Integration of principal components analysis and cellular automata for spatial decisionmaking and urban simulation

LI Xia (黎 夏)[1] & YEH Gar-On (叶嘉安)[2]

1. Guangzhou Institute of Geography, Guangzhou 510070, China;
2. Centre of Urban Planning and Environmental Management, The University of Hong Kong, Hong Kong SAR, China
Correspondence should be addressed to Li Xia (email: lixia@graduate.hku.hk)

Received November 8, 2001

Abstract This paper discusses the issues about the correlation of spatial variables during spatial decisionmaking using multicriteria evaluation (MCE) and cellular automata (CA). The correlation of spatial variables can cause the malfunction of MCE. In urban simulation, spatial factors often exhibit a high degree of correlation which is considered as an undesirable property for MCE. This study uses principal components analysis (PCA) to remove data redundancy among a large set of spatial variables and determine 'ideal points' for land development. PCA is integrated with cellular automata and geographical information systems (GIS) for the simulation of idealized urban forms for planning purposes.

Keywords: principal components analysis, cellular automata, geographical information systems, urban simulation.

Cellular automata (CA) were first introduced in 1948 by von Neumann and Ulam to model complex dynamic systems, such as biological reproduction and crystal growth. Although CA models only use very simple rules, they can generate very complex behavior and global structures. In this way, the role of local rules can be compared to that of DNA in life sciences. CA models have been increasingly used in the simulation of complex systems, such as biological reproduction, chemically self-organizing systems, propagation phenomenon, and human settlements[1, 2].

CA are quite suitable for the simulation of land use changes and evolution of urban systems because of their powerful spatial modeling capabilities. In recent years, many studies on urban CA models have been reported with interesting outcomes[2−7]. CA models can be used for testing hypotheses and theories, such as fractal properties and the evolution of dynamic systems. The integration of GIS and CA can help to solve complex decision problems as they can benefit from each other. A series of constraints can be defined and obtained from GIS to address environmental concerns so that sustainable cellular cities can be simulated[5, 8]. Multiple criteria evaluation techniques (MCE) can be incorporated into CA models to deal with various complex spatial variables in urban simulation[4].

Numerous complex and conflicting factors are involved in spatial analysis and decisionmaking processes. Multicriteria evaluation techniques (MCE) can be employed to handle a number of criteria in decisionmaking[9]. MCE techniques began to emerge to solve decisionmaking and planning problems in the early 1970s[10]. The planning process is becoming more complicated in technical, physical, social and economic aspects. MCE can be used for analyzing the complex trade-

offs between different alternatives. MCE typically requires that the evaluation criteria be independent of each other. A high degree of correlation between evaluation criteria is considered as an undesirable property for decisionmaking[11].

This paper discusses the issues about the correlation of spatial variables in urban simulation. Principal components analysis (PCA) is used to remove data redundancy. PCA is among the most widely used methods for spatial data handling, owing to its simplicity and straightforward interpretation. It can transform a set of correlated variables into uncorrelated orthogonal variables. This paper examines the integration of PCA and CA models in reducing data redundancy among a large set of spatial variables for urban planning.

1 Principal components analysis and cellular automata for urban simulation

It is difficult to determine weights when many factors are involved. It is inadequate to carry out CA simulations based on the direct use of MCE when there are correlated spatial variables. The correlation of factors may result in the malfunction of the weighting for MCE by 'double counting' similar variables. Principal components analysis (PCA) can be integrated in CA simulation to tackle the problem of correlation among many layers of spatial data. PCA is a linear transformation of data which rotates the axes of variable space along lines of maximum variance. The transformation is based on the following equation[12]:

$$pc_{ij} = \sum_{k=1}^{n} X_{ik} E_{kj}, \tag{1}$$

where pc_{ij} is the component score of the jth principal component for cell i, X_{ik} is the value of the kth criterion or layer for cell i, and E_{kj} is the element of the eigenvector matrix at row k and column j.

The eigenvectors and eigenvalues for the linear transformation are mathematically derived from the covariance matrix by the following equation:

$$E\ \mathrm{Cov}E^{\ T} = V, \tag{2}$$

where Cov is the covariance matrix, V is the diagonal matrix of eigenvalues, E is the matrix of eigenvectors, and T is the transposition function.

