INT. J. GEOGRAPHICAL INFORMATION SCIENCE
VOL. 18, NO. 8, DECEMBER 2004, 723–744

Research Article

Data mining of cellular automata's transition rules

XIA LI

School of Geography and Planning, Sun Yat-sen University, 135 West
Xingang Rd., Guangzhou 510275, P.R. China; e-mail: lixia@graduate.hku.hk;
gplx@zsu.edu.cn

and ANTHONY GAR-ON YEH

Centre of Urban Planning and Environmental Management, The University
of Hong Kong, Pokfulam Road, Hong Kong SAR, P.R. China; e-mail:
hdxugoy@hkucc.hku.hk

(*Received 23 May 2003; accepted 13 February 2004*)

Abstract. This paper presents a new method to discover knowledge for geographical cellular automata (CA) by using a data-mining technique. CA have the ability to simulate complex geographical phenomena. Very few studies have been carried out on how to determine and validate the transition rules of CA from observed data. The transition rules of traditional CA are usually expressed by mathematical equations. This paper demonstrates that the explicit transition rules of CA can be automatically reconstructed through the rule induction procedure of data mining. The explicit transition rules are more intuitive to decision-makers. The transition rules are obtained by applying data-mining techniques to spatial data. The proposed method can reduce the uncertainties in defining transition rules and help to generate more reliable simulation results.

1. Introduction

In recent years, there has been an increasing number of studies on the development of geographical cellular automata (CA) for simulating complex systems. CA have been applied to the simulation of wildfire propagation (Clarke *et al.* 1994), population dynamics (Couclelis 1988), and urban evolution and land-use changes (White and Engelen 1993, Batty and Xie 1994). They are also used to generate idealized or optimized urban forms for land-use planning (Li and Yeh 2000, Yeh and Li 2001a). CA were originated from the research of self-reproducing systems done by Ulam and Von Neumann in the 1940s (White and Engelen 1993). The 'game of life' developed in 1970 by the mathematician Conway can also be regarded as an explicit CA game (Gardner 1971, Portugali 2000). However, much influence on later developments of CA techniques can be attributed to Wolfram's (1984) studies, which demonstrated that the origins of the complexity in natural systems could be investigated through dynamic models called 'cellular automata'.

CA have great potential for geographical applications because of their strong

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810410001705325

modelling capabilities. Tobler (1979) was perhaps the first to recognize the advantages of CA in solving geographical problems (White and Engelen 1993). In his cellular space model, the state of a cell is determined by the states of a set of 'neighbour' cells according to some uniform location-independent rules. The basic principle of such models is to use a cell-space representation to realize spatial dynamics. The simulation of population dynamics is a good demonstration of CA's capabilities in modelling complex natural systems. Couclelis (1988) has successfully generated a variety of different spatio-temporal structures of rodent population by using a very simple one-dimensional cellular automaton. Her research indicated that the whole range of complex and apparently bizarre population dynamics can be easily reproduced by the simple cellular automaton.

Urban simulation may be the most successful example of the use of CA techniques in solving geographic problems (White and Engelen 1993, Batty and Xie 1994, Clarke and Gaydos 1998, Wu and Webster 1998, Li and Yeh 2000). Couclelis (1985, 1997) carried out some early research on urban simulation using cellular automata. She showed that CA might be used as an analogue or metaphor to study how a variety of urban dynamics might arise. Batty and colleagues (Batty and Xie 1994, 1997, Batty *et al.* 1999) also carried out some interesting research on urban CA in the early 1990s. They used CA to model the growth of built-up areas using diffusion-limited aggregation (DLA) (Batty *et al.* 1989). However, they later developed a general class of CA which emerged through insights in computation and biology (Batty and Xie 1994). Their models are very similar to the *Game of Life* in which each cell can only take on one of two states (dead or alive).

CA are flexible and transparent when they are used to solve geographical problems. It is relatively easy to define transition rules. CA can have a variety of applications, such as testing hypotheses of urban theories (Webster and Wu 1999), simulating urban forms and land-use dynamics (Clarke and Gaydos 1998), and generating development alternatives for conserving land resources (Li and Yeh 2000). It is also possible to incorporate planning objectives in simulating alternative urban forms and densities for urban planning (Yeh and Li 2001a, 2002).

The definition of transition rules in geographical CA is strongly dependent on domain knowledge and individual preferences. In urban simulation, the transition rules are usually given according to the intuitive understanding of the process of urban growth. Transition rules can be defined using a variety of mathematical expressions, such as nested neighbourhood spaces and distance decay functions (Batty and Xie 1994), predefined parameter matrices (White and Engelen 1993), linear equations of multicriteria evaluation (MCE) (Wu and Webster 1998), logistic models (Wu 2002), grey-cell or fuzzy states (Li and Yeh 2000), and neural networks (Li and Yeh 2002). However, most of these transition rules are not explicit because they use mathematical equations instead of using explicit transition rules.

A critical issue in CA modelling is how to obtain domain knowledge and determine transition rules in an objective way. The number of ways of defining transition rules seems to be virtually unlimited. The calibration of geographical CA becomes very difficult because a large number of rules have been used. Usually, transition rules consist of many variables and parameters, but there are many uncertainties in determining parameter values. Urban CA are very sensitive to transition rules and their parameter values (Wu and Webster 1998, Li and Yeh

2002, Wu 2002). The calibration of CA is very important for producing realistic urban simulation.

There are very limited studies on the calibration of geographical CA. Most of the existing CA are based on the so-called 'trial and error' approach. The visual test is the main method for validating urban CA in early studies (Clarke *et al.* 1997, White *et al.* 1997, Ward *et al.* 2000). There have been several attempts to develop more elaborate methods to tackle the problems of uncertainties in defining transition rules and parameter values. Computer search algorithms have been proposed to derive optimal parameter values according to the best fit between the observed data and various simulated results (Clarke and Gaydos 1998). This method involves intensive computation by comparing numerous possible combinations of parameter values. Artificial neural networks have been incorporated into urban CA for deriving parameter values automatically (Li and Yeh 2002). However, it is difficult to comprehend the meanings of these parameter values because of the back-box approach of neural networks. Wu (2002) provides a method to estimate the global development probability by using a logistic regression model. The initial global probability is calibrated according to historical land-use data. It seems to be easy to understand the meanings of the coefficients in the logistic regression equation. However, logistic equations cannot provide explicit rules. Moreover, mathematical equations are sometimes difficult to capture the complexity of relationships.

In this study, a new method based on knowledge discovery or machine learning is proposed to reconstruct the transition rules of geographical CA. Existing CA have adopted the heuristic approach in defining transition rules. The approach is associated with uncertainties because it is subject to the influence of individual knowledge and preferences. Moreover, the simulation of complex geographical phenomena often involves the processing of a large set of spatial data. Automatic knowledge discovery from spatial data should provide significant improvement on the performance of geographical CA.

2. Knowledge discovery of transition rules for geographical cellular automata

The process for acquiring domain knowledge is tedious and time-consuming. Although experts are capable of using their knowledge to solve problems, they cannot guarantee that the knowledge is explicitly expressed in a systematic, correct and complete form. A well-known problem when creating expert systems is often called the 'knowledge acquisition bottleneck' (Huang and Jensen 1997).

