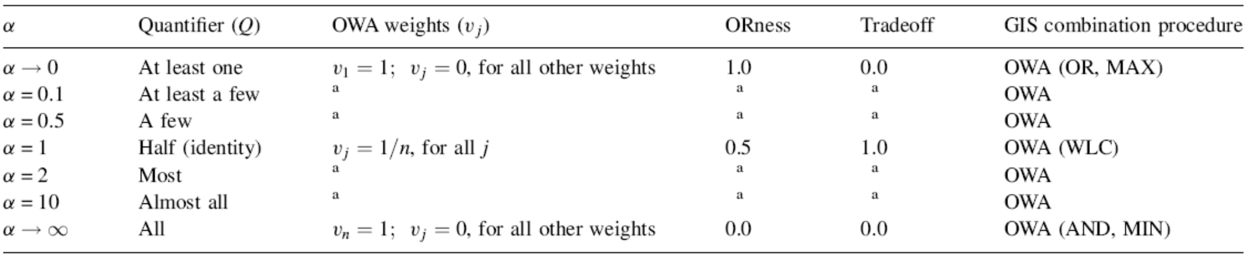
Given the criterion weights, w_j, and order weights, v_j, the quantifier-guided OWA is defined as follows:

$$OWA_i = \sum_{j=1}^{n} \left(\left(\sum_{k=1}^{j} u_k \right)^{\alpha} - \left(\sum_{k=1}^{j-1} u_k \right)^{\alpha} \right) z_{ij} \qquad (4)$$

3. CASE STUDY

3.1.Study area

The study area, Shavoue, lies in the Northern of Khouzestan province,Iran. It is located within coordinate of latitude 31°37'30'' and 32°30'00'' North and longtitude 48°15'00'' and 48°40'40'' East with the area of 77404/23 ha (hectar). (figure1.)



α	Quantifier (Q)	OWA weights (v_j)	ORness	Tradeoff	GIS combination procedure
$\alpha \to 0$	At least one	$v_1 = 1;\ v_j = 0$, for all other weights	1.0	0.0	OWA (OR, MAX)
$\alpha = 0.1$	At least a few	a	a	a	OWA
$\alpha = 0.5$	A few	a	a	a	OWA
$\alpha = 1$	Half (identity)	$v_j = 1/n$, for all j	0.5	1.0	OWA (WLC)
$\alpha = 2$	Most	a	a	a	OWA
$\alpha = 10$	Almost all	a	a	a	OWA
$\alpha \to \infty$	All	$v_n = 1;\ v_j = 0$, for all other weights	0.0	0.0	OWA (AND, MIN)

Table 1. Some properties of the RIM quantifiers for selected values of the a parameter

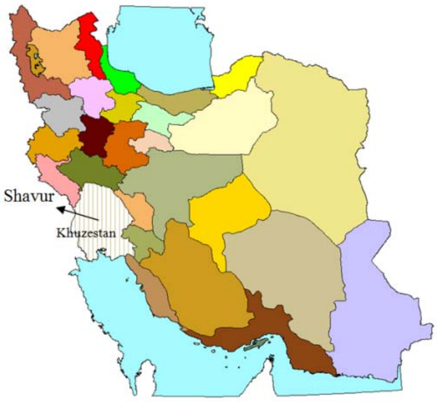


Figure 1. Location of study area

3.2.Criteria Evaluation

For all criteria that are seen as map layer, the criterion values are generated. The causative factors for the land suitability evaluation are EC, pH, ESP, $CaCO_3$, Gypsium, wetness, texture, slope, depth and topography. The data in this study is validated by Power Ministry of Khouzestan.

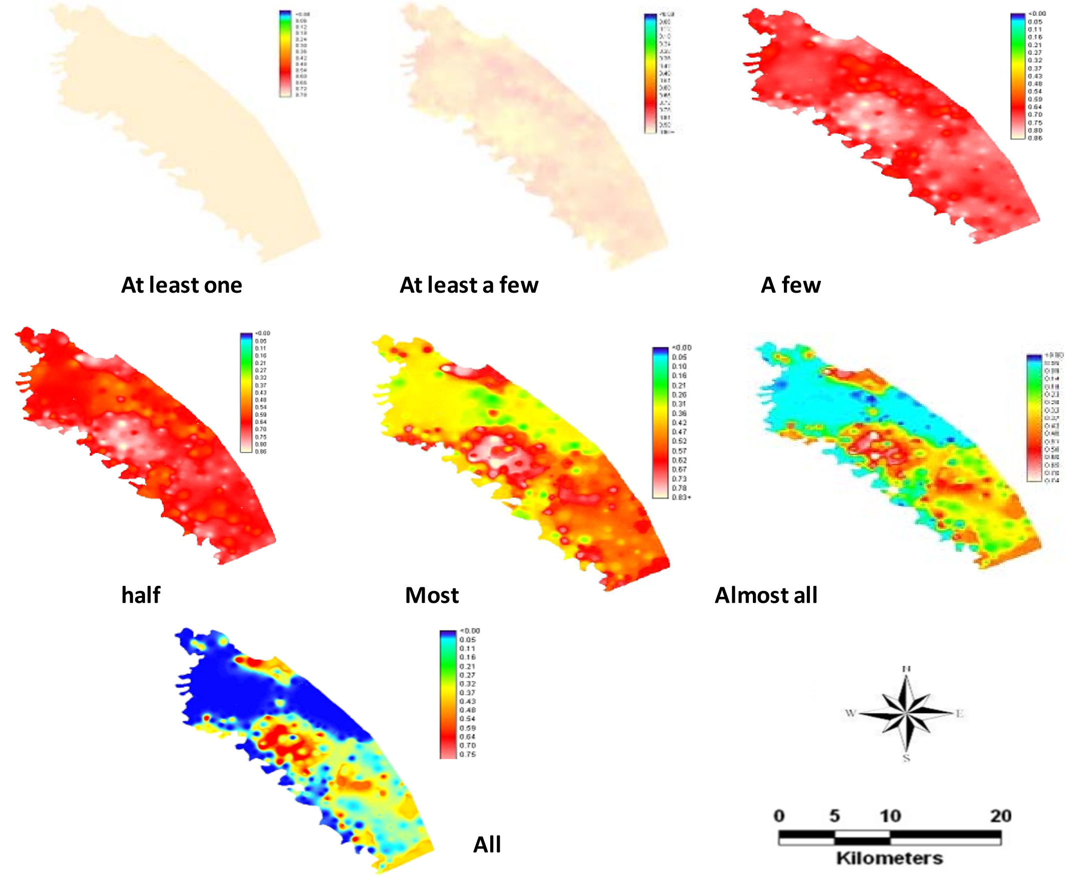
3.3.Assigning Criteria Weights

The purpose of the criterion weighting is to express the importance of each criterion relative to other criteria. The more important criterion had the greater weight in the overall evaluation. Using Eq. (3), the following estimated values for the criterion weights of EC & ESP, topography, wetness, texture, PH, $CaCO_3$, Gypsium and depth are 0.222, .028, 0.194, 0.083, 0.111, 0.167, 0.139 and 0.056. Given the standardized criterion maps and corresponding criterion weights, we apply the OWA operator using Eq. (4) for selected values of fuzzy quantifiers: at least one, at least a few, a few, identity, most, almost all, and all. Each quantifier is associated with a set of order weights that are calculated according to Eq. (3). Fig. 2 shows the seven alternative land suitability patterns.

4.CONCLUSIONS

The optimal use of reserved land resources for agriculture is a complex problem that involves subjective assessments with multiple criteria. This paper has presented a GIS-based multicriteria land suitability evaluation using Ordered Weight Averaging with fuzzy quantifier approach for effectively solving this problem. An empirical study in Shavour, Iran has been conducted using the approach presented.

The fuzzy-quantifier-based OWA approach is capable of capturing qualitative information the decision maker or analyst may have regarding his/her perceived relationship between the different evaluation criteria. It is in this effort one can see the benefit of the fuzzy quantifier approach to GIS-based multicriteria analysis. This is especially true in situations involving a large number of criterion maps. In such situations, it is impractical or even impossible to specify the exact relationships between evaluation criteria. The OWA approach provides a mechanism for guiding the decision maker/analysis through the multicriteria combination procedures. It allows him/her to explore different decision strategies or scenarios. Consequently, the approach facilitates a better understanding of the alternative land-use suitability patterns


At least one

At least a few

A few

half

Most

Almost all

All

Figure 2. Land suitability maps of OWA results for selected fuzzy linguistic quantifiers in Shavur

REFERENCES

Antonie, J., Fischer, G. and Makowski, M., 1997. Multiplecriteria land use analysis. Applied Mathematics and Computation. Vol.85 pp.195-215.

Asproth, V., Holmberg, S.C., Ha ˚kansson, A., 1999. Decision Support for spatial planning and management of human settlements.In: International Institute for Advanced Studies in Systems Research and Cybernetics. In: Lasker, G.E. (Ed.), Advances in Support Systems Research, vol. 5. Windsor, Ont., Canada, pp. 30–39.

Banai, R., 1993. Fuzziness in geographic information systems: contributions from the analytic hierarchy process. International J. Geogr. Inform. Syst. 7 (4), 315–329.

Barredo, J.I., 1996. Sistemas de Informacion Geografica y Evaluacio Multicriterio en la Ordenacion del Territorio. Ra-Ma, Madrid.

Barredo, J.I., Benavidesz, A., Hervhl, J., van Westen, C.J., 2000. Comparing heuristic landslide hazard assessment techniques using GIS in the Tirajana basin, Gran Canaria Island,Spain. International J. Appl. Earth Observ. Geoinform. 2 (1), 9–23.

Beedasy, J.,Whyatt, D., 1999.Diverting the tourists: aspatial decisionsupport system for tourism planning on a developing island. J. Appl. Earth Observ. Geoinform. 3/4, 163–174.

Bojorquez-Tapia, L.A., Diaz-Mondragon, S., Ezcurra, E., 2001. GIS-based approach for participatory decision-making and land suitability assessment. Int. J. Geogr. Inform. Sci. 15 (2), 129–151.

Collins, M.G., Steiner, F.R., Rushman, M.J., 2001. Land-use suitability analysis in the United States: historical development and promising technological achievements. Environ. Manage. 28 (5), 611–621.

Dai, F.C., Lee, C.F., Zhang, X.H., 2001. GIS-based geo-environmental evaluation for urban land-use planning: a case study. Eng. Geol. 61 (4), 257–271.

FAO. 1983. Guidelines: Land evaluation for rainfed agriculture. Soils Bulletin 52. FAO, Rome, 237 pp.

Geoinformation. 8: 270–277.

Heywood, I., Oliver, J., Tomlinson, S., 1995. Building an exploratory multi-criteria modelling environment for spatial decision support. In: Fisher, P. (Ed.), Innovations in GIS, vol. 2. Taylor & Francis, London, pp. 127–136.

Hopkins, L., 1977. Methods for generating land suitability maps: a comparative evaluation. J. Am. Inst. Planners 34 (1), 19–29.

Jankowski, P., 1995. Integrating geographical information systems and multiple criteria decision making methods. International J. Geogr. Inform. Syst. 9 (3), 251–273.

Jankowski, P., Richard, L., 1994. Integration of GIS-based suitability analysis and multicriteria evaluation in a spatial decision support system for route selection. Environ. Plann. B 21 (3), 326–339.

Jiang, H., Eastman, J.R., 2000. Application of fuzzy measures in multi-criteria evaluation in GIS. Int. J. Geogr. Inform. Syst. 14, 173–184.

Joerin, F., 1995.Me ´thode multicrite `re d'aide a ` la de ´cision et SIG pour la recherche d'un site. Rev. Int. Ge ´omatique 5 (1), 37–51.

Joerin, F., The ´riault, M., Musy, A., 2001. Using GIS and outranking multicriteria analysis for land-use suitability assessment. Int. J. Geogr. Inform. Sci. 15 (2), 153–174.

Lin, H., Wan, Q., Li, X., Chen, J. and Kong, Y., 1997. GIS-based multicriteria evaluation for investment environment. Environment and Planning B: Planning and Design, v:24, pp:403-414

Makropoulos, C., Butler, D., Maksimovic, C., 2003. A fuzzy logic spatial decision support system for urban water management. J. Water Resour. Plann. Manage. 129 (1), 69–77.

Malczewski, J. (2006). Ordered weighted averaging with fuzzy quantifiers: GIS-based multicriteria evaluation for land-use suitability analysis. International Journal of Applied Earth Observation and Geoinformation. 8: 270–277.

Malczewski, J., 1996. A GIS-based approach to multiplecriteria group decision making. International Journal of Geographical Information Systems 10(8), 955-971.

Malczewski, J., 1999. GIS and Multicriteria Decision Analysis. John Wiley & Sons Inc., New York.

Malczewski, J., 2004. GIS-based land-use suitability analysis: a critical overview. Progr. Plann. 62 (1), 3–65.

Malczewski, J., Chapman, T., Flegel, C., Walters, D., Shrubsole, D., Healy, M.A., 2003. GIS-multicriteria evaluation with ordered weighted averaging (OWA): case study of developing watershed management strategies. Environ. Plann. A 35 (10), 1769–1784.

Malczewski, J., Rinner, C., 2005. Exploring multicriteria decision strategies in GIS with linguistic quantifiers: a case study of residential quality evaluation. J. Geogr. Syst. 7 (2), 249–268.

Mohamed, A.B.A.A., Sharifi, M.A., van Keulen, H., 2000. An integrated agro-economic and agro-ecological methodology for land use planning and policy analysis. Int. J. Appl. Earth Observ. Geoinform. 2 (2), 87–103.

quantifiers: GIS-based multicriteria evaluation for land-use suitability

Sys, C. (1985) Land evaluation, State University of Ghent, Ghent; The Netherlands

Van Ranst, E., H. Tang, R. Groenemans, and S. Sinthura hat. 1996. Application of fuzzy logic to land suitability for rubber production in peninsular Thailand. Geoderma 70:1–19.

Yager, R.R., 1988. On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans. Syst. Man Cybernet. 18 (1), 183–190.

Yager, R.R., 1996. Quantifier guided aggregation using OWA operators. Int. J. Intell. Syst. 11, 49–73.

Zadeh, L.A., 1983. A computational approach to fuzzy quantifiers in natural languages. Comput. Math. Applic. 9, 149–184.

IDENTIFICATION OF MUNICIPAL POLICIES THAT INFLUENCE THE DISTRIBUTION OF GREEN COVER ACROSS METROPOLITAN REGIONS

S. J. Lee [a], T. Longcore [a], J. P. Wilson [a]

[a] Dept. of Geography, University of Southern California, 3620 S. Vermont Ave, KAP 444, Los Angeles, CA, 90089-0255, USA - (sujinlee, longcore, jpwilson)@usc.edu

KEY WORDS: Green Cover, Policy, Aerial Photography, Feature Analyst

ABSTRACT:

Nature's services provided by green cover are important to environmental conditions in cities and their ability to adapt to climate change. Researchers using geospatial technologies have dramatically increased the spatial and temporal resolution of knowledge about the distribution of tree and shrub cover in cities. Much of the current research on tree cover in cities has concentrated on individual preferences and associations between socioeconomic characteristics and environmental conditions. To complement existing research and provide planners with the practical tools they need to maintain the benefits of urban nature, this study focuses on the public policy factors that influence tree and other green cover at the lot and neighbourhood scales, concentrating on single family neighbourhoods. Green cover is classified using an object-oriented method with high spatial-resolution aerial imagery and GIS techniques. Landscape and property information were extracted from Los Angeles County Assessor Office files at a parcel scale for 20 cities in Los Angeles County. The extracted variables included lot size, floor-area ratio, residential landscape standards, tree protection ordinances, and street tree programs and were used along with average temperature and rainfall information in multiple regression models to explain the distribution and character of green cover across different neighbourhoods.

1. INTRODUCTION

The need for and benefits of green cover and especially forests within US cities has been well documented (McPherson and Rowntree, 1993; Nowak, 1993; McPherson, et al., 2005; Barbosa, et al., 2007). These benefits include, for example, increased groundwater percolation and recharge, improved air quality, increased carbon sequestration and biodiversity, reduced urban heat island impacts and energy consumption for air conditioning, and stormwater runoff reductions (Simpson and McPherson, 1996; McPherson and Simpson, 1999; Akbari, et al., 2001; Akbari, 2002; Xiao and McPherson, 2003; Carver, et al., 2004; Donovan and Butry, 2009). Researchers have investigated the green cover effects on energy use (Bengston, et al., 2004; Ewing and Rong, 2008) and aesthetics and neighbourhood character (Szold, 2005; Nasar, et al., 2007), but the consequences for ecosystem services and biodiversity have not yet been adequately described.

The green cover has been maintained by tree planting programs that often are directed at publicly owned lands such as parks or easements along streets. The potential ecosystem services and biodiversity benefits cannot be fully realized only on public land, but rather require involvement of private landowners. The largest single land use in which such actions can take place is low density residential development (Wu, et al., 2008). Although researchers have investigated various socioeconomic correlates of landscape characteristics within residential neighborhoods, theses efforts have been geographically limited (Martin, et al., 2004; Grove, et al., 2006; Troy, et al., 2007) and not yet linked to policy decisions (e.g., tree preservation ordinances, zoning and building codes) that could influence them.

There are several noteworthy trends in urban morphology and social norms that influence both the prospects for provision of ecosystem services within residential neighborhoods and the function of these neighborhoods as ecological spaces within the city, as illustrated by the following three examples.

First, the size of the average single-family dwelling has almost doubled over the past 50 years (Szold, 2005). In some regions, these houses are disparagingly called "monster homes" (Szold, 2005) or "McMansions" (Nasar, et al., 2007), because they are extended to the minimum legal setbacks and despite their size, they are occupied by fewer residents than smaller homes on average (Breunig, 2003).

Second, access to parks and green space is unequally distributed among the poor and people of color (Loukaitou-Sideris, 1995; Wolch, et al., 2005). This pattern reinforces itself because real estate prices correlate positively with surrounding green cover (Conway, et al., 2008) and urban green spaces are disproportionately found in wealthy areas (Iverson and Cook, 2000). As a consequence, green space and its ecological functions can be characterized as an outgrowth of socioeconomic characteristics that may seem to be beyond the control of planners. This creates a negative feedback loop wherein disadvantaged communities are disproportionately denied access to both urban forest amenities and natural open space.

Third, the increasing proportion of the US population that lives in cities decreases the access that the average resident has to nature in general. The human relationship with the planet's natural ecosystems increasingly depends on the lessons learned through interaction with urban nature. The experiences of nature, especially as children, are important factors leading to environmental sensitivity as adults (Tanner, 1980; Chawla, 1999). Therefore, this study investigated the factors that influence green cover and natural values within residential neighborhoods and the policies that can change them across a sample of cities in Los Angeles County, California.

2. METHODS

A series of single family neighbourhoods (SFNs) was randomly selected in 20 of the 24 cities in Los Angeles County (LAC), California with populations of at least 80,000 and used to examine the impact of the presence and character of city policies on urban tree cover (Figure 1). Four cities were excluded – Los Angeles because of its great size and diversity in terms of environmental and socio-economic conditions and Lancaster, Palmdale, and Santa Clarita because of their locations to the north of Angeles National Forest and the increased aridity that characterizes these environmental settings (Figure 2). The remaining 20 cities varied tremendously in terms of green cover and the socio-economic and environmental characteristics that have routinely been used to explain this variability.

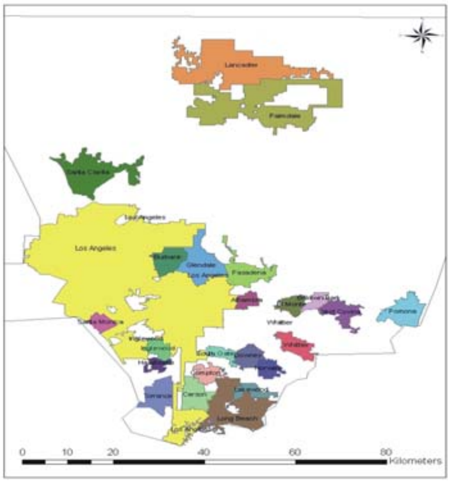


Figure 1. Cities in Los Angeles County with populations greater than 80,000.

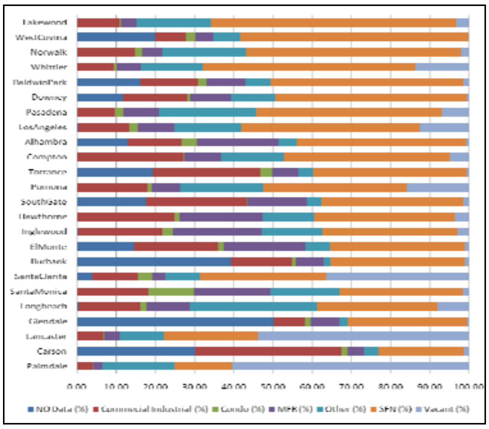


Figure 2. Land uses in the cities in Los Angeles County with populations greater than 80,000.

We selected census block groups within these cities with at least 222 SFNs and that covered at least 37% of the block group area. A total of 656 census block groups and 224,861 SFNs were selected using these criteria. The chosen SFNs covered 54.7% of the census block group areas on average.

The four subsections that follow describe how the various datasets were acquired, analyzed and interpreted.

2.1 Data acquisition and pre-processing

Green cover in the study areas was identified from 2006 color orthoimagery that was downloaded from the USGS in uncompressed, georectified, and tagged image file format (TIFF) at a spatial resolution of one foot (i.e. 0.3048 m) and saved in an ArcGIS geodatabase.

Parcel boundaries and attributes were extracted from the LAC Assessor's office and used to identify the SFNs in each city and compile house characteristics for the neighbourhoods that were chosen and used in our analysis. Figure 2 shows the ratio of land uses in each city from highest to lowest in terms of the proportion of the land area devoted to SFNs and confirms the point made earlier – that these residential areas are important given that they occupy 36% of the land area on average across the 24 cities listed here. Census information at the block group level was obtained from the US Census Bureau website and used to characterize the residents once the sample neighbourhoods were chosen.

Information about city policies with the potential to influence green cover – tree, landscape, water and zoning ordinances – was collected from city websites and phones calls to the appropriate city offices. Seven of the cities – Burbank, Glendale, Lakewood, Pasadena, Pomona, Santa Monica, and Whittier – had earned the designation "Tree City USA®" and three of these cities (Glendale, Pasadena, and Pomona) and one other (West Covina) had passed tree protection ordinances within the past 10–15 years.

2.2 Image Classification

We mosaicked and saved the images for each city as raster catalogs in a geodatabase file. We then used the object-based classification approach in Feature Analyst (Visual Learning Systems (VLS), Missoula, Montana) to digitize the green cover in the SFNs. This software uses a training dataset for which the user manually digitizes green cover and has been successfully used to classify urban land uses and land cover types (Zhou and Wang, 2007; Yuan, 2008; Miller et al., 2009).

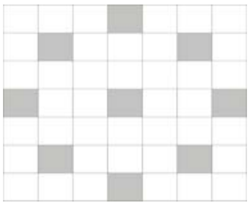


Figure 3. Bull's Eye 3 search pattern used to identify individual trees, shrubs, and other natural features.

For this study, we used the Feature Analyst with the following seven step procedure: (1) add the aerial image with true color (red, green, and blue) to ArcMap 9.3; (2) digitize the training sites; (3) set the feature type to natural feature to extract individual trees, shrubs, and other natural features; (4) set input red, green, and blue bands as reflectance; (5) set the input representation as Bull's Eye 3 (Figure 3) because this is the best model to identify natural features such as trees and shrubs; (6) set the masking tabs to select the regions of interest; and (7) set the learning options to help select parameters for aggregating areas, smoothing shapes, or filling background features. The minimum search area was specified as a 3 x 3 window (0.85 m^2) at this last step.

We then conducted an accuracy assessment using 500 random sites from the study areas (Congalton 1991), which showed that the method identified trees, shrubs and other natural features with > 90% accuracy. Figure 4 shows the classified green cover (green and orange areas combined) and the portion of this green cover that overlapped the parcels in our sample SFNs (shown in green on top of parcel with red boundaries in this graphic).

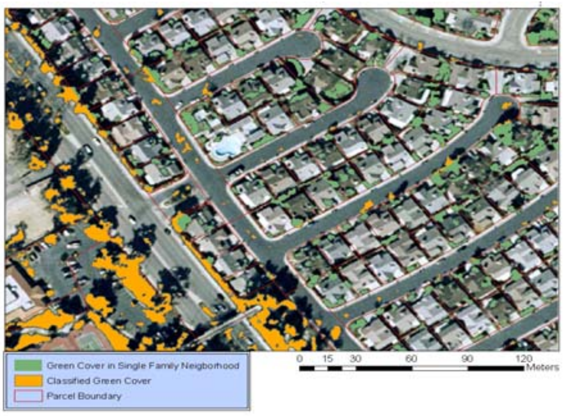


Figure 4. Green cover classification performed with Feature Analyst.

2.3 City policy analysis

For this aspect of the study we distinguished tree, landscape, water, and zoning ordinances similar to Hill et al. (2010). Our first task was to identify those cities that had earned the "Tree City USA®" designation. This program is sponsored by the Arbor Day Foundation in cooperation with the USDA Forest Service and National Association of State Foresters (Arbor Day Foundation, 2009). The many benefits of being a Tree City include creating a framework for action and education, a positive public image, and citizen pride. To earn this designation, a city must have: (1) a tree board or department; (2) a tree care ordinance: (3) a community forestry program with an annual budget of at least $2 per capita; and (4) an Arbor Day observance and proclamation. We also recorded how many years the seven cities had the "Tree City USA®" designation, and whether the cities had a public or street tree ordinance, a specific tree protection ordinance, and how many types of trees were protected by the aforementioned ordinances.

Residential areas are subject to hundreds of zoning and building regulations, but for the purposes of our study, we limited our attention to those that could affect tree canopy cover. Many of these regulations specify numerical minima and maxima, such as the minimum front, side, rear yard setbacks, maximum building height, minimum lot area, maximum lot coverage, maximum floor area ratio (FAR), and minimum floor area. Among these, we selected the minimum front, side, and rear yard setbacks as well as the minimum lot area since these indicate the spaces that can be used to plant new trees and/or maintain existing trees.

Last, we also checked whether or not the cities have a specific residential landscape ordinance or a water efficient landscape ordinance (since the latter may encourage homeowners to practice water conservation, plant specific types of native or drought resistant plants, and/or adhere to limited watering and irrigation hours) since these ordinances can also affect the level and character of the green cover present in SFNs.

2.4 Regression Modeling

We built numerous linear regression models to identify the relationship between green cover and various independent variables (representing city policies, environmental parameters, house characteristics, and occupant characteristics; Figure 5) across the 20 cities. Our study hypothesis is that green cover is related to one or more of the characteristics that are described in more detail in Appendix I.

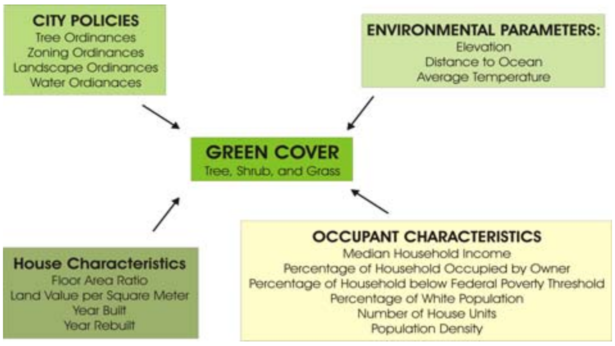


Figure 5. Variables utilized in multiple regressions.

The variables were collected at various scales (pixel, parcel, block group, city) and had to be spatially joined using ArcGIS 9.3. These variables were then checked for normality, homogeneity, multicollinearity, transformed if necessary, and the Akaike (1974) information criterion used to evaluate model performance in STATA 11. We conducted three different multi-regressions (stepwise forward and backward, and OLS) using the parcel and census block group as the unit of analysis. From the parcel analysis unit, the adjusted R-square was below 0.2, in past because we obtained over three million observations. Hence, we draw on concentrated block group areas to improve analysis. Throughout processing, we excluded insignificant variables, and finally we obtained the independent variables listed in Table 2.

3. RESULTS

Table 1 shows the percentage of SFNs and green cover in the single family neighbourhoods we examined in each city. Overall, the results in Table 1 show that the cities with the greenest SFNs are Pasadena, Santa Monica, Torrance, Norwalk, and West Covina.

These results point to the complicated set of drivers that determine green cover in residential settings. Burbank and Santa Monica, for example, have been Tree City USA® designees for 32 and 19 years, respectively without specifying a single protected tree type. In addition, Burbank is third from the bottom in terms of GC in SFNs. In contrast, Pasadena has 11 tree species as well as landmark and strategic trees that are protected in both public and private areas. West Covina, Pomona, and Glendale have five, four, and three trees protected by tree ordinances, and yet Glendale has relatively sparse GC in

SFNs (5.4%). Torrance, Norwalk and Inglewood, on the other hand, have expansive GC but lack tree policies.

City	SFN (%)	GC (%)	City	SFN (%)	GC (%)
Alhambra	55.5	6.1	Lakewood	54.5	3.6
Baldwin Park	52.6	5.4	Long Beach	56.6	5.7
Burbank	57.3	3.3	Norwalk	53.1	9.4
Carson	53.8	1.2	Pasadena	52.1	11.7
Compton	48.5	3.4	Pomona	61.9	7.5
Downey	53.8	5.3	Santa Monica	56.8	10.8
El Monte	45.9	2.3	South Gate	56.6	7.3
Glendale	55.5	5.4	Torrance	52.9	10.1
Hawthorne	57.7	4.6	West Covina	58.5	8.9
Inglewood	53.4	8.8	Whittier	54.5	4.6

Table 1. Percentage of SFNs and GC in single family neighbourhoods by city.

These kinds of contradictions also explain why we searched for additional variables (see Figure 5 for a complete list) and constructed multiple regression models linking the underlying neighbourhood characteristics and green cover. Table 2 lists the eight independent variables that were significant in explaining the variability of green cover in SFNs across the 20 cities. The level of GC in these neighbourhoods was negatively correlated with average floor area ratio (i.e. the bigger the house area, the lower the GC), number of house units (i.e. the larger number of units, the lower the GC), elevation (i.e. higher elevations correlated with less the GC), and population density (i.e. the higher the density, the lower the GC). Number of protected tree species, land values, minimum lot size, and the average percentage of households occupied by owners were positively correlated with the level of GC in SFNs.

Independent Variables	Coefficient	T - value
Protected tree species	61.887	(12.27)**
Number of house units	–2,234.31	(11.56)**
Land value per m^2	0.00093	(9.33)**
Floor area ratio	–333.574	(7.80)**
Elevation	–0.125	(6.71)**
Population density	–0.43	(4.25)**
Minimum lot size	0.001	(3.96)**
Household occupied by owner	0.34	(3.12)**
Constant	307.702	(16.06)**

Absolute value of t statistics in parentheses
* significant at 5%; ** significant at 1%

Table 2. Results of multiple regression model.

The eight variables listed in Table 2 were all significant and explained 55.5% of the variability in GC across 551 census block groups (Akaike's information criterion: 9.674, Bayesian information criterion: 1891.652). The coefficients show that the proportion of GC in a SFN is positively correlated with the number of tree ordinances, land values, lot size, zoning ordinances, and the percentages of owner occupied units – all variables that might be directly or indirectly influenced by city policies.

4. DISCUSSION AND CONCLUSIONS

Our main goal has been explore the role of city policies in determining green cover in single family neighborhoods. Several of the significant variables in our model have shown up in earlier work. For instance, Landry and Pu (2009) found that residential tree cover in the City of Tampa, Florida was correlated with the proportion of parcels regulated by tree protection ordinances, median building age, median building cover, median market value, proportion of White and Hispanic, median age of persons, housing unit density, and proportion vacant housing units. Troy et al. (2007) examined predictors of vegetative cover on private lands in Baltimore, Maryland using population density, lot coverage, and building density in low-income areas. The results significantly indicate how social stratification is related to vegetation cover. Finally, Heynen (2006) investigated the relationship between changes in median household income and changes in urban forest canopy cover in Indianapolis, Indiana.

Our results extend the earlier work because we concentrated specifically on identifying city policies that are correlated with GC extent. Two of the variables identified by Landry and Pu (2009) and Troy et al. (2007) were retained in our final model: lot coverage and the proportion of parcels regulated by tree protection ordinances.

By concentrating on attributes of SFNs that can be regulated, and in some instances, changed by city decisionmakers, we have identified a useful path for planners and regulators seeking to maintain and increase ecosystem services in residential neighborhoods. Although our models include some attributes over which managers have no control — elevation, land value, owner occupancy — others can be regulated at the planning stage of development or even changed in existing SFNs. At the planning stage, planners might consider the adverse effects of small minimum lot sizes on resulting green cover and weigh it against the benefits of affordable housing from smaller lots. Ordinances actually protecting tree species turn out to be important in maintaining green cover, consistent with previous studies (Landry and Pu 2009; Troy et al. 2007).

Floor area ratio, which can be regulated through zoning action, is also an important predictor of GC and may be the best tool that municipalities have against mansionization of existing SFNs. The ecosystem services provided by GC on generously sized parcels are quickly lost when new homes are constructed to fill the entire area within lot line setbacks. The loss of these services has an effect on society as a whole, which should provide a public interest rationale to ensure that zoning codes cap the floor area ratio allowed in SFNs. Keeping floor area ratios restrained also counterbalances the effects of larger minimum lot sizes by keeping homes at a more modest size.

Future research should quantify the magnitude of ecosystem services provided by SFNs, given their large proportion of city area shown here. It should also trace out magnitude and rate of the loss of those services to mansionization — e.g., water management, buffering against climate change, and urban biodiversity (Tratalos et al. 2007). Such losses could be described for the past and potential losses modelled for the future under various policy scenarios. But even as these research routes are pursued, the current study indicates policy options for cities desiring to maintain trees and green cover in their residential neighborhoods.

REFERENCES

Akaike, H., 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic control*, 19(6), pp. 47–56.

Akbari, H., 2002. Shade Trees Reduce Building Energy Use and Co2 Emissions from Power Plants. *Environmental Pollution*, 116(S1), pp. S119–S136.

Akbari, H., Pomerantz, M., and Taha, H., 2001. Cool Surfaces and Shade Trees to Reduce Energy Use and Improve Air Quality in Urban Areas. *Solar Energy*, 70(3), pp. 295–315.

Arbor Day Foundation., 2009. Tree City USA Program, http://www.arborday.org/programs/treeCityUSA/index.cfm (accessed 10 Sep. 2009)

Barbosa, O., Tratalos, J.A., Armsworth, P.R., Davies, R.G., Fuller, R.A., Johnson, P., and Gaston, K.J., 2007. Who Benefits from Access to Green Space? A Case Study from Sheffield, United Kingdom. *Landscape and Urban Planning*, 83(2), pp. 187–195.

Bengston, D.N., Fletcher, J.O., and Nelson, K.C., 2004. Public Policies for Managing Urban Growth and Protecting Open Space: Policy Instruments and Lessons Learned in the United States. *Landscape and Urban Planning*, 69(2-3), pp. 271–286.

Breunig, K., 2003. *Losing Ground: At What Cost?* Changes in Land Use and Their Impact on Habitat, Biodiversity and Ecosystem Services in Massachusetts. Massachusetts Audubon Society. Lincoln, Massachusetts.

Carver, A.D., Unger, D.R., and Parks, C.L., 2004. Modeling Energy Savings from Urban Shade Trees: An Assessment of the CITYgreen Energy Conservation Module. *Environmental Management*, 34(5), pp. 650–655.

Chawla, L., 1999. Life Paths into Effective Environmental Action. *Journal of Environmental Education*, 31(1), pp. 15–26.

Congalton, R. G., 1991. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment*, 37(1), pp. 35–46.

Conway, D., Li, C.Q., Wolch, J., Kahle, C., and Jerrett, M., 2008. A Spatial Autocorrelation Approach for Examining the Effects of Urban Greenspace on Residential Property Values. *Journal of Real Estate Finance and Economics*, DOI 10.1007/s11146-008-9159-6.

Donovan, G.H., and Butry, D.T., 2009. The Value of Shade: Estimating the Effect of Urban Trees on Summertime Electricity Use. *Energy and Buildings*, 41(6), pp. 662–668.

ESRI., 2009. *Arcgis 9.3 Desktop Help*. ESRI. Redland, CA.

Ewing, R., and Rong, F., 2008. The Impact of Urban Form on U.S. Residential Energy Use. *Housing Policy Debate*, 19(1), pp. 1–30.

Grove, J.M., Cadenasso, M.L., Burch Jr, W.R., Pickett, S.T., Schwarz, K., O'Neil-Dunne, J., Wilson, M., Troy, A., and Boone, C., 2006. Data and Methods Comparing Social Structure and Vegetation Structure of Urban Neighborhoods in Baltimore, Maryland. *Society and Natural Resources*, 19(2), pp. 117–136.

Heynen, N., 2006. Green Urban Political Ecologies: Toward a Better Understanding of Inner-City Environmental Change. *Environment and Planning A*, 38(3), pp. 499–516.

Hill, E., Dorfman, J.H., and Kramer, E., 2010. Evaluating the Impact of Government Land Use Policies on Tree Canopy Coverage. *Land Use Policy*, 27(2), pp. 407–414.

Iverson, L.R., and Cook, E.A., 2000. Urban Forest Cover of the Chicago Region and Its Relation to Household Density and Income. *Urban Ecosystems*, 4(2), pp. 105–124.

Landry, S., and Pu, R., 2010. The Impact of Land Development Regulation on Residential Tree Cover: An Empirical Evaluation Using High-Resolution IKONOS Imagery. *Landscape and Urban Planning*, 94(2), pp. 94–104.

Loukaitou-Sideris, A., 1995. Urban Form and Social Context: Cultural Differentiation in the Uses of Urban Parks. *Journal of Planning Education and Research*, 14(2), pp. 89–102.

Martin, C.A., Warren, P.S., and Kinzig, A.P., 2004. Neighborhood Socioeconomic Status Is a Useful Predictor of Perennial Landscape Vegetation in Residential Neighborhoods and Embedded Small Parks of Phoenix, AZ. *Landscape and Urban Planning*, 69(4), pp. 355–368.

McPherson, E.G., and Rowntree, R.A., 1993. Energy Conservation Potential of Urban Tree Planting. *Journal of Arboriculture*, 19(6), pp. 321–331.

McPherson, E.G., and Simpson, J.R., 1999. *Carbon Dioxide Reduction through Urban Forestry: Guidelines for Professional and Volunteer Tree Planters*. United States Department of Agriculture, Albany, CA.

McPherson, G., Simpson, J.R., Peper, P.J., Maco, S.E., and Xiao, Q.F., 2005. Municipal Forest Benefits and Costs in Five US Cities. *Journal of Forestry*, 103(8), pp. 411–416.

Miller, J. E., Nelson, S.A.C., and Hess, G.R., 2009. An Object Extraction Approach for Impervious Surface Classification with Very-High-Resolution Imagery. *The Professional Geographer*, 61(4), pp. 250–264.

Nasar, J., Evans-Cowley, J., and Mantero, V., 2007. McMansions: The Extent and Regulation of Super-Sized Houses. *Journal of Urban Design*, 12(3), pp. 339–358.

Nowak, D.J., 1993. Atmospheric Carbon Reduction by Urban Trees. *Journal of Environmental Management*, 37(3), pp. 207–217.

Simpson, J.R., and McPherson, E.G., 1996. Potential of Tree Shade for Reducing Residential Energy Use in California. *Journal of Arboriculture*, 22(1), pp. 10–18.

Szold, T., 2005. Mansionization and Its Discontents: Planners and the Challenge of Regulating Monster Homes. *Journal of the American Planning Association*, 71(2), pp. 189–202.

Tanner, T., 1980. Significant Life Experiences. *Journal of Environmental Education*, 11(4), pp. 399–417.

Tratalos J., Fuller, R.A., Warren, P.H., Davies, R.G., Gaston, K.J., 2007. Urban Form, Biodiversity Potential and Ecosystem Services. *Landscape and Urban Planning* 83(4), pp. 308–317.

Troy, A.R., Grove, J.M., O'Neil-Dunne, J.P.M., Pickett, S.T.A., and Cadenasso, M.L., 2007. Predicting Opportunities for Greening and Patterns of Vegetation on Private Urban Lands. *Environmental Management*, 40(3), pp. 394–412.

Wolch, J., Wilson, J.P., and Fehrenbach, J., 2005. Parks and Park Funding in Los Angeles: An Equity-Mapping Analysis. *Urban Geography*, 26(1), pp. 4–35.

Wu, C., Xiaoa, Q., and McPherson, E.G., 2008. A Method for Locating Potential Tree-Planting Sites in Urban Areas: A Case Study of Los Angeles, USA. *Urban Forestry and Urban Greening*, 7(2), pp. 65–75.

Xiao, Q., and McPherson, E.G., 2003. Rainfall Interception by Santa Monica’s Municipal Urban Forest. *Urban Ecosystems*, 6(4), pp. 291–302.

Yuan, F., 2008. Land Cover Change and Environmental Impact Analysis in the Greater Mankato Area of Minnesota Using Remote Sensing and GIS Modeling. *International Journal of Remote Sensing*, 29(4), pp. 1169–1184.

Zhou, Y., and Wang, Y.Q., 2007. An Assessment of Impervious Surface Areas in Rhode Island. *Northeastern Naturalist*, 14(4), pp. 643–650.

ACKNOWLEDGEMENTS

We would like to gratefully acknowledge the summer support for this research provided to the first author by HSSBC, and to thank Chelsea Buckland for collecting all the city policy information.

APPENDIX I

City Policy	**Applied Ordinances**	**City Policy**	**Applied Ordinances**	
Tree Protection Ordinance	Tree City USA? (Y/N)	**Zoning Ordinance**	Front Yard Setbacks (FYS)	FYS + SYS + RYS
	Years as a Tree City USA		Side Yard Setbacks (SYS)	
	Public/ Street Protection Tree Ordinance? (Y/N)		Rear Yard Setback (RYS)	
	Specific Tree Ordinance with number of specific types of trees protected		Max Height	
	Applicable areas (public, private, both, or none)		Min Lot Area	
Landscape Ordinance	Residential Landscape Requirements (Y/N)		Min Lot Width	
Water Rates and Ordinance	Water Efficient Landscape Policy (Y/N)		Max Lot Coverage	
	Monthly Water Cost		Max FAR (Floor Area Ratio)	

US Census Block Group						
Population	Population Density	Median Household Income		Population Under Poverty Level		
White	African American	Asian	Hawaiian	American Indian	Hispanic	Other
Block Group area	Average family size	Number of house units	Number of vacant houses	Household occupied by owner	Household occupied by renter	

Parcel from Los Angeles County Assessor’s Office				
Area of Parcel	Land Value (per m^2)	Size of building	Year built	Year of rebuild

Environmental Parameters		
Elevation	Average temperature	Distance to ocean

THE CALCULATION OF TVDI BASED ON THE COMPOSITE TIME OF PIXEL AND DROUGHT ANALYSIS

*Lingkui Meng [a], Jiyuan Li [a], *, Zidan Chen [b], Wenjun Xi e [a], Deqing Chen [b], Hongwei Duan [a]*

[a] School of Remote Sensing and Information Engineering , Wuhan University, 129 Luoyu Road, Wuhan 430079, P. R. China – lkmeng@whu.edu.cn, lijiyuan_521@163.com, whuxwj@gmail.com
[b] Water Resources Information Center, MWR, Beijing 100053, P.R. China – (zdchen, chendq)@mwr.gov.cn

KEY WORDS: Remote Sensing, Precipitation, Soil Moisture, Temperature-Vegetation Dryness Index, Drought, Time, Composite

ABSTRACT:

Temperature-Vegetation Dryness Index (TVDI) is one of the agriculture drought indexes. This paper presents a data composite method which improves the calculation of TVDI through taking the time of pixel into consideration, and the adaptability of TVDI in drought assessment has also enhanced significantly. First, the Normalized Difference Vegetation Index (NDVI) data series are composed by using maximum value composite (MVC) method, and the Land Surface Temperature (LST) data series are composed to construct NDVI-Ts feature space. Then, the wet and dry sides of NDVI-Ts feature space are fixed by a number of ways to build new TVDI, and we note it as T-TVDI, for assessing the drought condition. To verify our proposed method, TVDI in time scale of ten-days is established for Chongqing region in China, and the results coincide with the actual situation. Finally, the T-TVDI and TVDI of Chongqing region in 2008 are calculated and compared. The correlations of them and Soil Moisture are analyzed as well as Precipitation. It shows that T-TVDI has the advantages of stability and high accuracy in the short term. It is feasible to use T-TVDI to evaluate drought in proper region and reasonable crop growth period.

1. INTRODUCTION

The water stress indicator (TVDI) proposed from NDVI-Ts feature space reflecting the surface soil moisture well, especially in large areas of vegetation coverage, is used to assess the drought condition locally.

In the NDVI-Ts feature space, the expression of TVDI calculation is as follows:

$$\mathrm{TVDI} = \frac{T_s - T_{s-\min}}{\mathrm{T}_{\text{s-max}} - T_{s-\min}} \tag{1}$$

Where $T_{s-\min}$ = minimum of land surface temperature When the NDVI is equal to a particular value
$T_{s-\max}$ = maximum of land surface temperature When the NDVI is equal to a particular value

We can see from the above equation (1), NDVI and LST data are the bases for the TVDI calculation. The NDVI is calculated from near-infrared and red bands of the multispectral image, and the LST data is able to get by split-window algorithm. Drought is a complex phenomenon, the formation and development of its strength go through a process in gradual accumulation which is so slow that it is difficult to detect during beginning period. In the process of TVDI calculation, NDVI and LST data are composed by time-series data accumulated in a certain observing period. The rationality of composite algorithm is directly related to the quality and accuracy of TVDI.

There are a number of conventional composite methods of NDVI and LST, such as MVC, CV-MVC and BRDF (These methods will be gave explanatory notes in what follows). Whichever method is used, composites of NDVI and LST data are carried out separately. Although TVDI is the statistical values of the period, there are no real spatial and temporal consistency of geography and time between the vegetation condition and surface temperature which reduces the calculation accuracy and evaluation effects of drought. This paper analyzes the principles and characteristics of TVDI firstly, and then composes the NDVI and LST based on the time of pixel; finally carry out T-TVDI with a good evaluation of drought in the region.

2. THE CALCULATION OF T-TVDI

2.1 Introduction to the theory of TVDI

Temperature Vegetation Drought Index uses the relationship between the surface temperature and soil moisture (relative soil moisture) to reflect degree of drought.

TVDI comprehensively considers relation and changes between the NDVI and LST. From the physical mechanism, it is certainly hysteretic to using NDVI as Water Stress Index. Temperature is time-sensitive as indicator of water stress, but is apt to be affected by vegetation coverage when using temperature method to monitor soil moisture. TVDI integrates vegetation indices and surface temperature to monitor soil moisture with the ability to composite information of visible, near infrared and thermal infrared bands of light, so it has a wider range of applicability. Figure 1 shows the simple NDVI-Ts feature space.

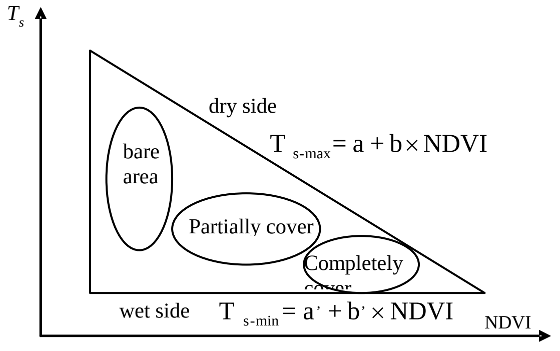

T_s
dry side
$T_{s\text{-max}} = a + b \times NDVI$
bare area
Partially cover
Completely cover
wet side
$T_{s\text{-min}} = a' + b' \times NDVI$
NDVI

Figure 1. Simple NDVI-Ts feature space

Temperature Vegetation Drought Index uses the relationship between the surface temperature and soil moisture (Relative soil moisture) to reflect degree of drought.

TVDI comprehensively considers relation and changes between the NDVI and LST. From the physical mechanism, it is certainly hysteretic to using NDVI as Water Stress Index. Temperature is time-sensitive as indicator of water stress, but is apt to be affected by vegetation coverage when using temperature method to monitor soil moisture. TVDI integrates vegetation indices and surface temperature to monitor soil moisture with the ability to composite information of visible, near infrared and thermal infrared bands of light, so it has a wider range of applicability.

The NDVI-Ts feature space is able to be simplified to a triangle,

$T^{s-\max}$ and $T^{s-\min}$ are linearly fit at the same time, and result is as follow:

$$\begin{aligned} T_{s\text{-max}} &= a + b \times NDVI \\ T_{s\text{-min}} &= a' + b' \times NDVI \end{aligned} \qquad (2)$$

$$TVDI = \frac{[T_s - (a' + b' \times NDVI)]}{[(a + b \times NDVI) - (a' + b' \times NDVI)]} \qquad (3)$$

where a，b，a$'$，b$'$ = coefficients of dry side and wet side fitting equation

The range of TVDI is [0, 1], TVDI = 1 on the dry side, TVDI = 0 on the wet side. The greater the value of TVDI, the lower the soil moisture, and the higher the level of drought will be.

2.2 The composite of NDVI and LST

The purpose of composite is to choose the best observational data. This method should be able to ensure spatial and temporal consistency of vegetation index values. At present, there are following common ways of composite:

MVC(*maximum value composite***):** Select the maximum NDVI value of observed pixel as the vegetation value in composite period. It is the best method for information with no atmospheric correction because of minimizing the selection of cloudy and heavy aerosol pixels.

CV-MVC(*constraint view angle maximum value composite***):** The CV-MVC compares the two highest NDVI values and selects the observation closest to nadir view to represent the 16-day composite cycle. This helps to reduce spatial and temporal discontinuities in the composite product.

BRDF(*bidirectional reflectance distribution function***):** The BRDF scheme is a more elaborate and constrained technique in which all bidirectional reflectance observations, of acceptable quality, are utilized to interpolate to their nadir-equivalent band reflectance values from which the VI is computed and produced.

Because composites of NDVI and LST data are carried out separately by the methods above, although TVDI is the statistical values of the period, there are no real spatial and temporal consistency of geography and time between the vegetation condition and surface temperature which reduces the calculation accuracy and evaluation effects of drought.

In this paper, MVC and CV-MVC method are applied in accordance with the following priority sequence of composite methods:

1. CV-MVC within limited perspective: If in the period of composition, the number of days with no cloud is less than 30%, and more than 2, choose the maximum of two vegetation indexes within smallest perspective.
2. Calculating the vegetation index directly: If there is only one day without cloud, choose the vegetation index of the day directly.
3. MVC: If all days observed is not sunny, choose the maximum of al the vegetation index values in the composite period.

While composing the NDVI, the time information obtained of every composite pixel in ten-days are stored in a band. Then choose the corresponding pixel to compose LST from LST data series referring to the time raster chart.
The specific algorithm is stated as follows:

```
For x = 0 to Xsize
   For y=0 to Ysize
      For i = 0 to dayCount
      If the count radio of sunny day is under 30%:
         If the count of sunny equal one
            NDVComposite[x,y] = NDVI[i]
            Day[i] = the date of NDVI[i]
         Else
            NDVComposite[x,y] get the maxNDVI
            day[i] = the date of maxNDVI
        Endif
      Else
         NDVComposite[x,y] = maxNDVI of two values
within smallest perspective
         day[i] = the date of maxNDVI
    Endif
     EndFor
     dateComposite[x,y] = day[i]×100
     LSTComposite[x,y] = LST[i] of day[i]
   EndFor
```

```
  EndFor
EndFor
```

NDVComposite[x,y], dateComposite[x,y] and LSTComposite[x,y] are separately composite data of NDVI, date image and LST. Day[i], LST[i] and NDVI[i] respect separately pixels of date image, LST and NDVI series in ten-days. The figure 2 shows the visualization of this method.

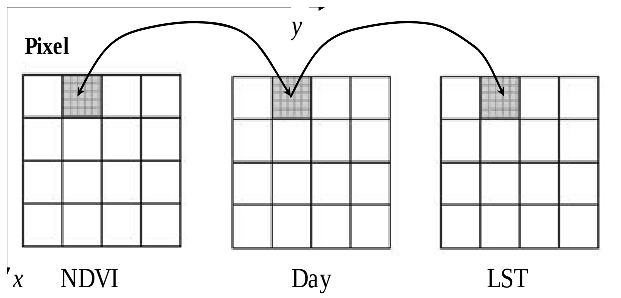


Figure 2. Composite of LST referring to the time raster chart

2.3 Fixing of wet and dry sides

The maximum and minimum temperature corresponding to NDVI as the thermal and cold edges of NDVI-Ts feature space are obtained by method of maximum and minimum. The thermal and cold edges obtained by the method are direct-viewing, clear and easy-to-linear fitting with rapidly handling and showing. But sometimes the pixels of the sides are scattering and shapes are irregular. Then we will consider using the scatter plot to fix the thermal and cold edges.

Fix the wet side referring to the average of cold edge (base on the cloud removing), water surface temperature and average in the same period of many years in the region. According to the histogram of image, remove the pixels at the end of the thermal edge with fixing the boundaries at 1% of the total number of pixels, and obtain the parameters of dry side by linear fit. Due to irregular thermal edge, the parameters may be not good. Then, we set up a standard feature space as background reference picture according to the relationship between wet and dry sides. It can be used to adjust and fix the parameters of wet and dry sides. Standard feature space previously divides the LST in appropriate interval in feature space, and then establishes the equations of all corresponding dry sides. This method remains an error. But with the smaller the intervals, the error is smaller. The parameters of dry side are replaced by corresponding contour in reference picture of standard feature space as following figure.

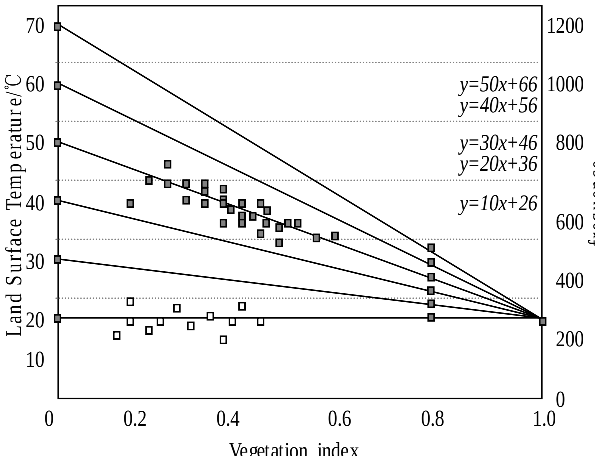


Figure 3. Picture of standard NDVI-Ts feature space

3. EXPERIMENT OF T-TVDI CALCULATION

In this paper, Chongqing region in China is selected as the experimental area where the typical southern dry farming areas are with frequently occurrence of drought.

The data using in the paper include remote sensing data and measured data. Remote sensing data with 1-km resolution is provided by MODIS (Moderate Resolution Imaging Spectroradiometer), and measured data (soil moisture and rainfall) in the experiment are from all the hydrological stations located in various parts of Chongqing region (Figure 8 & Figure 9). Soil moisture data are from calculation of measured data in the depth of 10cm, 20cm, and 50cm of soil.

First of all, calculate the LST through split-window algorithm using NDVI dataset. Then, two methods of TVDI calculation are adopted. One is the method using MVC to compose NDVI and LST series of 10 day. The other is the methods composing NDVI and LST data by the way mentioned above, extracting the minimum and maximum LST in different climatic zones and every ten-days under the different conditions of NDVI with a small step size.

The NDVI-Ts feature spaces from two methods are as follow (data of ten days are acquired in mid-May 2008):

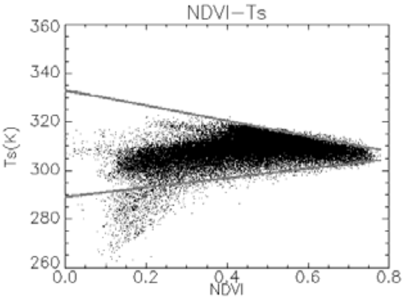


Figure 4. NDVI-Ts feature space based on the Composite Time of pixel

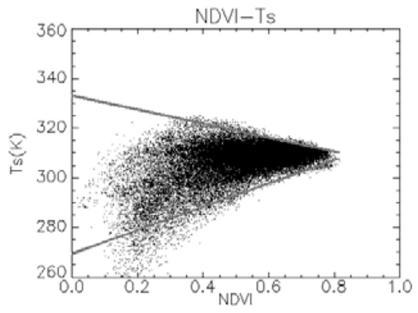


Figure 5. NDVI-Ts feature space based on MVC

As can be seen from the figures above, angle of triangle in the second feature space is bigger than the first one, because most of LST concentrate in high-temperature region, especially in the wet side. It is difficult to fix the boundaries of LST. Slope of dry side is less than 0, indicating that LST is decreasing with the vegetation coverage increase. On the contrary, the slope of wet side is greater than 0, indicating that there is increasing trend for LST while vegetation coverage decreases.

