Error Propagation and Model Uncertainties of Cellular Automata in Urban Simulation with GIS

Anthony Gar-On Yeh[1] and Xia Li[1,2]

[1]Centre of Urban Planning and Environmental Management, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, P.R. China (hdxugoy@hkucc.hku.hk)
[2] School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, P.R. China (xlib@gis.sti.gd.cn; lixia@graduate.hku.hk)

Abstract

Errors and uncertainties are important issues in most geographical analyses and modelling processes. It is well known that errors popularly exist in GIS data. A variety of methods have been developed to measure error propagation in GIS. Cellular automata (CA) have been increasingly used for modelling geographical phenomena, such as urban development. 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 the approximation of reality. These uncertainties also have impacts on the outcome of urban simulation. Identification and evaluation of these errors and uncertainties are crucial for the understanding and utilization of the simulation results of CA. This paper attempts to address these important issues which have not been well dealt with before. This can help urban modelers and planners to understand more clearly the meanings and implications of CA modelling.

1. Introduction

Errors and uncertainties are important issues in 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 complexity of nature. GIS databases are the approximations of real geographical variations with very limited exceptions (Goodchild et al., 1992). Understanding of errors and uncertainties of GIS is needed 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 for simulating complex systems in physics, chemistry and biology. Recently, a series of urban CA have been developed for modeling complex urban systems with the integration of GIS. The application of CA in urban modeling can give insights into a wide variety of urban phenomena. Urban CA have better performance in simulating urban development than conventional urban models that use mathematical equations. Urban CA have much simpler forms, but produce more meaningful and useful results than mathematical-based models. Temporal and spatial complexities of urban development can be well simulated by properly defining transition rules in CA models. CA are capable of providing important information for understanding urban

theories, such as the emerging and evolution of forms and structures. They are also used as planning model for plan formulation.

Although many studies have been reported in urban simulation, the errors and uncertainties of CA have not attracted much attention so far. This issue should be important because a huge volume of geographical data is usually used in urban simulation, especially in modelling real cities. Spatial variables are usually retrieved from GIS and imported to CA modeling processes. It is well known that most GIS data are affected by a series of errors. Like many GIS models, urban CA simulation is not without problems because of the inherent 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. Although there are many studies on error types and error propagation in GIS, little research has been carried out to examine the issue in CA simulation. This paper attempts to evaluate the influences of errors and uncertainties in urban simulation by carrying out some experiments. The study is expected to provide useful information about the adequacy of urban CA for urban planning and spatial decision-making. It can contribute to the recent development of urban CA because the issue has not been addressed.

2. Errors and Uncertainties in Urban CA

Spatial modeling with GIS is an important topic in researches and applications in geography. It uses GIS powerful functions to simplify mathematical representation of reality. In recent years, a class of dynamic spatial modeling is developing very rapidly by the integration of CA with GIS. CA are dynamic spatial models which have powerful capabilities in modeling complex systems in physics, chemistry, biology and geography. Particularly, CA and GIS have been used to simulate urban systems for testing urban theories (Wu and Webster, 1998; Webster and Wu, 1999) and formulating development plans for urban planning (Li and Yeh, 2000; Yeh and Li, 2001; Yeh and Li, 2002).

A major concern for urban CA models is their errors and uncertainties if they are applied to real cities. CA models for geography and urban simulation are significantly different from Wolfram's deterministic CA models (Wolfram, 1984). Wolfram's models have strict definitions and use very limited data. This allows CA models to produce stable outputs without any error and uncertainty. However, urban CA models usually need to input 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 (Figure 1).

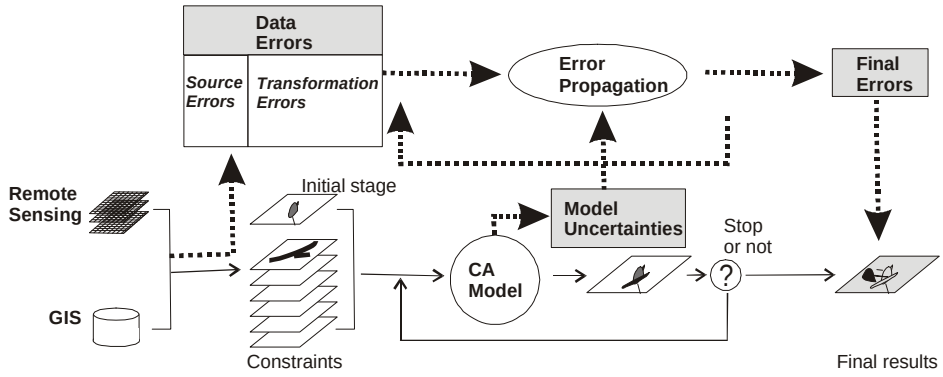


Figure 1. Data errors, model uncertainties, and error propagation in cellular automata

