PARTIAL LEAST SQUARES REGRESSION BASED CELLULAR AUTOMATA MODEL FOR SIMULATING COMPLEX URBAN SYSTEMS

Y.J. Feng [a, *], X.H. Tong [a], M.L. Liu [a]

[a] Department of Surveying and Geo-informatics, Tongji University, Shanghai 200092, China

Commission WG II/1

KEY WORDS: Partial Least Squares Regression, Cellular Automata, GIS, Urban Simulation

ABSTRACT:

A Cellular Automata model based on Partial Least Squares Approaches is proposed for simulating complex urban systems. The core part of Geo-CA model is the transition rules and a mass of independent spatial variables are involved in the process of creating CA model. Studies have focused on eliminating correlation using Multi-Criteria Evaluation (MCE) and Principal Component Analysis (PCA), but there is no a thoroughly solution with a reasonable result for the issue. Using Partial Least Squares Regression integrated with Geo-CA and GIS, a novel CA model is created for better urban expansion simulation. The model has been successfully applied to the simulation of urban development in Jiading District, Shanghai.

1. INTRODUCTION

Cellular Automata (CA), invented in the 1940's by the mathematicians John von Neuman and Stanislaw Ulam, while they were working at the Los Alamos National Laboratory in northern central New Mexico, is a discrete dynamic model defined on time and space. It consists of a regular grid of cells, each in one of a finite number of states, and used for simulating and analyzing various spatial phenomena (Zhou, C., Sun, Z. & Xie, Y., 1999). CA has been widely used in natural science, e.g. urban and geography (Batty, M. & Xie, Y., 1994, Batty, M., & Xie, Y., 1997, Li, X.& Yeh, A Gar-On, 2002), disaster (Lv C.S., Weng W.G. et al, 2007), traffic flow (Wang Y.M., Zhou L.H., Lv Y.B., 2008). In the field of geographical modelling (or spatial and urban modelling), CA arrested attention from international geographers and yet had many research outcomes (WU F, 2002, WHITE R & ENGELEN G, 1993, COUCLELIS H., 1997). Batty M. & Yichun Xie, White R., Couclelis H. introduced CA into urban (geography), and established the fundamental of CA modelling (Batty, M. & Xie, Y., 1994, Batty, M., & Xie, Y., 1997, WU F, 2002, WHITE R & ENGELEN G, 1993). Batty M. & Yichun Xie proposed GeoCA-Urban model (Zhou, C., Sun, Z. and Xie, Y., 1999, Batty, M. & Xie, Y., 1994, Batty, M., & Xie, Y., 1997), and Keith C. Clarke proposed SLEUTH model (Clarke, K. C., Gaydos, L. J. and Hoppen, S., 1997), WU Fulong proposed MCE-CA model (Wu, F.L., 2002), Ward et al. proposed Constrained CA model (D.P. Ward, A.T. Murray, S.R. Phinn a., 2000), Li Xia & Yeh A G O proposed Neural-network-based CA (Li, X., Yeh, A Gar-On, 2002). At the same time, some researchers at home have systemically studied the Geo-CA (LI Xia, YEH A G O, 2002, ZHANG X.F., CUI W.H., 2001, Feng Y.J., Tong X.H., Liu M.L., 2007).

The core part of the Geo-CA model is the transition rule (Zhou, C., Sun, Z. and Xie, Y., 1999) and how to consider various restraints (factors) and set its weight in Geo-CA model is the principal issue of its Algorithms. Complex factors affecting urban development are too many, and some times which include several hundred independent factors. And then how to eliminate the correlation of the spatial variables for setting better parameters of CA is a key problem. Wu have proposed logistic regression to determine parameter of CA model (Wu F., Webster C J, 1998). But if spatial variables are terribly independent, the model is not suited for geographical modelling. Wu also pointed out that MCE could be used for analyzing and solving conflict of spatial variables, but MCE requires the dependent of spatial variables (Carver S J, 1991, Nijkamp P van Delft, 1977). And others proposed PCA method for eliminating correlation of independent variables (Li Xia, YEH A G O, 2001), PCA Subset chosen to explain independent variables rather than dependent variables, and so, nothing guarantees that the principal components, which explain independent variables, are relevant for dependent variables.

