Urban Growth Pattern Modeling Using Logistic Regression

NONG Yu1, DU Qingyun1, 2
1. School of Resource and Environmental Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
2. Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China

© Wuhan University and Springer-Verlag Berlin Heidelberg 2011

Abstract  Transformation of land use/land cover change occurs due to the numbers and activities of people. Urban growth modeling has attracted substantial attention because it helps to comprehend the mechanisms of land use change and thus helps relevant policies made. This paper tends to apply logistic regression to model urban growth in the Jiayu county of Hubei province, China. It is applied in a GIS environment to calculate variables and, then, in SPSS to discover the relationships between urban growth and the driving forces. The relative operating characteristic (ROC) shows the modeling accuracy with the curve 0.891 with standard error 0.001. A probability map is generated finally to predict where urban growth will occur as a result of the computation. The result shows the model simulates urban growth well in the county scale.

Keywords  logistic regression; urban growth; modeling

CLC number  P208

Introduction

A profound transformation of the Earth’s environment that is currently underway is primarily due to the numbers and activities of people. During the last 50 years, the human population has risen from two and a half to over six billion, and economic activity has increased tenfold. Human transformation of the ecosystems and landscapes are the largest source of change in the natural systems on Earth, affecting the ability of the biosphere to sustain life. Intensification and diversification of land use have led to rapid changes in landscape dynamics. Land use and cover change study is core part of global land plan sponsored by IGBP and IHDP.

Urban growth and land use change is dominated by human activities with complex spatio-temporal dynamics, which is a great interest to planners, conservationists, ecologists, economists, and resource managers.[1-7] Over the past decades, research in these disciplines has sought to develop models of these processes for forecasting future development, evaluating future plans, and identifying endangered natural areas. Spatial modeling are popular in the last two decades due to increased computing power, improved availability of spatial data, and the need for innovative planning tools to help decision making.[8-10] There are mainly two groups according to the key mechanisms to simulate the process of land use change:[11,12] rule-based/process-based models and empirical-statistic models.[13, 14] Rule-/ process-based models imitate processes and often address the interaction of components forming a system, of which
there are cellular automata (CA) with the great capability to handle temporal dynamics\cite{15-18} and artificial intelligence model.\cite{19-21} CA focuses on microspatial pattern,\cite{22} but it is difficult to reflect macrochanges affected by social and economic factors. In contrast, empirical-statistical models locate land-cover changes by applying multivariate regression techniques to relate historical land use changes to spatial characteristics and other potential drivers. Statistic and multivariable urban growth models reflect complexity of land use change by applying a broad range of social and economic factors and also imply potential changes.

Different methods of statistical modeling enable decision makers or land managers to answer different questions. One approach is to apply the predictive statistical models used widely to predict the distribution of species or habitat types, such as regression techniques. Logistic regression, one of empirical-statistical methods, is an example of discrete outcome modeling techniques. Such models are appropriate where transitions starting from one cover type have different possible end states. Binary logistic regression can be used to predict a dependent variable on the basis of continuous and categorical independent variables and to determine the percent of variance in the dependent variable explained by the independent variables; to rank the relative importance of the independent variables; to assess interaction effects; and to understand the impact of covariate control variables. It applies maximum likelihood estimation after transforming the dependent variable into a logit variable. The impact of predictor variables is usually explained in terms of odds ratios. The unique value of dependent variable and the ability of odds prediction estimate the probability of a certain event occurring. It has been applied in the research of biological species distribution,\cite{23,24} landslide hazard prediction,\cite{25} and also urban growth study.\cite{12,26,27} Most logistic regression of urban growth are based on a single scale, and the raster resolution is determined by the spatial detail in data.\cite{28}

This study tends to model urban expansion based on vector features in a GIS environment and to discover the relationship between urban growth and the driving forces of which biophysical and social-economic factors are selected as independent variables.

