FUZZY EVIDENCE THEORETIC APPROACHES FOR KNOWLEDGE DISCOVERY IN SPATIAL UNCERTAINTY DATA SETS

Binbin He[1,2] Tao Fang[1] Dazhi Guo[2a]

[1]Institute of Image Processing & Pattern Recognition, Shanghai Jiao Tong University, No.1954 Huashan Road, Shanghai, China 200030, binb_he@163.com, tfang@sjtu.edu.cn
[2]Department of Environment & Spatial Informatics, China University of Mining and Technology, Xuzhou, JiangSu, China 221008, guodazhi@pub.xz.jsinfo.net

KEY WORDS: Data mining, Reasoning, Algorithms, Transformation, Representation, Visualization

ABSTRACT:

Although uncertainties exist in spatial knowledge discovery, they have not been paid much attention to. In the past years, the most researches of spatial knowledge discovery focused on the methods of data mining and its algorithms. In this paper, uncertainty and its propagation of spatial data are discussed and analysed firstly. Then, uncertainties at various stages of spatial knowledge discovery are briefly analysed, including data selection, data preprocessing, data mining, knowledge representation and uncertain reasoning. Thirdly, a method of spatial knowledge discovery in conjunction with uncertain reasoning by means of fuzzy evidence theory is proposed. Herein, the framework for uncertainty handling in spatial knowledge discovery is constructed, and the fundamental issues include soft discretization of spatial data, fuzzy transformation between quantitative data and qualitative concept, reasoning under uncertainty and uncertain knowledge representation.

1. INTRODUCTION

Spatial Knowledge Discovery (SKD) is to extract the hidden, implicit, valid, novel and interesting spatial or non-spatial patterns, rules and knowledge from large-amount, incomplete, noisy, fuzzy, random, and practical spatial databases, which include spatial data mining and uncertain reasoning. In recent years, the term, "spatial data mining and knowledge discovery" (SDMKD) has been connectedly used, in which data mining is a key step or technique in the course of spatial knowledge discovery. With an efficient and rapid improvement of automatic obtaining technologies of spatial data, the amount of data in spatial database have been increased in index movement. But the deficiency of analysis functions in geographic information systems (GISs) induces a sharp contradiction between the magnanimity data and useful knowledge acquisition, in the other words, “The spatial data explode but knowledge is poor” (Li, 2002). At present, spatial knowledge discovery mainly concentrated on the principles and methods of data mining. Another important issue –uncertainty in spatial knowledge discovery –have not been paid much attention to. On the one hand, spatial data itself lies in uncertainty, and on the other hand, many uncertainties will be reproduced in spatial knowledge discovery process, even propagated and accumulated, it lead to the production of uncertain knowledge. These characteristics had not been considered, and the knowledge discovered had been regarded as an entirely useful and certain knowledge in traditional spatial data mining and knowledge discovery. The role that uncertainty can play in spatial knowledge discovery probably is more significant than those in many other research fields, because of the native of knowledge discovery (which is to find hidden knowledge patterns from data). It is to convenient to study spatial knowledge discovery by starting from perfect spatial data with perfect result. However, spatial data are usually far from perfect, and the spatial knowledge discovery process itself is full of various kinds of uncertainty. Spatial knowledge discovery incorporating uncertainty is important, because it puts the study of spatial knowledge discovery in more realistic setting. So the research on the uncertainty of spatial knowledge discovery have become a very important issue.

Furthermore, uncertain reasoning, as a traditional research area of artificial intelligence is aimed at developing effective reasoning method involving uncertainty, namely, to derive what is behind data even data is incomplete, inconsistent, or with other problems. Many uncertain reasoning methods, such as fuzzy set theory, evidence theory, and neural networks, are powerful computational tools for data analysis and have good potential for data mining as well. But traditional spatial data mining and knowledge discovery did not pay attention to these characteristics. In this paper, on the basis of analysis of uncertainty in spatial data, uncertainties at various stage of spatial knowledge discovery were analysed briefly. Especially, a method of spatial knowledge discovery in conjunction with uncertain reasoning by means of fuzzy evidence theory is proposed.

