Urbanization prediction with an ART-MMAP neural network based spatio-temporal data mining method

Weiguo Liu[a], Karen C. Seto[b] and Zhanli Sun[c]

A: ACI Worldwide Inc., 100 Amaral St., Riverside, RI 02915, Email: wgliu@crsa.bu.edu
B: Institute for International Studies, Encina Hall, Room E413, Stanford University, Stanford, CA 94305-6055. Email: kseto@stanford.edu.
C: Department of Urban and Regional Planning, University of Illinois at Urbana and Champaign, 611 Taft Drive, Champaign, IL 61820 Email: sun2@uiuc.edu

KEY WORDS: ART-MMAP, data mining, urban growth prediction, land use change, neural network

ABSTRACT:

Data mining methods have been widely and successfully used in many fields in the last decade. And geographic knowledge discovery and spatial data mining also have attracted more attentions recently. This paper presents an ART-MMAP neural network based spatio-temporal data mining method to simulate and predict urban expansion. The spatial matrices derived from different urban related features, i.e. transportation, land use, topography, were directly used as inputs to the neural network model for learning. The trained network was then applied to research region to predict the land use change to urban. The learning and prediction process are automatic and free of intervention. The method has been successfully validated with the urban growth prediction at St. Louis region at Missouri, USA.

1. INTRODUCTION

The huge amount of data has posed great challenges to traditional data analysis method for information and knowledge extraction. Data mining, referring to the application of low-level algorithms for revealing hidden information in a database (Klosgen and Zytkow, 1996), has been emerging as a new research field and a new technology in the last decade. Data mining represents the interdisciplinary research of several fields, including machine learning, neural network, statistics, database, visualization and information theory (Koperski et al., 1996).

Similar to the fields using relational and transactional databases, geography has changed from a data-poor and computation poor to a data-rich and computation-rich environment. Traditional spatial analytical methods can't be used to discover the hidden information from huge amount spatial related dataset. Spatial data mining has attracted attentions in recent research (Miller and Han, 2001). It refers to the discovery and extraction of implicit information, spatial relationships or spatial distribution patterns from spatial databases (Koperski et al., 1996). Spatial data mining can be used to understand spatial data, discover spatial relationships and relationships between spatial data and non-spatial data etc. It has wide applications in geographical information systems, remote sensing and many other areas related to spatial data.

Among different spatial data mining algorithms, spatial classification aims to assign an object to a class from a given set of classes based on both spatial and non-spatial attribute values of the object. Decision tree classifiers (Ester et al., 1997) and neural networks (Gopal et al., 2001; Liu et al, 2001) have been widely used as base classification methods which can handle both spatial and non-spatial data. Different to traditional classification methods, spatial classification will explicitly involve spatial related features or metrics (Koperski et al., 1996; Ester et al., 2001). In this paper, we present an ART-MMAP neural network based spatial classification method to process the multi-temporal urban growth data for predicting future urban expansion.

2. URBAN ANALYSIS REVIEW

Urban growth and the resultant sprawling patterns of US communities are causing social, economic and environmental strain (Schmidt, 1998), which raise great concern from federal government to local public sectors. However, urban growth is a complex issue, which can be best understood through dynamic modelling. While land use change is generally considered as the signature of urban growth, land use change modelling has been the focus of urban growth research. Very recently, computer-based urban system simulation models are increasingly employed to forecast and evaluate land-use change (Batty and Xie, 1994; Birkin, 1994; Engelen et al., 1995; Landis, 1994). This spatial dynamic modelling approach enables planners to view and analyse the future of their decisions and policies even before they are put in action. Therefore, it can help improve our fundamental understanding and communication of the dynamics of land-use transformation and the complex interactions between urban change and sustainable systems (Deal, 2001). Nowadays, spatial dynamic modelling techniques are considered essential to Planning Support System (PSS) after being overshadowed by GIS applications since the 1980s (Hopkins, 1999; Kammeier, 1999).