Independent compressed components can be produced by PCA and used for CA simulation. This can help to solve the problems for general MCE methods in dealing with correlated variables. PCA can be integrated with CA for better urban simulation. Standard cellular automata may be given by the neighbourhood function[8].

$$S^{\ t+1} = f(S^{\ t}, N), \tag{3}$$

where S is a set of all possible states of the cellular automata, N is a neighbourhood of the cells providing input values for the function f, and f is a transition function that defines the change of the state from time t to $t+1$.

CA models usually use discrete states for simulation. Traditionally, CA simulation only uses a binary value to address the status of conversion based on the estimated probability. The prob-

ability of conversion is calculated based on some kind of neighborhood function. Usually, the probability is further compared with a random value to decide whether a cell is converted or not (1 for converted and 0 for non-converted). In our model, the status of a cell has a continuous 'grey value' between 0—1 to represent the stepwise selection or conversion process. A cell will not be suddenly 'selected' or converted for land development. The 'grey value' is calculated based on the cumulative equation.

$$G_i^{t+1} = G_i^t + \Delta G_i^t, \tag{4}$$

where G^t is the 'grey value' for development which falls within the range of 0—1 at time t, and i is the location of the cell. The simulation will stop when t reaches the final time T^0. A candidate cell will not be regarded as a developed cell until its 'grey value' reaches 1.

The increase of the 'grey value' is based on the neighborhood function and the similarity between a candidate cell and the 'ideal point'. The first part is the traditional neighborhood function which counts the number of developed cells in the neighborhood. There is a higher probability for conversion when a cell is surrounded by a larger number of developed cells[13]. The second part is related to the similarity between a candidate cell and the 'ideal point'. The 'ideal point' can produce the best benefit if it is developed. Development suitability can be obtained based on various criteria using land evaluation[14]. The 'ideal point' should achieve the maximum scores for all criteria. A cell with a larger value of similarity with the 'ideal point' means that the cell is more similar to the 'ideal point' and a higher growth rate of 'grey value' should be applied to the cell proportionally.

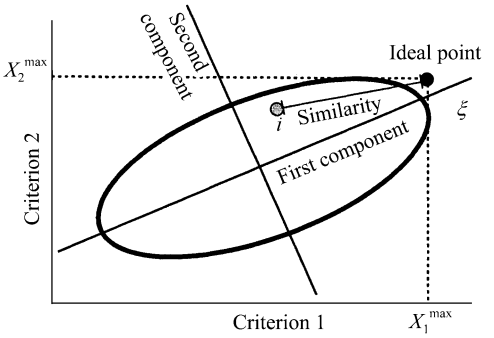

Fig. 1. Principal components transformation and the 'ideal point'.

The 'ideal point' should have the best criterion scores for all criteria (fig. 1). The 'ideal point' in the variable space can be expressed as

$$\xi = (X_1^{\max}, X_2^{\max}, ..., X_j^{\max} ..., X_K^{\max}), \tag{5}$$

where $X_j^{\max}$ is the maximum score for the jth criterion.

In fact, the 'ideal point' is a virtual point. Its transformed coordinate in components space can be obtained using eq. (1). A series of factors for environmental protection and sustainable development can be incorporated in the model by using the 'ideal point' approach. A candidate cell that is more similar to the 'ideal point' in terms of site attributes will have a faster rate of urban growth. This can ensure that greater benefits can be achieved. As mentioned before, the attributes have been compressed into a few major principal components, but they still contain the most original information. The principal components are then used to calculate the similarity based on a form of Euclidean 'distance' given by

$$d_{i\xi} = \sqrt{\sum_{j}^{m} w_j^2 (pc_{ij} - pc_j^0)^2}, \tag{6}$$

where $d_{i\zeta}$ is the 'distance' between cell i and 'ideal point' ξ based on the attributes of m components, pc_{ij} is the value of jth component for cell i, w_j is the weight for the jth component, and pc_j^0 is the transformed score of the 'ideal point' for the jth principal component.

The similarity (SIM) is given by

$$\mathrm{SIM} = 1 - \frac{d_{i\xi}}{d_{i\xi}^{\max}}, \tag{7}$$

where $d_{i\xi}^{\max}$ is the maximum value of $d_{i\zeta}$.

The increase of 'grey value' should be proportional to the neighborhood function and the similarity. There holds

$$\begin{aligned} \Delta G_i^t &= f_i(q^t, N) \times \mathrm{SIM}^t \\ &= \frac{q^t}{\pi\, l^2} \times \left(1 - \frac{d_{i\xi}^t}{d_{i\xi}^{\max}}\right)^k, \end{aligned} \tag{8}$$

where q^t is the total amount of developed cells in the neighborhood N at time t, l is the radius of the circular neighborhood, and k is the parameter for power transformation.