2. UNCERTAINTIES OF SPATIAL DATA

2.1 The Types and Origins of Uncertainty in Spatial Data

It is said that the uncertainty within spatial data is the major components and forms for the evaluation of spatial data quality. Spatial data quality includes lineage, accuracy, completeness, logical consistency, semantic accuracy and currency (FGDC, 1998). All types of spatial data are subjected to uncertainty, since it is impossible to create a perfect representation of the infinitely complex real world (Goodchild, 2003). Error refers to the discrepancy between observation results and true value, which has statistic characteristics. Uncertainty is more broadly-defined error concept continuation, measuring the discrepancy degree of the surveying objects’ knowledge. Uncertainties in spatial data can be classified: error, vagueness, ambiguity and discord (Fisher, 2003).

The obtaining process of spatial data includes cognition, surveying, interpreting, data input, data processing and data representation. The uncertainties of spatial data stem from two parts. On the one hand, they stem from instability of natural phenomena and incompleteness of men's cognition. On the other hand, the process of spatial data capture and handling bring a lot of uncertainty. In addition, these uncertainties can be propagated from the former phase into the latter one, and accumulated in different laws (Figure 1).

2.2 Uncertainty Measurement and Propagation of Spatial Data

At present, a great deal of research have been developed in some areas, such as positional uncertainty and its propagation, especially the uncertainty of points, lines, polygons or areas. Therein, the uncertainty of points and lines is the basis of polygons and areas. Some uncertainty models have been constructed, including standard ellipse model (Mikhail and Ackerman, 1976) and circle normal model (Goodchild, 1991) of point position, Epsilon-band model (Chrisman, 1982) and error band model (Dutton, 1992) of line position. For last years, the study of spatial data quality control data put emphasis on the spatial positional uncertainty, but little on attribute uncertainty. In recent years, some scholars studied the attribute uncertainty of GIS data (Liu, 1999; Ehlschlager, 2000; Shi, 2002). Usually, positional uncertainty and attribute uncertainty were studied respectively. Shi (2000) constructed "S-band" model that combine positional uncertainty with attribute uncertainty. Zhang (1999) constructed field model that position uncertainty and attribute uncertainty are described in uniform.

The spatial uncertainty propagation problem can be formulated mathematically as follows:

$$Y(\bullet) = Opt(D_1(\bullet), \cdots, D_m(\bullet)) \quad (1)$$

Let $Y(\bullet)$ be the output of GIS operation $Opt(\bullet)$ on the m spatial data sets. The operation $Opt(\bullet)$ may be one of the various operations in GIS such as intersection of data sets. The principle of the spatial uncertainty propagation analysis is to determine the spatial uncertainty in the output $Y(\bullet)$ given the operation $Opt(\bullet)$ and spatial uncertainties in the data sets. The spatial uncertainty propagation is relatively easy when the operation $Opt(\bullet)$ is a linear function, which can be performed by error propagation law. However, few of operation $Opt(\bullet)$ were linear or could be solved by simple calculation. However, in general, rigorous methods and functions will be very troublesome. The Monte Carlo method (Openshaw, 1989) uses an entirely different approach to determine the uncertainty of geospatial objects. In this method, the results of Equation (1) are computed repeatedly, with input value $D = [D_1, D_2, \cdots, D_m]$ that are randomly sampled from their joint distribution. The outputs of the equation construct random samples, in which their distribution parameters, such as mean value and variance, can be estimated. The Monte Carlo method may be a general method for uncertainty handling, and can be applied to the computation processing of spatial or attribute data. An outstanding advantage of Monte Carlo method is able to provide the entire distribution of output data at an arbitrary level of accuracy. The other advantages of this method are easy implementation and more general application. However, this method is more intensive computationally.