According to scatter gram of the first feature space, fit the wet and dry sides as follows:

$$\begin{aligned} T_{\text{s-max}} &= 5.6 + 34.5 \times \text{NDVI} \\ T_{\text{s-min}} &= 13.5 - 40.5 \times \text{NDVI} \end{aligned} \tag{4}$$

Finally the T-TVDI result from calculation shows that heavy-dry appeared in the middle, southeast and small part of east of Chongqing region. RSM(*relative soil moisture*) of northern areas is suitable for the growth of crops, and light-dry appeared in south-western regions. From the experimental results, detection is generally consistent with the reality.

The figure 6 shows the result of T-TVDI calculation:

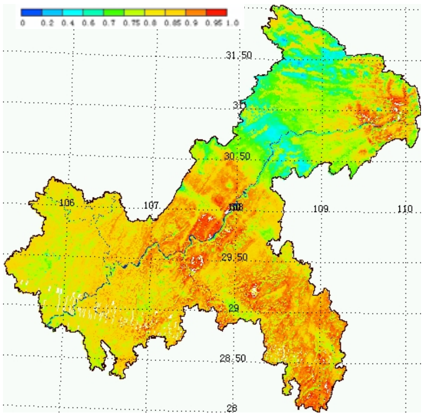


Figure 6. T-TVDI of Chongqing in mid-May 2008

4. DROUGHT ANALYSIS

Soil moisture and rainfall are two normal indexes to evaluate soil drought situation, they can better reflect the intensity and duration of drought situation. Soil moisture is the field soil moisture and the corresponding crop water status, as one of the indicators of drought situation. It plays an important role in the exchange between the water and energy exchange of surface and atmospheric. Relative Soil Moisture can be offset the impact of soil texture to some extent. Figure 7 shows the relative soil moisture and stations of Chongqing in mid-May 2008. As can be seen, spatial distribution of relative soil moisture is uneven that in the north is less than in the south.

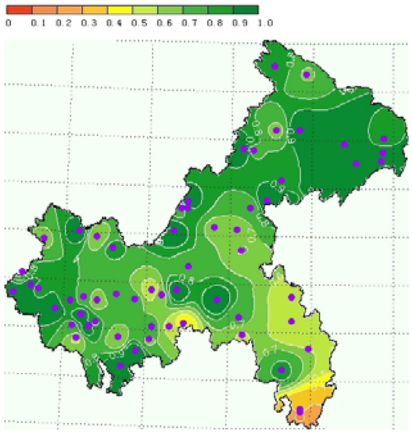


Figure 7. Relative soil moisture and stations of Chongqing in mid-May 2008

Precipitation is the main factors affecting drought situation, the amount of precipitation reflects the weather conditions basically. Standard Precipitation Index (SPI) is easy to compute and access to necessary information, and because of not involving a specific mechanism of drought situation, the space-time adaptability is more suitable. The figure 8 shows standard precipitation index and stations of Chongqing in mid-May 2008. The graph shows that the rainfall in most parts of the province ten days is less than normal. The west of Chongqing seriously lack of rainfall.

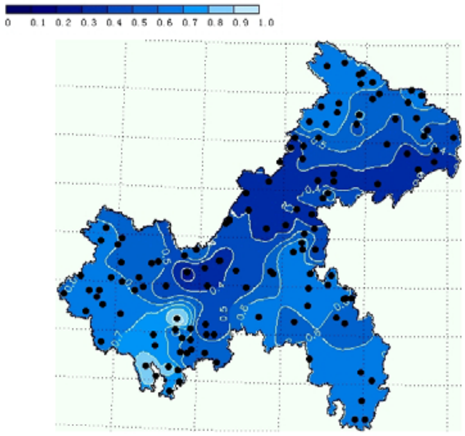


Figure 8. Standard precipitation index and stations of Chongqing in mid-May 2008

The scatter plot of T-TVDI and RSM from Chongqing in mid-May 2008 is given in the following, as well as the scatter plot of T-TVDI and RSM.

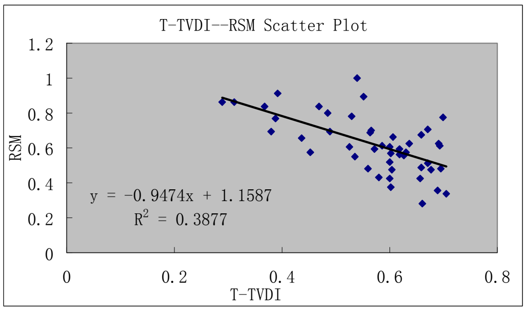


Figure 9. Scatter plot of T-TVDI and RSM in mid-May 2008 of Chongqing

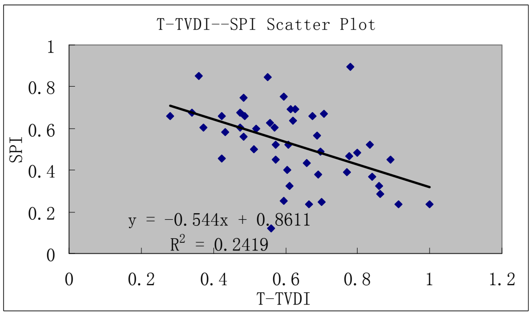


Figure 10. Scatter plot of T-TVDI and SPI in mid-May 2008 of Chongqing

As can be seen from figure 9 and figure 10, it is obvious that scattered points distribute nearby on both sides of a straight line and show a good linear distribution that indicating a good linear model. The correlation arrives 0.62 and the linear model has passed the significance level F test with reliability of 0.01. At the same time, T-TVDI also has a good correlation with SPI, indicating T-TVDI has a good effect in evaluating the situation of the regional drought.

To carry out further analysis and evaluation of drought situation, this paper uses two kinds of TVDI data calculating from MODIS data covering Chongqing region in every ten-days to analyze the adaptive real-time indices. The analysis of adaptive real-time indices is respect that doing correlation analysis between the index and RSM in the ten-days. It emphasizes adaptability of index at the time to RSM, and provides a basis for choosing real-time index for drought monitoring. As a reference, two types of indexes and the precipitation index are analyzed in the same way.

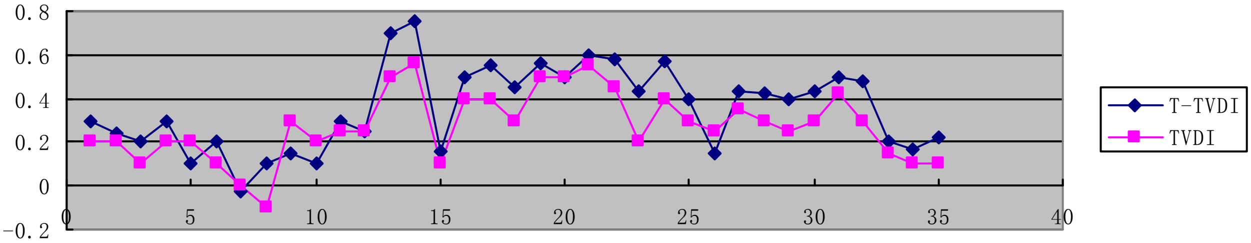


Figure 11. Correlations gram between RSM and two indexes of Chongqing in 2008

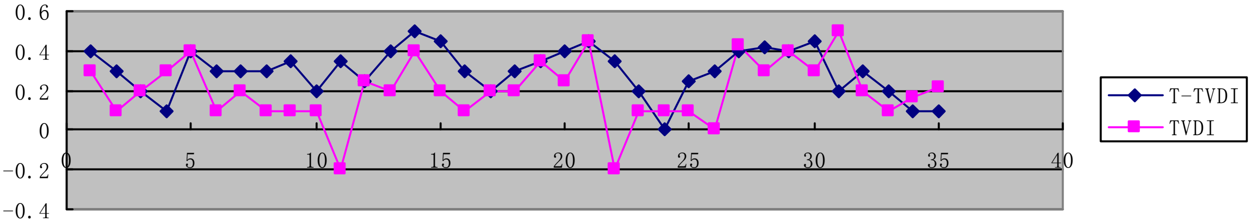


Figure 12. Correlations gram between SPI and two indexes of Chongqing in 2008

The figure 11 shows that in the winter the two indexes are not too high correlating with RSM, this is because of the low vegetation cover in winter, the TVDI sensitivity is not high for the ground. In the crop growing period, the correlation is increasing with the crop growing,. The correlation of T-TVDI and RSM is better than TVDI.

The figure 12 shows that T-TVDI and TVDI are not good correlating with SPI, but the trend of T-TVDI was relatively stable, so it is an acceptable range of fluctuation for the evaluation of drought situation.

5. CONCLUSION

This paper uses two different composite methods to obtain NDVI and Ts in ten-day scales from MODIS data in 2008, and then constructs the NDVI-Ts feature space. On the basis of this NDVI-Ts feature space, TVDI and T-TDVI that reflecting the drought situation are created and used to evaluate the drought situation. Comparing with the surface soil moisture data and precipitation data getting from hydrological observation sites , the result shows that , TVDI basing on the composite time of pixel can better reflects the correlation with soil moisture, more suitable for monitoring and evaluating drought situation.

TVDI only adapts to a relatively high vegetation coverage area, the higher of vegetation coverage, and the greater of the correlation with the RSM. Because of the significant difference of wet and dry side in different regions, the evaluated range should not be too big by using TVDI.

Because satellite view affects the amount of information received by sensors, thus vegetation growth status shows in images further. This study does not take into account the impact of satellite view to the NDVI and LST.

6. ACKNOWLEDGEMENT

The work presented in this paper is supported by National 948 project of China (200610) which is named Hydraulic information service of grid system, armed at introducing advanced foreign technology. Thanks for the Leadership and colleagues in Water Information Center of The Ministry of Water Resources in China who provided the MODIS L1B data, thanks all the same to the Leadership and colleagues in Beijing Golden-Water Info-Tech Ltd. who supply the necessary experimental equipment and the environment.

7. REFERENCES

Kogan F N. 1995. Application of vegetation index and brightness temperature for drought detection. Adv. Space Res. 15(11), pp. 91-100.

Cuizhen Wang, Jiaguo Qi，Susan Moran，Robin Marsett. 2004. Soil moisture estimation in a semiarid rangeland using ERS-2 and TM imagery Preliminary results. Remote Sensing of Environment，90, pp. 178-189.

Sandholt et al. 2002. Rasmussen K, Andersen J.A Simple Interpretation of the Surface Temperature/Vegetation Index Space for Assessment of Surface Moisture Status. Remote Sensing of Environment. 79, pp. 213-224.

Njoku, E.G. Li, L. 1982. Retrieval of Lnad Surface Parameters Using passive microwave measurements at 6-18GHz. IEEE Transanctions on Geoscience and Remote Sensing， 20(4), pp. 468-475

England AW, Galantowkz J F. 1992. Seheretter.The Radiobrightness Themral Inertia Measure of Soil Moistuer. IEEE Trans Geosci. Remote Sensing, 30(1), pp. 132-139.

Carlson T N, et al. 1990. Remote sensing estimation of Soil moisture availability and fractional vegetation cover for agricultural fields. Agri.& Fore. Meteo. 52, pp. 45-69.

Nemani, R.R., et al. 1993. Developing satellite derived estimates of surface moisture status. Journal of Applied Meteorology, 32, pp. 548-557.

Lingkui Meng, Liang Tao, Jiyuan Li and Chunxiang Wang. A System for Automatic Processing of MODIS L1B Data. Proceeding of the 8th International Symposium on Spatial Accuracy Assessment in National Resources and Environmental Sciences, Shanghai, P.R.China, June 25-27, 2008, pp.335-343.

Gengming Jiang, Zheng Niu, Weili Ruan, Changyao Wang,. Cloud-free composition of MODIS data and algorithm realization. Remote sensing for land and resource, No.2, Jun, 2004, pp. 11-15.

Members of the MODIS Characterization Support Team, MODIS Level 1B Product User’s Guide, NASA/Goddard Space Flight Center, December 1,2003, pp. 15-50.

OBJECT-BASED IMAGE CLASSIFICATION UTILIZING BACKGROUND KNOWLEDGE: A CASE STUDY OF LAND USE CLASSIFICATION

T. Zhang, X.M.Yang, C.H.Zhou, F.Z.Su, J.M.Gong, Y.Y.Du

State Key Laboratory of Resources and Environmental Information System, IGSNRR, CAS, Beijing, China –
(zhangt, yangxm, zhouch, sufz, gongjm, duyy)@lreis.ac.cn

KEY WORDS: object-based classification, background knowledge, change detection, land use, CBERS

ABSTRACT:

Object-based image classification and information extraction approaches are rapidly developing from the beginning of this century, but the automatic procedure for land use mapping are still problematic facing with land use complexity. This paper presents a method of incorporating background knowledge into the object-based image classification procedure, intending to improve the classification accuracy and boundary consistency of land use data. Two forms of knowledge are used in this paper: the expert interpreted land use polygons and the land use change rules. The idea of this paper is tested on the platform of Definiens Developer 7 (trial version). The proposed approach mainly contains three parts: segmentation supported by land-use thematic layer, classification and change detection supported by land-use change rules. The experiment result shows that the proposed procedures have good potential for automatic land use mapping.

1. INTRODUCTION

Object-based remote sensing imagery classification and information extraction approach have been widely tested and proved to be a promising method for automatic image classification and interpretation. The object-based approach has two main advantages over the traditional pixel-based classification method: one is reducing the salt-pepper effect, and the other is utilizing multi features extracted from image objects to improve the classification accuracy. These features include the spectral features, shape features, texture features, spatial relational features and semantic relations multi-level image object hierarchy and class hierarchy.

Many efforts have been endeavoured into the application of object-based approach into image classification and information extraction, including vegetation classification using very high resolution satellite imagery or airborne imagery (Yu *et al.*, 2006; Mathieu *et al.*, 2007; Mallinis *et al.*, 2008), mangrove and tree mortality mapping from IKONOS imagery (Wang *et al.*, 2004; Guo *et al.*, 2007), vehicles detection from high resolution aerial photography (Holt *et al.*, 2009), burn area and fire type mapping from IKONOS imagery and NOAA-AVHRR data (Gitas *et al.*, 2004; Mitri and Gitas, 2006).

In the field of land use and land cover mapping, as the land use and land cover pattern of certain region (like the coastal region) change rapidly and fiercely, it is of great importance to develop automatic procedures to do the time consuming land use and land cover mapping work. The object-based approaches show great potential of automatic land information extraction in imagery, and several attempts have been made for land use mapping and change detection (Walter, 2004; Stow *et al.*, 2007; Rahman and Saha, 2008), but it still has a long way to go before reaching the goal of automatic procedures of land use mapping and change detection.

When using the object-based approach for land use mapping, it is hard to develop a consistent threshold of multi features for certain land use type, and features may be time variant for certain land use type. Take the land use type “cropland” as an example, the spectral features are probably quite different before and after the crop harvested. In addition, land parcel boundaries are generally vague in image, and it is hard to get consistent segmentation result for certain land parcel. Furthermore, the same land parcel may have different boundaries, because of the difference of the radiation characteristics among sensors and the time variation influence.

Intending to improve the object-based image classification result for land use mapping, this paper introduces an idea of incorporating background knowledge into the object-based land use classification procedures. The knowledge presented in this paper mainly refers to two kinds: one is the knowledge persisted by expert in the process of manual interpretation, the other is the land use change pattern knowledge.

Traditionally, land use was mapped through field investigation. With the remote sensing technique developing, land use can be mapped from remotely sensed imagery by the means of manual interpretation by domain experts. Experience shows that the manually interpretation by trained interpreters with field investigation experience can obtain good classification accuracy. There is certain knowledge an interpreter possess for image interpreting, including spectral, texture, shape, spatial relations, semantic relations, and other features unknown. It is justifiable to think that part of the unutterable features and rules remains in the land use data which the interpreter produced. This research is intend to seek for a procedure to utilize the interpreted land use as background knowledge to aid the object-based classification and information extraction, and it is the first kind of knowledge used in this research.

The second kind of knowledge refers to the land use change pattern. Land use types do not change randomly in space, but they change following certain rules in certain regions. For example, “built-up land” is not inclined to change to other land use types in a temporal scale of year. Land use change rules may vary from region to region, and the techniques to extract these rules from data sources are still under development. We employ simple land use change rules in this study, just for the illustration of the procedure and framework of knowledge aided object-based automatic land use mapping.

2. DATASETS

2.1 Study area

The study area is located in west coastal area of Chinese Pearl River Estuary (figure 1). It lies in the west of Hongkong and in the administration zone of Zhongshan city. Rapid urbanization undergoes in this area since the China's reform and opening up.

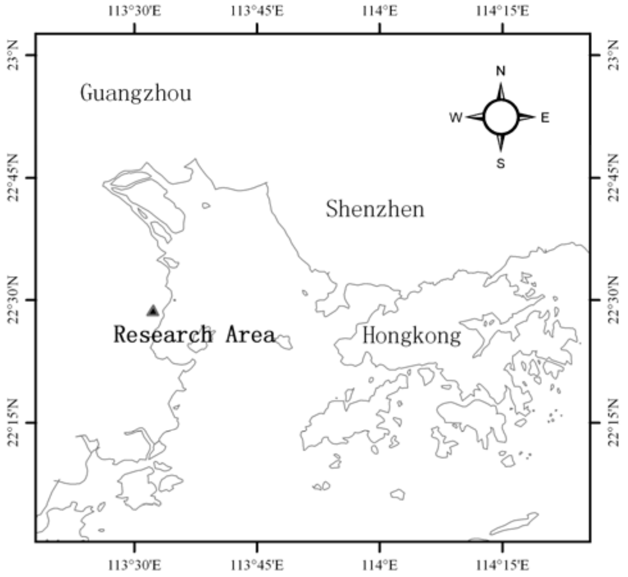


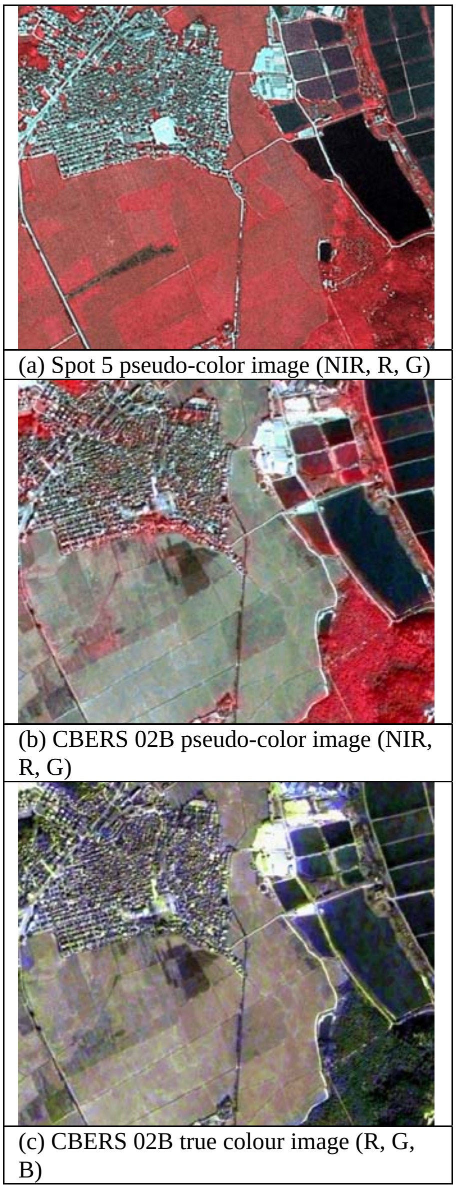
Figure 1. Location of the study area

2.2 Imagery

Two image sources are selected for this research: one is the SPOT 5 image and the other is CBERS 02B image. The SPOT 5 image is a fusion image of SPOT 5 panchromatic (2.5m) and multispectral image (10m). The fusion algorithm is Pansharpen implemented in PCI software package, and the final image resolution is 2.5m. Three bands are included in the SPOT 5 fusion image, NIR, Red, and Green. The pseudo colour combined image (NIR, R, and G) is shown in figure 2(a). The imaging time is October 23, 2003, and SPOT image is the source image of the old land use polygon obtained by expert manual interpretation.

The CBERS image is a fusion image of CBERS 02B High Resolution (HR) (2.36m) and multispectral image (20m). The fusion algorithm is also Pansharpen. The final image resolution is 2.5m. Four bands are included in the CBERS fusion image, namely NIR, Red, Greed, and Blue. The combined colour image is shown in figure 2(b, c). The scene date is January 5, 2009. The CBERS imagery is distributed by China Centre for Resources Satellite Data & Application. The CBERS image is used to generate the new land use classification.

By manually interpreting images in figure 2, one can see that SPOT image have better detailed texture information than the CBERS image, one possible reason is that too much resolution difference between the CBERS panchromatic (2.36) and multispectral image (20m). The relatively lower image quality of the CBERS image place some challenges for the effective land use extraction.



(a) Spot 5 pseudo-color image (NIR, R, G)

(b) CBERS 02B pseudo-color image (NIR, R, G)

(c) CBERS 02B true colour image (R, G, B)

Figure 2. Remote sensing imagery of the study area

2.3 Land use polygon data

The land use thematic data is from the Chinese Coast and Island Remote Sensing Investigation Project (Sun, 2008). The land use is mapped through manually image interpretation from SPOT 5 imagery and other ancillary data, the scale is about 1:50 000. The verification result from field investigation shows that this dataset have more than 91% classification accuracy of level 2 land use type. Five land use type appear in the study area, namely cropland, shrub and grassland, forest land, built-up land, and aquaculture (figure 3).

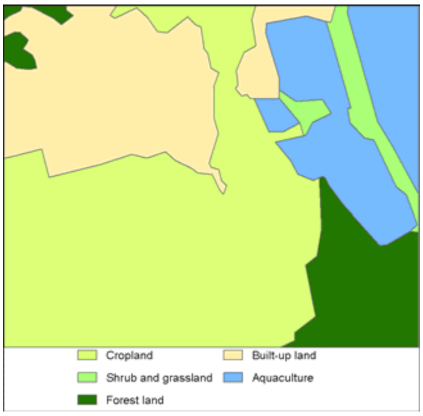


Figure 3. Land use data of the study area

3. METHODS

3.1 Background knowledge aided image segmentation

The knowledge aided image segmentation methods is achieved by incorporating land use thematic layer into image segmentation, and force the segmented image object to have the edge which the land use thematic layer delineated.

This research employs multiresolution segmentation algorithm (Baatz and Schäpe, 2000) implemented by Definiens (Definiens, 2007) to do the segmentation. Through the multiresolution segmentation, individual pixels are perceived as the initial regions, which are sequentially merged pairwise into larger ones with the intent of minimizing the heterogeneity of the resulting objects. The sequence of the merging objects, as well the size and shape of the resulting objects, are empirically determined by the user. Initially, the layers, as well as their weight, the parameters for homogeneity/heterogeneity, and the crucial scale parameters are specified by the analyst (Benz *et al.*, 2004).

The homogeneity criterion is calculated as a weighted combination of color and shape properties of both the initial and the resulting image objects of the intended merging. The color homogeneity is based on the standard deviation of the spectral colors. The shape homogeneity is based on the deviation of a compact or a smooth shape. The computation of homogeneity criterion *f* is illustrated in equation(1):

$$f = (1-w_1)\sigma_{color} + w_1((1-w_2)\sigma_{smooth} + w_2\sigma_{compact}) \qquad (1)$$

Where σ_{color} refers to the color homogeneity, σ_{smooth} refers to the smoothness shape homogeneity, and $\sigma_{compact}$ refers to the compactness shape homogeneity.

Segmentation on several scales with different scale parameters can be carried out leading to the formation of a hierarchical network of objects. This procedure is constrained so that spatial shape of objects in one level fits hierarchically into objects of another level enabling consideration of sub-objects and super-objects and their relationships in the classification step.

As in the introduction section explained, expert knowledge is contained in manually interpreted land use data. This research uses the land use polygon data as a thematic layer to help delineate proper land parcels. Detailed procedures are in figure 4. Firstly, image is segmented at a coarse level (level 1) to delineate the old land use type which is indicated by the background thematic layer. Then, based on the coarse level objects, more detailed land parcels at a fine level (level 2) are segmented. Level 2 objects are copied and they make a new level at level 3. Level 3 objects are the sub-objects of level 2 objects, but in fact they hold the same image pixels. In this research, level 3 objects are used for change detection.

The good segmentation result and the definite land use attribute the level 1 objects hold will greatly benefit the classification procedure in section 3.2.

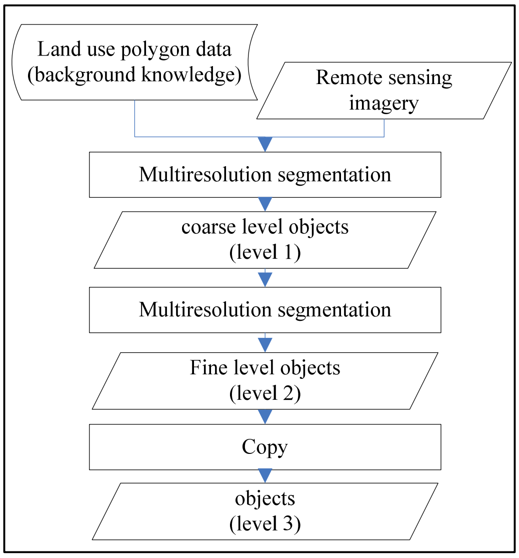

Land use polygon data (background knowledge)
Remote sensing imagery
Multiresolution segmentation
coarse level objects (level 1)
Multiresolution segmentation
Fine level objects (level 2)
Copy
objects (level 3)

Figure 4. Image segmentation procedure utilizing background knowledge.

3.2 Knowledge aided image classification

In this approach, image objects are classified not only by the traditional multi features, but also by land use change rules.

First, the level 1 image objects land use are classified based on the background land use thematic layer attribute "land use type", and the level 1 image objects classification represents the old land use type. Then, the level 2 image objects are classified with the land use change rules and multi features. The land use change rules used in this research is described in table 1, and the features used for classification is shown in table 2. The feature brightness and vegetation index in table 2 refer to the object mean value.

Table 1. Description of land use change rules.

Old land use	*New land use*	*Condition*
Cropland	Built-up land Aquaculture Cropland	BT > 50, VI < 0.1, Entropy >6 BT < 45, VI <0.05 Other condition
Shrub and grassland	Built-up land Aquaculture Shrub and grassland	BT > 50, VI < 0.1, Entropy >6 BT < 45, VI <0.05 Other condition
Forest land	Built-up land Aquaculture Forest land	BT > 50, VI < 0.1, Entropy >6 BT < 45, VI <0.05 Other condition
Aquaculture	Built-up land Shrub and grassland Aquaculture	BT > 50, VI < 0.1, Entropy >6 NDVI > 0.15 Other condition
Built-up land	Built-up land	All condition

Table 2. Description of the various object features used in the study.

Abbreviation of the feature	*Name of the feature*	*Computation method (or reference)*
BT	Brightness	(Red + Greed + Blue) / 3
VI	Vegetation index	(Nir – Red) / (Nir + Red)
Entropy	GLCM entropy	Refers to (Definiens, 2007)

3.3 Change detection

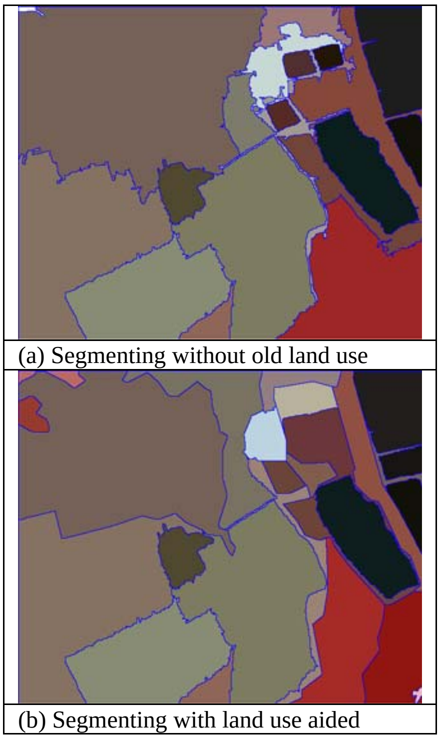
Change detection is achieved by comparing the classification of objects at level 1 and that of level 2. Classification at level 1 refers to the Old land use, and classification at level 2 refers to the new land use. If the land use type of a level 2 object is different with the land use type of its super object at level 1, then the sub-object at level 3 of this image object is classified as “changed”, otherwise, the sub-object at level 3 is classified as “unchanged”.

4. EXPERIMENTS AND RESULTS

4.1 Knowledge aided image segmentation

4.1.1 Comparison of segmentations without and with land use polygons

To demonstrate the effect of segmentation method supported by land use thematic layer, a test is conducted to compare the segmenting result with or without the old land use polygons. Segmentations are performed at CBERS layer 1-4, with scale parameter 100. Result is shown in figure 5. The segmentation result without land use delineate land parcels too fine boundaries, and it does not in consistent with the map generalization rule; but the segmentation result supported by land use polygon preserves the old land parcel boundaries very well, and some possible type changed land parcel are delineated as well without conflicting with the old land boundary.



(a) Segmenting without old land use

(b) Segmenting with land use aided

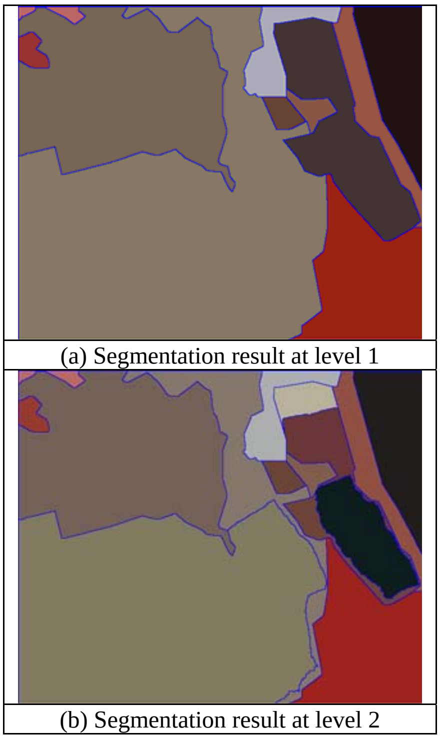
Figure 5. Comparison of segmenting result of CBERS image without and with land use polygons.

4.1.2 Segmentation result at multi level

As described in the method section 3.1, the coarse level objects (level 1) are segmented first, then based on the old land use polygon, fine level objects (level 2) are segmented. Level 3 objects are used for change detection, and they are copied from level 2 objects. The detailed parameters used for segmentation is shown in table 3. The segmentation result is shown in figure 6. The boundaries of level 1 object are consistent with the land use map shown in figure 3.

Table 3. Parameters used for segmenting four image layer of the multi-scale object

level	**scale**	**bands**	**thematic**	**shape**	**compactness**
Level 1	500	Band 1-4	yes	0.1	0.5
Level 2	200	Band 1-4	no	0.1	0.5
Level 3	200	Band 1-4	no	0.1	0.5



(a) Segmentation result at level 1

(b) Segmentation result at level 2

Figure 6. Segmentation result at level 1 and level 2.

4.2 Knowledge aided image classification

Land use change rules are employed in the process of classification. Objects on level 2 are classified according to the rules described in table 2. The features used for classification are described in table 1. Take the classification procedure of land use type “aquaculture” for example, the detailed classification tree is shown in figure 7. Classification procedure of other land use types is about the same of type “aquaculture”, only the land use change rules varied. The final classification result is shown in figure 8.

4.3 Change detection

By comparing the land use type of level 2 objects (new land use) with the land use type of their super-object (old land use), the land use change status for each land parcel in level 3 are defined. The final result is as figure 9 shows, one land parcel changed the land use type.

By manually interpreting the image in figure 1, the spectral feature of land use type “cropland” in the CBERS image is quite different from that in the SPOT image. Generally used object-based change detection approaches are hard to tell this kind of difference, normally they result in false change detection in this research condition. The pixel-based change detection approach are not suitable for this condition, because the radioactive characteristics of the sensors (SPOT 5 and CBERS 02B) are quite different.

The result of the automatic procedure of our approach is quite consistent with manual interpretation. As described in section 2, the image quality of SPOT 5 image is better than that of CBERS 02B image, but high price of SPOT image limit the access of research community for such data. The procedure proposed in this paper also shows good potential for classification and change detection for less quality images.

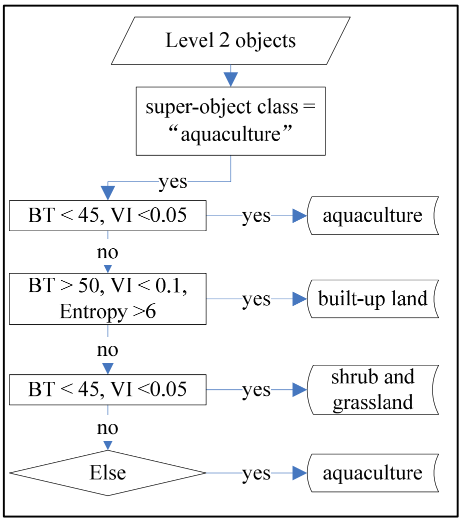


Figure 7. Classification tree for land use type “aquaculture”.

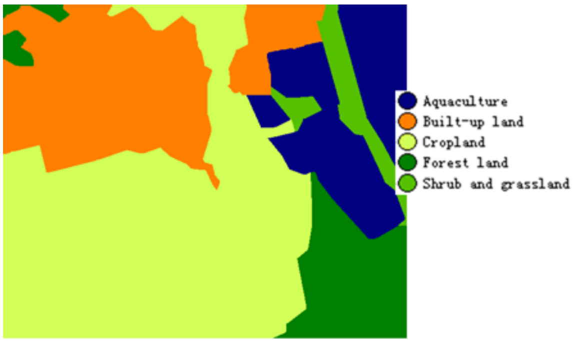


Figure 8. Classification result for level 2 objects.

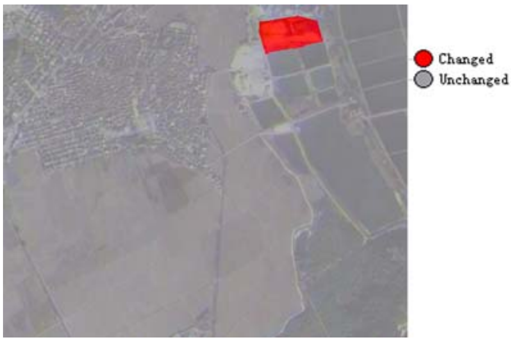


Figure 9. Change detection result.

5. CONCLUSION

Object-based approaches are promising for automatic land mapping, but there is still a long way to go to reach the goal automatic. This paper introduced an object-based approach for land mapping and change detection by utilizing background knowledge. In the case study, the old manually interpreted land use thematic layer and user defined expert land change rules are employed as background knowledge, and aiding the segmentation, classification, and change detection in the proposed procedures. The experiment result shows that there is good potential for the proposed knowledge aided object-based classification approach to become automatic procedure for land mapping.

However, it has to acknowledge that the framework proposed in this paper is only a prototype, and the exact knowledge used in this paper is problematic when applying for other regions. Further work has to be done to reach full automatic procedure for land mapping by knowledge aided object-based approach. Firstly, proper knowledge discovering techniques have to be developed to extract effective knowledge, either in the form of rules or patterns, which can be used for automatic land classification. Secondly, the optimal feature space for the classification of certain land type need to be discovered. Thirdly, knowledge and feature variation among different regions need to be considered for further automatic procedure for large area (regional) land mapping.

ACKNOWLEDGMENTS

The research is supported by “863” Project of China (No.2008AA121706, No.2009AA12Z148, No.2007AA12Z222). We are grateful to the anonymous referees for their comments and suggestions, and thanks are also due to Definiens Company for supporting Definiens Developer trial version software package.

REFERENCES

Baatz, M., and Schäpe, A., 2000, Multiresolution Segmentation-an optimization approach for high quality multi-scale image segmentation. In Angewandte Geographische Informations-Verarbeitung XII, Karlsruhe: Wichmann Verlag, pp. 12-23.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., and Heynen, M., 2004, Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *Isprs Journal of Photogrammetry and Remote Sensing*, 58(3-4), pp. 239-258.

Definiens, 2007, *Definiens Developer 7 - Reference Book*, Germany: München.

Gitas, I.Z., Mitri, G.H., and Ventura, G., 2004, Object-based image classification for burned area mapping of Creus Cape, Spain, using NOAA-AVHRR imagery. *Remote Sensing of Environment*, 92(3), pp. 409-413.

Guo, Q.H., Kelly, M., Gong, P., and Liu, D.S., 2007, An object-based classification approach in mapping tree mortality using high spatial resolution imagery. *Giscience & Remote Sensing*, 44(1), pp. 24-47.

Holt, A.C., Seto, E.Y.W., Rivard, T., and Gong, P., 2009, Object-based Detection and Classification of Vehicles from High-resolution Aerial Photography. *Photogrammetric Engineering and Remote Sensing*, 75(7), pp. 871-880.

Mallinis, G., Koutsias, N., Tsakiri-Strati, M., and Karteris, M., 2008, Object-based classification using Quickbird imagery for delineating forest vegetation polygons in a Mediterranean test site. *Isprs Journal of Photogrammetry and Remote Sensing*, 63(2), pp. 237-250.

Mathieu, R., Aryal, J., and Chong, A.K., 2007, Object-based classification of ikonos imagery for mapping large-scale vegetation communities in urban areas. *Sensors*, 7(11), pp. 2860-2880.

Mitri, G.H., and Gitas, I.Z., 2006, Fire type mapping using object-based classification of Ikonos imagery. *International Journal of Wildland Fire*, 15(4), pp. 457-462.

Rahman, M.R., and Saha, S.K., 2008, Multi-resolution Segmentation for Object-based Classification and Accuracy Assessment of Land Use/Land Cover Classification using Remotely Sensed Data. *Photonirvachak-Journal of the Indian Society of Remote Sensing*, 36(2), pp. 189-201.

Stow, D., Lopez, A., Lippitt, C., Hinton, S., and Weeks, J., 2007, Object-based classification of residential land use within Accra, Ghana based on QuickBird satellite data. *International Journal of Remote Sensing*, 28(22), pp. 5167-5173.

Sun, X.Y., 2008, Analysis of Exploitative Intensity of Coastal Zone Area - A Case Study on the Coastal Zone of Eastern Part of GuangDong (Ph.D Thesis): Graduate School of Chinese Academy of Science, Beijing.

Walter, V., 2004, Object-based classification of remote sensing data for change detection. *Isprs Journal of Photogrammetry and Remote Sensing*, 58(3-4), pp. 225-238.

Wang, L., Sousa, W.P., and Gong, P., 2004, Integration of object-based and pixel-based classification for mapping mangroves with IKONOS imagery. *International Journal of Remote Sensing*, 25(24), pp. 5655-5668.

Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., and Schirokauer, D., 2006, Object-based detailed vegetation classification. with airborne high spatial resolution remote sensing imagery. *Photogrammetric Engineering and Remote Sensing*, 72(7), pp. 799-811.

THE APPLICATION ON SUSTAINABLE LAND USE EVALUATION BY '3S' TECHNOLOGY

ZHAO Bin [a,b,c], ZHAO Wen-ji[a,b,c*], LI Jia-cun[a,b,c]

[a]College of Resources Environment and Tourism, Capital Normal University, 100048,Beijing, China
[b]Laboratory of 3D Information Acquisition and Application, 100048, Beijing, China
[c]Beijing Municipal Key Laboratory of Resources Environment and GIS, 100048, Beijing, China

KEY WORDS: sustainable land use, evaluation, '3S' technology, indicator system, spatialization, statistical indicator, geospatial indicator, statistical unit

ABSTRACT:

With land resource lack and environment pollution, land conflict between supply and demand problems has affected human survival environment sustainable development badly. Sustainable land use evaluation could explore the land resource exploitation direction, level and mechanism, it provides gist for land resource planning and management to make it in favour of social environment development. '3S' technology is useful in land resource evaluation, GPS provides the precise space position, RS provides land cover image frequently and accurately, and GIS provides a tool for spatial analysis and cartography. This paper introduced the process and method of sustainable land use evaluation, mainly about the indicator spatialization. The indicator can be divided into statistical indicator and geospatial indicator. Take Guyuan county, Kangbao county and Zhangbei county in Bashang region as examples, an sustainable land use evaluation indicator system was set up. Evaluating the sustainable land use in Bashang region in 2003, the result distributes continuous. It illuminates that the sustainable land use level is low in Bashang region. From the land use view, the distribution of farm land has higher level than water and useless land.

1. INTRODUCTION

With the rapid development of economic construction, shortage and abuse of land resources have become important factors that affecting social development and land management. Land resource evaluation analysis determines whether the requirements of land use are adequately met by the properties of the land (Bandyopadhyay, S. et al., 2009). It is the basis of land resource plan and management, play an important part on land resource evaluation in a long term.

Along with the rapid advance of '3S' technology, integrate '3S' technology to evaluate the sustainable land use capacity proves great potential capacity. GPS provides the precise space position, RS provides the image of land cover condition frequently and accurately, GIS provides a tool for spatial analysis and cartography. Therefore, it's available to use '3S' technology on the evaluation research of sustainable land use, it has important theoretical significance and practical value.

The object of this paper was to build a sustainable land use evaluation system with '3S' technology to evaluate the land resource level. The approach in this paper was based on '3S' technology to collect and calculate the indicators, especially about spatializing the indicators to let them distributed continuous.

2. STUDY AREA

Bashang region is generally called as a highland in north Hebei Province, located in the transition region among Inner Mongolia Plateau, Yanshan Mountains and North China Plain. It was located between 41°13′ to 40°57′N, and 114°50′ to 116°05′E, and the total area is 18,202 square kilometres. It is a narrow nearly east-west direction belt, continental monsoon and arid climate, north temperate broad-leaved forest with grasslands, agricultural and pastoral areas transition zone (Zhou, 2004). The maximum ground elevation in Bashang region is 1200 to 1500 meters above sea level. The mean yearly temperature is -0.3 to 3.5 ℃, the mean annual rainfall is about 400 mm, and it's also the ecological barrier between Beijing and Tianjin, even North China (Zhou, 2004). The agricultural activities depend on rainfall, the water are mainly from small tanks and underground water. The soils of the area are chestnut soil, poor adhesion, easy physical weathering, with low organic, predominantly yellowish to brown in colour. Bashang region including Zhangjiakou Guyuan county, Zhangbei county, Kangbao county, and part of Shangyi county, Yixian County, Fengning county, and Weichang county (Yuan, 2006). This paper takes Guyuan County, Kangbao County, and Zhangbei County as a study area.

The data mainly used contain Landsat-7 ETM + images on September 10, 2008, 1:50000 scale DEM, Statistical Yearbook, and detail vector map. After get the data, image pre-process was first step, including geometric correct with GPS point gathered from the field, mosaic, clip the images to the study area. Take the typical land use/cover types noted from field investigations as interesting area to classify the image. The result classified into farmland, woodland, grassland, water, urban and industrial land, useless land (bare land, saline soil, sand) total 6 land-use types.

[*] Corresponding author. Email:. zhwenji1215@163.com

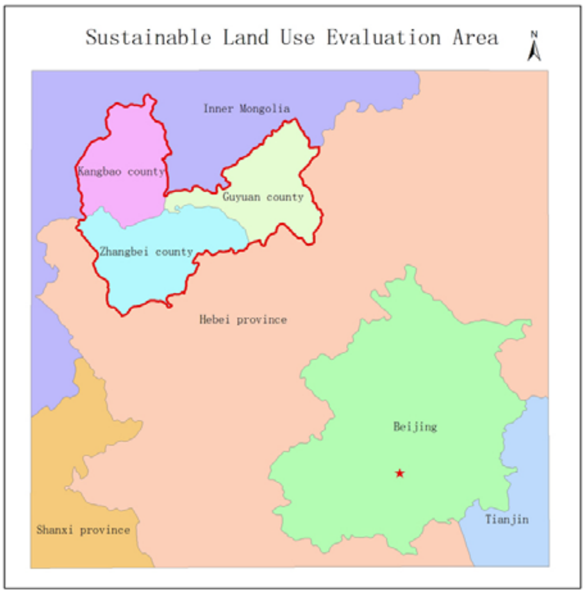


Figure 1. study area in Bashang region chart

3. METHOD

The work flow of this paper is: accord with the sustainable evaluation target, as well as spatial and temporal scales characteristic built a multi-level evaluation indicator system; based on available, acquire and compositive principles selected indicators; then quantified it with acquire method, selected threshold value and standardized method for evaluating indicator standardization; displayed it in the map step which called spatialization, mainly related to spatial interpolation methods; select a appropriate weight assignment methods on evaluation index system; at last, establish an GIS evaluation model to get the evaluation result(Figure 2).

3.1 Build Evaluation Indicator System

The evaluation indicator system was empirical assessment systems and based on the knowledge and understanding of the study area. So the factor involves in land evaluation either single or multiple parameters converted to an integrated indicators (Guo et al., 2005).

Accord to comprehensive operability, and dynamic principle, classed the indicators from three levels: objective level, criterion level and indicators level. There aer many factors influence the land resource condition, So selected FAO promoted from the production, productivity, stability, environmental protection, economic viability, social acceptability of five aspects to establish evaluation indicator system of sustainable land use (FAO, 1990). The indicator includes slope, vegetation coverage, crop productivity index, multiple cropping indexes and so on of 13 indicators. The indicators were collected from remote sensing, GIS analysis, yearbook and field survy.

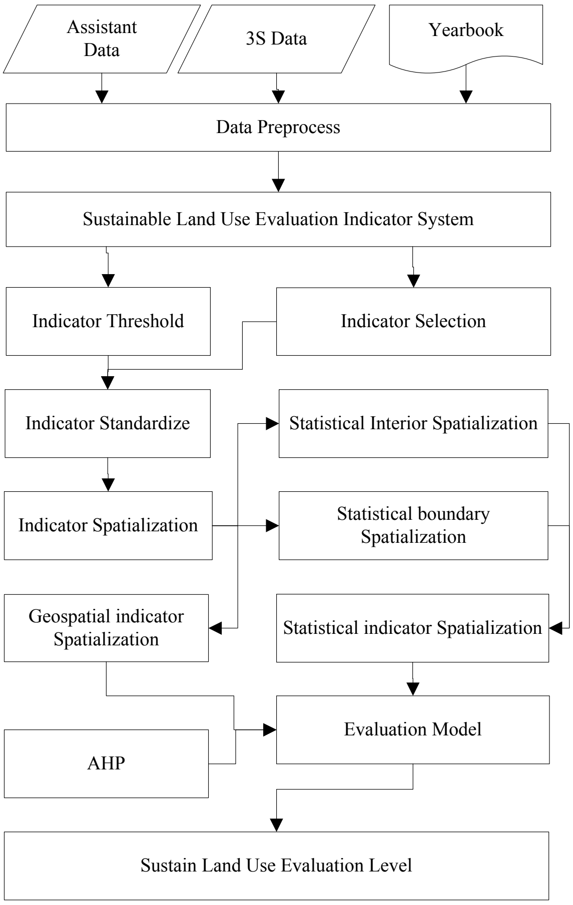

Assistant Data
3S Data
Yearbook
Data Preprocess
Sustainable Land Use Evaluation Indicator System
Indicator Threshold
Indicator Selection
Indicator Standardize
Statistical Interior Spatialization
Indicator Spatialization
Statistical boundary Spatialization
Geospatial indicator Spatialization
Statistical indicator Spatialization
Evaluation Model
AHP
Sustain Land Use Evaluation Level

Figure 2. Sustainable Land Use flow chart

3.2 Standardize Indicator

Select the indicator's threshold is difficult, it always in accordance with the regulations and standard. Some indicators didn't have clear threshold, such as multiple cropping indexes; and some were hard to quantitative, such as land deal value index. So we have to base on the local condition to select a proper threshold.

Due to different units, the Quantitative evaluation indicator can not be compared with each other, it's a necessary step to standard it. It contains three steps: indicator type and standardized methodology. The general relationship of the indicators was divided into three categories: positive type, negative type, and moderate-type (Zhou, 2004) Standardized methods include extreme-minus method, percentage ratio method, fuzzy mathematics method, power conversion method and so on.

Positive type：

$$Ri = 100 \times r_i = \begin{cases} 1 & ; \quad (x_i > z_i) \\ x_i / z_i & ; \quad (0 < x_i \le z_i) \\ 0 & ; \quad (x_i \le 0) \end{cases} \tag{1}$$

Negative type：

$$Ri = 100 \times r_i = \begin{cases} 0 & ; \quad (x_i > z_i) \\ 1 - x_i / z_i & ; \quad (0 < x_i \le z_i) \\ 1 & ; \quad (x_i \le 0) \end{cases} \tag{2}$$

Moderate type：

$$Ri = 100 \times r_i = \begin{cases} x_i / z_i & ; \quad (x_i > z_i) \\ 1 & ; \quad (x_i = z_i) \\ z_i / x_i & ; \quad (0 \le x_i < z_i) \end{cases} \tag{3}$$

Where: x_i = actual index value
z_i = threshold
r_i = standardized index
R_i = ri × 100

3.3 Spatialize Method

Expression of Sustainable land evaluation result on map was the direct form to monitor the land use situation; and indicator spatialization was its core method. From the '3S' technology content, Spatialization was the excellent representation to express its advantage. The spatialization of population was pointed out at the beginning of the study before the widely used of '3S' technology. It's mainly using population gravity, potential models, Lorenz curve and spatial autocorrelation theory etc (Liu, 2002). These traditional methods tried to combine qualitative methods with quantitative methods to reflecting the population spatial distribution (Li, 2008), however, it didn't meet with the population distribution in the physical geography. With '3S' technology improves, researchers has been developed a lot of complex models and methods to simulate population's distribution precisely. Jiang (2002) based on '3S' technology summarized the population's spatialization development. Now, spatialize methods mainly composed of correlation analysis, interpolation, grid spatialization (Li, 2008).

Interpolation method was based on the known value of the region to estimate the value of the unknown region (Li, 2000). It was a crucial technique in analyzing spatial data and had been used in a wide range of disciplines (Lam, 2009). The interpolation of spatial data had been considered in many different forms. The various forms of kriging and inverse distance weighting (IDW) were among the best known in the earth sciences (Myers, 1994).

Spatial interpolation is widely used to create continuous surfaces from discrete data points (Wang, 2003).Evaluation indicator mainly included statistical indicator and geospatial indicator. Statistical indicators (such as population density, per capital GDP etc.) had an accurately statistical value, but there was no geographic coordinates and projection information, therefore it's difficult to locate on map accurately, and its value was homogeneous interior the statistical unit and exist gaps between them. Geospatial indicator (such as vegetation cover, slope, etc.) with precise location and projection information inherently that easy to map, however, the different spatial resolution was difficult to topological and overlay analysis. Spatialization focus on these problems to establish a continuous spatial distribute interpolation method (Pan, 2002; Lu, 2008).

3.4 Spatialize Statistical Indicator

Spatialize statistical indicator is based on the values collected by statistical units to establish the relationship between statistics value and models. It was feasible on the premise that the statistical indicators uniformly distributed in the region that didn't match the physical geography (Lv, 2002). Statistical indicators spatialization was to break the statistical unit boundary and make it close to the real situation farthest (Ma, 2008). Grid cell size's areal weighting interpolation method (GCAWIM) could use grid to spatialize statistical unit boundary. The basic idea was taken small grid to instead big statistical unit, use grid area as power recalculated its value. Its key was selected a proper grid size.

Use GCAWIM only can spatialize the statistical unit boundary, and spatialize the interior statistical unit is also important. Inverse distance weighting (IDW) method was a common deterministic spatial interpolation method. Its general idea was based on the assumption that the attribute value of an unknown point is the weighted average of known values within theneighborhood, and the weights are inversely related to the distances between the prediction location and the sampled locations (Lu, 2008).it can use the land use types as weight to spatialize the statistical interior. At last, integrated the boundary and interior spatialize result to get the indicator spatial distribution.

3.5 Spatialize Geospatial Indicator

Geospatial indicator was spatialized through resample and band calculator. RS index can be used as parameters in the evaluation directly. Vegetation coverage index (VCI) was an important parameter to reflect the extent of vegetation cover. It's one of important ecological indicators about environmental change, the higher value, the more vegetation is flourishing, It could be got from DNVI based on the pixel and NDVI moiety model to calculate the vegetation coverage (Chen, 2008). NDVIs and NDVIv are representing the pure bare land and pure vegetation's NDVI values; they can be got from field measurement or image. The simple method was to choose the minimum and maximum of NDVI directly act as them respectively (Sun, 2006). A more precise method was select 1% and 99% of the total pixels' NDVI as the corresponding value. Compared with the former, the Latter has fewer extreme points and bad lines, stripes.

$$VCI = \frac{NDVI - NDVI_s}{NDVI_v - NDVI_s} \tag{4}$$

Where: NDVIs = pure bare land's NDVI
NDVIv = pure vegetation's NDVI
VCI= Vegetation coverage index

According to this method we can spatialize other indicators. The indicators associate with population, such as population density in each village to spatialize per capital GDP.

The weight assignment methods can divide into subjective and objective kinds, the former lean to experience but the later mathematics. Compared with the living method, Analytic Hierarchy Process (AHP) integrates both advantages, so use

AHP method to give each indicator a suitable weight based on their sensitivity. Build the integrate evaluation model based on GIS to get the sustainable land resources level distribution (Figure 3).

$$S = \sum_{i=1}^{5} (R_i \times W_i) \tag{5}$$

Where: Ri = Standardized indicator value
Wi = Indicator weight
S= Integrate evaluation value

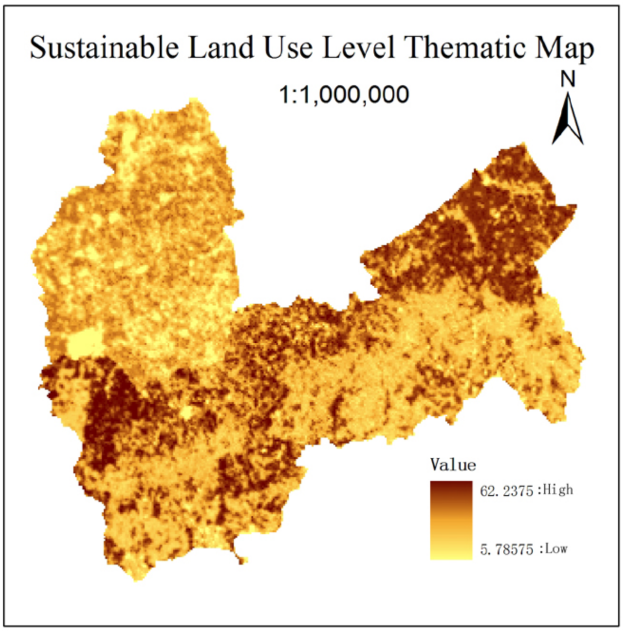


Figure 3. sustainable land use level in Bashing region 2003 chart

4. CONCLUSION

This paper attempts to develop an index system integrating the collected information using '3S' technology. The result distributes continuous in the study area, and it indicated that the sustainable land use level in study area was low. The land resources condition in Kangbao county and north of Zhangbei county are better than other regions. From the land-use perspective, the three counties damaged relatively serious, especially about agriculture and stockbreeding land. Bashang region only suitable for drought-hardy plants growth, land productivity is low, typically of extensive cultivation agriculture. Land desertification is an important factor in the decline of soil quality. Data analysis showed that the desertification was positive correlation with farmland, and negative with forest and grassland (Zhou, 2004). Increase farmland and desertification caused land degradation affected sustainable land use condition. Therefore, change existing leanness farmland to reconstruct forest or grasslands is very necessary.

5. DISCUSSION

This paper establishes a sustainable land use evaluation system in Bashang region, including statistical indicators and geographical indicators. The statistical indicators spatialized by GCAWIM combine with land use weight interpolation method. Geospatial indicators mainly come from '3S' indicator. Evaluate the Guyuan county, Kangbao county, and Zhangbei county sustainable land resources. The result indicated that the sustainable land use level in study area was low. Kangbao county and north Zhangbei county's land resources condition are better than other regions; farmland has higher level than the other land use types.

Sustainable land use evaluation has been proposed for a few years, a lot of researchers pay great effort study on it. Some difficulties still exist, such as indicator selection veracity, indicator threshold selection, spatialization method.

'3S' technology provides precise and quantitative parameters and analysis methods to improve the quality of sustainable land use evaluation. There are many ways to select the indicator systems, how to avoid their disadvantages and integrates their advantages is still have to study. Now, we select the indicators mainly based on the easy quantitative ones; ignore those hard to quantitate.

Spatialization is about areal interpolation, but until now we have to use point interpolation and line interpolation to instead. With the '3S' technology development, more indexes contain clear physical meaning can be applied to the evaluation index of the space. In the future a appropriate spatialization method should be founded.

References from Journals:

Bandyopadhyay, S. et al., 2009. Assessment of Land Suitability Potentials for Agriculture Using a Remote Sensing and GIS Based Approach. International Journal of Remote Sensing, 30(4), pp. 879 - 895.

Chen, H. F. et al., 2008. Advances in Researches on Application of Remote Sensing Method to Estimating Vegetation Coverage. Remote Sensing For Land & Resources, (1), pp. 13-18.