1 Material and methods

1.1 Study area

We study the urban growth in a small scale, taking Jiayu County as case. This county, which has eight towns, is located in Hubei Province in the middle China (Fig.1). With rather flatness of elevation and rich irrigation from Yangzi River, the county has a long history of agriculture and a key farmland conservation area of China. In this agriculture-domain region, most people are mainly engaged in agriculture, so agriculture factors need to be included. Over these years, industry begins to develop, and many companies set up their own factories there, which have fueled the urban sprawl.

1.2 Variables design

The map of the whole county was dispersed into regular cells representing the actual 100 m×100 m tessellation each. All points will get the attributions from demographic, economical, and geographical
driving forces through spatial analysis. There are a total of 102301 cells, and after excluding the cells with land use attribute of water, 75711 cells remain for the next step.

The choice of economic and biophysical variables conforms to most dynamic simulation modeling practices, which usually consider the determining factors of SLEUTH (slope, land use, exclusion, urban extent, transportation, hill shade) as in Clarke’s SLEUTH model. In this study, independents selected referred to SLEUTH. Population density ($X_1$) is often established as land use determinants to indicate labor availability, accessibility, or presence of local markets. Gross industrial output value ($X_3$) is the total volume of final industrial products produced and industrial services provided during a given period. It reflects the total achievements and overall scale of industrial production during a given period. Gross agricultural output value ($X_5$) statistics agriculture, forestry, animal husbandry, and fishery. The population and economic data all came from Jiayu County Annuals. The variables reflecting the spatial influences of major roads and economic centers were from the 2003 topographic map. $X_4$ and $X_5$ were not just simply distance measured, and they were standardized as comprehensive index, ranging from 0 to 100. The closer to the roads or economic centers, the larger the index number. The slope ($X_6$) variable was created from DEM data with five classifications. According to the investigation, if the slope is above $6^\circ$, the farm land intensity makes discount.

The response variable represented was prepared using land use map digitalized from aerial photographs and field inventory and aggregating them into two classes: urban land and nonurban land. The time span from between the two assessments was 5 years due to the limitation of land use map resource from local authority. Samples from land use map of 2003 were used to calculate the correlation coefficients of $Y$ and $X_i$, and then, the whole area was simulated. The tabular statistic data of 2007 was used to verify the simulation of quantity, while the 2007 map certified the location where urban growth occurs. The 2003 census data were used for the social variables in model calibration. All the six variables chosen are listed in Table 1.

Correlations may exist between those variables. Logistic regression calibration should check for collinearity. Table 2 shows the correlation coefficients of independents. The collinearity of the predictor variables was checked by using a hierarchical cluster analysis with a Pearson correlation coefficient. As shown in Table 2, the collinearity is not strong. After multicollinearity test, the spatial analysis will be done in ArcView GIS 3.2 with Avenue Script.

| Variable | Broad definition | Unit | Nature of variable |
|----------|------------------|------|-------------------|
| $Y$      | 0. no urban growth; 1. urban growth |      | Dichotomous       |
| $X_1$    | Population density | Population (km$^2$) | Continuous |
| $X_2$    | Gross industrial output value | Billion Yuan (town) | Continuous |
| $X_3$    | Gross agricultural output value | Billion Yuan (town) | Continuous |
| $X_4$    | Index of distance to economic center | - | Continuous |
| $X_5$    | Index of distance to the major road | - | Continuous |
| $X_6$    | Slope ($5$ categories) | Degree | Design |

1.3 Logistic regression

A logistic regression model was used to associate the urban growth with demographic and econometric driving forces and generate an urban growth probability map. This regression model is useful for situations that prediction of the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. The nature of urban growth of a cell is dichotomous, either the presence of urban

| $X_1$   | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---------|-------|-------|-------|-------|-------|
| $X_1$   | 1     | -0.283| 0.003 | 0.360 | -0.069| -0.033|
| $X_2$   | -0.283| 1     | 0.116 | -0.115| -0.100| -0.018|
| $X_3$   | 0.003 | 0.116 | 1     | 0.182 | -0.031| 0.277 |
| $X_4$   | 0.360 | -0.115| 0.182 | 1     | 0.171 | 0.094 |
| $X_5$   | -0.069| -0.100| -0.031| 0.171 | 1     | -0.057|
| $X_6$   | -0.033| -0.018| 0.277 | 0.094 | -0.057| 1     |