To date, however, spatial dynamic urban modelling is still in its infancy. Due to the extreme complexity of urban system, few, if any, models have been built which truly represent the dynamics of urban growth and which can provide consistent results with what we know about such changes (Maria de Almeida et al., 2003). Consequently, such models are barely operational and therefore are rarely used to assist urban planning practice. Nevertheless, Cellular Automata (CA) and Agent Based Models, representing a different approach to the traditional top-down approaches, are emerging as promising model tools. Agent Based Models possess characteristics that are analogous to those of cellular automata. Most of Agent Based Models can be considered as extended CA models with “smart” cells, which can communicate with other cells and environment, make decisions based on information received and sometimes move across the space. Therefore, there is no essential difference between CA models and Agent Based Models. Here we focus our reviews on CA based land use change models.

Cellular Automata are discrete dynamical systems whose behavior is completely specified in terms of a local relation. It is embedded with a spatial dynamic feature, which makes CA a natural tool for spatial modelling. CA application in geographic modelling dates back to the spatial diffusion model developed by Hagerstrand (Hagerstrand, 1967), which essentially a stochastic CA although he didn’t even use the term CA. Geographer Tobler (1979) first defined CA as geographical models although he believed that some CA are too simple to be usefully applied. Later on, the implication of CA to geographic modelling, including advantages and theoretical obstacles of applying CA to geographic modelling, was explored theoretically (Batty et al., 1997; Couclelis, 1985; 1987; 1997). CA is very appealing to geographic modellers because 1) CA based model is simple and intuitive, yet capable to simulate self-organizing complex system; 2) The natural born spatial dynamic feature enables modelling spatial dynamic system in extreme spatial detail and spatial explicitly; 3) The cellular structure of CA has natural affinity with raster data format of Remote Sensing images and GIS grid map. CA model can be easily integrated with GIS through generalization of map algebra (Takeyama and Couclelis, 1997); 4) The bottom-up approach of CA provides a new strategy of geographic modelling; 5) CA is computational model running in parallel which fits the high-performance geo-computation.
Since then, CA application in geography has been experiencing exponential growth, especially in urban land-use simulation. Batty was one of the earliest geographer who sketched the general framework of CA-based urban models (Batty and Xie, 1994). An integrated platform, named DUEM, designed for geographic CA exploration was also developed by Batty and his group (Batty et al., 1999). Engelen used CA to model urban land-use dynamics to forecast climate change on a small island setting (Engelen et al., 1995). Wu presented a model that also included user-decisions to determine model outcomes (Wu and Webster, 1998). White’s St. Lucia model (White and Engelen, 1997) is an example of high-resolution CA modelling of urban land-use dynamics and an attempt to use the standard non-spatial models of regional economics and demographics, as well as a simple model of environmental change for predicting the demand for future agricultural, residential, and commercial/industrial land-uses. An urban growth model of the San Francisco Bay Area (Clarke and Gaydos, 1998) is another example of using relatively simple rules in the CA environment to simulate urban growth patterns. Li and Yeh integrated neural-network and CA in GIS platform and successfully applied to urban land-use change simulation in Guangdong, China (Li and Yeh, 2002).

Although a large number of models have been proposed and built over the last twenty years, CA based land-use modelling technique is still far from being mature. Despite the flexibility of the CA approach, limitations remain (Torrens and O'Sullivan, 2001). The hypothetical urban forms emerging from CA models with surprisingly simple local transition rules are certainly plausible. However, urban system evolves in a much more complex way in reality. The current CA-based urban models are just too simple to capture the richness of urban systems. Consequently, very few CA models are operational and are used as productive tool to support regional planning practice.

To build useful models, modellers try to extend the concept of CA, and also integrate a diversity of models, such as traditional regional social-economic models (White and Engelen, 1997; Wu and Martin, 2002). In this paper, instead of using CA based models, we present a neural network based model to learn the urban growth patterns based on historical urban data and predict the future urban growth.