Food and Agricultural Organization of the United Nations (FAO), 1990. Guidelines for Soil Profile Description (Rome, Italy: FAO).

Guo, X. D., et al., 2005. Land Degradation Analysis Based on the Land Use Changes and Land Degradation Evaluation in the Huan Beijing Area. In: Remote Sensing for Environmental Monitoring, GIS Applications, and Geology V, Proceedings of SPIE, M. Ehlers and U. Michel (Eds), 5983, pp. 598319.

Jiang, D. et al., 2002. Study on Spatial Distribution of Population Based on Remote Sensing and GIS. Advance in Earth Sciences, 17(5), pp. 734-738.

Lam, N. S. et al., 2009. Spatial Interpolation, International Encyclopedia of Human Geography. Elsevier, Oxford, pp. 369-376.

Li, M. J. et al., 2008, Discussing and Using the Method of Population Density Spatialization of Liao cheng. City Journal of Guangzhou University (Natural Science Edition), 7(2), pp. 71-74.

Liu, D. Q., Liu, Y., and Xue, X. Y., 2002. Spatial Distribution and Autocorrelation Analysis of Population in China. Remote Sensing Information, (6), pp. 1-6.

Lu, G. Y., David, W. W., 2008. An Adaptive Inverse-distance Weighting Spatial Interpolation Technique. Computers & Geosciences, 34(9), pp. 1044-1055.

Lv, A. M., Liu, H. Q., and Li, C. M., 2002. Population Density Algorithm Based on Areal Interpolation. Journal of China Agricultural Resources and Regional Planning, 23(1), pp. 734-738.

Ma, J., Jiao W. X., 2008. A Review on Pixelizing of Social Statistical Data. Future and Development, (3), pp. 25-28.

Myers, Donald, E., 1994. Spatial Interpolation: An Overview. Geoderma, 62, (1-3), pp. 17-28.

Pan, Z. Q, Liu G. H., 2002. The Research Progress of Areal Interpolation. Progress in Geography, 21(2), pp. 152-146.

Sun, J. H. et al, 2006. Estimation of Vegetation Fraction in Bei Yunhe District by Remote Sensing. Research of Soil and Water Conservation, 13(6), pp. 97-99.

Wang, S. W., Armrtrong, Marc P., 2003. A Quadtree Approach to Domain Decomposition for Spatial Interpolation in Grid Computing Environments. Parallel Computing, 29,(10), PP.1481-1504.

Yuan, J.G. et al., 2006. Land Degradation and Ecological Reconstruction of Eco-fragile Region in Bashang of Hebei Province. Journal of Arid Land Resources and Environment, 20(2), pp. 139-143.

Zhou, X.C. et al., 2004. Study on Dynamic Monitoring Land Use/Cover Change in Bashang Area Based on RS and GIS Technique. Research of Soil and Water Conservation, 11(3), pp. 17-20.

Zhou, X. C. et al., 2004. Study on the RS-based Dynamic Monitoring of LandUse/cover Change in the Bashang Rigion, Heibei Province. Arid Zone Research, 21(4), pp. 408-410.

AN ASSESSMENT OF THE EFFICIENCY OF LANDSAT, NIGERIASAT-1 AND SPOT IMAGES FOR LANDUSE/LANDCOVER ANALYSES IN EKITI WEST AREA OF NIGERIA

Ojo A G[a] , Adesina F A[b]

[a]African Regional Centre for Space Science and Technology Education, PMB 019 OAU Campus, Ile-Ife.
ojobayous@yahoo.com[a]
[b] Department of Geography Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria
faadesin@oauife.edu.ng[b]

KEYWORDS: **L**and-use, Accuracy Assessment, Landsat TM, SPOT XS, NigeriaSat-1, Classification

ABSTRACT:

Several remote sensing data types are now available for environmental studies. The variety has increased as many nations including some African countries invest in satellite remote sensing. However, each data type has its own peculiar features that may limit or enhance its relevance to capture data for specific range of information. This study used geo-information techniques based on multi-source imageries to enhance the utilization of images with coarser resolutions in landuse analysis in Ekiti west area of south western Nigeria. The objective of the study is to evaluate the variations in landuse characterization with multi-source satellite data sets. The remotely sensed data sets used included Landsat TM 1986, SPOT XS 1995 and NigeriaSat-1 2007 satellite images. To make the images comparable, they were georeferenced, re-sampled and enhanced for visualization in a GIS environment. The tonal values recorded in the images with the features on the ground were validated by ground truthing. The data from ground truthing were combined with visual image interpretation for "supervised" classification. The classes defined and analyzed included "built-up area'', "bare rock'', "farmland'', secondary forest regrowth'' and "water body''. The results show that each image has certain relative advantage over the other. For instance, while NigeriaSat-1 image was efficient in the analysis of information within the visible portion of the electromagnetic spectrum, SPOT image was better in the Near Infrared. Information from Landsat image was rather weak at both portions (Visible and NIR) of the Electromagnetic Spectrum. The study also shows that SPOT image has the lowest level of data redundancy of the three image providers. The study confirms the relevance of the growing interest in the use of geo-information techniques for landuse analysis.

1. INTRODUCTION

Remotely sensed imageries are one of the most important sources of spatial data for environmental studies. They are data obtained via remotely placed sensors which may be located at heights sometime several hundred of kilometres in space to make it possible for the sensor to "see'' a large portion of the earth's surface at the same time. Such images can also be obtained from low flying aircrafts equipped with suitable cameras to track earth-based features. These data sets allow earth-based phenomena such as landuse and landcover characteristics to be rapidly mapped, if needed repetitively and at relatively low costs. With increasing capacity to rapidly generate maps of large areas, planners in the rural and urban areas are getting more empowered to address issues associated with landuse analysis such land misuse and various forms of incursion into properties and trespassing.

Some of the most commonly used remote sensing data sets for mapping landuse and landcover are those from Landsat, SPOT (Système Probatoire d'Observation de la Terre), IRS (Indian Remote Sensing), ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer), MODIS (Moderate Resolution Imaging Spectrometer), JERS-1 (Japanese Earth Resources Satellite), and recently, NigeriaSat-1 satellites. The Landsat data have greater spectral resolution (Gastellu-Etchegorry, 1990) and a longer time series, while SPOT provides better spatial resolution but with shorter historical records. Newer satellite imaging systems ar commonly equipped with enhanced instruments to generate additional data that permit more accurate mapping and analysis. Landuse/landcover analyses usually proceed from classification of the area of study. The classified units can be further analysed in terms of their characteristics particularly size.

Factors that may influence classification accuracy include a sensor's spatial, radiometry and spectral resolutions. Spatial resolution describes the size each pixel represents in the real world (Cushnie, 1987). For example, a satellite with 30 metre resolution produces pixels that measure a 30x30 metre area on the ground. Radiometric resolution, on the other hand, is the smallest difference in brightness that a sensor can detect. A sensor with high radiometric resolution would therefore have very low "noise''. The "noise'' is described as any unwanted or contaminating signal competing with the desired signal. Spectral resolution is the number of different wavelengths that a sensor can detect. A sensor that produces a panchromatic image alone has a very low spectral resolution, while one that can distinguish many shades of each colour has a high spectral resolution (Jensen, 2007).

Generally, spatial resolution is the most important factor of the three for landuse and landcover definition. For example Gastellu-Etchegorry (1990), in Indonesia studied landuse with SPOT and Landsat images. He showed that SPOT Multispectral (XS) images are better than Landsat Multispectral Scanner (MSS) images for mapping of heterogeneous near-urban landcover because of SPOT's superior spatial resolution. The link between spatial resolution and classification accuracy, however, is sometimes weak (Jensen, 2005). In heterogeneous areas, such as residential areas, it has been shown that classification accuracies may improve even as spatial resolution decreases (Cushnie, 1987). This occurs due to the potentials of urban features to blend together to form composite distinctive “urban signals” that can be distinguished from other landcovers.

In this study, the capabilities of three satellite imageries including Landsat, NigeriaSat-1 and SPOT-XS for landuse studies are evaluated with a view to amplifying the understanding of their specific potentials, advantages and limitations for environmental studies. This is borne out of the growing recognition of the need to develop “appropriate technology” for developing countries to enable researchers based in these countries to accomplish desirable level of sophistication in earth’s resources analyses even with their limited access to spatial data.

2. THE STUDY AREA/ LOCATION EXTENT

The study area is made up of Ekiti west, Ado-Ekiti, Irepodun/ Ifelodun and Ekiti south-west. Local government areas in Ekiti State of western Nigeria (Figure1). The State lies within Longitudes 4^{o} $5^{'}$ and 5^{o} $45^{'}$ East of the Greenwich Meridian and Latitudes 7^{o} $15^{'}$ and 8^{o} $5^{'}$ North of the Equator. It is about 6,353 square kilometers in size. It is bounded in the north by Kwara and Kogi States, in the West by Osun State and Ondo State in the East and in the South. The State has 16 Local Government Areas. By 1991 Census, its population was 1,647,822. The estimated population upon its creation on October 1st 1996 was 1,750,000 with the capital located at Ado-Ekiti (Ekiti Investors Handbook, 2002). The current estimated population based on 2006 census, was 2384212 million people (NPC, 2006).

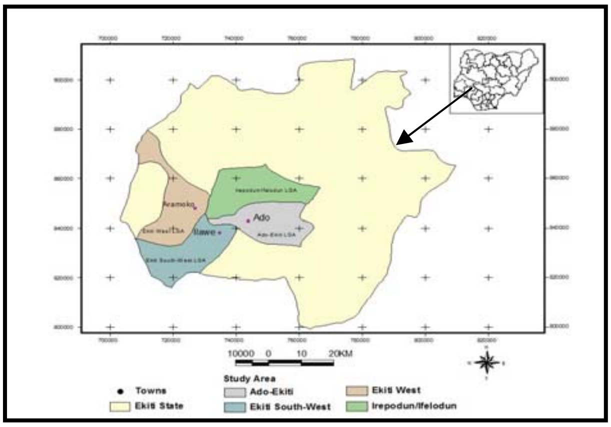


Figure 1: Map of Ekiti State and the LGAs used for the study

The State is mainly an upland area, rising generally around 250 meters above the sea level **(**Ekiti Investors Handbook, 2002). The landscape consists of ancient plains broken by steep-sided outcropping dome rocks. These rocks occur singly or in groups or ridges and the most notable of these are to be found in Efon-Alaaye, Ikere-Ekiti and Okemesi-Ekiti (EKSG, 1997). The area is underlain by metamorphic rock of the basement complex. The State is dotted with hills of varying heights. The notable ones among them are Ikere-Ekiti Hills in the southern, Efon-Alaaye Hills in the western and Ado-Ekiti Hills in the central parts.

The State enjoys the tropical climate with two distinct seasons. These are the rainy (April-October) and the dry (November-March) seasons with an annual rainfall of around 1150mm. Temperatures range between 21^{o} and 28^{o}C with high humidity. The South-Westerly wind and the North-Easterly Trade winds blow in the rainy and dry (harmattan) season respectively. The tropical forest originally covered this part of Nigeria. However, as a result of exploitation for many centuries of exploitation, the original vegetation has been removed and now replaced with anthropic covers. Forest is now confined largely in the south, while savanna dominates the natural landscape in the north (Kayode, 1999 and EKSG, 1997).

3. METHODOLOGY

3.1 Data Acquisition and Image Preprocessing

Relevant data were collected on the physical attributes of the five landuse types, i.e. farmland, built-up areas, forest regrowth, bare lands and water bodies, which were the dominant landuse in the area. Sample sites for data collection were determined from the remotely sensed imageries .For each landuse, five sample sites measuring 100 x 100m were selected and were fully described during field observations, variables examined for site characterization included dominant, concrete among others and drainage characteristics of the site . The coordinates of the sample sites were tracked with Global Positioning System (GPS) receiver. Secondary data used were Landsat TM, SPOT- XS NigeriaSat-1 covering the study area. Table1 shows key parameters of the data used for the study. Images were obtained from the National Centre for Remote Sensing Jos, Forest Monitoring, Evaluation and Coordinating Unit (FORMECU) Abuja and Global Land Cover Facility (GLCF) an Earth Science Data Interface. The topographic map covering the study area was collected from the Ministry of Land and Surveys, Lagos.

The satellite data were extracted one after the other as sub-scenes from the original datasets. For the purpose of landuse/cover assessment, a common window covering the same geographical coordinates of the study area was extracted from the scene of the images obtained. This made the band Red, Blue, Green (RGB-123) colour combination. For SPOT XS data, Channel 3 was assigned red plane, Channel 2 to green and channel 1 to blue plane. The band combination then consisted of Blue, Green and Red (BGR-321) colour combination. For NigeriaSat_1 data set, colour combination, channel 1 was assigned to red plane, channel 2 to green plane and 3 to blue plane. This puts the band combination as Red, Green and Blue (RGB-123).

The geometric errors were corrected using ground control points (GCP). The process of georeferencing in this study started with the identification of features on the image data, which can be clearly recognized on the topographical map of the study area and whose geographical locations were clearly defined. Stream intersections and the intersection of the highways were used as ground control points (GCP). The latitude and longitude of the GCPs of clearly seen features obtained in the base map were used to register the coordinates of the image data used for the study. All the images were georeferenced to Universal Transverse Mercator projection of WGS84 coordinate system, zone 31N with Clarke 1866 Spheroid. Nearest-neighbor re-sampling method was used to correct the data geometrically.

Table. 1 Attributes of the images used for the Study

Image Name	Source	Sensor	Acquisition Date	Band	Resolution	Spectral Range (um)	Area (km^2)
Topographic Map	MLSL Lagos	-	1966	-	-	-	-
Landuse Map	NCRS, Jos		1995	-	-	-	-
Land sat TM	GLCF	TM	1986	1-7	30m	0.45-0.90	185 x185
SPOT	FORMECU, Abuja	MSS	1995	1-3	10m	0.45-0.69	60x 80
NigeriaSat-1	NCRS, Jos	Imager	2007	1-3	32m	0.52-0.90	600x 580

3.2 Landuse Classes

Based on the knowledge of the study area, reconnaissance survey and additional information from previous studies in area, a classification scheme was developed after Anderson *et al.*, (1976). The scheme gives a broad classification where each of the land use/ land cover was identified by a class (Table 2). These classes are apriori well defined on the three images used for the study.

Table. 2 Landuse classification scheme (after Anderson *et al* 1976)

LANDUSE/LANDCOVE CATEGORIES	DESCRIPTION OF THE LANDUSE/LANDCOVER
Built-up Area	Roads, buildings, open spaces
Bare Rock	Bare soil, bare land
Farm Land	Shrubs, fallow, cropped land.
Secondary forest	Agro forest, riparian forest, advanced bush re-growth
Water Body	Dam, rivers streams.

3.3 LandCover/ Landuse Analysis

For Landuse/Landcover analyses, the satellite images were classified using the supervised classification method. The combined processes of visual image interpretation of tones/colours, patterns, shape, size, and texture of the imageries and digital image processing were used to identify homogeneous groups of pixels, which represent various land use classes already defined. This process is commonly referred to as “training’’ sites because the spectral characteristics of those known areas are used to “train’’ the classification algorithm for eventual land use/ cover mapping of the remaining parts of the images.

A Map of the study area was produced and was used to locate and identify features both on ground and on the image data. The geographical locations of the identified features on the ground were clearly defined. These were used as training samples for supervised classification of the remotely sensed images. The five categories of land uses/ land covers were clearly identified during ground truthing. Locations were tracked with the GPS to facilitate transference of the field information onto the images.

3.4 Classification

In this study, the satellite images were classified using supervised classification method. The combined process of visual image interpretation of tones/colours, patterns, shape, size, and texture of the imageries and digital image processing were used to identify homogeneous groups of pixels, which represent various land use classes of interest. The study engaged in ground truthing to the four Local Government Area of the study area. These are Ekiti west, Ado-Ekiti, Irepodun/ Ifelodun and Ekiti south-west Local government areas in Ekiti State (Figure1). Before the ground truthing, map of the study area was printed and was used as guide to locate and identify features both on ground and on the image data. The geographical locations of the identified features on the ground were clearly defined. These were used as training samples for supervised classification of the remotely sensed images. Five categories of land uses and land covers were clearly identified during ground truthing. These are secondary re-growth forest, water body, bare rocks, built-up areas and farm land. The processed images were subject to band correlation analysis to assess the nature and strength of the relationship among the bands in the imageries.

4 RESULTS

4.1 Comparison of Basic Features among the Three Sensor Data

Table 3 summarizes the correlation analysis of bands with each other within each of the three sensors. In the NigeriaSat-1 image, the Near-Infrared (NIR) band was negatively correlated with the visible bands (Green and Red) ($-0.16, \leq r \geq -0.04$; $p < 0.05$). In the Landsat TM image, the NIR band positively correlated with visible bands ($0.02 \leq r \geq 0.22$; $p < 0.05$). For the SPOT image, the NIR band also positively correlated with the visible bands ($0.53 \leq r \geq 0.63$; $p < 0.05$). The relationship between the visible bands were strongest in SPOT ($r = 0.98$), relatively strong in NigeriaSat-1 images ($r = 0.53$) and relatively low in Landsat TM ($r = 0.22$).

Table. 3 Correlation matrix analysis results for the three sens data

Sensor	Bands	Green	Red	NIR
Landsat	Green	1.00	0.22	0.02
	Red	0.22	1.00	0.93
	NIR	0.02	0.22	1.00
NigeriaSat-1	Green	1.00	**0.95**	-0.04
	Red	0.95	1.00	-0.16
	NIR	-0.04	-0.16	1.00
SPOT	Green	1.00	0.98	0.63
	Red	0.53	1.00	0.98
	NIR	0.63	0.53	1.00

Level of significance (p) <0.05

The results imply that the SPOT image is likely preferable to either of the other image types for the study of earth base features at the Visible and Near Infrared portions of the Electromagnetic Spectrum. On the other hand, NigeriaSat-1 imageries could give better information at the visible portion while Landsat imageries could be better in the Visible and Near Infrared portions of the spectrum. The results indicate that the strength of the correlation among the bands increases with increase in the spectral resolution of the imageries. This corresponds with what many authors have observed. For example, Kuplich *et al.* (2000) have suggested based on their studies, that high correlation between spectral bands is indicative of high degree of information. Spectrally adjacent bands in a multispectral remotely sensed image are often highly correlated. Multiband visible/near-infrared images of landuse areas will show negative correlations between the near-infrared and visible red bands and positive correlations among the visible bands because the spectral characteristics of land use are such that as the vigour or tone of the feature increases, the red reflectance diminishes and the near-infrared reflectance increases.

4.2 Landuse Classes

Table 4, 5, and 6 contain summaries of the results accuracy assessment generated from the three images on the five land use. The overall, user as well as producer accuracies of individual classes were consistently high for the three imageries. The accuracies for Landsat, NigeriaSat-1 and SPOT were 66.5%, 81.2% and 82.8% respectively. The Kappa statistics were 0.87, 0.97 and 0.89 respectively. The user and producer accuracies of individual classes in Landsat ranged from 51.9% to 87.3%, whereas in NigeriaSat-1, they varied from 60% to 97.3% and in SPOT from 74.7% to 89%. On Landsat image for the 'built-up area' category of land use, the producer accuracy is 82.8% and the user is 87.3%. This means that more than 80% of the built-up area in the image was correctly defined and mapped. The other four land use classes i.e. bare rocks, farm land; secondary forest regrowth forest and water body had relatively low producer and user accuracies. For instance secondary forest regrowth has a producer accuracy of 66.5% and user accuracy of 51.9%. The four categories are thus not as well defined as the built-up areas. The 'built-up area' of NigeriaSat-1 had the highest accuracy on producer accuracy (97.3%), it also had user accuracy of 96.9%. On the SPOT image, the statistics on built-up area category are 81.98% for producer accuracy and 91.58% for user accuracy.

It appears that Landsat has lower value for accuracies than either SPOT or NigeriaSat-1. Different reasons may be responsible for this outcome. One reason may have to do with the intrinsic characteristics of the images. For instance Landsat has a spatial resolution of 30metres, NigeriaSat-1 32 metres and the SPOT 10metres. Chen, et al., (2004) has shown that these can variously have effect on the levels of accuracies obtained from the images. Another reason could be that, spectral characteristics among the different land cover types (e.g. built-up, bare rock) are similar, while spectral variation within the same land cover type or even within the same image might be high (Cushine, 1987).

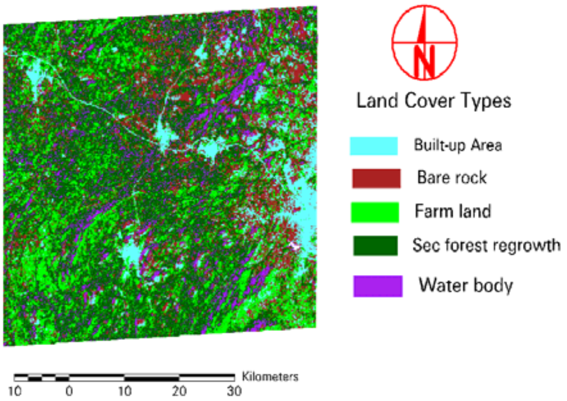


Fig2: Classified print of the Landsat TM image

TABLE. 4 ACCURACY ASSESSMENT OF LANDSAT TM IMAGERY

Satellite Image	Classified Data	Reference Data					Ref. Totals	Class Totals	Number Correct %	PA %	UA % Kappa	
		Built-up Area %	Bare Rock %	Farmland %	Sec. Forest %	Water-body %						
Landsat	Built-up Area	**82.79**	5.60	4.67	1.15	0.67	100	94.88	82.79	82.79	87.26	1.0000
	Bare rock	8.26	**58.84**	12.13	8.21	4.52	100	91.96	58.84	58.84	63.98	1.0000
	Farmland	7.02	19.14	**52.15**	12.29	0.90	100	91.50	52.15	52.15	56.99	1.0000
	Sec. forest re-growth	1.79	9.41	29.21	**66.49**	21.14	100	128.04	66.49	66.49	51.93	0.8601
	Water body	0.15	8.02	1.84	11.87	**72.76**	100	94.64	72.76	72.76	76.89	1.0000

Overall Landsat Classification Accuracy = 66.47% (i.e. 82.79+58.84+52.15+66.49+72.76),
Overall Kappa Statistics = 0.8714
501.02

TABLE. 5 ACCURACY ASSESSMENT OF NIGERIA SAT-1 IMAGERY

Satellite Image	Classified Data	Reference Data					Ref. Totals	Class Totals	Number Correct %	PA %	UA %	Kappa
		Built-up Area %	Bare Rock %	Farmland %	Sec. Forest %	Water-body %						
Nigeriasat-1	Built-up Area	**97.32**	3.03	0.02	0.05	1.16	100	100.42	97.32	97.32	96.91	1.0000
	Bare rock	1.89	**77.27**	0.23	3.09	35.07	100	117.55	77.27	77.27	65.73	1.0000
	Farmland	002	0.39	**88.07**	12.12	0.00	100	100.60	88.07	88.07	87.55	1.0000
	Sec. forest re-growth	0.04	7.64	11.37	**82.38**	3.77	100	105.20	82.38	82.38	78.31	0.9230
	Water body	0.73	11.67	0.31	2.37	**60.00**	100	75.08	60.00	60.00	79.92	1.0000

Overall NgeriaSat-1 Classification Accuracy = 81.20% (i.e. 97.32+77.27+88.07+82.38+60.00),
Overall KappaStatistics = 0.9712
498,85

TABLE . 6 ACCURACY ASSESSMENT OF SPOT IMAGERY

Satellite Image	Classified Data	Referer					Ref. Totals	Class Totals	Number Correct %	PA %	UA %	Kappa
		Built-up Area %	Bare Rock %	Farm land %	Sec. Forest %	Water-body %						
SPOT	Built-up Area	**81.98**	4.65	2.86	0.03	0.00	100	89.52	81.98	81.98	91.58	1.0000
	Bare rock	11.15	**88.99**	2.17	4.26	3.91	100	110.48	88.99	88.99	80.55	1.0000
	Farmland	5.72	2.10	**83.40**	4.98	3.91	100	100.11	83.40	83.40	83.31	0.8681
	Sec. forest re-growth	1.05	3.27	11.19	**78.11**	10.94	100	104.56	78.11	78.11	74.70	0.8714
	Water body	0.10	0.98	0.37	12.62	**81.25**	100	95.32	81.50	81.50	85.50	0.9781

Overall Spot Classification Accuracy = 82.75% (i.e. 81.98+88.99+83.40+78.11+81.25),
Overall Kappa Statistics = 0.8718
499.99

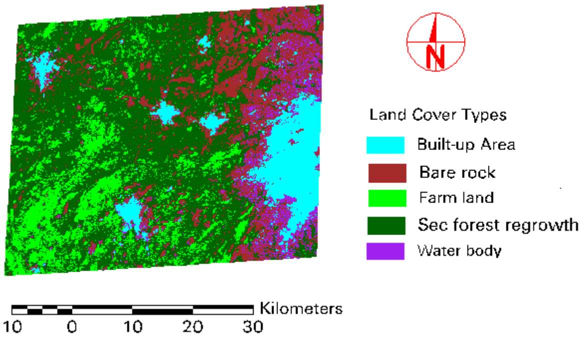


Fig3: Classified print of the NigeriaSat-1 image

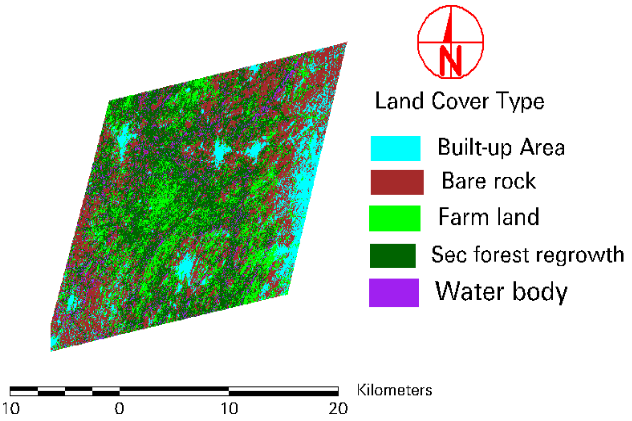


Fig4: Classified print of the SPOT image

The SPOT image had been rectified but was resampled to 32m in order to have the three imageries in the same spatial resolution.

4.3 Land use characterization of the study

Table 4.3 Shows the area covered in hectares, by the five landuse classes' namely built-up areas, bare rocks, farmland and secondary forest regrowth and water body. On the Landsat image, "built-up area'' covered 4,388.15 hectares which represented 9.6% of the total area of land under study. "Bare rock'' covered 8,925.8 hectares (19.4%); "farmland'' 9,461.96 hectares (20.6%); "secondary forest regrowth'' 19,881.9 hectares (43.3%) and water body 3,241.56 hectares (7.1%). On NigeriaSat-1, "built-up area'' covered an area of 4,547.69 hectares which represented 10% of the total land area under study. "Bare rock" was 8,016.2 hectares (18%) "farmland" 6,682.6 hectares (15%); "secondary forest regrowth'' 23,266.60 hectares (52%) and "water body'' 2,326.32 hectares (5%).

With respect to SPOT image, "built-up area'' covered 4,891.85 hectares which was 12% of the total land area under study. "Bare rock'' was 12,887.70 hectares (30%); "farmland" 7,565.11 hectares (18%), "secondary forest regrowth'' 14,665.7 hectares (34%) and "water body'' covered 2,642.84 hectares (6%). In general, the land cover characterizations of the three imageries are over comparable area cover, despite the differences in the spatial resolutions of the images. "Secondary forest regrowth'' is the largest in the three images and farmlands are comparable in size. The main difference is with respect to bare rock which was shown to cover a larger area in SPOT. The difference may be related to the time of imaging. Images taken in dry seasons would reveal more bare surfaces than those taken in the wet season.

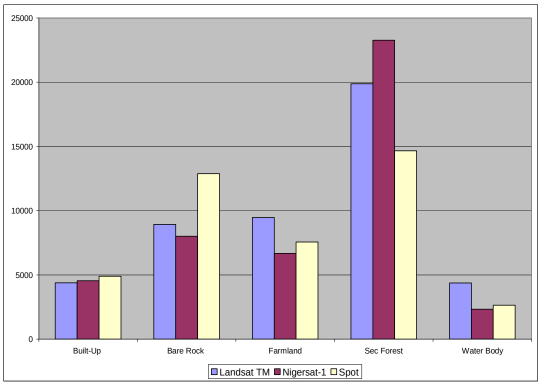


Figure . 5 Histogram showing the three imageries with their respective landuses types.

4.4. Discussion: Variation in land use characterization and Accuracy

The most basic output in remote sensing applications to landuse/landcover studies is land use characterization. This is information about the sizes of various landuse. When such information is available over a long period of time, it allows an assessment of landuse dynamics. Such characterizations are invaluable for the monitoring and managing of land resources and are increasingly vital for natural and regional development in Africa.

Landuse characterization has been most used particularly in the developing countries, (Adesina and Amamoo 1994; Oyinloye and Adesina 2006). Differences in the spatial and spectral characteristics of features make a differentiation of landuse/landcover types possible.

As a result of fundamental differences in the characteristics of the images used for this research, the study has proceeded with the assumptions that land use characterization will some how be different from one image to the other. Thus, SPOT, because of its higher resolution, would be more able to capture information than the two others. Also, Landsat would generate more details than NigeriaSat-1.

In general, the point raised above was demonstrated in this study. For example SPOT image had higher correlation of important wave bands with each other than both Landsat and NigeriaSat-1. This implies that, SPOT captures more information than either of Landsat and NigeriaSat-1. Also with respect to landuse recognition, built-up areas were more efficiently identified and defined on the SPOT image than on the two others. However, the three satellite images were relatively inefficient in defining the other land use categories. It appears from the results that the superiority of enhanced spatial resolution declines with the size of the land use category to be defined.

Table. 7 Areas of Landuse categories on the image

LANUSE/LAND COVER CATEGORIES	LANDSAT		NIGERIASAT-1		SPOT	
	AREA (Ha)	AREA %	AREA (Ha)	AREA %	AREA (Ha)	AREA %
BUILT-UP AREA	4388.15	9.56	4547.69	10.14	4891.85	11.47
BARE ROCK	8925.80	19.44	8016.18	17.88	12887.70	30.22
FARM LAND	9461.96	20.62	6682.62	14.90	7565.11	17.74
SEC. FOREST REGROWTH	19881.90	43.31	23266.60	51.89	14665.70	34.38
WATER BODY	3241.56	7.07	2326.32	5.19	2642.84	6.20
TOTAL	**45899.37**	**100**	**44839.41**	**100**	**42653.20**	**100**

Different reasons may be responsible for the differences in accuracies. One reason may have to do with accurate spatial registration of the intrinsic characteristics of the images. For instance Landsat has a spatial resolution of 30metres, NigeriaSat-1 32 metres and the SPOT 10metres. Chen, et al., (2002) has shown that these can variously have effect on the levels of accuracies obtained from the images. Another reason could be that, spectral characteristics among the different land cover types (e.g. built-up, bare rock) are similar, while spectral variation within the same land cover type or even within the same image might be high (Cushine, 1987).

5. CONCLUSION

This study was conducted with the intention of evaluating the difference in landuse characterization, relative accuracy of feature definitions and the usage of spatial data with Landsat, NigeriaSat -1 and SPOT images. The result of the study supports the knowledge that each image has certain relative advantage over the other. For stance, while NigeriaSat-1 images are shown to be very efficient in the analysis of information within the visible portion of the electromagnetic spectrum, SPOT images are better in the Near Infrared. Information from Landsat images was rather weak at both portions (Visible and NIR) of the Electromagnetic Spectrum. The study also showed that SPOT images have the lowest level of data redundancy of the three image providers. This observation is similar to that of Kuplich, et al (2000), which concluded that SPOT has relatively smaller data redundancy than 'some' other image producers. This is because high correlation between spectral bands is indicative of high degree of information. Spectrally adjacent bands in a multispectral remotely sensed image are often highly correlated. Multiband visible/near-infrared images of landuse areas will show negative correlations between the near-infrared and visible red bands and positive correlations among the visible bands because the spectral characteristics of land use are such that as the vigour or tone of the feature increases, the red reflectance diminishes and the near-infrared reflectance increases.

The study also showed that SPOT images have higher level of accuracy (> 97%) than Landsat and NigeriaSat-1. The reasons for this may be the intrinsic characteristics of the images. Another reason of course, is that spectral characteristics among the different land cover types (e.g. built-up, bare rock) could be similar.

In addition, it was also revealed that the distinguishing spectral characteristics between "farmland'' and "rock surfaces'' and "farmland'' and "secondary forest'' in the Landsat images were relatively poor. However, seamed data sets of NigeriaSat-1 and Landsat images on the one hand, and NigeriaSat-1 and SPOT on the other produce landuse classifications of better accuracies (>80%) than the individual images i.e. SPOT, NigeriaSat-1 or Landsat especially when representing land uses such as built-up area, bare rock, water body and farmland.

Finally, the images differ in their ability to reveal landuse characteristics and differences in spatial resolution may not be a challenge to accuracy and details of reporting depending on the subject of interest.

ACKNOWLEDGEMENT

We are grateful to the National Centre for Remote Sensing and FORMECU for releasing of the satellite images. The authors wish also to thank the department of geography of Obafemi Awolowo University for intellectual support during this research work. Also an appreciation to African Regional Centre for Space Science and Technology Education for their supportive ways to see to the completion of this research work. The authors and the researchers of the literatures we referred to should also get our appreciations. We extend our sincere appreciation to the editor and anonymous reviewers, whose comments are valuable in improving the quality of the article.

REFERENCES:

Adesina, F. A. and Amamoo O. E., (1994): "Land Cover Characterization with SPOT Satellite imagery in the forest areas of Nigeria. *The Nigeria Geographical Journal, New series*, 1, 70- 90.

Anderson, J.R., Hardy, E.T., Roach, J.T. and Witmer, R.E., (1976): A land use and land cover classification system for use with remote sensing data, Professional Paper 964, U.S. Geological Survey, and Washington D.C. Appraising the anatomy and spatial growth of the Bangkok Metropolitan area

Chen, D., Stow, D., (2002): The effect of training strategies on supervised classification at different spatial resolutions. Photogrammetric Engineering and Remote Sensing 68, 1155–116

Cushnie, JL. (1987): "The Interactive Effect of Spatial Resolution and Degree of Internal Variability within Land-Cover Types on Classification Accuracies." *Photogrammetric Enginee Remote Sensing* 8, 1 (1987): 15-29.

Ekiti Investors Handbook, 2002

EKSG (Ekiti State Government) (1997): First Anniversary Celebrations of Ekiti State Government. Government Press, Ado-Ekiti. 22pp.

Gastellu-Etchegorry, J.P., (1990): "An Assessment of SPOT XS and Landsat MSS Data for Digital Classification of Near-Urban Land Cover." *International Journal of Remote Sensing* 11, 2 (1990): 225-235.

Jensen, J.R. (2005): *Digital Image Processing: a Remote Sensing Perspective*, 3rd ed., Prentice Hall.

Jensen, J.R. (2007): *Remote sensing of the environment: an Earth resource perspective*, 2nd ed., Prentice Hall.

Kayode, J. (1999): Phytosociological investigation of composite weeds in abandoned farmlands in Ekiti State, Nigeria. Composite Newsletter 34, 62 – 68

Kuplich, T. M., (2000): Estudo da complementaridade de imagens o´ ticas (Landsat/ TM) e de radar (ERS-1/SAR) na discriminac¸a˜o tema´tica de uso da terra. INPE-5608-TDI/554, MSc dissertation, INPE, Sa˜o Jose´ dos Campos, Brazil

NPC, (2006): Population census and post enumeration survey chart National Population Commission (NPC), Abuja, Nigeria.

Oyinloye, R.O and Adesina F. A., (2006): "Some Aspects of the Growth of Ibadan and Their Implications for Socio-economic Development'' Ife Social Sciences Review Vol 20, No. 1, pp. 113-120

LAND USE DATA GENERALIZATION INDICES BASED ON SCALE AND LANDSCAPE PATTERN

Y.L. Liu [a, b,] *, L.M. Jiao [a, b], Y.F. Liu [a, b]

[a] School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan, 430079 China - (yaolin610, lmjiao027, yfliu610)@163.com
[b] Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan, 430079 China - (yaolin610, lmjiao027, yfliu610)@163.com

KEY WORDS: Land use database, Landscape pattern, Land use data generalization, Indices of land use data generalization, Scale

ABSTRACT:

This paper studies the index system of land parcel generalization which is crucial in land use data generalization. We discuss the macro and micro indices of land use data generalization with consideration of spatial scales and landscape pattern. To quantitatively relate the indices and scale and landscape pattern metrics, land use data samples have been collected at multiple spatial scales in various land use regions across China. Based on statistic analysis, we then generate both macro and micro control rules for land use data generalization at various spatial scales and patterns. Finally, we prove the proposed method to be effective with sample data at county level.

1. INTRODUCTION

Indices for parcels generalization are critical for generating multi-scale land use maps and databases. China is conducting its second nationwide land investigation. The land investigation produces land use maps at county level (1:10,000), which are then generalized into a series of land use maps and databases at smaller spatial scales. These smaller spatial scales range from 1:50,000 to 1:500,000. However, there are not nationwide criteria for land use data generalization.

Previous studies on spatial data generalization usually focused on general threshold values for terrain mapping (Butterfield and McMaster, 1991; Muller and Wang, 1992; Oxenstierna, 1997; Lee, 2001), whereas these research generally do not consider indicators for land parcels generalization. Liu (2002) and Liu et al. (2003) proposed a framework of land use database generalization based on models and rules, and provided some basic criteria, such as the maintenance of area proportion of land use types. Ai and Wu (2000) and Ai et al. (2001, 2002) studied the operators of parcels generalization, and discussed parcel merging based on neighbourhood analysis. Gao et al. (2004) derived certain thematic knowledge for land use data generalization in the form of production rules. Several studies concentrated on indicators of land use data generalization, such as minimum parcel area in land use map, but were limited in local area (Liu, 2005; Chen, 2005; Zhang, 2006). Nevertheless, these works do not take regional landscape pattern in consideration. Previous literatures suffer from two major setbacks. First, determination of threshold values for multi-scale land use data generalization in a large area, such as a nation, remains subjective. Second, there is a general ignorance of landscape pattern in land use data generalization.

This paper develops the index system of multi-scale land use database generalization from macro and micro aspects, whereby estimates the relationship between indices and scale and the relationship between indices and landscape pattern metrics based on typical samples dataset.

2. DATA AND METHODOLOGY

2.1 Data sampling and preprocessing

Two datasets are used in this paper. A nationwide dataset is used to derive model estimates, while a local dataset serves as a case study. The local dataset is land use data of Zigui county in Hubei province, China, at 1:10,000 scale.

The nationwide land use dataset should be representative in terms of both land use pattern and spatial scale. Land resource in China is zoned into twelve land use regions (Li, 2000). These land use regions can be further divided according to different geomorphologic properties (Resource zoning committee of Chinese academy of science 1959). Therefore, we derive our samples based on two principles as follows: a) The amount of samples in each land use region is proportional to the region area. b) Each geomorphologic zone in a land use region should have at least one sample. Finally we selected land use maps of 51 counties at the scale of 1:10,000, 1:50,000, 1:100,000, and 1:250,000. We also incorporate another 23 subdivided maps for land use at the scale of 1:500,000 in our sample dataset. Our sample counties, which distribute in different land use regions across China, are shown in Figure 1.

We conduct data preprocessing because that our samples are stored in different formats with different reference systems. These data preprocessing includes verification of database, transformation of data formats, transformation of coordinate system, and normalization of coding system for land use types.

* Corresponding author.

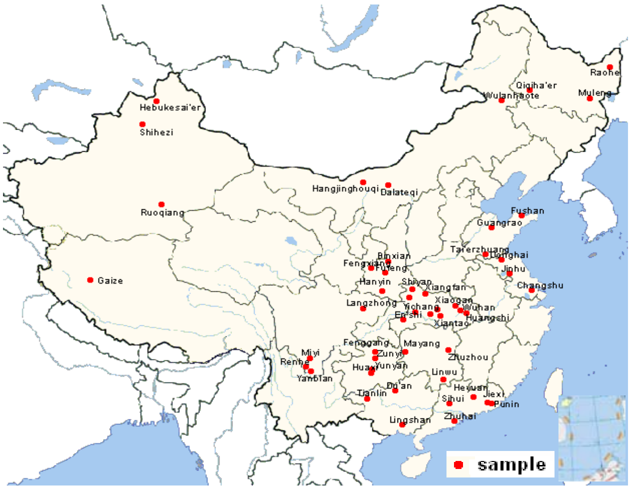


Figure 1 Map of land use sample dataset

2.2 Index system of land use data generalization

Traditionally, researchers employ indicators for land use data generalization at micro scale, such as minimum parcel area, and minimum distance between parcels. Nevertheless we should incorporate certain indices at macro level, such as area proportion of land use types, and spatial distribution characteristics. These macro indices can be used to control generalization operations and to evaluate the generalization result. Hence we build the index system of land use data generalization from both macro-perspective and micro-perspective.

Macro indices for land use data generalization include map load, area proportion of different land use types, and semantic characteristics. Map load describes the map content from a macro perspective. There are at least three kinds of map loads in land use maps: maximum map load, optimum map load, and features map load. The area proportion of land use types serves as an important threshold in land use data generalization because that generalization of parcels will lead to change in area proportion. Spatial contrast of land use types also needs to be maintained since land use maps are primarily used to express the spatial pattern of land use types. Additionally, semantic characteristics are important for parcel merging control in land use data generalization. The hierarchy of land use types is often used to analyze the similarity between two land use types.

Micro thresholds for land use data generalization include minimum parcel area, minimum distance between parcels, and minimum bend diameter. Minimum parcel area reflects the importance of land use type and landscape pattern. Hence different land use types can have different area thresholds at the same spatial scale.

2.3 Scale effect on indices of land use data generalization

Map scale and land use pattern exert influence on indices of land use data generalization at different strength. Map scale has determinant influence on generalization indices, thus the maps with different scales in the same area will have significantly different indices. Land use pattern has relative smaller impacts than map scale, and its effects can be perceived in a large region with various terrains at the same scale.

We employ exploratory statistics and non-linear regression model to evaluate the relationship between generalization indices and map scale. For convenience, we replace map scale with the scale's denominator. A logarithm function was chosen as a non-linear regression function type according to the dataset.

2.4 Land use pattern effect on indices of land use data generalization

In landscape ecology, patch-based indices are developed to quantify landscape characteristics, such as index of landscape diversity, index of landscape dominance, index of landscape homogeneity, and index of landscape fragmentation. Comparing land use parcels with ecology patches, land use pattern indices are defined as follows.

2.4.1 Design of the land use pattern indices

We define land use indices according to the comparison between land use parcels and landscape patches. In addition to aforementioned indices employed in landscape ecology, we propose two new indices to describe a specific land use type: dominance index of land use type and fragmentation index of land use type.

①Index of land use diversity (H)

This index describes the diversity of land use types based on informatics.

$$H = -\sum_{k=1}^{m} P_k Iog_2(P_k) \qquad (1)$$

where H is the index of land use diversity , m is the number of land use types, and p_k is the percentage of land use type k . If $m = 1$, then $H = 0$, the minimum. When m goes to infinity, H reaches its maximum.

②Index of land use dominance (D)

The index of land use dominance measures the degree of how the land use was dominated by one or two types.

$$\begin{aligned} D &= H_{\max} + \sum_{k=1}^{m} P_k Iog_2(P_k) \\ &= \log_2(m) + \sum_{k=1}^{m} P_k Iog_2(P_k) \end{aligned} \qquad (2)$$

where D refers to the index of land use dominance, H_{max} represents the maximum of the index of land use diversity, and p_k is the percentage of land use type k .

③Index of land use homogeneity (E)

This index describes the homogeneity of land use types in a land use pattern (Wang, 2003), which is given by:

$$E = H/H_{\max} = -\lg\left[\sum_{k=1}^{m}(P_k)^2\right] \Big/ \lg(m) \qquad (3)$$

④Index of land use fragmentation(C)

This index describes the degree of fragmentation of land use pattern,

$$C = N / A \qquad (4)$$

where N and A are the number of land use parcels and the total area of the studied region, respectively.

⑤Dominance index of land use type（D_t）

This index measures how much the land use is dominated by one or several large parcels, and is expressed as:

$$\begin{aligned} D_t &= H_{t\max} + \sum_{i=1}^{n} P_i Iog_2(P_i) \\ &= \log_2(n) + \sum_{i=1}^{n} P_i Iog_2(P_i) \end{aligned} \qquad (5)$$

where D_t is the dominance index of land use type t, n is the number of parcels belonging to land use type t, $H_{t\max}$ is the maximum of the index of diversity of land use type t, and p_i is the area percentage of parcel i of land use type t.

⑥Fragmentation index of land use type（C_t）

This index captures the fragmentation degree of the distribution of land use types,

$$C_t = N_t / A_t \qquad (6)$$

where N_t represents the number of land use parcels belonging to land use type t, and A_t is the total area of the land use type t in the studied region.

2.4.2 Analysis of the effect of land use pattern metrics on the indices of land use data generalization

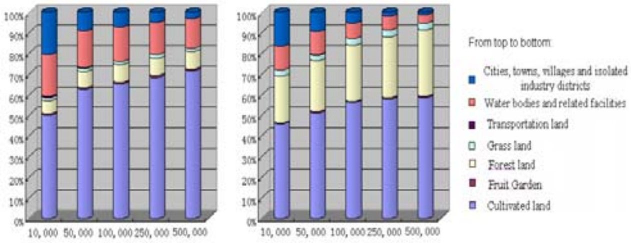
We use correlation analysis to find the major explanatory factors of the thresholds for land use data generalization. Consequently we can set up the regression model between the threshold and major explanatory variables. We can then determine the indices for land use data generalization appropriately based on model estimates.

3. RESULT AND ANALYSIS

3.1 Scale effect on indices of land use data generalization in China

3.1.1 Scale effect of macro thresholds

There are three driving-forces for changes in area proportions of land use types in generalization. The first cause is collapse of parcels. For example, polygonal residential area is simplified as points, polygonal roads or rivers are simplified as lines during generalization. The second is boundary simplification of parcels whereas a third cause is the elimination, aggregation, amalgamation or exaggeration of parcels. Figure 2 illustrates the changing of area proportion of land use types in Middle and Lower Reaches of Changjiang River at different spatial scales.



(1) Plain area (2) Hilly and mountainous area

Figure 2 Multi-scale changing of area proportion of land use types in Middle and Lower Reaches of Changjiang River

Figure 2 shows that the map area of cities, towns, villages, isolated industrial districts and the area of water bodies and water resource facilities decrease considerably in both plain regions and hilly and mountainous regions. The area of the map objects of transportation land decreases as well. However, the decrease is not obvious in the figure since the total area of transportation land is relatively small. On the contrary, the area of the map objects of cultivated land in plain region and the area of the ones of forest land in hilly region increase. We produce the ranges of area proportion changes in land use data generalization based on our nationwide samples. The result is summarized in Table 2.

Table 1 Changing rate of area proportion of land use types in generalization (%)

Scales of before and after generalization	Changing rate of area proportion of land use types
1:10k~1:50k	12~20
1:50k~1:100k	4~6
1:100k~1:250k	4~7
1:250k~1:500k	4~6
1:10k~1:500k (accumulative)	15~30

Map load is related to map scale and land use pattern. When map scale decreases, map area and map content increase, and hence the map load increases. Even at the same scale, Map load is larger in regions with more land use types and fragmentary land use distribution. Therefore map load is correlated with land use fragmentation index at the same spatial scale. Land use fragmentation index increases along with the decrease of map scale, whereas the ratios among land use fragmentation index in different areas are almost constant. At the scale of 1:10000, land use fragmentation index can be categorized into three classes: low-level fragmentation (>0.5), medium-level fragmentation (0.3~0.4), and high-level fragmentation (<0.3). We analyze the scale effect of map load in three subdivisions of land use fragmentation. The appropriate total map load of land use maps and map load of parcel features in land use maps are shown in Table 3 and Table 4 respectively. These ranges and averages can be used as benchmarks of area proportion control and map load control in generalization of land use maps.

Table 2 Range of suitable total map load of land use maps (%)

Fragm-entation	1:10k	1:50k	1:100k	1:250k	1:500k
Low	3~7	13~17	17~21	22~26	26~30
Medium	5~9	15~19	20~24	26~30	31~35
High	5~9	17~21	23~27	30~34	35~39

Table 3 Average of suitable map load of parcel features (%)

Fragmentation	1:10k	1:50k	1:100k	1:250k	1:500k
Low	1.100	2.227	2.712	3.353	3.838
Medium	1.200	2.327	2.812	3.453	3.938
High	1.700	2.666	3.082	3.631	4.047

3.1.2 Scale effect of microscopic thresholds

There are four factors influencing the minimum parcel area in generalization. The first one is the precision required by mapping purpose. The second is the resolution determined by map scale. The third is the importance of land use types, while the fourth is the spatial pattern of land use. We employ non-linear regression to fit the relationship between minimum parcel area and map scale. Taking cultivated land as an example, the scatterplot of the samples is shown in Figure 3. The samples shown in the figure are tidied by eliminating the outliers with two times of variance method. Logarithm function is selected to fit the relationship between minimum parcel area and scale.

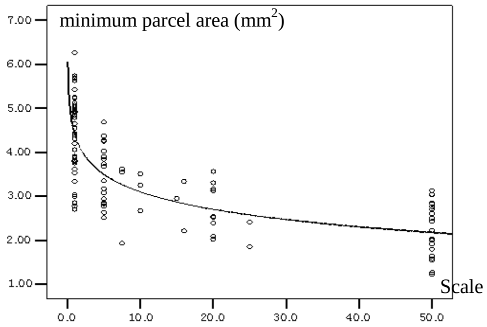


Figure 3 Regression analyses between minimum parcel area of cultivated land and map scale

The regression function is

$$Y=-0.574*\ln(X)+4.424 \qquad (7)$$

where Y is the minimum parcel area of cultivated land, X is the denominator of map scale. The independent variable explains 61.1 of the variations of minimum parcel area ($R^2 = 0.611$). The F-ratio of 155.310,indicates that the model is well-fitted. Moreover, we generate the regression functions for other land use types with the same routine, and the model estimates and associated statistics are presented in Table 5.

Table 4 Regression results between minimum parcel area and map scale

Land use type	Regression (Y =)	R	F
Cultivated land	-0.574*ln(X)+4.424	0.781	155.310
Fruit Garden	-0.875*ln(X)+6.300	0.760	102.530
Forest land	-1.240*ln(X)+8.977	0.800	159.526
Grass land	-1.187*ln(X)+8.860	0.761	109.999
Transportation land	-2.151*ln(X)+7.571	0.806	89.084
Water bodies	-1.626*ln(X)+8.979	0.766	140.662
Others	-1.670*ln(X)+9.860	0.719	77.208
cities, towns, villages, industry districts	-0.626*ln(X)+4.035	0.802	176.121

3.2 Land use pattern effect on indices of land use data generalization

3.2.1 Land use pattern effect of macro indices

The changing of area proportion of land use types is influenced by land use pattern. Regions with higher fragmentation values usually have larger changes in the area proportion. Thus as for Table 2, we should employ the lower limit in regions with a lower fragmentation, and the upper limit in regions with a higher fragmentation. Map load is influenced primarily by the index of land use fragmentation, see Table 4.

3.2.2 Land use pattern effect of microscopic indices

Minimum parcel area correlates with not only map scale but also land use pattern. We employ correlation analysis to explore the land use pattern indices which influence the minimum parcel area. Then we use regression analysis to quantify these influences. We take minimum parcel area of cultivated land in 1:50,000 map as an example to describe the analytical procedure, and the correlation analysis results are demonstrated in Table 6.

Table 5 Correlation analysis between minimum parcel area of cultivated land and land use pattern indices (Scale 1:50,000)

Indices	Pearson coef.	Sig. (2-tailed)
Diversity index (H)	-0.510	0.052
Dominance index (D)	0.440	0.101
Homogeneity index (E)	-0.487	0.065
Fragmentation index (C)	-0.678	0.024
Dominance index of land use type (D_t)	0.325	0.237
Fragmentation index of land use type (C_t)	-0.799	0.018

Table 6 reveals that fragmentation index of cultivated land and general index of land use fragmentation have the most significant influences on the minimum parcel area of cultivated land at 1:50,000 scale. In contrast, the diversity index, predominant index, homogeneity index and dominance index of land use type are not significant. Therefore fragmentation index of cultivated land is adopted as explanatory variable in the regression analysis at the second stage, and the linear regression takes the following form:

$$Y=-18.843* X+4.532 \qquad (8)$$

Where Y is the minimum parcel area threshold for cultivated land, X is the fragmentation index of cultivated land. The Correlation coefficient R^2, F-ratio, and p-ratio are 0.638. 5.112, and 0.008 respectively, all of which are statistically significant and show that the model is well-fitted. The model estimates reveals that one percent change in the fragmentation index has a marginal effects of 0.19mm^2 in the minimum parcel area threshold. The range of the fragmentation index of cultivated land is between 0.25% and 5.00%, and therefore the theoretical range of the minimum parcel area threshold for cultivated land is from 3.6 mm^2 to 4.5 mm^2.

Table 6 changing of minimum parcel area with the fragmentation index of land use type

Land use type	Decrease of minimum parcel area for each percent increase of the fragmentation index of land use type (mm2) / The average of fragmentation index of land use type (%)			
	1:50k	1:100k	1:250k	1:500k
Cultivated land	0.19/ 1.62	0.17/ 2.51	0.14/ 3.41	0.12/ 5.45
Fruit Garden	0.20/ 1.86	0.17/ 2.56	0.13/ 3.53	0.10/ 4.76
Forest	0.11/ 1.08	0.10/ 2.02	0.09/ 2.72	0.06/ 3.24
Grass	0.13/ 1.04	0.12/ 2.01	0.10/ 2.71	0.07/ 3.12
Transportation	0.24/ 0.65	0.22/ 1.05	0.19/ 1.65	0.17/ 1.72
Water	0.20/ 2.03	0.20/ 2.44	0.18/ 2.94	0.15/ 3.24
Cities, towns, villages and industry districts	0.25/ 2.15	0.23/ 2.65	0.21/ 3.40	0.19/ 3.87
Others	0.10/ 0.98	0.10/ 1.68	0.08/ 2.38	0.06/ 2.43

3.3 Experiment of land use data generalization

We use a local land use database at 1:10,000 scale, as mentioned before, in our case study. The study area has a relative large fragmentation index of land use pattern (0.53%). The control range of area proportion changing of land use types are set to upper values in Table 2. The minimum parcel area of cultivated land for 1:50,000 map is 4.0mm2 according to the regression functions in Table 3. The number of cultivated parcels in 1:50,000 map is estimated by fractal selection method, and the fragmentation index of cultivated land (3.43%) is computed using equation 6. Therefore the minimum cultivated parcel area in 1:50,000 map is adjusted to 3.7mm2 according to Table 5. The other thresholds of minimum parcel area for other land use types and other map scales can be estimated in the same way. We present the minimum parcel area for different land use types at various spatial scales in Table 8.

Table 7 The minimum parcel area in multi-scale land use maps (mm^2)

Land use type	1:50k	1:100k	1:250k	1:500k
Cultivated land	3.7	3.0	2.5	2.2
Fruit Garden	4.3	3.2	2.7	2.4
Forest	8.0	7.2	6.5	6.0
Grass	7.8	6.5	6.1	5.8
Transportation	4.5	3.5	—	—
Water	7.0	6.1	5.3	4.7
Cities, towns, villages and industry districts	3.3	2.8	2.3	2.0
Others	8.5	7.7	7.1	6.5

The land use data generalization in the study area is implemented based on these indices. Some results of a part of the area are shown in Figure 4. Macroscopic indices prior to and after generalization are shown in Table 9.

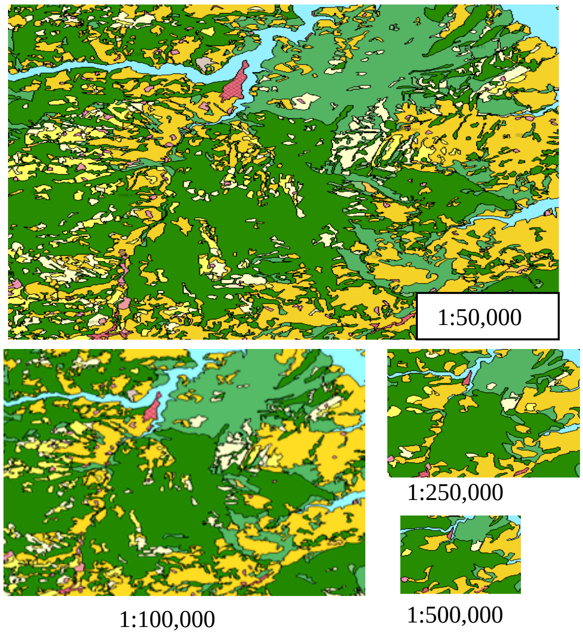


Figure 4 Results of land use data generalization

Table 8 Changing of macroscopic indices of land use data generalization

Index	1:10k	1:50k	1:100 k	1:250 k	1:500 k
Total map load (%)	7.9	19.8	25.6	31.0	37.2
Map load of parcels (%)	1.5	2.6	3.1	3.7	4.1
Maximum extent of the range of area proportion change	—	16%	6%	6%	5%

The comparison of the observations in the experiment and theoretical indices are shown in Figure 5 to 7.