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:

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 recoded location. Euclidean distance can be used for the measurement. 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. It is calculated in both x- and y-directions.

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 the band and the distribution are uniform. However, recent studies show that both the band and distribution might be non-uniform (Caspary and Scheuring, 1993; Veregin, 1999).

2.1.2 Attribute Errors

Attribute 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 also contribute to the uncertainty of urban simulation when these attributes are used in CA models.

There are concerns on how data source errors contribute to uncertainties in the results of GIS operations and computational models. Yet it is only relatively recently that much attention has been paid to the problems of data source errors and error propagation in GIS (Heuvelink et al., 1989). There are attempts to use quantitative models to determine the magnitude of error propagation in GIS. For example, Heuvelink et al. (1989) present detailed methods to derive error propagation equations using Taylor series. The advantages of quantitative models are that they are able to yield analytical expression 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.

Unlike Wolfram's traditional CA models, urban CA models are usually implemented in heterogeneous space by importing spatial information from GIS. The simulation of real cities needs to use many spatial data and the simulation results are very sensitive to data errors. Spatial data quality should be a major concern in urban simulation. There are a number of errors that may be present in source data. Although GIS data are stored in digital formats, they are not error-free. GIS database are often created by means of digitizing paper maps. Errors will be introduced during the transfer of the source map to the digital database. Most GIS databases do not have information about the error of the source maps. When no accuracy information is kept, it becomes extremely difficult to evaluate the true accuracy of the final results of modeling (Heuvelink, 1998).

2.2 Transformation or Operation Errors

Besides data source errors, there are also new errors from common GIS transformation. Some standard GIS operations have to be carried out to generate additional or specific information that is not already stored in GIS as inputs to CA models. GIS databases only contain basic data for storage efficiency. User-specific information may be produced by standard GIS operations, e.g. data conversion, adding, buffering and masking. The transformation 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 implemented using a 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 first 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 the loss of spatial details.

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

The 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 and Webster, 1998). These operations can 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 urban development (Li and Yeh, 2000). The purpose is to prohibit uncontrolled urban development according to the constraint information provided by GIS. A series of resource and environmental factors can be retrieved from GIS database and imported to CA models as site attributes. These factors may include topography, land use types, proximity and agricultural productivity (Li and Yeh, 2000). Constraint scores can be calculated by using GIS linear or non-linear transformation functions. However, there are uncertainties in defining the forms of transformation functions.

Errors can also be produced during proximity analysis or buffer analysis in 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. Urban development probability is estimated based on the attractiveness for urban development. It is more attractive for urban development if a site has better accessibility to major transport networks or

facilities. Some distance variables are used to represent the attractiveness, including various distances to roads, railways, town centers, hospitals and schools. These variables can be conveniently calculated from GIS layers of corresponding points and lines. 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 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 for 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 generally be measured by comparing ground truths with classification results. A confusion matrix can be constructed to indicate the percentages of correctly and wrongly classified points.

The existence of mixed pixels also brings about the uncertainty in remote sensing classification. It is well known that remote sensing and other raster data are subject to the errors caused by 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 assumes the following conditions (Fisher and Pathirana, 1990):

- All pixels are purely occupied by a land use type;
- Any one single pixel is just entitled to one land use type;
- Different land use types should generate a distinct signature.

In reality, these assumptions are not true because of the existence of mixed pixels. 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 for urban simulation may be obtained from classification of remote sensing images. Classification errors can significantly influences the simulation of urban growth because the errors can propagate through the simulation process.

