Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers, Environment and Urban Systems
30 (2006) 10–28
Computers, Environment and Urban Systems
www.elsevier.com/locate/compenvurbsys

Errors and uncertainties in urban cellular automata

Anthony Gar-On Yeh [a,*], Xia Li [b]

[a] Centre of Urban Planning and Environmental Management, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, PR China
[b] School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, PR China

Received 22 February 2002; accepted in revised form 12 May 2004

Abstract

Errors and uncertainties are important issues in most geographical analyses and modelling processes. Cellular automata (CA) have been increasingly used for modelling geographical phenomena, such as the evolution of urban systems. Urban simulation frequently involves the inputs of a large set of spatial variables from GIS. The errors of data source in GIS can propagate through CA modelling processes. Moreover, CA models themselves also have modeling uncertainties because they are just an approximation to reality. These uncertainties have impacts on the outcome of urban simulation. Identification and evaluation of these errors and uncertainties are crucial for understanding and implementing the simulation results of urban CA modelling. It is found that some of the characteristics of errors and uncertainties in urban CA are quite unique which are not present in traditional GIS models. The study can help urban modelers and planners to understand more clearly the characteristics and implications of CA modelling.

Keywords: Uncertainty; Cellular automata; Urban simulation; GIS

* Corresponding author. Tel.: +852 2855 9481; fax: +852 2559 0468.
E-mail addresses: hdxugoy@hkucc.hku.hk (A. Gar-On Yeh), gplx@zsu.edu.cn, lixia@graduate.hku.hk (X. Li).

0198-9715/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2004.05.007

1. Introduction

Errors and uncertainties are important issues in the GIS literature. Compared to traditional methods (e.g. manual overlay), GIS provides more powerful functions and accurate information based on computer technology. However, GIS are not free of errors and uncertainties because of human errors, technical limitations and the complexity of nature. GIS databases are approximations to real geographical variations with very limited exceptions (Goodchild, Sun, & Yang, 1992). Understanding of errors and uncertainties of GIS is important for successful applications of GIS techniques. There are two main types of GIS errors: (a) data source errors that exist in GIS databases; and (b) error propagation through the operation performed on the data by using GIS functions.

There is a growing trend of using cellular automata (CA) to study geographical phenomena. CA were originally developed in digital computing and have been widely used for simulating complex systems in physics, chemistry and biology. Recently, a number of urban CA have been proposed to model complex urban systems with the integration of GIS (Batty & Xie, 1994). Urban CA have much simpler forms, but produce more meaningful and useful results in simulating urban dynamics than mathematical-based models. Temporal and spatial complexities of urban development can be well simulated by appropriately defining transition rules in CA models. The application of CA in urban modeling can give insights into a wide variety of urban phenomena. CA are capable of providing important information for understanding urban theories, such as the emergence and evolution of forms and structures (Webster & Wu, 1999; Wu & Webster, 1998). They are also used as planning models for formulating development scenarios (Li & Yeh, 2000; Yeh & Li, 2001, 2002).

Although there are many studies on urban CA, however, the errors and uncertainties of CA have not attracted much attention. Only a few studies have been carried out by examining the 'sensitivity' issues of CA (Benati, 1997). Huge volume of geographical data is usually used in urban CA simulation, especially in modelling real cities. Spatial variables can be retrieved from GIS and imported to CA modeling processes. Like other GIS models, urban CA also has problems of data errors and model uncertainties. These errors will propagate in CA simulation and affect the simulation outcomes. This requires the evaluation of the influences of source errors and error propagation on simulation results. This paper attempts to examine the influences of errors and uncertainties on urban CA simulation. This can help urban planners to be aware of these issues when CA are used for projecting and modeling future development in urban planning.

2. Uncertainties in urban CA

Urban CA models are subject to errors and uncertainties when they are applied to real cities. It is because urban CA models are quite different from Wolfram's deterministic CA models (Wolfram, 1984). Wolfram's models have strict definitions and

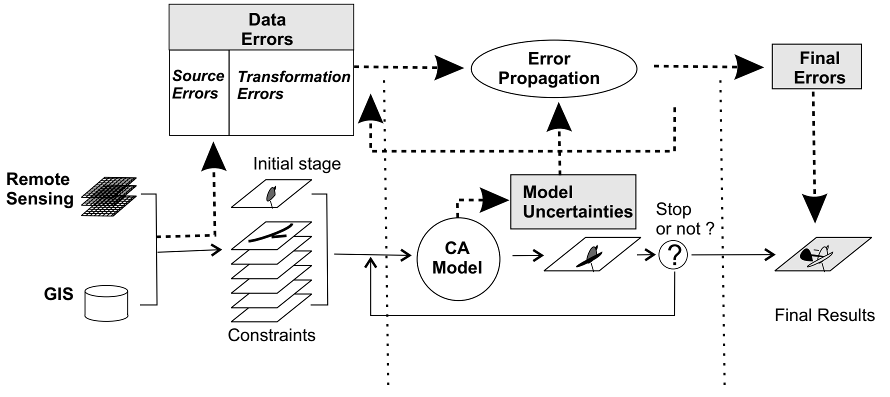

Fig. 1. Data errors, model uncertainties, and error propagation in cellular automata.

use very limited data. This allows CA models to produce stable outputs. However, urban CA models usually require the input of a large set of spatial data for realistic simulation. The outcome of CA models will be affected by a series of errors and uncertainties from data sources and model structures (Fig. 1). The structure of dynamic looping is quite different from the simple GIS operations (e.g. overlay) which can derive strict mathematical equations to estimate error propagation in modeling process.

2.1. Errors from data sources

When spatial data are used in urban CA, the simulation is affected by a variety of data source errors, such as investigation errors, mapping errors, and digitization errors in building GIS databases. The first step is to identify the types of errors from data sources. Two major types of source errors can be identified: Positional errors and attribute errors.