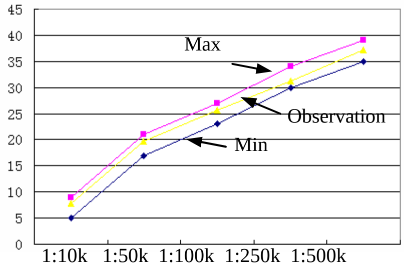


Figure 5 Total map load

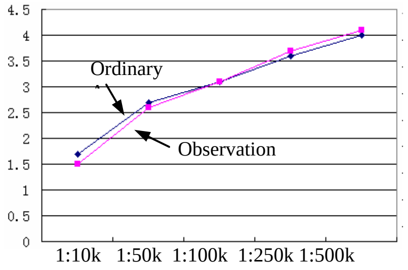


Figure 6 Map load of parcels

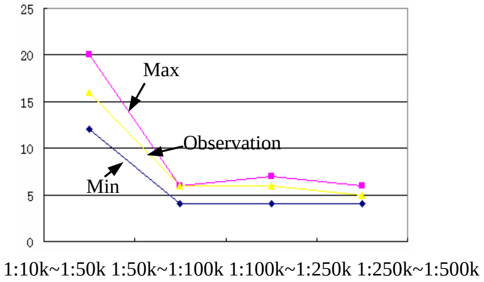


Figure 7 Change of area proportion of land use types

Both the total map load and the parcels map load increase as the scale increases, however the increase rate decreases gradually. There is a significant change of area proportion of land use types when land use data is generalized from 1:10,000 to 1:50,000, which is about 16%. In contrast, the changes of area proportion in generalization among other scales ranges from 5% to 6%. These observations in the experiment coincide with the aforementioned rules in this paper.

4. CONCLUSION

Land use database generalization indices are restricted by visual discrimination ability, and influenced by subjective factors, map usage, importance of land use types, etc (Zhu, 2004). It is hard to formulate the relationships between land use database generalization indices and these factors individually. The paper introduces land use pattern metrics into the determination of land use generalization indices, and proposes an analysis framework of scale and land use pattern effects on the indices. The paper highlights an empirical method to determine land use database indices by examining the relationships between the indices and map scale and land use pattern metrics. Taking China as an example, our major findings based on countrywide land use samples include: Map scale dominates the determination of land use database generalization indices. The map area proportions of cities, towns, villages and isolated industrial districts, water bodies and related facilities, and transportation land decrease significantly with map scale, while the map area of cultivated land in flat regions and forest in hilly and mountainous regions increase gradually. Total map load and parcels map load increase with progressive land use database generalization. The minimum parcel area was observed to decrease with map scale, and the logarithm functions between minimum parcel area and map scale for different land use types were formulated. Land use pattern exerts impact on the land use database generalization indices when at the same map scale. It was observed that higher land use fragmentation increases the land use area proportion change in generalization and map load at the same scale. It was found that the minimum parcel area is correlated significantly with land use type fragmentation index, and the changing rule of minimum parcel area against land use type fragmentation index was generated at each main map scale. As a result, the thresholds of the ratio of land use area proportion change were generated, and the thresholds of total map load and suitable parcels map load during generalization with consideration of different land use fragmentation levels were also produced. The minimum parcel area at each main scale can be calculated based on our quantified rules. The experiment of land use database generalization indicates that our indices are applicable. At the same time, the experiment also proves implicitly the methodology of this paper and shows the results to be reasonable.

Further research should include rules for computer-aided land use data generalization and land use data auto-generalization.

ACKNOWLEDGEMENTS

This study was funded by National Science Foundation, Project number: 40771168

REFERENCES

Ai T.H., Guo R.Z., Chen X.D. 2001. Simplification and Aggregation of Polygon Object Supported by Delaunay Triangulation Structure. Journal of Image and Graphics, 7, pp. 703~709.

Ai T.H., Liu Y.L., 2002. Aggregation and Amalgamation in Land-use Data Generalization, Geomatics and Information Science of Wuhan University, 27(5), pp. 486~492.

Ai T.H., Wu H.H., 2000. Consistency Correction of Shared Boundary between Adjacent Polygons, Journal of Wuhan Technical University of Surveying and Mapping, (5), pp. 426~431.

Buttenfield B.P., McMaster R.B., 1991. Map generalization, pp. making rules for knowledge representation. London, Longman.

Chen X.W., 2005. Structural models and algorithms in land use database generalization, dissertation of Wuhan University.

Gao W.X., Gong J.Y., Li Z.L., 2004. Thematic knowledge for the generalization of land use data. Cartographic Journal, 41(3), pp. 245-252.

Lee D., 2001. Generalization in the new generalization of GIS. Proceedings of ICC, Beijing, China.

Li Y., 2000. China Land Resource, China Land Press, Beijing.

Liu X.H., 2005. Method and Practice of land use data generalization, dissertation of China Agriculture University.

Liu Y.L., 2002. Categorical database generalization in GIS. Wageningen University, ITC Dissertation.

Liu Y.L., Molenaar, M., Ai T.H., Liu Y.F., 2003. Categorical database generalization aided by data model, The 21st International Cartographic Conference, Durban.

Muller J.C., Wang Z., 1992. Area_patch Generalization: A Competitive Approach. The Cartographic Journal, 29(2), pp. 137~144.

Oxenstierna, A., 1997. Generalization rules for database-driven cartography. In: Proceedings ICC, Stockholm.

Qi Q.W., Jiang L.L., 2001. Research on the Index System and Knowledge Rules for Geographic-Feature-Oriented Generalization, Progress in Geography, 20, pp. 1~12.

Resource zoning committee of Chinese academy of science, 1959. China Geomorphology zoning, Science Press, Beijing, China.

Wang Q., Wu H.H., 1996. Fractal method in generalization of polygon groups, Journal of Wuhan Technical University of Surveying And Mapping, 1, pp. 59~63.

Wang S.Y., Zhang Z.X., Zhou Q.B., Liu B., Wang, C.Y., 2003. Analysis of landscape patterns and driving factors of land use in China, Acta Ecologica Sinica, 23(4), pp. 649~656.

Zhang W., 2006. Research on methodology of land use map generalization, dissertation of Zhejiang University.

Zhu G.R., 2004. Cartography, Wuhan University Press, Wuhan, China.

EVALUATION AND FORECAST OF HUMAN IMPACTS BASED ON LAND USE CHANGES USING MULTI-TEMPORAL SATELLITE IMAGERY AND GIS: A CASE STUDY ON ZANJAN, IRAN (1984-2009)

Mohsen Ahadnejad *[a], Ali Reza Rabet[b]

[a] Assistance Professor, Dept. of Geography, Zanjan University, Iran- E-mail :ahadnejad@gmail.com

[b] PhD student in Geography and Rural planning, Peyam-e-Nour University, Tehran Branch- E-mail :Rabet2001IRAN@yahoo.com

Key Words : Fuzzy ARTMAP, Cellular Automata, Markov Chain, Land-use Change Detection

ABSTRACT:

Land use and land cover change due to human activities in a time sequence .Detection of such changes may help decision makers and planners to understand the factors in land use and land cover changes in order to take effective and useful measures .Remote sensing and GIS techniques may be used as efficient tools to detect and assess land use changes.

In recent years, a considerable land use changes have occurred in the greater Zanjan area .In order to understand the type and rate of changes in this area, Landsat TM images captured in 1984 and 2009 have been selected for comparison.

First, geometric correction and contrast stretch are applied .In order to detect and evaluate land use changes, image differencing, principal component analyses and Fuzzy ARTMAP classification method are applied .Finally, the results of land cover classification for three different times are compared to reveal land use changes .Then, combined Cellular Automata with Markov Chain analysis is employed to forecast of human impacts on land use change until 2020 in Zanjan area

The results of the present study disclose that about 36 percents of the total area changed their land use, e.g., changing agricultural land, orchard and bare land to settlements, construction of industrial areas and highways .The crop pattern also changes, such as orchard land to agricultural land and vice versa .The mentioned changes have occurred within last 25 years in Zanjan city and its surrounding area.

1.Introduction:

The protection of global environment is one of the most critical problems and it is related to several factors, such as population increase, depletion of natural resources, environmental pollution and land use planning .Presently unplanned changes of land use have become a major problem .Most of the land use changes occur without clear and logical planning, paying no attention to their environmental impacts .Floods and air pollution in large cities as well as deforestation, urban growth, soil erosion and desertification are all consequences of mismanaged planning without considering environmental impacts.

Many researchers have employed satellite imagery for land use mapping as well as change detection .Sunarar (1998) has compared the results of five different techniques :band combination, subtraction, band division, principal component analysis and classification, in Ekitally, Turkey .This study revealed that the principal component analysis (PCA) shows better results comparing with classification results .Gupta and Parakash (1998) used a combined method of colour composite, band subtraction, band division and supervised classification to prepare a land-use map for change detection in a coal-mining district in India .They concluded that the supervised classification gives better results for detecting changes .Ahandejad (2002) used PCA, image differencing and classification methods for change detection in Maragheh region, Iran .He concluded that a crosstab method and a comparison image classification method are very suitable for land use change assessment .Neshat (2002) employed Markov Chain to detect the change of forest areas to urban use in Golestan province, Iran .

In the present research, supervised classification based on Fuzzy Artmap is employed to detect land use changes occurred in the Zanjan area, Iran .For forecasting human impacts on land-use change until 2020, both Cellular Automata and Markov Chain are employed.

2.Study Area and Methods of Study:

The study area is located between 36° 38′ 56” to 36° 42′ 22” N and 48° 25′ 42” to 48° 33′ 05” E .The area covers Zanjan city and its surrounding area with 7180 hectares .The study area comprises two topographic units’ foothill and plain . Zanjan population in 1986 was about 215,458 people and its population has been reached to 349,713 people in 2006, the population growth rate in this period was about 3.93 percent . The main reason to select this area is that considerable land-use changes have occurred due to urban developments, rural developments, and industrial developments in the east, west and south areas, and that major changes in the crop pattern are ongoing.

3.Material and methods:

In this paper, Landsat TM images captured in 1984 and 2009 are employed for digital image processing .Figure 1 shows Landsat TM image were used in this study .Also Figure 2 shows the flowchart of this study.

*- Corresponding author

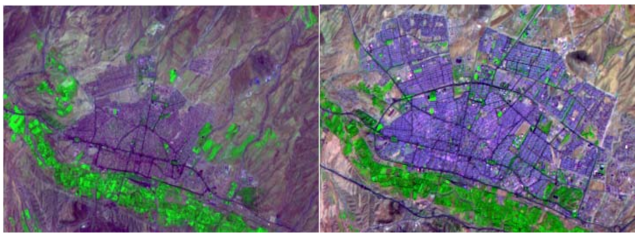


Figure1 .Landsat TM image from case study area in 1984(left) and 2009(right)

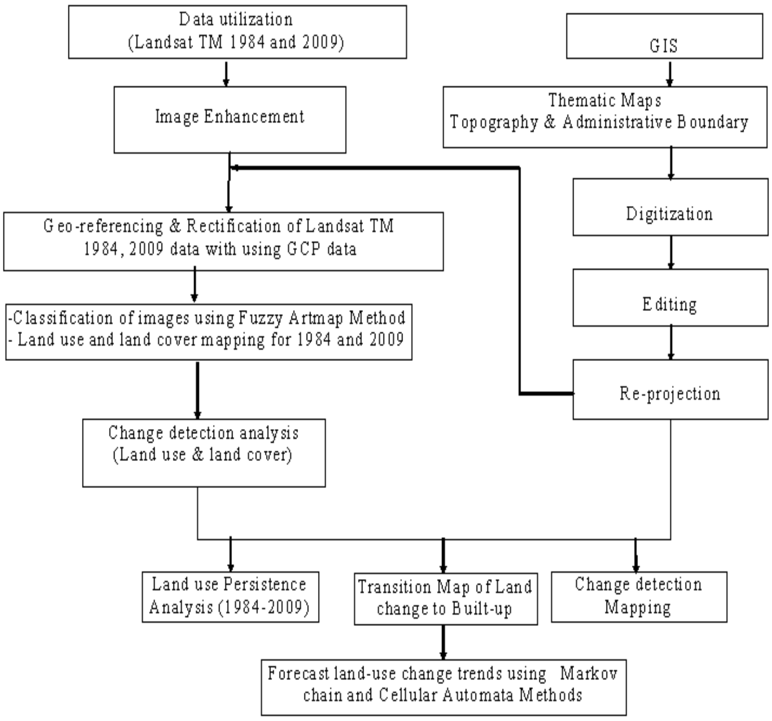

Data utilization
(Landsat TM 1984 and 2009)
GIS
Image Enhancement
Thematic Maps
Topography & Administrative Boundary
Digitization
Geo-referencing & Rectification of Landsat TM 1984, 2009 data with using GCP data
Editing
-Classification of images using Fuzzy Artmap Method
- Land use and land cover mapping for 1984 and 2009
Re-projection
Change detection analysis
(Land use & land cover)
Land use Persistence Analysis (1984-2009)
Transition Map of Land change to Built-up
Change detection Mapping
Forecast land-use change trends using Markov chain and Cellular Automata Methods

Figure 2 . Flow chart showing the major steps of this research

4.Classification:

Various methods have been employed for classification of satellite imagery .Recently, artificial fuzzy methods are used widely because they show very high accuracy in comparison with the conventional ones like Maximum Likelihood Classification (MLC), Minimum Distance Classification, and Parallelepiped Classification.

In this paper, the fuzzy adaptive resonance theory (Fuzzy Artmap(is employed for image classification. First, 741 RGB color composites of Landsat images were prepared .Then, training areas were selected for 6 land use and Land cover classes, which are built-up area, orchards, irrigated agriculture land, dry farming, water, regolith and waste land . These training areas were determined, referring to aerial photographs and GIS thematic maps. To assess the accuracy of classification, topographic maps and aerial photos were employed .Overall accuracy was estimated to be around 96 .% Figures 3 and 4 shows the results of land use classification and Table 1 shows the summary of the classification.

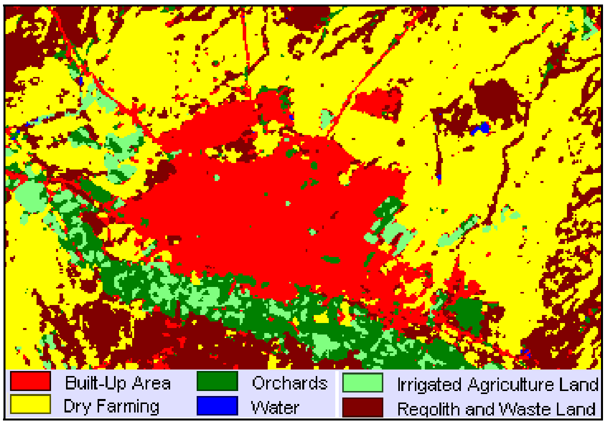


Figure 3 :Result of land use classification for Zanjan, Iran using Landsat TM image captured in 1984

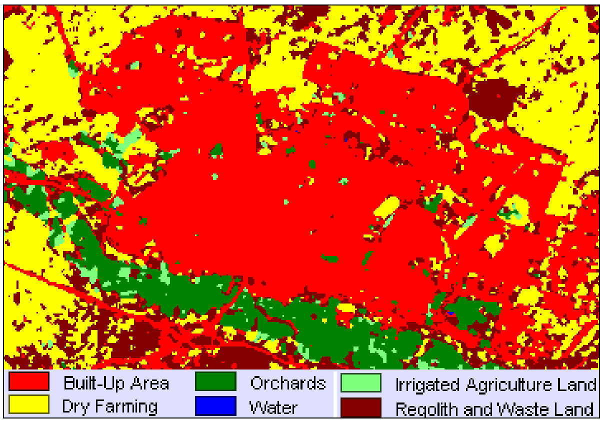


Figure 4 .Result of land use classification using Landsat ETM + image captured in 2009

Class	Land use Type	1984	2009
1	Built-Up Area	1417.06	3377.88
2	Dry Farming	3355.60	1987.65
3	Orchards	614.12	609.21
4	Water	2.14	0.9
5	Irrigated Agriculture Land	355.65	124.65
6	Regolith and Waste Land	1435.72	1080
	Total-Hectare	7180.29	7180.29

Table 1 .Summary of image classification performed in this study(Hectare)

5.Comparison of classification results:

The classification results for the two different times revealed that the land use of the target area has changed about 36 % during the period of 1984-2009 .Table 2 shows the estimated land use transitions based on the comparison of the classification results for the 1984 and 2009 images .Figure 4 shows the areas whose land use has changed to built-up ones in these periods .More than 60 %of the area that belongs to built-up changes to dry farming and waste areas .Dry land farming attains the least changes 27.68 %in this period .

The results also show that built-up area changed from 1417.06 hectare in 1984 to 3377.88 hectare in 2009 .The increase is mainly due to the needs of settlements in Zanjan City because its population has increased from 215,458 in 1986 to 349,713 in 2006 .New suburban areas, such as Sayan, Elahieh,Amir Kabir,GolShahr and Kazemieh, have also developed in the period.

Land use& Land Cover	Built-Up Area	Dry Farming	Orchard	Water	Irrigated Agriculture Land	Regolith and Waste Land	Total	Change
Built-Up Area	1417.06	0	0	0.54	0	0	1476.27	20
Dry Farming	1444.32	1380.06	30.33	0.27	27.54	448.74	3331.26	46.39
Orchards	21.15	107.19	384.93	0	27.27	69.12	609.66	8.48
Water	2.7	3.6	0	0	0	2.25	8.55	0.1
Irrigated Agriculture Land	57.51	37.26	161.28	0	58.68	38.34	353.07	4.92
Regolith and Waste Land	376.47	459.54	32.67	0.09	11.16	521.55	1401.48	19.52
Total	3377.88	1987.65	609.21	0.9	124.65	1080	7180.29	
Change	47.04	27.68	8.48	0.01	1.74	15.04		100

(Row related to 1984 land use and Column related to 2009 land use)

Table 2 .Estimated land use transitions in Zanjan area between 1984 and 2009 (Hectare)

6.Analysis of Land use Transition to Built-Up Area

The Results of Land use changes analysis show that in case study area dry farming and regolith and waste land have most change to built-up area that respectively 1444.32 and 376.47 hectares .Also water body and orchards have minimum changes to built-up area that respectively 2.7 and 21.15 hectares .

In totally in Land use and Land cover changes in 1984-2009, Built-up area have maximum changes with 47.04 percent and minimum changes related to water body with 0.01 percent changes .Figure 5 show the areas that have changed to built-up ones in the period of 1984-2009.

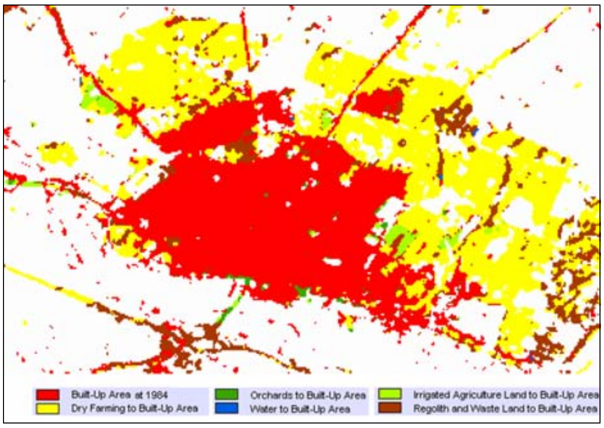

🟥 Built-Up Area at 1984
🟨 Dry Farming to Built-Up Area
🟩 Orchards to Built-Up Area
🟦 Water to Built-Up Area
🟩 Irrigated Agriculture Land to Built-Up Area
🟫 Regolith and Waste Land to Built-Up Area

Figure 5 .The areas that have changed to built-up ones in the period of 1984-2009

7.Land use Persistence and Changes Analysis

One of other analysis in this paper related to land use persistence in the period of 1984-2009 in our case study .It means that how much of land use and land cover and what areas have persistence in during of study periods and has not changes .According to analysis in this case study area about 3848.87 hectares of land use and land cover have not any changes and 3383 hectare of land use and land cover has been changed in the study period 1984-2009.

In between land use built-up area with 1486 hectare has most persistence in comparing with another land use and Irrigated Agriculture Land with 59 hectare has lowest persistence in our case study area .Also dry farming with 1961.6 hectare has maximum changes and orchards with 223.38 hectare have minimum changes in Zanjan area between 1984 and 2009 . Table 3 and figure 6 and 7 shows that spatial distribution map of land use and land cover persistence and Land use changes in the period of 1984-2009.

ID	Type	Persistence	Changes
1	Built-Up Area	1417.51	0.00
2	Dry Farming	1390.14	1961.60
3	Orchards	387.74	223.38
5	Irrigated Agriculture Land	59.11	291.54
6	Regolith and Waste Land	525.36	907.36
	Total	3779.87	3383.88

Table3 .Estimated land use persistence and changes in Zanjan Area between 1984 and 2009

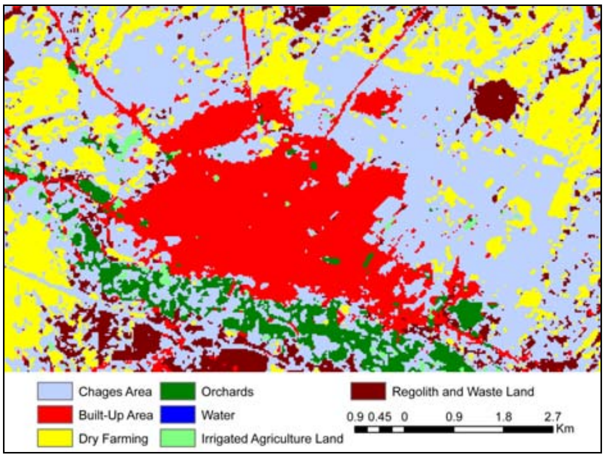


Figure 6 .The areas have that Land use persistence between 1984 -2009

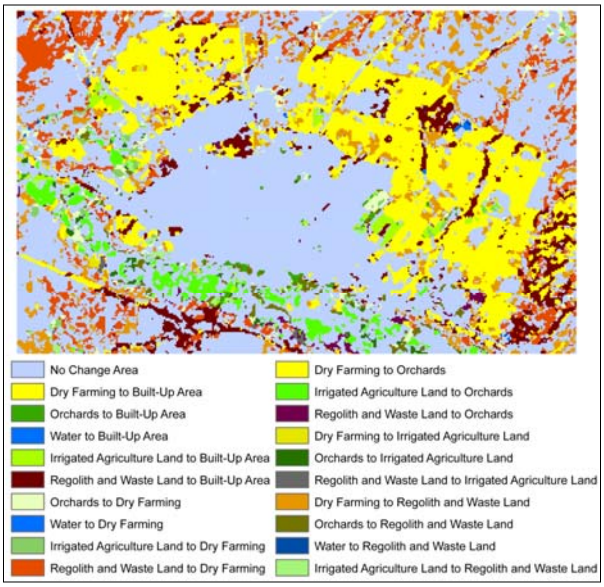

No Change Area
Dry Farming to Built-Up Area
Orchards to Built-Up Area
Water to Built-Up Area
Irrigated Agriculture Land to Built-Up Area
Regolith and Waste Land to Built-Up Area
Orchards to Dry Farming
Water to Dry Farming
Irrigated Agriculture Land to Dry Farming
Regolith and Waste Land to Dry Farming
Dry Farming to Orchards
Irrigated Agriculture Land to Orchards
Regolith and Waste Land to Orchards
Dry Farming to Irrigated Agriculture Land
Orchards to Irrigated Agriculture Land
Regolith and Waste Land to Irrigated Agriculture Land
Dry Farming to Regolith and Waste Land
Orchards to Regolith and Waste Land
Water to Regolith and Waste Land
Irrigated Agriculture Land to Regolith and Waste Land

Figure 7.The areas have that Land use changes between 1984 -2009

8.Prediction of the trends of land use changes :

The other object of this paper is to predict the trend of land use changes in the future .Many methods can be applied to predict the trend .In this paper, two methods are used.

8.1 Markov chain

The Markov chain method analyzes a pair of land cover images and outputs a transition probability matrix, a transition area matrix, and a set of conditional probability images .The transition probability matrix shows the probability that one land-use class will change to the others . The transition area matrix tells the number of pixels that are expected to change from one class to the others over the specified period .

The conditional probability images illustrate the probability that each land cover type would be found after a specific time passes .These images are calculated as projections from the two input land cover images .The output conditional probability images can be used as direct input for specification of the prior probabilities in Maximum Likelihood Classification of remotely sensed imagery (such as with the MAXLIKE and BAYCLASS modules) .A raster group file is also created listing all the conditional probability images .

In this study, a series of image processing was performed to predict the trend of land use change in 2020 (Table 4). The result shows that the probability to change to Built-up area is highest .Figure 8 shows the probability that the area will be converted to Build-up area in 2020.

	Built-Up Area	Dry Farming	Orchards	Water	Irrigated Agriculture Land	Regolith and Waste Land
Built-Up Area	0.9998	0	0	0.0002	0	0
Dry Farming	0.2715	0.5708	0.0024	0	0.0094	0.1458
Orchards	0	0.1072	0.7806	0	0.0411	0.0711
Water	0.1297	0.53	0	0	0	0.3403
Irrigated Agriculture Land	0.1027	0.0372	0.4981	0	0.2724	0.0896
Regolith and Waste Land	0.0978	0.3549	0.0163	0	0.0071	0.5238

(Row related to 2009 and Column related to 2020)

Table 4 .The probability of land use changes based on Markov Chain in the period of 2009-2020

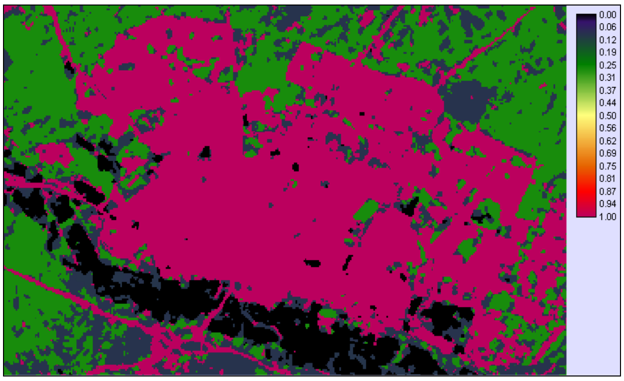


Figure 8.The probability to remain/change to built-up areas by 2020 obtained by Markov Chain

8.2 Combination of Cellular Automata and Markov Chain

To know the changes that have occurred in the past may help to predict future changes .Combination of Cellular Automata and Markov Chain is often employed to predict land cover change estimation .

In order to predict the trends of land use changes, first 1984 and 2009 land use map were analyzed with Markov Chain . Then, combined method of Cellular Automata and Markov Chain was used for forecasting land use change in 2020 . According to the results (Figure 9 and Table 5), built-up areas increase from 3377.88 hectare in 2009 to 4034.79 hectare in 2020 and the probability that the areas will change to built-up one is highest .

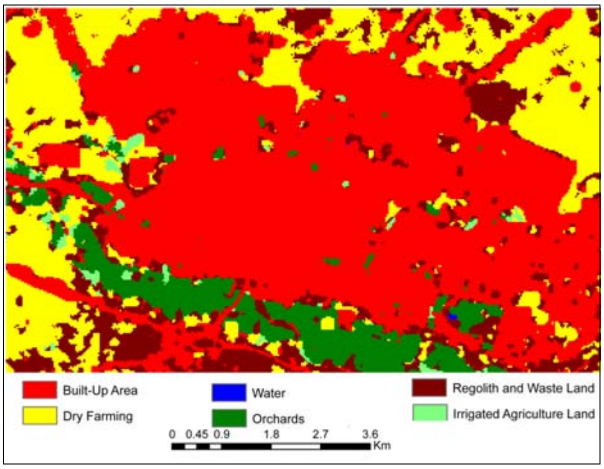


Figure 9 .Predicted result of land-use change in 2020 by the combination of Cellular Automata and Markov Chain

	Built-Up Area	Dry Farming	Orchards	Water	Irrigated Agriculture Land	Regolith and Waste Land
Built-Up Area	0.9998	0	0	0.0002	0	0
Dry Farming	0.2715	0.5708	0.0024	0	0.0094	0.1458
Orchards	0	0.1072	0.7806	0	0.0411	0.0711
Water	0.1297	0.53	0	0	0	0.3403
Irrigated Agriculture Land	0.1027	0.0372	0.4981	0	0.2724	0.0896
Regolith and Waste Land	0.0978	0.3549	0.0163	0	0.0071	0.5238

Table5 .The result of prediction of land use in 2020 by the combination of Cellular Automata and Markov Chain

9.Conclusions :

In this paper, using Landsat Satellite images in 1984 and 2009, land use changes in Zanjan city area, Iran were evaluated .For classification of the images, Fuzzy Artmap classification method was applied, which has very high confidence comparing with other classification methods .In addition, combined Cellular Automata with Markov Chain method was employed to forecast human impacts on land use change until 2020 for the study area.

The results revealed that the land use change has occurred for the area of about 3383.88 hectares in the period 1984-2009 . These changes due to developments of settlements on orchards and agriculture lands, which occurred mostly in the urban fringe of Zanjan city, are recognized as highly impacted areas from the environmental point of view.

According to Cellular Automata and Markov Chain Forecasting model, built-up areas will increase from 3377.88

hectare in 2009 to 4034.79 hectare in 2020 .The continuation of such a trend may endanger the surrounding land as well as the agricultural lands and orchards in the area .Hence, it is recommended to protect these critical areas.

The results of this study also revealed that agricultural land around major towns and settlements are recognized as critical regions in terms of land use changes, and special protection measures are needed to be taken .In case of improper planning, these regions will be changed to settlements in a very short time, which is totally in contradiction to sustainable development.

10. References

Ahadnejad, M ,2002 .Environmental land use Chang detection and assessment using with multi -temporal satellite imagery, Mapasia2002, Bangkok, Thailand.

Carpenter, G.A., Grossberg, S., and Reynolds, J.H ,1991. ARTMAP :Supervised Real-Time Learning and Classification of Non stationary Data by a Self-Organizing Neural Network .Neural Networks, 4, 565-588 .

Carpenter, G.A, 1989 .Neural Network Models for Pattern Recognition and Associative Memory .Neural Networks, 2, 243-257.

Eastman, J .Ronald, 2006 .IDRISI Andes Tutorial .Clark Labs .Clark University.

Gong, P, 1993 .Change Detection Using Principal Component Analysis and Fuzzy Set Theory, Can , J . Remote Sensing .19(1)22-29.

Jensen, J.R, 1996 .Introduction to Digital Image Processing :A Remote Sensing Perspective, Englewood Cliffs, New Jersey :Prentice-Hall.

Muchoney, D .and J .Williamson, 2001 .A Gaussian Adaptive Resonance Theory Neural Network Classification Algorithm Applied to Supervised Land Cover Mapping Using Multi Temporal vegetation Index Data, IEEE Transaction September .

Neshat, A, 2002 .Analysis and evaluation land use and land-cover changes using remote sensing data and geographic information systems in Golestan province, Msc Thesis, Tarbat Modaress University, Tehran, Iran.

Pontius, Jr, Robert Gilmore, Olufunmilayo Thontteh and Hao Chen, 2008 .Land Change Modeling with GEOMOD, Clark University .

Prakash, A .and Gupta, R .P, 1998 .Land use mapping and change detection in a coal mining area, a case study in the Jharia coalfield, India”, Int .J .Remote sensing Vol .19.

Stehman, S.V, 1996 .Estimating the Kappa Coefficient and Its Variance under Stratified Random Sampling . Photogrammetric Engineering and Remote Sensing .62, 401, 407.

Sunar, F, 1998 .An Analysis of changes in a Multi -data set; a case study in the Ikitelli area Istanbul Turkey, Int, J, Remote Sensing, Vol.19 .

Tung-Hsu Tong HOU and Ming-Der Pern, 2000 .A New Shape Classifier by Using Image Projection and a Neural Network, International Journal of Pattern Recognition and Artificial Intelligence, Vole 14, NO.2.

COMPARISON OF SPATIAL COMPACTNESS EVALUATION METHODS FOR SIMPLE GENETIC ALGORITHM BASED LAND USE PLANNING OPTIMIZATION PROBLEM

CAO Kai[a, *], HUANG Bo[a]

[a] Department of Geography and Resource Management, The Chinese University of Hong Kong Shatin N.T., Hong Kong

KEY WORDS: Compactness, Land Use Planning, Optimization, Spatial Autocorrelation

ABSTRACT:

As one of the most important objectives for land use planning towards sustainability, the compactness could not only decrease threat to species survivability and the energy consumption, but also improve the accessibility of city and the social equity towards sustainability et al. Although there have existed several methods to evaluate compactness, the spatial autocorrelation methods have not been applied in raster based land use planning optimization problem, which is one kind of spatial optimization problem and of great complexity and generally operated by heuristic methods, such as Genetic Algorithm (GA), Simulated Annealing (SA) et al. Besides, there has not been comprehensive comparison of these methods including linear, non-linear, or spatial statics methods during the optimization process. In this research, most of these methods related are reviewed, furthermore, three of these representative methods including the non-linear neighbour method, shape index and Moran's I have been compared based on simple GA on hypothesis data. The non-linear neighbour method with the simplest principle yields the best effect and efficiency. On the other hand, Moran's I method shows another angle to evaluate the compactness although the result is not very good. Furthermore, the mono Moran's I and comprehensive Moran's I also have been compared, compared to the worse result of mono Moran's I, the comprehensive Moran's I did better while it is also worse than the neighbour methods. The effect clearly shows us one possible combination of compactness and other objectives, such as compatibility, so as to improve the efficiency of the whole land use planning optimization process.

1. INTRODUCTION

1.1 Background

Compact land use is desired in various planning domains, such as forest management and reserve design et al. Promoting compactness/controlling fragmentation thus has been a common and important goal of land use planning towards sustainability, which is a hot topic nowadays.

On the contrary to compactness, urban sprawl is a widespread problem affecting much of the urban development that has occurred in the past fifty years. Environmentally, there are two main concerns related to urban sprawl: the extent it is consuming the landscape, and the air pollution that such a high level of automobile reliance is causing (Williams, 1999; Guiliano and Narayan, 2003). This also leads to the destruction of natural habitat for many species, which as a result have become endangered. Besides, sprawl is also the consumption of land resource with so much highly inefficient form, which is harmful to the provision of services and infrastructure by local governments. Furthermore, it also causes the need of more transportation facilities with much worse air pollution and energy consumption.

As for the social aspect, while the societal effects of urban sprawl are very difficult to measure accurately, there are also obvious evidences of its unsustainability. Reduced social equity, negative health impact, a loss of community, segregation, polarisation and an inability to adapt to changing lifestyles and family structures are just some of the ways in which urban sprawl is said to adversely affect social sustainability (Gillham, 2002; Kelly-Schwartz et. al., 2004). Furthermore, social equity is negatively impacted in many detailed ways below: limiting transport options of the poor due to the high costs of car ownership and poor public transport; increasing the likelihood of poor people living in less desirable neighbourhoods; increasing fear and anxiety generated by high traffic volumes; greater exposure to air pollution and resulting poor health; and losing 'a sense of community' as most people travel beyond the local neighbourhood to conduct their daily activities (Hillman, 1996).

As talked above, the negative environmental, economic and social effects of land use sprawl are widespread, diverse and clearly at odds with the concept of sustainability. The concept of compactness attempts to provide a more sustainable alternative style of land use sprawl. It should be one of the important objectives to plan a sustainable city.

Although there may be consensus that the compact land use is clearly distinct from urban sprawl, and is very essential to pursue the final objective of sustainability. There still remain many questions surrounding exactly how to evaluate the compactness of the land use during the land use planning optimization process, which not only require the effect of compactness but also the efficiency to evaluate amounts of land use planning scenarios as one kind of complicated spatial optimization problem. In this research, the methods used to evaluate the compactness will be reviewed and systematically compared during the land use planning optimization process based on simple GA, which might be very meaningful to supply the raster based land use planning optimization possibility to

[*] Corresponding author (caokaisqtv@gmail.com)

evaluate the compactness more effective and efficient. Besides, as one kind of spatial optimization, the research will also be spread to other similar applications inside spatial optimization problem.

1.2 Review of Existing Measures

Though the objective of encouraging land compactness is apparent, there exists no common accepted the best measures of spatial compactness. Herein, for different land use types, there has been several kinds of measures listed below or similar ones on "compactness" for raster based spatial problem:

1) Non-linear integer program-neighbour method;
2) Linear Integer program-neighbour method;
3) Linear Integer Program using Buffer cells;
4) Linear IP using "Aggregated Blocks"/Minimization of the number of clusters per land use types;
5) Minimization of Shape Index (Aerts 2002; Stewart 2004);
6) Spatial Autocorrelation (Wardoyo and Jordan 1996; Kurttila et al. 2002).

The first one is the most direct explanation to compactness of land use, which only takes advantage of the neighbours of each cell to evaluate the compactness by sum. The second one is equivalent linear reformulation of the first model, at the expense of including additional integer variables. The third one (Wright, Revelle, and Cohon 1983; Williams and Revelle 1998), was described as a problem where one selects parcels and each reserve (one land use type) consists of core cells, and surrounding buffer zone. Compactness is indirectly obtained through minimizing the number of buffer cells around the core areas. The fourth idea is to aggregate individual cells to blocks and develop a model that minimize the number of blocks that contain only one land use type in the final allocation result. In other words, the target is to minimizing the number of clusters according to each land use type. The fifth one is to compute the shape index of each cluster, which sounds very complex but effective to represent the compactness. The last method is from the perspective of spatial statistic, by taking advantage of Moran's I, Geary's C et al.

Most utilization of compactness as the objective are included or deviated from the six models talked above. Aerts, Stewart, and Janssen (2003; 2004; 2007) have combined the fourth model and the sixth model with the definition of maximization of the large cluster of each land use type to pursue the target of compactness of each scenario and the effect is good within the small research area. Aerts and Erwin have compared the anterior four models in 2002, according to the result on the testing area (8*8 grid). From the comparison of the four models by them, we can clearly know that the efficiency and the effect of the first measure is the best. It is the fastest way to get the global optimum while it is not a linear method. Although the spatial autocorrelation methods have been used to evaluate the characteristic of compactness, there has not been the comparison of the spatial autocorrelation with the other methods on the effect and efficiency to pursue the compactness of land use planning during the optimization process.

In this research, one new application of Moran's I index to evaluate the compactness will be tested to compare with the methods non-linear neighbour method with the best efficiency and the shape index method with the best explanation of compactness.

2. METHODS

2.1 Non-linear Neighbour Method

The first measure can be described in terms of recording for each cell, the number of neighbouring cells which have the same land use. In this sense, the "neighbouring" cells to (i, j) are the (i-1, j), (i+1, j), (i, j-1), (i, j+1), (i-1, j-1), (i+1, j+1), (i-1, j+1), (i+1, j-1) (ignoring cells outside the region). At this definition, it can be shown as follows:

Minimize:

$$-\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M} \mathrm{A}_{ijk} x_{ijk} \tag{1}$$

Where

$$\mathrm{A}_{ijk} = \mathrm{x}_{\mathrm{i-1jk}} + \mathrm{x}_{\mathrm{i+1jk}} + \mathrm{x}_{\mathrm{ij\text{-}1k}} + \mathrm{x}_{\mathrm{ij+1k}} + \mathrm{x}_{\mathrm{i-1j\text{-}1k}} + \mathrm{x}_{\mathrm{i+1j+1k}} + \mathrm{x}_{\mathrm{i\text{-}1j+1k}} + \mathrm{x}_{\mathrm{i+1j\text{-}1k}}$$
$$\forall \mathrm{k} = 1,...,\mathrm{K}, \mathrm{i} = 1,...,\mathrm{N}, \mathrm{j} = 1,...,\mathrm{M} \tag{2}$$

Herein we understood a neighbourhood of eight cells (top, down, left, right, left-top, right-down, right-top, left-down), but alternatively there are some other smaller or larger neighbourhood can be defined, such as four neighbours etc. It can be clearly seen that minimization process can give birth to solutions in which neighbouring cells have the same land use type.

2.2 Shape Index Method

The shape index method can be calculated using the following equations:

$$Shape_{total} = \sum_{\mathrm{k}=1}^{\mathrm{K}}\sum_{c=1}^{\mathrm{C}} \frac{\mathrm{P}_{kc}}{\sqrt{R_{kc}}} \tag{3}$$

Where $\mathrm{P_{kc}}$ stands for the perimeter of one cluster c for land use type k. And $\mathrm{R_{kc}}$ represents the area of each cluster c for land use k. The values for the perimeters of cluster A, B, C and D are 20, 20, 18, and 14. The values for cluster A, B, C and D are 16, 13, 12, and 7, so the value for this shape index is 21.04.

For some special situation, such as the single cell as a cluster in an optimal result, through the minimization of the Shape total, the function can prove both the shape of each cluster and the number of the clusters. Of course, the complexity is also obvious.

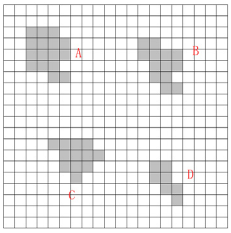


Figure 1 Shape Index of the Four Clusters

2.3 Moran's I Method

Moran's I is a measure of spatial autocorrelation developed by Patrick A.P. Moran. Like autocorrelation, spatial autocorrelation means that adjacent attribute for the same phenomenon that is correlated. Spatial autocorrelation is about proximity in (two-dimensional) space. Spatial autocorrelation is more complex than autocorrelation because the correlation is two-dimensional and bi-directional.

Moran's I is defined as

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ji}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{4}$$

Where n is the number of spatial units indexed by i and j. x is the variable of interest; $\bar{x}$ is the mean of x; and W_{ij} and W_{ji} are matrices of spatial weights.

The expected value of Moran's I is

$$\mathrm{E(I)} = \frac{-1}{\mathrm{N}-1} \tag{5}$$

Its variance equals

$$Var(I) = \frac{\{n[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0]\} - \{k[(n^2 - n)S_1 - 2nS_2 + 6S_0{}^2]\ \}}{(n-1)(n-2)(n-3)S_0{}^2} - E(I)^2 \tag{6}$$

Where

$$S_0 = \sum_{i=1}^{n}\sum_{j}^{n} Wij \tag{7}$$

$$S_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} (Wij + Wji)^2 \Big/ 2 \tag{8}$$

$$S_3 = \sum_{i=1}^{n} (Wi. + W.i)^2 \tag{9}$$

($W_{i.}$ and $W_{.i}$ mean i row and i column of the related matrix)

$$k = \frac{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 \Big/ n\right]}{\left[\sum_{j=1}^{n}\left(x_i - \bar{x}\right)^2 \Big/ n\right]^2} \tag{10}$$

According to the steps talked above, the Moran's I will be between -1 and 1, if the index is greater than 1, it means that the correlation is positive; if less than 0, it means negative, and the more, the larger of the correlation, and vice versa. And if the value is near to 0, it represents random distribution.

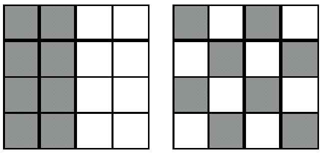


Figure 2 The Representations of Positive and Negative Correlation for value of 1 and -1

3. COMPARISON ON HYPOTHSIS DATA

In the previous section, three good methods encouraging compactness are presented. In this section we will evaluate and compare these models on their effect and the efficiency for use. The comparison will be operated on 20 by 20 grid with 5 assumed land use types and PC with an Intel Xeon CPU @2.33GHZ processor, 3GB RAM base on simple GA. The performance of the methods used can be evaluated by two criteria. Criterion 1 evaluates the computation time against the achieved degree of compactness. Thereafter, the characteristics of the compactness created by these methods should also be considered as the criterion 2.

3.1 Comparison of Efficiency

Table 1 Comparison of the CPU time based on Three Models (the unit is second, and Model-1, Model-2, and Model-3 are separately related to non-linear neighbour method, shape index method and Moran's I method)

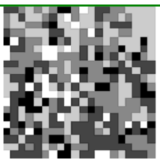
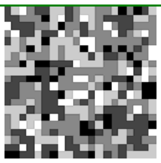
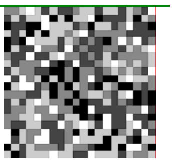
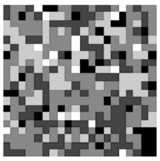
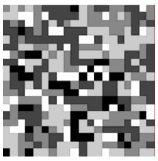
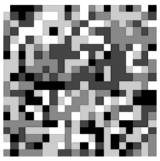
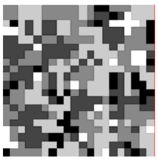
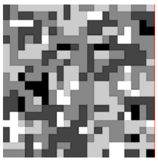
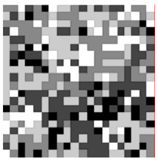
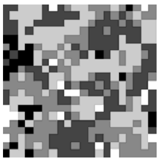
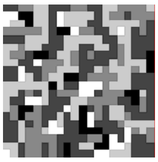
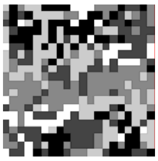
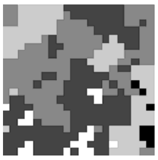
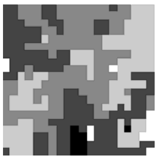
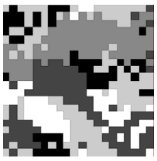
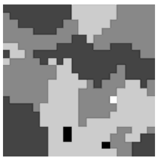
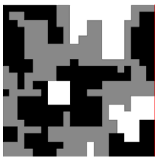
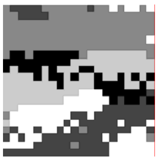
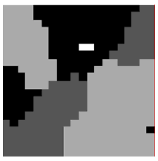
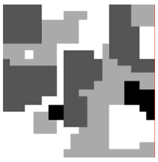
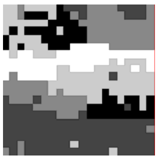
Iteration Number	Model-1	Model-2	Model-3
100	8.39	33.34	279.17
200	10.32	45.62	387.67
500	19.46	75.64	755.04
1000	35.17	127.16	1402.36
5000	156.83	442.9	5877.64
10000	316.13	780.47	12930.07
50000	1526.12	3367.32	56595.42

From the Table 1 above, it is evident that the Model-1 (non-linear neighbour method) is the fastest model to achieve the compactness than the other two models: shape index and Moran's I. On the contrary, the Moran's I method is the worst with the much more time spending than the neighbour model.

3.2 Comparison of Effect

From the comparison of the effect in the Table 2, the effect of the each model is good enough to operate the optimization process. The effect of the non-linear neighbour model is smoother than the other two models, which is more suitable to reflect the reality of the real change of the land use. For the other two models, shape index and Moran's I, the effect are also good enough to satisfy the need of optimization, however, the time spending are too much for the process of optimization if the research area is larger.

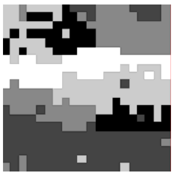
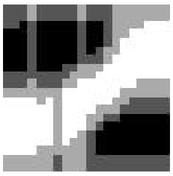
Table 2 Comparison of the Effect based on the Three Models

Iteration	Model-1	Model-2	Model-3
100			
200			
500			
1000			
5000			
10000			
50000			

3.3 Comparison of Mono-Moran's I and Comprehensive Moran's I

While from the Table 3 below, when considering the comprehensive Moran's I, which is considering the five land use types at one time, not only the compactness but also the correlation among different land use types can be reflected from the effect, the result is evidently better than the Mono Moran's I using less than 1/4 time. It means that, we can take use of this model to achieve two or more objectives (such as the compatibility et al) at one time to promote the efficiency and effect, of course, the weighting of the compactness and the compatibility might be the problem.

Table 3 Comparison of efficiency and effect of Mono and Comprehensive Moran's I Index (Mono Moran's I means only considering one land use on one time; Comprehensive Moran's I means considering five land use types together, and the unit is second)

	Mono Moran's I	Comprehensive Moran's I
Time Spending	56595.42	10791.35
Result		

4. CONCLUSION

Compactness is one of the most important objectives for land use planning, which could decrease threat to species survivability, decrease the energy consumption, improve the accessibility of city and the social equity towards sustainability et al. However, during the land use planning optimization process, which is generally very complicated and solved by heuristic methods, it lacks of efficient quantified evaluation methods to pursue the compactness of the land use as one important objective. Although there have existed several methods to evaluate compactness, there have not been application of autocorrelation methods on spatial optimization problem, which not only require the effect but also the efficiency, besides, the comprehensive comparison of these methods is also meaningful to the development of quantification evaluation of compactness on land use planning optimization. In this research, all the methods to quantified evaluation of compactness are reviewed, furthermore, three of these representative methods including the neighbour method, shape index and Moran's I have been compared using the same simple GA environment on hypothesis data. Obviously that the neighbours method with the simplest principle yields the best effect and efficiency; Moran's I method brings its potential to evaluate the compactness although the result is not very good. Besides, the mono Moran's I and comprehensive Moran's I also have been compared, compared to the worse result of mono Moran's I, the comprehensive Moran's I does better while it is also worse than the neighbour methods. The result also suggests one possible combination of compactness and other objectives, such as compatibility, which might improve the effect and efficiency of the whole optimization process.

References

Aerts, J. C. J. H., Eisinger, E., Heuvelink, G. B. M., & Stewart, T. J. (2003). Using Linear Integer Programming for Multi-Site Land-Use Allocation. Geographical Analysis, 35(2), 148-169.

Aerts, J. C. J. H., Herwijnen, M. v., Janssen, R., & Stewart, T. J. (2005). Evaluating Spatial Design Techniques for Solving Land-use Allocation Problems. Journal of Environmental Planning and Management, 48(1), 121-142.

Aerts, J. C. J. H., Herwijnen, M. v., & Stewart, T. J. (2003). Using Simulated annealing and spatial goal programming for solving a multi site land use allocation problem. Lecture notes in computer science 2632, 448-463.

Aerts, J. C. J. H., & Heuvelink, G. B. M. (2002). Using simulated annealing for resource allocation. International Journal of Geographical Information Science, 16(6), 571-587.

Arthur, J. L., & Nalle, D. J. (1997). Clarification on the Use of Linear Programming and GIS for Land-Use Modelling. *International Journal of Geographical Information Science 11*(4), 397- 402.

Barber, G. M. (1976). Land Use Plan Design via Interactive Multo-Objective Programming. *Environmental and Planning A, 8*, 625-636.

Burton, E. (2000). The Potential of the Compact City for Promoting Social Equity. In *Achieving Sustainable Urban Form* (pp. 19-29). London: Taylor & Francis.

Chuvieco, E. (1993). Integration of Linear Programming and GIS for Land-use Modelling. *International Journal of Geographical Information Science, 7*(1), 71-83.

Cova, T. J., & Church, R. L. (2003). Contiguity Constraints for Single-Region Site Aearch Problems. *Geographical Analysis, 32*(4), 306-329.

Cromley, R. G., & Hanink, D. M. (2003). Scale-Independent Land-Use Allocation Modeling in Raster GIS. *Cartography and Geographic Information Science, 30*, 343-350.

Gilbert, K. C., D.Holmes, D., & Rosenthal, D. E. (1982). A Multiobjective Discrete Optimization Model For Land Allocation. *Management Science, 31*(12), 1509-1522.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning* Boston: Addison-Wesley Longman Publishing.

Guiliano, G., & Narayan, D. (2003). Another look at travel patterns and urban form: the US and Great Britain. Urban Studies, 40(11), 2295-2312.

Hillman, M. (1996). In Favour of the Compact City. In M. Jenks, E. Burton & K. Williams (Eds.), The Compact City: a sustainable urban form? (pp. 36-44). London: E&FN Spoon.

Kelly-Schewartz, A., Stockard, J., Doyle, S., & Schlossberg, M. (2004). Is sprawl unhealthy? A multi-level analysis of the relationship of metropolitan sprawl to the health of the individual. Journal of Planning Education and Research, 24, 184-196.

Kurttila, M., Pukkala, T., & Loikkanen, J. (2002). The performance of alternative spatial objective types in forest planning calculations: a case for flying squirrel and moose. Forest Ecology and Management, 166, 245-260.

Leung, Y., Li, G., & Xu, Z.-B. (1998). A Genetic Algorithm for the Multiple Destination Routing Problems. *IEEE Transactions on Evolutionary Computation, 2*(4), 150-161.

Malczewski, J. (1999). *GIS and Multicriteria Decision Analysis*. New York: Wiley.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Berlin: Springer.

Nalle, D. J., Arthur, J. L., & Sessions, J. (2002). Designing Compact and Contiguous Reserve Networks with a Hybrid Heuristic Algorithm. *Forest Science, 48*(1), 59-68.

Stewart, T. J., Janssen, R., & Herwijnen, M. v. (2004). A genetic algorithm approach to multiobjective land use planning. Comput. Oper. Res., 31(14), 2293-2313.

Shirabe, T. (2005). A Model of Contiguity for Spatial Unit Allocation. *Geographical Analysis, 37*, 2-16.

Wardoyo, W., & Jordan, G. A. (1996). Measuring and assessing management of forested landscapes. Forestry Chronicle, 72, 639-645.

Williams, J. C., & Revelle, C. S. (1998). Reserve assemblage of critical areas: A zero-one programming approach. European Journal of Operational Research, 104(3), 497-509.

Williams, K., Burton, E., & Jenks, M. (1996). Achieving the Compact City through Intensification: an acceptable option? In M. Jenks, E. Burton & W. K (Eds.), The Compact City: a sustainable urban form? (pp. 83-96). London: E&FN Spoon.

Williams, K. (1999). Urban intensification policies in England: problems and contradictions. Land Use Policy, 16(3), 167-178.Wright, J., Revelle, C., & Cohon, J. (1983). A multiobjective integer programming model for the land acquisition problem. *Regional Science and Urban Economics, 13*(1), 31-53.

Wright, J., Revelle, C., & Cohon, J. (1983). A multiobjective integer programming model for the land acquisition problem. Regional Science and Urban Economics, 13(1), 31-53.

AUTOMATIC DERIVATION OF LAND-USE FROM TOPOGRAPHIC DATA

Frank Thiemann [a], Monika Sester [a], Joachim Bobrich [b]
[a] Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Appelstraße 9a,
30167 Hannover, Germany {frank.thiemann, monika.sester}@ikg.uni-hannover.de
[b] Federal Agency for Cartography and Geodesy, Richard-Strauss-Allee 11,
60598 Frankfurt am Main, Germany, joachim.bobrich@bkg.bund.de

KEY WORDS: Generalization, Aggregation, Processing, Large Vector Data, CORINE Land Cover

ABSTRACT:

The paper presents an approach for the reclassification and generalization of land-use information from topographic information. Based on a given transformation matrix describing the transition from topographic data to land-use data, a semantic and geometry based generalization of too small features for the target scale is performed. The challenges of the problem are as follows: (1) identification and reclassification of heterogeneous feature classes by local interpretation, (2) presence of concave, narrow or very elongated features, (3) processing of very large data sets. The approach is composed of several steps consisting of aggregation, feature partitioning, identification of mixed feature classes and simplification of feature outlines.

The workflow will be presented with examples for generating CORINE Land Cover (CLC) features from German Authoritative Topographic Cartographic Information System (ATKIS) data for the whole are of Germany. The results will be discussed in detail, including runtimes as well as dependency of the result on the parameter setting.

1. INTRODUCTION

1.1 Project Background

The European Environment Agency collects the Coordinated Information on the European Environment (CORINE) Land Cover (CLC) data set to monitor the land-use changes in the European Union. The member nations have to deliver this data every few years. Traditionally this data set was derived from remote sensing data. However, the classification of land-use from satellite images in shorter time intervals becomes more cost intensive.

Therefore in Germany the federal mapping agency (BKG) investigates an approach of deriving the land cover data from topographic information. The BKG collects the digital topographic landscape models (ATKIS Base DLM) from all federal states. The topographic base data contains up-to-date land-use information. But there are some differences between ATKIS and CLC.

1.2 CORINE Land Cover (CLC)

CORINE Land Cover is a polygon data set in the form of a tessellation: polygons do not overlap and cover the whole area without gaps. The scale is 1:100000. Each polygon has a minimum area of 25 hectare. There are no adjacent polygons with the same land-use class as these have to be merged.
Land cover is classified hierarchically into 46 classes in three levels, for which a three digit numerical code is used. The first and second level groups are:

1. artificial (urban, industrial, mine)
2. agricultural (arable, permanent, pasture, heterogeneous)
3. forest and semi-natural (forest, shrub, open)
4. wetland (inland, coastal)
5. water (inland, marine)

CLC has a detailed thematic granularity concerning vegetation objects. In the agricultural group, there are also some aggregated classes for heterogeneous agricultural land-use. Such areas are composed of small areas of different agricultural land-use, e.g. class 242 which is composed of alternating agricultural uses (classes 2xx).