Table 2 Correlation coefficients of independents
growth or no urban growth. It is assumed that the probability of a cell changing to urban use follows the logistic curve as described by the logistic function. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model. It is applicable to a broader range of research situations than discriminant analysis. For the dependent represents urban growth results, \( Y \) has a binary value of 1 and 0 for Yes and No, respectively. Actually, the probability reflects in what extend \( Y \) will change into 1, as shown in Eq. (1).

\[
P(Y = 1 | X_1, X_2, \ldots, X_n) = \frac{1}{1 + e^{\sum \beta_i X_i}} \tag{1}
\]

Where \( P(Y = 1 | X_1, X_2, \ldots, X_n) \) is the probability of \( Y \) given by \( X_i (i = 1, 2, \ldots, n) \) and changes from non urban land to urban land. Moreover, \( 1 - P \) is the probability of presence of no urban growth. Through Logit transformation, as shown in Eq. (2), we have

\[
\ln \left( \frac{P}{1 - P} \right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{2}
\]

After the maximum likelihood estimate, the coefficients of independents, which could interpret how driving forces affect urban sprawl, could be computed by iteration. All statistic procedure will be performed in SPSS 11.5 for Windows.

2 Results and discussion

2.1 Statistic validation

The goodness of fit classification shows the predicted value given by all the independents in the 2003 land use map, of which the cut value is 0.500 in SPSS. There were 2437 points of actual value 1 classified into value 0, with an accuracy of 60.5% (see Table 3). It seems that the urban growth occurs more than what was expected.

| Observed \( Y \) | Predicted \( Y \) | Percentage correct |
|-----------------|-----------------|-------------------|
| 0               | 68347           | 1197              | 98.3 |
| 1               | 2437            | 3730              | 60.5 |
| Overall Percentage |                 |                    | 95.2 |

The variables’ coefficients and Wald value shows in Table 4 that \( X_1, X_2, X_3, X_4, X_5, \) and \( X_6 \) have a significant relationship with \( Y \), the dependent variable. In these six variables, only slope has negative coefficients, due to the designed classification with larger numbers representing abrupt slope. It turned out that population density, index of distance to economic distance, index of distance to roads, and slope have a strong relationship with urban growth.

| Coefficient | Standard error | Wald | Df | Sig.(Pr > chi-square.) |
|-------------|----------------|------|----|------------------------|
| \( X_1 \)   | 0.023          | 0.002| 116.928 | 0.000         |
| \( X_2 \)   | 0.002          | 0.000| 97.834  | 0.000         |
| \( X_3 \)   | 0.001          | 0.000| 1.718   | 0.019         |
| \( X_4 \)   | 0.040          | 0.003| 173.902 | 0.000         |
| \( X_5 \)   | 0.022          | 0.001| 1074.271| 0.000         |
| \( X_6 \)   | -0.051         | 0.018| 7.978   | 0.005         |
| Constant    | -1.873         | 0.051| 1329.625| 0.000         |

2.2 Map validation

Relative operating characteristic (ROC) is a plot of the probability of having true positive identified urban growth versus the probability of the false positive identified as the cut-off probability varies. In this study, it was used to testify the logistic regression compared to actual 2007 land use map to measure the simulated and real change. The area under the ROC curve is particularly important for evaluating how good the decision making is at discriminating between stable versus unstable areas. An idea model would have an area of 1.000, and this logistic model has the largest area under the curve 0.891 with standard error 0.001 (Fig. 2). A plot of ROC curve for the logistic model is shown in Fig. 2. ROC analysis evaluates the sensitivity and specificity of decision procedures. Each point on the ROC curve is associated with a specific decision criterion. The ROC assessed
that the pair of maps agrees the location of points being urbanized well.

