Environ Model Assess (2007) 12:55–61
DOI 10.1007/s10666-006-9052-8

Modeling the spatial pattern of farmland using GIS and multiple logistic regression: a case study of Maotiao River Basin, Guizhou Province, China

Qiu-Hao Huang · Yun-Long Cai · Jian Peng

Received: 14 June 2005 / Accepted: 21 April 2006 / Published online: 22 June 2006

Abstract Land use change is an important topic in the field of global environmental change and sustainable development. Land use change modeling has attracted substantial attention because it helps researchers understand the mechanisms of land use change and assists regulatory bodies in formulating relevant policies. Maotiao River Basin is located in the province of Guizhou, China, which has a developed agricultural industry in the karst mountain areas. This paper selected biophysical and social–economic factors as independent variables, and constructed a multiple logistic regression of farmland spatial distribution probability by random sampling. Then, by using GIS technology and integrating the 2000 data, this study predicted the farmland spatial pattern. When the predicted map was compared with the actual farmland map for 2000, we noted that 71% of the simulation is in accordance with the 2000 farmland pattern. The result satisfactorily proves the reasonability and applicability of our model.

Keywords land use pattern · farmland spatial pattern model · GIS · multiple logistic regression · Maotiao River Basin · China

Q.-H. Huang · Y.-L. Cai (✉) · J. Peng
Department of Resources, Environment and Geography,
Center for Land Study, Peking University,
Laboratory for Earth Surface Processes,
Ministry of Education,
Beijing, China, 100871
e-mail: caiyl@urban.pku.edu.cn

1. Introduction

Land use change is an important research subject in global environmental change and sustainable development [11]. Modeling land use change has attracted considerable attention because it could explain the mechanisms and causes of land use change, and help governments formulate relevant policies [4, 24].

Currently, there are two kinds of land use change models: non-spatial land use change models and spatial land use change models. The non-spatial land use change models include Markov chains model [21] and dynamic (process-based) simulation models [20]. A difficulty in applying these models to land-use change studies lies in their inability to deal with spatial variability in the processes of land use change. Spatial land use change models, such as CA [12, 27] and CLUE models [23, 25], can successfully predict the quantity of change with the location of land use change over long periods. But the problem lies in that these models are more complicated to construct.

In fact, multiple logistic regression (MLR) is also a good tool to model land use change. It has already been successfully used in wildlife habitat studies [1, 16, 17], prediction of forest fires [22], and deforestation analyses [13, 15].

Combining GIS technology and MLR, this study focused on the Maotiao River Basin as an example; we selected the 1995 biophysical and social– economic factors affecting the farmland spatial pattern, sampled 2400 points (1200 in farmland, 1200 out of farmland) randomly, and constructed a logistic model of the farmland spatial pattern. Using the model, we simulated the farmland spatial pattern by the biophysical and social–economic factors for the year 2000.

This study was conducted (1) to identify the main factors affecting the farmland spatial pattern in the Karst Mountain Ecosystem and (2) to offer a new method for land use pattern simulation.

2. Study area

The study area, Maotiao River Basin, covers about 2,883 km^2 and is located in the middle of Guizhou province, southwestern China (figure 1). The province has a warm–moist climate with an annual temperature ranging from 14 to 16°C, and elevation ranging between 700 and 3800 m above sea level. Annual precipitation is about 1300 mm, most of which falls during the summer months of June and September [2].

The valley, which has a long history of cultivation, has a moderate population density (average of 170 inhabitants per km^2); the main industry is agriculture.

We opted to study area this because it has the best biophysical conditions for cultivation in Guizhou province and is one of the most important agricultural areas in the province. Modeling the farmland spatial pattern may help the local government to formulate relevant agricultural policies and develop local agriculture.