2.3 Model Uncertainties in Urban CA Modelling

The error problems of CA models are further exacerbated by taking into account model uncertainties. There are other types of errors which are not produced physically during the process of data capture. These errors come from models themselves due to poor human knowledge, complexity of nature and limitation of technology. In CA simulation, not only input errors propagate through the simulation process, but model errors as well. Like any computer models, CA models could disagree with reality even when the inputs were completely error-free. CA models are only 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.

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

- Discrete entities in space and time;

- Neighborhood definitions (types and sizes);
- Model structures and transition rules;
- Parameter values;
- Stochastic variables

3. Evaluation of Errors and Uncertainties of Urban CA

3.1 Error Propagation in CA Modeling

Assessment of error propagation in 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 simulation results. Unexpected features can emerge during CA simulation because of the interplay of various local actions. CA simulation may become meaningless if the behavior of the automation is completely unstable and unrepeatable. Fortunately, it is found that CA simulation can produce stable results at 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. There are many studies to show how such errors propagate in GIS manipulation, such as the common overlay operation (Veregin, 1994). 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.

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. CA models adopt relatively complicated configuration by using neighborhood and iterations. The simulation is a dynamic process in which very complex features can arise according to transition rules. The transition of states is influenced by the states in neighborhood. It is almost impossible to develop strict mathematical equations for the error propagation in the dynamic process. It can be seen from Figure 1 that error propagation in CA models is quite complicated because of the use of 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 establish the effects of error in database on analytical outcomes in general GIS analysis (Lodwick, 1989; Fisher, 1991). Monte Carlo simulation is often used to sensitize spatial data, and then the sensitized data are used to determine 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 the study of error propagation in CA simulation. Standard error propagation theory cannot be used in some models which involve complicate operations (Heuvelink and 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 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 Monte Carlo simulation should have more advantages because explicit mathematical equation cannot be built for urban CA models.

A simplest realization of noise is to use the uncontrolled perturbation which assumes no knowledge exists about the errors. The perturbation can be carried out to simulate attribute errors for the following spatial data that are used as the main inputs to urban CA models: a) land use types; b) initial urban areas; c) suitability analysis.

The Monte Carlo method can be used to assess the influences of attribute errors of cells on the simulation results. Urban simulation is based on the attributes of each cell. These attributes may or may not be correct due to data source errors. Initial land use types for urban simulation are usually obtained from the classification of remote sensing data. Classification errors have effects on the final outcome of urban simulation. For example, some of the initial urban areas may be wrongly located because of misclassification. This means that the original states have errors and the errors can be propagated in urban simulation. The final results can be influenced by the errors inherent in the data sources. Since CA simulation is based on neighborhood functions, the functions should control the process of error propagation. The error propagation in CA should be examined to ensure the successful application of CA in urban simulation.

The following experiment is to evaluate the impacts of the 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 and 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. 20% errors were randomly generated in the classified remote sensing images since there is no prior knowledge about the spatial locations of the errors. Then, a very simple urban CA was used to examine the error propagation. The use of too complicated CA cannot isolate the effects of model uncertainties. The model is based the following rule-based structure (Batty 1997):

IF any cell $\{x\pm1, y\pm1\}$ is already developed
THEN $P_d\{x,y\}=\sum_{ij \in \Omega} P_d\{i,j\}/8$
&
IF $P_d\{x,y\}$ > some threshold value
THEN cell$\{x,y\}$ is developed with some other probability $\rho\{x,y\}$

where $P_d\{x,y\}$is urban development probability for cell $\{x,y\}$, cell $\{i,j\}$ are the all cells which from the Moore neighbourhood Ω including the cell $\{x,y\}$ itself.

Dongguan in the Pearl River Delta is chosen for carrying out the experiment to examine error propagation during urban simulation. The experiment is very simple by just running the model twice and then comparing the errors by overlay analysis. The baseline is the simulation based on the initial urban areas without error perturbation. It is compared with that with 20% errors perturbed to the initial urban areas. Figure 2 shows the error propagation during the simulation. The simulation without error perturbation also has errors for the simulation results. Higher simulation errors were obtained when the initial urban areas were perturbed with 20% errors at original land use types. However, the increased errors are much less than 20%. The increased errors only amount to about 5%. This means that the perturbed errors (20%) have been significantly reduced during the simulation process. It is because CA adopt neighborhood functions which have average effects to reduce the errors. The analysis also indicates that all the errors will be reduced with time. This is

because land available for development will be reduced as the urban areas grow. The simulation will then be subject to more constraints which minimize the chance of producing errors.