In this paper, a Cellular Automata model based on Partial Least Squares Regression Approaches is proposed for simulating complex urban systems. Generalizing and combining features from principal component analysis and multiple regressions, PLS is particularly useful when we need to predict a set of dependent spatial variables from a large set of independent spatial variables. PLS regression finds components from independent variables that are also relevant for dependent variables. CA model integrated with PLS (PLS-CA) could markedly improve the accuracy of simulating results and avoid the disadvantages of MCE and PCA approaches.

2. THEORIES AND METHODS

It is very difficult to define the weight of various spatial factors (variables) when the variables are too many. Furthermore, the weights determined by current methods of Geo-CA models are not accurate when there are serious correlations between spatial factors (Li Xia, YEH A G O, 2001). The introduction of PLS could resolve the tough issue. The components found from independent variables (factors) can best explain dependent

* Corresponding author: fengyongjiu@126.com; phone +86-(0)21-65988851.

variable (probability of urban development). Replacing original variables by new variables (a set of dependent spatial variables) and used in CA simulation can avoid the irrationality of weight of MCE (Li Xia, YEH A G O, 2001), and can correct the poor explanation of subset variables to explained variables. By using PLS, a more wide range of spatial variables of urban can be adopted for improving precise of CA models.

Generally, CA defines the state of a cell at *t*+1 as a function of the state of the cell and its neighbourhood at time *t* in accordance with a set of transition rules (Wu F, 1996):

$$S_{ij}^{t+1} = f(S_{ij}^{t}, \Omega_{ij}^{t}, N) \tag{1}$$

Where S_{ij}^{t+1}, S_{ij}^{t} = sates of cell *ij* at time *t* and *t*+1 respectively
t = operation time
Ω_{ij} = the sate of neighbourhoods
N = the number of cell
f = transition function.

Suppose that study area consists of *n*×*m* cell, and then CA neighbourhoods can be given by:

$$\Omega^{t}{}_{ij} = \{s^{t}{}_{i-k,j-l}, ..., s^{t}{}_{i+k,j+l}\} \tag{2}$$

Where *k*, l=1, 2,…, <<n, m.

Generally, we adopt Moore neighborhood, so CA neighbourhoods can be further given by:

$$\Omega^{t}{}_{ij} = \frac{\sum_{3\times3} con(S_{ij}{=}urban)}{3\times3-1} \tag{3}$$

Thus, the global probability is:

$$p_c{}^{t} = p_g con(s_{ij}^{t} = suitable)\Omega_{ij}^{t} \tag{4}$$

Where p_c = combined probability
p_g = global probability of a cell
Con() = restrained function, its vale is 0 or 1.

In formula (4), p_g can be determined by transition probability and factors (spatial variables) of various years. Suppose *Y* is explained variable (transition probability, simple variable), thus the set of various variables is $X = [x_1, x_2, .., x_i, ., x_n]$.

Note F_0 is the standard variable of variable *y*, then:

$$F_{0i} = \frac{y_i - \bar{y}}{s_y}, i = 1,2,..,n \tag{5}$$

Where $\bar{y}$ = the expectation of *y*
s_y = the standard deviation of *y*.

And suppose E_0 is the standard deviation of independent variable set *X*.