The parameter k is used to generate more discriminated growth results[4,5,8]. A stochastic disturbance term is also added to represent unknown errors during the simulation. This can allow the generated patterns to be more close to reality[3]. The error term (RA) can be given by

$$RA = 1 + (-\ln \gamma)^{\alpha}, \tag{9}$$

where γ is a uniform random variable within the range {0, 1}, and α is a parameter to control the size of the stochastic perturbation. α can be used as a dispersion factor in this simulation.

Finally, by adding eq. (9) to the model, eq. (8) is revised as

$$\begin{aligned} \Delta G_i^t &= RA \times \frac{q^t}{\pi\, l^2} \times \left(1 - \frac{d_{i\xi}^t}{d_{i\xi}^{\max}}\right)^k \\ &= (1 + (-\ln \gamma)^{\alpha}) \times \frac{q^t}{\pi\, l^2} \times \left(1 - \frac{d_{i\xi}^t}{d_{i\xi}^{\max}}\right)^k. \end{aligned} \tag{10}$$

At each iteration, the increase of 'grey value' will be calculated to determine urban growth. The cells will be converted into urban areas when their 'grey values' reach 1. Complex urban systems can be simulated by the iterations of CA simulation.

2 Model implementation and results

The model is applied to the simulation of urban development in Shenzhen and Dongguan in

the Pearl River Delta of southern China. The first step was to obtain and examine the spatial factors that play an important role in influencing urban development. Distance-based variables can be used to represent spatial influences. The amenities for urban development may be measured by the proximities to urban major centres, sub-centres, roads, expressways, railways, parks and rivers. Distance gradient functions can be used for the estimation of such influences[15]. There is a larger amount of benefits for a closer distance to these types of influences. However, a spectrum of environmental suitability could also be used as constraints for CA simulation to reduce development costs. Environmental suitability can be defined using distance decay functions according to various objectives, such as the protection of drinking water (reservoirs), cropland, orchard, vegetable land, fishpond, forest and wetland. A closer distance to these types of influences can bring about a larger amount of costs.

Remote sensing and GIS can be used to obtain spatial variables. The first set of six spatial variables was identified to address the benefits that can be obtained from closer distance to sources of development attraction. They are a) Distance to the major urban center (city proper); b) distance to town sub-centres (town centers); c) distance to railways; d) distance to expressways; e) distance to roads; f) distance to rivers.

A closer distance to these sources of attraction is more beneficial to urban development because energy and construction costs can be saved. These spatial variables (X_{ik}) can be defined using the negative exponential function.

$$X_j = e^{-\beta_j \mathrm{dist}_j}, \tag{11}$$

where X_j is the spatial variable for the positive criterion j, dist_j is the distance to the source of development attraction for criterion j, and β_j is its respective parameter of the distance decay function. The second set of variables includes these negative factors, a) distance to cropland; b) distance to orchard; c) distance to vegetable land; d) distance to fishpond; e) distance to reservoir (drinking water); f) distance to forest; g) distance to wetland.

A closer distance to these sources will create disturbances or negative effects for environmental and resource protection. These spatial variables can be defined using the following negative exponential function:

$$X_j = 1 - e^{-\beta_j \mathrm{dist}_j}. \tag{12}$$

These spatial variables are usually used as the site attributes for general GIS site selection and urban simulation. However, these spatial variables are usually correlated with each other. There are problems for using these spatial variables for MCE. It is difficult to provide weights when the number of spatial variables could be as many as several hundreds[16]. The PCA analysis should be incorporated in CA simulation to remove data redundancy.

Table 1 lists the principal components created from the thirteen layers of distance variables for Shenzhen and Dongguan. It is found that the first 5 components account for more than 90% of the variance of the original thirteen variables (93.9% for Shenzhen and 92% for Dongguan). Even

the first three components contain more than 80% of the total variance (88.8% for Shenzhen and 81.4% for Dongguan). Therefore, severe data redundancy exhibits in these spatial distance variables. PCA should be carried out to remove the data redundancy in the CA simulation which deals with a lot of spatial variables.