2.1.1. Positional errors

Positional errors in GIS can affect the accuracy of urban simulation. Such errors can cause mistakes in estimating conversion probability which is related to proximity variables. Positional accuracy has been widely discussed in many GIS studies (Goodchild, 1991; Veregin, 1999). The positional errors for points can be measured by the discrepancy between the actual location and recorded location. The spatial error for a set of points has been commonly represented by root mean squared error (RMSE), which is computed as the square root of the mean of the squared errors.

The position errors for lines can be represented using some variant of the epsilon band (Veregin, 1999). There is a certain probability of observing the 'actual' line within the band. The simplest one is to assume that the band and the distribution are uniform. However, recent studies show that both the band and distribution might be non-uniform (Caspary & Scheuring, 1993; Veregin, 1999).

2.1.2. Attribute errors

Attribute errors in data source can affect urban simulation results when these data are used as the inputs to CA. These errors convey that something is wrong for labeling at each location. Conventional surveying maps have errors that are associated with human errors (e.g. reading errors) and instrumental errors (e.g. unstable conditions). For example, a site labeled as vegetables on a map may turn out to be grass on the ground. A DEM derived from contours is also susceptible to the errors of interpolation. These errors will contribute to the uncertainty in determining the initial state and calculating the constraint of each cell.

2.2. Transformation or operation errors

Besides data source errors, common GIS operations or transformations can also bring about uncertainties to CA modeling. In preparing the inputs to CA models, some standard GIS operations have to be carried out to generate additional information that is not stored in GIS. GIS databases only contain basic data for storage efficiency. User-specific information may be produced by standard GIS operations, e.g. data conversion, map algebra, buffering and masking. For example, development suitability may be an important variable in estimating development probability (Wu & Webster, 1998). The calculation of development suitability involves the use of a series of GIS operations. Various layers of spatial variables are overlaid in most situations. These operations are subject to modelling uncertainties. The transformation of GIS data may include:

- vector–raster transformation;
- raster–raster transformation (e.g. resampling);
- overlay or buffer operations;
- other complex operations (e.g. classification).

There are two major types of GIS data formats—vector and raster. The conversion between vector and raster format is a common task in GIS operations. Urban CA models are usually implemented using raster format—cells. Therefore, the inputs of GIS data to CA models should be prepared in raster format. Vector data have to be converted into raster data before spatial data can be handled by most of urban CA. It is apparent that the conversion of vector data into raster data will result in a loss of spatial detail (Fig. 2).

Even for raster data, raster–raster transformation is required for two purposes—registration of different layers of data and conversion of data from one spatial resolution to another. Registration of different sources of raster data is an important procedure of using geographical data. Geo-referencing of maps is usually done by using affine transformation or polynominal transformation. The transformation will resample data with the method of nearest neighbor, bilinear interpolation, or cubic convolution. It is possible that new errors may be created because of the mistakes in registration or resampling. Conversion of data by changing cell size can allow them

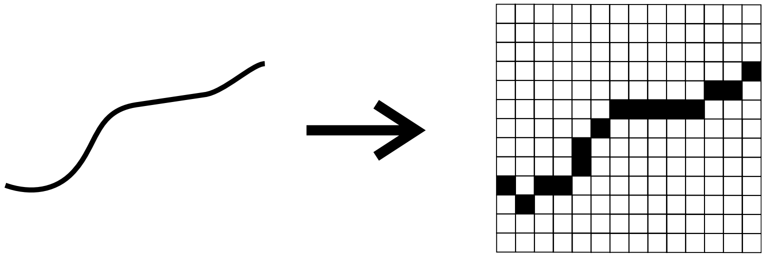

Fig. 2. Loss of spatial information using discrete cells.

to be comparable. However, when raster data are converted from a higher spatial resolution to a lower spatial resolution, there is a loss of information.

Transformation related to GIS overlay can be implemented by 'cartographic algebra' (Burrough, 1986). Sometimes, multicriteria evaluation (MCE) may be required when a number of spatial factors are involved in urban simulation (Wu & Webster, 1998). These operations may generate new errors during the process of data handling. GIS operations are in effect a computational model which is merely an approximation to reality (Heuvelink, 1998). Model errors can be introduced in GIS database when such operations are carried out.

Environmental factors or constraints are usually incorporated in urban CA. This type of information is obtained by using ordinary GIS operations, such as overlay analysis or transformation. For example, constrained CA models may be developed to simulate planned development (Li & Yeh, 2000). The purpose is to prohibit uncontrolled development using constraint information provided by a GIS. A series of resource and environmental factors can be defined in GIS and imported to CA models as the attribute of each cell. These factors may include topography, land use types, proximity and agricultural productivity (Li & Yeh, 2000). Constraint scores can be calculated by applying GIS linear or non-linear transformation functions. However, there are uncertainties in defining the forms of transformation functions.

Errors may also be created during proximity analysis or buffer analysis of GIS. In urban simulation, a common procedure is to calculate urban development probability. Urban development probability decides whether land development can take place during the simulation process. The probability is estimated based on the attractiveness to urban development. It is more attractive to urban development if a site has closer distance to major transport networks or facilities. Some distance variables are used to represent the attractiveness (accessibility), including various distances to roads, railways, town centres, hospitals and schools. These variables can be conveniently defined based on GIS buffer analysis with the use of corresponding point and line layers. A major problem is that there may be positional errors in representing points and lines in GIS layers. These errors can originate from human errors (e.g. mis-registration) or model errors (e.g. limitations of pixel size). These positional errors can cause uncertainties in determining development probability for urban simulation.