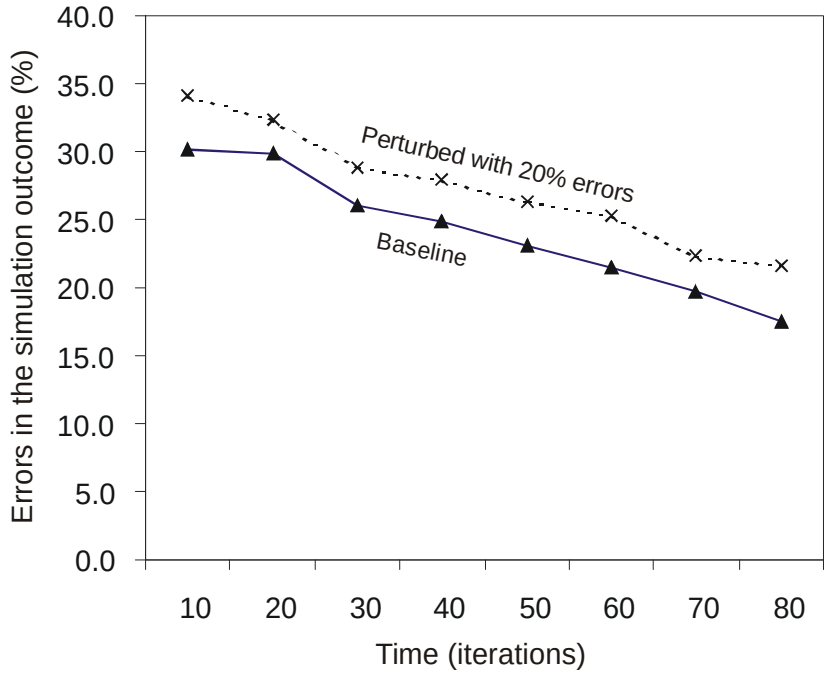


Figure 2. Error propagation of CA with 20% error perturbed to original initial urban areas