3. Materials

3.1 Basic data

Land use maps were obtained from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science (IGSNRR, CAS). In the 1995 farmland maps, areas within the farmland were coded 1, while areas out of the farmland were 0. The binary response variable data layer (i.e., presence/absence) could then be used to investigate the relationship between response probability and the explanation variable. The slope and aspect maps were derived from the DEM. The road network and villages were manually digitized from the local topographic map at a scale of 1:50,000. Population density maps and GDP per capita maps were derived from the China population census of 1995 and 2000 [7].

All data were brought together in a raster GIS and resampled to a common spatial resolution of 30 m. The GIS software used in this study was ARCGIS 8.2 and the statistical software was SAS 8.0.

3.2 Dependent variable

Farmland maps were extracted from the land use maps. The 1995 farmland map was used to calculate the coefficients of the model, whereas the 2000 farmland map was used to verify the simulation.

3.3 Independent variable

Elevation. In Guizhou Plateau, elevation is the main factor that affects the farmland spatial pattern [2]. Usually, temperature drops by 0.5 – 0.6°C and rainfall increases by 92 mm for every altitude increase of 100 m.

Slope. Slope is another important factor influencing farmlands [6]. In Guizhou Plateau, the average slope is 21.15°; areas with a slope ≥18° cover about 64% of the total area of Guizhou province, and farmlands with a slope ≥15° accounts for about 50% and those with a slope ≥25° represents 20%.

Aspect. Because rainfall and sunlight are correlated with aspect, while rainfall and sunlight are necessary for crops, then aspect may also have some association with farmland spatial pattern [28].

Population density. Two population layers were used in the models: population density at sub-location level for 1995 and 2000.

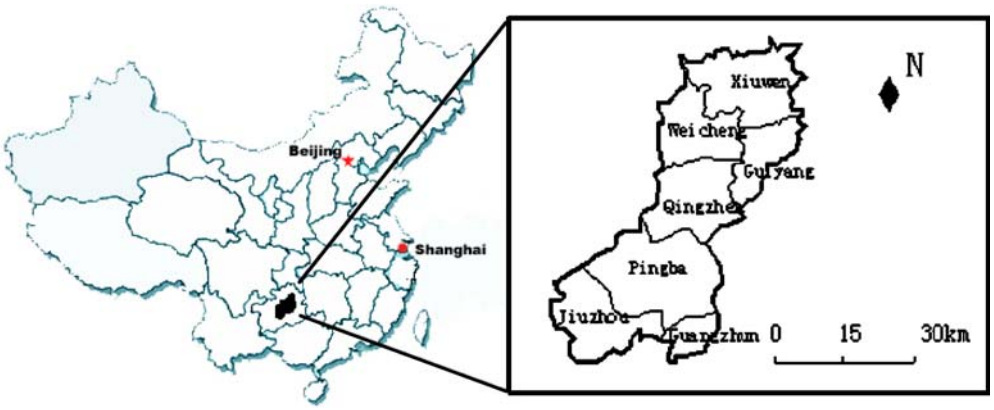

Figure 1 Location of Maotiao River Basin, China

Table 1 Description of variables.

Variables	Type	Unit
Dependent variables		
Farmland	Binary	0–1
Independent variables		
Elevation	Continuous	m
Slope	Continuous	°
Aspect	Continuous	°
Population density, 1995	Continuous	per km^2
Population density, 2000	Continuous	per km^2
GDP per capita, 1995	Continuous	per yuan
GDP per capita, 2000	Continuous	per yuan
Distance to road	Continuous	km
Distance to villages	Continuous	km

GDP per capita. This variable illustrates the regional economic conditions. In the province of Guizhou, the higher GDP per capita, the more developed is the agriculture and the larger the farmland area [3].

Distance to road, distance to towns. These two variables were used to reveal the relationship between farmland distribution and accessibility.

All variables are listed in table 1.