According to the approaches and process of PLS, we can know that of *h*th component of PLS is:

$$w_h = \frac{E_{h-1}^{/} F_0}{||E_{h-1}^{/} F_0||} \tag{6}$$

$$t_h = E_{h-1}^{/} w_h \tag{7}$$

$$p_h = \frac{E_{h-1}^{/} t_h}{||t^h||^2} \tag{8}$$

$$E_h = E_{h-1} - t_h P_h^{/} \tag{9}$$

Here, we obtain components $t_1, \ldots, t_m$, so F_0 can be given by:

$$\hat{F}_0 = r_1 t_1 + ... + r_m t_m \tag{10}$$

And we have $t_h = E_{h-1} w_h = E_0 w_h^*$, thus $\hat{F}_0$ can be represented as expression of E_0:

$$\hat{F}_0 = r_1 E_0 w_1^* + ... + r_m E_0 w_m^* = E_0 [\sum_{h=1}^{m} r_h w_h^*] \tag{11}$$

Also the expression can be given by:

$$\hat{y} = \alpha_0 + \alpha_0 x_1 + ... + \alpha_p x_p \tag{12}$$

Therefore, in the PLS-CA model, p_g is given by:

$$p_g = \alpha_0 + \alpha_0 x_1 + ... + \alpha_p x_p \tag{13}$$

Where p_g = global probability
α_0, α_0,…, α_0 = weight of the spatial variables.

In order to get more actual simulation result, stochastic factors are introduced as a part of the PLS-CA model (Wu F L, 2002):

$$RA = 1 + (-\ln \gamma)^{\alpha} \tag{14}$$

Where γ = a random number in the range of (0, 1)
α = a parameter for controlling the effect of stochastic factors, range from 0~10.

Thus, the final development probability is given by:

$$P^{t} = P_{c}^{t} \times RA \tag{15}$$

An iteration formula for the PLS-CA model is therefore as follows:

$$S_{ij}^{\ t+1} = \begin{cases} Urban, \ p_{ij}^{t} > P_{\text{threshold}} \\ NonUrban, \ p_{ij}^{t} \le P_{\text{threshold}} \end{cases} \tag{16}$$

From formula (16) we know that if $p_{ij}^{\ t} > P_{\text{threshod}}$, the cell at *ij* transforms to urban, while if $p_{ij}^{\ t} < P_{\text{threshod}}$, the cell at *ij* keeps the current state.

Using ArcGIS Engine for processing spatial data, PLS-CA was developed in VS.NET environment. ArcGIS provides abundant functions for pre-processing vector and grid data, and we can easily use ArcGIS Engine to call various functions of ArcGIS to create a GIS-based CA model. The simulation process of PLS-CA for complex urban is as Figure.1.

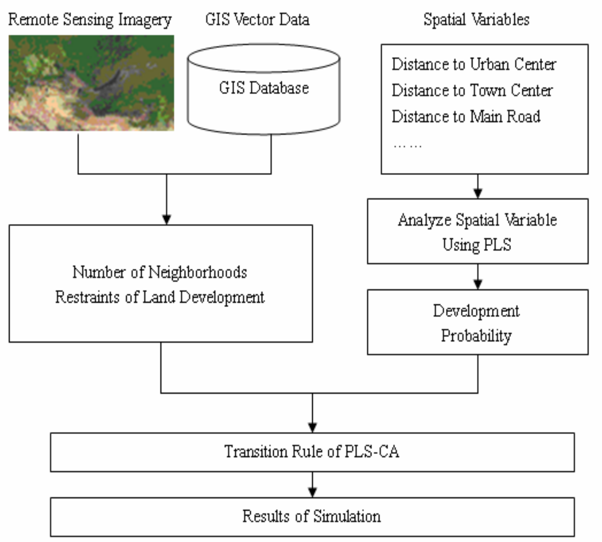


Figure.1 Partial Least Squares Approaches Based Geo-CA Model for the Simulation of Complex Urban System Changes

3. APPLICATION AND RESULTS ANALYSIS

Jiading District, a fast urbanization area in Shanghai, is selected as a case study area for testing the PLS-CA model. Jiading District, adjoining to Kunshan, Jiangsu Province, is located at Northwest (far-suburb) of Shanghai, Yangtze River Delta. In the past ten years, Jiading District have entered into a fast developing period with drastic land use/ land cover changes, immoderately expansion of urban, population explosion, great changes of economy and employment. To research urban system of Jiading District using PLS-CA can beneficial to understand mechanism, rules and development characteristics of urban in China like Jiading District.