We propose using data mining to solve the problems of the difficulties and uncertainties in knowledge solicitation in defining the transition rules of CA. Data mining involves discovering and capturing knowledge embedded in a large data set automatically. This is usually done through machine learning. There are a number of machine-learning algorithms available for data mining, such as ID3 (Quinlan 1986), C4.5 (Quinlan 1993), CART (Breiman *et al.* 1984), IB1, IB2, MPIL1, and MPIL2 (Romaniuk 1993).

Quinlan (1979, 1993) carried out pioneering studies on rule discovery by induction and machine-learning procedures. The first inductive learning program was called C4.5 and has been widely used in many applications (Berry and Linoff 1997). See5 for Window and its Unix counterpart, C5.0, are the most updated version of C4.5. The series of C4.5 and See5/C5.0 systems have been used to reconstruct the rules for classifying remote-sensing imagery (Defries and Chan 2000). Studies have demonstrated that machine learning can provide an accurate and efficient tool for land-cover classification using remote-sensing data (Friedl *et al.* 1999). Recently, there have also been several attempts to apply these systems to soil analysis by using GIS data (Eklund *et al.* 1998, Moran and Bui 2002).

The simulation of geographical phenomena usually involves a vast volume of spatial data. Techniques for constructing the transition rules of CA need to be automated as much as possible. A number of advantages can be identified by using machine-learning systems (e.g. C4.5 and See5/C5.0) to reconstruct the transition rules of geographical CA:

- Decision-tree learning is the most efficient form of inductive learning (Huang and Jensen 1997).
- These systems can automatically determine threshold values and create a knowledge base from observation data.
- They can be conveniently integrated with GIS for using spatial data.
- CA are simultaneously calibrated during the rule-induction process from data mining.
- The retrieved rules are explicit for easier understanding and implementation.

This study applies data-mining techniques to the automatic reconstruction of transition rules for geographical CA using urban simulation as an example (figure 1). A data-mining tool, the See5 system, is used for discovering transition rules. It is based on the 'information gain ratio' to determine the splits at each internal node of the decision tree (Quinlan 1993). The information gain measures

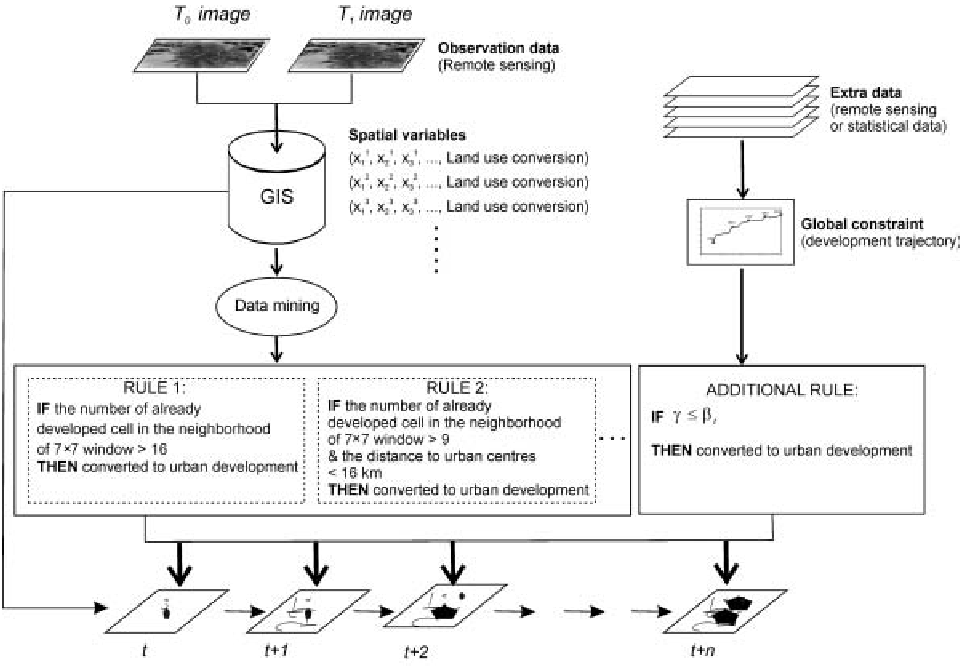
T_0 image
T_1 image
Observation data
(Remote sensing)
Extra data
(remote sensing
or statistical data)
Spatial variables
$(x_1^1, x_2^1, x_3^1, \ldots,$ Land use conversion)
$(x_1^2, x_2^2, x_3^2, \ldots,$ Land use conversion)
$(x_1^3, x_2^3, x_3^3, \ldots,$ Land use conversion)
GIS
Global constraint
(development trajectory)
Data mining
RULE 1:
IF the number of already developed cell in the neighborhood of 7×7 window > 16
THEN converted to urban development
RULE 2:
IF the number of already developed cell in the neighborhood of 7×7 window > 9
& the distance to urban centres < 16 km
THEN converted to urban development
ADDITIONAL RULE:
IF $\gamma \leq \beta$,
THEN converted to urban development
t
t+1
t+2
t+n

Figure 1. Data mining for reconstructing the transition rules of geographical CA.

the reduction in entropy in the data produced by a split. At each node, the tree is divided based on the subdivision that maximizes the reduction in entropy of the descendant nodes. First, imagine selecting one case at random from a training data set S and announcing that it belongs to some class C_j. This message has the probability:

$$\frac{freq(C_j, S)}{|S|} \tag{1}$$

where $freq(C_j, S)$ is the number of cases in S belonging to class C_j, and $|S|$ is the total number of observations in S.

The information from such a message (entropy) is calculated by:

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2 \frac{freq(C_j, S)}{|S|} \tag{2}$$

Consider that S has been partitioned into n outcomes for a test X. The expected information is:

$$info_x(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times info(S_i) \tag{3}$$

The information gained by splitting S using X equals:

$$gain(X) = info(S) - info_x(S) \tag{4}$$

The bias inherent in the gain criterion with a large number of splits should be corrected by normalizing *gain*(X) using *split info*(X) (Quinlan 1993):

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2\left(\frac{|S_i|}{|S|}\right) \tag{5}$$

Then,

$$gain\ ratio(X) = gain(X)/split\ info(X) \tag{6}$$

The ratio can avoid the bias with too many splits during the rule induction procedure. S will be recursively split to ensure that the gain ratio is maximized at each node of the tree. This procedure continues until each leaf node contains only observations from a single class, or there is no gain in information by further splitting. The tests for the continuous attributes are also simple by partitioning each attribute into two outcomes at each node using a threshold. The optimal threshold is also determined according to the gain ratio. The values of an attribute are first sorted, and the midpoint of each interval is used as the representative thresholds (Quinlan 1993). A number of threshold values may be used for the partition. The threshold value with the greatest gain ratio value is selected at each node (DeFries and Chan 2000).

The above procedure automatically creates decision trees or rule sets based on the criterion of 'information gain ratio' (Quinlan 1993). The same procedure can be applied for reconstructing transition rules in simulating geographical phenomena. Many existing urban CA do not provide concrete transition rules but use mathematical equations to estimate conversion probability. In fact, decision-makers are more familiar with explicit rules. For example, it is much easier for them to comprehend the following explicit rules:

Rule 1:

IF *Land-use types = forest or wetland*
THEN *No development is allowed*

Rule 2:

IF *Land-use types = cropland*
Distance to urban centres < 10 km
Number of developed cells in the neighbourhood > 16
THEN *Development is allowed*

GIS and remote sensing can provide spatial data for discovering the transition rules of geographical CA (figure 1). In this study, remote-sensing images of two different years are treated as the observed data for extracting the transition rules. The transition rules based on the two images reveal the relationship between spatial variables and land-use conversion for the observation interval (ΔT) (figure 2). CA use discrete time to update the state of each cell step by step. There are many iterations before the final results are obtained in urban simulation. The transition rules from data mining can be applied to all iterations for simulating urban development based on the past development trend. The assumption is that the relationship between spatial variables and land-use changes do not change.