3. ART-MMAP NEURAL NETWORK

The Adaptive Resonance Theory (ART) family of pattern recognition algorithms was developed by Carpenter and Grossberg (1991). ART is a match-based learning system, the major feature of which is its ability to solve the ‘stability-plasticity dilemma’ or ‘serial learning problem’, where successive training of a network interferes with previously acquired knowledge. Among the ART family models, fuzzy ARTMAP is a supervised learning system that has been used widely in many fields. A comprehensive description of the model is detailed in Carpenter et al. (1992).

ART-MMAP, an extension of ARTMAP, decreases the effect of category proliferation in the testing process for mixture analysis. The ART-MMAP model keeps the learning process of the ARTMAP model and changes the testing process. During the testing process, ARTMAP selects a category to each test sample using the Winner-Take-All (WTA) rule. Instead of picking one winner, ART-MMAP selects winners (ART_a) based on one predefined threshold parameter - τ. The categories with activation value larger than the threshold value are selected. If none of them are selected, the WTA rule is activated. This winner selection strategy provides an enhanced interpolation function which is based on a weighted summation operator. ART-MMAP model overcomes the limitation of class category of the ARTMAP model and increases the prediction accuracy as well (Liu et al., 2004).

4. DATA

4.1 Study area description

Spanning parts of the states of Missouri and Illinois on both side of Mississippi Rive, the great St. Louis metropolitan region (Figure 1) includes ten counties. This area is about 120 miles from east to west and about 90 miles from north to south. It accounts for a little more than 30 million grid cells at 30m * 30m spatial resolution. Like most other older metropolitans, St. Louis faces great challenge of sustainable growth. With relatively slow population growth, even negative growth in urban core, city is continuing to sprawl. St. Louis metropolitan region is already the third largest in the amount of land that it covers while ranks 14th in term of population. Under such a condition, the prediction and planning of urban growth becomes very important for St. Louis region.

4.2 Land Use Factors

To simulate the urban growth and then make a prediction, we need several important spatial related features extracted from land use factors, also called land use drivers, consisting of social, economic, transportation and biophysical factors affecting land use change. In reality, there are numerous factors affecting urban land use change more or less. Apparently, it is impossible to incorporate all these factors in one land use model. Some most significant factors, which can be varied in different study areas, are selected. Besides the non-spatial factors including most of social economic factors, which define the regional demand, the following factors are incorporated in our model to describe the land use transition possibility of each cell:

1. Cities Attractor: A gravity model is used to emulate the cities attractiveness.

An attractiveness index of each cell is calculated as:

$$A_{cities} = \sum_i {Pop_i}/{TT_i} \tag{1}$$

Where A_{cities} is the attractiveness index of a cell; Pop_i is the population of city i; TT_i is the travel time from a cell to city i.

2. Employment Attractor: Similar to cities attractor, employment attractor represent the attractiveness impact of employment centers.
3. Neighbour: It describes the number of urbanized cells in the neighbourhood (i.e. 3X3 window). The new growth more likely happens near developed or developing neighbourhood.
4. DEM/Slope: Suitability of development differs on various degrees of slope.
5. Water Proximity: distance to lakes, rivers.
6. Forest Proximity: distance to forest area.
7. Transportation Proximity Factors: travel times to transportation facilities.
   a. Proximity to interstate ramps
   b. Proximity to state highways
   c. Proximity to major roads
   d. Proximity to major road intersections

With the extracted spatial features, each cell of the study area was assigned one vector that consists of the spatial value of each factor, land use type in 1992 and land use type in 2000 (urban or non-urban land use type). The vectors will be used to describe the urban growth pattern through space and time. For the purpose of learning, we first selected 15931 samples from whole study area with a sample ratio 900:1. Then whole research area was used as testing data.