1.3 ATKIS Base DLM

The *Base Digital Landscape Model* (DLM) of the *Authoritative Topographic Cartographic Information System* (ATKIS) is Germany's large scale topographic landscape model. It contains polygon and also poly-line and point data. The scale is approx. 1:10000. The minimum area for polygons is one hectare. The data set is organized in thematic layers, which can also overlap. The land cover information is spread among these different layers.

Each object has a four digit class code[1] and different attributes consisting of a three character key and a four digit alphanumeric attribute. The classes are also organized hierarchically in three levels. The seven first levels groups are:

1. presentation
2. residential
3. traffic (street, railway, airway, waterway)
4. vegetation
5. water
6. relief
7. areas (administrative, geographic, protective, danger)

Table 1 gives a summarized comparison of the two data sets.

Data set	ATKIS Base DLM	CORINE LC
scale	1:10000	1:100000
source	aerial images, cadastre	satellite images
min. area size	1 ha	25 ha
topology	overlaps, gaps e.g. between the divided carriageways	tessellation
feature classes	90 relevant (155 with attributes)	44 (37 in Germany)
agricultural feature classes	5 relevant (9 with attributes)	11 (6 in Germany) 4 (2) heterogeneous classes

Table 1: Comparison of ATKIS and CLC

[1] In the new AAA model, which is currently being introduced, there is a 5-digit object code.

1.4 Automatic derivation of CLC from DLM

The aim of the project is the automated derivation of CLC data from ATKIS. This derivation can be considered as generalization process, as there it requires both thematic selection and reclassification, and geometric operations due to the reduction in scale. Therefore, the whole workflow consists of two main parts. The first part is a model transformation and consists of the extraction, reclassification and topological correction of the data. The second part, the generalization part, which will be described in more detail in this paper, is the aggregation and simplification for the smaller scale.

The first part consists of the following steps: after the extraction of the relevant features from the DLM the topological problems like overlaps and gaps area solved automatically using appropriate algorithms. The reclassification is done using a translation table which takes the ATKIS classes and their attributes into account. In the cases where a unique translation is not possible, a semi-automatic classification from remote sensing data is used. The derived model is called DLM-DE LC.

In the second part the high level information from the DLM-DE is generalized to the small scale of 1:100000 of the CLC. For that purpose a sequence of generalization operations is used. The operators are dissolve, aggregate, split, simplify and a mixed-class filter.

1.5 Main Challenges

One of the main challenges of the project is the huge amount of data. The DLM-DE contains ten million polygons. Each polygon consists in average of thirty points, so one has to deal with 300 million points, which is more than a standard PC can store in the main memory. Therefore a partitioning concept is needed that allows processing the data sequentially or in parallel. Fast algorithms and efficient data structures reduce the required time.

Another challenge is the aggregation of agricultural heterogeneous used areas to a group of 24x-classes in the case that a special mixture of land-uses occurs. The difficulty is to separate these areas from homogeneous as well as from other heterogeneous classes.

2. RELATED WORK

CORINE Land Cover (Büttner et al. 2006) is being derived by the European States (Geoff et al. 2007). The Federal Agency of Cartography and Geodesy attempts to link the topographic data base with the land-use data. To this end, transformation rules between CLC and ATKIS have been established (Arnold 2009). As described above, the approach uses different generalization and interpretation steps. The current state of the art in generalization is described in Mackaness et al. (2007). The major generalization step needed for the generalization of land-use classes is aggregation. The classical approach for area aggregation was given by Oosterom (1995), the so-called GAP-tree (Generalized Area Partitioning). In a region-growing fashion areas that are too small are merged with neighboring areas until they satisfy the size constraint. The selection of the neighbor to merge with depends on different criteria, mainly geometric and semantic constraints, e.g. similarity of object classes or length of common boundary. This approach is implemented in different software solutions (e.g. Podrenek, 2002). Although the method yields areas of required minimum size, there are some drawbacks: a local determination of the most compatible object class can lead to a high amount of class changes in the whole data set. Also, objects can only survive the generalization process, if they have compatible neighbors. The method by Haunert (2008) is able to overcome these drawbacks. He is also able to introduce additional constraints e.g. that the form of the resulting objects should be compact. The solution of the problem has been achieved using an exact approach based on mixed-integer programming (Gomory, 1958), as well as a heuristic approach using simulated annealing (Kirkpatrick 1983). However, the computational effort for this global optimization approach is very high.

Collapse of polygon features corresponds to the skeleton operation, which can be realized using different ways. A simple method is based on triangulation; another is medial axis or straight skeleton (Haunert & Sester, 2008).

The identification of mixed classes is an interpretation problem. Whereas interpretation is predominant in image understanding where the task is to extract meaningful objects from a collection of pixels (Lillesand & Kiefer, 1999), also in GIS-data interpretation is needed, even when the geo-data are already interpreted. E.g. in our case although the polygons are semantically annotated with land-use classes, however, we are looking for a higher level structure in the data which evolves from a spatial arrangement of polygons. Interpretation can be achieved using pattern recognition and model based approaches (Heinzle & Anders, 2007).

3. APPROACH

3.1 Data and index structures

Efficient algorithms demand for efficient data and search structures. For topology depending operations a topologic data structure is essential. For spatial searching a spatial index structure is needed; furthermore, also structures for one-dimensional indexing are used.

In the project the we use a extended Doubly Connected Edge List (DCEL) as topologic structure and grids (two-dimensional hashing) as spatial index.

3.1.1 Extended DCEL

The doubly connected edge list (DCEL) is a data structure for polygonal meshes. It is a kind of boundary representation. The topological elements (and their geometric correspondence) are faces (polygon), edges (lines) and nodes (points). All topologic relations (adjacencies and incidences) are expressed by explicit links (see Figure 1).. For efficient iteration over all nodes or edges of a face or all incident edges of a node the edges are split into a pair of two directed half-edges. Each half-edge links its origin (starting point), its twin, the previous and next half-edge and the incident face. The node contains the geometric information and a link to one of the incident half-edges. The face contains a link to a half-edge from the outer loop and if the polygon has holes also, a half-edge from each inner loop respectively.

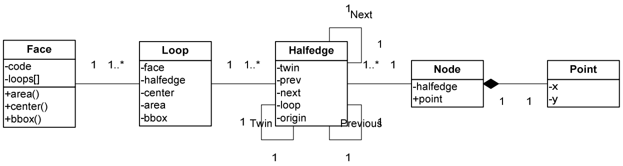


Figure 1: UML Diagram of the extended DCEL

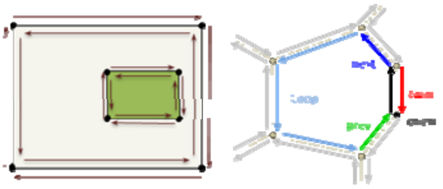


Figure 2: Left: A Face with an inner face in DCEL Right: Topological relations of a half-edge

For reasons of object orientated modeling loops were placed between the faces and half-edges, as one can often find in 3D data-structures (e.g. ACIS). The loop represents a closed ring of half-edges. This ring can be an inner or outer border of a face (see Figure 2). Algorithms for calculation the area (the area of inner loops is negative) or the centroid are implemented as member function of the loop. Because of the linear time complexity the values will be stored for each loop. For efficient spatial operations also the bounding box of the loop is stored. The land-use code is attached to faces.

3.1.2 GRID

As spatial index for nodes, edges and faces we use a simple two dimensional hashing. We put a regular grid over the whole area. Each cell of this grid contains a list of all included points and all intersecting edges and faces, respectively. This simple structure can be used, because of the approximately equally distributed geometric features.

For the DLM-DE a grid width of 100 m for points and edges (<10 features per cell) and 1000 m for faces (40 faces per cell) leads to nearly optimal speed. Experiments with a KD-tree for the points lead to similar results.

3.2 Topological cleaning

Before starting the generalization process, the data have to be imported into the topological structure. In this step we also look for topological or semantic errors. Each polygon is check for a valid CLC class. Small sliver polygons with a size under a threshold of e.g. 1 m will be rejected. A snapping with a distance of 1 cm is done for each inserted point. With a point in polygon test and a test for segment intersection overlapping polygons are detected and also rejected. Holes in the tessellation can be easily found by building loops of the half-edges which not belong to any face. Loops with a positive orientation are holes in the data set. The largest loop with a negative orientation is the outer border of the loaded data.

3.3 Generalization operators

3.3.1 Dissolve

The dissolve operator merges adjacent faces of the same class. For this purpose the edges which separate such faces will be removed and new loops are built. Besides the obvious cases which reduce the number of loops, there are also cases which generate new inner loops (see Figure 3).

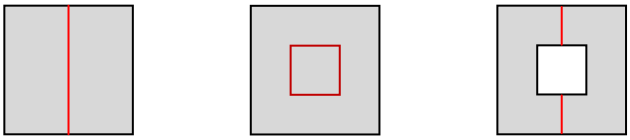


Figure 3: Beside the obvious cases (left and middle) of a merge, where the number of loops is reduced there are also cases which produce new inner loops (right).

3.3.2 Aggregate

The aggregation step aims at guarantying the minimum size of all faces. The aggregation operator in our case uses a simple greedy algorithm. It starts with the smallest face and merges it to a compatible neighbor. This fast algorithm is able to process the data set sequentially. However, in some cases it may lead to unexpected results, as shown in Figure 4. This is due to the fact that the decisions are only taken locally and not globally.

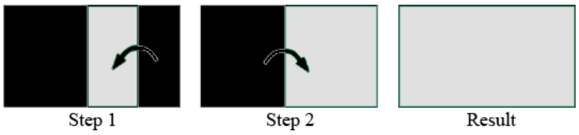
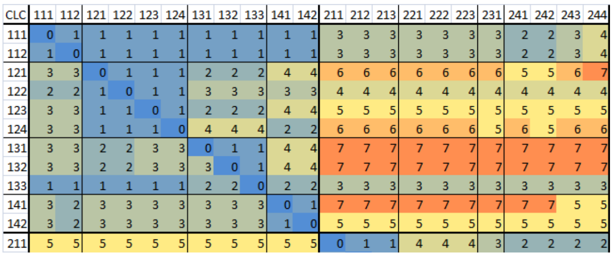


Figure 4: The sequential aggregation can lead to an unexpected result: The black area is dominating the source data set, but after aggregation the result is grey (according to Haunert (2008))

There are different options to determine compatible neighbors. The criterion can be:

- the semantic compatibility (semantic distance),
- the geometric compactness
- or a combination of both.



CLC	111	112	121	122	123	124	131	132	133	141	142	211	212	213	221	222	223	231	241	242	243	244
111	0	1	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2	3	4
112	1	0	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2	3	4
121	3	3	0	1	1	1	2	2	2	4	4	6	6	6	6	6	6	6	5	5	6	7
122	2	2	1	0	1	1	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4
123	3	3	1	1	0	1	2	2	2	4	4	5	5	5	5	5	5	5	5	5	5	5
124	3	3	1	1	1	0	4	4	4	2	2	6	6	6	6	6	6	5	6	5	6	6
131	3	3	2	2	3	3	0	1	1	4	4	7	7	7	7	7	7	7	7	7	7	7
132	3	3	2	2	3	3	3	0	1	4	4	7	7	7	7	7	7	7	7	7	7	7
133	1	1	1	1	1	1	2	2	0	2	2	3	3	3	3	3	3	3	3	3	3	3
141	3	2	3	3	3	3	3	3	3	0	1	7	7	7	7	7	7	7	7	7	5	5
142	3	2	3	3	3	3	3	3	3	1	0	5	5	5	5	5	5	5	5	5	5	5
211	5	5	5	5	5	5	5	5	5	5	5	0	1	1	4	4	4	3	2	2	2	2

Figure 5: Small extract of the CLC priority matrix

The **semantically** nearest partner can be found using a priority matrix. We use the matrix from the CLC technical guide (Bossard, Feranec & Otahel 2000) (Figure 5). The priority values are from an ordinal scale, so their differences and their values in different lines should not be compared. The matrix is not symmetric, as there may be different ranks when going from one object to another than vice versa (e.g. settlement -> vegetation). Priority value zero is used if both faces have the same class. The higher the priority value, the higher is the semantic distance. Therefore the neighbor with the lowest priority value is chosen.

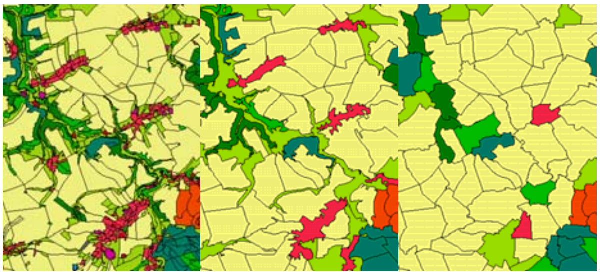


Figure 6: (left to right) Original situation, the result of the semantic and geometric aggregation.

As **geometric** criterion the length of the common edge is used. This leads to compact forms. Compactness can be measured as the ratio of area and perimeter. A shorter perimeter leads to better compactness. So the maximum edge length has to be reduced to achieve a better compactness.

The effects of using the criteria separate are shown in a real example in Figure 6. The semantic criterion leads to non-compact forms, whereas the geometric criterion is more compact but leads to a large amount of class change.

The combination of both criteria allows merging of semantically more distant objects, if the resulting form is more compact. This leads to Formula 1.

$$distance(A,B) = \frac{b^{priority}}{length} \qquad (1)$$

The formula means that a b-times longer shared edge allows a neighbor with the next worse priority. The base b allows to weight between compactness and semantic proximity. A value of b=1 leads to only compact results, a high value of b leads to semantically optimal results. Using the priority values is not quite correct; it is only a simple approximation for the semantic distance.

Another application of the aggregation operation is a special kind of dissolve that stops at defined area size. It merges small faces of the same class to bigger compact faces using the geometric aggregation with the condition that only adjacent faces of the same class are considered.

3.3.3 Split

In addition to the criterion of minimal area size also the extent of the polygon is limited to a minimum distance. That demands for a collapse operator to remove slim, elongated polygons and narrow parts. The collapse algorithm by Haunert & Sester (2008) requires buffer and skeleton operations that are time consuming. Therefore - as faster alternative - a combination of splitting such polygons and merging the resulting parts with a geometric aggregation to other neighbors is used.

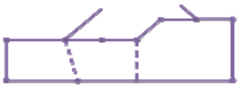


Figure 7: The operator splits the polygon at narrow parts if there is a higher order node or a concave node. An existing node is preferred if it is close to the orthogonal projection.

The split operator cuts faces at narrow internal parts. First, the concave or higher order node with the smallest distance to a non-adjacent edge is calculated. A new node will be inserted at the orthogonal projection if there is no existing node nearby. An edge is inserted if it fulfills the conditions being inside and not intersecting other edges. Else the next suited node is chosen (see Figure 7). After the split operator the aggregate operator merges too small pieces to other adjacent faces.

3.3.4 24x-Filter

In CORINE land-cover there is a group of classes which stands for heterogeneous land-uses. The classes 242 and 243 are relevant for Germany. Class 242 (complex cultivation pattern) is used for a mixture of small parcels with different cultures. Class 243 is used for land that is principally occupied by agriculture, with significant areas of natural vegetation.

Heterogeneous classes are not included in the ATKIS schema. To form these 24x-classes an operator for detecting heterogeneous land-use is needed. The properties of these classes are that smaller areas with different, mostly agricultural land-use alternate within the minimum area size (actually 25 ha in CLC). For the recognition of class 242 only the agriculture areas (2xx) are relevant. For 243 also forest, semi- and natural areas (3xx, 4xx) and lakes (512) have to be taken into account.
The algorithm calculates some neighborhood statistics for each face. All adjacent faces within a distance of the centroid smaller than a given radius and with an area size smaller than the target size are collected by a deep search in the topological structure. The fraction of the area of the majority class and the summarized fractions of agricultural areas (2xx) and (semi-)natural areas (3xx, 4xx, 512) are calculated. In the case the majority class dominates (>75%) then the majority class becomes the new class of the polygon. Otherwise there is a check, if it is a heterogeneous area or only a border region of larger homogeneous areas.

For that purpose the length of the borders between the relevant classes is summarized and weighted with the considered area. A heterogeneous area is characterized by a high border length, as there is a high number of alternating areas. To distinguish between 242 and 243 the percentage of (semi-natural) areas has to be significant (>25%).

3.3.5 Simplify

The simplify-operator removes redundant points from the loops. A point is redundant, if the geometric without using this point is lower than an epsilon and if the topology do not change.

We implemented the algorithm of Douglas & Peucker (1973) with an extension for closed loops and a topology check. The algorithm is running over all loops, between each pair of adjacent topological nodes (degree > 2). If the loop contains no topological nodes, the first one is chosen. The algorithm tries like Douglas-Peuker to use the direct line between the two end nodes and searches for the farthest point of the original line to this new line. The first extension is for the case, that both end points are the same nodes. Then the point to point distance is used instead of point to straight line distance. If the distance of the farthest point is larger than the epsilon-value then the point is inserted in the new line and the algorithm processes both parts recursively. If the distance is smaller than epsilon the Douglas-Peucker algorithm would remove all points between the end nodes. Here the second extension is done to checks the topology. All points in the bounding-box spanned by the two nodes are checked for switching the side of the line. If a point switches the side, the farthest point is inserted to the line (i.e. treating it as if it were too far).

3.4 Process chain

In this section the use of the introduced operators and their orchestration in the process change is shown. The workflow for a target size of 25 ha is as follows:

1. import and data cleaning
2. fill holes
3. dissolve faces < 25 ha
4. split faces < 50 m
5. aggregate faces < 1 ha geometrically (base 1.2)
6. reclassify faces with 24x-filter (radius 282 m)

7. aggregate faces < 5 ha weighted (base 2)
8. aggregate faces < 25 ha semantically
9. simplify polygons (tolerance 20 m)
10. dissolve all

During the import step (1) semantic and topology is checked. Small topologic errors are resolved by a snapping. The hole-fill step (2), searches for all outer loops and fills gaps with dummy objects. These objects will be merged to other objects in the later steps.

A first dissolve step (3) merges all faces with an adjacent face of the same CLC class which are smaller than the target size (25 ha). The dissolve is limited to 25 ha to prevent polygons from being too large (e.g. rivers that may extend over the whole data set). This step leads to many very non-compact polygons. To be able to remove them later, the following split-step (4) cuts them at narrow internal parts (smaller than 50 m = 0.5 mm in the map). Afterwards an aggregation (5) merges all faces smaller than the source area size of 1 ha to geometrically fitting neighbors.

The proximity analysis of the 24x-filter step (6) reclassifies agricultural or natural polygons smaller than 25 ha in a given surrounding as heterogeneous (24x class).
The next step aggregates all polygons to the target size of 25 ha. First we start with a geometric/semantically weighted aggregation (7) to get more compact forms, second only the semantic criterion is used (8) to prevent large semantic changes of large areas.

The simplify step (9) smoothes the polygon outlines by reducing the number of nodes. As geometric error tolerance 20 m (0.2 mm in the map) is used. The finishing dissolve step (10) removes all remaining edges between faces of same class.

4. RESULTS

4.1 Runtime and memory

The implemented algorithms are fast but require a lot of memory. Data and index structures need up to 160 Bytes per point on a 32 bit machine. With 6 GB free main memory on a 64 bit computer we were able to process up to 30 million points at once, which corresponds to the tenth part of Germany.

The run-time was tested with a 32 bit 2.66 GHz Intel Core 2 processor with a balanced system of RAM, hard disk and processor (windows performance index 5.5). The whole generalization sequence for a 45 x 45 km data set takes less than two minutes. The most time expensive parts of the process are the I/O-operations which take more than 75% of the computing time. We are able to read 100000 points per second from shape files while building the topology. The time of the writing process depends on the disk cache. In the worst case it is the same as for reading.

The time of the operations highly depends on the data. The most expensive one is the split operation that is quadratic with the number of points per polygon. At the introduced position in the process chain the split operation takes the same time as the reading process.

The other operations are ten and more times faster than I/O operations. The aggregation operator processes one million points per second. The line generalization with 0.7 million points per second is a bit slower, but it works on the reduced data set at the end of the generalization process.

4.2 Semantic and geometric correctness

To evaluate the semantic and geometric correctness we did some statistics comparing input, result and a CLC 2006 reference data set, which was derived from remote sensing data.

Data set	DLM-DE	Result	CLC 2006
Polygons	91324	1341	878
Points $^{\text{per Polygon}}$	24	104	77
Area $^{\text{per Polygon}}$	2.3 ha	155 ha	238 ha
Perimeter $^{\text{per Polygon}}$	0.6 km	9.4 km	10.1 km
Avg. Compactness	50%	24%	33%

Table 2: Statistic of the test data set Dresden (45 x 45 km)

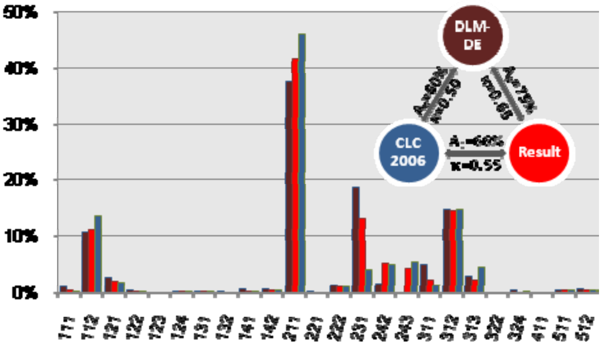


Figure 8: Percentage of area for each CLC class (bars) and percentage of match (A_0) and κ-values for the Dresden data set.

Figure 9 shows the input data (DLM-DE), our result and the CLC 2006 of the test area Dresden. The statistics in Figure 8 verifies that our result matches with DLM-DE (75%) better than the reference data set (60%). This is not surprising as for CLC 2006 different data sources were used. Because of the removing of the small faces our generalization result is a bit more similar to CLC 2006 (66%) than CLC 2006 to our input.

Table 2 shows, that our polygons are only a bit smaller and more complex and less compact than the CLC 2006 polygons. The percentage of the CLC classes is similar in all data sets (Figure 9). There are some significant differences between the DLM-DE and CLC 2006 within the classes 211/234 (arable/grass land) and also between 311/313 (broad-leaved/mixed forest) and 111/112 (continuous/discontinuous urban fabric). We assume that it comes from different interpretations. The percentages in our generated data set are mostly in the middle. The heterogeneous classes 242 and 243 are not included in the input data. Our generalization generates a similar fraction of these classes. However, the automatically generated areas are mostly not at the same location as in the manually generated reference data set. We argue though, that this is the result of an interpretation process, where different human interpreters would also yield slightly different results.

Input (DLM-DE) and the result match with 75%. This means that 25% of the area changes its class during generalization process. This is not an error; it is an unavoidable effect of the generalization. The κ-values 0.5-0.65 which stand for a moderate up to substantial agreement should also not be interpreted as bad results, because it is not a comparison with

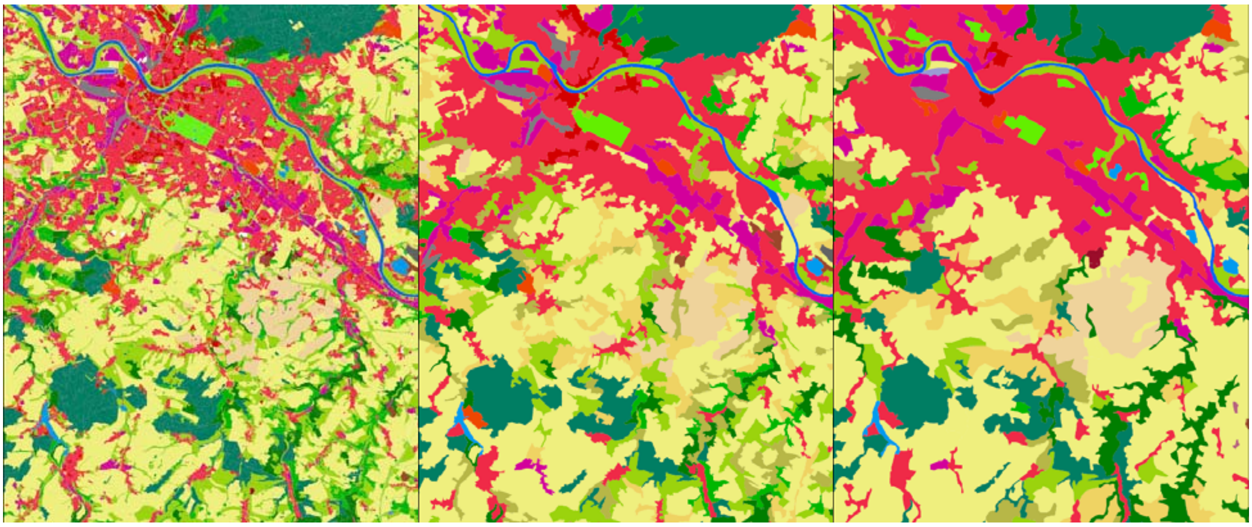


Figure 9: Extract (20 x 25 km) of test data set Dresden from left to right: input DLM-DE, our result and CLC 2006 as reference.

the real truth, or with a defined valid generalization, respectively.

5. OUTLOOK ON FUTURE WORK

With non-high end PC's it is not possible to process the whole of Germany at once. Theoretically it is possible to process all data sequentially. But most operators need a spatial environment for each polygon. Reading polygons and environments object by object leads to a high I/O-traffic, which is the bottleneck of the algorithms. This communication time would even get worse in a database implementation. Therefore we currently try some partitioning concepts which allow working on bigger areas than single polygons. The partitioning may also allow for a parallel processing of the data. However, the borders of the partitions should have only a very small effect on the generalization result.

Because of the aggregation algorithm there may be a chaotic effect - small changes causes a big changes of the result. We want to study and quantify these effects. Also the problem of the heterogeneous classes should be studied to find out, if it is possible to get more certain results.

Our next project aim is to derive land cover changes from the historic versions of CLC. The problem is to divide pseudo from the real changes. In account, that it is not possible to get real 5 ha changes from a 25 ha data set; we think about getting and aggregating the changes from the high resolution DLM-DE.

6. REFERENCES

Arnold, S., 2009. Digital Landscape Model DLM-DE – Deriving Land Cover Information by Intergration of Topographic Reference Data with Remote Sensing Data. *Proceedings of the ISPRS Workshop on High-Resolution Earth Imaging for Geospatial Information,* Hannover.

Bossard, M., Feranec, J. & Otahel, J., 2000. EEA CORINE Land Cover Technical Guide – Addendum 2000. – Technical Report No. 40, Kopenhagen.

Büttner, G., Feranec, G. & Jaffrain, G., 2006. EEA CORINE Land Cover Nomenclature Illustrated Guide – Addendum 2006. – European Environment Agency.

Douglas, D. & Peucker, T., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, *The Canadian Cartographer* 10 (1973) 112-122.

Geoff, B. et al. 2007. UK Land Cover Map Production Through the Generalisation of OS MasterMap®. *The Cartographic Journal*, 44 (3). 276-283.

Gomory, R., 1958. Outline of an algorithm for integer solutions to linear programms, *Bulletin of the American Mathematical Society*, 64(5), 274-278.

Haunert, J.-H., 2008. Aggregation in Map Generalization by Combinatorial Optimization, Vol. Heft 626 of *Reihe C*, Deutschen Geodätische Kommission, München.

Haunert, J.-H. & Sester, M., 2008. Area collapse and road centerlines based on straight skeletons, *GeoInformatica*, vol. 12, no. 2, p. 169-191, 2008.

Heinzle, F. & Anders, K.-H., 2007. Characterising Space via Pattern Recognition Techniques: Identifying Patterns in Road Networks, in: W. Mackaness, A. Ruas & L.T. Sarjakoski, eds, *Generalization of geographic information: cartographic modelling and applications*, Elsevier, Oxford, pp. 233-253.

Kirkpatrick, S., Gelatt, C. D. Jr., & Vecchi, M. P., 1983. Optimization by Simulated Annealing. In: *Science 220* (4598), 671. 13 May 1983.

Lillesand, T. M. & Kiefer, R. W., 1999. *Remote Sensing and Image Interpretation*, 4th edn, John Wiley &\& Sons.

Mackaness, W., A., Ruas, A. & Sarjakoski, L.T., 2007. *Generalisation of Geographic Information - Cartographic Modelling and Applications*, Elsevier Applied Science.

Pondrenk. M, 2002. Aufbau des DLM50 aus dem Basis-DLM und Ableitung der DTK50 – Lösungsansatz in Niedersachsen. In: *Kartographische Schriften, Band 6, Kartographie als Baustein moderner Kommunikation*, S.126-130, Bonn.

van Oosterom, P., 1995. The GAP-tree, an approach to 'on-the-fly' map generalization of an area partitioning, in: J.-C. Müller, J.-P. Lagrange & R. Weibel, eds, *GIS and Generalization - Methodology and Practice*, Taylor & Francis, pp. 120-132.

EXTRACTING THE SPATIAL-TEMPORAL RULES OF THE MESOSCALE OCEAN EDDIES IN THE SOUTH CHINA SEA BASED ON ROUGH SETS

Qi Guangya[a], Du Yunyan[a], Cao Feng[a]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China – (qigy, duyy, caof)@lreis.ac.cn

KEY WORDS: Mesoscale Ocean Eddy; Spatial-Temporal Relationship; Rough Sets; Rules Extraction; South China Sea

ABSTRACT:

Many different types of Mesoscale Ocean Eddies have been found in China's coastal and offshore since the 1970s. Domestic and foreign scholars are holding ongoing in-depth investigation and research in the South China Sea, especially since the TOPEX / Poseidon (T / P) data have been widely used. Due to the complex causes and numerous affecting factors of the Mesoscale Ocean Eddies in the South China Sea, the methods such as numerical simulation, quantitative statistics post limitations in analyzing the spatial-temporal relationships. This paper adopts rough sets theory to express the spatial-temporal relationships of the Mesoscale Ocean Eddies in the South China Sea, without adding any a priori information. Firstly, the paper extracts spatial-temporal rules of the Mesoscale Ocean Eddies in the South China Sea, by using the extracted eddy data from the remote sensing image. The decision making attributes respectively are sea area, time, and the eddy type. Then, the paper describes specific characters of the Mesoscale Ocean Eddies respectively from time and space, as well as the types. The results suggest this method effectively extracted the spatial-temporal rules of the Mesoscale Ocean Eddies from multi-source data sets, thus efficient support an in-depth understanding of the phenomenon of Mesoscale Ocean Eddies.

1. INTRODUCTION

Ocean eddies, the breakthrough understanding of the ocean environment in recent decades, play an important role in the impact of the exchange of the material and energy flux in the ocean. There is abundance of Mesoscale Ocean Eddies in the South China Sea, which has great significance in the country's military, production and environment and has attracted much attention of scholars home and abroad. The scholars (Huang Q Z, 1992; Guan B X, 1997; Sun X P, 1997; Li Y C, 2002; Lan J, 2006; Guan B X, 2006; Li L, 2002; Cheng X H, 2008) studied the Mesoscale Ocean Eddies in the South China Sea by using the quantitative methods as following: Numerical Simulation, Yang (Yang Q, 2000) simulated and analyzed the multi-eddy system in the northern South China Sea in winter by using a modified eddy-resolving the Princeton University Ocean Model (POM). Qian (Qian Y P, 2000) used the POM to numerically simulate the mechanisms of the formations of cold and warm eddies under the joint effects of the wind stress in the South China Sea. The numerical simulation method obtains the information of the flow field from particular sea area with boundary condition, and sets the affect conditions to indirectly analyze the eddies. This method could successfully simulate the currents, seabed topography, coastline, wind stress and other factors, but still restricted by the complex formation causes of eddies. When the spatial resolution is high, this method receives the computer capacity and speed limits. Remote Sensing Image Information Extraction, Ge (Ge Y, 2007) used a multifractal filtering technology to extract the ocean eddies. The extracted information contained shape, size, spatial distribution patterns and the direction of energy flow of eddies. This method is superior to the traditional extraction methods, but there are still shortcomings in the boundary effect problem. Using remote sensing image information extraction methods can extract a transient moment of the characteristics information of the Mesoscale Ocean Eddies, but it still relatively weak in extracting information of the Mesoscale Ocean Eddies as a whole moving process. Quantitative Statistics, Gu (Gu J S, 2007) tracked the mesoscale eddies in the northeastern South China Sea, using the data of sea surface height anomaly (SSHA) observed by TOPEX/POSEIDON (T/P) satellite altimeter and the altimeter optimum interpolation data in the modular ocean data assimilation system (MODAS), and statistically analyzed the characteristic values of the eddies. Cheng (Cheng X H, 2005) used the 11-yr (1993-2003) T/P, Janson and ERS1/2 altimeter data to acquire the temporal and spatial distribution characteristics of mesoscale eddies in the South China Sea. The seasonal and interannual variabilities as well as the forming mechanism of mesoscale eddies in the South China Sea were studied. Lin (Lin P F, 2007) identified and traced the mesoscale eddies in the South China Sea from 1993 to 2001 using T/P merged ERS1/2 altimeter data through several criteria, and statistically analyzed their space-time variation characteristics. This method limits by the observational data which can only quantitative analyze the localized Mesoscale Ocean Eddies during a particular time. There is a certain defect of model-based analysis of the Mesoscale Ocean Eddies. Due to the limitations of the quantitative methods mentioned above, the further study on the rules of spatial-temporal behaviour of the Mesoscale Ocean Eddies in the South China Sea is still needed.

Rough sets theory, whose distinct characteristic is not required any a priori information outside of the processed data (Wang G Y, 2001), is an approach of researching presentation, learning, concluding of the incomplete, uncertain knowledge and data (Miao D Q, 2008). This study adopts rough sets theory to express the spatial-temporal relationships and extracts the spatial-temporal rules of the Mesoscale Ocean Eddies in the South China Sea, and using the eddy data extracted from remote sensing image (Nov. 2003 to Jun. 2009) as an example. Firstly, the raw data is obtained from the U.S. Naval Research Laboratory, which includes sea surface height anomaly (SSH), sea surface temperature (SST) and the current field

(Current/Speed) data. The typical cases of the Mesoscale Ocean Eddies are derived from these raw data by experts. Secondly, two types of the Mesoscale Ocean Eddies attributes are calculated as the conditional attributes of rough sets decision-table, one is the Mesoscale Ocean Eddies's own characteristics and the other is characteristics of the spatial-temporal relationships. Different decision-tables are made with the different decision-making attribute, such as the area of occurrence, the time of occurrence, and the eddy type. The condition attributes and the decision-making attributes above composes the decision-table. Finally, this study applies Boolean discrete algorithms to discretize the decision-making table, and uses genetic algorithm to reduce the decision-making table and extracts the rules. A total of three tables are obtained to show the spatial-temporal relationship rules of the Mesoscale Ocean Eddies in the South China Sea.

2. METHODOLOGY

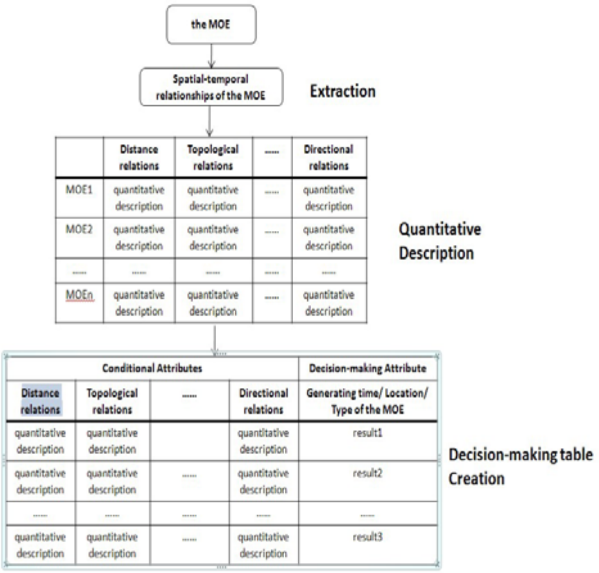
2.1 Expression of Spatial-Temporal Relationships of the Mesoscale Ocean Eddies based on Rough Sets

To extracting the spatial-temporal relationships of the Mesoscale Ocean Eddies, firstl express the spatial-temporal relationships of the Mesoscale Ocean Eddies quantitatively, replace them with the form of decision-making table by rough sets. Figure 1 shows the flow of expression of spatial-temporal relationships of the Mesoscale Ocean Eddies based on rough set.

(1) Selecting the spatial-temporal relationships of the Mesoscale Ocean Eddies: According to the prior knowledge, selecting the specific spatial-relationships of the Mesoscale Ocean Eddies as the object of study. For example, the distance relations/ topological relations/ direction relations between the aim eddy and the nearest one, whose generating time is the nearest to the aim one; the distance relations between the aim eddy and the mainland coastline.

(2) Quantitatively describing the spatial-temporal relationships of the Mesoscale Ocean Eddies: Using appropriate quantitative methods to describe the spatial-temporal relationships of the Mesoscale Ocean Eddies. For example, the topological relations can be described by the RCC-8 model (Randell DA, 1992; Randell, 1989).

(3) Creating the decision-making table of the spatial-temporal relationships of the Mesoscale Ocean Eddies: The rows of the decision table represent the instances of the Mesoscale Ocean Eddies. The columns are divided into conditional attributes part which represent the spatial-temporal relationships of the Mesoscale Ocean Eddies and decision-making attribute part which represents the results. The value of the each row is the results of the spatial-temporal relationships of the Mesoscale Ocean Eddies described by different methods (not including the decision attribute).


the MOE
Spatial-temporal relationships of the MOE
Extraction
Quantitative Description
Decision-making table Creation

	Distance relations	Topological relations	......	Directional relations
MOE1	quantitative description	quantitative description	......	quantitative description
MOE2	quantitative description	quantitative description	......	quantitative description
......	......	......	......	......
MOEn	quantitative description	quantitative description	......	quantitative description

Conditional Attributes				Decision-making Attribute
Distance relations	Topological relations	......	Directional relations	Generating time/ Location/ Type of the MOE
quantitative description	quantitative description		quantitative description	result1
quantitative description	quantitative description	......	quantitative description	result2
......	......	......	......	......
quantitative description	quantitative description	......	quantitative description	result3

Figure.1 Flow chart of expression of the relationship of the Mesoscale Ocean Eddies based on rough sets

2.2 Extraction of the Spatial-Temporal Rules of the Mesoscale Ocean Eddies

(1) Expressing the spatial-temporal relationships of the Mesoscale Ocean Eddies based on rough sets: Representing the spatial-temporal relationships of the Mesoscale Ocean Eddies in the form of decision-making table showed by Figure 1.

(2) Discretizing the decision-making table: Using rough sets theory to deal with the decision-making table, the values of which are represented as discrete data (eg integer, string, enumeration). If certain conditional attributes or decision-making attributes are continuous range values (such as float), they must be discretized first. Therefore, discretizing the decision-making table which (1) got.

(3) Reducing the spatial-temporal relationships: Reducing the decision-making table in order to extract the high fitness rules of the decision-making table. After reduction of the decision-making table, calculating the Coverage (Wang G Y, 2001) and Confidence (Wang G Y, 2001) of the rules from the rules table of the spatial-temporal relationships.

The rough sets rules could be expressed as A⇒ B.
The Coverage of the rules:

$$\alpha = \frac{|X \cap Y|}{|X|} \tag{1}$$

The Confidence of the rules:

$$\beta = \frac{|X \cap Y|}{|Y|} \tag{2}$$

Which

$$X = \{x \mid x \in U \wedge A_x\}$$
$$Y = \{x \mid x \in U \wedge B_x\}$$

A_x indicates that the value of the conditional attributes of x satisfied with A; B_x indicates that the value of the conditional attributes of x satisfied with B. Set X is the instances whose conditional attributes satisfied with A; Set Y is the instances whose conditional attributes satisfied with B. The Confidence of the rules of the spatial-temporal relationships of the Mesoscale Ocean Eddies represents the credibility of the rules, and the Coverage represents the degree of support of the rules.

3. APPLICATION DEMONSTRATION

3.1 About the Experimental Area

The South China Sea is a semi-enclosed marginal sea located in 98.5 ° E -122.5 ° E, 0 ° N-24.5 ° N. Its area is about 3.5km × 106km, the average water depth is up to 1800m, and the maximum water depth of about 5000m (Wang G H, 2005). There are complex seabed terrain and lots of islands in the South China Sea, the water of which is shallow in the northwest and southwest part and deep in the central and eastern part. Through a number of straits, the South China Sea links with the ocean and the adjacent sea.

The South China Sea locates in the monsoon climate zone where the strong northwest winds are prevailing in winter and southwest monsoon in summer. In general, it's winter monsoon from October to March of the next year, summer monsoon from June to August, spring monsoon transition period from April to May, and autumn monsoon change period in September (Wang G H, 2005).The study have shown that the upper circulation is mainly affected by the monsoon-driven (Wang G H, 2005). In winter, the surface circulation of the South China Sea is in the cyclone-type situation; in summer, the surface circulation shows anti-cyclonic circulation trend (Huang Q Z, 1992; Li L, 2002; Wang J, 2003).

In the South China Sea, there are many active Mesoscale Ocean Eddies, which change seasonally and greatly influenced by the monsoon and heat exchange on the sea (Guan B X, 2006). The studies have shown that the Mesoscale Ocean Eddies mainly occurred in the southwest of Taiwan Island, and off the west coast of Luzon and the east sea of Vietnam (Wang G H, 2004). The regions of the Mesoscale Ocean Eddies are mainly located in the line of east of the southern part of Vietnam to the southwest of Taiwan, showing the northeast - southwest distribution (Lin P F, 2007; Wang G H, 2004). The warm eddies is more than the cold eddies(Lin P F, 2007; Wang G H, 2004). During the winter monsoon period, the Mesoscale Ocean Eddies of the South China Sea generate the most (Wang G H, 200 4), and very few of them come from Northwest Pacific (Lin P F, 2007). 80% of the Mesoscale Ocean Eddies move westwards with the change "Σ" type distribution in latitude (Lin P F, 2007). Therefore, the Mesoscale Ocean Eddies in the South China Sea distribute in certain amount of time and space laws, which need to be further in-depth quantitatively studied.

This research expresses the spatial-temporal relationships and extracts the spatial-temporal rules of the Mesoscale Ocean Eddies in the South China Sea based on the above method from November 2003 to June 2009. The raw data is obtained from the U.S. Naval Research Laboratory, which includes sea surface height anomaly (SSHA), sea surface temperature (SST) and the current field (Current/Speed) data. The SSHA data are assimilated from data of the ENVISAT, GFO and JASON-1, etc. The SST data are assimilated from IR data. The time resolution is one day, and the spatial resolution is (1 / 32) °. The typical cases of the Mesoscale Ocean Eddies are derived from these raw data by experts. The typical cases of the South China Sea is totally 391, in which warm eddies are 291, and the cold eddies are 100. Figure 2 shows a warm eddy case and the corresponding three kinds of environment elements field data.

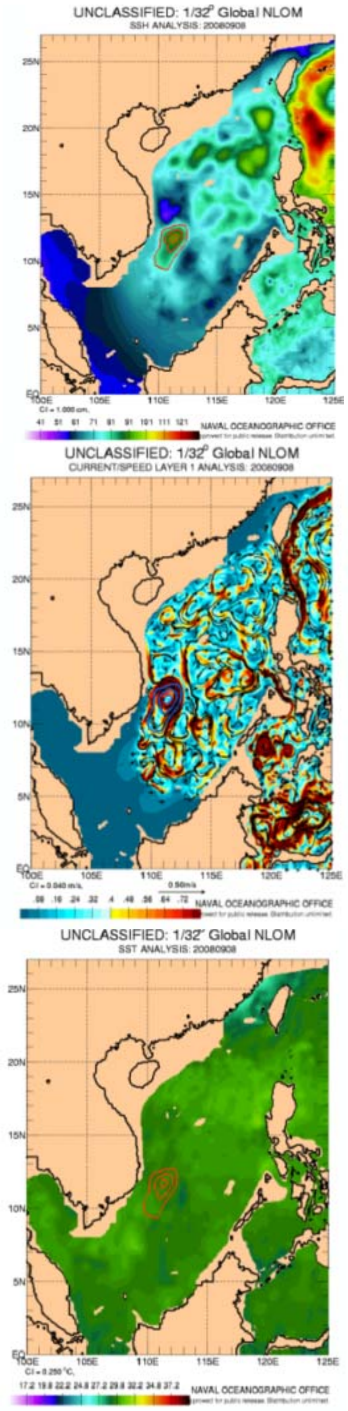

UNCLASSIFIED: 1/32° Global NLOM
SSH ANALYSIS: 20080908
25N
20N
15N
10N
5N
EQ
100E
105E
110E
115E
120E
125E
NAVAL OCEANOGRAPHIC OFFICE
UNCLASSIFIED: 1/32° Global NLOM
CURRENT/SPEED LAYER 1 ANALYSIS: 20080908
NAVAL OCEANOGRAPHIC OFFICE
UNCLASSIFIED: 1/32° Global NLOM
SST ANALYSIS: 20080908
NAVAL OCEANOGRAPHIC OFFICE

Figure.2 The example of the mesoscale ocean eddies in the South China Sea vector expression and the background field data

3.2 Expression of the Spatial-Temporal Relationships in the Experimental Area based on Rough Sets

Use the above-mentioned method to express the spatial-temporal relationships. Focus to different study of the Mesoscale Ocean Eddies, there are different choices of the decision-making attributes. For example, when the research emphasis on the rules of spatial-temporal relationships in different regions, the decision-making attribute is need to be the location of the Mesoscale Ocean Eddies; and if the research is focus on the rules of spatial-temporal relationships in different types of the Mesoscale Ocean Eddies, the decision-making attribute is need to be the type of the Mesoscale Ocean Eddies. In this research, the three decision-tables are made with the different decision-making attribute, which is the area of occurrence, the time of occurrence, and the eddy type.

Two types of the Mesoscale Ocean Eddies attributes are calculated as the conditional attributes of rough sets decision-table, one is the Mesoscale Ocean Eddies's own characteristics and the other is characteristics of the spatial-temporal relationships.

(1) Location of the Mesoscale Ocean Eddies where it generated (EddyZone): according to different physical characteristics of the marine environment of the South China Sea, divided the South China Sea into four sea areas, northeast part, central part, southeast part and southwest part.

(2) Time of the Mesoscale Ocean Eddies where it generated (EddyTime): divided by season, spring from March to May, summer from June to August, autumn from September to November, winter from December to February.

(3) Type of the Mesoscale Ocean Eddies (EddyType): the warm eddies and the cold eddies.

(4) Intensity of the Mesoscale Ocean Eddies (EddyIntensity): the amplitude difference of the center and the periphery, the unit is meter.

(5) Vorticity of the Mesoscale Ocean Eddies (Vorticity): the unit is s^{-2}.

$$\varsigma = {\partial v}/{\partial x} - {\partial u}/{\partial y} \approx {8gM}/{fD^2} \qquad (3)$$

where M is the intensity of the Mesoscale Ocean Eddies

D is the diameter of the Mesoscale Ocean Eddies

f is Coriolis parameter

(6) Horizontal scale of the Mesoscale Ocean Eddies (Horizontal): half of the sum of east-west diameter and south-north diameter, the unit is meter.

(7) Temperature of the center of the Mesoscale Ocean Eddies (CenterTemp): the unit is degree Centigrade.

(8) Temperature difference of the Mesoscale Ocean Eddies (EddyTemp): the unit is degree Centigrade.

(9) Depth of the sea water (depth): the average of the sea water in the central part of the Mesoscale Ocean Eddies, the unit is meter.

(10) Distance relations (distance): the distance from aim eddy to the reference eddy[1], represented by Euclidean, the unit is meter.

(11) Directional relations (direction): the directional relation to the reference eddy, represented by eight directions.

(12) Topological relations (topology): the topological relation to the reference eddy, represented by RCC-8 Model (Randell DA, 1992; Randell, 1989).

Using ArcGIS secondary development VBA to achieve the above 12 attributes of the 391 typical Mesoscale Ocean Eddies in the South China Sea, the above-mentioned indicator (1), (2), (3), respectively, be the decision-making attributes, other indicators as conditional attributes of the rough sets decision-making table. Table 3 is the example of the decision-making table, whose decision-making attribute is location of the Mesoscale Ocean Eddies.

3.3 Spatial-temporal rules extraction

Using the specific software, Rosetta (Qhrn A, 1999), which cooperative R & D by the Department of Computer and Information Science of Norwegian University of Science and Technology and Institute of Mathematics of University of Warsaw Poland (Wang G Y, 2001), calculate and extract the rules.

(1) Import the decision-making table of the spatial-temporal relationships of the Mesoscale Ocean Eddies into Rosetta, whose decision-making attribute is location. (Table 3)

(2) Discretize the continuous range of values of the table, using the discrete method of combination of Boolean and the rough sets theory.

(3) Reduce the spatial-temporal relationships and extract the rules of the decision-making table which has been discretized, using the genetic algorithm. The specific method of implementation can be found in paper (Qhrn A, 1999).

(4) Similarly, changing the decision-making attributes and taking the above steps, obtain the spatial-temporal rules corresponding to the attributes.

[1] The reference eddy is the nearest eddy to the aim eddy in time. With the aim eddy as the center and radius as 1.83km × 60km, search the eddy whose time before and most neighboring the aim eddy within the buffer zone.

Conditional attributes											**Decision-making attribute**
Time	**Type**	**Intensity**	**Vorticity**	**Horizontal scale**	**Temperature of center**	**Temperature difference**	**Distance relations**	**Directional relations**	**Topological relations**	**Depth**	**Location**
Winter	Warm Eddy	0.0912	$0.692*10^{-6}$	150762.08	23.31	0.028	77340	East	Disjoint	2028	Central of the South China Sea
Summer	Cold Eddy	0.1845	$0.973*10^{-6}$	200655.8	29.94	0.0753	20074	Southeast	Touch	802	Southwest of the South China Sea
……	……	……	……	……	……	……	……	……	……	……	……

Table 3 Decision table of spatial-temporal of the Mesoscale Ocean Eddies,whose decision attribute is the location

4. RESULTS

Analyze the spatial-temporal relationships of the Mesoscale Ocean Eddies respectively from location, time and type by convert the rules table, as shown in Table 4 to Table 6. In Table 4, from a regional point of view, due to the winter monsoon the Mesoscale Ocean Eddies generate mostly in in winter and spring in the northeast of the South China Sea, both warm and cold eddies appear, their temperature of the center is relatively low and vorticity is greater, and horizontal scale is in the low level. In the southeast of the South China Sea, there are more warm eddies, which generate mostly in winter (The warm eddies have taken place in the winter in Southwestern Luzon (Wang G H, 2004).); the temperature of the center is higher, the vorticity, intensity, horizontal scale and temperature difference are low; and the topological relation between the aim eddy and the reference eddy is overlap. In the southwest of the South China Sea, the horizontal scale of the Mesoscale Ocean Eddies is high, but the intensity is low; the Mesoscale Ocean Eddies generate mostly in where the water depth is lower than 1756m, and basically appear in the southern of the reference eddy.

Characteristic Attributes	**Northeast of the South China Sea**	**Southeast of the South China Sea**	**Southwest of the South China Sea**
Temperature of the center	<26.92°	[26.92°,28.53°)	
Vorticity/ s^{-1}	$>0.7*10^{-6}$	$<0.7*10^{-6}$	
Type	Warm, Cold Eddy	Warm Eddy	
Time	Winter, Summer	Winter	Summer
Horizontal scale/km	<172.6	<172.6	>194.9
Intensity/m		<0.08 or [0.08,0.12)	[0.08,0.11)
Temperature difference		[0.03°,0.16°)	
Average depth of the water/m		<1756 or >1897	<1756
Topological relations between the aim and the reference eddy		Overlap	
Directional relations between the aim and the reference eddy			South

Table 4 Rules of spatial-temporal of mesoscale eddies in different zones

In Table 5, from a time point of view, in winter, the temperature of center, the temperature difference and the intensity of the Mesoscale Ocean Eddies is low, but the vorticity is high; there are more warm eddies, which generate mostly in the central South China Sea; the topological relation between the aim and the reference eddy is overlap, and the Mesoscale Ocean Eddies basically appear in the northeast of the reference eddy. In autumn, the temperature of the center of the Mesoscale Ocean Eddies is relative low, but the intensity is high. There are mostly warm eddies which generate in the central South China Sea. In summer, there are more warm eddies, whose temperature difference is low, and they generate in the southwest of the South China Sea.

Characteristic Attributes	**Winter**	**Autumn**	**Summer**
Temperature of the center	<26.22°	[26.22°, 28.72°)	>29.59°
Type	Warm Eddy	Warm Eddy	Warm Eddy
Location	Central South China Sea	Central South China Sea	Southwest of the South China Sea
Topological relations between the aim and the reference eddy	Overlap		
Directional relations between the aim and the reference eddy	Northeast		
Vorticity/ s^{-1}	$>0.6*10^{-6}$		
Intensity/m	[0.08, 0.11)	>0.16	
Temperature difference	<0.18°		<0.18°
Average depth of the water/m	[1723, 1881)or [1881, 2092)		[1723,1881)

Table 5 Rules of spatial-temporal of mesoscale eddies in different times

In table 6, from a type point of view, the warm and cold eddies show a very clear spatial-temporal characteristics. The vorticity and the intensity of the warm eddies are relatively low, but they are high of cold eddies. The temperature of the center of the warm eddies is high, mostly higher than 29.02°, but the temperature of the center of the cold eddies is relatively low, in [27.72°, 29.02°) or lower. The warm eddies generate mostly in where the water depth is lower than 1741m, or [1723m, 1881m), but the cold eddies generate in where the water depth is higher than 2105m, or [1879m, 2105 m). The warm eddies is near the reference eddy, but the cold eddy is far away from the reference eddy. Also, the warm eddies mostly generate in the southeast of the South China Sea and central South China Sea in winter, but the cold eddies mostly generate in the southwest and northeast of the South China Sea.

Characteristic Attributes	**Warm Eddy**	**Cold Eddy**
Vorticity/ s^{-1}	$<0.7*10^{-6}$	$>0.9*10^{-6}$
Temperature of the center	>29.02°	<27.72° or [27.72°, 29.02°)
Distance relations between the aim and the reference eddy /m	<48.17	>48.17
Intensity/m	<0.1	>0.17

Average depth of the water /m	<1741 or [1741, 1879)	[1879, 2105) or >2105
Time	Winter	Autumn, Summer
Location	Southeast of the South China Sea, Central South China Sea	Southwest of the South China Sea, Northeast of the South China Sea

Table 6 Rules of spatial-temporal of different types of mesoscale eddies

5. CONCLUSION

According to the extracted spatial-temporal rules, following conclusions can be drawn. The warm eddies were produced in winter (which is same as Lin's (Lin P F, 2007) statistic result) and generated mostly in the southeast and middle of the South China Sea, where the place is relatively shallow. Their intensity and vorticity are relatively low, the temperature of their central region is high, and they only move a short distance. On the other hand, the cold eddies are produced in spring and autumn, and are generated mostly in the southwest and northeast of the South China Sea (which is same as Lin's (Lin P F, 2007) statistic result), where the place is relatively deep. Their intensity and vorticity are relatively high, the temperature of their central region is low (Consistent with the characteristics of the cold eddies), and they move a long distance.

6. DISCUSSION

This study adopts rough sets theory to express the spatial-temporal relationships and extracts the spatial-temporal rules of the Mesoscale Ocean Eddies in the South China Sea, by using the data extracted from the raw data (Nov.2003-Jun.2009) obtained from the U.S. Naval Research Laboratory. These rules not only describe the spatial-temporal relationships, but also specifically describe the characteristics of the two types of Mesoscale Ocean Eddies in the South China Sea. The results suggest this method effectively extracted the spatial-temporal rules of the Mesoscale Ocean Eddies from multi-source data sets. There are different choices of the decision-making attributes focus on different aims, and it's more flexible to extract the spatial-temporal rules, with the feasibility of practical application. However, the method requires a priori knowledge in the selection of the spatial-temporal relationships from the Mesoscale Ocean Eddies and the specific quantitative description of them. Selecting different spatial-temporal relationships, different results will obtain.

Besides, the experimental data is identified by digitizing the remote sensing data. The horizontal scale of some cold eddies is lower and the cycle is shorter, thus the number of the cold eddies is low. It is also the impact of the results. This work can also be augmented through the following means: □ increase in experimental data; □ select factors that can better reflect spatial-temporal relationships; □use other discretization and reduction algorithms based on rough sets theory and compared the results.

REFERENCE

Cheng X H, 2008. Distribution and propagation of mesoscale eddies in the global ocean learnt from altimetric data. *Advances In Marine Science*, 26(4), pp. 447-453.

Cheng X H, 2005. Seasonal and interannual variabilities of mesoscale eddies in South China Sea. *Journal of Tropical Oceanography*, 24(4), pp. 51-59.

Ge Y, 2007. Multifractal filtering method for extraction of ocean eddies from remotely sensed imagery. *ACTA Oceanologica Sinica*, 29(5), pp.40-47.

Guan B X, 1997. Warm eddy in the open sea east of Hainan Island. *Journal of Oceanography of HuangHai & BoHai Seas*, 15(4), pp.1-7.