4. Methodology

4.1 Multiple logistic regression models

The technique used in this study is MLR [9]. It is designed to estimate the parameters of a multivariate explanatory model in situations where the dependent variable is dichotomous, and the independent variables are continuous or categorical. The MLR technique yields coefficients for each independent variable based on a sample of data. These coefficients are interpreted as weights in an algorithm that generates a map depicting the probability of a specific category of land use change for all sampling units.

MLR identifies the role and intensity of explanatory variables X_n in the prediction of the probability of one state of the dependent variable, which is defined as a categorical variable P (0 or 1). Suppose $\boldsymbol{X}$ is a vector of explanatory variables and p is the response probability to be modeled with, in the case of a dichotomous dependent variable, $P = Pr\ (x/1 + x)$, with $P = 0$ (absence of farmland) and $P = 1$ (presence of farmland). The logistic model has the form

$$P = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)} \quad (1)$$

where α is the intercept and β_n are slope parameters.

4.2 Sampling procedure

Before MLR was performed, a stratified random sampling procedure was used to select N observation points distributed throughout the study area. Random sampling of observations was used due to the presence of spatial auto-correlation in the data. For every sample observation, the values of dependent variables and the set of independent variables were recorded. A random sample of 2,400 observations was selected, with an equal number of 0 and 1 observations of the dependent variable. Unequal sampling rates do not affect the estimation of the coefficients of the explanatory variables in logistic models [14], but it does affect the constant term. When using the model to run simulations, the constant term is corrected by adding ln p_1–ln p_2, where p_1 and p_2 are the proportions of observations chosen from the two groups for which the dependent variable is 1 and 0, respectively [14]. Because in MLR, multicollinearity may affect the coefficients greatly, the independent variables in the samples were tested [18]. The linearity of the bivariate

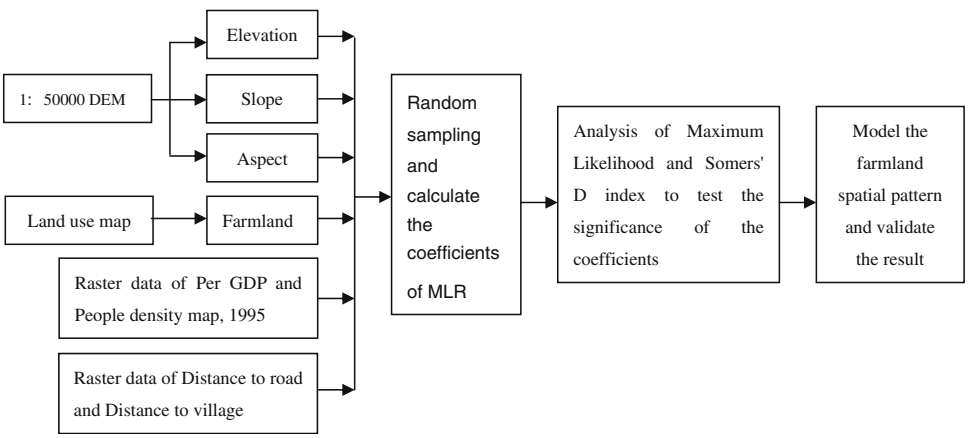

Figure 2 Flow chart of farmland spatial pattern model

Table 2 Analysis of maximum likelihood estimates.

Variable	Individual model $Pr > \chi^2$	Stepwise model $Pr > \chi^2$	Full model $Pr > \chi^2$
Elevation	<0.0001	<0.0001	<0.0001
Slope	<0.0001	<0.0001	<0.0001
Aspect	0.9588	–	0.5190
Population density, 1995	<0.0001	<0.0001	<0.0001
GDP per capita, 1995	<0.0001	<0.0001	0.0005
Distance to road	0.0052	–	0.8461
Distance to village	0.0829	0.0004	0.0014