The projection from these two images can be used to estimate future land consumption. This assumes that the rate of urban growth is constant. However, the rate of urban growth may not be the same because of changes in economic, social and political factors. One way to capture the growth trend is to use more than two years of satellite images (figure 3). These extra remote-sensing data can be used to provide the aggregated information about the development trajectory for simulating future urban development. If extra remote-sensing data are not available, the aggregated information about the growth trend can be estimated by using other sources of data, such as statistical yearbooks.

There is a discrepancy between the iteration interval, the observation interval, and the simulation interval (figure 2). It may be ideal if the observation interval (ΔT) is equal to or close to the iteration interval (Δt) so that the mined transition rules can be directly used in urban simulation. The acquisition of such observation data is subject to the availability of data. The observation interval of remote-sensing data is usually yearly based, while the iteration interval of CA is much smaller. It is impractical to collect data within the iteration interval of Δt. Moreover, the observed data cannot comprehend the long-term trend if the observation interval is too short. ΔT equal to 2–3 years may be practical in most situations. Observed data with the interval of several years have been commonly

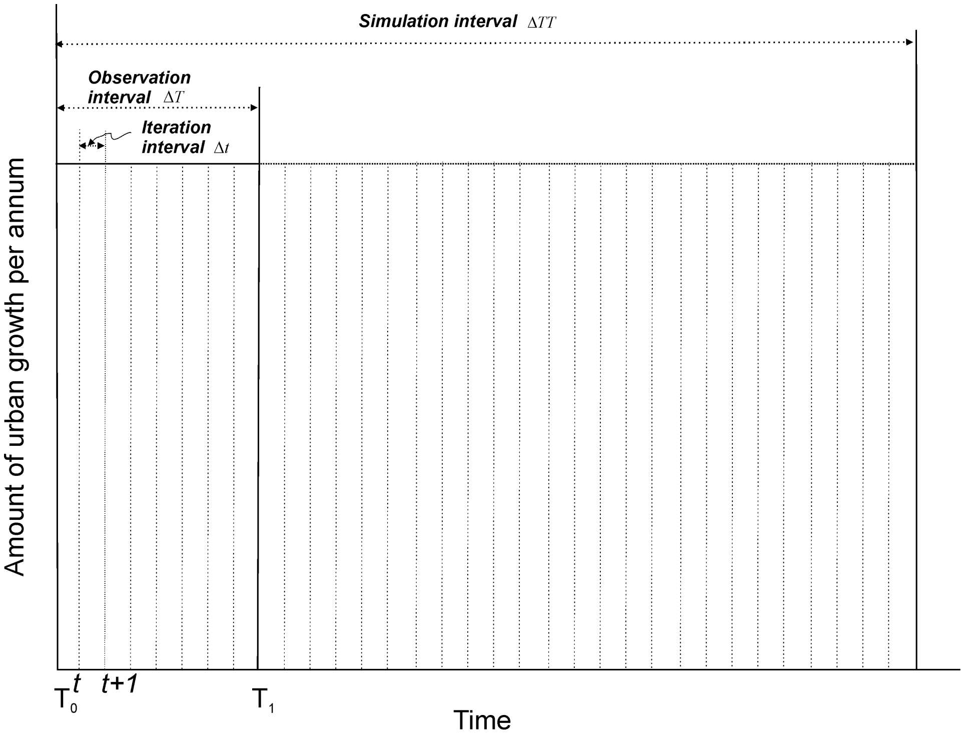
Simulation interval ΔTT
Observation interval ΔT
Iteration interval Δt
Amount of urban growth per annum
t $t+1$
T_0
T_1
Time

Figure 2. Reconstructing transition rules using two years of observation data.

used in the calibration of CA. For example, Li and Yeh (2002) calibrate CA using two satellite images with the interval of 5 years. The calibration in Wu's study (2002) is even based on the land-use maps with an interval of 20 years.

There is a need to discuss how the extracted rules from the observed data are applied to each iteration of urban simulation. First, the relationship between the number of iterations (K), the iteration interval (Δt) and the observation interval (ΔT) is as follows (figure 2):

$$K = \Delta T / \Delta t \tag{7}$$

where ΔT is the observation interval for the two remote-sensing images, Δt is the iteration interval between t and $t+1$, and K is the number of iterations.

Transition rules from data mining only determine whether land-use conversion will take place for the larger interval of ΔT. However, it is possible to estimate the proportion of land-use conversion between t and $t+1$. The estimation can be obtained by using the following equation:

$$\Delta q_t = \Delta Q_t / K \tag{8}$$

where ΔQ_t is the amount of land-use conversion for the observation interval, and Δq_t is the amount of land-use conversion for the iteration interval.

When $\Delta T > \Delta t$, the land-use conversion in Δt only amounts to a small proportion of that in ΔT. Moreover, there is no way to identify the exact locations that will have land-use conversion in the smaller period. Δq_t can decide the system birth rate at each step of the iterations. A random variable (γ) is then used to

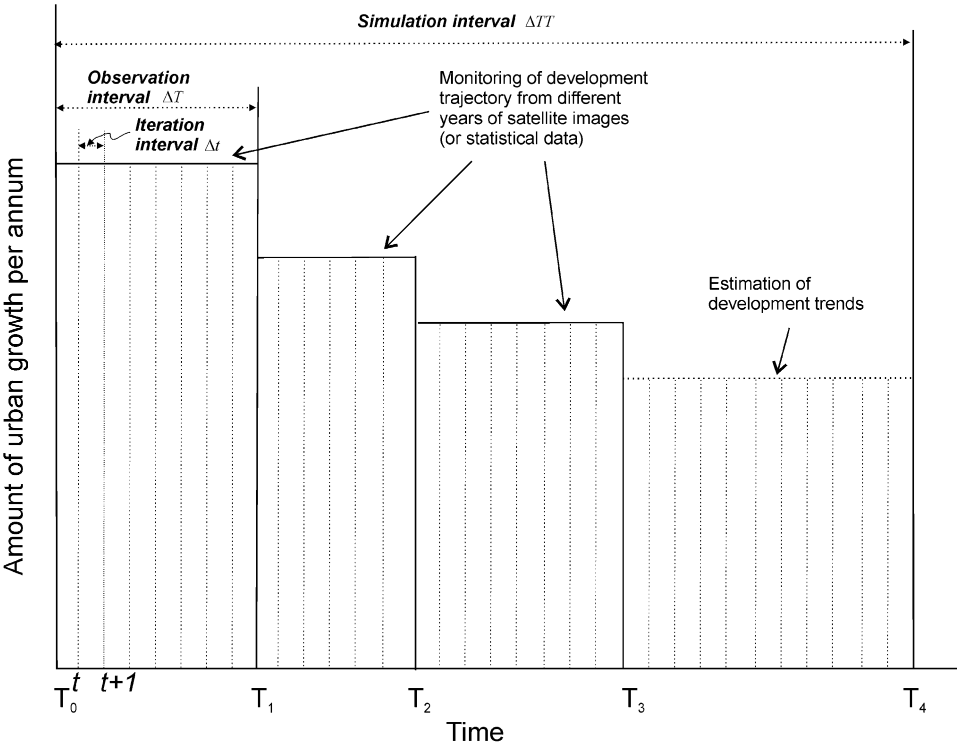
Simulation interval ΔTT
Observation interval ΔT
Iteration interval Δt
Monitoring of development trajectory from different years of satellite images (or statistical data)
Estimation of development trends
Amount of urban growth per annum
t
t+1
T0
T1
T2
T3
T4
Time