Other operations on spatial data can also bring about uncertainties. An example is that attribute errors may come from the classification of remote sensing data. Remote sensing classification is mainly based on spectral characteristics. Sensors' noises, atmospheric disturbance, and limitations of classification algorithms are all liable to classification errors. For example, some pixels may be misclassified for their land use types by employing classification techniques to remote sensing data. These errors can be generally measured by comparing ground data with classification results. A confusion matrix is usually constructed to indicate the percentages of correctly or wrongly classified points.

The existence of mixed pixels also causes uncertainty for remote sensing classifications. It is well known that remote sensing and other raster data are subject to the errors caused by spatial resolution limitations. Remote sensing images are made up of pixels. Each pixel corresponds to a basic sampling unit which records ground information. Conventional remote sensing classification usually assumes the following conditions (Fisher & Pathirana, 1990):

- any one single pixel has exactly one land use type;
- different land use types should have distinct signatures in remote sensing imagery;
- the same land use type should maintain homogenous and stable spectral properties.

In reality, these assumptions are not realistic partly because of the existence of mixed pixels in remote sensing images. A mixed pixel indicates that there is more than one type of land use occupying a single pixel. General methods may have errors in classifying mixed pixels. There are uncertainties when these data are stored in GIS and further used for urban simulation. For example, initial urban areas, which are the key input to urban simulation, may have errors from classification of remote sensing images. Classification errors significantly influence the simulation of urban growth because the errors in initial urban areas can propagate through the simulation process.

2.3. Model uncertainties in urban CA modelling

The error problems of CA models are further exacerbated by model uncertainties. There are other types of errors which are not produced during the process of data capture. These errors come from models themselves due to limited human knowledge, complexity of nature and limitation of technology. In CA simulation, not only do input errors but model errors propagate through the simulation process. Like any computer models, CA models could disagree with reality even when the inputs were completely error-free. CA models are only an approximation to reality. Most of the existing CA models are just loosely defined and a unique model does not exist. Various types of CA models have been proposed according to individuals' perception and preference, and requirements of specific applications. The simulation results are hard to repeat when different CA models are used for the same data set.

3. Evaluation of uncertainties in urban CA

The problem and characteristics of data source errors are well researched (Fisher, 1999; Heuvelink, 1998). The following sections mainly examine the characteristics of error propagation and uncertainties in urban CA modeling.

3.1. Error propagation in urban CA modeling

The assessment of error propagation in urban CA modeling is important for understanding the results of simulation. In urban simulation, initial conditions, parameter values and stochastic factors play important roles in influencing the simulation results. Unexpected features may emerge during CA simulation because of the interactions of various local actions. The simulation could become meaningless to urban planners if the behavior of the automation were completely unstable and unrepeatable. Fortunately, it is found that CA simulation is able to produce stable results at the macroscopic level (Benati, 1997). The general shape of CA simulation remains the same although the configuration may be changed. However, the behaviors of CA simulation are unpredictable to a certain extent at the microscopic level.

Error and uncertainty can propagate through the modeling process. The original errors may be amplified or reduced in the modeling process. All the errors inherent in individual GIS layers can contribute to the final errors of the output during the overlay of these layers. There are many studies to show how such errors propagate in GIS manipulation, such as the common overlay operation (Veregin, 1994). Heuvelink, Burrough, and Stein (1989) present detailed methods to derive error propagation equations in GIS using Taylor series. The advantages of quantitative models are that they are able to yield analytical expressions of error propagation and the computation is not intensive. Another method to analyze error propagation is Monte Carlo simulation which has been widely used in many applications. The advantages include easy implementation and generally applicability, but the disadvantage is the lack of an analytical framework.

Error propagation in CA models is different from that of GIS overlay operations. In GIS operations, mathematical expressions can be given to calculate the errors presented in simple overlay using the logical *AND* and *OR* operators. However, CA models adopt relatively complicated configurations by using neighborhood and iterations. Simulation is a dynamic process in which very complex features can arise according to transition rules. The conversion of the state of a central cell is influenced by the states of its neighborhood. It is almost impossible to develop strict mathematical equations to represent the error propagation in dynamic process. It can be seen from Fig. 1 that error propagation in CA models is quite complicated because of using dynamic looping.

A convenient way to examine error propagation in urban simulation is to perturb spatial variables and assess the error terms in the outcome of simulation. Sensitivity analysis has been used to estimate the effects of error in a database on analytical outcomes in general GIS analysis (Fisher, 1991; Lodwick, 1989). Monte Carlo simulation is often used to perturb spatial data, and then the sensitized data are used to estimate the accuracy of outcomes. Fisher (1991) has presented two algorithms to perturb categorically mapped data, as exemplified by soil map data, and to assess the error propagation.

The Monte Carlo method seems to be most suitable for analyzing error propagation in CA simulation. Standard error propagation theory cannot be used in some models which involve complicated operations (Heuvelink & Burrough, 1993). The Monte Carlo method is a convenient way to study error propagation when mathematical models are difficult to define. Although the Monte Carlo method is very computationally intensive, increasingly this is less problematic because of the advancement of computer technology. When the Monte Carlo method is used, perturbations will be inserted in spatial variables so that the sensitivities of the perturbations in urban simulation can be examined.

The simplest realization of noise is to use the uncontrolled perturbation when detailed knowledge about the errors is unavailable. The perturbation can be carried out to simulate attribute errors for the following spatial data: (a) land use types; (b) initial urban areas; and (c) suitability analysis.

The following experiment is to evaluate the effects of attribute errors on the simulation results. The initial images have two major types of land use—urban areas and non-urban areas. It is expected that the initial image may be subject to classification errors for these two land use types. There is only some general information about the classification errors in most situations. The accuracy of land use classification from satellite remote sensing usually falls within the range of 80–90% (Li & Yeh, 1998). However, the detailed locations of classification errors are not available in most situations.