2.3 Model interpretation

Logistic regression modeling was used to identify and improve our understanding of the demographic, economic, and biophysical forces that have driven the urban growth and to find the most probable sites of urban growth in Jiayu County in a small scale. This test result shows that urban development tends to occur in the area of higher road accessibility, economic centers neighborhoods, and higher population density areas, which is the first step of urbanization happening in most developing countries. Like most Chinese cities, urban sprawl and suburbanization are in low-density urban development. The predicted spatial patterns of future urban areas are the compromised outcomes of the above driving forces.

In this case, ROC area goes to 0.891. The main reason is that intervals between simulated year and origin year are five years only. Besides, the national basic farmland protection in Jiayu goes well, restricting urbanization expanding into good-quality cropland. The growth map was generated, as shown in Fig.3. The urban growth likely happens near major roads and economic centers and also areas with larger density.

In this case, ROC area goes to 0.891. The main reason is that intervals between simulated year and origin year are five years only. Besides, the national basic farmland protection in Jiayu goes well, restricting urbanization expanding into good-quality cropland. The growth map was generated, as shown in Fig.3. The urban growth likely happens near major roads and economic centers and also areas with larger density.

Fig. 3 The actual pattern and the predicted pattern in 2007 with legend showing the probability

3 Conclusion

Logistic regression does not focus too much on the complexity of the system, making it easier to understand the inner driving forces that change landscape. The probability predicted can successfully capture points that urban growth occurs with less demand of computation resource. Vector-feature-based spatial analysis has higher accuracy than raster-based calculation, for raster cells always have to be calibrated at a single resolution.

Logistic regression does not rely on distributional assumptions in the same sense that discriminant analysis does. However, the solution may be more stable if predictors have a multivariate normal distribution. In addition, as with other forms of regression, multicollinearity among the predictors, although it is not strong, can lead to biased estimates and inflated standard errors. The procedure is most effective when group membership is a truly categorical variable. Social and economic factors are chosen for calculation after correlation matrix analysis; still, multicollinearity cannot be eliminated completely. Finally, the spatial autocorrelation in geographical phenomenon also has influence in the final result carried out in logistic regression. In the further study, those shown above are what need solutions.

References

[1] Henríquez C, Azócar G, Romero H (2006) Monitoring and modeling the urban growth of two mid-sized Chilean cities[J]. Habitat International, 30(4): 945-964

[2] Vliet Jv, White R, Dragicevic S (2009) Modeling urban growth using a variable grid cellular automaton[J]. Computers, Environment and Urban Systems, 33(1): 35-43

[3] Luo J, Wei Yhd (2009) Modeling spatial variations of urban growth patterns in Chinese cities: the case of Nanjing[J]. Landscape and Urban Planning, 91(2): 51-64

[4] Herold M, Goldstein N C, Clarke K C (2003) The spatiotemporal form of urban growth: measurement, analysis and modeling[J]. Remote Sensing of Environment, 86(3): 286-302

[5] Haack B N, Rafter A (2006) Urban growth analysis and
modeling in the Kathmandu Valley, Nepal[J]. Habitat International, 30(4): 1056-1065

[6] Eppink F V, van den Bergh JCJM, Rietveld P (2004) Modelling biodiversity and land use: urban growth, agriculture and nature in a wetland area[J]. Ecological Economics, 51(3-4): 201-216

[7] Barredo J I, Demicheli L (2003) Urban sustainability in developing countries’ megacities: modelling and predicting future urban growth in Lagos[J]. Cities, 20(5): 297-310

[8] Konijnendijk C C, Thorsen B J, Tyrväinen L, et al.(2007) Decision-support for land-use planning through valuation of urban forest benefits[J]. Allgemeine Forst Und Jagdzeitung, 178(4): 74-84

[9] Matthews K B, Sibbald A R, Craw S (1999) Implementation of a spatial decision support system for rural land use planning: integrating geographic information system and environmental models with search and optimisation algorithms[J]. Computers and Electronics in Agriculture, 23(1): 9-26