Figure 3. Monitoring of development trajectory using additional remote-sensing data.

determine the locations of land-use conversion at the smaller interval (figure 4). The following additional rule is used to obtain the smaller portion of land-use conversion at each iteration:

IF $x(i, j)$ should be converted according to the original transition rules that are obtained from the observation interval of ΔT
& $x(i, j)$ have not developed at $t-1$
& $\gamma \leqslant \beta_0$

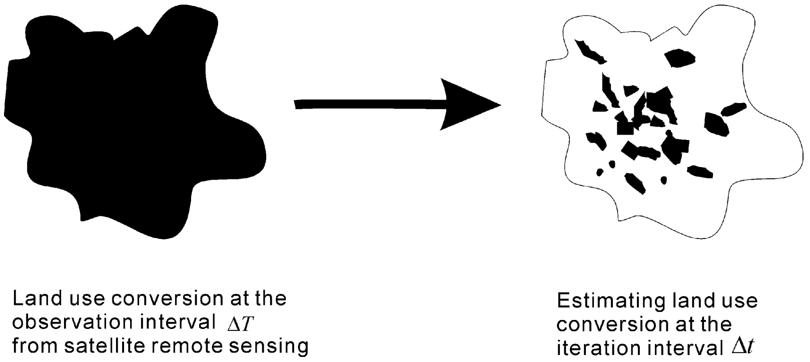
Land use conversion at the observation interval ΔT from satellite remote sensing
Estimating land use conversion at the iteration interval Δt

Figure 4. Assigning land-use conversion at the iteration interval.

THEN $x(i, j)$ will be developed at t

$$\beta_0 = \frac{\Delta q_0}{\Delta Q_0} = \frac{1}{K} \tag{9}$$

where $x(i, j)$ is the cell at location (i, j), ΔQ_0 is the amount of land-use conversion retrieved from the two images, and Δq_0 is the amount of land-use conversion for its iteration interval.

β_0 can be considered a global birth rate or a development probability for the smaller interval of Δt. The use of random variables is important for addressing the stochastic perturbation or uncertainties in complex geographical phenomena. This can help to achieve more realistic simulation results. The final land-use conversion is usually determined by the comparison of a development probability with a random variable in many urban CA (Batty and Xie 1994, Wu and Webster 1998). For example, a cell will be developed if the calculated development probability is greater than a random value (Wu 2002).

Since the amount of land-use conversion is changing with time, it should be calculated from various years of satellite images instead of just using two images. The above additional transition rule should have the following generic form by replacing β_0 with β_t:

IF $x(i, j)$ should be converted according to the original transition rules that are obtained from the observation interval of ΔT
& $x(i, j)$ have not developed at $t-1$
& $\gamma \leqslant \beta_t$
THEN $x(i, j)$ will be developed at t

$$\beta_t = \beta_0 \times \frac{\Delta Q_t}{\Delta Q_0} = \frac{1}{K} \times \frac{\Delta Q_t}{\Delta Q_0} \tag{10}$$

β_t is used as the global constraint to reflect the fluctuation in the amount of land-use conversion (ΔQ_t). The additional rule is important for capturing development trajectory. There is an issue about the uncertainty in deciding land development at the iteration interval. The random variable (γ) may introduce the 'noise' in the simulation process because of the uncertainty in determining the locations of development for the smaller iteration interval. Uncertainty may be minimized by making the observation interval (ΔT) close to the iteration interval (Δt). However, this can be a problem, because the long-term trend cannot be captured. There is a need to balance the trade-off by choosing a suitable interval for the observation data (e.g. 2–3 years). CA are based on discrete time, and a sufficient number of time steps (iterations) are needed to ensure simulation accuracy. However, there is no agreement on how many time steps should be used. Many CA usually run from 100 to several hundred iterations.

3. Implementation and results

3.1. *Training data*

The proposed model has been tested in Dongguan, a city in the Pearl River Delta of Southern China. Dongguan has an area of 2465 km^2 with a city proper and 29 towns. Rapid urban expansion has been witnessed in the Pearl River Delta because of fast economic development (Li and Yeh 1998). A number of CA models have been developed for simulating land development in the study area. The first model uses 'grey' states to simulate the continuous conversion process of urban development (Li and Yeh 2000). The constraints retrieved from GIS are imported to CA to explore various possible development scenarios. Little attention has been paid to the calibration of the model. The second model is to simplify the procedure of defining transition rules and facilitate the calibration of CA by using neural networks (Li and Yeh 2002). However, the transition rules of this model are not transparent because of the back-box approach of neural networks.

This proposed model is the successor of previous models, but it has made significant improvements with a strong capability of rule discovery. In this study, the transition rules are automatically reconstructed from the GIS databases. A series of spatial data are used for the data mining. The data include the layers of urban development, proximity variables, neighbourhood conditions, and physical attributes. Studies indicate that urban development probability is decided by these spatial variables (Wu and Webster 1998, Li and Yeh 2000).

The proposed model is integrated with a GIS for the convenient access to its spatial data and geoprocessing functions. Table 1 lists the spatial variables used for the data mining. The target variable is the urban development in 1988–1993, which was obtained from change detection using the 1988 and 1993 TM images.

Proximity variables were obtained by calculating the distances to roads, expressways, railways, and urban centres. These variables play an important role in determining land-use conversion. For example, a higher development probability is often associated with a site with a closer distance to major transportation routes and urban centres. Proximity variables can be conveniently derived by using the distance functions of GIS.

Neighbourhood conditions are essential to the determination of state conversion of each cell during the iterations of CA. There is a higher development probability if a site has a larger number of surrounding developed cells. The surrounding developed cells are counted at each iteration.

The physical attributes of a site also influence the development probability in an urban simulation. The first variable is land-use types. For example, the development probability in wetland will be different from that in cropland, even though other conditions are the same. Information about land-use types was obtained from the classification of the remote-sensing image. The second variable is related to the agricultural suitability which represents the productivity of a site. It was obtained through GIS land evaluation (Yeh and Li 1998). The third variable is associated with terrain features which pose constraints to urban development. The layer of the slope was generated from the DEM model of GIS.

All these spatial data were converted to raster format to facilitate the calculation and simulation. The resolution was fixed at a ground resolution of $30\,m^2$ to match the resolution of the satellite TM images. Data mining was used to reconstruct the rules that reveal the relationships between these spatial variable and urban development. The simulation assumes that the spatial relationship will not change, although the global constraint (the amount of land-use conversion) may be subject to fluctuations.

Our previous study used only two satellite images to train the neural-network-CA model for simulating future urban development (Li and Yeh 2001). It assumes that the rate of urban growth is constant for all periods. This assumption may not

Table 1. Spatial variables used for data mining.