The first step of the experiment was to perturb the classified satellite images with some errors. The size of the perturbation was based on the above expert knowledge about the classification error of remote sensing. Twenty percent of errors were used for the random perturbation of the classified remote sensing images. A very simple constrained urban CA was used to examine error propagation. The use of too complicated CA will mix the source error with model uncertainties. The model is based the following rule-based structure (Batty, 1997):

IF any cell $\{x \pm l, y \pm l\}$ is already developed
THEN $N\{x,y\} = \sum_{ij\in\Omega} D\{i,j\}$ ($D\{i,j\} = 1$ for a developed cell,
Otherwise $D\{i,j\} = 0$)
&
IF $N\{x,y\} > T_1$ and $R > T_2$
THEN The cell $\{x,y\}$ is developed

where $N\{x,y\}$ is the total number of the developed cell in the neighborhood, T_1 and T_2 are the threshold values, R is a random variable, and cell $\{i,j\}$ are all the cells which from the Moore neighborhood Ω including the cell $\{x,y\}$ itself.

A more sophisticated model is to use the development probability rather than the simple total amount of already developed cell ($N\{x,y\}$) to decide land use conversion. The development probability is usually compared with a random variable to decide if the conversion will take place. Development probability is calculated by the combination of a series of spatial variables, including the number of already developed cells in the neighborhood, distances to urban centres, and distances to roads, etc. However, the definition of the parameter values for these variables is not an easy job. To avoid this uncertainty, our experiments are based on the above simple rule-based structure.

The experiment, which examines error propagation during urban simulation, was carried out in Dongguan in the Pearl River Delta. It simulated the land development in 1988–1993 when rapid urban expansion took place in the region. In this study, the parameter l for the neighborhood size was set to 3. The threshold values of T_1 and T_2 determine how many cells can be converted into urban areas at each time step (iteration). Lower values of T_1 and T_2 allow a larger number of cells to be developed. The value of T_2 falls within the range [0, 1]. It controls the size of perturbation for the random variable. A larger size of perturbation will be introduced when a larger value of T_2 is used.

If the objective is to use the same amount of land consumption, lower values of T_1 and T_2 will require fewer time steps to finish the simulation. Therefore, the values of T_1 and T_2 can be defined according to the amount of land consumption and the time steps to complete the simulation.

The first experiment was to examine the influences of source errors on simulation outcomes. The values of T_1 and T_2 were set to 10 and 0.90 respectively. The model was used to simulate 23,330.5 ha of land development in Dongguan in 1988–1993. The land area of Dongguan is 2465 km^2. The model used a grid of 709×891 cells with a resolution of 50 m^2 on the ground for the simulation. The model run two times—one with the input of original initial urban areas, and the other with the input perturbed by 20% errors. The baseline test was to simulate urban growth using the initial urban areas without error perturbation. The simulation was compared with that of 20% errors perturbed to the initial urban areas. The errors were computed by comparing the simulated results with the actual land development obtained from remote sensing.

Fig. 3 shows some characteristics of error propagation in the simulation process. It is found that the simulation without error perturbation also had errors in the simulation results. It is expected that the higher simulation errors are obtained when the initial urban areas are perturbed with 20% errors at the original land use types. However, the error difference in the outcome is much less than the expected increased error (20%). Fig. 3 shows that 35% error ($t = 10$) is produced in the result when 20% error is inserted, compared to 30% ($t = 10$) when no error is inserted. Therefore, 20% error perturbed only increases 5% error at the result instead of 20%. The conclusion is that the 20% error will not totally pass through the model and source error is significantly depressed in the outcome. It is because the neighborhood functions of CA can reduce the size of error propagation. These neighborhood functions usually sum up the states in the neighborhood and the averaging effects will be introduced in the simulation. The analysis also indicates that all the errors are reduced as the time steps increase in simulation. This is because land available for development becomes

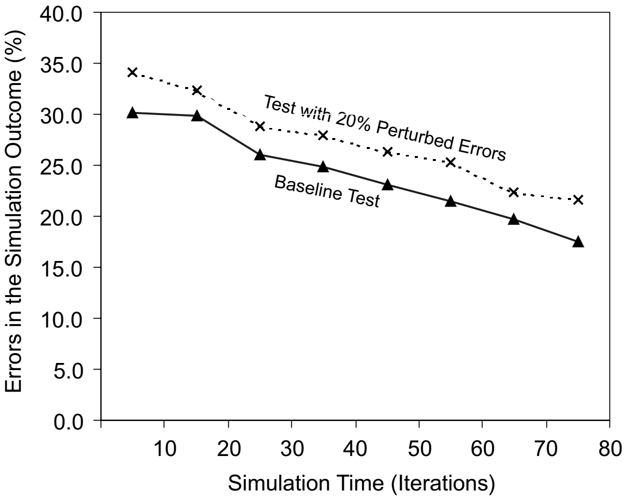

Fig. 3. Error propagation of CA with 20% error perturbed to original initial urban areas.

less as the urban areas grow in size. The simulation is then subject to more constraints which minimize the chance of producing errors.

3.2. Uncertainties within CA models

A series of inherent model errors and uncertainties can be identified for CA models. They are related to the following aspects:

- transition rules;
- neighborhood configuration;
- simulation time;
- stochastic variables.

3.2.1. Transition rules

Transition rules are used to represent a process to be modeled. However, the definition of transition rules is affected by researchers' understanding of the nature. There are many ways to define transition rules in CA models. Different model forms will have impacts on the outcome of CA simulation. A variety of urban CA models has been proposed to tackle specific problems in urban simulation. Model variations are usually dependent on individual preferences and specifications of applications. It is essential to define transition rules while there is no unique way to do so. Substantially different methods have been proposed in defining transition rules. They include:

- the use of five controlling factors (*Diffusion*, *Breed*, *Spread*, *Slope*, and *Road* factors) (Clarke, Gaydos, & Hoppen, 1997);

- estimating development probability based on the analytical hierarchy process (AHP) of multicriteria evaluation (Wu & Webster, 1998);
- defining transition rules with fuzzy sets (Wu, 1999);
- calculating transition potentials using a predefined parameter matrix (White & Engelen, 1993);
- simulating urban conversion using 'grey-values' (Li & Yeh, 2000);
- incorporating planning objectives in urban simulation (Yeh & Li, 2001);
- simulating urban development with neural networks (Li & Yeh, 2002a);
- automatically discovering transition rules through data mining (Li & Yeh, in press).