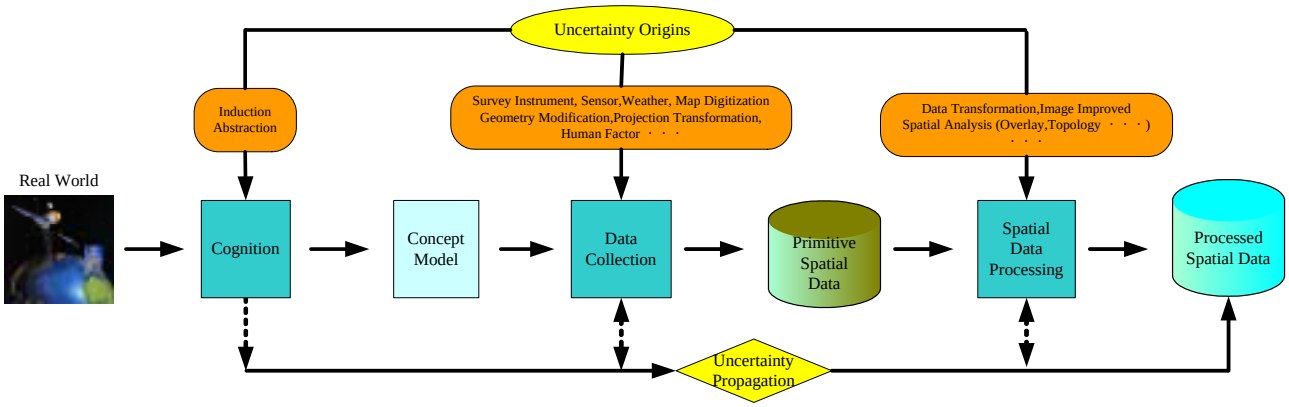

Figure 1. Uncertainty origins and uncertainty propagation of spatial data

3. UNCERTAINTY ANALYSIS IN SPATIAL KNOWLEDGE DISCOVERY

The uncertainties in spatial knowledge discovery may exist in the process of spatial data selection, spatial data preprocessing, data mining, knowledge representing and uncertain reasoning. The study on the uncertainties of spatial data themselves are very important for spatial knowledge discovery, for the original data of spatial knowledge discovery stem from uncertain spatial database or uncertain spatial data sets being analyzed. Moreover, uncertainties in spatial data may directly or indirectly affect the quality of spatial knowledge discovery (Miller and Han, 2001). At the same time, a lot of uncertainties exist in spatial knowledge discovery. Moreover, uncertainties will be propagated and accumulated in spatial knowledge discovery process (Figure 2). The uncertainties of every phase will be analyzed briefly as follows:

At the phase of spatial data selection, Uncertainties mainly stem from a subjectivity of selecting object data according to the task of spatial knowledge discovery, including what data should be collected, and how much data is enough, also these spatial data necessarily embody some kinds of errors or uncertainties.

Spatial data preprocessing mainly include data cleaning, data transformation and data discretization, in which many uncertainties will be produced if we do not adopt appropriate uncertainty handling methods. Data discretization is to divide a

given continuous attribute data into discrete values, and this operation may be a main origins of uncertainties in the whole process of spatial knowledge discovery. At this phase, a lot of uncertainties may be eliminated by uncertainty handling techniques but never completely, even some new uncertainties will be produced in handling process due to impropriety of the techniques.

Uncertainties from data mining mainly refer to the limitation of mathematical models, and mining algorithm may further propagate, enlarge the uncertainty during the mining process.

Spatial knowledge representation exists in uncertainties, including randomness, fuzziness and incompleteness. To a same knowledge, it may be represented by different methods. Most of spatial knowledge discovered by spatial data mining is qualitative knowledge and the best way to represent them is the natural language.