Spatial variables	Acquisition methods	Value ranges
1. Target variable		1: converted to urban areas
Urban development in 1988–1993	Classification of satellite TM images	0: non-converted
2. Proximity variables		
Distance to the city proper (PropD)	*Eucdistance* of ARC/INFO GRID	0 ~ 60 km
Distance to town centres (TownD)		0 ~ 30 km
Distance to roads (RoadD)		0 ~ 20 km
Distance to expressways (ExprD)		0 ~ 60 km
Distance to railways (RailD)		0 ~ 60 km
3. Neighbourhood function		
Number of developed cells in the 7 × 7 neighborhood (Nsum)	*Focalsum* of ARC/INFO GRID	0–49
4. Physical attributes of a site		1: crop
Land use types (Land)	Classification of satellite TM images	2: bared soil
		3: construction sites
		4: orchard
		5: built-up areas
		6: forest
		7: water
Agricultural suitability (Agsu)	Land evaluation of GIS	0 ~ 1
Slope (Slope)	DEM of GIS	1 ~ 90°

be true because the dynamics of economy, policy and resource supply will result in changes in the land-use conversion process As discussed above, a time sequence of satellite images can help to comprehend the long-term trends of urban development. Satellite TM images in four years, namely 10 December 1988, 24 December 1993, 29 August 1997, and 20 November 2001, were used to calibrate the CA model. These satellite images can provide useful information about the trend of land-use changes in the region.

3.2. *Data sampling and data mining*

The ability to make accurate predictions is important for applying decision trees or rule sets to classification. The accuracy of a classifier should not be judged by measuring how well it does on the cases used in its construction. It is more reasonable to assess the performance of the classifier on new cases. In this study, the empirical data from GIS and remote sensing were divided into two separate sets. One was the training data set for deriving the rules, and the other was the test data set for confirming the performance of the classifier.

Sampling techniques were also used to serve two main purposes in this study. First, sampling techniques can help to explore a huge set of spatial data. It is inefficient to process the entire set of spatial data for data mining. Even though See5 is relatively fast, a much longer time is needed for building decision trees, especially when options such as boosting are employed. Second, it is undesirable to use a whole set of data for mining because of spatial autocorrelation. Bias will be introduced to the analysis results if the training data have a severe correlation.

The use of a smaller set of training cases may reduce the classifier's predictive accuracy. We first construct a classifier from the sample and then assess the classifier on a test data set. Figure 5 shows the relationships between the increase in sampling points and prediction error. It is clear that the prediction error can be significantly improved by using more sampling points within the range of 0–8% of the training data. The prediction error is 35.2% by using 1% of the data, and it is reduced to 25.0% by using 10% of the data. The improvement rates are insignificant

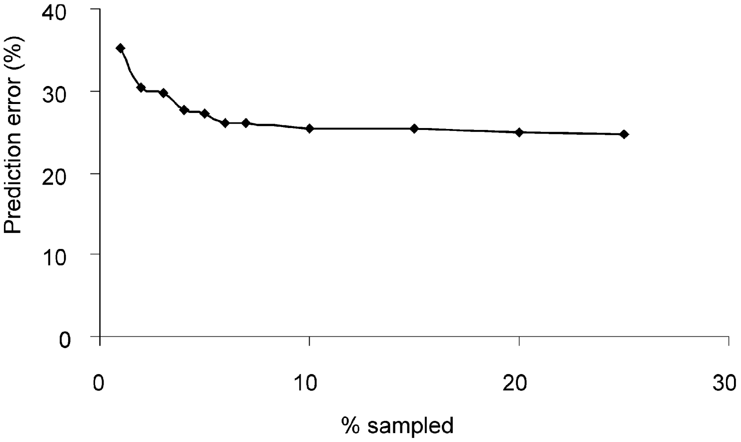

Figure 5. Sampling rate and prediction error.

after the first 10% of the data. Therefore, this study only used 20% sampled data to derive the transition rules.

A technique known as 'boosting' has been developed in the field of machine learning. See5/C5.0 has incorporated this technique, adaptive boosting, to generate several classifiers rather than one. First, a tree is generated as usual, which may make mistakes in some cases in the data. When the second classifier is constructed, more attention is paid to these cases in an attempt to get them right. This is repeated for *n* trials. When a new case is to be classified, each classifier votes for its predicted class, and the votes are counted to determine the final class. Boosting can reduce bias and avoid over-fitting of decision trees (Haruno *et al.* 1999)

Lower error rates are expected by using this technique. The effect of boosting is assessed by comparing the predictive errors from the boosting method and non-boosting method. The prediction error is 21.2% after boosting is applied to the 10% sampled data set. The error rate for the test cases was reduced by about 12% compared with that of the original classifier.

A very large and complex decision tree is often produced, and the tree may overfit the training data. If the training data contain errors, overfitting the tree to the data can lead to poor performance (Friedl *et al.* 1999). The original tree must be pruned to minimize such a problem. See5 provides the pruning option for simplifying decision trees but maintaining sufficient accuracy. A large tree is first grown to fit the data closely and then pruned by removing the unnecessary parts which are predicted to have a relatively high error rate. A pruning rate of 25% was used to consider the trade-off between tree accuracy and size.

3.3. *Simulation results*

The proposed model was tested by simulating the urban development in the study area in 1988–2005. The observation data mainly include the 1988 and 1993 satellite TM images. The 1997 and 2001 images are just used for capturing the urban development trend. The initial urban areas were based on the classification of the 1988 TM image. Figure 6 shows the development trajectory of the study area in 1988, 1993, 1997 and 2001 according to the classification of satellite images.

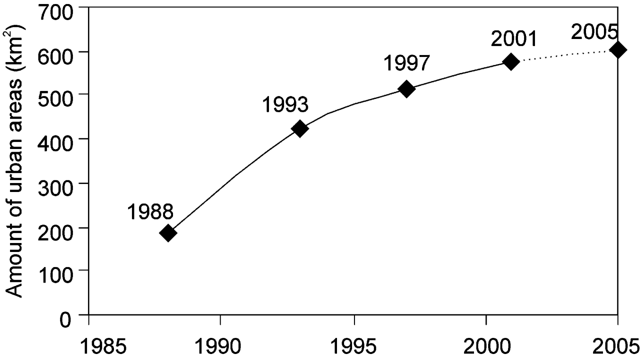

Figure 6. Monitoring of the development trajectory of Dongguan using the 1988, 1993, 1997, and 2001 TM images.

Table 2. Iterations, intervals, and amount of urban growth for each period.

	1988–1993	1993–1997	1997–2001	2001–2005
K (iterations)	200	200	200	200
ΔT (year)	5	4	4	4
Δt (year)	1/40	1/50	1/50	1/50
ΔQ_t (km^2)	233.3	90.6	62.9	25.0
Δq_t (km^2)	1.167	0.453	0.315	0.125
β_t	0.0050	0.0019	0.0013	0.0005

Urban expansion was astonishing in 1988–1993 as the urban areas were more than double during this short period. The rate of urban expansion decreased in the later periods because of government intervention. If the projection of urban development is based only on the 1988 and 1993 observation data, the simulated urban areas will be much larger than the actual areas for the later periods. Therefore, the total amount of urban areas in each period can be used as the global constraint for urban simulation. This can ensure that the amount of the simulated urban areas is equal to the amount of the actual urban areas. Given n iterations, interpolation was carried out to derive the amount of conversion for each step from t to $t+1$ (figure 3).