Guan B X, 2006. Overview of studies on some eddies in the China seas and their adjacent seas-I. The South China Sea and the region east of Taiwan. *ACTA Oceanologica Sinica*, 28(3), pp.1-16.

Gu J S, 2007. Statistics of the mesoscale eddies on both sides of the Luzon strait. *Advances In Marine Science*, 25(2), pp.139-148.

Huang Q Z, 1992. General situations of the current and eddy in the South China Sea. *Advances In Earth Science*, 7(5), pp.1-5.

Lan J, 2006. Seasonal variability of cool-core eddy in the western South China Sea. *Advances In Earth Science*, 21(11), pp.1145-1152.

Li L, 2002. A review on mesoscale oceanographical phenomena in the South China Sea. *Journal Of Oceanography In Taiwan Strait*, 21(2), pp.562.

Lin P F, 2007. Temporal and spatial variation characteristics on eddies in the South China Sea. *ACTA Oceanologica Sinica*, 29(3), pp.14-22.

Li Y C, 2002. Observation of mesoscale eddy fields in the sea southwest of Taiwan by TOPEX/POSEIDON altimeter data. *ACTA Oceanologica Sinica*, 24(Supp.1), pp.163-170.

Li Y C, 2003. Seasonal and interannual variabilities of mesoscale eddies in northeastern South China Sea. *Journal of Tropical Oceanography*, 22(3), pp.61-70.

Miao D Q, 2008. *Rough sets theory algorithms and applications*. Beijing, TsingHua University Press, pp.V.

Qhrn A, 1999. *Discernibility and rough sets in medicine, pp.tools and application*. Norway, Norwegian University of Science and Technology, pp.41-51

Qian Y P, 2000. Numerical modelings of the wind forced cold and warm gyres in the South China Sea. *Chinese Journal of Atmospheric Sciences*, 24(5), pp.625-633.

Randell DA, 1989. Modelling topological and metrical properties in physical processes. Proceedings of the 1st International Conference on the Principles of Knowledge Representation and Reasoning. San Francisco, Morgan Kaufmann Publishers, pp.55-66.

Randell DA, 1992. A spatial logic based on regions and connection. Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning. San Francisco, pp.Morgan Kaufmann Publishers, pp. 165-176.

Sun X P, 1997. Analysis on the cold eddies in the Sea Area Northeast of Taiwan. *Marine Science Bulletin*, 16(2), pp.1-10.

Wang G H, 2004. Discussion on the movement of mesoscale eddies in the South China Sea. Qingdao: Ocean University of China.

Wang G H, 2005. Advances in Studying Mesoscale Eddies in South China Sea. *Advances In Earth Science*, 20(8), pp.882-886.

Wang G Y, 2001. *Rough sets theory and knowledge acquisition*. Xi'an, Xi'an JiaoTong University Press, pp.1.

Wang J, 2003. Characteristics of sea surface height in South China Sea based on data from TOPEX/Poseidon. *Journal of Tropical Oceanography*, 22(4), pp.26-33.

Yang Q, 2000. Numerical study about the mesoscale multi-eddy system in the northern South China Sea in winter. *ACTA Oceanologica Sinica*, 22(1), pp.27-34.

THE EFFECT OF DISTANCE CORRECTION FACTOR IN CASE-BASED PREDICTIONS OF VEGETATION CLASSES IN KARULA, ESTONIA

M. Linder*, L. Jakobson, E. Absalon

Institute of Ecology and Earth Sciences, University of Tartu, Estonia 51014, Vanemuise 46 – madlili@ut.ee

KEY WORDS: Spatial Autocorrelation, Case-based Predictions, Predictive Vegetation Mapping, Machine Learning, Remote Sensing and Map Data, Distance-related Similarity Correction

ABSTRACT:

The aim of this study was to investigate the applicability of the distance correction parameter (DCP) integrated to the case-based prediction system CONSTUD to reduce the effect of spatial autocorrelation of training data in machine learning process. To achieve this, calculated similarity between observations is decreased by the so-called distance correction value (DCV – the quotient of DCP and distance between two observations). 50 machine learning iterations were carried through in the case of different DCP-s from 0 to 15 000 m using random samples generated from 450 training observations from southern Estonia (Karula National Park and its vicinity). Independent validation samples were used to estimate the effects of the use of each DCP. Machine learning results showed that the Cohen’s kappa index of agreement decreased in accordance with the increase of DCP-s. The correspondences of field observations and predicted values followed the same trend. The explanation would be that with the increase of DCP-s successively more observations were rejected as useful ones. Conversely, no considerable decrease in correspondences of the predictions was recognized when DCP was increased. In our case, probably the most useful exemplars were chosen and the less useful ones were left beyond. As a result, scattered and probably spatially and thematically highly representative sample of observations remained. The border might be drawn at DCP from which the number of the in-between distances started to decrease considerably, but the correspondence in validation sample estimations as well as in training sample estimations remained relatively stable.

1. INTRODUCTION

This paper is related to the issues of autocorrelation of observations in training samples and to the spatial and thematic representativeness of training data, and also to the overtraining problems in predictive vegetation mapping.

Spatial autocorrelation occurs when locations close to each other have more similar values than those further apart (the values of variables are not independent from each other). Autocorrelation of ecological phenomena may arise for different reasons (see Sokal & Oden, 1978). Positive spatial autocorrelation in moderate distances may accrue from spatial and temporal synchrony of certain abiotic factors that shape particular landscape patterns, e.g., blotched configuration of landscape components. In farther distances, positive autocorrelation may originate from regular variation of environmental gradients and habitat patches. Populations and species may be spatially aggregated due to their dispersal limits caused by different environmental and historical as well as intrinsic organism-specific factors. Among other reasons causing spatial autocorrelation in (predictive) models, is omitting an important variable from the model, observation biases (variance in data collection, sampling and mapping) (Dormann et al., 2007), etc.

Spatial autocorrelation may be interpreted as intrinsic feature of a phenomenon providing additional information for spatial analysis. When spatial autocorrelation occurs, the values of variables are predictable on the basis of the values of the same variable in other locations. Luoto et al. (2005) found that performance of species–climate models depends on geographical attributes of the species, including spatial autocorrelation. They also found that butterfly species with more aggregated occurrence pattern (expressing high spatial autocorrelation) were better predicted compared to the species with scattered distribution (exhibiting low autocorrelation).

However, the presence of spatial autocorrelation is frequently a disadvantage for hypothesis testing and prediction, because it violates one of the main assumptions of standard statistical analyses that residuals are independent and identically distributed (Dormann et al., 2007). The presence of positive spatial autocorrelation in model residuals (spatial dependency) may bias parameter estimates and can increase the likelihood of type I statistical error (Betts et al., 2006).

Since spatially close locations/observations tend to be similar due to spatial autocorrelation, they predict each-other with great accuracy. As a result, deceptively high prediction accuracy (overtraining) occurs and the application of this set of training observations for more distant locations would not be so reliable.

Accounting for spatial autocorrelation should increase prediction versatility. Taking into account autocorrelation is crucial when image data are classified, because estimations of classification accuracy that compare prediction and actual situation pixel by pixel, tend to overestimate results due to the autocorrelation of the pixel values (Muchoney & Strahler, 2002).

A variety of widespread statistical tools have been developed to correct for the effects of spatial autocorrelation in species distribution data. Dormann et al. (2007) presented different statistical approaches that efficiently accounted for spatial autocorrelation in analyses of spatial data. Most of the spatial modeling techniques they tested on spatially autocorrelated simulated data showed good type I error control and precise

[*] Corresponding author.

parameter estimates. Accounting for autocorrelation via autologistic models has become common (e.g., Augustin et al., 1996; Osborne et al., 2001; Luoto et al., 2005). It has been shown that including spatial autocovariates improves model prediction success (Augustin et al. 1996; Osborne et al., 2001; Knapp et al., 2003; Betts et al., 2006). Generalized estimating equations (GEE – an extension of generalized linear models) have been used by Augustin et al., 2005, Carl & Kühn 2007, etc. The use of GEE models reduced the autocorrelation of the residuals considerably indicating effectively removed spatial dependency. Though, Diniz-Filho et al. (2003) concluded that ignoring spatial autocorrelation does not cause problems necessarily in all analyses.

The aim of this study was to introduce and test the usability of the distance correction parameter integrated to the case-based prediction system CONSTUD for extenuating the effect of the autocorrelation in predictions of spatial phenomena (in this case – vegetation classes).

2. METHODOLOGY

2.1 Study Area and Field Data

450 training observations from southern Estonia (Karula National Park and its vicinity; Figure 1) were gathered mostly during the summers 2007 and 2008, partly during the inventories from 2001 to 2007. EUNIS classes (Davies et al., 2004) were used as a predictable variable.

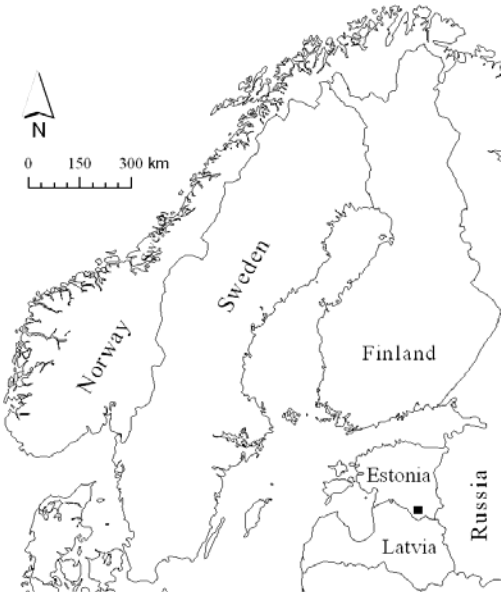


Figure 1. Location of the study area (black square).

2.2 Remote Sensing and Map Data

Data layers of explanatory variables were: rasterized 1: 10 000 Estonian base map and 1: 10 000 digital soil map, Landsat 5 TM satellite images (scenes 186-19 and 186-20) of 21st of May 2007 and 9th of August 2007, and the orthophotos from the year 2005. Layers for red, green, blue, yellow, hue, saturation and lightness were derived from orthophotos. In the case of satellite images, the NDVI layers were derived from the near-infrared and red channel values. In addition, the Baltic SRTM30 (*Shuttle Radar Topography Mission*) elevation model was used. The data layers were prepared according to the prerequisites for the application of CONSTUD (CONSTUD, 2009) using ArcGIS 9, Idrisi Andes, LSTATS (LSATS, 2009) and an original application for rasterizing the soil map.

2.3 Case-based Prediction System CONSTUD

The case-based (Aha, 1998; Remm, 2004) machine learning and prediction system created in the University of Tartu by Kalle Remm was used (more details and case studies in Linder et al., 2008; Remm & Remm, 2008; Remm et al., 2009; Remm & Remm, 2009; Tamm & Remm, 2009; CONSTUD, 2009). CONSTUD was used for: 1) calculating the pattern indices in training locations from map and image data (explanatory variables), 2) machine learning – iterative search for the best set of feature weights of the observations and the best observations (exemplars), 3) predictions of vegetation classes.

Decisions are made on the basis of similarity between studied cases and predictable sites in CONSTUD. Similarity between observations is calculated as a weighted average of partial similarities of single features (further details: Linder et al., 2008; Remm, 2004). During machine learning process, goodness-of-fit of predictions is estimated using leave-one-out cross validation (the predicted value for every observation is calculated using all exemplars except this particular one), and in the case of multinomial variable (like vegetation classes), Cohen's kappa index of agreement is used to measure the correspondence of predictions to observations.

2.4 Distance-related Similarity Correction

Distance correction parameter (DCP) is integrated to the system CONSTUD to reduce the effect of spatial autocorrelation in training data in machine learning process. DCP regulates the extent of reciprocal prediction of close observations by decreasing the calculated similarity between observations in proportion to the inverse distance between them. The extent of decrease is regulated by DCP (in meters) chosen by the user. Distance correction value (DCV) is calculated as the ratio of DCP and distance between two observations. Then calculated similarity between these values is corrected. Corrected similarity value (CSV) is gained as calculated similarity (from 0 to 1) minus DCV. If the distance between two observations is equal to or less than DCP, then CSV is set to 0 even if the calculated similarity is 0.9. The closer are observations, the higher is the rate the similarity between them is corrected. In this study, DCP-s from 0 to 20 000 m were tested.

2.5 Calculations

First, explanatory variables (spatial pattern indices) from image and map data layers were calculated. Then, 50 machine learning iterations were carried through in the case of each selected DCP in two stratified random samples generated from 450 training observations. One of the samples was first used as a training sample and the other as a validation sample. Then the roles were exchanged and the results were averaged. Independent validation samples were used to give the estimation for the use of each particular DCP. Finally, the correspondences for predictions (the proportion of coincident observations among all observations) were calculated.

3. RESULTS

The results showed that the Cohen's kappa index of agreement continually decreased with the increase of the DCP-s (Figure 2).

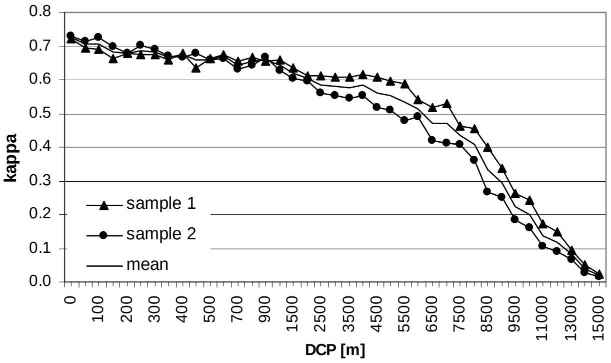


Figure 2. Kappa values gained during machine learning iterations of two random samples using different distance correction parameters (DCP-s).

The correspondences of field observations and predicted values in machine learning followed the same trend (Figure 3). The reason is probably the fact that with the increase of DCP lower weights were attributed to successively more observations.

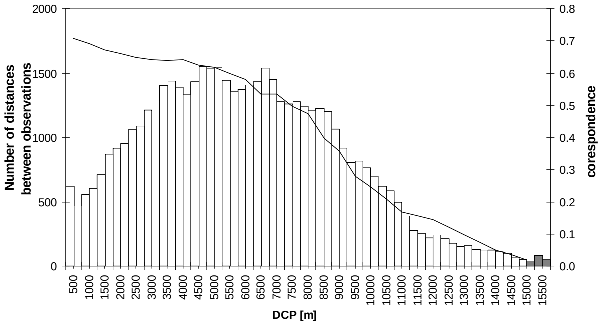


Figure 3. Line – correspondence of field observations and classes estimated during machine learning iterations (mean of two samples). White columns – distribution of all the distances between all observations (total of 50 400). Grey columns – distances between observations used in the last samples (in the case of the highest DCP – 15 000 m – 179 distances, i.e., 0.36% of all distances were comprised).

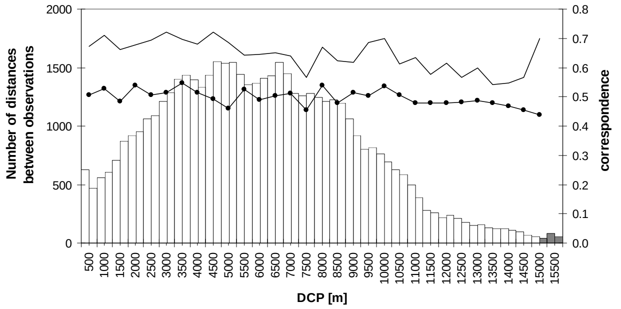


Figure 4. Line with dots – correspondences when estimating validation samples (mean of the two samples). Regular line – correspondences when estimating machine learning samples (mean of the two samples). White and grey columns – see Figure 3 caption.

Conversely, in the case of predictions, no considerable decrease in correspondences was recognized when DCP was increased (Figure 4). Furthermore, the lines of correspondences approached to each other (Figure 4, 5).

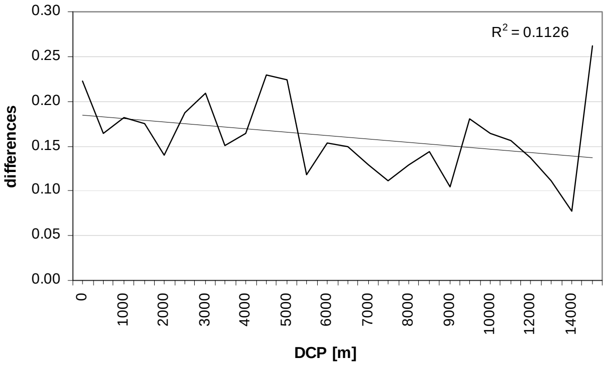


Figure 5. Differences between correspondences of estimated samples that were fitted during machine learning iterations and those of validation samples.

Samples from within the range of spatial autocorrelation may be inefficiently large (e.g., unduly time-consuming), because observations with spatially autocorrelated values will probably add little independent information (Dormann et al., 2007). It might be suggested that into the samples of our study, the most useful exemplars were chosen and the less useful ones were left beyond by CONSTUD. As a result, dispersed and probably highly (spatially and thematically) representative compact samples of observations remained (Figure 6). The border might be drawn at DCP from which the number of the in-between distances started to decrease considerably, but the correspondence in validation sample estimations as well as in training sample estimations remained relatively stable – in this case, at DCP of somewhere between 8500 and 9000 m (Figure 4).

4. CONCLUSIONS

Relying upon the results of this study, the use of distance correction parameter in case-based prediction and machine learning system CONSTUD gives a presumably thematically and spatially representative training sample which in turn reduces or removes the effect of autocorrelation and noise in data. This enables reducing the time expended on calculations of predictions. However, as far as only 450 observations were used (furthermore, these were divided into training and validation samples), wider interpretations could be biased. Using higher amount of field observations might give converse results, or might just increase the time for calculations. Also, expanding the area from where field data are gathered from could have unpredictable effects, due to the unique character of different landscape regions or for any other reason.

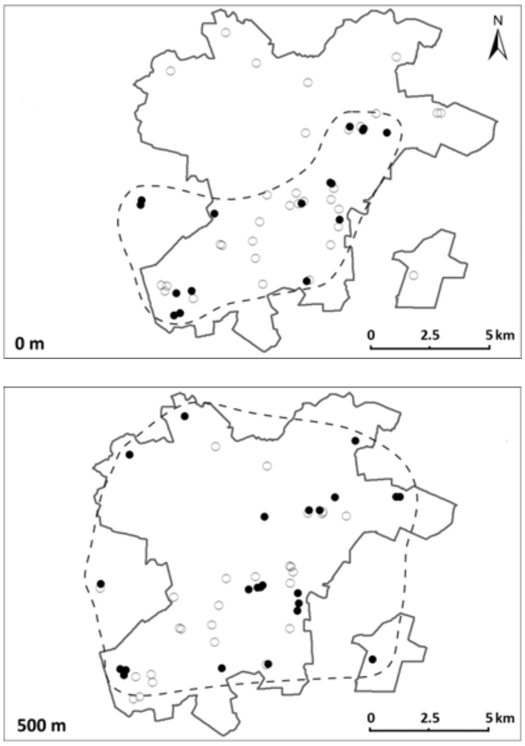


Figure 6. The effect of the use of distance correction parameter (DCP) in the case of coniferous forest observations in and around Karula National Park (within grey boundary line). Black dots – used exemplars, transparent dots – cases that turned out to be not very useful during machine learning iterations. Upper figure – DCP = 0 m, bottom figure – DCP = 500 m. Compared to the case when DCP was not used, the used exemplars are more dispersed (area within dashed line) (Linder et al., 2009, modified).

REFERENCES

Aha, D. W., 1998. The omnipresence of case-based reasoning in science and application. *Knowledge-Based Systems*, 11, pp. 261–273.

Augustin N.H., Kublin E., Metzler B., Meierjohann E., von Wuhlisch G., 2005. Analyzing the spread of beech canker. *Forest Science*, 51, pp. 438–448.

Augustin N.H., Mugglestone M.A., Buckland S.T., 1996. An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, 33, pp. 339–347.

Betts, M. G., Diamond, A.W., Forbes, G.J., Villard, M.-A., Gunn, J.S., 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling*, 191, pp. 197–224.

Carl, G., Kühn, I., 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modelling*, 207, pp. 159–170.

CONSTUD, 2009. http://www.geo.ut.ee/CONSTUD (accessed 13 Dec. 2009).

Davies, C. E., Moss, D., Hill, M. O., 2004. EUNIS Habitat Classification. Revised 2004. http://eunis.eea.europa.eu/upload/EUNIS_2004_report.pdf (accessed 15 Oct. 2009).

Diniz-Filho J.A.F., Bini L.M., Hawkins B.A., 2003. Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography*, 12, pp. 53–64.

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30(5), pp. 609–628.

Knapp, R.A., Matthews, K.R., Preisler, H.K., Jellison, R., 2003. Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. *Ecological Applications*, 13, pp. 1069–1082.

Linder, M., Remm, K., Absalon, E., 2009. The utility of the machine learning and prediction system CONSTUD. In: Mander, Ü., Uuemaa, E., Pae, T. (Eds.). *Uurimusi eestikeelse geograafia 90. aastapäeval. Publicationes Instituti Geographici Universitatis Tartuensis*, 108, pp. 52–62. Tartu: Tartu University Press.

Linder, M., Remm, K., Proosa, H., 2008. The application of the concept of indicative neighbourhood on Landsat ETM+ images and orthophotos, using circular and annulus kernels. In: Ruas, A., Gold, C. (Eds.). *Proceedings of the 13th International Symposium on Spatial Data Handling*, Montpellier, France, 23rd-25th June, pp. 147–162. Springer.

LSTATS, 2009. http://www.geo.ut.ee/LSTATS (accessed 13 Dec. 2009).

Luoto M., Poyry J., Heikkinen R.K., Saarinen K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, 14, pp. 575–584.

Muchoney, D. M., Strahler, A. H., 2002. Pixel- and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment*, 81, pp. 290–299.

Osborne P.E., Alonso J.C., Bryant R.G., 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, 38, pp. 458–471.

Remm, K., 2004. Case-based predictions for species and habitat mapping. *Ecological Modelling*, 177(3-4), pp. 259–281.

Remm, K., Linder, M., Remm, L., 2009. Relative density of finds for assessing similarity-based maps of orchid occurrence. *Ecological Modelling*, 220(3), pp. 294–309.

Remm, K., Remm, L., 2009. Similarity-based large-scale distribution mapping of orchids. *Biodiversity and Conservation*, 18(6), pp. 1629–1647.

Remm, M., Remm, K., 2008. Case-based estimation of the risk of enterobiasis. *Artificial Intelligence in Medicine*, 43(3), pp. 167–177.

Sokal, R. R., Oden, N., 1978. Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, 10, pp. 199–228.

Tamm, T., Remm, K., 2009. Estimating the parameters of forest inventory using machine learning and the reduction of remote sensing features. *International Journal of Applied Earth Observation and Geoinformation*, 11(4), pp. 290–297.

ACKNOWLEDGEMENTS

The authors are especially grateful to the creator of the system CONSTUD – Kalle Remm – for his helpful comments. We also thank all the people whose field observations were used in this study and Olivia Till from Karula National Park who provided us with additional observations from inventories. The maps and orthophotos were used according to the Estonian Land Board licenses (107, 995, 1350). The research was supported by the Estonian Ministry of Education and Research (SF0180052s07, SF0180049s09) and by the University of Tartu (BF07917).

EXPLORE MULTIVARIABLE SPATIO-TEMPORAL DATA WITH THE TIME WAVE CASE STUDY ON METEOROLOGICAL DATA

Xia Li [a,b,]*, Menno-Jan Kraak [a]

[a] ITC- International Institute for GeoInformation Science and Earth Observation
PO Box 6, 7500 AA Enschede, the Netherlands-@itc.nl
[b] College of Earth Science And Resources, Chang'an University

Commission II

KEY WORDS: Time wave, Temporal exploration, Time space

ABSTRACT:

Traditionally the GIScience community is well able to deal with the locational and attribute component of spatio-temporal data. However, the methods and techniques to deal with the data's temporal component are less developed. This paper introduces a conceptual framework that combines user tasks, available temporal data and visualization theories to discuss temporal visualization. Two limitations of existing method are improved by introducing the time wave environment which is a close combination of temporal graphic representation and temporal interactive tools, and operates in so-called time space. A case study based on meteorological data illustrates the approach.

1. INTRODUCTION

Many of the most important challenges our society is facing today, such as global climate change, economic development and infectious diseases depend on spatio-temporal data to detect and analyze changes as well as trends to support problem solving. Especially the temporal component of the data should be studied carefully to understand the changes and their impacts. Current data collection techniques offer a wide variety of thematic data in many different spatial and temporal resolutions. From a temporal perspective, earth observation techniques provide data with temporal resolutions varying from weeks, days, hours, to even minutes. The challenge faced is how to process, manage, and use these continuous streams of data to support problem solving and decision making. The application of graphic representations in a dynamic and interactive geovisualization environment is part of the solution.

Geovisualization integrates approaches from disciplines including cartography with those from scientific visualization, image analysis, information visualization, exploratory data analysis and GIScience (Dykes, MacEachren et al. 2005). The graphic representations, mostly maps, are used to stimulate (visual) thinking about geospatial patterns, relationships and trends. This is strengthened by looking at the data in a number of alternative ways. Playing with the data using these multiple representations without constraints (traditions) will trigger the mind of the users, and lead to an increase in their knowledge. This does improve our understanding of how to use visualization to get a better insight into spatial data, but not necessarily in temporal data. This is partly due to the fact that most methods and techniques used to solve the geo-problems are from either location or attribute perspective. This paper will look at the geo-problems from a temporal angle, from what is called time space (Li and Kraak 2008). Time space is a visualization space to represent time and answer the temporal questions; for example, what is the temporal distribution of precipitation per month, season or year? How is it linked to location space (e.g. maps) and attributes space (e.g. diagrams).

In the wider context of GIScience research on temporal data analysis and modelling (Goralwalla, Ozsu et al. 1998) and temporal visualization are receiving increased attention (Mackinlay, Robertson et al. 1991; Harrison 1994; Allen 1995; Brown 1998; Harris, Hetzler et al. 2000). This includes papers that offer an overview over existing visualization methods (Andrienko, Andrienko et al. 2003; Aigner, Miksch et al. 2007). Most of these focus on applications in the information visualization field or on discussions of specific aspects of temporal visualization such as static and dynamic representations (Muller 2003), linear time (Silva and Catarci 2000) and the visualization process (Chi 1998). Others take a more overall approach. Aigner et al (Aigner, Miksch et al.) discussed temporal visualization based on a time-oriented framework. Andrienko (Andrienko, Andrienko et al.) focused on spatio-temporal exploration and considers the time visualization from both data and user tasks. However, compared with location space, and attribute space, time space hasn't been studied very well.

In this paper, a proposed solution is to take a closer look at the integration of several visualization theories accepted in GIScience and information visualization, and see how to extend these with a specific temporal component. This is done by analyzing user tasks to structure the temporal data. Hereby the limitations of existing temporal visualization method will be discussed and a specific graphic representation will be suggested. To structure time space the suggested graphic representation should allow the user to represent the temporal data, to display different views on time (e.g. linear or cyclic), and last but not least should allow for interaction. For this the time wave was introduced (Li and Kraak 2008). In the next section, a commonly accepted visual problem solving approach is discussed with specific attention for the nature of the data, the

* Corresponding author.

user task and the visualization environment. After this background information a temporal conceptual framework for visualization is presented, followed by a case study that demonstrates the alternative approach.

2. VISUALIZATION THEORY AND TIME SPACE

2.1 Visualization theory

A common approach to support problem solving with visualization is shown in the scheme in Figure 1. A set of tasks, translated into questions is executed in an appropriate visualization environment with suitable data to solve the problem at hand. Three keywords are of importance: user tasks, data framework and visualization framework. The visualization framework includes the graphic representation and the functional tools to 'play' with the graphics.

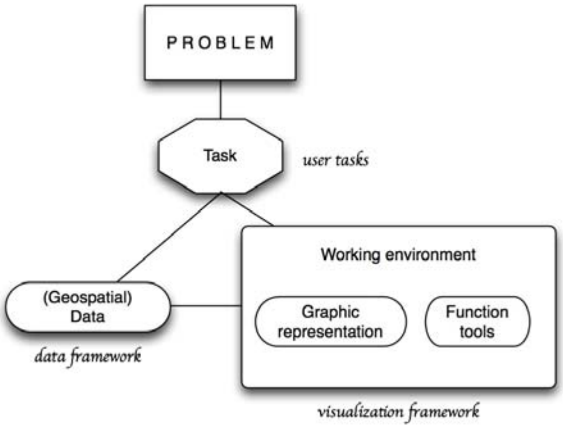


Figure 1. Visual problem solving: The relation between user tasks, a data framework and a visualization framework

Relevant questions are: How to abstract questions based on the data? How to address the data in a visualization environment? How to decide on the right graphic representation(s) and the required functions to answer the questions? Most cartographic and geovisualization theories are studied by looking at the interrelationship of these tightly coupled aspects, such as Bertin (1983), Macrachren (1995), Peuquet (2002) and Andrienko (2006).

2.2 Time space

Time space includes temporal graphic representations and temporal interactive tools. Here the requirements of time space will be discussed from temporal data and temporal user tasks perspectives.

2.2.1 Temporal data and temporal visualization

Compared with "attribute" and "location", time is considered special. Attribute and location can change over time, but the reverse is not possible. The notion of time is difficult to grasp. This has led to many different views on time. Examples relevant in GISciense are found in Frank(1994), Goralwalla, Ozsu et al. (1998) Peuquet (2002) and Aogner, Miksch et al. (2007) . It can be said time passes continuously, but it is observed and measured in discrete points or intervals. This results in a view of continuous time versus discrete time. Time can also be considered as relative (e.g. last week) or absolute (e.g. May 27th). The continuous nature of time makes it linear (before, after) but it obviously holds cyclic characteristics as well (day, season). The granularity of time is defined by the different units in which time is expressed, and can be used to define the scale of time. These granularities inherit from both linear and cyclic characteristics (hours, seasons).

From the temporal data aspect, the views on time such as linear or cyclic, continuous or discrete, single scale or multiple scale, should be considered and addressed in time space properly. In relation to the task each view might have different requirements for temporal visualization. For example, the question 'how did the city expand?' could be answered with discrete multiple views showing a set of snapshots of the city's extent. Alternatively an animation could be used to answer the same question in continues view. The dynamic nature of the animation could also attract attention to other aspects of the expansion. Based on the character of the data, a linear or cyclic representation might by appropriate. Both could express either continuous or discrete time. Although there might be a straightforward solution in the selection process of the visual representation, it could be useful to try other visualizations as well, because alternative view might reveal patterns or aspects of the data that could remain hidden in the straightforward solution.

2.2.2 Temporal user tasks and temporal visualization

Even though the selection of a graphic representation is based upon the nature of the data, the user task plays an important role as well, since at the end, the graphic representation has to answer questions. For instance, the question 'When did countries get their independence?' might require different representations depending on the focus of the question. A timeline with names and years along will do if the temporal distribution is the focus. If one is also interested in the spatial distribution then a world map with labels indicating the year of independence could the most suitable representation. If both are of interest, a space-time-cube (Hägerstrand 1967) could be a possible solution. When one would also be interested in the season of independence the graphic representation should be able to handle both linear and cyclic time and the time wave might be a solution.

Shneiderman's visual information seeking mantra is widely accepted in the visualization field and can be applied in time space. First, the temporal graphic representation is used to locate the time and show the temporal overview, and then temporal interactive tools are required to carry out temporal zooming and temporal filtering options to get the details-on-demand. This process is often an interactive and iterative process, which is supported by a close combination of representation and interactive tools in time space. However, this close combination has not been realized in temporal visualization. There are many temporal graphic representations which could display a temporal pattern of data, such as ThemeRiver (Havre, Hetzler et al. 2000), Stacked bar chat (Harris 1999), People garden (Xiong and Donath 1999), MultiCombs (Tominski, Abello et al. 2003) and etc. To assist answering the questions, temporal interactive tools should allow one to identify single or multiple points in time, identify points at periodic intervals, or define intervals of certain length in both the linear and the cyclic format. Some have experience with this approach, like Koike et al. (1997) with TimeSlider and Edsall et al (1997) experimented with a time wheel query device in the TEMPEST system. Above temporal visualization methods are either a temporal representation or a temporal interactive tool only.

3. TEMPORAL VISUALIZATION AND TIME WAVE: AS A NEW APPROACH IN TIME SPACE

Most temporal representation are either controlled by a timeline (linear time) (Mackinlay, Robertson et al. 1991; Plaisant, Milash et al. 1996; Silva and Catarci 2000; Wijk 2002) or a time wheel (cyclic time) (Mackinlay, Robertson et al. 1994; Carlis and Konstan 1998; Harris 1999; Daassi, Fauvet et al. 2002). However, this does not always result in satisfying solutions because many phenomena have both linear and cyclic characteristics. The time wave (Figure 2; Li and Kraak 2008) is one potential solution for representing both the linear and cyclic nature of data. It is a combination of the timeline and the time wheel and offers an alternative view by its own. With the timeline and the time wheel it is difficult to show multiple time scales. However, the time wave can show multiple time scales by nesting different waves with different wavelengths and amplitudes based on their temporal scale.

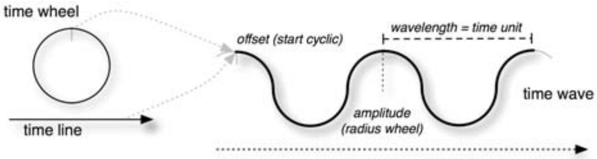


Figure 2. The time wave as a construct of the time line and the time wheel combining the linear and cyclic nature of time (Li and Kraak 2008)

Time wave not only could show the linear and cyclic time temporal pattern of data as a temporal representation, but also carry out zooming and filter as an interactive tools. For temporal interactive tools, the time wave can do both linear and cyclic operation as well. In addition, the interactive time wave also support one to represent some locational and attribute characteristics in relation with the phenomenon’s temporal distribution which can be explored.

Furthermore, the time wave allows users to interact with the temporal reference or temporal representation. Supported by Coordinated Multiple View (CMV) (Roberts 2005; Roberts 2008) principle, one could identify or select any time instant or interval at various scale, or select a temporal pattern, to show the corresponding information in the location space or attribute space. Figure 3 shows an interpretation of time space with the time wave combining the representation and manipulation functions to carry out the interactive and iterative exploration process. The figure also demonstrates the link between time space, location space and attribute space. With a temporal question, one starts in time space and depending on the nature of the question, the answer can be given in time space itself or one ‘jumps’ out of time space into location or attribute space. For instance, the question related to year of independence of the countries is an example where one moves to location space when one needs to see spatial patterns next to temporal patterns. In answering complex question one might have to jump from space to space in an extensive iterative process to identify and compare temporal, spatial and attributer patterns.

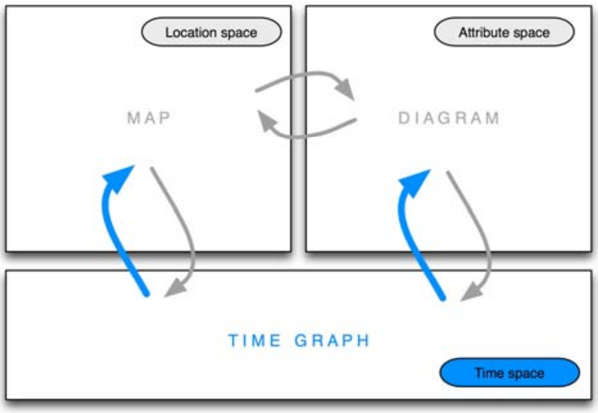


Figure 3. The interactive options to move from time space into location and attribute space supported by CMV principle

4. CASE STUDY: TIME SPACE AND METEOROLOGICAL DATA

To see how to create a proper time graph based on the analysis of temporal data, user tasks, a case study is discussed in this section. The data set contains observations of nineteen meteorological stations in Beijing, China. For each station, known at location (x, y, z), the temperature and the dominant land use information are given. The temperature was measured at minute interval over the months in July 2007 and July 2008. A data overview is shown in Figure 4. The location data is point data in context of geographical units (districts in Beijing city); the attribute data is quantitative (temperature) and qualitative (land use); the temporal data has linear characteristics (measurements over a month) and multi-scale cyclic characteristics (July for two year and days and nights for 24 hour periods).

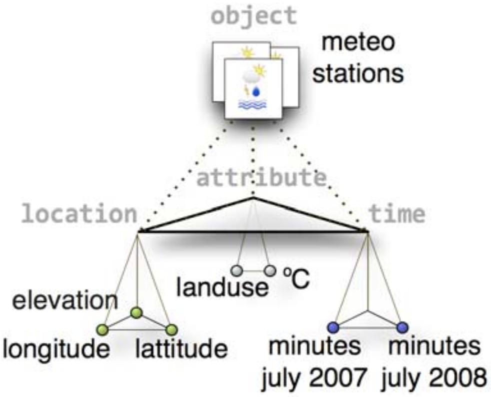


Figure 4. Case study data: the characteristics of the meteorological stations in the Beijing area.

The user task is to find out when the temperature reaches its daily maximum value at the different meteorological stations. Is the pattern related to the longitude, latitude and elevation of each meteorological station? What are the other impact factors, such as land use? Is there any difference between those two years? The temporal questions, such as when, how often, in what order, synchronization and trends, always have a relation to other data components. For instance, from object perspective (station name, station ID), location perspective (height; east/west; north/south), attribute perspective (low, high, or average temperature; land use category like urban, vegetation, water, or other).

Based on figure 4, figure 5 gives a schematic overview of the coordinated multiple view environment in which time space plays a key role. The time wave is selected as graphic representation since it can handle the temporal data requirements of the case study: linear and cyclic multi scale time. The wave length could be months, weeks, days or any other time unit one would need. It also allows interaction and is directly linked to location and attribute space which is needed if one has to switch spaces while executing the user tasks. In Figure 5 this is represented by the in and out arrows. Location space show the map with the position of the meteorological stations and attribute space shows a scatterplot of the station's temperature versus land use. Figure 6 gives a snapshot of the actual software. The time wave is a plug-in for the Udig open source GIS. The time wave can also include a visual representation of the data's attribute and location characteristics. In Figure 5 symbols related to the meteorological stations, the temperatures observed, as well as land use could be plotted on the wave. Figures 7 and 8 give examples related to the case study.

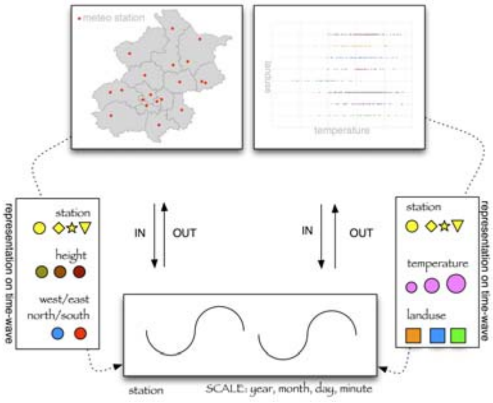


Figure 5. The working environment of the time wave, with options to visually represent characteristics from location and attribute space on the wave

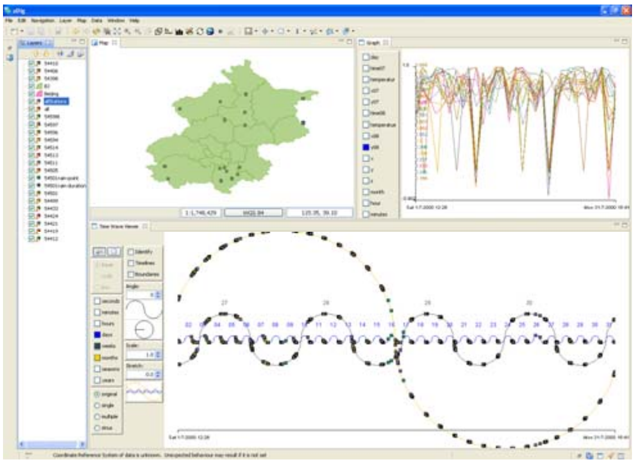


Figure 6. Snapshot of the time space/time wave environment as implemented in the open source GIS uDig

Figure 7 gives two examples of how time space works in practice. In the upper section of the Figure an overview of the month July 2007 is given. The wave length represents a day. For each day and each meteorological station, a symbol is located on the wave at the time when the highest temperature was observed. It is obvious that a wave like this will only offer and overview. A first glance at the data shows that the highest temperature is reached a few hours after 12 noon (the top of the wave). However, both at the beginning and the end of the month an anomaly in the temporal pattern can be observed. These two anomalies have been further explored in the Figure. In both cases one can first zoom in the wave to see more details. At the beginning of the month (the left of the Figure) two stations are rather late in reaching their maximum temperature and at the end of the month (the right of the Figure) there is a single day on which all stations are late in reaching their maximum temperature. In the first situation it makes sense to get more details on those two particular stations, and switch to location space to see if their location is special. The map doesn't reveal anything special, since stations with similar geographic condition can be found. For the second situation it was decided to switch to attribute space and have a closer look at the temperature ranges of all stations at that particular day and the days before and after. It can be observed that the maximum temperature for that particular day was also considerably lower than the days before and after. The available data do not give a good explanation. Other meteorological parameters, for example cloud cover, may be introduced to give an explanation. The experts will have to continue exploring the data and perhaps obtain additional information in order to be able to explain the anomalies.

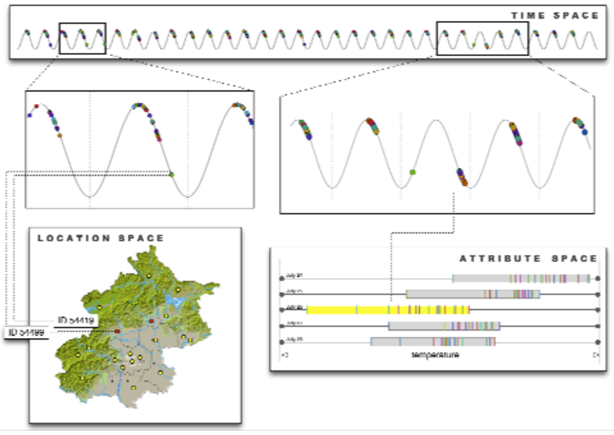


Figure 7. Working with the time wave: identify and explain temporal patterns. At the top the overview of July 2007 showing all stations at the moment these observe the highest temperature. In the middle zooming in on two clear anomalies in the monthly temporal pattern. At the lower left a map in location space and the lower right a temperature diagram in attribute space.

Figure 8 shows another useful function of the time wave. The top wave is the normal time wave, here showing again the moment that the daily temperature reaches its maximum at each meteorological station for three days. It is possible to create a set of parallel waves (Figure 8b). For each meteorological station a separate wave is created to obtain a better view on the patterns. This approach has been derived from the parallel coordinate plot. The wave in Figure 8c shows the data sorted based on the values of July 1st. This result in having the station with the earlier highest temperature in the lower wave. Keeping this sort order the following days show a different pattern. It means the order of each stations reach its maximum temperature varies over different day. It is possible to sort the data based on any other available variables. For instance, one could sort based on the height of the stations (Figure 8d), or sort from east to west or north to south. After sorting, one might observe patterns that will stimulate one to jump to location or attribute space, to look for more detail, to switch back to overview mode or even to retrieve more data. Working with the time wave to explore temporal patterns is clearly an iterative process. In the next section the specific advantages and possible limits of the approach will be discussed.

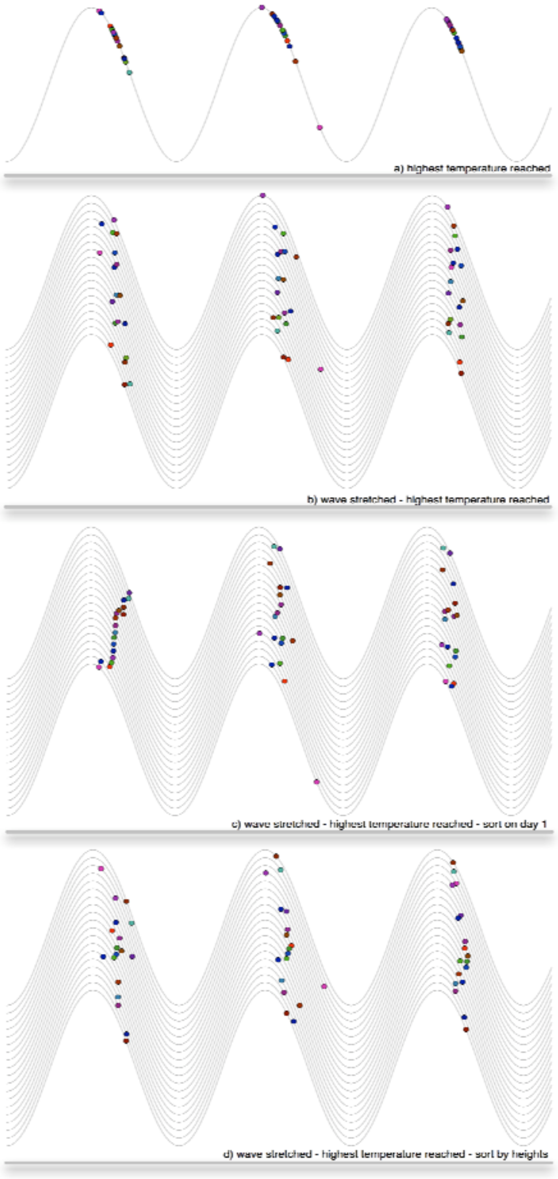


Figure 8. Working with the time wave: comparing temporal patterns. a) an overview of the first three days of July 2008 showing when the maximum temperature was observed for each station; b) temporal stretch by creating a set of parallel waves to un-clutter the pattern as seen in the upper wave; c) sorted wave based on the temperature values of July 1, 2008; d) sorted by station's elevation (highest elevation in lowest parallel wave).

5. CONCLUSIONS AND DISCUSSIONS

In this paper an alternative approach to work with spatio-temporal data has been presented based on a visual problem solving approach. The nature of the temporal data and the user task at hand are the driving forces to select a suitable graphic representation. In time space the graphic representation focuses on the temporal aspect of the data. Here the time wave is introduced. It not only combines linear and cyclic time, but also combines temporal data representations and interaction, and allows a limited representation of attribute and location data. The exploratory activities are guided by the visual information seeking mantra of Shneiderman. Depending on the user task, one can switch from one space to the other, and into any of the information seeking modes (overview, zoom / filter, and details on demand).

A case study based on the data observed at meteorological station in Beijing has demonstrated the time space framework and the capabilities of the time wave. Temporal patterns and distribution are studied from time space switching to location space (maps) and attribute space (diagrams) when required. The time wave is a good example of how an alternative view on the data might reveal patterns not always obvious from traditional graphic representations. In overview mode it is a very suitable visualization to provide an impression of the nature in data, being linear or cyclic. In this mode the wave easily reveals anomalies as well, as for instance in figure 7. This effect can be strengthened by moving the horizontal line (x-axis) vertically through the wave.

It is not argued that the time wave is the only visual representation in time space. Graphics based on the time line or time wheel only, would be preferred in certain situations. This very much depends on the nature of the data and the task at hand. However, it is claimed that the optimal solution space is time space interactively linked with location and attribute space. This allows the user to flexibly tackle the problem from many different perspectives.

Future work will deal with a detailed validation of the time wave in other case studies and its relation to graphic representation in location and attribute space. Known data sets will be used to see if and how the time wave might discover known patterns and possibly reveal new temporal patterns as well, based on tasks executed by users. In relation to the task space (see figure 3) further work is ongoing in which existing temporal visualization methods are be analyzed on their strength and weakness in the context of the actual temporal problem of users. If successful this could lead to a kind of advisory system that assist the user to select a suitable graphic representation in time space dedicated to the user task one is dealing with.

ACKNOWLEDGEMENTS

This research Supported by the Programme of Introducing Talents of Discipline to Universities, "111 project" , No: B08039

REFERENCE

Aigner, W., S. Miksch, et al., 2007. Visualization Time-oriented Data--- A Systematic View. *Computers and Graphics*, **31**, pp.401-409.

Allen, R. B., 1995. Interactive Timelines as Information System Interfaces. *Symposium Libraties*, Japan.

Andrienko, G. and N. Andrienko, 2006. *Exploratory Analysis of Spatiall and Temporal Data*. Springer, Berlin, Germany.

Andrienko, G., N. Andrienko, et al., 2003. Interactive maps for visual exploration of grid and vector geodata. *ISPRS Journal of Photogrammetry and Remote Sensing*, **57**(5-6), pp.380-389.

Andrienko, N., G. Andrienko, et al., 2003. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, **14**(6), pp.503-541.

Bertin, J., 1983. *Semiology of graphics : diagrams, networks, maps*. University of Wisconsin Press, Madison.

Brown, I. M., 1998. A 3D Interface for Visualization of Serial Periodic Data. *ACM GIS'98*, Washington DC USA.

Carlis, J. V. and J. A. Konstan, 1998. Interactive Visualization of Serial Periodic Data. *ACM UIST'98*, San Francisco.

Chi, E. H., et al., 1998. Visualizing the Evolution of Web Ecologies. *ACM CHI'98*, CA USA.

Daassi, C., M.-C. Fauvet, et al., 2002. Multiple Visual Representation of Temporal Data. *Proceedings of the 13th International Conference on Database and Expert Systems Applications* Springer-Verlag.

Dykes, J., A. M. MacEachren, et al., Eds. 2005. Exploring Geovisualization. Elsevier Science & Technology Books

Edsall, R., M.-J. Kraak, et al., 1997. Assessing the Effectiveness of Temporal Legends in Environmental Visualization *GIS/LIS '97 Annual Conference and Exposition* Cincinnati, Ohio

Frank, A., 1994. Different Types of "Times" in GIS. *GIS and Computational Science Perspectives.*

Goralwalla, I. A., M. T. Ozsu, et al., 1998. An Object-Oriented Framework for Temporal Data Models. *Temporal Database-Research and Practice*. O. Etzion, S. Jajodia and S. Sripada. Berlin Heidenberg, Springer. **1:** pp.1-35.

Hägerstrand, T., 1967. *Spatial Process*. University of Chicago, Chicago.

Harris, R. L., 1999. *Information graphics: a comprehensive illustrated reference* Oxford University Press US New York

Harris, S., B. Hetzler, et al., 2000. ThemeRiver: Visualization Theme Changes Over Time. *IEEE Symposium on Information Visualization*, USA, IEEE Computer Society.

Harrison, B. L., R.Owen, R.M.Baecker, 1994. Timelines: An Interactive System for the Collection of Visualization of Temporal Data. *Graphics Interface' 1994*, Canadian.

Havre, S., B. Hetzler, et al., 2000. ThemeRiver: Visualizing theme changes over time. *IEEE Symposium on Information Visualization*, Los Alamitos, USA, IEEE Computer Society.

Koike, Y., A. Sugiura, et al., 1997. TimeSlider: An Interface to Specify Time Point *Proceedings of UIST '97.*

Li, X. and M.-J. Kraak, 2008. The Time Wave. A New Method of Visual Exploration of Geo-data in Time–space. *The Cartographic Journal*, **45**(3), pp.1-9.

MacEachren, A. M., 1995. *How Maps Work: Representation, Visualization, and Design*. Guilford press, New York, USA.

Mackinlay, J., G. Robertson, et al., 1994. Developing calendar visualizers for the information visualizer. *Proceedings of the 1990 ACM conference on Computer-supported cooperative work,* Los Angeles,.

Mackinlay, J. D., G. G. Robertson, et al., 1991. The Perspective Wall: Detail and Context Smoothly Integrated. *ACM CHI'91*, New York.

Muller, W., Hetdrun Schumann, 2003. Visualization Methods for Time-dependent data-Overview. *Winter simulation conference 2003*, New Orleans USA.

Peuquet, D. J., 2002. *Representations of Space and Time*. Guilford, New York.

Plaisant, C., B. Milash, et al., 1996. LifeLines: visualizing personal histories. *Conference on Human Factors in Computing Systems* Vancouver, Canada ACM

Roberts, J. C., 2005. Exploratory Visualization with Multiple Linked Views. *Exploring Geovisualization*. J. Dykes, M. A.M. and M.J.Kraak. London, Elsevier. **8:** pp.159-180.

Roberts, J. C., 2008. Coordinated Multiple Views for Exploratory GeoVisualization. *Geographic Visualization: Concepts, Tools and Applications*. M. Dodge, M. Mcderby and M. Turner. Chichester, John Wiley & Sons Inc. **3:** pp.25-48.

Silva, S. F. and T. Catarci, 2000. Visualization of Linear Time-oriented data: a survey. *Web Information Systems Engineering, 2000*.

Tominski, C., J. Abello, et al., 2003. Interactive Poster: Axes-Based Visualizations for Time Series Data. *Poster Compendium of IEEE Symposium on Information Visualization*, IEEE.

Wijk, J. v., 2002. Image based flow visualization. *SIGGRAPH*, San Antonio, USA.

Xiong, R. and J. Donath, 1999. PeopleGarden: Creating Data Portraits for Users. *12th Annual ACM Symposium on User Interface Software and Technology (UIST '99)*, NC; USA.

RESEARCH OF SPATIAL AND TEMPORAL VARIATIONS OF WETLAND IN PEARL RIVER ESTUARY (1978 ~ 2005)

GAO Yi[a,b,c,d], SU Fenzhen[a], SUN Xiaoyu[a], XUE Zhenshan[a],He Yawen[b]

([a.]Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; [b].Yantai Institute of Coastal Zone Research for Sustainable Development, CAS,Yantai, 264003, China; [c] Graduate School of Chinese Academy of Sciences, Beijing 10039, China; [d] South China Sea Institute of Oceanology, CAS, Guangzhou 510301,China)

KEY WORDS: Pearl River Estuary, Coastal Wetlands, Change Range，Degree of Dynamic Change, Centroid displacement

ABSTRACT: During the period of 1978~2005, substance changes have taken place of coastal wetland in Pearl River Estuary as the result of interaction of human and natural factors. To investigate the spatial and temporal variations of coastal wetland in Pearl River estuary, Landsat-MSS images in 1978 and Landsat-TM images of 1995 & 2005 year were processed, and wetland database of the three periods was established based on these remote sensing images. The result showed that: (1) area of natural wetlands decreased while the area of construct wetland increased distinctively; (2) in the period of 1978~1995, coastal wetland area decreased by 3.3%, while it decreased by 18.3% in period of 1995~2005; (3) Among 6 administrative districts in the study area, the wetland dynamic degree can be sorted in a descending order as: Zhuhai, Shenzhen, Macao, Dongguan,Zhongshan and Guangzhou; (4) the centroids of wetland in Pearl River Estuary coastal zone moved to north by a linear distance of 0.6 km during 1978~2005, and move to south east during 1995~2005 by a linear distance of 6.1 km; (5) according to the changes of coastlines in Pear River Estuary, area of sea reclamation during 1978~2005 along coastline inland in Zhuhai, Macao, Zhongshan, Guangzhou Dongguan, Shenzhen are 12439.29，502.60，2946.45，5372.21，1815.96，6317.88 hectares respectively.

1. INTRODUCTION

Wetlands which are honoured as "kidney of the globe", "species gene pool" and “the cradle of humanity”(SUN,2000). They are the most productive ecosystems in the world compared to terrestrial ecosystems and marine ecosystems. They provide important ecosystem functions such as nursery habitats for fish and crustaceans, resting and feeding area for migratory birds, they also support biodiversity, filter containments, dissipate water energy, and offer intrinsic values such as aesthetics and education.(GOODWIN,2001) Coastal wetland is an important kind of wetland and natural landscape with abundant natural resources and unique environmental effect (AN,2007), but about 30% to 50% of the area of earth's major coastal environments has been degraded during past decades(VALIELA,2008). It become a common sense that coastal wetland is the most vulnerable resource which affected by the sea level rise (SLR), so it is of great importance to learn the changes of coastal wetland. Wetland is one of the most important natural resources of Pearl River delta. During the period of 1978~2005, substance changes have taken place in the coastal wetland of Pearl River Estuary as the result of natural and anthropogenic factors. Many researchers choose one part of Pearl River delta for research (LIU,2005, LI,2006, WANG,2007). However, few studies have focused on the spatial and temporal changes of coastal wetland in Pearl River Estuary. Based on GIS and RS, the spatial and temporal changes of coastal wetland in Pearl River Estuary in the past three decade are researched.

2. STUDY AREA

The study area, Pearl River Estuary coastal zone covers an inland area of 5 km buffers from the costal line, and extends to the -6m depth line in shallow sea, and the area is about 390,000 hectares. It is located at the middle south of Gongdong Province in China (Figure 1.). The bell-shaped Pearl River Estuary receives and carries most of the outflow from the Pearl River, eventually flow into the South China Sea. The climate here is summer and winter monsoon alternating, year-round high temperatures. Since economic liberalization was adopted in the late 1970s, the area has become one of the leading economic regions and a major manufacturing center of China. Including two Special Economic Zone: Zhuhai and Shenzhen; an Open Coastal City: Guangzhou.