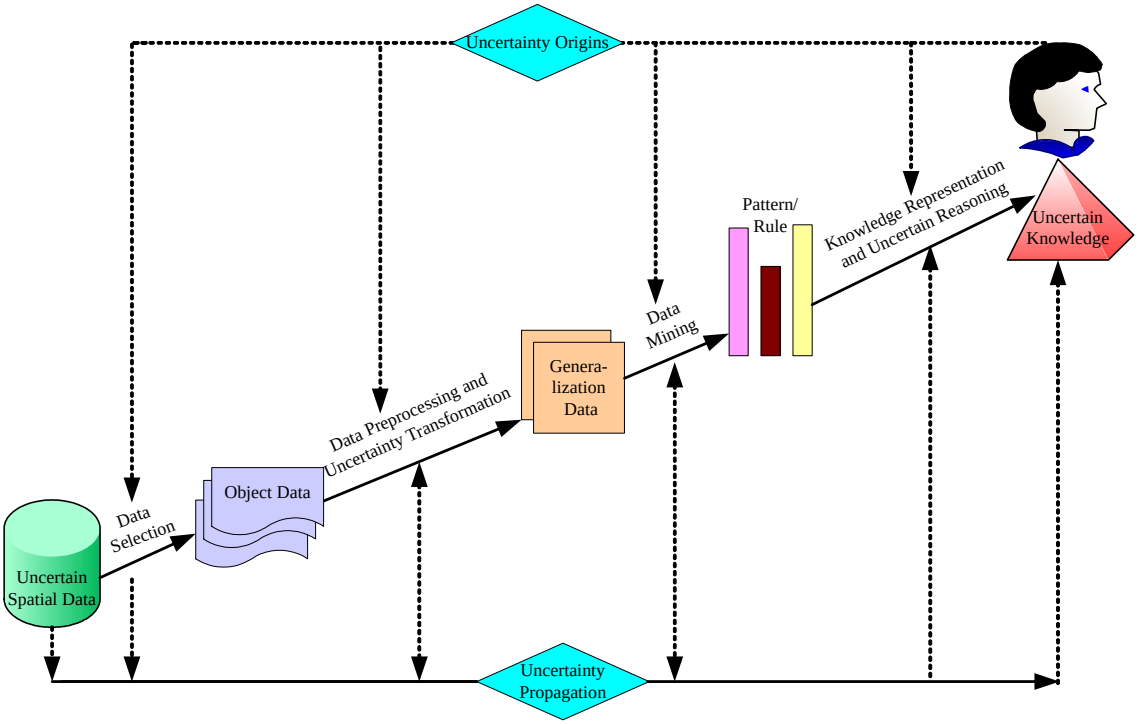

Figure 2. Uncertainties and its propagation in the process of spatial knowledge discovery

4. SPATIAL KNOWLEDGE DISCOVERY BASED ON FUZZY EVIDENCE THEORY

4.1 About the Evidence Theory

Evidence theory, namely Dempster-Shafer (D-S) theory, aims to provide a theory of partial belief, which extend traditional probability theory. Firstly, we should briefly introduce the evidence theory.

The frame of discernment, Θ, is the set of mutually exclusive and exhaustive propositions of interest. Defined on the set of subsets of Θ is the basic probability assignment or mass function, *m*, that associates with every subset of Θ a degree of belief that lies in the interval [0, 1]. Mathematically, m is defined as follows:

$$m : 2^{\Theta} \rightarrow [0,1] \tag{2}$$

such that:

$$m(\Phi) = 0 \tag{3}$$

$$\sum_{x \subseteq \Theta} m(x) = 1 \tag{4}$$

A *Belief function*:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{5}$$

A *Plausibility function:*

$$Pl(A) = 1 - Bel(\neg A) = \sum_{B \cap A \neq \Phi} m(B) \tag{6}$$

Thus, at any given time the interval [$Bel(A), Pl(A)$] defines the uncertainty associated with A. While $Bel(A)$ is the definite support for A, $Pl(A)$ is the extent to which the evidence at that present time fails to refute A.

When identify an object, all evidences associated with the object must be combined. The Rule for the combination of evidence (the Orthogonal Sum, $\oplus$):

$$(m_1 \oplus m_2)(c) = \frac{\sum_{A \cap B} m_1(A) m_2(B)}{1 - \sum_{A \cap B = \Phi} m_1(A) m_2(B)} \tag{7}$$

Evidence theory has been applied abroad in artificial intelligence field. Anand (1996) applied evidence theory to knowledge discovery by combination operator. More contents about evidence theory may refer to Shafer (1976).

4.2 Fuzzy Evidence Theoretic Approaches for Spatial Knowledge Discovery

Evidence theory can only process the uncertainty cased by randomness. In fact, spatial data and knowledge include both randomness and vagueness. When considering randomness and vagueness simultaneously of spatial data and knowledge, it may take account into combining fuzzy theory with probability theory. Fuzzy evidence theory can process the two kinds of uncertainty integrating uncertain reasoning.

Herein, we consider spatial knowledge discovery as uncertain reasoning process based on fuzzy evidence theory, which include soft discretization of spatial data and uncertainty transformation between quantitative data and qualitative concept by applying Gaussian fuzzy function, uncertain knowledge discovery and representation by fuzzy D-S belief structure and uncertain reasoning.

A fuzzy D-S belief structure is one of D-S belief structure that the focus element is the fuzzy sets. When apply combination operator to combine two fuzzy belief structures, only to apply fuzzy sets operation. For example, m_1 and m_2 are the two fuzzy D-S belief structures in he frame of discernment, Θ. Thus, the new fuzzy belief structure is:

$$m = m_1 \cup m_2 \tag{8}$$

where the focus element is: $F_k = A_i \cup B_j$, the membership function is $\max(\mu_{Ai(x)}, \mu_{Bj(x)})$ and $m(F_k) = m_1(A_i) * m_2(B_j)$.