Development probability is usually defined as a function of a series of spatial variables. These spatial variables can be measured using GIS tools. There are no agreements on how to choose spatial variables for urban simulation. When a series of variables are present, it is not easy to judge which variable is valid for estimating development probability. The selection of variables is a matter of experience. These spatial variables may be correlated and the use of more or less number of variables will affect the outcome of CA simulation (Li & Yeh, 2002b).

Moreover, the ways to measure and standardize these variables will also affect simulation results. GIS are often used to obtain proximity variables, which reflect the influences of sources (centres) on urban growth. For example, a closer distance to a utility (market centre) will have a higher score of attractiveness for urban development. The attractiveness of a centre will decrease as the distance increases. It is straightforward to use the Euclidean distance to indicate the influences of centres. However, a transformed form (e.g. a negative exponential index) may be more appropriate to represent the actual influences of centres. It can more appropriately represent the situation that the influences from centres do not decrease in a linear form as distance increases. A problem with the negative exponential function is that there are uncertainties in defining parameter values.

CA model errors are also introduced by mistakes in assigning parameter values. There are problems on how to determine parameter values. CA models need to use many spatial variables and thus many parameters. For example, White, Engelen, and Uijee (1997) present a CA model to simulate urban dynamics. Their models need to determine as many as $21 \times 18 = 378$ parameter values. Parameter values should be defined before CA models are executed. They have critical influences on the outcome of CA simulation (Wu, 2000). It is quite tedious to define proper parameter values when the number of variables is large. A very simple method to find suitable parameter values is to use the so-called visual test (Clarke et al., 1997). It is based on the trial and error approach in which the impact of each parameter is assessed by changing its value and holding other parameters constant. Wu and Webster (1998) provide another method that uses analytical hierarchy process (AHP) of multicriteria evaluation (MCE) techniques to determine parameter values. The pairwise comparison was used to recover weight vector by which land suitability can be computed. However, the comparison will become very difficult when there is a large set of variables. Moreover, the weights cannot be properly given when variables are correlated.

The above methods have uncertainties because parameter values are decided with subjective influences. Consistent methods should be developed to remove the uncertainties. There is some work on finding optimal parameter values using exhaustive computer search. Clarke and Gaydos (1998) develop a method to find suitable parameter values based on computer search algorithms. It tests various trials of parameter combinations and calculates the difference between the actual data and simulated results for each trial. The parameter values can be found according to the best fit of the trials. The computation is extremely intensive as the possible combinations are numerous. It usually needs a high-end workstation to run hundreds of hours before finding the best fit. It is practically impossible to try all the possible combinations. Computation time will even increase exponentially when there are a larger number of parameters. An alternative to determine parameter values of CA is to train neural networks by using the observation data of remote sensing (Li & Yeh, 2002a). This can significantly reduce the uncertainties in defining the parameter values of CA.

3.2.2. Neighborhood configuration

Neighborhood configuration is a problem of implementing the transition rules into computation models. Transition rules are expected to be independent of the models themselves as much as possible. For example, the distance of influences should not be affected by the resolution of cells. However, uncertainties will arise when computation models are used to implement transition rules. One of these problems is how to configure neighborhood. Cells, which are in the form of discrete space, are the basic unit of CA models. Discrete cells are only the approximation to the continuous space with loss of spatial detail (Fig. 2). There are questions on how to choose proper cell size and cell shape. A large cell size can reduce data volume, but it may lead to the decrease of spatial accuracy. Uniform cells are commonly used because they are simple for calculation. However, irregular cells may be more suitable under particular circumstances (O'Sullivan, 2001). An example is to use irregular cells to represent land parcels or planning units.

It can be calculated by summing or averaging the attribute values of cells within a neighborhood. A simple example is to estimate the conversion probability based on the summation of the total number of a state (e.g. development cells) in a 3×3 window. It is easy to know that the original data errors will be reduced if a large size of neighborhood is used. However, the use of a larger window will also be accompanied by the reduction of spatial detail because of the averaging effects.

The shape of neighborhood can affect the results of CA simulation. There are two common types of neighborhoods—von Neumann neighborhood and the Moore neighborhood. A way to examine neighborhood effects is to see how cities grow under different neighborhood influences. The Moore neighborhood will lead to exponential urban growth which is different from actual growth patterns. The von Neumann neighborhood can be used to reduce the growth rate. However, the two neighborhoods are generally in a rectangle form which has side effects in urban simulation. Instead, a circular neighborhood has better performance than rectangle ones because it treats all directions equally (Li & Yeh, 2000).

3.2.3. Simulation time

CA models adopt discrete time steps for the simulation of urban growth. The discrete time is different from the actual continuous time. There are problems on how to decide the interval of discrete time or the total number of iterations (time steps). The larger the interval of the discrete time is, the smaller the number of time steps becomes. The discrete simulation time of CA is different from continuous real time. For the same amount of land development, different time steps can be used to implement the simulation by properly defining the parameter values. However, the simulation using 100 time steps are not the same as that using 10 time steps for non-linear models.

There is a need to assess the influences of discrete time steps on CA simulation. Temporal errors can be introduced in CA because of using approximate discrete time steps. There is an issue on how to recalculate the parameter values when the time steps are changed. If the development probability is used, the standard practice for changing the transition probabilities related to time steps is as follows: if the old time step is t_s and the new one is T_s ($T_s = n \times t_s$), then the matrix of transition probabilities $\|P\|$ should be substituted by $\|P\|^n$. Since this study is based on the use of a threshold (T_1) to decide land use conversion, the threshold should be changed to allow the same amount of land consumption for the simulation. This threshold can be easily found out through a couple of trials.