There are many iterations of simulation before the outcome is obtained. A shorter interval between t and $t+1$ means that a larger number of iterations are required. Although there is no consensus on exactly how many iterations should be used, 100–200 of iterations are quite normal for producing a realistic simulation. The subtle patterns cannot be produced if there are too few iterations. This is because local interactions only take place at each iteration of urban simulation.

Table 2 lists the parameter values that were used in the simulation. There are 200 iterations in the simulation of urban growth for each period. The amount of urban growth (ΔQ_t) for each period was obtained from the change detection of remote sensing. The global constraint factor (β_t) was calculated according to equation (10).

The rule sets were obtained from the data-mining procedure using the See5 system. The following is part of the rule sets discovered by applying the data mining to the GIS and remote-sensing data:

Rule 1:

IF $PropD < 40$
$RoadD <= 5$
$Nsum > 18$
$Agsu < 0.8$
$Land = 1$

THEN Converted to urban development [confidence: 0.92]

Rule 2:

IF $PropD > 12$
$PropD < = 55$
$TownD > 11$
$RoadD < = 3$
$ExprD < = 45$
$RailD < = 12$
$Nsum > = 8$

THEN Converted to urban development [confidence: 0.82]

Rule 3:

IF $PropD < = 25$
$TownD > 7$
$Nsum > = 12$
$Agsu < = 0.5$
$Land = 4$
$Slope < = 6^{\circ}$

THEN Converted to urban development [confidence: 0.86]

Rule 4:

IF $PropD < = 48$
$TownD > 13$
$RoadD > 1$
$RoadD < = 5$
$Nsum > = 9$
$Agsu > 0.2$
$Agsu < = 0.4$

THEN Converted to urban development [confidence: 0.90]

Each applicable rule votes for its predicted class with a voting weight equal to its confidence value. The confidence value is also automatically obtained by See5 during the data-mining process. The votes are summed up, and the class with the highest total vote is chosen as the final prediction. It is straightforward to use these rule sets obtained from data mining as the transition rules for urban simulation. These rule sets are very similar to the practical rules used by planners and decision-makers. It is much easier to understand these rule sets than mathematical equations.

The amount of land-use conversion changes with time. β_t is used to reflect the changes in urban growth. The following additional rule is jointly used to decide the final land-use conversion at each iteration from t to $t+1$:

Additional rule:

$$\text{IF}\quad \gamma \leqslant \begin{cases} 0.0050 \text{ (in } 1988-1993) \\ 0.0019 \text{ (in } 1993-1997) \\ 0.0013 \text{ (in } 1997-2001) \\ 0.0005 \text{ (in } 2001-2005) \end{cases}$$

THEN Converted to urban development

The model simulates the urban growth of the study area in the period of 1988–1993, 1993–1997, and 1997–2001. The initial stage is based on the 1988 actual urban areas detected from remote sensing. Figure 7 compares the simulated and actual urban development in 1988–1993, 1993–1997, and 1997–2001. Figure 8 is a prediction of urban development in 2001–2005 based on the development trajectory. The rate of urban expansion is much lower in this period than in previous years.

3.4. *Examining the validity of the model*

It is unrealistic to reproduce the exact patterns of a natural phenomenon because of its complexity and modelling limitations. However, the assessment of goodness of fit is still required to give a general indication of roughly how good the simulation is compared with the actual development. A simple method to assess the goodness of fit is based on the spatial overlay between the actual and simulated urban development (Li and Yeh 2002). In this study, the actual urban areas in 1993, 1997 and 2001 were obtained from the classification of satellite TM images. The simulated urban areas were compared with the actual urban areas using an overlay analysis. Table 3 lists the overall accuracy obtained from the cross-tabulation of the overlay analysis. The overall accuracy is 82.0% for simulating the urban growth in 1988–1993. It becomes 74.8% and 72.4% for simulating the urban growth in 1993–1997 and 1997–2001, respectively. Although the transition rules were derived from the 1988–1993 images, we are still able to obtain a high simulation accuracy in the simulation of urban development in 1993–1997 and 1997–2001. This is because the urban development process captured by the transition rules has not changed much.

The assessment of the goodness of fit from spatial overlay is just based on a cell-by-cell approach. It cannot provide any information about the morphology of the urban spatial structures, such as connectivity, fractals, and compactness. Many urban applications are usually concerned about the characteristics of spatial structures. A visual comparison may sometimes provide more meaningful results for calibrating CA models (Clarke *et al.* 1997, Ward *et al.* 2000). The visual comparison of the actual with simulated urban development indicates that the model is able to generate plausible simulation results (figure 7).

It is better to use robust and consistent methods for the assessment based on quantitative indicators. These indicators should be able to describe the characteristics of spatial patterns and provide useful insights about urban morphology. However, there is no agreement on which indicator is most suitable for capturing the characteristics of urban structures because of its complexity. A variety of aggregated indicators have been proposed for this purpose, including

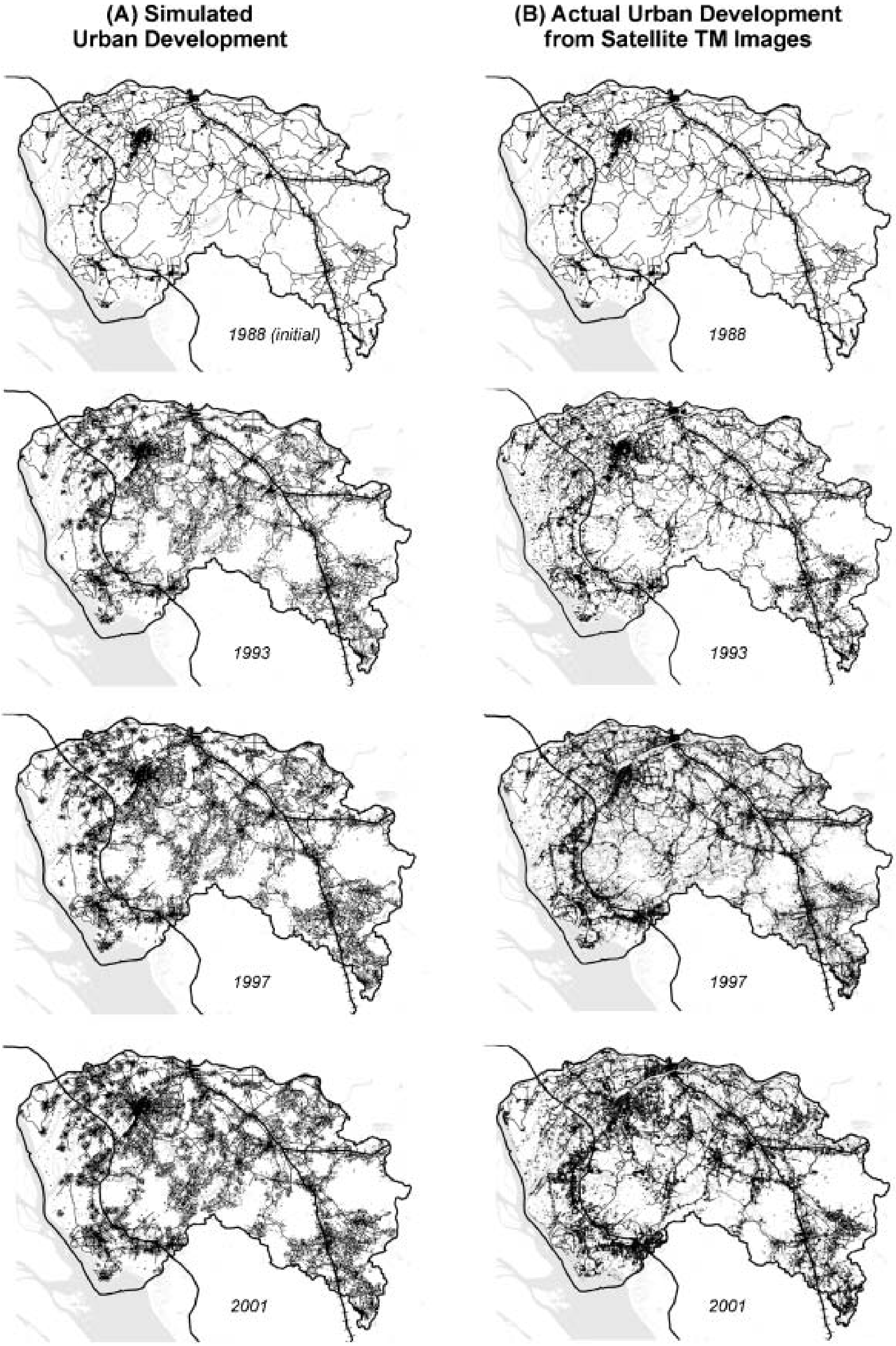
(A) Simulated Urban Development
(B) Actual Urban Development from Satellite TM Images
1988 (initial)
1988
1993
1993
1997
1997
2001
2001