3.2 Model Uncertainties in CA Modeling

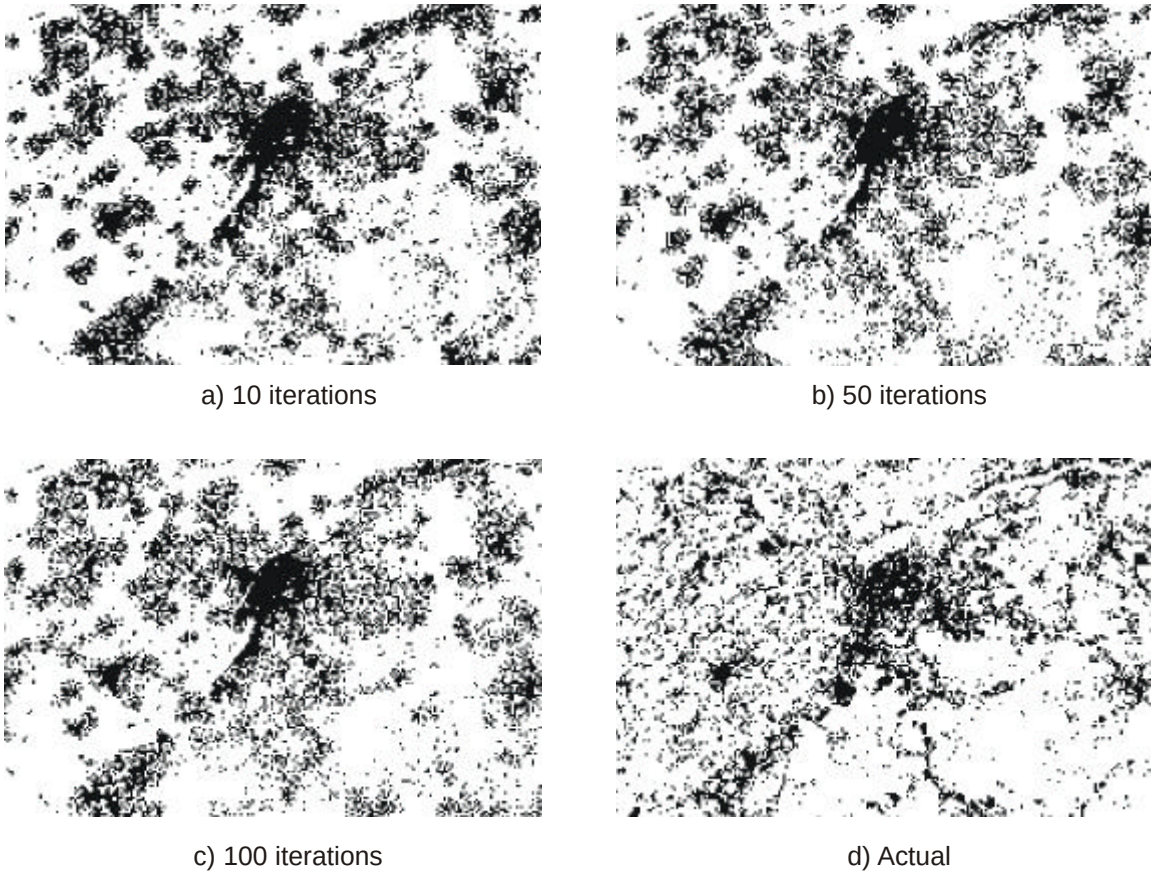
3.2.1. Discrete Space and Time

CA models are implemented using discrete space and time. Cells, which are in the form of discrete space, are the basic unit of CA models. However, discrete cells are only the approximation to the continuous space with loss of spatial details. There are questions on how to choose proper cell size and cell shape. A large cell size may be preferable for reducing data volume, but it may reduce 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.

CA models adopt discrete time (iterations) to represent actual time in simulation. There are problems on how to decide the interval of discrete time (the total number of iterations). The larger the interval of the discrete time is, the smaller number of iterations becomes. The discrete simulation time of CA is different from the continuous real time. The outcome of simulation from 100 iterations is not the same as that from 10 iterations. There is a need to assess the influences of discrete time on CA simulation. Temporal errors can be introduced in CA because of the use of approximate discrete time.

Figure 3 is the experiment results which clearly show the effect of discrete time on urban simulation. Figure 3a only uses 10 iterations to generate the simulation result. It is much different from the actual urban form that is obtained from remote sensing in Figure 3d. It is because local interactions

are important for generating realistic urban forms. Too few iterations cannot allow spatial details to emerge during the simulation process. The increase in the number of iterations can help to generate more accurate simulation results (Figure 3b and 3c).



a) 10 iterations b) 50 iterations

c) 100 iterations d) Actual

Figure 3. The influences of discrete time on simulation accuracies

3.2.2 Neighborhood Configuration

The most important notion of cellular automata is the so-called neighborhood function. The neighborhood function is defined to estimate the conversion probability from one state to another. The neighborhood function is based on a series of neighborhood operations. It is usually obtained by summing or averaging the values of input attributes at the neighborhood. A simple example is to estimate the conversion probability based on the summation of the total number of a state (e.g. land use type) in a 3×3 window. It is easy to perceive that the original data errors will be reduced for a large size of neighborhood. However, this will also lose the spatial details due to the averaging effects.

Neighborhood configuration affects the results of CA simulation. There are two common types of neighborhoods – von Neumann neighborhood and the Moore neighborhood. A way to examine the 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.

3.2.3 Model Structures and Transition Rules

It is expected that different model forms will have impacts on the outcome of CA simulation. A variety of urban CA models have been proposed to tackle specific problems in urban simulation. Model variations are usually related to individual preferences and the requirements of applications. It is essential to define transition rules which are the core of CA models. However, it is quite relax to determine transition rules because there is no unique way to do so. Substantially different methods have been proposed for defining transition rules. They include:

- Simulating the births, survivors and death of cells with the notion of the game of *Life* (Batty and Xie, 1994);
- Estimating development probability based on the analytical hierarchy process (AHP) of multicriteria evaluation (Wu and Webster, 1998);
- Defining transition rules with fuzzy sets (Wu, 1999);
- Calculating transition potentials using a predefined parameter matrix (White and Engelen, 1993);
- Simulating urban conversion using 'grey-values' (Li and Yeh, 2000);
- Incorporating planning objectives in urban simulation (Yeh and Li, 2001);
- Calibrating and simulating urban development with neural networks (Li and Yeh, 2001)

Development probability is the function of a series of spatial variables. These spatial variables are usually measured using GIS tools. There is also a controversial issue on how to choose variables. When a series of variables are present, there is difficulty to judge which variable should be selected or removed from CA models. The selection of variables is a matter of experiences. The use of more or less number of variables will affect the outcome of CA simulation.