An experiment with the same data set of our previous studies (Yeh & Li, 2001) and the same amount of land development for 1988–1993 was carried out to examine the influences of using different time steps (Fig. 4). Fig. 4a uses only 10 time steps to generate the simulation result. It is much different from the actual urban form obtained from remote sensing in Fig. 4d. It is because local interactions are important for generating realistic urban forms. Too few time steps cannot allow spatial details to emerge during the simulation process. An increase in the number of time steps can help to generate more accurate simulation results (Fig. 4b and c). This characteristic is different from that of linear models which do not depend on the choice of the time steps.

3.2.4. Stochastic variables

Most urban CA are not deterministic in simulating complex urban systems. Deterministic models may have problems in representing many geographical phenomena. These phenomena have manifested some unpredictable features which cannot be explained by independent variables because of the complexity of nature. It is almost impossible to forecast exact future patterns by using any kind of computer models. Frequently, urban CA models have to incorporate stochastic variables to represent the uncertainty of nature. Some 'noises' are artificially added to urban CA models by using controlled stochastic variables to produce 'realistic' simulation (White & Engelen, 1993). In the transition rules, calculated development probability is compared with a random number to decide whether the transition is successful or not (Wu & Webster, 1998). This can allow a certain degree of randomness to be inserted in urban simulation. However, there are questions when these models are used for urban planning. It is because each simulation will generate different results

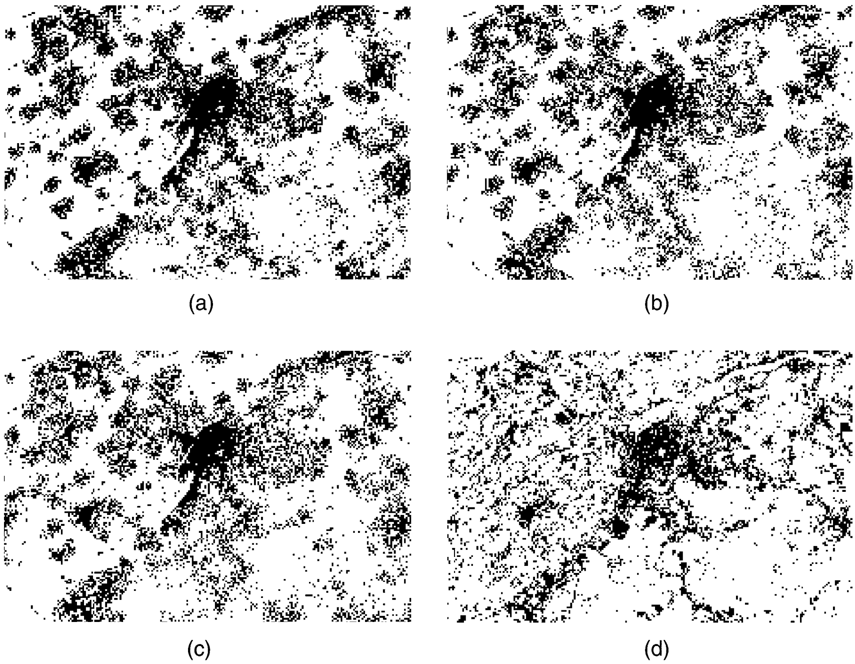

Fig. 4. The influences of discrete time steps on simulation accuracies: (a) simulation from 10 time steps (iterations); (b) simulation from 50 time steps (iterations); (c) simulation from 100 time steps (iterations); (d) actual development.

although the inputs are the same. A planner may be in a dilemma as to which result is suitable for the planning. There is a concern on the repeatability of urban simulation when CA are used for urban planning. The repeatability is crucial if CA are used for preparing development plans.

Two experiments were undertaken to examine the uncertainty of stochastic CA. First, a very simple experiment is to run the CA model twice repeatedly and examine the overlapping percentage of the two simulations. In the overlay analysis, the urban areas are coded with 1, and non-urban areas are coded with 0. If CA are deterministic, the urban areas and non-urban areas in the two different simulations should be the same. They should be 100% overlapping in the overlay. The overlay will only yield two values—2 for urban areas and 0 for non-urban areas. However, the stochastic CA will not generate the same simulation results. The two simulations will not completely overlap in the overlay, yielding three values in the hit count. The hit count of 2 corresponds to urban areas while the value of 0 corresponds to non-urban areas. However, the hit count of 1 just represents the areas of uncertainty. It is urban in one simulation, and becomes non-urban areas in another simulation.

The areas with uncertainty in the simulations should be within a small percentage of the total simulated urban area. Otherwise, the simulations are meaningless. It is interesting to see that the uncertainties mainly exist at the fringe areas of each urban cluster, and consistent simulation results are found in the large part of urban clusters. This means that stochastic CA can maintain stability at the macro-level while they may have subtle changes at the micro-level for each simulation. This characteristic is useful for urban planners to understand the implications of CA urban simulation.

A further experiment is to repeat the simulations 10 times and examine the overlapping of the simulations (Fig. 5). The hit count of 10 corresponds to the urban areas that exist in all 10 repeated simulations. It is also clear that the major uncertainties only exist in the fringe areas of urban clusters. In Fig. 5, the cells with hit count from 1 to 10 are the simulated urban areas with different probability. The cells with a larger value of hit count (e.g. 10) have higher confidence to be urban areas in the simulation.