There are too many factors impacting urban development (especially urban land use change), including institutional factors, policy factors, population factors, economy factors, and various geographical factors. Because of the serious correlation of geographical factors (spatial variables), a suited method should be adopted to eliminate correlation and obtain more reasonable simulation result. Using PLS to gain principal components of spatial variables is a very important approach for improving accuracy of CA models.

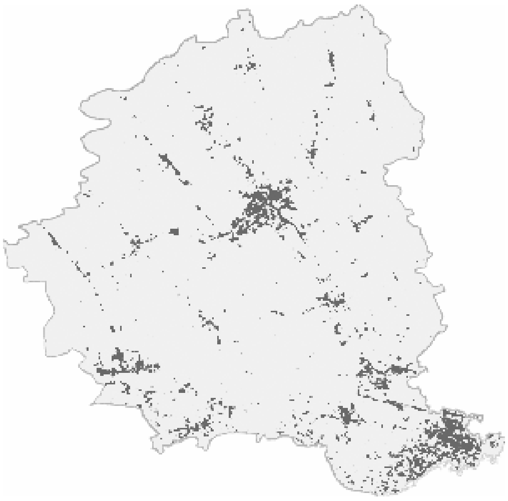
In order to simulate the urban development, we should firstly prepare spatial data for mining rules of PLS-CA, including remote sensing imagery and GIS data. TM image of two years 1989 and 1995 are prepared to obtain global probability using the cell changes from 1989 to 1995, and gain the spatial variables for rules using imagery of 1989. There are two methods to acquire spatial variables, one is to extract road, water, plowland and urban land from remote sensing imagery, the other is to read data of road, water, plowland and urban land from GIS vector database. The first method is adopted in this paper, and the spatial variables obtained from remote sensing imagery are as Table 1.

Positive Factors	Distance to Urban Center Distance to District Center Distance to Town Center Distance to Main Road Distance to Main River of Development	Using Eucdistance function of Arc/INFO GRID
Negative Factors	Distance to Water Distance to Plowland Distance to Kaleyard Distance to Urban Forest Distance to Everglade	Calculate number of cells In (7×7) neighborhood

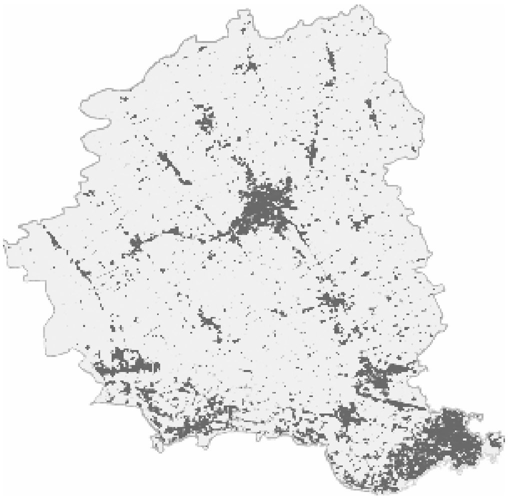
Table 1. Spatial variables in the partial least square based CA model

Generally, topography gradient data should be extracted and act as an important part of the spatial variables in Geo-CA model. When the gradient is too large it is difficult to develop land for urban under such environment, therefore, it is necessary embed this factor in the model. However, in this case we test PLS-CA model in Jiading District, Shanghai, a very level area of Yangtze River Delta, thus we can overlook the gradient factor in this case.

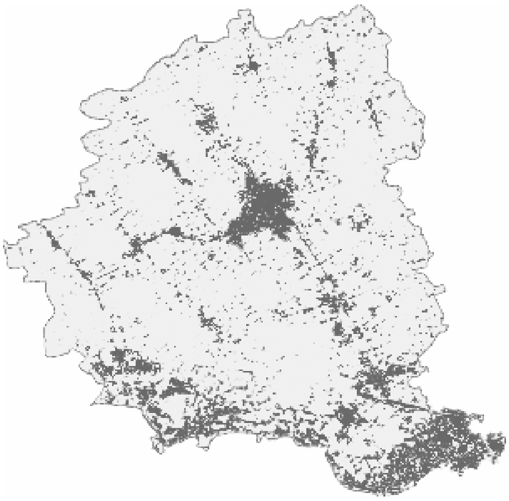
Before the simulation we should randomly create sample coordinate of 20% and obtain spatial variables through these points. And then using PLS approaches, we can extract principal components having best explanation for explained variable. Six principal components are listed by order: distance to main road, distance to district center, distance to town center, distance to main river of development, distance to plowland and distance to Kaleyard. The other four factors are having not obvious impact on urban development through PLS regression analysis. Figure.2 are the simulating results of urban development (expansion) in Jiading District using PLS-CA in 1998, 2002, 2008 respectively.