Moreover, the ways to measure and standardize these variables will also affect simulation results. An example is to obtain proximity variables using GIS functions. GIS functions are used to calculate the influences of a source (centre). 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 better represent the situation that the influences from centres do not decrease in a linear form as distance increase. The problem is that there are uncertainties in defining parameter values for the negative exponential function.

3.2.4. Parameter Values

CA model errors are also introduced by mistakes in assigning parameter values. It is a debatable issue on how to define parameter values. CA models need to use many spatial variables and thus many parameters. For example, White et al. (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 can be executed. Parameter values 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 decide parameter values. The pairwise comparison was used to recover weight vector by which suitability of the land can be computed. However, the comparison will become much difficult when there is a large set of variables. Moreover, the weights cannot be properly given when there are relevant variables.

These methods have uncertainties because parameter values are decided with subjective influences. Objective methods should be used to remove the uncertainties. There is some limited work for finding optimal parameter values using exhausted computer search. Clarke and Gaydos (1998) develop a relative robust 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 exponentially increase when there are a larger number of parameters. A more robust way is to train neural network and find out the parameter values of CA by using the observation data of remote sensing (Li and Yeh, 2001). This can significantly reduce the uncertainties in defining the parameter values of CA.

3.2.5 Stochastic Variables

Most urban CA are not deterministic for 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 due to 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 and 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 and 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 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 using stochastic variables and the use of urban simulation results in preparing land use 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 from an overlay analysis (Figure 4). 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 percent of 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 be totally overlapping and the overlay will yield three values in the hit count. The hit counts of 2 and 0 in the overlay represent the urban areas and non-urban areas respectively. However, the hit count of 1 is urban areas in one simulation while non-urban areas in another simulation. Therefore, the hit count of 1 represents the areas with uncertainty.

The areas with uncertainty in the simulations should be within a small percentage of the total simulated urban area. Otherwise, the simulations are meaningless. According to the experiment, it is

interesting to see that the uncertainties only mainly exist at the fringe areas of each urban cluster. Stochastic CA can produce consistent simulation results in the core urban areas. This means that stochastic CA can maintain stability at macro-level while they may have subtle changes at micro-level for each simulation. This characteristic should be useful for urban planners to understand the implication of CA urban simulation.

A further experiment is to repeat the simulations ten times and examine the overlapping of the simulations (Figure 5). The hit count of 10 corresponds to the urban areas that exist in all ten repeated simulations. It is also clear that the major uncertainties only exist in the fringe areas of urban clusters. In Figure 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.

Figure 6 further shows the distribution of the simulated urban areas among the different hit counts. Different values for the random variable (R) are also used to examine the effect of disturbance on the simulation. It can be found that 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. This means that 71.8% of the total simulated urban area can be repeated with a chance of 70% when the CA model is run gain. The simulation results also indicate that the use of a higher value of random variable (R) in the stochastic simulation will result in more uncertainties in the fringe areas. The distribution curve tends to be more flatten. As the result, there are less percentages of the total simulated urban areas for the higher hit counts.

The above experiments provide useful implications for developing urban-planning CA models. If a planner would like to use urban CA simulation to prepare land development plan, the model should be run a least a couple of times. 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.