Fig. 6 further shows the cumulative percentage of overlapping among different hit counts. Different threshold values for the random variable (R) were used in the transition rules to examine the effects of the size of perturbation on the simulation results. The cells with low values of hit count only amount to small percentage of the total simulated urban areas. The cells with hit counts greater than 7 (70% of hits) can amount to as high as 71.8% of the total simulated urban area (Fig. 6). This means that 71.8% of the total simulated urban area can be repeated with a chance of 70% when the CA model runs gain. The simulation results also indicate that the use of a higher threshold value of random variable (R) in the stochastic simulation will result in more uncertainties—lower percentages for higher hit counts.

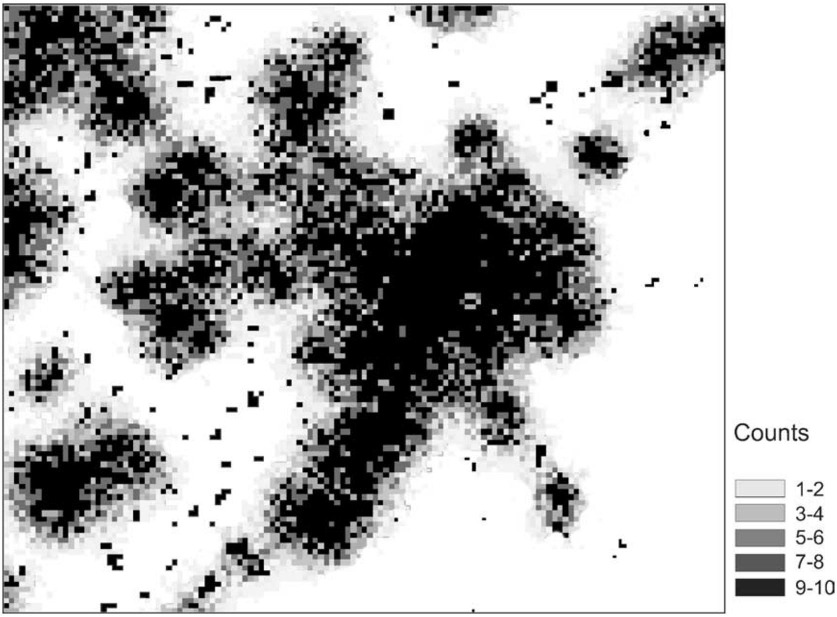

Fig. 5. Overlay of the simulation results by repeatedly running the stochastic CA 10 times.

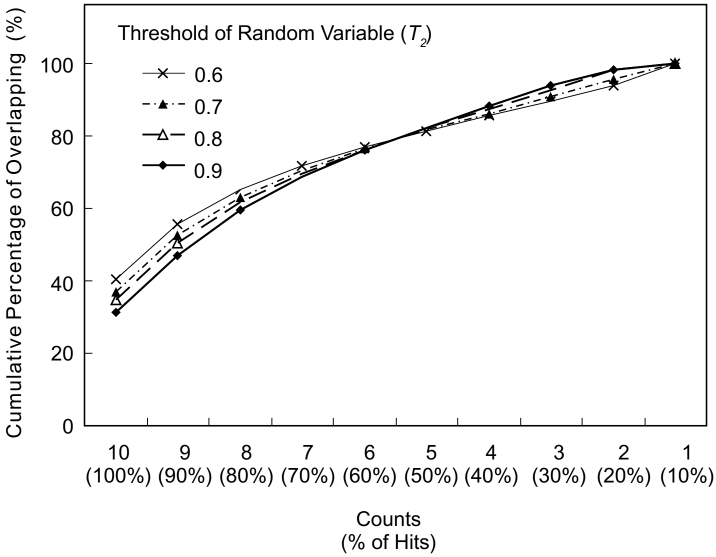

Fig. 6. Cumulative percentages of overlapping areas for 10 repeated simulations using stochastic CA with different sizes of perturbation.

The above experiments provide useful implications for the implementation of planning CA models. If a planner would like to use urban CA simulation to prepare development plans, the model should run a number of times to obtain the probability of simulated urban areas. This can allow them to identify the potential development sites with a high confidence, i.e. only selecting the sites with high hit counts greater than 70%. This method should be useful for producing more reliable simulation results for urban planning.

4. Conclusion

Like many GIS models, urban CA have inherent problems related to data errors and model uncertainties. The issues of data errors and model uncertainties have been well addressed in GIS literature. Although there are many studies on data errors and error propagation in GIS analysis, very few researches have been carried out to examine these issues in urban CA simulation. In this study, experiments have been carried out to examine the influences of errors and uncertainties on urban CA simulation so that it can be better used for projecting and modeling future development in urban planning.

GIS data provide the main inputs to most urban CA models in urban simulation. A large amount of GIS data is usually required for producing realistic urban simulation. It is well known that most GIS data are subject to a series of errors. There are many possibilities of creating errors in spatial data as the errors can come from source maps and even be created during data digitizing. New errors can also be generated in GIS operations. All these errors will propagate in CA simulation and affect the simulation results. There is concern whether urban CA models can produce meaningful results, especially when they are applied to urban planning. Although some researchers may be aware that errors can propagate through CA simulation, they rarely pay much attention to this problem in practice because of its complexity. When GIS data are used as key inputs to CA models, the source errors will propagate and affect the outcomes of simulation. A particular example is the errors in labeling land use types during land use classification. The study shows that errors in the data source can propagate through CA simulation. However, the errors are much reduced in the simulation because of the averaging effects of neighborhood functions and the use of iterations in CA. The error reduction is also caused by the constraints of decreasing land available for development as urban areas grow in size.