Table 1 Principal components and their variance

Principal components	Shenzhen		Dongguan	
	eigenvalues	percentage of variance (%)	eigenvalues	percentage of variance (%)
I	90.4	64.1	62.9	44.4
II	25.9	18.4	38.9	27.5
III	8.8	6.2	13.5	9.5
IV	3.7	2.6	8.5	6.0
V	3.6	2.5	6.5	4.6
VI	3.1	2.2	3.2	2.3
VII	1.8	1.3	2.6	1.9
VIII	1.2	0.9	19	1.4
IX	1.0	0.7	1.7	1.2
X	0.5	0.4	0.9	0.7
XI	0.5	0.4	0.5	0.3
XII	0.4	0.3	0.3	0.2
XIII	0.1	0.1	0.1	0.1

Table 2 is the component loadings for the thirteen spatial variables for Dongguan. It is easy to see that the first component is mainly related to agriculture and ecology, such as fishpond, vegetable land and wetland. The second component is mainly related to transport conditions, such as expressways, roads and rivers. The third component is mainly related to centers, such as city proper and town centres. There are a couple of advantages for the principal components transformation. The transformation can allow similar variables to group together with a large proportion of loadings in the same component. Suitable weights can be easily defined since principal components are independent of each other. This can avoid the repeated counting that may take place in general MCE.

The ‘ideal point’ is used to address economic, environmental and resource factors in CA simulation. These factors are represented by principal components to reduce data redundancy. The ‘ideal point’ is a virtual point having the maximum criteria scores for each criterion with regard to development suitability. It is the best point as the reference to urban development. The ‘ideal point’ for urban development is therefore (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1).

Only the first six principal components are used to calculate the similarity because the components contain 94.3% of the original information. According to the PCA transformation, the transformed ‘ideal point’ obtained by using the six principal components becomes (1.2, 2.6, 1.9, –0.2, –0.4, 0.1).

Table 2 Component loadings for the thirteen spatial variables

Distance variables	Principal components												
	I ecology and agri-culture	II transport	III urban centers	IV river	V xpress-ay	VI crops	VII	VIII	IX	X	XI	XII	XIII
City proper	–0.10	0.07	0.47	–0.50	0.02	0.04	–0.03	–0.07	0.06	0.07	–0.07	–0.17	–0.69
Town centres	–0.15	0.05	0.45	–0.52	–0.06	–0.05	0.01	–0.11	–0.04	–0.03	0.06	0.15	0.67
Railways	0.16	0.17	–0.07	–0.15	–0.72	0.15	0.27	0.53	–0.04	0.04	0.00	–0.13	0.01
Expressway	–0.26	0.62	–0.11	–0.09	0.51	0.03	0.09	0.50	–0.10	–0.05	0.03	0.05	0.01
Roads	–0.07	0.64	–0.34	–0.08	–0.29	–0.07	–0.10	–0.59	0.07	0.06	–0.02	–0.01	–0.02
Rivers	–0.43	0.21	0.54	0.63	–0.24	0.11	–0.05	0.00	0.07	–0.06	–0.02	0.01	0.00
Cropland	0.18	0.06	0.06	0.03	0.11	0.74	0.05	–0.22	–0.58	0.05	–0.01	–0.07	0.03
Orchard	0.23	0.10	0.15	0.11	0.18	0.00	0.85	–0.22	0.30	0.03	0.06	–0.01	0.02
Vegetable land	0.49	0.20	0.17	0.05	0.07	0.01	–0.18	0.06	0.16	–0.32	–0.71	0.07	0.08
Fishpond	0.48	0.19	0.17	0.05	0.03	0.10	–0.31	0.05	0.25	–0.25	0.68	0.03	–0.03
Reservoir	0.21	0.09	0.14	0.08	–0.10	–0.34	0.08	0.01	–0.45	0.14	0.06	0.71	–0.21
Forest	0.16	0.09	0.15	0.10	0.02	–0.52	0.06	–0.05	–0.49	–0.17	0.07	–0.62	0.06
Wetland	0.25	0.12	0.15	0.11	0.11	–0.06	–0.19	0.09	0.12	0.87	–0.06	–0.17	0.14

Table 3 Weights for various development objectives

Principal Compo-nents	Planning objectives				
	urban-center-based (city proper and town centers) development	transport-based (expressway, roads and rivers) development	cropland-conservation development	ecology and agriculture -conservation (vegeta-ble, fishpond, orchard, reservoir and wetland) development	economic environmental development
I Ecology and agri-culture	0.25	0.25	0.25	1.00	1.00
II Transport	0.25	1.00	0.25	0.25	1.00
III Urban centers	1.00	0.25	0.25	0.25	1.00
IV River	0.25	0.25	0.25	0.25	0.50
V Expressway	0.25	0.25	0.25	0.25	0.50
VI Crops	0.25	0.25	1.00	0.25	1.00
Weights: most important——1.00; very important——0.75; important——0.50; less important——0.25; not important——0					

Weights should be provided for different components according to their importance in simulation. There are different combinations of weights for various planning objectives. This can result in different simulation results. It is very difficult to provide weights when there are many variables in the simulation. However, the problem can be solved by using PCA because the number of variables can be much reduced.