The fuzzy rule based on fuzzy D-S belief structure is as follows:

$$R(r): If\ (X_1 is A_r^1)\ and\ (X_2 is A_r^2)\ \cdots and\ (X_n is A_r^n)$$
$$\text{then } Y \text{ is } m_r \tag{9}$$

where m_r is a fuzzy belief structure with focus element $B_{rj} \in \{B_1, B_2, \cdots, B_p\}(j = 1, \cdots p)$, it is a fuzzy partition of output space. $m_r(B_{rj})$ is the basic probability assignment of B_{rj}, which indicate that the belief degree (probability) of the $r^{th} \subset B_{rj}$. Therefore, the output of rules is uncertain. This kind of rule form should take account into the propagation of evidences in knowledge discovery integrating uncertain reasoning.

Suppose that $X_i = x_i, i = 1, \cdots n$ is a group of input values. Then the knowledge discovery and reasoning process based on D-S belief structure is as follows:

(1) Compute the activation degree of every rule τ_r:

$$\tau_r = \bigwedge_i [A_r^i(x_i) \text{ or } \prod_i [A_r^i(x_i)] \tag{10}$$

(2) Make certain the output of single rule according to activation degree and rule consequent:

$$\hat{m}_r = \varphi(\tau_r, m_r) \tag{11}$$

where φ is containing operator; $\hat{m}_r$ is a fuzzy belief structure on V and its focus element is F_{rj}; F_{rj} is the fuzzy subset of the output space V and its definition is:

$$\mu_{Frj}(y) = \tau_r \wedge \mu_{Brj}(y) \text{ or } \mu_{Frj}(y) = \tau_r * \mu_{Brj}(y) \tag{12}$$

B_{rj} is a focus element of m_r; the basic probability associated to F_{rj} is:

$$\hat{m}_r(F_{rj}) = m_r(B_{rj}) \tag{13}$$

(3) Output the combination rules, adopt no-null combination operator to combine fuzzy belief structure:

$$m = \bigoplus_{r=1}^{M} \hat{m}_r \tag{14}$$

to every set $F_k = \{F_{1, j_1^k}, \cdots, F_{M, j_M^K}\}$, where F_{r, j_r^K} is a focus element of $\hat{m}_r$, which lies in a focus element:

$$E_k = \nabla_{r=1}^{M} F_{r, j_r^K} \tag{15}$$

When operator is *average*, E_k may be defined as:

$$\mu_{E_k}(y) = \frac{1}{M} \sum_{r=1}^{M} \mu_{Fr, j_r^k}(y) \tag{16}$$

and its basic probability is:

$$m(E_k) = \prod_{r=1}^{M} \hat{m}_r(F_{r, j_r^k}) \tag{17}$$

So the output is a fuzzy D-S belief structure m with focus element E_k ($k = 1, \cdots p^M$).

(4) Anti-fuzzy to fuzzy belief structure m:

$$\bar{y} = \sum_{k=1}^{p^M} \bar{y}_k m(E_k) \tag{18}$$

where $\bar{y}_k$ is anti-fuzzy value of focus element E_k:

$$\bar{y} = \frac{\sum y \mu_{E_k(y)}}{\sum \mu_{E_k(y)}} \tag{19}$$

Here, we adopt Gaussian function as the membership function of fuzzy sets in input and output space.

$$\exp \frac{(x-c)^2}{-2\sigma^2} \tag{20}$$

Suppose that $[l, u]$ is discussion field of variable and l, u is minimum and maximum value respectively of every dimension

in training spatial data sets. If $[l,u]$ is divided into N fuzzy areas $A_i (i=1,\cdots n)$ and fuzzy area is represented by Gaussian membership function, A_i is defined as:

$$\mu_{A_i}(x) = \exp\frac{(x-c_i)^2}{-2\sigma_i{}^2},\ \forall x \in [l,u] \qquad (21)$$

Where $c_i = l + \frac{(i-1)(u-l)}{N-1}$, $\sigma_i = \frac{u-l}{2(N-1)\sqrt{\ln 4}}$,

$\mu_{A_i}(c_i) = 1$ and $\mu_{A_i}(\frac{c_i + c_i + 1}{2}) = \mu_{A_{i+1}}(\frac{c_i + c_i + 1}{2}) = 0.5$

For example, if the discussion field is [0,1], Figure 3 is an instance that the discussion field is partitioned to three Gaussian fuzzy subsets.