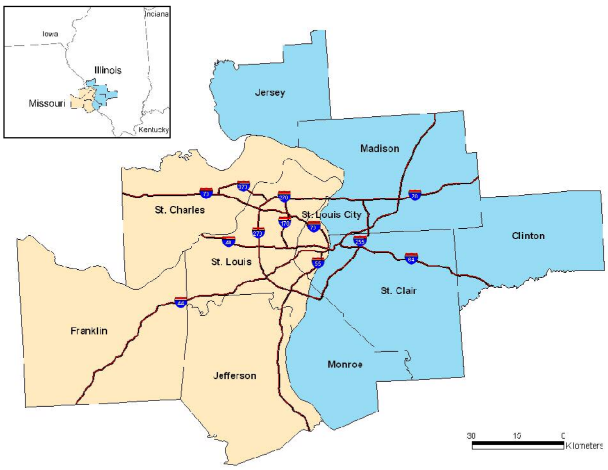


Figure 1. Study area: large metropolitan ST. Louis

5. RESULTS AND DISCUSSION

With this dataset, the parameters of the ART-MMAP model were set as: ρ_a =0.9, ρ_b =1.0 and τ = 0.98 (threshold value for selecting winning F_2 nodes). After learning the 15931 training samples, the ART-MMAP network was applied to the whole research area for prediction accuracy validation.

5.1 Model performance and validation

The ART-MMAP network based predictive model assesses the likelihood of urbanization by assigning individual pixels a score ranging from 0 to 100. Higher scores correlate to an increased probability of changing from other land use type to urban. The predicted class label can be assigned through setting a score threshold for the model. Scores equal to or greater than the score threshold are flagged as urbanized pixels. The choice of score threshold determines the number of pixels to be predicted as urban. As the score threshold is lowered, both the total number of pixels predicted as urban and the number of urban pixels predicted correctly increases. The performance of a predictive model is characterized by plots of the percentage of urban pixels detected versus the false positive ratio. Here the false positive ratio is calculated as the ratio between the number of non-urbanized samples who were classified as urban and the number of detected urbanized samples at certain score threshold.