The first six components were used to calculate the similarity. The factor loadings were examined according to table 1. This helps to define the weights reflecting various planning objectives. Weights are usually decided by expert's experience according to the importance of each factor. For example, if the planning objective is to protect agriculture and ecology, component I should be assigned with the highest value of 1. This study only uses five planning objectives to illustrate the methodology (table 3).

It is easy to make various development plans for different planning objectives. Fig. 2 is the simulation of transport-based development for Shenzhen in 1988—1997. A higher weight was used for the second component which has a large proportion of loadings for the variables of ex-

pressways, roads and rivers. Fig. 3(a) is the simulation of urban-center-based (city proper and town centers) development for Dongguan. The third component has a large proportion of loadings for the variables of city proper and town centers. There is a large amount of land development in the northwest part of the flood plain because it is close to the urban centers. Cropland conservation can be realized by putting a higher weight on the sixth component having a large proportion of loadings for the cropland variable. Cropland will be best protected if this alternative is realized (fig. 3(b)). The first component has a large proportion of loadings for the variables of vegetable, fishpond, orchard, reservoir and wetland. Greater concerns for ecological and agricultural protection can be built up by putting a higher weight for the first component (fig. 3(c)). There are severe conflicts between economic development and environmental conservation in most cases. A compromised objective will help to find an acceptable solution for both environmental conservation and economic development. The CA model is able to find suitable locations for reducing the conflicts as many as possible by balancing the weights for different components (fig. 3(d)).

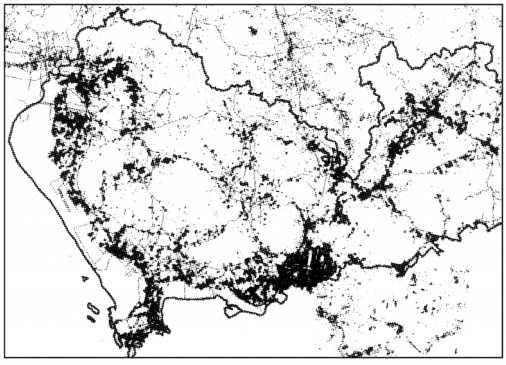

Fig. 2. The simulation of transport-based development for Shenzhen in 1988—1997.