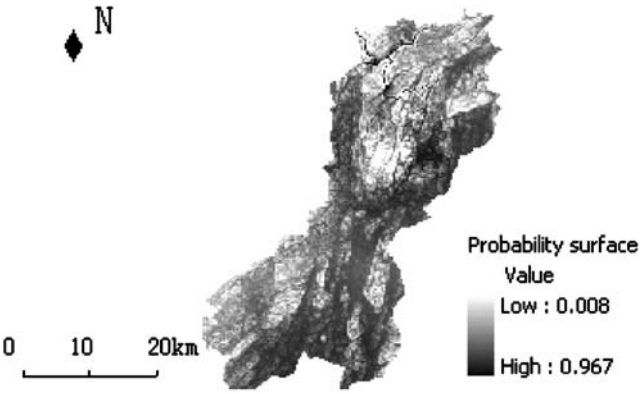

Figure 3 Probability map of farmland in Maotiao River Basin

relationship between each of the independent variables and the dependent variable was also tested for avoiding interaction [5, 10]. Figure 2 summarizes the overall analysis procedure followed in this study.

5. Results and discussion

5.1 Predicted model

After applying the logistic analysis, we obtain the model of farmland spatial pattern

$$p = \frac{\exp[10.488 - 0.0063\,(\text{elevation}) - 0.1355\,(\text{slope}) - 0.0002\,(\text{aspect}) + 0.000364\,(\text{GDP per capita, 1995}) - \frac{0.0106\,(\text{Population density,1995}) - 0.0029\,(\text{distance to road}) + 0.01\,(\text{distance to village})]}{0.0106\,(\text{Population density,1995}) - 0.0029\,(\text{distance to road}) + 0.01\,(\text{distance to village})]}}{1 + \exp[10.488 - 0.0063\,(\text{elevation}) - 0.1355\,(\text{slope}) - 0.0002\,(\text{aspect}) + 0.000364\,(\text{GDP per capita, 1995}) - \frac{0.0106\,(\text{Population density,1995}) - 0.0029\,(\text{distance to road}) + 0.01\,(\text{distance to village})]}{0.0106\,(\text{Population density,1995}) - 0.0029\,(\text{distance to road}) + 0.01\,(\text{distance to village})]}} \quad (2)$$

The analysis of maximum-likelihood estimates ($Pr >$ chi-square) shows that the predictor variables elevation, slope, population density (1995), GDP per capita (1995) are significant in predicting the probability of farmland spatial pattern in the three kinds of models: Individual Model, Stepwise Model, Full Model. However, the variables aspect, distance to road, and distance to village are not significant (table 2).

The SAS LOGISTIC procedure also tests the association of predicted probabilities and the observed responses using a series of rank correlation indices [19]. This statistical tool assesses the predictive ability of a model by using Somers' *D*, Goodman–Kruskal Gamma, Kendall's tau-a and Kendall's tau-c. All correlation indices for the full model reflected a high degree of association between predicted and observed responses. The Somers' *D* index is given in table 3 to illustrate the high degree of correlation between the predicted and observed responses. The statistical analysis of the model, therefore, reflects its application to geographic information dataset, which would produce an accurate probability surface of the farmland spatial pattern.

In table 3, the correlation indices of *aspect*, *distance to road*, and *distance to village* are quite low, indicating that these variables have little influence on farmland spatial pattern. In contrast, the correlation indices of GDP per capita show certain association with farmland spatial pattern.

Table 3 Association of predicted probabilities and observed responses.

	Elevation	Slope	Aspect	Population density, 1995	GDP per capita, 1995	Distance to road	Distance to villages	Stepwise	Full
Somers' *D*	0.6	0.52	−0.015	0.43	0.47	0.045	0.073	0.75	0.74

Table 4 Simulation accuracy of the logistic model for the validation data set (km^2).

True categories	Predicted categories		Accuracy (%)
	Farmland	Non-farmland	
Farmland	795	328	71
Non-farmland	470	1,288	73

Based on what has been discussed above, it is reasonable to cancel aspect, distance to road, and distance to village from the equation. The reason why distance to road and distance to village were not significant in the equation may be because, in the Guizhou Plateau, bio-physical and social–economic factors affect the farmland spatial pattern considerably, whereas accessibility exerts little influence.