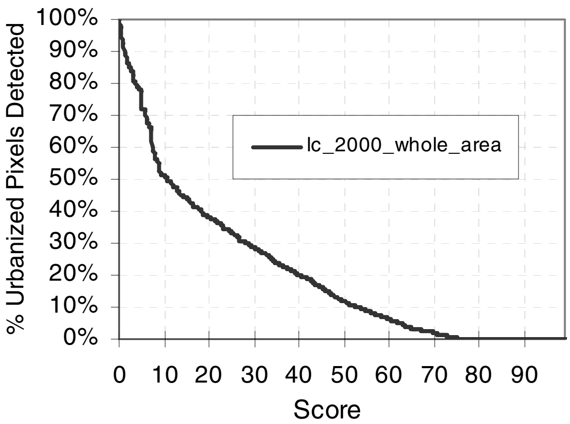


Figure 2. Score vs. percentage of urbanized pixels detected

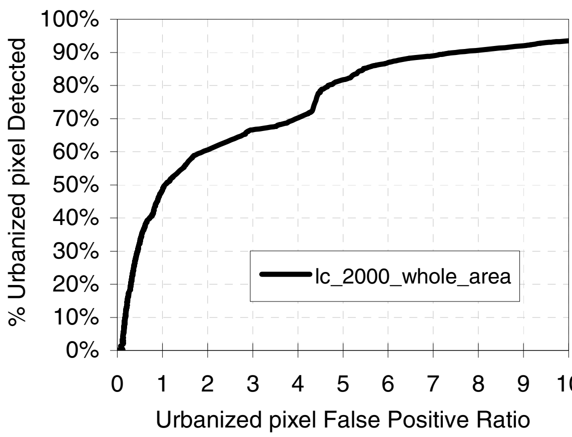


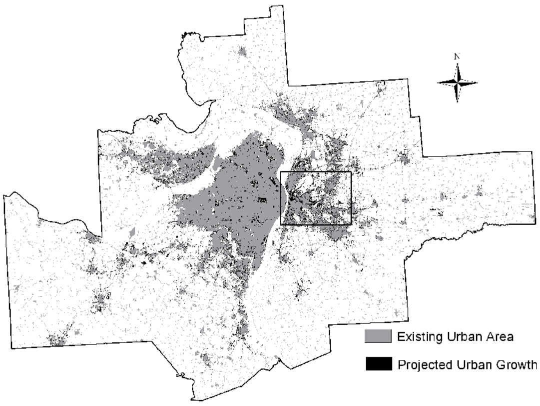
Figure 3. Urbanized pixel false positive ratio vs. percentage of urbanized pixel detected

Figure 2 shows the relationship between the cutting off score threshold and the percentage of urbanized samples detected. For example, if we set up the threshold score for urbanization is 12, approximately 50% of urbanized pixels are detected. Figure 3 shows the relationship between the percentage of detected urbanized samples and false positive ratio.

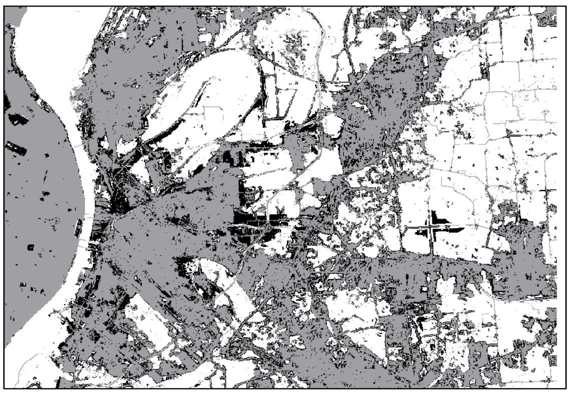
Combining Figure 3 and 2, we can figure out that at curtain score threshold, the percentage of detected urbanized pixels and the number of false classified non-urbanized pixels. To view the performance with traditional way, the classification error matrix built at score threshold 12 is shown in Table 1. With a total prediction accuracy of 94%, the ART-MMAP based data mining model successfully predicted 50% of urbanized pixels with a false prediction of similar number of non-urbanized cells (437027). Changing the cutting off score threshold, we can get different accuracy. In order to detect more urbanized pixels, we may reduce the cutting off threshold score. However, this will introduce more misclassified non-urbanized pixels as urban.

Class	Urbanized	Non-urban	Total pixels	Producer's accuracy
Urbanized	402520	396410	798930	0.504
Non-urban	437027	12632090	13069117	0.967
Total pixels	839547	13028500	Total Accuracy: 0.94	
User's Accuracy	0.479	0.970		

Table 1. Classification error matrix of the whole study area



(a) Regional urban growth overview



(b) Urban growth zoom in

Figure 4. Model simulation result (2000-2008)

5.2 Simulation results of year 2008

Based on the land use map of year 2000, we applied the trained ART-MMAP neural network model to predict the urban growth of year 2008 to evaluate its predictive performance. Since transportation, social and economic factors did not change much from year 1992 to year 2000, we keep the same value of these relative factors for each cell. The neighborhood value of urbanized pixels was recalculated with land use data of year 2000.

Figure 4 displays the predicted urban growth area in year 2008. The overall spatial pattern of the projected urban growth is quite reasonable. Most growth takes place around city peripheral region, and is clustered around highway and major road intersections. Those areas are under growth pressure according to the local planners.

The transition statistics (in Table 2) shows the number of pixels of each non-urban class changing to urban area from year 2000 to 2008. Approximately 5% of herbaceous planted area changes into urban area, which is the majority of the urbanized pixels. Land use type Barren and Forested upland are another two important types changing into urban.
each cell was recalculated with land use data of year 2000.