Figure 7. Simulated and actual urban development of Dongguan in 1988, 1993, 1997, and 2001.

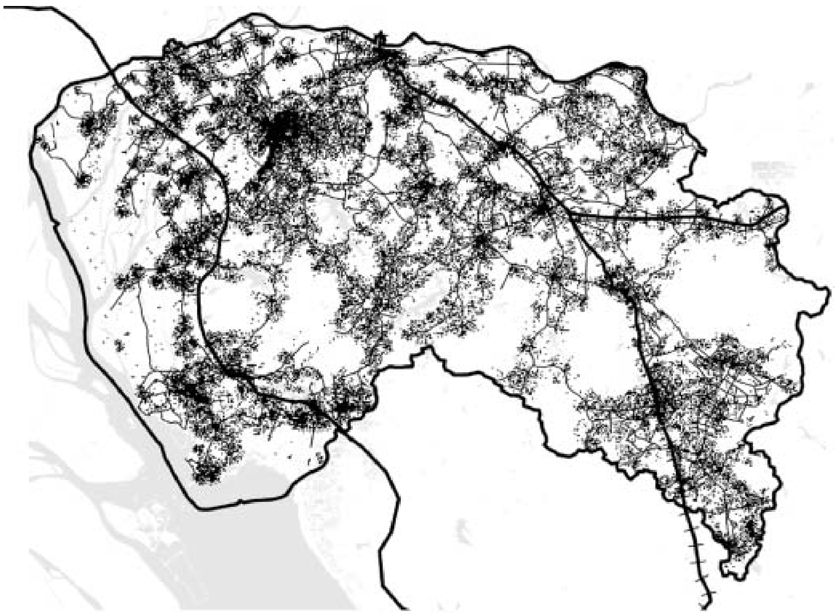

Figure 8. Prediction of future urban development in 2005 based on the development trend.

Table 3. Overall accuracy of simulation compared with the actual urban development obtained from satellite images in 1993, 1997, and 2001.

Year	1993	1997	2001
% Correct	82.0	74.8	72.4

entropy (Yeh and Li 2001b), compactness index (Li and Yeh 2000), fractal dimensions (White and Engelen 1993, Batty and Longley 1994), and Moran's I (Wu 2002).

In this study, the indicator of Moran's I was chosen for the assessment of the aggregated patterns because of its simplicity. It is quite easy to calculate the Moran's I values in the GIS package, ARC/INFO GRID. Moran's I is a useful spatial indicator that can reveal the degree of spatial autocorrelation (Goodchild 1986). The indicator is able to estimate how close the simulated land-use pattern is to the actual urban development (Wu 2002). The maximum value is one which indicates absolute concentration of land use. A smaller value, which can be below zero, indicates a more even distribution of land use.

Table 4 shows Moran's I for the actual and simulated urban development in

Table 4. Comparison of Moran's I between the actual and simulated urban development in 1993, 1997, and 2001.

	1993	1997	2001
Actual urban development	0.44	0.66	0.76
Simulated urban development	0.42	0.58	0.71

1993, 1997 and 2001, respectively. There is a good conformity between the actual and simulated urban development because they have similar values of Moran's I for each period. The analysis is consistent with the visual comparison. Urban development sites in the earlier stage (1993) are relatively isolated because of the prevailing urban sprawls. Urban developments tend to be more compact in the later years as they continue to grow. According to Moran's I, the simulated patterns appear to be slightly more dispersed compared with the actual patterns for all periods. This is probably because the simulation is affected by some randomness. Observation data always have a larger time interval than that of simulation. Interpolation is carried out to yield training data on a finer timescale. A random variable is used to decide the location of land-use conversion at each iteration. The problem may be alleviated by using a shorter interval for satellite images. However, this is subject to the availability of data.

Compared with the neural-network-based CA (Li and Yeh 2002), this model also has some improvements in terms of accuracy. The overall accuracy is 0.79, and Moran's I is 0.40 for the previous model. This is probably because the explicit transition rules are more easily adapted to complex relationships than mathematical equations. Simulation accuracy also depends on the degree of complexity of the study area. It is easy to understand that CA can have a better performance in a more uniform and smaller area, such as a monocentric city or a highly developed city. The geographical settings of this study area are fairly complicated, with various geomorphologic features (e.g. mountains, alluvial plains, and rivers) and many suburban centres (29 towns). A perfect simulation is impossible because of the complexity. There are several difficulties in applying the heuristic approach to the definition of transition rules in the areas of complex environmental settings. However, data mining seems to be a much better option for reconstructing transition rules under complex situations. In particular, it can provide explicit transition rules which are more easily understood by planners and decision-makers than the mathematical equations derived by other methods.

4. Conclusion

Data mining, which is a rapidly expanding field, can be applied to the discovery of transition rules of CA. Research on modelling geographical phenomena in various disciplines using CA is growing. Effective reconstruction of transition rules is important for simulating complex natural systems. Reliable simulation results cannot be achieved if the transition rules are not defined in a systematic and consistent way. This study demonstrates the potential of using data-mining techniques in automatically deriving the transition rules of CA. The benefits of this method include a faster rule-base construction, convenient calibration, and transparent rule structures.

We have attempted to use an innovative method for directly deducing explicit transition rules of CA based on data-mining techniques. Various transition rules have been proposed from different studies. These transition rules are not straightforward because they are mainly represented in the form of mathematical equations (e.g. conversion matrices, linear equations, and logistic equations). They are difficult to understand and comprehend by planners and decision-makers. Sometimes, complex relationships cannot be modelled by rigid mathematical equations. The calibration of CA is also difficult when complex mathematical equations are adopted to estimate the probability of land-use conversion.