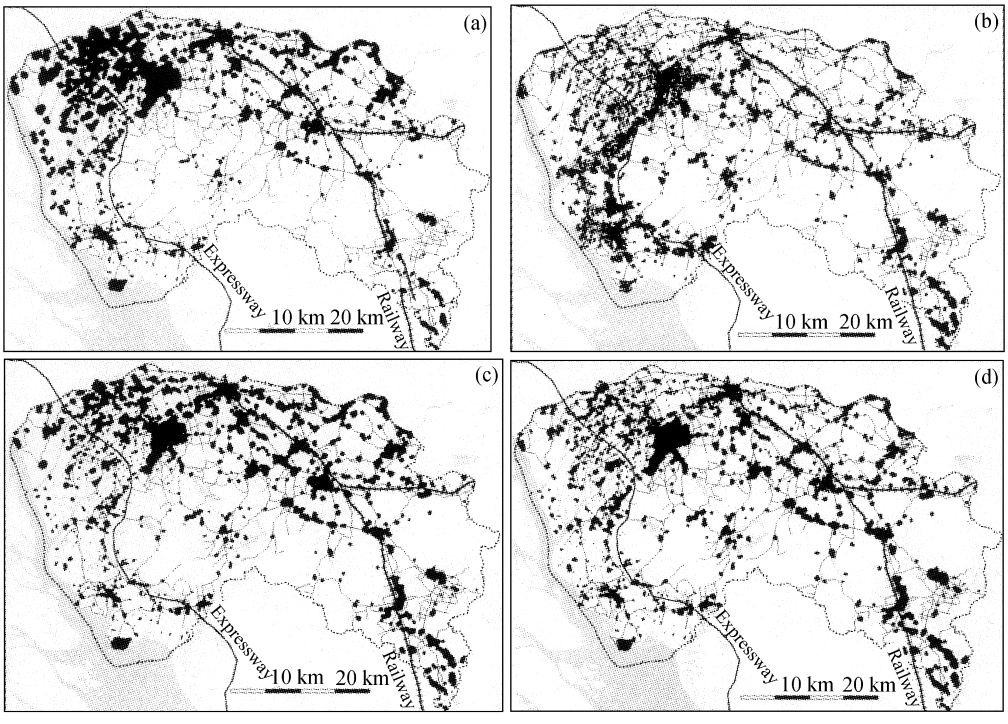

Fig. 3. CA simulation for urban development in Dongguan. (a) Urban-center-based; (b) cropland-conservation-based; (c) ecological and agricultural protection; (d) environmental conservation and economic development.

3 Conclusion

A large set of spatial variables is used in MCE during spatial decisionmaking. These spatial criteria can be retrieved from GIS. This study shows that there is high correlation between these criteria according to the principal components analysis. There are problems when MCE is used to deal with these correlated spatial variables. The correlation of spatial variables violates the principles of MCE because of repeatedly counting some variables. The study proposes the use of PCA and the 'ideal points' approach to deal with the common problems of spatial correlation. The PCA-CA model provides a useful planning tool for exploring various possible urban forms based on a large set of environmental constraints that could be considered in land use planning. It is easy to incorporate planning objectives in the urban simulation. Further studies are required to incorporate more factors, such as development density in the model for more realistic simulation.

Acknowledgements This project was supported by the National Natural Science Foundation of China (Grant No. 40071060) and the Croucher Foundation of Hong Kong (Grant No. 21009619).

References

1. Binder, P., Evidence of lagrangian tails in a lattice gas, in Cellular Automata and Modeling of Complex Physical Systems (eds. Manneville, P., Boccara, N., Vichniac, G. Y. et al.), Berlin: Springer-Verlag, 1989, 155—160.
2. Batty, M., Xie, Y., From cells to cities, Environment and Planning B: Planning and Design, 1994, 21: 531—548.
3. White, R., Engelen, G., Uijee, I., The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics, Environment and Planning B, 1997, 24: 323—343.
4. Wu, F., Webster, C. J., Simulation of land development through the integration of cellular automata and multicriteria evaluation, Environment and Planning B, 1998, 25: 103—126.
5. Li Xia, Yeh, G. O., Constrained cellular automata for modelling sustainable urban forms, Acta Geographica Sinica (in Chinese), 1999, 54(4): 289—298.
6. Li Xia, Yeh, G. O., Zoning for agricultural land protection using cellular automata, Chinese Environmental Science (in Chinese), 20(4): 318—322.
7. Zhou Chenghu, Sun Zhanli, Xie Yichun, Geo-cellular Automata (in Chinese), Beijing: Science Press, 1999, 1—163.
8. Li, X., Yeh, G. O., Modelling sustainable urban development by the integration of constrained cellular automata and GIS, International Journal of Geographical Information Science, 2000, 14(2): 131—152.
9. Carver, S. J., Integrating multi-criteria evaluation with geographical information systems, International Journal of Geographical Information Systems, 1991, 5(3): 321—339.
10. Nijkamp, P., van Delft, A., Multi-Criteria Analysis and Regional Decision-Making, The Netherlands: H.E. Stenfert Kroese B.V., 1977.
11. Malczewski, J., GIS and Multicriteria Decision Analysis, New York: John Wiley & Sons, Inc., 1999.
12. Gonzalez, R. C., Wintz, P., Digital Image Processing. Reading and Massachusetts: Addison-Wesley Publishing Company, 1977.
13. Batty, M., Cellular automata and urban form: A primer, Journal of the American Planning Association, 1997, 63(2): 266—274.
14. Yeh, G. O., Li, X., Sustainable land development model for rapid growth areas using GIS, International Journal of Geographical Information Science, 1998, 12(2): 169—189.
15. Batty, M., Xie, Y. C., Sun, Z. L., Modeling urban dynamics through GIS-based cellular automata, Computer, Environment and Urban Systems, 1999, 23: 205—233.
16. Bauer, V., Wegener, M., A Community information feedback system with multiattribute utilities, in Conflicting Objectives in Decisions (eds. Bell, D. E., Keeney, R. L., Raiffa, H.), West Sussex: John Wiley & Sons, Inc., 1977, 323—357.