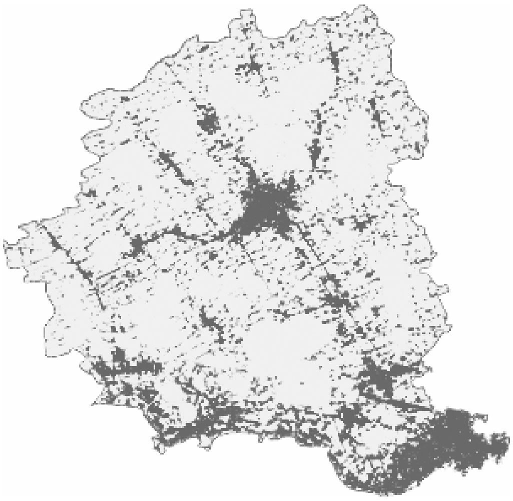
a. 1989（Original）



b. 1998（Simulation）



c. 2002（Simulation）



d. 2008（Simulation）

Figure.2 Simulating Complex Urban System Using Partial Least Squares Approaches Based Geo-CA

Figure.2 show that the urban of Jiading District mainly expands along the main roads, and close to the areas of center of Jiading District (Jiading Town) and Shanghai City (Putuo District). The results greatly accord with actual urban development of Jiading District. In the 1990’S the Chinese government launched a great program to develop Pudong, a coastal area of Shanghai, and drove the economy of Shanghai to a higher stage. Economy development effectively pushed urban expansion, therefore, the built-area spread at a striking speed and reached the far-suburban area (e.g. Jiading District, Baoshan).

In 1989 the urban morphology of Jiading was comparatively compact and mainly expanded on the basis of the primary built-area of Jiading (Jiading New City, Jiading Industry Zone, Jiaqiao Town and Zhenxin Street). However, with the construction of Shanghai International Automobile City, there were several developing zones appeared round Huangdu Town and Anting Town. In recent years, urban expansion in Jiading District sprawled to hinterland and appeared many scattered built-area.

It is necessary to evaluate the efficiency of a simulating model by its simulation precise. To test the model precise, we respectively measured the precise of MCE-CA and PLS-CA by each point (see Table.2). We can see from Fig.2 that the precise of PLS-CA are higher than that of MCE-CA.

Geo-CA	Year: 2002	Simulating Non-urban	Simulating Urban	Precise
MCE-CA	Actual Non-urban	456770	172390	72.60%
	Actual Urban	85205	246331	74.30%
PLS-CA	Actual Non-urban	491374	137786	78.10%
	Actual Urban	80232	251304	75.80%

Table 2. The result of Simulation based on MCE-CA and PLS-CA

4. CONCLUSIONS

Urban is an open complex giant system whose development is impacted by various spatial factors (variables) with serious correlation. Integrated with the analysis functions of GIS, we can use MCE and PCA to evaluate the weight of these factors. However, MCE can not eliminate the correlation of spatial variables. While PCA can eliminate the correlation to a certain degree but the principal components found in the independent variables set can not completely explain the dependent variables. Comparing with MCE and PCA, PLS can extract components with better explanation for dependent variables. This paper proposed a Geo-CA model based on PLS for simulating complex urban system.

Using TM imagery and GIS database, this model was applied in Jiading District, Shanghai to simulate urban development in 1998, 2002, and 2008.