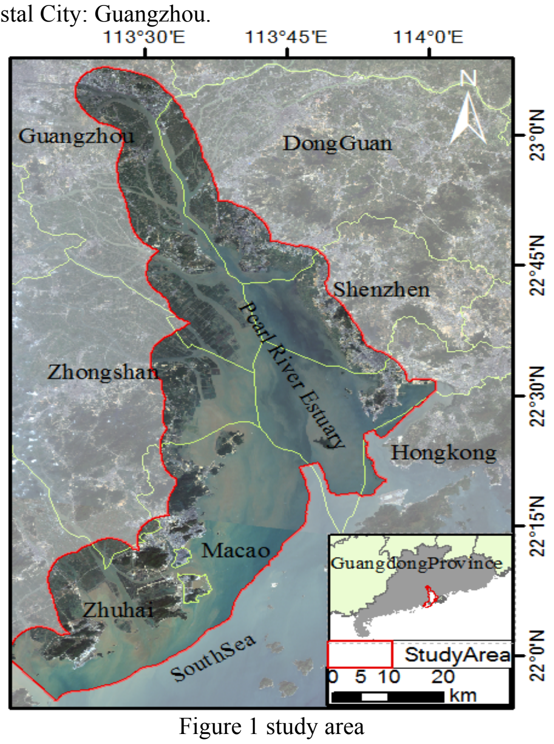


Figure 1 study area

3. DATA and METHODS

3.1 Data

In order to study the spatial and temporal changes of Pearl River Estuary in the past three decades, three periods data sources are collected, including Landsat MSS images in 1978; Landsat TM in 1995 and 2005(the remote sensing images were downloaded from the USGS website: http://glovis.usgs.gov). Each scene Landsat MSS image has 4 bands with 78m spatial resolution; and each scene Landsat TM images has 7 bands with 30m spatial resolution.

3.1.1 Data Processing

Remote sensing images processing, including pre-processing and false colour composite, and this was carried out by means of PCI Geomatica10.0. To deal with the unsystematically geometric errors, geometric rectification is needed, in this research all the remote sensing images were rectified to WGS-84 coordinate system, and the spatial rectify error is less than 1.5 pixels. Because of the uncertainty of the supervised or unsupervised classification, the wetland data of the year 1978, 1995 and 2005 were extracted by visual interpretation based on these geometric rectified remote sensing images. To ensure the interpretation accuracy, I carried on a field investigation with my research group in 2006 January, collected 135 sample sites , and record the position by GPS device, and interpretation symbol library was set up by the collected wetland type and position point according to landsat TM image of 2005. A field-prove work was carried out in 2008 January, the interpreted precision of 2005 was higher than 93%, and 1995 and 1978 was 90% and 88% respectively. The data processing packages mainly included ArcGIS9.0, SPSS 16.0 and PCI Geomatica10.0.

3.1.2 Wetland classification system:

According to the ecological classification principles and the classification of China and abroad (COWARDIN,1979, SADER,1995, SCOTT,1995, YANG,2002, SCHMIDT,2003), wetland in Pearl River Estuary is divided into natural wetland and constructed wetland according to the interfering degree of human activities. Natural wetland also divided into mud flat, mangrove, lake, river and shallow sea; and constructed wetland is divided into reservoir, aquaculture water, paddy field and pond.

3.2 Methods

In order to research the changes of wetland in the Pearl River Estuary, land use changes methods, land use transition matrix and land use dynamic degree model is applied.

3.2.1 Spatial and temporal dynamic of wetland:

To determine the change rate of coastal wetland in Pearl River Estuary change, the study period 1978~2005 is divided into two sub-periods and the wetland changes of the two sub-periods are compared. The period of 1978~1995 calls earlier stage, and the period of 1995~2005 calls later stage. In order to understand the change rate of each wetland type, wetland dynamic degree is calculated by the rate of land use change model as follows(SHENG-HE,2002):

$$S =(A_i - UA_i) / A_i (T_2 - T_1) \times 100\% \qquad (1)$$

Where S is the rate of the ith type wetland during the monitoring period T_1 to T_2;A_i is the area of the ith type wetland at the beginning of the monitoring period; and UA_i is the area of the ith type wetland that remains unchanged during the monitoring period. (A_i- UA_i) is the changed wetland area during the period, *i.e.* the total area of the ith type wetland converted into the other types of wetland or non-wetland

In order to understand the rate of regional wetland changes and their characteristics differences, the wetland dynamic degree was calculated by administrative districts in Pearl River Estuary. Regional difference in wetland change characteristic can be determined by using the land use dynamic degree model that could be mathematically expressed by the following relationship(LIU,2000):

$$S = \sum_{ij}^{n} (\Delta S_{i-j} / S_i) \times (1/t) \times 100\% \qquad (2)$$

Where S is the wetland dynamic degree over time t; S_i is the ith type wetland area at the beginning of the monitoring period, n is the number of the wetland types, and □S_{i-j} is the total area of the ith type wetland that is converted in to others (other wetland types or non-wetland).

3.2.2 Spatial changes of wetland

Spatial changes of wetland can be described by the centroid of land use resource distribution(JUNHONG,2008). The cenctrids of wetland distributions in 3 periods can be calculated as follows:

$$X_t = \sum_{i=1}^{n} (C_{ti} \times X_i) / \sum_{i=1}^{n} C_{ti} \qquad (3)$$

$$Y_t = \sum_{i=1}^{n} (C_{ti} \times Y_i) / \sum_{i=1}^{n} C_{ti} \qquad (4)$$

Where X_t and Y_t are the abscissa and ordinate centroid of wetland distribution in t period respectively. X_i and Y_i are the abscissa and ordinate of the centroid of wetland type i in the same period; C_{ti} denotes the area of each wetland patch; $\sum_{i=1}^{n} C_{ti}$ denotes the total wetland area in t period.

3.2.3 Coastline changes of Pearl River Estuary

In the past three decades, due to natural and man-made factors, especially the role of man-made factors, significant changes of coastlines in Pearl River Estuary have taken place. There area many successful studies to monitor coastline changes by using multi-temporal satellite image (HE,2006, ALESHEIKH,2007, SESLI,2009). In this study coastlines of 1978 and 2005 in Pear River Estuary were both extracted from satellite images by using PCI Geomatica10.0. In order to keep same spatial scale, Landsat-MSS images was sampled into 30-meter pixel size. Since boundaries of land and water are clear on band7 of MSS. And in the band TM5, water reflection rate is almost zero, so by take these advantages, coastline can be extracted. Water boundaries of 1978 and 2005 were extracted by unsupervised via ISODATA method, and then processed on ArcGIS9.0.

4. RESULT

4.1 Wetland distribution and dynamic degree

The spatial-temporal changes of wetland in the three periods are showing in Figure.2 and table1. It can be concluded that the wetland coverage rate in 1978, 1995 and 2005 is 81.6%, 78.9% and 64.5% respectively. All the spatial data and is processed on the platform of ArcGIS9.0, and the area data is calculated by SPSS16.0. As table1 shows, among the earlier and later stages, the area of mud flat, mangrove, shallow sea and paddy field wetland types continue to decrease, while the area wetland types of pond and aquaculture water continued to increase. Total in all, the wetland lost during the later stages was five times more than that lost during the earlier stage. Wetland changes and conversation during the two stages are difference from each other, which can be reflected by table1, table2 and table 3.

According to the spatial and temporal changes of coastal wetland in Pearl River Estuary, wetland dynamic degree of each wetland type in the earlier and later stages can be calculated by equation (1), dynamic degree of each wetland type in the earlier and later stages as figure 3 shows.

The conversion of natural wetland and constructed wetland reflects that it undergone a large-scale movement of enclosing tideland for cultivation in the earlier stage, but it slowed down in the later stage; And the dynamic degree of all wetland types (except for shallow sea and reservoir) in the later stage is higher than the earlier stage.

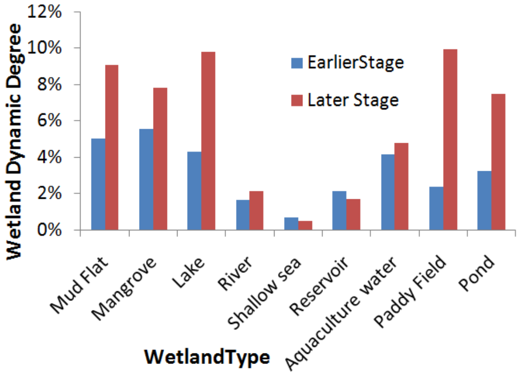


Figure 3. Dynamic degree of each wetland types in both earlier and later stages.

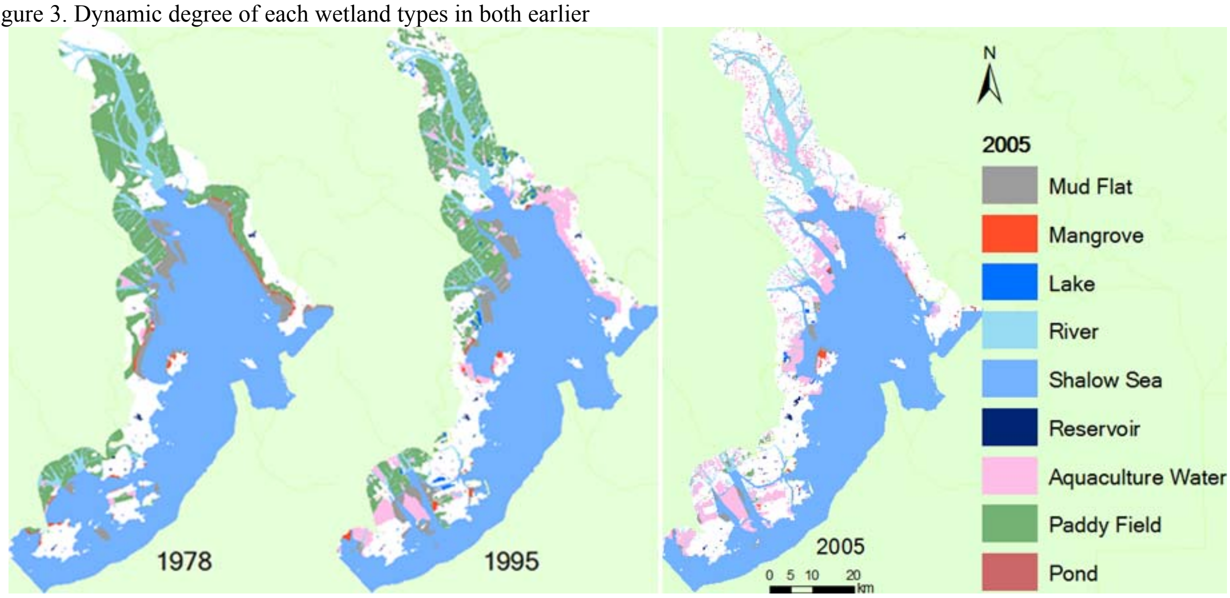


Figure2. Distribution of wetland in the periods of 1978, 1995 and 2005

Table1. composition of wetland in different periods and wetland changes in the two stages. (Units: ha)

	1978		1995		2005		Wetland change	
	Area (ha)	Percent (%)	Area (ha)	Percent (%)	Area (ha)	Percent (%)	1978~1995 (ha)	1995~2005 (ha)
Mud Flat	15300.19	4.79	11961.30	3.87	4851.57	1.92	-3338.89	-7109.73
Mangrove	2675.19	0.84	1120.03	0.36	580.34	0.23	-1555.16	-539.68
Lake	201.54	0.06	2563.96	0.83	740.82	0.29	2362.41	-1823.14
River	22851.44	7.15	23110.12	7.48	22391.22	8.87	258.68	-718.89
Shallow Sea	210860.72	65.99	192339.57	62.26	184808.25	73.22	-18521.14	-7531.32
Reservoir	394.62	0.12	410.27	0.13	663.20	0.26	15.64	252.94
Aquaculture Water	3954.71	1.24	23686.43	7.67	35930.89	14.24	19731.71	12244.46
Paddy Field	63181.82	19.77	53069.66	17.18	809.16	0.32	-10112.16	-52260.50
Pond	92.72	0.03	684.67	0.22	1617.04	0.64	591.95	932.37
Total	319512.95	100.00	308946.01	100.00	252392.50	100.00	-10566.95	-56553.51

Table2. The transition matrix of wetland types (units: ha): 1978~1995 in the study area. Value in bold: wetland types without change during 1978~1995.

		1995										
	Wetland type	Mud Flat	Mangrove	Lake	River	Shallow Sea	Reservoir	Aquaculture water	Paddy Field	Pond	Total	Non-wetland
1978	Mud Flat	**2224.95**	264.15	189.43	524.65	5520.95	0	2735.43	3008.70	35.01	14503.28	796.91
	Mangrove	52.20	**155.83**	53.06	30.98	92.07	0	1512.99	152.48	9.67	2059.27	615.92
	Lake	0	0	**53.68**	0	0	1.90	0	0.12	0	55.70	145.84
	River	205.35	0	82.05	**16434.70**	36.73	0	821.19	3590.26	16.21	21186.50	1664.94
	Shallow sea	9294.74	651.48	482.29	844.01	**186086.36**	0	6752.22	1573.10	255.10	205939.29	4921.43

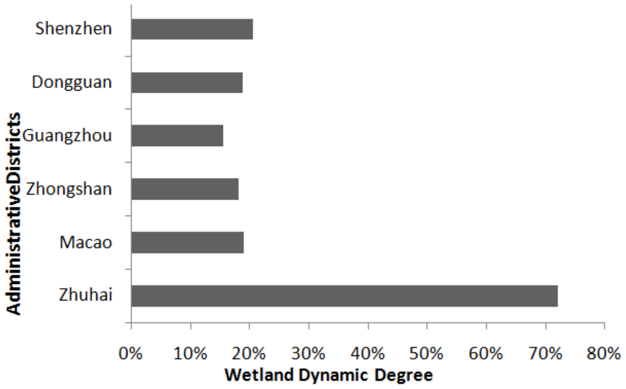
Reservoir	0	0	12.54	0	0	**252.72**	0	0	0	265.26	129.36
Aquaculture water	2.54	0.14	33.46	327.84	9.48	0	**1162.10**	1848.72	4.24	3388.51	566.20
Paddy Field	41.45	13.17	798.40	4218.32	40.52	0	9065.04	**37487.22**	126.1	51790.22	11391.60
Pond	0	0	8.02	0	0.17	0	0	0	**41.24**	49.43	43.29
Total	11821.23	1084.78	1712.9	22380.5	191786.27	254.62	22048.98	47660.59	487.6	299237.5	20275.49

Table 3 the transition matrix of wetland types (units: ha): 1995~2005 in the study area. Value in bold: wetland types without change from 1995~2005.

		2005										
	Wetland type	Mud Flat	Mangrove	Lake	River	Shallow Sea	Reservoir	Aquaculture water	Paddy Field	Pond	Total	Non-wetland
1995	Mud Flat	**1096.43**	90.06	174.47	564.53	1181.27	0	5290.14	37.92	46.24	8481.07	3480.23
	Mangrove	85.98	**243.63**	3.84	31.17	47.78	0	593.92	0	15.92	1022.26	97.77
	Lake	13.35	0	**47.71**	62.55	3.96	31.00	988.02	68.87	36.74	1252.21	1311.75
	River	539.29	1.46	0.25	**18214.13**	72.38	0	1060.54	5.94	40.90	19934.90	3175.22
	Shallow sea	1931.81	163.10	28.09	550.12	**182411.77**	0	3702.48	0.36	283.18	189070.90	3268.67
	Reservoir	0	0	6.08	0	0	**340.31**	0	0	0	346.39	63.88
	Aquaculture water	437.78	42.05	36.03	722.27	413.96	0	**12306.40**	0	151.03	14109.52	9576.91
	Paddy Field	345.95	0.01	10.64	1599.95	129.29	0.12	10015.54	**375.76**	317.29	12794.55	40275.11
	Pond	70.20	7.86	12.32	4.12	6.13	0.41	120.16	0	**171.93**	393.14	291.53
	Total	4520.78	548.17	319.44	21748.84	184266.54	371.85	34077.21	488.86	1063.25	247404.93	61541.08

4.2 Regional difference

Coastal wetlands in Pearl River Estuary are divided into six regions: Zhuhai, Macao, Zhongshan, Guangzhou, Dongguan and Shenzhen. The dynamic degree during the period of 1978~2005 in each administrative district as figure 4 shows. It is obviously that the dynamic degree of coastal wetland in the district of Zhuhai is highest, while it is lowest in Guangzhou district.



Figture4. Dynamic degree of coastal wetland in each administrative district in the period of 1978~2005.

4.3 Spatial changes of wetland

Spatial and temporal changes of coastal wetland in the two stages are processed by ArcGIS9.0, the results as figure5 shows. Compared by the spatial changes of wetland in the two stages, there was more wetland changed into other wetland types in the early stage than that in the later stage while there was more wetland changed into non-wetland in the later stage than that in the early stage. The wetland changes mainly along the coastline in the early stage while changes became widely distribute inland in the late stage.

The centroids of each wetland patch were extracted from wetland vector data based on ArcGIS9.0, and the centroids coordinates of each periods were calculated by equation (3) and (4). The centroids of wetland in the three different periods as figure6 show. The cenrtoids of coastal wetland moved 0.005° to the north direction and 0.001° to the east direction from 1978~1995, by a linear distance of 0.6km, which was related to the large-scale movement of enclosing tideland for cultivation. However, during 1995~2005, the wetland centroids moved 0.055° to the south direction and 0.01° to the east direction, by a linear distance of 6.1km, which was mainly because of inland wetland changed into non-wetland.

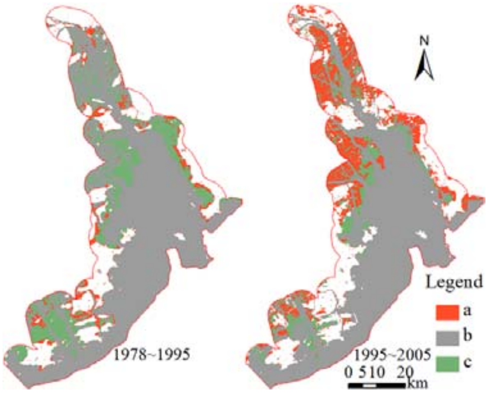


Figure5. Spatial changes of wetland in the two stages. a: wetland that changed into non-wetland; b: wetland that unchanged; c: wetland that changes into other types of wetland.

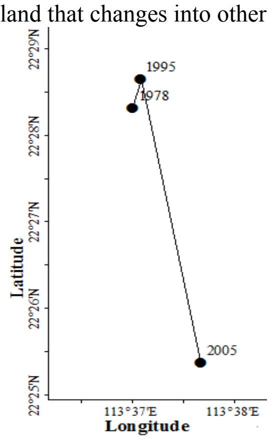


Figure6. Spatial changes of coastal wetland in Pearl River Estuary.

4.4 Coastline changes

The coastlines of Pearl River Estuary in 1978 and 2005 as figure7 show. The sea reclamation can be processed by ArcGIS9.0, and the area of sea reclamation (1978~2005) along the inland coastline in Zhuhai, Macao, Zhongshan, Guangzhou, Dongguan and Shenzhen are 12439.29，502.60，2946.45，5372.21，1815.96，6317.88 hectares respectively.

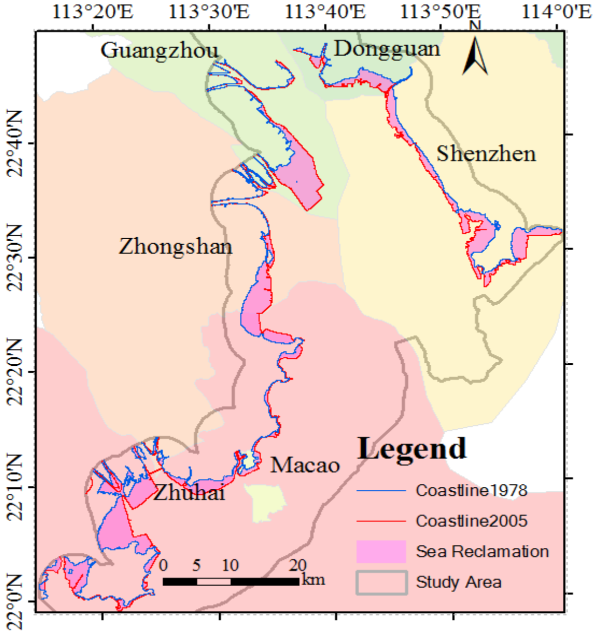


Figure7. Changes of coastlines inland and the distribution of sea reclamation of 1978~2005 in Pearl River Estuary.

5. DISCUTION

After the analysis of coastal wetland spatial and temporal change in the nearly three decades, and the comparison between early and late stages, we can see that coastal wetlands in Pearl River Estuary have been significantly changed. This change is mainly caused by wetlands changed into non-wetland (such as construction site or factories). These changes reflect an accelerating depletion of wetland. Coastal wetland depletion disturbed the ecological functions of wetland, which will cause a series of environmental problems, such as flood, coastal erosion, loss of habitat for aquatic species and so on. Wetland protection measures and long-term planning for the study area is an urgent task.

References

Sun, G.,2000. Development and prospect of wetlands science in China. Advance in Earth Science, **15**(6): p. 666-672.

Goodwin, P., A. Mehta, J. Zedler,2001. Coastal wetland restoration: An introduction. Journal of Coastal Research, **27**: p. 1-6.

An, S., H. Li, B. Guan, et al.,2007. China's natural wetlands: past problems, current status, and future challenges. AMBIO: A Journal of the Human Environment, **36**(4): p. 335-342.

Valiela, I.,S. Fox,2008. ECOLOGY: Managing Coastal Wetlands. Science, **319**(5861): p. 290.

Liu, K., X. Li, S. Wang,2005. Monitoring of the changes of mangrove wetland around the Zhujiang Estuary in the past two decades by remote sensing. ACTA SCIENTIARUM NATURALIUM UNIVERSITATIS SUNYATSENI, **25**(2).

Li, X., A. Yeh, K. Liu, et al.,2006. Inventory of mangrove wetlands in the Pearl River Estuary of China using remote sensing. Journal of Geographical Sciences, **16**(2): p. 155-164.

Wang, S., X. Li, K. Liu, et al.,2007. Dynamic Analysis of the Wetland Resource Changes in the Estuary of the Pearl River Delta Using Remote Sensing. ACTA SCIENTIARUM NATURALIUM UNIVERSITATIS SUNYATSENI, **46**(2).

Cowardin, L., V. Carter, F. Golet, et al., 1979.*Classification of wetlands and deepwater habitats of the United States*: US Department of the Interior/Fish and Wildlife Service.

Sader, S., D. Ahl, W. Liou,1995. Accuracy of Landsat-TM and GIS rule-based methods for forest wetland classification in Maine. Remote Sensing of Environment, **53**(3): p. 133-144.

Scott, D.,T. Jones,1995. Classification and inventory of wetlands: A global overview. Plant Ecology, **118**(1): p. 3-16.

Yang, Y.,2002. New knowledge on the progress of international wetland science research and priority field and prospect of Chinese wetland science research. Advance in Earth Sciences, **17**(4): p. 508-514.

Schmidt, K.,A. Skidmore,2003. Spectral discrimination of vegetation types in a coastal wetland. Remote Sensing of Environment, **85**(1): p. 92-108.

Sheng-he, L.,H. Shu-jin,2002. A spatial analysis model for measuring the rate of land use change [J]. Journal of Natural Resources, **5**.

Liu, J.,A. Buhe,2000. Study of spatial-temporal feature of modern land-use change in China: using remote sensing techniques. Quaternary Sciences, **20**(2): p. 229-239.

Junhong, B., O. Hua, C. Baoshan, et al.,2008. Changes in landscape pattern of alpine wetlands on the Zoige Plateau in the past four decades. Acta Ecologica Sinica, **28**(5): p. 2245-2252.

He, Q.,2006. Monitoring the change of the coastline of the Yellow River delta by integrating remote sensing(RS) and GIS. Geology in China, **33**(5): p. 1118-1123.

Alesheikh, A., A. Ghorbanali, N. Nouri,2007. Coastline change detection using remote sensing. International Journal, **4**(1): p. 61-66.

Sesli, F., F. Karsli, I. Colkesen, et al.,2009. Monitoring the changing position of coastlines using aerial and satellite image data: an example from the eastern coast of Trabzon, Turkey. Environmental Monitoring and Assessment, **153**(1): p. 391-403.

GIS TECHNIQUES FOR MAPPING URBAN VENTILATION, USING FRONTAL AREA INDEX AND LEAST COST PATH ANALYSIS

M. S. Wong [a], J. E. Nichol [a], E. Y. Y. Ng[b], E. Guilbert [a], K. H. Kwok [a], P. H. To [a], J. Z. Wang [a]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University
Kowloon, Hong Kong
[b] School of Architecture, The Chinese University of Hong Kong

m.wong06@fulbrightmail.org

Commission VI, WG VI/4

KEY WORDS: Frontal area index, Landscape roughness parameter, Least cost path analysis, Wind ventilation

ABSTRACT:

This paper presents the urban wind ventilation mapping from building frontal area index on the example of urbanised city, Hong Kong. The calculations of frontal area index for each uniform 100m grid cell are based on three dimensional building databases at eight different wind directions. The frontal area index were then correlated with land use classification map, and the results indicate that commercial and industrial areas were found to have higher values as compared to other urban land use types, primary because these areas have densely high rise buildings. However, with using the frontal area index map, the potential ventilation paths created using least cost path analysis of the city can be located, and the "in-situ" measurements were suggested the existence and accuracy of these ventilation paths. These high ventilation paths could play significant roles in relieving the urban heat island formation and increasing the urban wind ventilation, planning and environmental authorities may use the derived frontal area index and ventilation maps as objective measures of environmental quality over a city, especially human comfort of the urban climate.

1. INTRODUCTION

Most of the data included in wind and air quality studies are from ground level instrument. The gathering of data over large regions therefore is a major challenge to these studies. Wind tunnel model provides an alternative for visualizing the local wind direction and pollutant dispersion in a large scale over a city. Duijm (1996) experimented the wind tunnel model in Lantau island of Hong Kong in a large scale (1:4000) of a small area. Mfula et al. (2005) tested the pollution sources affect by buildings and analysed the wind and pollutant patterns of the surface, a very large building model scale at 1:100. Although wind tunnel studies of air ventilation in urban can provide accurate datasets of concentration fields, measured under hypothesis and constrained conditions, the small coverage, computer demanding requirements and high operational cost are prohibitive its usage. In recent years, a variety of numerical models has been developed for modelling air ventilation, such as PSU/NCAR mesoscale model (known as MM5) and Computational fluid dynamics (CFD) model. The MM5 model works for mesoscale phenomena such as air flow of sea breeze and mountain-valley flow (Dudhia et al., 2003) with large coverage and at coarse resolution, while the CFD model simulates urban wind flow at a larger scale. The CFD model is being widely used in engineering flow analysis, building and structural design, urban wind flow predictions (Baik and Kim, 1999), and air pollution dispersal modelling (Blocken et al., 2007; Huber et al., 2004; Kondo et al., 2006). The CFD model comprises a set of physical models which attempt to closely match the real geometry inside the urban areas and thus simulate the air flow. This model is highly site-specific and cannot be used for all meteorological conditions eg. other cities. Fine resolution products achieve higher accuracy are mainly used for monitoring at street and district levels, due to the higher computational requirements. Therefore, any task of wind ventilation model at city-scale, over densely urbanised regions with complex street and building structures will become more challenging.

However, GIS and remote sensing techniques then provide solutions with simplifying assumptions and numerical approximations. Wind modelling at near surface condition can always be simplified mathematically through the estimating of roughness parameters. There are many studies on modelling and retrieving surface roughness using GIS and remote sensing techniques, and several methods and parameters have been suggested on calculation of surface roughness: zero-plane displacement height (z_d) and the roughness length (z_0) (Lettau, 1969; Counihan, 1975), plan area density (λ_p), frontal area index (λ_f) (Grimmond and Oke, 1999; Burian et al., 2002), average height weighted with frontal area (z_h), depth of the roughness sub-layer (z_r) (Bottema, 1997; Grimmond and Oke, 1999) and the effective height (h_{eff}) (Matzarakis and Mayer, 1992) etc.

Among these urban morphological parameters, frontal area index is suggested as a good indicator for mesoscale meteorological and urban dispersions models (Burian et al., 2002). Frontal area index is the measurement of building walls facing the wind flow in a particular direction (frontal area per unit horizontal area) (Figure 1). It has a strong relationship with surface roughness z_0, and as a function with flow regime inside urban street canyons (Burian et al., 2002). More details of frontal area index will be illustrated in section 3.

The aim of this paper is to (i) analyse the frontal area index with different land use classification types in Hong Kong. (ii) In further, the occurrence frequency of ventilation paths generated from Least Cost Path (LCP) analysis over the study area was used to gain insights to the high ventilation locations, and prevailing transport pathways of airflow. The technique described provides high resolution raster maps which has more understanding on wind ventilation and its interactions with different geometries of buildings at city-scale. (iii) The validity of the mapped ventilation paths was then evaluated in the field with “in-situ” measurements. The deliverables are the maps of frontal area index and the maps of occurrence frequency of ventilation paths.

2. STUDY AREA

The Kowloon peninsula in Hong Kong was selected for the study. It consists of high-density residential areas, financial and commercial districts, and urban parks. The population within the study area is approximately 2 million. The topography is flat in the southern part and hilly terrain is located in the northern part where the elevation has a range from 3m to 300m.

3. METHODS

3.1 Calculation of frontal area index

The frontal area index (λ_f) is calculated as the total areas of building facets projected to plane normal facing the particular wind direction divided by the plane area (Equation 1).

$$\lambda_f = A_{facets} / A_{plane} \qquad (1)$$

where λ_f is the frontal area index, A_{facets} is the total areas of building facets facing the wind direction, and A_{plane} is the plane area.

Burian et al. (2002) used a similar approach for estimating the λ_f in Los Angeles. Here we modified the algorithm by eliminating the blocking areas facing the wind direction on the blocked buildings. Figure 1 illustrates an example of frontal area calculation. Those areas of facets on second buildings were not calculated. A program was written in ESRI® ArcGIS™ 9.2 software to estimate the total frontal areas in the projected plane normal to specific wind direction using digital data of building polygons from the Hong Kong Lands Department. The scale of the digital data is 1:5000. This program was first screened for particular wind direction and generated projected lines with 5m increment horizontally. If the projected lines hit the first facet and could not reach the second facet, only the frontal area of the first facet would be calculated. This process is important for irregular building groups and can reduce number of facets being calculated in computer memory. A splitting process was then applied to ensure building groups were split into individual buildings, and all building polygons were inside their corresponding plane polygons.

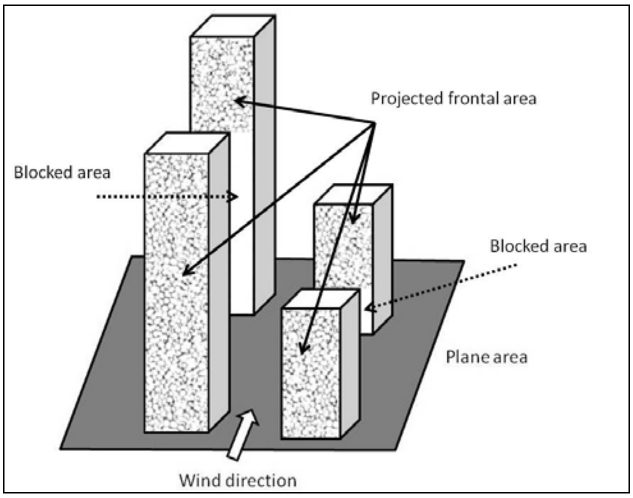


Figure 1. Example of frontal area calculation

For the plane area, we divided the study area into grid cells of 100m x 100m size. Nichol and Wong (2008) show that the resolution at 200m may indicate intra-urban differences at meso-scale between different land cover types but smaller areas such as green spaces with ca. 1 hectare (100m^2) can produce significantly lower surface temperature in urban areas. In addition, the Planning Department (2009) in Hong Kong also found that grid resolution at 100m was compatible to all variables in determining dynamic potential and thermal load contributions in urban climatic study. Thus, our study calculated λ_f at 100m grid resolution in a large 11km by 7km urban area, for eight different wind directions (north, northeast, east, southeast, south, southwest, west, and northwest).

3.2 Frontal area index in different land use types

In Hong Kong, as in most cities, different land use types support different building structures and characteristics, for example: high-rise buildings (>50 floors) in residential, 20-30 floorers in commercial districts, and 20-50 storey larger buildings in industrial district. Therefore, it was expected that these areas would have different λ_f properties. This study used a generalized land use map at 10m resolution acquired from the Hong Kong Planning Department for analysing the relationship between λ_f and different land use.

3.3 Ventilation paths: least cost path analysis

In order to evaluate the relevance of frontal area index to the fresh air corridors in the study area, Least Cost Path (LCP) analysis was undertaken to compare those pathways generated in different wind directions. The pathways represent routes of “high potential” of ventilation locations in the city, degree of connectivity between starting and ending points, and minimum Euclidean distance by considering the cumulative pixel values at each grid cell. The rationale of LCP is to identify the path of least resistance across a cost surface from a starting point to an ending point. This study adopted an approach by allocating variable weightings to the frontal area index of each pixel, eg. the higher λ_f, the higher the friction value. The friction values represent the percentage of obstruction of wind ventilation or air flow, these values can be varied according to the user.

First, the λ_f map was imported to IDRISI v.14.02 (Clark Labs., Worcester, MA, USA). The λ_f pixel values were reclassified into 5 classes and each class was given a weight as friction value. Table 1 shows the weights assigned to 5 classes. Second, starting points should be given, for example, there are fifty points allocated on the east coast of Kowloon peninsula representing the eastward wind transporting across the peninsula from east to west. The friction surface was then created by the IDRIS COST module which computes cost surfaces for the fifty starting points. Third, fifty ending points allocated on the west coast of Kowloon peninsula were then input to the PATHWAY module (Eastman, 2006) for generating LCPs. Finally, there are 2500 and 5186 LCPs for eastward and northeastward wind respectively (Figure 3). The occurrence frequencies for the grid cells in the study domain are calculated by counting the overlaid of LCP segments. Thus, grid cells with high occurrence frequencies are associated with the low friction values, and these cells are indicated as areas of "high potential" contribution of wind ventilation. All these processes were implemented in customization scripts in IDRISI, and the results are shown in section 4.3.

Allocated friction value	λ_f
20	<0.2
40	0.2-0.4
60	0.4-0.6
80	0.6-0.8
100	>0.8

Table 1. Friction values allocated to different classes of λ_f

4. RESULTS

4.1 Map of λ_f

Figure 2 shows the gross distribution of averaged λ_f in eight directions over study area. For the general downtown area, typical average of λ_f is 0.25. Isolated high buildings (>400m high) can be found near the west coast of Kowloon peninsula. It is one of the claims for causing "wall effect" which blocks the sea-breeze wind transporting into the downtown, thus inducing significant UHI effect.

Since the most densely built areas are devoid of vegetation, and elsewhere street planting is severely restricted by lack of space (Jim, 2004), urban vegetation is small and fragmented. Since urban vegetation has a small frontal area comparatively to artificial buildings, trees are not considered in the λ_f calculation. In addition, the urban topography in mainly is flat, terrain is also not considered.

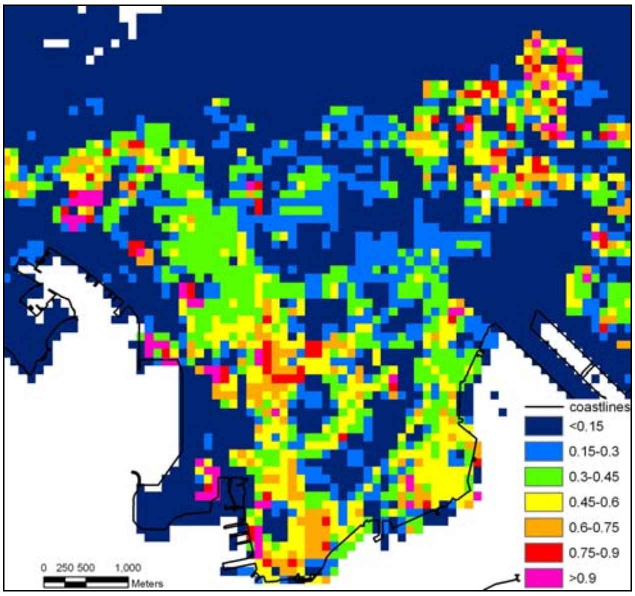


Figure 2. Map of frontal area index

4.2 Relationship between λ_f and different land use types

Table 2 shows the λ_f as a function of different land use types. The industrial and commercial types have the significant high λ_f (0.324 and 0.305) compared to the others. Residential classes have moderate high λ_f (~0.254), public transportation and warehouse have lower λ_f (~0.15). A direct relationship between the frontal area index and the building characteristics can be demonstrated in this study since a general trend can be observed with higher λ_f always associated with wider and taller buildings (eg. industrial and commercial districts). Burian et al. (2002) found that the λ_f in residential, commercial, industrial, public transportation are 0.176, 0.246, 0.095, 0.011respectively in Los Angeles. Grimmond and Oke (1999) studied the λ_f in major cities in north America, and they found the highest λ_f in city center in Vancouver, Canada (0.3) and suburban residential areas in Arcadia, United States (0.33). Although a direct comparison between our study and the others cannot be made due to different algorithms for calculating λ_f, we can still note that the urban areas in Kowloon peninsula have significantly high λ_f values.

Landuse types	λ_f
Private Residential	0.267
Public Residential	0.241
Commercial/Business & Offices	0.305
Industrial	0.324
Warehouse & Storage	0.155
Rural Settlements	0.056
Vegetation	0.075
Public transportation	0.15
Vacant Development Land	0.191

Table 2. Friction values allocated to different classes of λ_f

4.3 Ventilation paths

The eastward and northeastward λ_f maps gave somewhat similar results (Figure 3a, 3c), but different on the distributions of occurrence frequency of LCPs (Figure 3b, 3d). In Figure 3b and

3d, the cross and triangle symbols represent the starting and ending points of the LCPs. There are a total of 2500 and 5186 segments created from easterly and northeasterly directions. From the distribution of the occurrence frequency of LCPs, one can see that the "high potential" of easterly wind ventilation paths are mostly located on

(i) Boundary street (marked as A in Figure 3b). The route traverses between residential area with low-rise buildings across patches of vegetated areas. This street is a marked boundary between the southern and northern Kowloon, where the southern part was ceded to the United Kingdom in 1860 and the northern part was leased in 1898. The occurrence frequency of LCPs along this route is greater than 30% (750/2500).

(ii) Argyle street and Cherry street (marked as B in Figure 3b). This is the shortest route, but is a four-lane dual-way which connects the old airport in the east to the west coast. This route provides significant air corridor in east-west direction and the occurrence frequency of LCPs along this route is greater than 18% (450/2500).

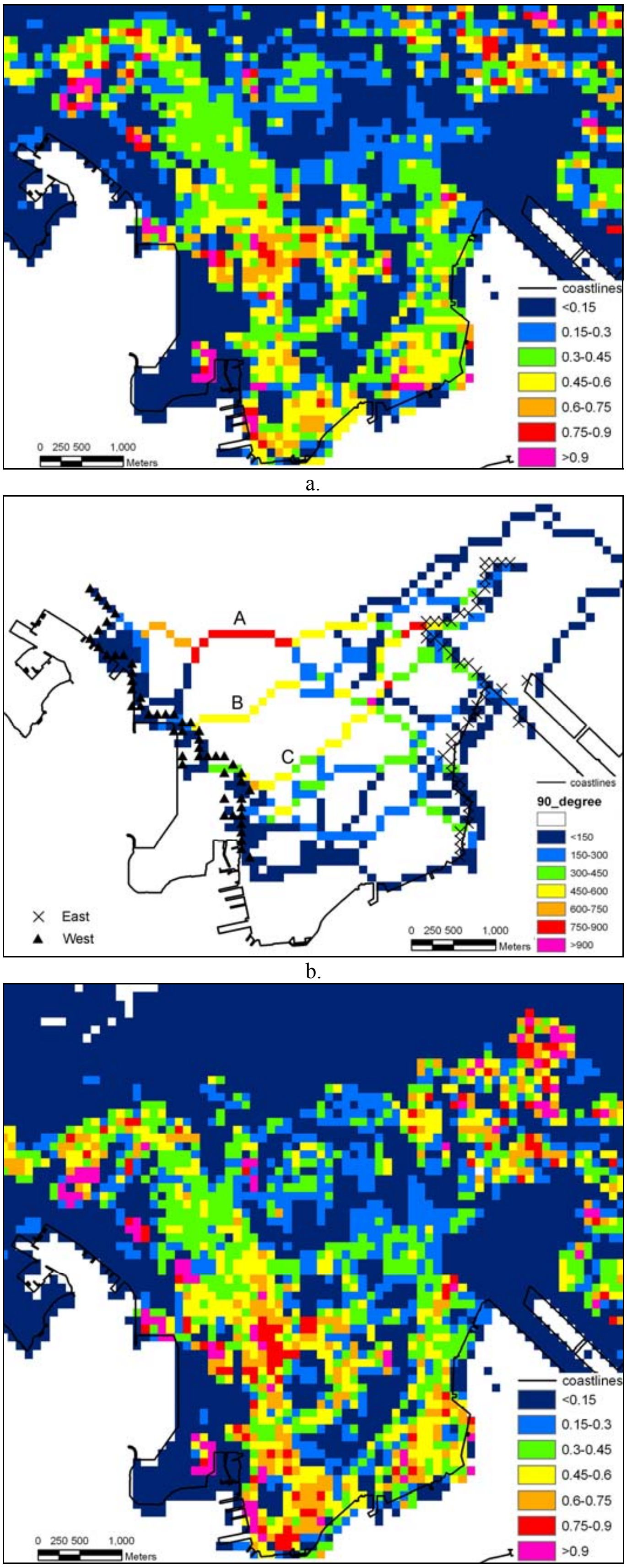
(iii) Ho Man Tin Hill road, King's park road and Gascoigne road (marked as C in Figure 3b). This route traverses large patch of urban vegetation, low density residential development with fragmented trees, and densely commercial areas. The occurrence frequency of LCPs along this route is only greater than 16% (400/2500) which is the least among three.

For these routes, flows are from the east and reached the new reclamation land in the west. These three routes are generally represented by occurrence frequency greater than 24%. However, the northeasterly wind ventilation path is located on

(iv) Princess Margaret road, Chatham road, Hong Chong road and Salisbury road (marked as D in Figure 3d). This route comprises of low rise residential area, urban park, university campus, commercial district and harbor walkway. Route D is the only significant route contains segments in north-south direction among all the 5186 paths. From this route, air flow is from the northeast to the south and turned westerly along the south coast of peninsula. This route is represented by occurrence frequency greater than 28% (1500/5186).

These four air flow pathways had the "higher potential" wind ventilation locations and accounted for the greater number of LCP segments. However, since the easterly and northeasterly winds dominate 66% of all the wind directions in Hong Kong, the north-south oriented street and building geometry reinforce the trapping of polluted air at the downtown. Due to the wind blocking by ridged mountains (eg. 900m height) on the north of peninsula, the fresh air corridors in north-south direction are rarely observed. This study only observes a significant segment of route D in north-south direction.

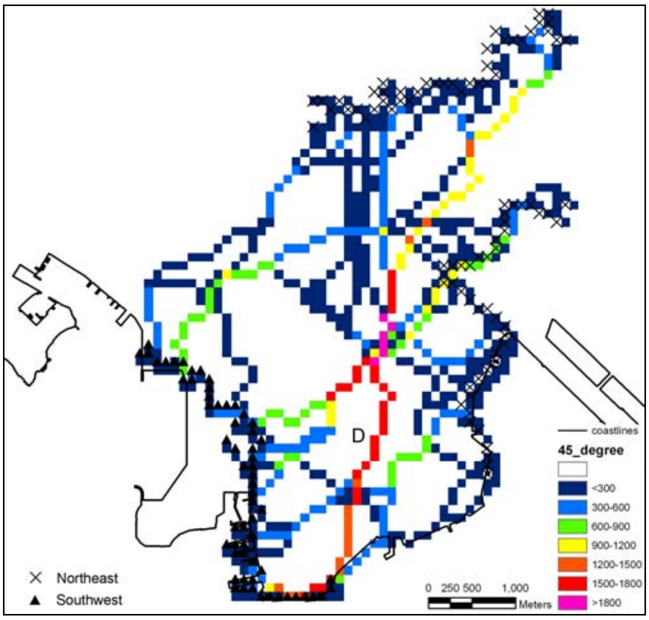
These maps of ventilation paths facilitate the visualization of wind ventilation and show the specific locations in the city e.g. those in red and purple colours, which appear to provide potential air flow corridors for dispersing of the urban pollution. These maps therefore promote better understanding of air ventilation at detailed, as well as in city scale.



a.

b.

c.

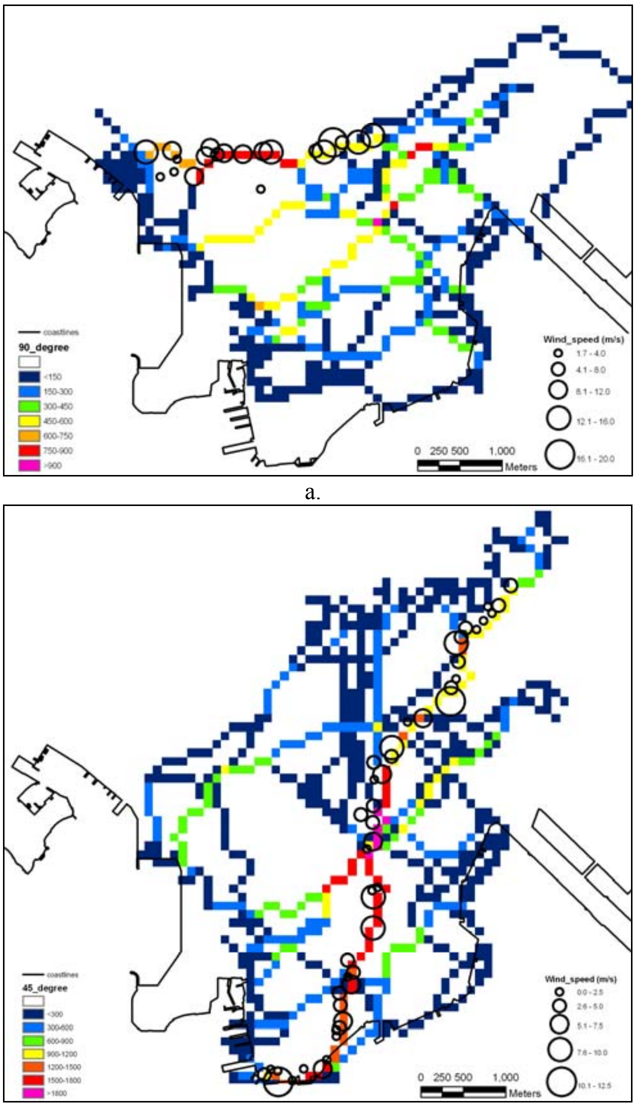


d.

Figure 3. a. Frontal area index map in east-west direction; b. Occurrence frequency of ventilation paths in east-west direction (total number of paths is 2500); c. Frontal area index map in northeast-southwest direction; b. Occurrence frequency of ventilation paths in northeast-southwest direction (total number of paths is 5186)

4.4 Validation strategy

In order to investigate the significance and functionality of the occurrence frequency of ventilation paths, fieldwork was undertaken on 09 Oct 2009, 13 Oct 2009, 15 Oct 2009. The general wind direction of these days are 45, 90, 45 degree, respectively. A slow walk was undertaken along the differing sections of the route A and D, recording the wind speed and locations by GPS along the routes. The wind speeds were measured on 4 occasions, and the results are averaged. These two routes were only chosen because their high occurrence frequencies of ventilation paths, and the total number of "in-situ" measurements along route A and D are 22 and 48 respectively. Figure 4 shows the "in-situ" measurements overlaid with occurrence frequency of ventilation paths, the dots' sizes represented the "in-situ" measurement are scaled with the wind speed. Along the route A, average and maximum wind speeds were observed with values of 9.3m/s and 17.8m/s respectively, and the background wind speed was 2m/s. The wind speed off the route was remarkably low (ca. 2.3m/s). About 55% of "in-situ" data with wind speed larger than 9.1m/s fall in the higher occurrence frequency of 24%, and about 32% of data with wind speed larger than 5.7m/s fall in moderate occurrence frequency of 18%. This may be indicated that route A is identified as a key connecting corridor link the air in the eastern and western sections of the peninsula. However, along route D, the average and maximum wind speeds are smaller than route A, they are 3.5m/s and 12.4m/s respectively. The background wind speed is 0.65m/s and that off the route is ca. 1.1m/s. Approximate 44%, 25% and 21% of data fall in higher occurrence frequencies of 28%, 23%, 17% respectively. However, the largest wind speed (12.4m/s) was observed on the harbor walkway on the south coast of peninsula, since the wind was not only from the north at that location, the sea breeze and offshore wind apparently were the source resulted in this high wind speed. In this validation study, the analysis of occurrence frequency of ventilation paths provided what appears to provide more realistic information on wind speeds and air corridors over the city, especially significantly ventilation paths are observed on route A and D. Since the model has potential for pinpointing the key buildings and refining the building geometry for better air ventilation, this can be useful for urban redevelopment and land reclamation.



a.

b.

Figure 4. Occurrence frequency of ventilation paths overlaid with "in-situ" measurements in a. east-west direction, b. northeast-southwest direction

5. CONCLUSION

This paper presents a comprehensive study of urban ventilation using the model of roughness parameter - building frontal area index on the example of a large urban area in Hong Kong. We calculated the frontal area index based on three dimensional building data under a refined algorithm, as a function of land use type. Most of the frontal area indices as a function of land use type were found to be similar to values computed for other studies, but commercial and industrial areas were found to have significantly high values because these areas are high rise in Hong Kong.

To evaluate the air ventilation in city scale over Hong Kong, LCP analysis was performed. However, whereas LCP analysis usually operates with single segment on purpose, this study adopted an innovative approach by overlaid the LCP segments to derive maps of occurrence frequency of LCP. Four major air ventilation pathways were identified from these maps. The pathways, routes A, B, C, D accounted for 30%, 18%, 16%, 24% respectively of all least cost paths, and they pass from easterly and northeasterly directions respectively. "In-situ" wind speeds were measured to evaluate the robustness and accuracy of the models, and the results apparently show consistency between modelled pathway and "in-situ" measurements.

In densely urbanised Hong Kong, resolution corridors of 100m width in this study may not appear to provide viable connecting corridors at street level, but it would be sufficient for city scale modelling. This study offers a more complete and relevant air ventilation study, and such detailed mapping at city scale has not been undertaken previously. The fuzzy querying on assigning the weights of friction values to the frontal area index pixels can be varied to facilitate upscaling to coarser resolution for regional scale study, eg. regional air dispersion model or wind ventilation models. However, given the high accuracy and high spatial resolution of deriving the wind ventilation model for a city, planning and environmental authorities may use the derived maps as an objective measure of air quality and wind ventilation over a whole city, for comparisons between places and cities and for monitoring changes over time.

Acknowledgments

The authors would like to acknowledge the Hong Kong Planning Department for land use and Tertiary Planning Unit data, and the Hong Kong Lands Department for building GIS databases. This project was funded by PolyU SURF project 1-ZV4V.

References

Baik, J.J. and Kim J.J., 1999. A numerical study of flow and pollutant dispersion characteristics in urban street canyons. Journal of Applied Meteorology, 38, pp. 1576–1589.

Blocken B., Carmeliet J. and Stathopoulos T., 2007. CFD evaluation of the wind speed conditions in passages between buildings – effect of wall-function roughness modifications on the atmospheric boundary layer flow. Journal of Wind Engineering and Industrial Aerodynamics, 95(9-11), pp. 941-962

Bottema, M., 1997. Urban roughness modelling in relation to pollutant dispersion. Atmospheric Environment, 31, pp. 3059-3075.

Burian S.J., Brown M.J. and Linger S.P., 2002. Morphological analysis using 3D building databases, Los Angeles, CA. LA-UR-02-0781, Los Alamos National Laboratory, USA.

Counihan J., 1975. Adiabatic atmospheric boundary layers: a review and analysis of data from the period 1880–1972. Atmospheric Environment, 9, pp. 871–905.

Dudhia J., Gill D., Manning K., Wang W. and Bruyere C., 2003. PSU/NCAR Mesoscale modeling system tutorial class notes and user's guide: MM5 modeling system version 3, NCAR.

Duijm N.J., 1996. Dispersion over complex terrain: wind-tunnel modelling and analysis techniques. Atmospheric Environment, 30(16), pp. 2839-2852.

Eastman R., 2006. IDRISI Andes Tutorial, Clark Labs, Worcester, USA.

Grimmond C.S.B. and Oke T.R., 1999. Aerodynamic properties of urban areas derived from analysis of surface form. Journal of Applied Meteorology, 34, pp. 1262–1292.

Huber A.H., Tang W., Flowe A., Bell B., Kuehlert K. and Schwarz W., 2004. Development and applications of CFD simulations in support of air quality studies involving buildings. 13th Joint Conference on the Applications of Air Pollution Meteorology with the Air & Waste Management Association, (CD-ROM) Vancouver, British Columbia, Canada, August 23-27, 2004.

Jim C.Y., 2004. Impacts of intensive urbanisation on trees in Hong Kong. Environmental Conservation, 25(2), pp. 146-159.

Kondo H., Asahi K., Tomizuka T. and Suzuki M., 2006. Numerical analysis of diffusion around a suspended expressway by a multi-scale CFD model. Atmospheric Environment, 40, pp. 2852-2859.

Lettau H., 1969. Note on aerodynamic roughness-parameter estimation on the basis of roughness-element description. Journal of Applied Meteorology, 8, pp. 828–32.

Matzarakis A. and Mayer H., 1992. Mapping of urban air paths for planning in Munchen. Wissenschaftlichte Berichte Institut for Meteorologie und Klimaforschung, The University of Karlsruhe, 16, pp. 13–22.

Mfula A.M., Kukadia V., Griffiths R.F. and Hall D.J., 2005. Wind tunnel modelling of urban building exposure to outdoor pollution. Atmospheric Environment, 39(15), pp. 2737-2745.

Nichol J.E. and Wong M.S., 2008. Spatial variability of air temperature over a city in a winter night. International Journal of Remote Sensing, 29(24), pp. 7213-7223.

Planning Department, 2009. Urban Climatic Map and Standards for Wind Environment - Feasibility Study. January 2009.

ESTIMATING ICE THICKNESS IN SOUTH GEORGIA FROM SRTM ELEVATION DATA

A P R Cooper[a*], J W Tate[b], A J Cook[a]

[a] British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK - (aprc, acook)@bas.ac.uk
[b] Dept. of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT UK - james.tate@ucl.ac.uk

KEY WORDS: South Georgia, Glaciology, ice thickness, SRTM, Climate change

ABSTRACT:

South Georgia is a glaciated island in the South Atlantic, which provides a primary nesting site for the albatrosses and petrels of the Southern Ocean. 60% of the island is covered by glaciers and ice fields, and the majority of the coastal glaciers are observed to be retreating. A small number of these glaciers are advancing, and others are retreating at anomalously fast rates. As the status of these glaciers is important for environmental management of South Georgia, potentially controlling the spread of invasive species into currently pristine regions, it is necessary to understand the pattern of glacier change in South Georgia. However, detailed study of the glaciology of South Georgia is hampered by lack of measurements of the thickness of the ice. Because of the logistic difficulties of operating on South Georgia, there are no conventional ice thickness measurements from drilling, radar or seismic techniques, and it is unlikely that these will be available in the near future.

This paper addresses this lack of basic information by using surface slope data to estimate the ice thickness of glaciers and ice fields in South Georgia. The surface slope data are derived using surface elevations from the Shuttle Radar Topographic Mission, which provides elevation measurements with a high relative accuracy. The estimate of ice thickness critically depends on assumptions about the conditions in the ice column and at the base of the ice mass, and areas where the estimate is clearly in error provide an insight into changed ice flow conditions or the environment at the ice/rock interface. These anomalous regions are then compared with glacier change data, providing insights into the reasons for the unusually rapid retreat or advance of certain glaciers.

The paper describes the methodology used to compute ice thickness values, with an estimate of accuracy and variability of the thickness measures under different assumptions. The paper then identifies regions with anomalous thickness measurements, and seeks to ascertain why the thickness measurement is unreliable in certain regions. Finally these anomalous areas are compared with coastal change data to suggest why certain glaciers are retreating or advancing more rapidly than the norm for South Georgia, and to make predictions concerning future glacier change.

[*] Corresponding Author

1. INTRODUCTION

The island of South Georgia is located in the South Atlantic centred on 54° S, 37° W (Figure 1). It is isolated, and heavily glaciated, with 60% of the island being covered by glaciers or ice sheets. Its location and isolation make it the primary breeding ground for many species of marine birds, such as albatrosses, petrels and penguins.