Class Name	ClassID	Total pixels	Pixel changed	Change percent
Water	1	404567	3141	0.78%
Urban	2	2650997	0	0.00%
Barren	3	102742	19161	18.65%
Forested Upland	4	4151988	87335	2.10%
Shrubland	5	11	0	0.00%
Herbaceous Upland Natural/Semi-natural Vegetation	6	161269	7809	4.84%
Herbaceous Planted/Cultivated	7	7690045	378031	4.92%
Wetland	8	578757	2132	0.37%

Table 2. Land use transition statistics (year 2008)

6. CONCLUSION

Along the changing from poor-data to rich-data environment in the field of geography, spatial data mining has become more interesting to many researchers. In this paper, we present an ART-MMAP neural network based sptatio-temporal data mining method to simulate the future urban expansion and further to predict urban growth. With the training based on the multi-temporal urban growth data, the ART-MMAP model can automatically predict the probability of urbanization of each pixel in the near future. Since the prediction is score based, we can get different urban expansion maps with setting different cut-off score threshold. Considering the goal of the urban prediction, the prediction accuracy of the St. Louis data set is pretty good although the model only detected 50% of urbanized pixels and also misidentified the similar amount of non-urban pixels. Realistically, no model will very accurately predict the urban growth of the future. The predicted urbanized area with this model will provide a base probability map for urban planning. The goal of this model is to help planners answer what-if questions by implementing planning scenarios. By exploring these scenarios, it will significantly enhance planners' insights into future land use and its impact. Although this model probably will not make decisions for planners, neither will make planners smarter, it certainly will help them make smarter decisions.

In this research, the training and testing data set were selected at the same time periods (1992 – 2000). To understand the model performance more accurately, the trained ART-MMAP network with one time period data (1992-2000) will be applied to later time period, i.e. 2000-2004, for performance validation in the future research. Also, the integration of the prediction model and Cellular Automata based models or Agent based models will be more interesting for realistic urban planning. It may also shed some lights on urban simulation and urban modeling research.

REFERENCES:

Batty, M., Xie, Y., 1994. From Cells To Cities. *Environment and Planning B-Planning & Design*, 21, S31-S38.

Batty, M., Couclelis, H., Eichen, M., 1997. Urban systems as cellular automata. *Environment and Planning B-Planning & Design*, 24, 159-164.

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

Birkin, M., 1994. Urban Landscape Dynamics - A Multilevel Innovation Process - Montanari,A, Curdes,G, Forsyth,L. *Environment and Planning A*, 26, 1480-1480.

Carpenter, G., Grossberg, S. (1991). *Pattern Recognition by Self-organizing Neural networks*. Cambridge, MA: MIT Press.

Carpenter, G., Grossberg, S., Markuzon, S., Martens, N., Reynolds, J., Rosen, D. (1992). Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps, *IEEE Transaction on Neural Networks*, 3, pp.698-713.

Clarke, K.C., Gaydos, L.J., 1998. Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. *International Journal ff Geographical Information Science*, 12, 699-714.

Couclelis, H., 1985. Cellular Worlds - A Framework For Modelling Micro-Macro Dynamics. Environment And Planning A 17, 585-596.

Couclelis, H., 1987. Cellular-Dynamics - How Individual Decisions Lead To Global Urban Change. *European Journal ff Operational Research*, 30, 344-346.

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

Deal, B., 2001. Ecological urban dynamics: the convergence of spatial modelling and sustainability. *Building Research and Information*, 29, 381-393.

Engelen, G., White, R., Uljee, I., Drazan, P., 1995. Using Cellular-Automata For Integrated Modelling Of Socio-Environmental Systems. *Environmental Monitoring and Assessment*, 34, 203-214.

Ester, M., Kriegerl, H. P.,, Sander J. (1997). Spatial data mining: a database approach, in *Proceedings of the fifth international symposium on large spatial databases*, Berlin, Germany: 47:66.

Ester, M., Kriegerl, H. P.,, Sander J. (1997). Algorithms and applications for spatial data mining, in Miller and Han (eds) *Geographic data mining and knowledge discovery*. Taylor and Francis, London and New York. Pp.161-187.