Simulation uncertainty is further increased by model uncertainty. The relationship between errors and outcomes is complicated for dynamic models. CA have a series of inherent model uncertainties, which are related to a number of factors in defining CA models—the neighborhood, cell size, computation time, transition rules, and model parameters. Most CA models have incorporated stochastic variables in urban simulation. This has allowed some unpredictable features to be inserted in the simulation process. There are arguments that uncertainty is necessary for generating realistic urban features, such as the emergence of new urban centres during the simulation process. A simple overlay of two repeated simulations from stochastic CA can reveal the discrepancy between them. Fortunately, the discrepancy only exists in the fringe areas of urban clusters according to the experiments. This means that stochastic CA can generate relatively stable simulation results at the macro-level although there are variations at the micro-level. This characteristic is important in the application of stochastic CA models in simulating planning scenarios. Uncertainties are mainly located at the urban fringe but there are relative certainties in areas close to existing urban areas. Planners should run urban CA a number of times repeatedly to obtain probability maps so that potential development sites can be identified. The probability of simulated development can be obtained by overlaying these repeated simulations. Planners can then select the simulated development sites of high probability (confidence).

The issues of data errors, error propagation and model uncertainties are important but often neglected in urban CA models. This paper has examined and addressed some of these issues by carrying out experiments using GIS data. Many model errors are related to model configurations, i.e. how to define a proper model to reflect the real process of urban development. This study demonstrates that some of them, however, are quite unique to CA: (1) data source errors will be reduced during simulation because of the averaging effects of neighborhood functions; (2) simulation errors will decrease with time because the land available for urban development is reduced in constrained urban CA as the urban areas grow in size; (3) enough time steps (iterations) are required to ensure that spatial details can be simulated from CA; and (4) the major uncertainties of simulation are mainly found at the edge of simulated urban areas. These characteristics are quite different from those of general GIS modeling. The study shows that the errors and uncertainties are less severe than what is normally expected from a CA model. The uncertainties of the simulation will be reduced as the simulation continues with time. Moreover, the uncertainties are mainly located at the urban fringe.

The findings of the study can help urban modelers and planners to understand more clearly the characteristics of errors and uncertainties in urban simulation. This is important for preventing the misinterpretation of the modeling results. Further work is needed to develop a methodology for reducing the influences of errors so that more reliable simulation results can be achieved to provide efficient planning tools.

Acknowledgements

This study is supported by the funding from the Research Grants Council of Hong Kong (HKU 7209/02H). We would like to thank the anonymous referees for their very helpful comments and suggestions in improving the earlier version of the paper.

References

Batty, M. (1997). *Growing cities*. Working paper, Centre for Advanced Spatial Analysis, University College London.

Batty, M., & Xie, Y. (1994). From cells to cities. *Environment and Planning B: Planning and Design, 21*, 531–548.

Benati, S. (1997). A cellular automaton for the simulation of competitive location. *Environment and Planning B: Planning and Design, 24*, 205–218.

Burrough, P. A. (1986). *Principles of geographical information systems for land resource assessment*. Oxford: Clarendon Press.

Caspary, W., & Scheuring, R. (1993). Positional accuracy in spatial database. *Computers, Environment and Urban Systems, 17*, 103–110.

Clarke, K. C., Gaydos, L., & Hoppen, S. (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.

Clarke, K. C., & 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.

Fisher, P. F. (1991). Modeling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems, 5*(2), 193–208.

Fisher, P. F. (1999). Models of uncertainty in space data. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), *Geographical information systems* (pp. 177–189). New York: Wiley.

Fisher, P. F., & Pathirana, S. (1990). The evaluation of fuzzy membership of land cover classes in the suburban zone. *Remote Sensing of Environment, 34*(2), 121–132.

Goodchild, M. F. (1991). Issues of quality and uncertainty. In J. C. Müller (Ed.), *Advances in cartography* (pp. 111–139). Oxford: Elsevier Science.

Goodchild, M. F., Sun, G. Q., & Yang, S. H. R. (1992). Development and test of an error model for categorical data. *International Journal of Geographical Information Systems, 6*(2), 87–104.

Heuvelink, G. B. M. (1998). *Error propagation in environmental modelling with GIS*. London: Taylor & Francis.

Heuvelink, G. B. M., & Burrough, P. A. (1993). Error propagation in cartographic modeling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems, 7*(3), 231–246.

Heuvelink, G. B. M., Burrough, P. A., & Stein, A. (1989). Propagation of errors in spatial modeling with GIS. *International Journal of Geographical Information Systems, 3*(4), 303–322.

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

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

Li, X., & Yeh, A. G. O. (2002b). Urban simulation using principal components analysis and cellular automata for land-use planning. *Photogrammetric Engineering & Remote Sensing, 68*(4), 341–351.

Li, X., & Yeh, A. G. O. (in press). Knowledge discovery for geographical cellular automata by data mining. *International Journal of Geographical Information Science*.

Lodwick, W. A. (1989). Developing confidence limits on errors of suitability analysis in GIS. In M. F. Goodchild & S. Gopal (Eds.), *Accuracy of spatial databases* (pp. 69–78). London: Taylor & Francis.

O'Sullivan, D. (2001). Exploring spatial process dynamics using irregular cellular automaton models. *Geographical Analysis, 33*(1), 1–18.

Veregin, H. (1994). Integration of simulation modeling and error propagation for the buffer operation in GIS. *Photogrammetric Engineering & Remote Sensing, 60*(4), 427–435.

Veregin, H. (1999). Data quality parameters. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), *Geographical information systems* (pp. 177–189). New York: Wiley.

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

White, R., & 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., & 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 as models of complexity. *Nature, 31*(4), 419–424.

Wu, F. (1999). A linguistic cellular automata simulation approach for sustainable land development in a fast growing region. *Computer, Environment, and Urban Systems, 20*(6), 367–387.

Wu, F. (2000). A parameterised urban cellular model combining spontaneous and self-organising growth. In P. Atkinson & D. Martin (Eds.), *GIS and geocomputation* (pp. 73–85). New York, NY: Taylor & Francis.

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

Yeh, A. G. O., & Li, X. (2001). 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., & Li, X. (2002). A cellular automata model to simulate development density for urban planning. *Environment and Planning B, 29*, 431–450.