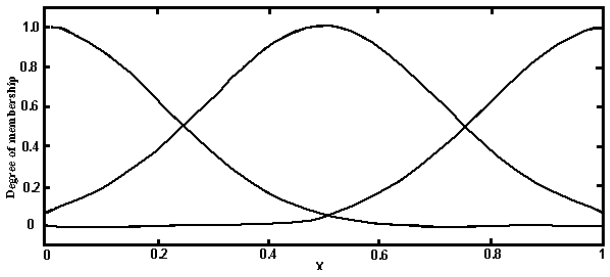

Figure 3. Gaussian fuzzy subsets

5. CONCLUSION

It is our mind in this research to achieve both of objectives. Firstly, the quality of spatial knowledge discovery can be improved by analyzing the uncertainties and its characteristics in each phase of spatial knowledge discovery and finding efficient method to reduce its uncertainties. Secondly, although the uncertainties of spatial knowledge discovery cannot be completely eliminated, the uncertainty of spatial knowledge discovery results can be represented in order to make use of the knowledge discovered in spatial knowledge discovery. In this paper, we briefly analyze the uncertainties in spatial data and spatial knowledge discovery. Then, the framework of spatial knowledge discovery based on fuzzy evidence theory was constructed. Further work aims at experimental study based upper theories and methods, visualization of spatial knowledge and uncertainty propagation law in spatial knowledge discovery is also our interesting.

ACKNOWLEDGES

The work described in this paper was supported by the funds from National Natural Science Foundation of China (Project No.60275021).

REFERENCES

Chrisman N. R., 1982. A theory of cartography error and its measurement in digital database. In: *Proceedings of Auto-Carto5*, pp.159-168.

Dutton G., 1992. Handling positional uncertainty in spatial database. In: *Proceedings of 5th International Symposium on Spatial Data Mining*, pp. 460~469.

Ehlschlager C.R., 2000. Representing uncertainty of area class maps with a correlated inter-map cell swapping heuristic. *Computers, Enviroment and Urban Systems*, 24, pp.451~469.

FGDC(Federal Geographic Data Committee),1998. *Content standard for digital geospatial metadata*. FGDC-STD-001-1998, National Technical Information Services, Computer Products Office, Springfield, Virginia, USA.

Fisher P.F., 2003. Data quality and uncertainty: ships passing in the light. In: *Proceedings of The 2nd International Symposium on Spatial Data Quality'2003*, Edited by Shi W.Z., Goodchild M.F., Fisher P.F, pp.17-22.

Goodchild M. F., 1991. Issuees of quality and uncertainty. In : *Advances In Cartography* ,edited by Muller J.C., London, Elsevier, pp.113-139

Goodchild M.F, 2003. Models for uncertainty in area-class maps. In: *Proceedings of The 2nd International Symposium on Spatial Data Quality'2003*, Edited by Shi W.Z., Goodchild M.F., Fisher P.F., pp.1-9.

G. Shafer, 1976. *A Mathematical Theory of Evidence.* Princeton University Press.

Li D.R., Wang S.L., Li D.Y. and Wang X.Z., 2002. Theories and technologies of spatial data knowledge discovery. *Geomatics and Information Science of Wuhan University*, 27(3), pp.221~233.

Liu W.B., Den M. and Xia Z.G., 1999. Analysis of uncertainties of attributes in vector GIS. *Acta Geodaetica et Cartographica Sinica*, 29(1), pp.76~81.

Mikhail E. M. and Ackerman F., 1976. *Observations and Least Squares*. IEP-A Dun-Donnelley Publisher, New York.

Miller, Harvey J., Han, Jiawei, 2001. *Geographic data mining and knowledge discovery*. Taylor & Francis.

Openshaw, S., 1989. Learning to live with errors in spatial database. In: *Accuracy of Spatial Database*, Taylor & Francis Ltd, New York, pp.263-276.

S S.anand, D A. Bell and J G. Hughes, 1996. EDM: A general framework for data mining based on evidence theory. *Data & Knowledge Engineering*, 18, pp.189-223.

Shi W.Z., Wang S.L., 2002. Further development of theories and methods on attribute uncertainty in GIS. *Journal of Remote Sensing*, 6(5) , pp.393~400.

Shi W.Z., 2000. *Theory and Methods for Handling Errors in Spatial data*. Science Press, Beijing.

Zhang J.X., Du D.S., 1999. Field based models for positional and attribute uncertainty. *Acta Geodaetica et* Cartographica Sinica, 28(3), pp. 244~249.