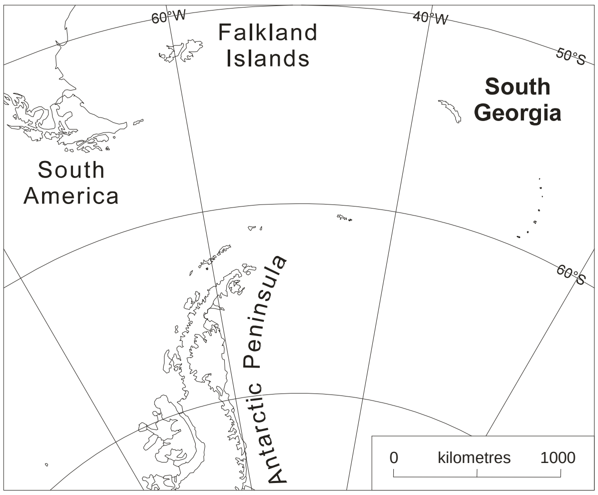


Figure 1: Location map for South Georgia

Recent studies (Cook et al., 2009) have shown that the majority of coastal glaciers on South Georgia are in retreat. A small minority are, however, advancing, and a further small number are retreating far more rapidly than the average. As Cook et al. (2009) show, these changes in glacier extent may have significant consequences on the breeding success of these iconic birds. Introduced terrestrial predators, in particular rats, are currently blocked from major breeding grounds by glacier barriers, and continuing retreat of glaciers threatens these breeding grounds. Comparison of areas occupied by rats and those not shows that most species of bird nesting in South Georgia cannot breed successfully in areas inhabited by rats.

In order to understand glacier dynamics, a crucial parameter is ice thickness. This can be measured in a variety of ways using standard techniques such as seismic sounding, ice-penetrating radar or even by drilling. However, all these techniques require substantial logistic support, which is not available in South Georgia. South Georgia is only accessible by ship; there is no landing ground for fixed-wing aircraft. While an over-snow expedition could potentially carry out a survey of ice thickness, it would be limited in its areal coverage compared with airborne survey. Aircraft equipped for ice-penetrating radar surveys do not have sufficient range to perform a survey over South Georgia after flying from the nearest airfield at Port Stanley in the Falkland Islands.

Fortunately, South Georgia was covered by the Shuttle Radar Topography Mission (SRTM) during February 2000, being just north of the southern limit of 56° S. The product used was DTED 1, 3 arc-second product (~90m post spacing). SRTM elevations have an absolute accuracy of 8 metres (90% probability) for islands, and a relative accuracy of 6.2 metres (Farr et al., 2007). Farr et al. also state that accuracies are worst over steep slopes, and better for flat areas, such as glaciers and ice fields, though the improvement of accuracy over flatter areas is not quantified. These relatively accurate elevation data permit an estimate of ice thickness to be made from the surface slope of the glaciers (Paterson, 1981, page 86).

In places, the ice thickness estimates are clearly substantially in error, giving unrealistically high estimates. This can be linked to changes at the base of the glacier, providing an insight into conditions that are relevant to the pattern of retreat of coastal glaciers.

2. METHOD

2.1 Theory

The surface slope of a glacier in a steady state is related to the ice thickness by the following relationship:

$$\tau = \rho . g . h . \sin \alpha \quad (1)$$

Where τ is the basal shear stress, ρ is the density of ice, g is the acceleration due to gravity, h is the ice thickness and α is the surface slope (Paterson, 1981, page 86).

This equation can be re-arranged to provide a relationship between ice thickness and surface slope, assuming a constant basal shear stress:

$$h = \frac{\tau}{\rho . g . \sin \alpha} \quad (2)$$

So, assuming that the retarding forces at the base of the glacier (τ) are unvarying, it is possible to estimate ice thickness using surface slope values alone, as all other terms in the equation are constant. The value of τ varies within the range 50 kPa to 150 kPa depending on a variety of factors including the temperature of the ice and the nature of the substrate; a reasonable assumption for its value in the absence of other information is therefore 100 kPa.

Glaciers in South Georgia are constrained by valley walls, so additional corrections are required to account for this.

$$h = \frac{\tau}{\rho . g . \sin \alpha . F} \quad (3)$$

Where F is a correction factor that depends on W, the ratio of the distance to the valley wall and the ice thickness on the centre-line of the glacier. GIS techniques detailed below allowed the distance to the valley wall to be computed accurately, and F was obtained from Table 1.

Given that much of South Georgia is covered by perennial snow or ice, determining the location of glacier margins is not trivial. A variety of techniques were tested, but the most reliable was clipping the slope data at a value of 17°. Methods based on image analysis using a composite Landsat ETM+ image failed due to snow cover on glaciers and heavy shadowing, but were used to eliminate areas of low slope that are not snow-covered (e.g. deglaciated areas in front of retreating glacier snouts). The second derivative of the surface (i.e. rate of change of slope) in many areas provided a good delineation of the edge of a glacier, but failed in areas where the glacier merged into snow-fields and at ice-falls. Having determined the glacier margins, the next step was to compute the glacier centrelines. This was done by computing the Euclidean distance from the glacier margins, the centre-line is then the trace of the maximum distances from the glacier margins. The distance to the glacier wall is then available at every point along the glacier centre-line, and the mean ice thickness can be computed by averaging over a small region along the centre-line. These parameters are used to compute the correction factor in Equation 3 (above).

W	F (Parabola)
1	0.445
2	0.646
3	0.746
4	0.806
∞	1.000

Table 1: Table of corrections for a glacier with a parabolic cross-section (Paterson, 1981, p 103)

2.2 Assumptions

In order to estimate ice thickness using the technique described above, certain assumptions are made. These are:

1. That the glacier is flowing by plastic deformation.
2. That ice motion is primarily horizontal.
3. That the limiting basal stress is constant.
4. That the value of the limiting basal stress is 100 kPa,
5. That the cross-section of the glacier is approximated by a parabola.

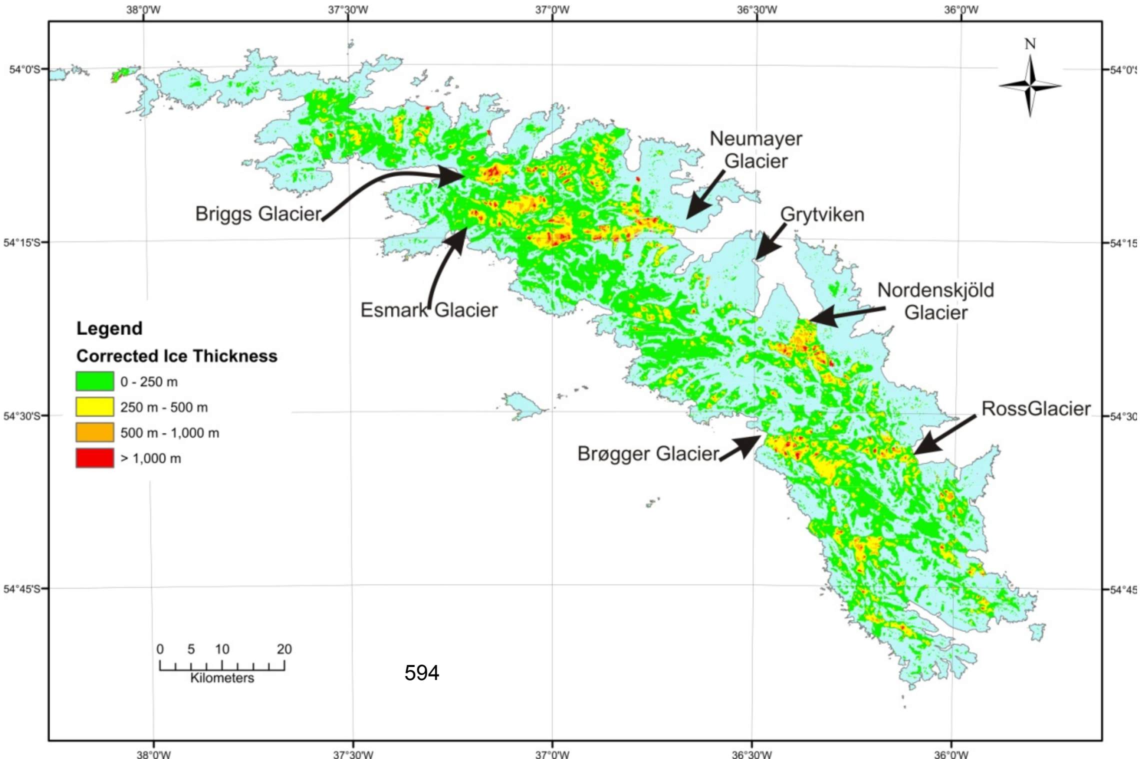
A further issue is that there is a likely (but nor precisely known) difference in the surface measured by the SRTM radar between ablation and accumulation areas of the glacier, caused by differences in the density of firn and glacier ice. However, this will not affect the overall results of this analysis, as the surface density of the ice changes slowly with horizontal location.

These assumptions all break down under certain circumstances, which will be discussed further below. However, assumption 4 is unavoidable; there is no way of estimating the actual limiting shear stress. 100 kPA is a reasonable value; measured values (computed for glaciers where the ice thickness is known) vary over a range from 50 kPa to 150 kPa (Paterson, 1981, page 86); values computed for the Antarctic ice sheets show lower values in certain areas where it is likely that the base of the glacier is lubricated (Drewry, 1983, Sheet 5). Assumption 5 is made for convenience; these parameters have only been computed for a limited range of cross-sectional shapes, and a parabola is the most realistic. However, computations of the correction factor for other cross-sectional profiles only vary by about 10% from each other; errors from breakdown of this assumption will be small compared with other sources of error.

3. RESULTS

The results of the estimation of ice thickness are shown in Figure 2 and summarized in Table 2. As breakdown of the assumptions detailed in Section 2.2 results in overestimates of ice thickness in almost all cases, high ice thicknesses are more likely to be incorrect than lower ice thickness estimates. Figure 3 shows an area where surface slopes are very low in the accumulation area of several glaciers; as can be seen from the 100 m contours of ice thickness, the apparent thickness gradient increases rapidly as the ice thickness increases, being so great when estimated ice thickness exceeds 1000 m that contours above 1000 m have been omitted. These high and rapidly changing slopes are not found in deglaciated terrain, and are unlikely to be real. Therefore, the likelihood of the estimate being correct decreases with increasing estimated ice thickness. Similarly, high ice thicknesses in the lower parts of certain glaciers, as shown in Figures 7-10, are also correlated with high thickness gradients, again suggesting that these high estimates are caused by breakdown of the assumption that the glacier is closely coupled to its bed and is moving by plastic deformation.

Figure 2 Estimated Ice Thickness, with locations mentioned in the text.



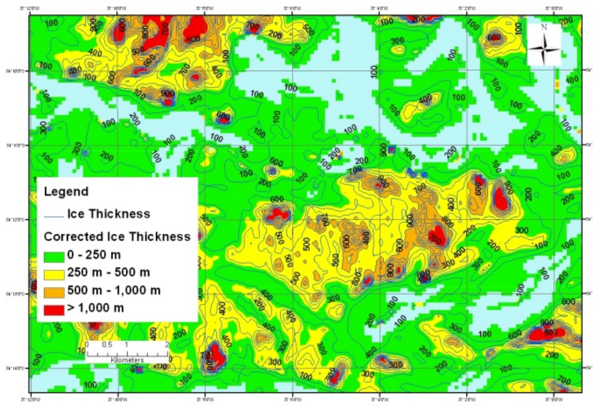


Figure 3 Ice thicknesses illustrating errors

In Figure 2, ice thicknesses have been divided into four colour bands corresponding to the likelihood of the results being correct. Inspection of ice thickness gradients suggests that ice thickness estimates between 0 and 250 m are likely to be an accurate reflection of the actual ice thickness (green on Figure 2), estimates between 250 m and 500 m (yellow on Figure 2) are probably subject to errors in some locations; estimates between 500 m and 1000 m (orange on Figure 2) are very likely to be incorrect, and estimates exceeding 1000 m (red on Figure 2) are highly unlikely to be correct.

Thickness Band	Number of pixels	Area (km^2)	% of ice area
0-250 m	104861	849	74.57%
250-500 m	26321	213	18.72%
500-1000 m	7365	60	5.24%
>1000 m	2071	17	1.47%

Table 2 Area of ice thickness classes.

Regions where the estimate is likely to be unreliable are concentrated in two general regions: inland zones crossing ice-sheds, and the coastal ends of some large glaciers. In both cases, the unreliability is caused by break-down of the basic assumptions of the technique.

Areas in the vicinity of ice-sheds are unreliable because ice motion in these regions has a strong vertical component, defeating the simple analysis of section 2.1.

Areas in the coastal regions of glaciers are almost certainly lubricated at the bed, and the limiting basal shear stress is substantially reduced from the nominal value of 100 kPa. Furthermore, the assumption of plastic flow may also be untrue in this area; the ice may be moving as a block sliding over the base. This is shown by examination of satellite images, which show evidence of rapid streaming flow evidenced by the presence of flow-lines (figure 4). The stream-lines can be better seen in this link (Google Maps), which provides access to high resolution copyright images.

4. DISCUSSION

The ice thicknesses presented here must only be regarded as estimates, with potential systematic errors of ±50% due to the unknown actual value of the limiting basal shear stress. However, with this proviso, it is very encouraging that at least 75% of the ice area provides results that appear to be reasonable, and perhaps another 10% of the area (50% of the 250 m – 500 m) band may well also be reliable. Comparison with ice thicknesses from similar sized but colder Svalbard glaciers (Dowdeswell et al., 1984) suggests that the ice thickness estimates of a few hundred metres are of the right magnitude.

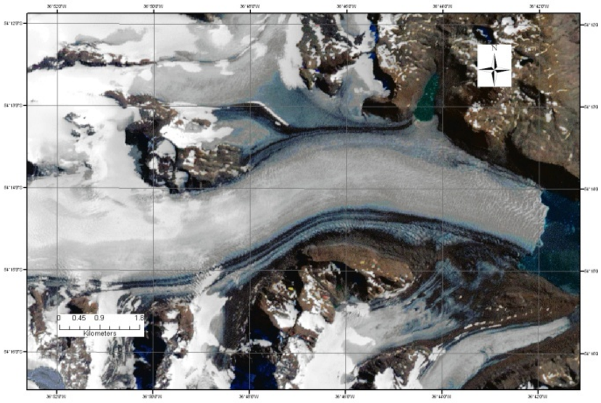


Figure 4 Landsat ETM Image of Neumayer Glacier showing flowlines

The ice thicknesses of Briggs Glacier and Esmark Glacier, coastal glaciers that may be critical to preventing the spread of invasive species (Cook et al., 2009), are shown in Figure 5 and Figure 6. These glaciers, which form barriers to the spread of rats into currently rat-free areas, both show ice-thicknesses that indicate that the glacier flow conforms to the assumptions of Section 2.2 above. This suggests that previously rapid retreat of these two glaciers may not continue in the future, and that the bed of these glaciers is not lubricated.

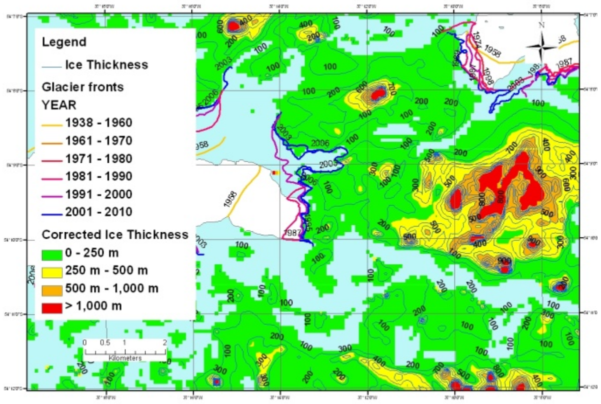


Figure 5 Briggs Glacier showing ice thickness estimates and glacier retreat

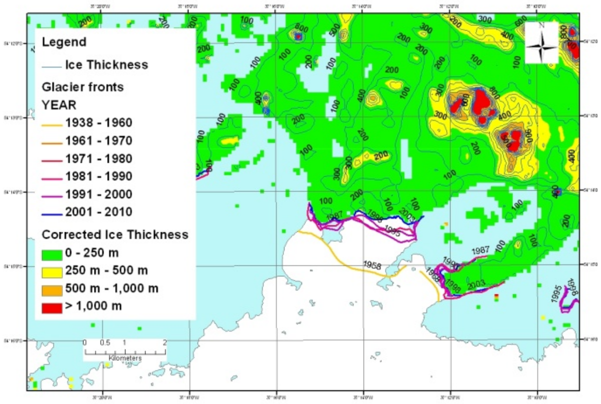


Figure 6 Esmark Glacier showing ice thickness estimates and glacier retreat

The regions where the estimates are too great fall into two distinct classes. First, regions at or near the ice-sheds, where the horizontal motion of the glacier is small, and the ice velocity vectors have a strong downwards component. In these regions, the ice motion does not correspond to the simple model of Section 2.1, and so the ice thickness estimates are too great. However, in these regions (where ice thickness is expected to be at its greatest), the estimates in the range 250 m - 500 m are more likely to be correct.

More interesting are the regions in the coastal extremities of certain large glaciers. As noted above, it is likely that in these regions the increased thickness estimates are due to decoupling of the glacier from the bed, reducing the limiting basal shear stress, and so causing an over-estimate of the ice thickness. This decoupling is likely to be caused by the presence of water at the bed of the glacier. The darker ice shown in **Error! Reference source not found.** indicates that the apparently high thickness region on Neumayer Glacier corresponds to the ablation zone of the glacier, as well as showing other evidence of bed decoupling in the form of stream-lines.

Of the major glaciers of South Georgia, two on the north coast (Neumayer Glacier (Figure 7) and Ross Glacier (Figure 8)) showing this bed decoupling also show rapid retreat. Neumayer Glacier has retreated by over 4 km and Ross Glacier by 3 km. Nordenskjöld Glacier (Figure 9Figure 8), in a similar setting, has retreated, but only by 850 m, a much smaller retreat. A fourth, Brøgger Glacier (Figure 10), on the south coast, shows no retreat at present.

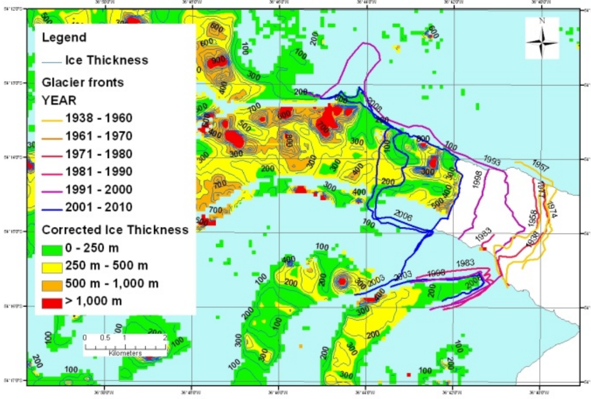


Figure 7 Neumayer Glacier showing ice thickness estimates and glacier retreat

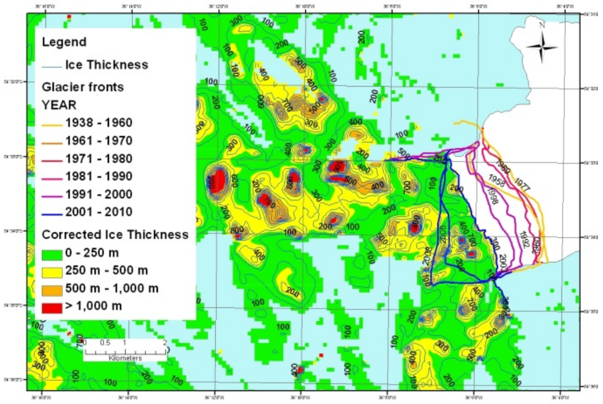


Figure 8 Ross Glacier showing ice thickness and glacier retreat

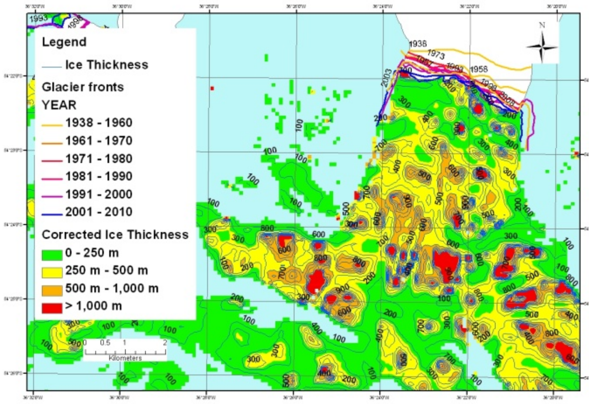


Figure 9 Nordenskjöld Glacier showing ice thickness and glacier retreat

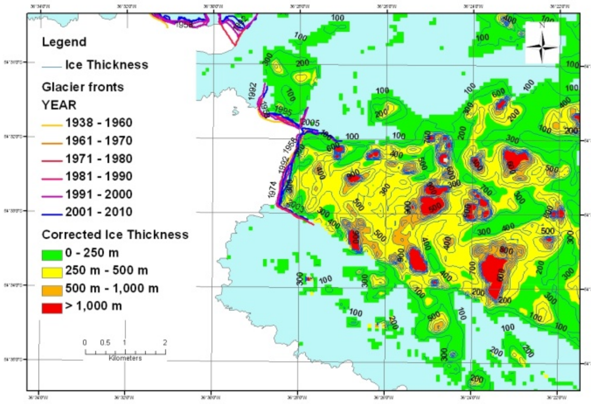


Figure 10 Brøgger Glacier showing ice thickness and glacier retreat

It is noteworthy that the three glaciers showing retreat are all on the north coast of South Georgia, and because of prevailing westerly wind direction, the orographic effect of the mountains forming the central spine of the island will cause a lower accumulation rate than that to the south of the central mountains. Any reduction in the accumulation rate or increase in the elevation of the equilibrium line due to increasing temperatures will cause rapid retreat, especially where the glacier velocity is high. The temperatures at Grytviken, on the north coast of South Georgia, show that summer (January) temperatures have increased significantly since observations began in 1905, and are nearly two degrees higher than the minimum temperatures observed in the 1930s (Figure 11). Winter temperatures show little, if any, change.

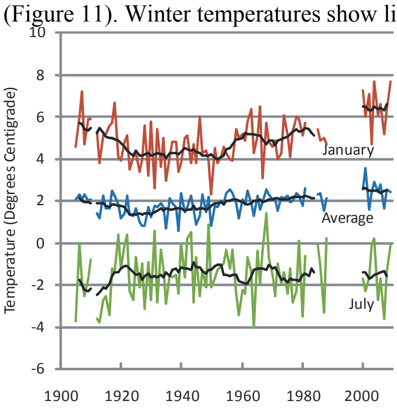


Figure 11 Temperatures recorded at Grytviken. The black lines are 10 year moving averages.

The anomalous ice thickness values for Brøgger Glacier suggest that the glacier is sliding over its bed. Therefore, it requires a high accumulation rate in the upper part of the glacier to sustain the rapid motion in the lower part of the glacier. We can predict that Brøgger glacier may retreat rapidly if the accumulation rate south of the central mountain ranges decreases, either because of increasing temperatures raising the equilibrium line of the glaciers, or because of reduced precipitation.

5. CONCLUSIONS

In the absence of other data, the analysis of this paper shows that surface slope derived from remotely-sensed surface elevation measurements can provide a useful estimator for ice thickness. In addition, the paper shows that useful conclusions can be drawn from regions where it is apparent that the underlying assumptions of the technique break down. In particular, we make the prediction that Nordenskjöld and Brøgger Glaciers will retreat rapidly at some point in the future, as increasing global temperatures reduce the accumulation available to drive these glaciers.

6. ACKNOWLEDGEMENTS

The authors wish to express their appreciation for input from their colleagues, especially Mr A H Fleming and Mr D J Herbert of BAS. Mr Tate wishes to express his thanks to Professor I Dowman, his supervisor at University College, London.

REFERENCES

A. J. Cook, S. Poncet, A. P. R. Cooper, D. J. Herbert, D. Christie (2010), Glacier retreat on South Georgia and implications for the spread of rats, Antarctic Science.

Dowdeswell, J. A., D. J. Drewry, O. LiestøL and O. Orheim (1984). "Radio echo-sounding of Spitsbergen glaciers: problems in the interpretation of layer and bottom returns." Journal of Glaciology 30(104): 16-21.

Drewry, D. J. (1983). Antarctica: Glaciological and Geophysical folio. Cambridge, Scott Polar Research Institute.

Farr, T. G., et al. (2007), The Shuttle Radar Topography Mission, Rev. Geophys., 45, RG2004, doi:10.1029/2005RG000183.

Paterson, W. S. B. (1981). The physics of glaciers. Oxford, Pergamon: 380 pp.

ON-SHORE WIND AND SOLAR POWER PLANTS AS ALTERNATIVE ENERGY SOURCES FOR VICTORIA

S. Margret-Gay [a], I. D. Bishop[a], C. Pettit [b]

[a] Dept. of Geomatics, The University of Melbourne, Victoria, 3010 - sophiemargret@gmail.com;
i.bishop@unimelb.edu.au;
[b] Dept. of Primary Industries, Future Farming Systems Research Division, Victoria, Melbourne, 3001 -
Christopher.Pettit@dpi.vic.gov.au

KEY WORDS: Solar, Wind, Renewable Energy, Victoria, Site Suitability, Landscape Visualisation

ABSTRACT:

This study investigates alternative energy scenarios for Victoria to minimise carbon-emissions. There are many renewable energy options for Victoria; however, the visual impact on the State's landscape is unknown. This project illustrates a Victoria powered solely by renewable energy: namely on-shore wind and solar power plants. Areas suitable for such power sources were identified using data analysis via geographic information systems. The visualisation of energy landscapes was produced using digital globe technologies. Landscape scale visualisation enables the State to be perceived in its totality. This will benefit decision-makers considering sustainable energy mixes by assisting with site selection and policy development.

1. INTRODUCTION

Due to heavy reliance on brown coal, green house gas emissions for Victoria are among the highest in the world per capita (State of the Environment Report, 2008). Sustainable energy options, such as solar and wind power, provide alternatives to reduce Victoria's carbon-footprint. Appropriate siting of alternative power options requires social, environmental and economic concerns to be addressed. Spatial analysis and visualisation assist with the selection of alternative energy options and understanding of their effect on the State's landscape. Visual portrayal of information further enables ideas to be widely shared.

1.1 Wind and Solar Power Potential and Feasibility for Victoria

The Victorian Wind Atlas (2003) indicates Victoria has world-class opportunities for wind farm sites; both inland and along the coastline. In addition, the Sustainability Victoria Website (2009) highlights solar energy as a potential option in Victoria. Manufacturing advances, in both wind turbine and solar cell development, have increased output potentials (Hoffmann, 2006; Edwards, 2008).

1.2 Environmental Concerns

Wind power uses kinetic energy from the wind to produce a clean form of energy without directly producing harmful emissions. Nonetheless, many concerns are raised regarding environmental effects. For instance, sites with high wind speeds may coincide with migratory paths of birds (Welch and Venkateswaran, 2009). No wind farm should be sited directly in an avian migratory path. Also, according to the United State's Federal Aviation Administration (FAA) analysis, the movement of the turbines can cause electromagnetic interference with radar that may result in blind spots for air traffic controllers. Ten kilometres is an acceptable distance to ensure wind farms are not a hazard to aircraft in flight (MTC, 2009).

Furthermore, wind turbines lead to concerns regarding noise. The Danish Wind Industry association reports that new technology has resulted in wind turbines becoming increasingly quiet (DWIA, 2003). It is generally recommended that a distance of 1km is sufficient to eliminate noise disturbance (Pedersen and Persson Waye, 2007).

Solar energy does not produce any direct emissions, pollutants, bi-products or noise. Nonetheless, both these renewable energy resources have a visual impact on the surrounding landscape.

1.3 Visual Impacts

Both wind and solar sources affect the aesthetics of the landscape. Wind farms, in particular, have been widely studied for their visual impact and met considerable local opposition in a number of locations. A variety of variables determine the extent and nature of the impact. Torres-Sibille et al. (2009a, 2009b) have devised indicators to quantify the visual impact of both solar and wind farms on the landscape. They identify factors which contribute to a person's perception of these power plants are: visibility, colour, fractality and movement. Consequently, it is necessary to consider the distribution, configuration and placement of such power plants across the Victorian landscape. Solar plants, on the other hand, are considered 'unobtrusive' (Renewable Energy Sources, 2009) although there has been very little research on the impact of their widespread use.

1.4 Objectives

This study identifies a sustainable energy scenario for Victoria, by determining a mix of on-shore wind farms and solar plants. The mix of resources generates sufficient power to meet current and projected electricity needs for the year 2030. Issues of power storage to cope with variations in production and demand are however not addressed here. Also, our research does not address economic considerations. Overall this study had three objectives:

1) To determine an appropriate energy mix scenario for on-shore wind and solar power plants to address the electricity demands for Victoria.
2) To identify optimal site locations for wind and solar plants under environmental and social constraints.

3) To present visual aids that illustrate the impact on the State’s landscape in its entirety should this energy scenario be realised. Visualisations are intended to communicate a general understanding of the extent of the impact (rather than provide an in situ experience).

2. METHOD

2.1 Determine Energy Scenarios and Corresponding Energy Demand

The energy consumption for Victoria for 2008/2009 and the projected consumption in the year 2030 were provided by the Australian Bureau of Agricultural and Resource Economics (ABARE, 2008). Additionally, the prospect of vehicles being solely powered by electricity by the year 2030, rather than other fuels, was considered. Calculations were based on figures provided by the Australia Bureau of Statistics (ABS, 2003), the Victorian Planning Provisions (VPP, 2008), and General Motors (2006). We defined two renewable production scenarios, with associate storage for load distribution, based on generation potential equivalent to 100% of current and projected (2030) energy demand:

Scenario 1: 7.2 GW
Scenario 2: 12.5 GW (including electric vehicles)

If all existing, approved and proposed solar and wind farms in Victoria were realised then they would generate 71% of estimated energy requirements for scenario 1 and 41% for scenario 2 (Department of Primary Industries, 2009). An additional power capacity of 2GW for 2009 levels and 7.5GW for 2030 levels is required under these scenarios.

Given necessary energy storage, this remaining demand is assumed to be met with 50% wind and 50% solar.

2.2 Identify Suitable Site Locations

Using ESRI’s ArcGIS geographic information system (GIS), suitable site locations for wind and solar farms were determined by considering all concerns raised by the literature.

Sufficient Energy Resources: Only areas with an average annual wind speed greater or equal to 6m/s are considered suitable for wind farms. Similarly, only areas of the State which benefit from 21 MJ/sq. m of solar exposure annually were selected as potential solar plant locations.

Planning Zone Considerations: The potential site locations were limited to the following planning zones: Farming Zone (FZ), Green Wedge Zone (GWZ), Rural Conservation Zone (RCZ), Rural Activity Zone (RAZ), Mixed Use Zone (MUZ) and Industrial Zones (IN*). Additionally, the identified areas were not located in national parks or areas of environmental, heritage or cultural significance. This was achieved by eliminating potential sites with the following planning overlays: Environmental Significance Overlay (ESO), Significant Landscape Overlay (SLO), Heritage Overlay (HO), Erosion Management Overlay (EMO), or Vegetation Protection Overlay (VPO).

Land Use – Protection of Flora: It was necessary to ensure that the proposed sites were not located in areas of dense vegetation. All areas of dense vegetation including native bush and commercial plantation land use areas were eliminated.

Distance from Townships: A one kilometre buffer was placed around towns to minimise aesthetic disturbance to communities.

Threatened Fauna: Point data revealing sightings of threatened fauna was examined. Potential site areas were then limited to areas more than five kilometres from sightings.

Airports: To avoid interference to aircraft, potential sites were limited to greater than ten kilometres from any airport.

High Voltage Power lines: To facilitate the connection of power plants to the State electricity grid, potential sites were preferred within 15 kilometres of high voltage power lines. This was feasible in the case of wind farms because of the large suitable area. Solar sites are more restricted and this preference was not applied.

This process resulted in a single mapping of optimal areas for solar plants. Available wind farm sites were divided into four classes according to wind speed (6-6.5m/s, 6.5-7m/s, 7-7.5m/s, 7.5-8m/s).

The site analysis addresses continuous factors in a binary inclusion/exclusion process; a comprehensive spatial analysis would use a weighted factor combination method. The focus of this study was to illustrate a potential impact of the transition to renewable power rather than provide precise spatial planning.

2.3 Identify New Sites

In order to visualise the energy mix, it was necessary to determine the final sites to be used from the potential sites previously identified. Selected sites were dispersed across the State. (Thus minimising the effect on any one community’s environment). Wind farm sites were limited to 20 km², resulting in 180 turbines. Solar plant sites were restricted to 10 km².

Polygons were created in ArcGIS to represent these sites’ size, shape and location. Additionally, the centroid of these polygons was used to create a point shapefile to depict the location of the sites.

2.4 Visualisation of the Energy Scenarios

The final objective was to visually represent the overall impact on the landscape of Victoria. The two scenarios were visualised using the Google Earth digital globe and Google Sketch-Up software packages, to produce a landscape scale depiction. These visualisations were reproduced in the form of maps and still images although flyovers were also created.

Wind turbines or solar panels are highly visible from ground level but tend to disappear rapidly at distance. Therefore two different approaches to infrastructure visualisation were followed.

- Three-dimensional (3D) polygons were created in Google Sketch-Up to illustrate the space occupied by solar and wind plants. The 3D polygons representing wind farms occupy 20km² and are 120 metres high; these are blue. 3D polygons representing solar plants occupy 10km² and are 20 metres high; these are coloured yellow. Bright colours were selected to maximise the visibility of the polygons. The polygons created in ArcGIS, of required dimensions and exact location of the power plants, were converted into KML (Keyhole Modelling Language) files. These were then imported into Google Earth to be used as guides. The 3D models were then manually placed over each of the 2D polygons.

- Using Google Sketch-Up, solar plant and wind farm realistic 3D models were created. The models were built to scale using the dimensions previously determined. The 3D models were exported as COLLADA (COLLAborative Design Activity) files. These files were then imported into Google Earth. The previously imported polygon KML files enabled the solar plant and wind farm models to be situated in locations determined as suitable through the spatial overlay method described above.

Each process was repeated until the entire Victorian landscape in Google Earth was populated with existing, approved and proposed wind farms and solar plants for both scenarios (1) and (2). Still images were taken in Google Earth at different locations and heights.

3. RESULTS

3.1 Optimal Site Locations for Wind and Solar Plants

The optimal sites for wind farm development were mainly distributed across western and southern Victoria (Figure 1). The optimal positions for solar plants were identified in the north-west of the State (Figure 2). Figures 3 and 4 show the complete distribution, including those currently existing or proposed, of power plants needed to produce the output levels of the two scenarios.

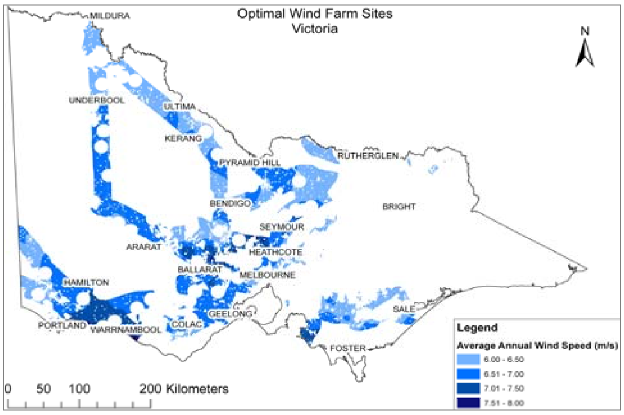


Figure 1: The optimal sites for wind farms across the State of Victoria

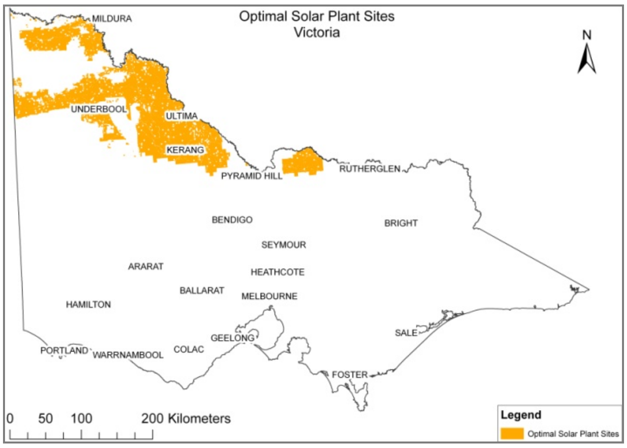


Figure 2: The optimal locations for solar plants across Victoria

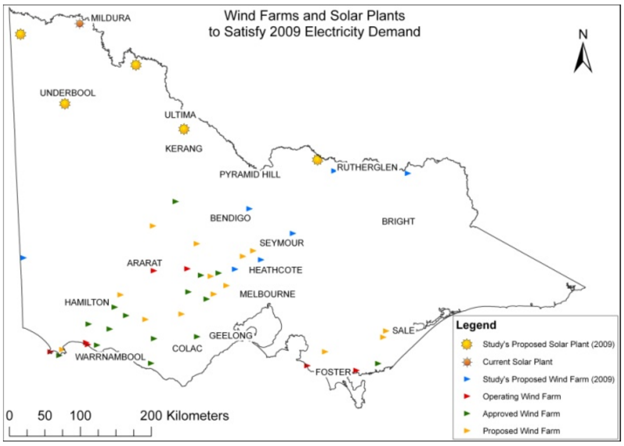


Figure 3: The number, distribution and location of the required solar and wind power plants to meet electricity demand for 2009

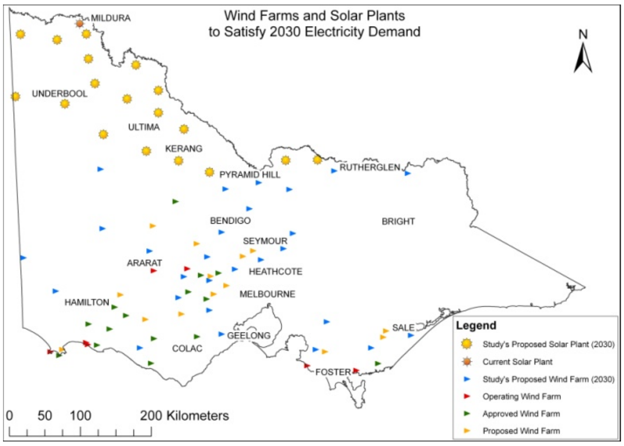


Figure 4: The number, distribution and location of the required solar and wind power plants to meet electricity demand for 2030

3.2 Visualising the Impact on the State's Landscape

Assuming solar and wind plants are visible from a distance of 20 kilometres and no greater, Figures 5 and 6 show how much of the State is visually affected by the developments. A more comprehensive analysis would take account of the hiding effect of the terrain. The analysis could also generate visibility mapping from the road network or other significant vantage points.

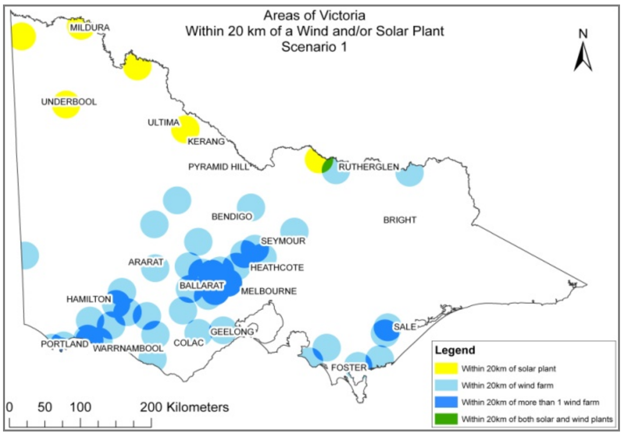


Figure 5: Regions of Victoria where a solar plant or wind farm are considered to be visible for scenario 1

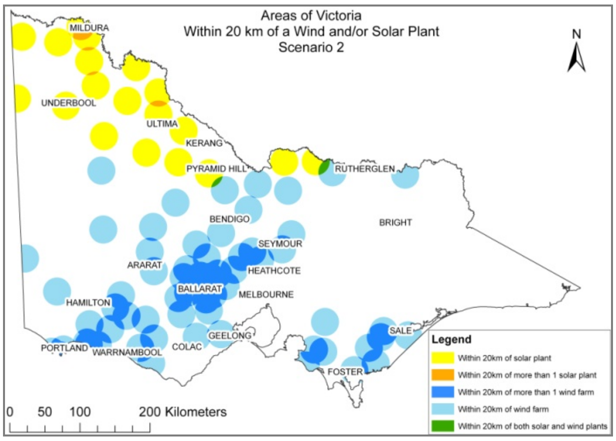


Figure 6: Regions of Victoria where a solar plant or wind farm are considered to be visible for scenario 2

While maps provide an overview of likely affected areas they give no impression of either the visual impact of a 180 turbine wind farm (20 km²) or a 10 km^2 solar plant. They also give little idea of the overall degree of impact on the broader Victorian landscape. Figures 7 to 10 endeavour to address these issues.

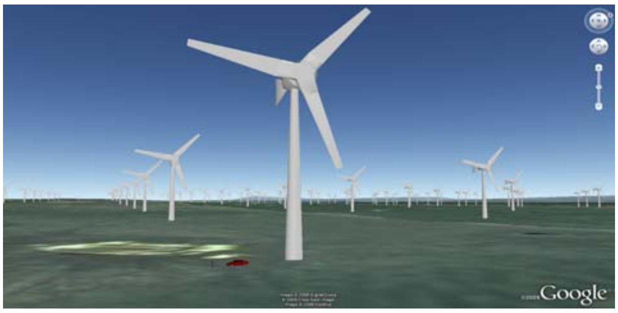


Figure 7: A 180 turbine wind farm on the immediate landscape. A car (5m in length) and a person (1.8m in height) show scale

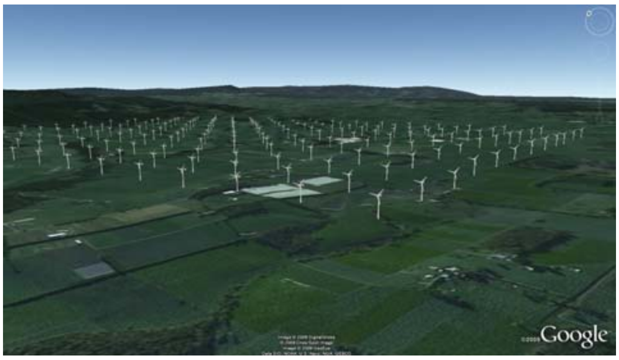


Figure 8: 180 turbine wind farm; eye altitude of 500m

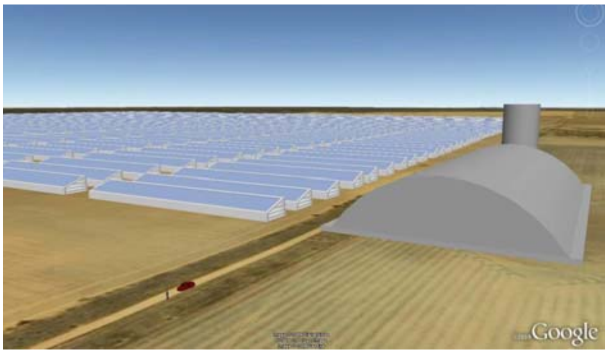


Figure 9: A close-up of a 10 square kilometre solar plant. A car (5m in length) and a person (1.8m in height) show the scale

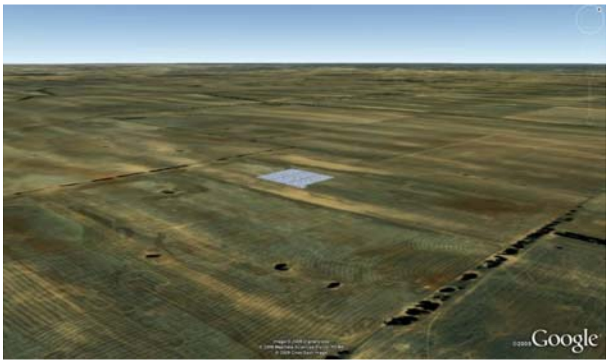


Figure 10: A 10 km² solar plant in north-west Victoria; eye altitude of 500m

If the camera is moved still higher to give a view of the relative distribution of plants, the individual turbines are no longer visible. Figure 11 shows the use of coloured 3D blocks to represent the location and dimensions of the wind and solar installations.

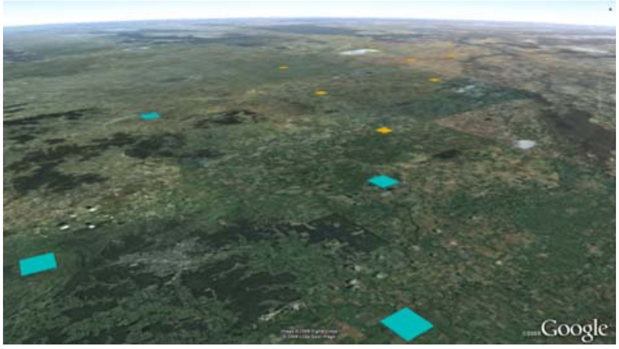


Figure 11: Image illustrating the distribution, size, shape and location of wind farms (blue) and solar plants (yellow); eye altitude of 40km

4. DISCUSSION

This study presents renewable energy mix scenarios to provide for Victoria’s current (2009) and projected future (2030) electricity demands. The proposed scenarios consist of a combination of wind and solar. Landscape scale visualisations illustrate the impact on the State’s landscape in its entirety.

4.1 Site Suitability and Energy Mix

In order to satisfy the two scenarios, suitable site locations were determined. Analysis of geographic information revealed that 9.8% of the State is optimal for solar plant development. In contrast, 30% of Victoria is appropriate for wind farms. The optimal positions for solar plants were identified in the north-westerly part of the State where the average annual solar exposure is 21 MJ/sq. m. Moreover, the average annual wind speeds available in suitable locations for wind farms range from six metres per second to eight metres per second. The windiest areas are situated along the coastline near Portland and Warrnambool and inland near Ballarat.

The existing, approved and currently proposed plants were also considered. Currently the only solar plant development in Victoria is estimated to have an installed capacity of 154 MW. In contrast, the estimated capacity of operating and planned wind farms in the State is 4 957 MW. Thus, 3% of the energy generated by these two sources is solar and the remaining 97% is wind. The extensive wind resources in Victoria are already being used to a far greater extent than solar power.

In general, wind farms are considered more intrusive to the landscape than solar plants (Renewable Energy Sources, 2009). This suggests that from a visual viewpoint it would be beneficial to maximise use of solar plants. However, as suitable sites for solar plants are restricted to the north-west corner of the State, this would result in a large number of solar plants in a limited region; such grouping is visually undesirable.

By satisfying the electricity demand with 50% solar resources and 50% wind resources, it was calculated that overall, for 2009 levels, solar plants would contribute 17% and wind farms the remaining 83%. This dramatically reduces the current discrepancy of distribution. For 2030 levels, solar plants would contribute 31% and wind farms the remaining 69%. This energy mix coincides with the availability of the respective resources and enables dispersion of the plants to reduce the visual impact.

4.2 Final Site Selection for Wind Farms and Solar Plants

Stanton (1996) highlights that the intrusiveness of a wind farm is not directly proportional to the number of turbines in an array, but rather is attributable to design aspects. For example, large wind plants may appear less dominating than a smaller development when the large wind plant is presented in a visually comprehensible method. For this study, it was decided to use fewer large sites for wind farms and solar plants to minimise visual interference; rather than a greater number of smaller plants more frequently distributed across the State.

To satisfy current electricity demands an additional eight wind farms and five solar plants would be required. To meet the projected electricity demand for the year 2030, an additional 27 wind farms and 18 solar plants are needed. The land area required for scenario 1 is 210 km² and for scenario 2 is 720 km².

Among the environmental and social concerns raised by the literature our mapping indicated that the dominant constraint was the visual impact on the environment; primarily to minimise the visual impact on any one community. This objective was achieved by dispersing the wind farms and solar plants across the State in the potential areas identified.

The most suitable locations for wind farms in Victoria cover an area from the coast near Portland, moving inland in a north-easterly direction, to Seymour. Despite the considerable number of current existing developments or projects in different stages of completion in this region, a number of additional developments were proposed within this zone (figures 3 and 4). Sites chosen aim to keep the density within reasonable limits. Furthermore, in order to minimise the visual impact of wind farms, this study selected no sites in coastal regions (within 15km of the coastline), despite the rich wind resources available. Research reveals that the impact on the landscape is significantly higher in areas of natural beauty, primarily coastal regions. In contrast, in areas of low natural beauty, wind farms actually improve the visual aesthetics (Lothian, 2008).

4.3 Visibility Analysis

Using the simplistic assumption that solar and wind plants are visible to the human eye from a distance of 20 kilometres and no greater; analysis was performed to identify the proportion of the State where a solar and/or wind farm could be seen. Figures 5 and 6 illustrate the areas where the power plants may be visible for each scenario.

For scenario 1, only 3% of the State is within visible range of a solar plant and 20% is within range of a wind farm. For scenario 2, 9% of the State can ‘see’ a solar plant and 27% can ‘see’ a wind farm. Unsurprisingly, the proportion of the State that can ‘see’ a solar plant in 2030 increases approximately three times, in accordance with the increase in the number of solar plants. However, the area where wind farms are visible only increases by 37% despite the number of additional turbines proposed by the study increasing by 237%. There are 56 wind farms in different stages of development in Victoria (Wind Projects in Victoria, 2009). This study proposes a further 27 wind farms in scenario 2: a 48% increase. However at 180 turbines each, these farms are considerably bigger than many of the existing farms which typically have 20 to 50 turbines. The area where wind farms may be visible increases more than the percentage of additional wind farms; this is because the wind farms proposed by this study are more dispersed than the existing developments. Whether this is the best strategy is an open question. Figure 6 shows that large areas are within viewing distance of more than one wind farm. A traveller from Portland to Seymour may never be out of sight of wind turbines.

4.4 Landscape Scale Visualisation Techniques

Three-dimensional models, created in Google Sketch-Up and then imported into Google Earth, provided a realistic illustration of the impact on the State’s landscape. The size of 3D models and their impact on the surrounding environment has been visualised. Still images were taken of the 3D models in Google Earth at different proximities and elevations. Figures 7 and 9 indicate the visual effect of a power plant on the landscape for a nearby person. These images give the viewer a sense of the extent of each power plant.

In contrast, figures 8 and 10 are images taken of the power plants from 500 metres altitude. These images are effective in illustrating the impact of individual power plants across the broader landscape. However, no two power plants were closer than 30 kilometres. Thus when visualising the landscape it was not possible to easily see the next closest power plant. Hence the viewer would be unable to perceive the visual impact of all the power plants on the Victorian landscape in their entirety from the still images. A secondary form of visualisation was required.

To provide a more representative visualisation of proximity of power plants to one another, 3D polygons built to scale highlight the size, shape, position and distribution of the plants (figure 11). These brightly coloured polygons are easier to see on the landscape compared to the wind turbines and solar cells.

The polygons clearly and accurately depict the distribution of power plants on the landscape. The bright colours of the models however, do not give a representative idea of the visual impact of power plants. These models have a greater impact on the landscape than the actual power plants would from the same viewpoint. Thus, the polygons are appropriate for highlighting the distribution of power plants on the landscape and their proximity to one another; the realistic 3D models are more representative of the actual visual impact as perceived from the viewpoint of a person travelling through the landscape.

5. CONCLUSION

This research examined an alternative renewable energy mix for Victoria. It identified optimal site locations for wind and solar power plants across the State of Victoria. In addition, a suitable energy mix of solar and wind resources was established to satisfy two energy demand scenarios. Scenario (1) satisfied 100% of current energy demands and scenario (2) satisfied 100% of projected energy demands for 2030.

The study employed landscape scale visualisation to convey the visual impact of the scenarios on the Victorian landscape. Maps and still images sought to illustrate the impact on the landscape in its entirety should these energy scenarios be realised. The success of the visualisation approach has not been tested. The ideal, in terms of providing a viewer with an accurate sense of the visual effects on the Victorian landscape would be to enable the viewer to take unrestricted virtual journeys through and over the landscape. A number of flyovers were produced but even these distance the viewer from the full impact of a landscape with frequent and substantial energy infrastructure. Another visualisation option would be to generate animations showing continuous views from major highways since these would give the best sense of the frequency of occurrence of the new power plants. Further exploration of these issues, in light of the demands arising from greenhouse mitigation strategies, is clearly warranted. Nevertheless, our simple approach could benefit decision-makers considering renewable energy mixes; assisting with appropriate site selection and policy development.

6. REFERENCES

ABARE, 2008. Australian Consumption of Electricity, by State, The Australian Bureau of Agricultural and Resource Economics (ABARE), Australia. http://www.abare.gov.au/interactive/energyUPDATE08/excel/Table_I_08.xls (accessed 20 Mar. 2009)

Australian Bureau of Statistics, 2003. Survey of Motor Vehicle Use, Australian Bureau of Statistics (ABS), Australia. http://www.abs.gov.au/ausstats/abs@.nsf/ProductsbyReleaseDate/3EAAB384EF8D2F62CA2570800072002D?OpenDocument (accessed 20 Mar. 2009)

DPCD, 2008. Victoria in Future 2008, Department of Planning and Community Development, Victoria, Australia. http://www.dse.vic.gov.au/DSE/dsenres.nsf/LinkView/B9023E3BAACA5A6ACA256EF60019E55806C7DF80826B65674A256DEA002C0DCA (accessed 22 Aug. 2009)

DWIA, 2003. Danish Wind Energy Association, Denmark. www.windpower.org/en/core.htm (accessed 10 May 2009)

Edwards, R., 2008. When it comes to turbines, bigger is accepted as better, *New Scientist*, 200(2677), p. 35.

General Motors, 1999. Performance Statistics - 1999 General Motors EV1 w/NiMH. United States Department of Energy Office of Energy Efficiency and Renewable Energy, USA. http://www1.eere.energy.gov/vehiclesandfuels/avta/pdfs/fsev/eva_results/ev1_eva.pdf (accessed 25 Apr. 2009)

Gipe, P. and Wiley, J.,1995. Wind energy comes of age. *Energy Policy*, 19(8), pp. 756-767.

Hoffmann, W., 2006. PV solar electricity industry: Market growth and perspective. *Solar Energy Materials and Solar Cells*, 90, pp. 3285-3311.

Lothian, A., 2008. Scenic Perceptions of the Visual Effects of Wind Farms on South Australian Landscapes. *Geographical Research*, 46(2), pp. 196 -207

MTC, 2009. Airspace Issues in Wind Turbine Siting, Massachusetts Technology Collaborative, Massachusetts, USA. http://www.masstech.org/rebates/Community_Wind/faaairspace.html (accessed 1 May 2009)

Pedersen, E. and Persson Waye, K., 2007. Wind turbine noise, annoyance and self-reported health and well-being in different living environments. *Occupational and Environmental Medicine*, 64, pp. 480-486.

Renewable Energy Sources, 2009. Technology Comparison, World Press, USA. http://www.renewable-energy-sources.com/2008/11/28/comparison-of-energy-sources/ (accessed 24 Sep. 2009)

Stanton, C.,1996. The Landscape Impact and Visual Design of Windfarms, School of Landscape Architecture, Heriot-Watt University, Scotland, United Kingdom. pp. 1-52.

State of the Environment Report, 2008. Commissioner Environmental Sustainability Victoria, Victoria, Australia. www.ces.vic.gov.au/CES/wcmn301.nsf/childdocs/-FCB9B8E076BEBA07CA2574F100040358?open (accessed 28 May 2009)

Sustainability Victoria Website, 2006. Sustainability Victoria, Victoria, Australia. www.sustainability.vic.gov.au/www/html/2119-interactive-maps.asp (accessed 20 May 2009)

Torres Sibille, A.d.C., Cloquell-Ballester, V.-A., Cloquell-Ballester, V.-A. and Darton, R., 2009b. Development and validation of a multicriteria indicator for the assessment of objective aesthetic impact of wind farms. *Renewable and Sustainable Energy Reviews*, 13(1), pp. 40-66.

Torres-Sibille, A.d.C., Cloquell-Ballester, V.-A., Cloquell-Ballester, V.-A. and Artacho Ramírez, M.Á., 2009a. Aesthetic impact assessment of solar power plants: An objective and a subjective approach. *Renewable and Sustainable Energy Reviews*, 13(5), pp. 986-999.

Victoria Planning Provisions, 2009. Department of Planning and Communtiy Development, Victoria, Australia. http://www.dse.vic.gov.au/planningschemes/VPPs/combinedPDFs/VPPs_All_Clauses.pdf (accessed 20 Aug. 2009)

Victorian Wind Atlas, 2003. Sustainable Energy Authority, Victoria, Australia.

Welch, J. and Venkateswaran, A., 2009. The dual sustainability of wind energy. *Renewable and Sustainable Energy Reviews*, 13(5), pp. 1121-1126.

Wind Projects in Victoria, 2009. Department of Primary Industries, Victoria, Australia. http://www.dpi.vic.gov.au/dpi/dpinenergy.nsf/childdocs/-384C1AC0F3D5716CCA25729D00102547-FD29EA297F0AB66DCA2573540007C7D6?open (accessed 24 Apr. 2009)