Gopal, S., Liu, W.,, Woodcock, C., 2001. Visualization based on the Fuzzy ARTMAP neural network for mining remotely sensed data, in Miller and Han (eds) *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, London, United Kingdom, pp.315-336.

Hagerstrand, T., 1967. *Innovation diffusion as a spatial process*. University of Chicago Press, Chicago.

Kammeier, H.D., 1999. New tools for spatial analysis and planning as components of an incremental planning-support system. *Environment and Planning B-Planning & Design,* 26, 365-380.

Hopkins, L.D., 1999. Structure of a planning support system for urban development. *Environment and Planning B-Planning & Design*, 26, 333-343.

Klosgen, W., Zytkow, J. M. (1996). Knowledge discovery in databases terminology, in Fayyar, E.M., Piatetsky-Shapiro, G., Smyth, P. and Ulthurusamy, R. (eds) *Avances in Knowledge discovery and data mining*. Combridge, MA: MIT Press. Pp.573-592.

Koperski, K., Adhikary J., Han, J. (1996). Knowledge discovery in spatial databases: progress and challenges. In *Proceedings of the SIGMID workshop on research issue in data mining and knowledge discovery*, Technical report 96-08. University of British Columbia, Vancouver, Canada.

Landis, J.D., 1994. The California Urban Futures Model - A New-Generation Of Metropolitan Simulation-Models. *Environment and Planning B-Planning & Design*, 21, 399-420.

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

Liu, W., Seto, K., Wu, E., Gopal, S., Woodcock, C., 2004. ART-MMAP: a neural network approach to subpixel classification, *IEEE Transactions on Geoscience and Remote Sensing*, 42(9), pp.1976-1983.

Liu, W., Gopal, S., Woodcock, C., 2001. Spatial data mining for classification, visualization and interpretation with ARTMAP neural network, in R. Grossman (eds) *Data Mining for Scientific and Engineering Applications*, Kluwer Academic Publishers, Netherlands, pp.205-222.

Maria de Almeida, C., Batty, M., Vieira Monteiro, A.M., Camara, G., Soares-Filho, B.S., Cerqueira, G.C., Pennachin, C.L., 2003. Stochastic cellular automata modelling of urban land use dynamics: empirical development and estimation. *Computers, Environment and Urban Systems*, 27, 481-509.

Miller, H., Han, J. (2001). *Geographic data mining and knowledge discovery*. Taylor and Francis, London and New York. pp.1-31.

Schmidt, C.W., 1998. The Specter of Sprawl. Environmental *Health Perspectives*, 106, 274-279.

Takeyama, M., Couclelis, H., 1997. Map dynamics: Integrating cellular automata and GIS through Geo-Algebra. *International Journal of Geographical Information Science*, 11, 73-91.

Tobler, W., 1979. Cellular geography. In: Gale, G., Olsson, S. (Eds.), *Philosophy in geography*, Reidel, Dordrecht, pp. 379-386.

Torrens, P.M., O'Sullivan, D., 2001. Cellular automata and urban simulation: where do we go from here? *Environment and Planning B-Planning & Design*, 28, 163-168.

Von Neumann, J., 1966. *Theory of Self-Reproducing Automata*. University of Illinois Press, Urbana, IL.

White, R., Engelen, G., 1997. Cellular automata as the basis of integrated dynamic regional modelling. *Environment and Planning B-Planning & Design*, 24, 235-246.

Wolfram, S., 2002. *A New Kind of Science*. Wolfram Media, Inc., Champaign, IL.

Wu, F., Webster, C.J., 1998. Simulation of land development through the integration of cellular automata and multicriteria evaluation. *Environment and Planning B-Planning & Design*, 25, 103-126.

Wu, F.L., Martin, D., 2002. Urban expansion simulation of Southeast England using population surface modelling and cellular automata. *Environment and Planning A*, 34, 1855-1876.