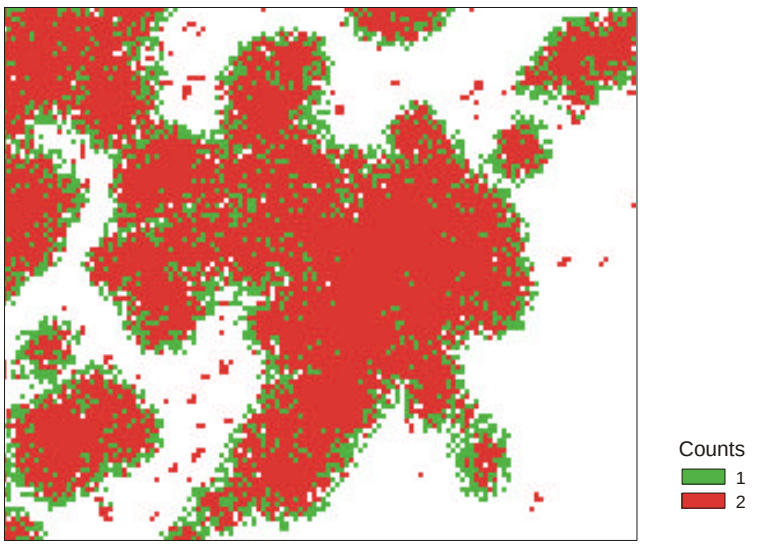


Figure 4. Overlay of the simulation results by repeatedly running the stochastic CA twice

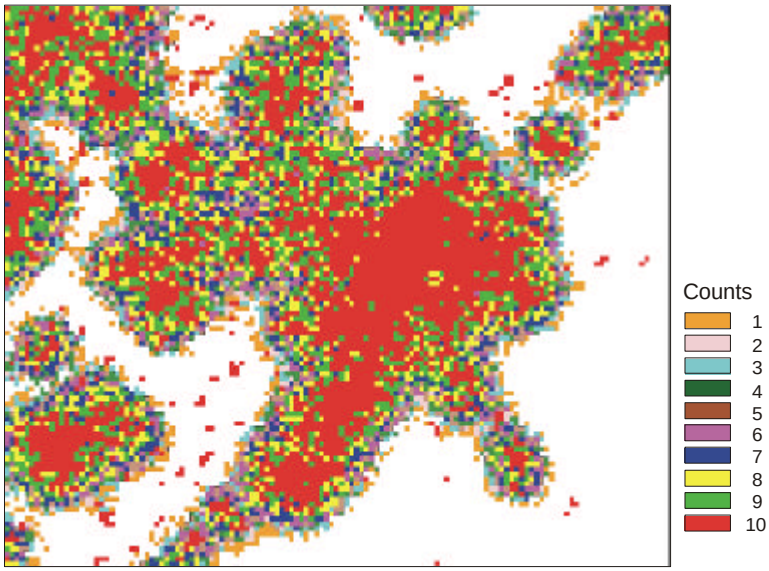


Figure 5. Overlay of the simulation results by repeatedly running the stochastic CA ten times