[10] Castella J-C, Pheng Kam S, Dinh Quang D, et al. (2007) Combining top-down and bottom-up modelling approaches of land use/cover change to support public policies: Application to sustainable management of natural resources in northern Vietnam[J]. Land Use Policy, 24(3): 531-545

[11] Heistermann M, Muller C, Ronneberger K (2006) Land in sight?: Achievements, deficits and potentials of global to local scale land-use modeling[J]. Agriculture, Ecosystems & Environment, 114(2-4): 141-158

[12] Hu Z, Lo C P (2007) Modeling urban growth in Atlanta using logistic regression[J]. Computers, Environment and Urban Systems, 31(6): 667-688

[13] Aspinall R (2004) Modeling land use change with generalized linear models—a multi-model analysis of change between 1860 and 2000 in Gallatin Valley, Montana[J]. Journal of Environmental Management, 72(1-2): 91-103

[14] Hietel E, Waldhardt R, Otte A (2007) Statistical modeling of land-cover changes based on key socio-economic indicators[J]. Ecological Economics, 62(3-4): 496-507

[15] Fang S, Gertner G Z, Sun Z, et al. (2005) The impact of interactions in spatial simulation of the dynamics of urban sprawl[J]. Landscape and Urban Planning, 73(4): 294-306

[16] White R, Engelen G (2000) High-resolution integrated modelling of the spatial dynamics of urban and regional systems[J]. Computers, Environment and Urban Systems, 24(5): 383-400

[17] Yeh AG-O, Li X (2001) A constrained CA model for the simulation and planning of sustainable urban forms by using GIS[J]. Environment and Planning B, 28: 733-753

[18] Yassemi S, Dragicevic S, Schmidt M(2008) Design and implementation of an integrated GIS-based cellular automata model to characterize forest fire behaviour[J]. Ecological Modelling, 210(1-2): 71-84

[19] Ligtgenberg A, Bregt A K, van Lammeren R (2001) Multiactor-based land use modelling: spatial planning using agents[J]. Landscape and Urban Planning, 56(1-2): 21-33

[20] Bolliger J(2005) Simulating complex landscapes with a generic model: sensitivity to qualitative and quantitative classifications[J]. Ecological Complexity, 2(2): 131-149

[21] Veldkamp A, Verburg P H (2004) Modelling land use change and environmental impact[J]. Journal of Environmental Management, 72(1-2): 1-3

[22] Liu X, Andersson C (2003) Assessing the impact of temporal dynamics on land-use change modeling[J]. Computers, Environment and Urban Systems, 28(1-2): 107-124

[23] Aspinall R J (2002) Use of logistic regression for validation of maps of the spatial distribution of vegetation species derived from high spatial resolution hyperspectral remotely sensed data[J]. Ecological Modelling, 157(2-3): 301-312

[24] Yang X, Skidmore A K, Melick D R, et al. (2006) Mapping non-wood forest product (matsutake mushrooms) using logistic regression and a GIS expert system[J]. Ecological Modelling, 198(1-2): 208-218

[25] Gorsevski PV, Gessler P, Foltz R B(2000) Spatial prediction of landslide hazard using logistic regression and GIS[C]. Proceedings of the 4th International Conference on Integrating GIS and Environmental Modelling (GIS/EM4), Banff, Alberta, Canada

[26] Rutherford G N, Guisan A, Zimmermann N E (2007) Evaluating sampling strategies and logistic regression methods for modelling complex land cover changes[J]. Journal of Applied Ecology, 44(2): 414-424

[27] Cheng J, Masser I (2003) Urban growth pattern modeling: a case study of Wuhan city, P. R. China[J]. Landscape and Urban Planning, 62(4): 199-217

[28] Verburg P H, de Koning G H J, Kok K, et al. (1999) A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use[J]. Ecological Modelling, 116(1): 45-61

[29] Dietzel C, Clarke K (2006) The effect of disaggregating land use categories in cellular automata during model calibration and forecasting[J]. Computers, Environment and Urban Systems, 30(1): 78-10