The data-mining procedure is convenient and efficient. The explicit transition rules can be instantly derived from a vast volume of geographical data by using data-mining techniques. Remote sensing and GIS provide basic information about spatial variables for data mining. This procedure can minimize the uncertainties and time consumed in defining and testing transition rules because they are automatically reconstructed by machine learning. Calibration is automatically carried out during the rule-induction process. This has significant improvements in the process of model building.

The experiments were carried out in a region of complicated environmental settings. There are a variety of terrain features and many suburban centres. The heuristic approach adopted by traditional CA has difficulties in defining transition rules and calibrating CA. The transition rules automatically induced from data mining have been successfully applied to the urban simulation in the region. The validity of the model has been assessed based on the visual comparison and the indicator of Moran's I. The assessment indicates good conformity between the actual and simulated urban development.

Further studies should examine the influences of discrete time steps on simulation outcomes. Experiments may be carried out on the search for the optimal intervals of iterations and observations. This is useful for generating more accurate simulation results. Model uncertainties should also be assessed for a better understanding of the implications of simulation.

Acknowledgement

This study is supported by funding from the Research Grants Council of Hong Kong (HKU 7209/02H).

References

BATTY, M., and LONGLEY, P. A., 1994,, *Fractal Cities: A Geometry of Form and Function* (London: Academic Press).

BATTY, M., and XIE, Y., 1994, From cells to cities. *Environment and Planning B: Planning and Design*, **21**, 531–548.

BATTY, M., and XIE, Y., 1997, Possible urban automata. *Environment and Planning B: Planning and Design*, **24**, 175–192.

BATTY, M., LONGLEY, P., and FOTHERINGHAM, S., 1989, Urban growth and form: scaling, fractal geometry, and diffusion-limited aggregation. *Environment and Planning A*, **21**, 1447–1472.

BATTY, M., XIE, Y. C., and SUN, Z. L., 1999, Modeling urban dynamics through GIS-based cellular automata. *Computer, Environment and Urban Systems*, **23**, 205–233.

BERRY, M. J. A., and LINOFF, G., 1997,, *Data Mining Techniques for Marketing, Sales, and Customers Support* (New York: Wiley).

BREIMAN, L., FREIDMAN, J. H., OLSHEN, R. A., and STONE, C. J., 1984,, *Classification and Regression Trees* (Belmont, CA: Wadsworth).

CLARKE, K. C., and GAYDOS, L. J., 1998, Loose-coupling a cellular automata model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. *International Journal of Geographical Information Science*, **12**(7), 699–714.

CLARKE, K. C., BRASS, J. A., and RIGGAN, P. J., 1994, A cellular automata model of wildfire propagation and extinction. *Photogrammetric Engineering & Remote Sensing*, **60**, 1355–1367.

Clarke, K. C., Hoppen, S., and Gaydos, L., 1997, A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design*, **24**, 247–261.

Couclelis, H., 1985, Cellular worlds: a framework for modeling micro–macro dynamics. *Environment and Planning A*, **17**, 585–596.

Couclelis, H., 1988, Of mice and men: what rodent populations can teach us about complex spatial dynamics. *Environment and Planning A*, **20**, 99–109.

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

Defries, R. S., and Chan, J. C. W., 2000, Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, **74**, 503–515.

Eklund, P., Kirkby, W. S. D., and Salim, A., 1998, Data mining and soil salinity analysis. *International Journal of Geographical Information Science*, **12**(3), 247–268.

Friedl, M. A., Brodley, C. E., and Strahler, A. H., 1999, Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience and Remote Sensing*, **37**(2), 969–977.

Gardner, M., 1971, On cellular automata, self-reproduction, the garden of Eden and the game of life. *Scientific American*, **224**, 112–117.

Goodchild, M. F., 1986,, *Spatial Autocorrelation: Concepts and Techniques in Modern Geography*, 47 (Norwich, UK: Geo Books).

Haruno, M., Shirai, S., and Ooyama, Y., 1999, Using decision trees to construct a practical parser. *Machine Learning*, **43**, 131–149.

Huang, X., and Jensen, J. R., 1997, A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. *Photogrammetric Engineering & Remote Sensing*, **63**, 1185–1194.

Li, X., and Yeh, A. G. O., 1998, Principal component analysis of stacked multi-temporal images for monitoring of rapid urban expansion in the Pearl River Delta. *International Journal of Remote Sensing*, **19**(8), 1501–1518.

Li, X., and Yeh, A. G. O., 2000, Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science*, **14**(2), 131–152.

Li, X., and Yeh, A. G. O., 2001, Calibration of cellular automata by using neural networks for the simulation of complex urban systems. *Environment and Planning A*, **33**, 1445–1462.

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

Moran, C. J., and Bui, E. N., 2002, Spatial data mining for enhanced soil map modeling. *International Journal of Geographical Information Science*, **16**(6), 533–549.

Portugali, J., 2000,, *Self-Organization and the City* (New York: Springer).

Quinlan, J. R., 1979, Discovering rules by induction from large collection of examples., In *Expert Systems in the Microelectronic Age*, edited by D. Michie (Edinburgh: University Press).

Quinlan, J. R., 1986, Induction of decision trees. *Machine Learning*, **1**, 81–106.

Quinlan, J. R., 1993,, *C4.5: Programs for Machine Learning* (San Mateo, CA: Morgan Kaufmann).

Romaniuk, S. G., 1993,, *Multi-Pass Instance Based Learning*, Technical Report trh3/93, Department of Information System and Computer Science, National University of Singapore, Singapore.

Tobler, W. R., 1979, Cellular geography., In *Philosophy in Geography*, edited by S. Gale and G. Olssen (Dordrecht: Reidel), pp. 379–386.

Ward, D. P., Murray, A. T., and Phinn, S. R., 2000, A stochastically constrained cellular model of urban growth. *Computers, Environment and Urban Systems*, **24**, 539–558.

Webster, C. J., and Wu, F., 1999, Regulation, land-use mix, and urban performance. Part 1: theory. *Environment and Planning A*, **31**, 1433–1442.

White, R., 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**, 1175–1199.

WHITE, R., ENGELEN, G., and UIJEE, I., 1997, The use of constrained cellular automata for high-resolution modelling of urban land use dynamics. *Environment and Planning B: Planning and Design*, **24**, 323–343.

WOLFRAM, S., 1984, Cellular automata: a model of complexity. *Nature*, **31**, 419–424.

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

WU, F., and WEBSTER, C. J., 1998, Simulation of land development through the integration of cellular automata and multicriteria evaluation. *Environment and Planning B*, **25**, 103–126.

YEH, A. G. O., and LI, X., 1998, Sustainable land development model for rapid growth areas using GIS. *International Journal of Geographical Information Science*, **12**(2), 169–189.

YEH, A. G. O., and LI, X., 2001a, A constrained CA model for the simulation and planning of sustainable urban forms by using GIS. *Environment and Planning B: Planning and Design*, **28**, 733–753.

YEH, A. G. O., and LI, X., 2001b, Measurement and monitoring of urban sprawl in a rapidly growing region using entropy. *Photogrammetric Engineering & Remote Sensing*, **67**(1), 83–90.

YEH, A. G. O., and LI, X., 2002, A cellular automata model to simulate development density for urban planning. *Environment and Planning B*, **29**, 431–450.