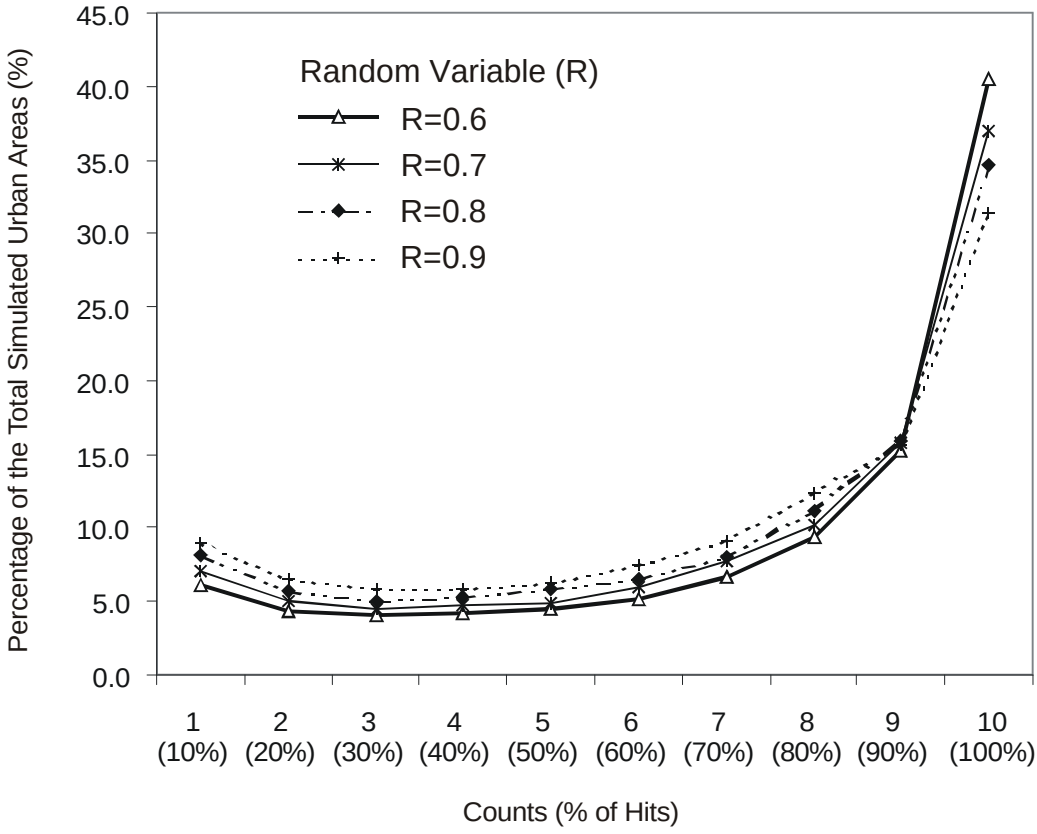


Figure 6. The distribution of the simulated urban areas among different hit counts by ten repeated simulations with different random variable (R) values

4. Conclusion

Like many GIS models, urban CA have the inherent problems related to data errors and model uncertainties. This study demonstrates that the outcomes of urban simulation are affected by the errors of GIS data and the structures of CA models. The errors and uncertainties will in turn affect planning and development decisions when the results of urban CA simulation are applied in the planning and development process. Although there are many studies emphasized on data errors and error propagation in GIS analysis, not much research has been carried out to examine these issues in urban CA simulation.

GIS data are the main inputs to most urban CA models. 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 original sources and even be produced in the process of data capturing. New errors can also be created during GIS operations.

These errors will propagate in CA simulation and affect the simulation results. There is concern whether CA models can produce reliable and repeatable results, especially when it is 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 the complexity. When GIS data are used as 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 experiment shows that the errors in the initial land use types can propagate through CA simulation. However, the errors will be much reduced in the simulation outcomes because of the average effects of the neighborhood and iterations of CA.

Simulation uncertainty is further worsened by model uncertainty. The relationship between errors and outcomes is much complicated for dynamic models. CA also have a series of inherent model uncertainties. These uncertainties 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 centers 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. It means that stochastic CA can generate stable simulation results at the macro-level although there are variations at the micro-level. This characteristic is important to ensure the applicability of stochastic CA in simulating planning scenarios. Therefore, planners should run urban CA a couple of times repeatedly when CA are used for selecting development sites. Planners can then select the simulated development sites with more certainty.

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 with GIS data. The analysis can help to understand the implications of CA simulation for urban planning. However, further work is needed to develop methodology for reducing the influences of errors and producing more reliable simulation results.

5. References

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

BATTY, M., and Y. XIE, 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., and SCHEURING, R., 1993. Positional accuracy in spatial database. *Computers, Environment and Urban Systems*, 17, 103-110.

CLARKE, K. C., and L. J. GAYDOS, 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., L. GAYDOS, and S. HOPPEN, 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.

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., and 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., and 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., and 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., and STEIN, A., 1989. Propagation of errors in spatial modeling with GIS. *International Journal of Geographical Information Systems*, 3(4), 303-322.

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.

----------, 2001. Calibration of cellular automata by using neural networks for the simulation of complex urban systems, *Environment and Planning A*, 33: 1445 -1462.

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., 1994. Data quality parameters. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind, *Geographical information systems* (pp 177-189). New York: Wiley.

WEBSTER, C. J., and Wu, F. 1999. Regulation, land-use mix, and urban performance. Part 2: simulation. *Environment and Planning A*, 31: 1529-1545.

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 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, *GIS and Geocomputation* (P. Atkinson, and D. Martin, editors), Taylor & Francis, New York, NY, pp. 73-85.

WU, F., and 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. and LI, X., 2002. A cellular automata model to simulate development density for urban planning, *Environment and Planning B*, 29: 431-450.

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