Analysis shows that the simulation results well accord with actual urban development of Jiading District. The simulation result of PLS was compared with that of MCE-CA, and we know that the simulating precise of PLS-CA higher than that of MCE-CA, and have more actual spatial pattern of urban development.

Although PLS-CA have unparalleled advantage in eliminating correlation of spatial variables, and can obtain excellent Geo-CA transition rules, but it requires the assist of other PLS analysis software. In order to improve the intelligence of CA rules mining, we should integrate PLS with intelligentizing algorithms in further researches of Geo-CA.

REFERENCES

Zhou, C., Sun, Z. and Xie, Y., 1999. Studies on Geo-Cellular Automata (Geographical Information Science Series). Science Press of China, Beijing, pp. 52-74.

Batty, M., Xie, Y., 1994. From cells to cities. Environment and Planning B, 21, pp.531-548.

Batty, M., Xie, Y., 1997. Possible urban automata. Environment and Planning B, 24, pp.175-192.

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

Lv C.S., Weng W.G. et al, 2007. Fire evacuation model based on motor schema and cellular automaton. Tsinghua University (Natural Science), 2007(12), pp.2163-2167.

Wang Y.M., Zhou L.H., Lv Y.B., 2008. Vehicle Equivalent Conversion of Cellular Automat Model Based on Traffic Flow. China Journal of Highway and Transport, 01, pp.114-117.

WU F, 2002. Calibration of Stochastic Cellular Automata: the Application to Rural-urban Land Conversion. International Journal of Geographical Information Science, 16(8), pp.795-8l8．

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, pp.1175-1199．

COUCLELIS H., 1997. From Cellular Automata to Urban Models: New Principles for Model Development and Implementation. Environment and Planning B, 24. pp.165-174.

Clarke, K. C., Gaydos, L. J. and Hoppen, S., 1997. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. Environment and Planning B, 24, pp.247- 61.

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

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

LI Xia, YEH A G O, 2002. Neural-network based Cellular Automata for Simulating Multiple Land Use Changes Using GIS. International Journal of Geographical Information Science, 16(4), pp.323-343．

ZHANG X.F., CUI W.H., 2001. Integrating GIS with Cellular Automata to Establish a New Approach for Spatiotemporal Process Simulation and Prediction. Acta Geodaetica et Cartographica Sinica, 30(2), pp.148-155．

Feng Y.J., Tong X.H., Liu M.L., 2007. Extended Cellular Automata Based Model for Simulating Multi-Scale Urban Growth Using GIS. In: Proceedings of the 2007 International Conference on Intelligent Systems and Knowledge Engineering, Chengdu, China, Vol.1, pp.1-7.

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, pp.103-126.

Carver S J, 1991. Integrating multi-criteria evaluation with geographical information systems. International Journal of Geographical information Systems, 50, pp.321-339.

Nijkamp P van Delft, 1977. A Multi-criteria analysis and regional decision-making. H E Stenfert Kroese BV, Netherlands.

Li Xia, YEH A G O, 2001. Application of PCA and Cellular Automata in Spatial Decision-making and urban simulation. China Science: Series D, 31(8), pp.683-690.

Wu F. A linguistic cellular automata simulation approach for sustainable land development in a fast growing region [J]. Computers, Environment and Urban, Vol. 20, No. 6, pp. 367-387, 1996.

Wu F L, 2002. Calibration of stochastic cellular automata: the application to rural-urban land conversions. Int. J Geographical Information Science, 16 (8): 795-818.

Wang H.W, 1999. Partial Least Squares Regression – Method and Applications. Defense Industry Press of China, Beijing, pp.186-216.

ACKNOWLEDGEMENTS

The work described in this paper was substantially supported by the National Natural Science Foundation of China (Project No. 40771174 and No. 40301043), the Program for New Century Excellent Talents in Universities (Project No. NCET-06-0381), the Foundation of Shanghai ShuGuang Scholarship Program (Project No. 07SG24), and the grants from the Doctoral Program of Higher Education of China (Project No. 20070247046).