Next, we recalculated the coefficient of the four variables: elevation, slope, population density (1995) and GDP per capita (1995); the logistic regression equation is

$$P = \frac{\exp[10.3346 - 0.00616\,(\text{elevation}) - 0.1353\,(\text{slope}) - 0.0102\,(\text{Pop density}) + 0.000356\,(\text{GDP per capita})]}{1 + \exp[10.3346 - 0.00616\,(\text{elevation}) - 0.1353\,(\text{slope}) - 0.0102\,(\text{Pop density}) + 0.000356\,(\text{GDP per capita})]} \quad (3)$$

According to the model, the location's probability for farmland increased with GDP per capita. The interpretation agreed with the fact that in this region the main industry is agriculture. Elevation, slope, and population density have negative impact on farmland spatial pattern.

5.2 Predicted probability surface

The coefficients derived for the model in equation (3) were applied to the GIS database of farmland spatial pattern to produce a probability surface of the farmland, with value ranging from 0 to 1 (figure 3).

5.3 Validation

Then we overlaid the 1995 farmland spatial map on the percent probability surface. Nearly 85% of the farmland was in regions of >85% probability. This value can therefore be used as the cut-off probability to model the farmland spatial pattern for 2000.

Using the model, we predicted the farmland map for the year 2000 by integrating the 2000 annual data. Next, we validated our results. A cell-by-cell comparison of the accuracy of the model was performed to evaluate the predicted versus actual farmland spatial pattern. The result shows that the model correctly predicted 71% of the farmland spatial distribution and

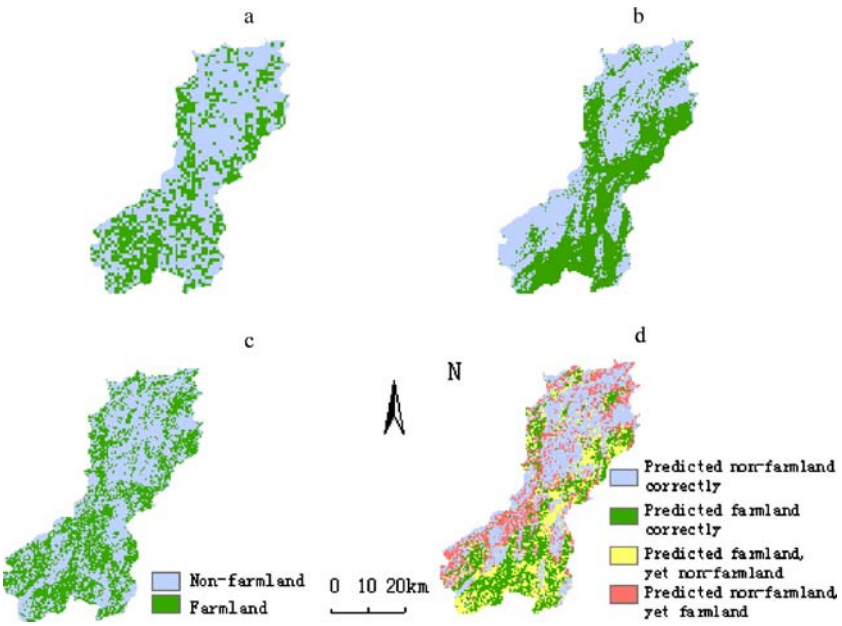

Figure 4 Comparison of simulated with the actual farmland spatial pattern of Maotiao River Basin. [(a) 1995 farmland spatial pattern; (b) predicted 2000 farmland spatial pattern; (c) 2000 farmland spatial pattern; (d) the comparison of simulated and actual farmland spatial pattern]

73% of the non-farmland spatial distribution (table 4 and figure 4).

We consider the result quite satisfactory for two reasons. Firstly, because the complexity of land use change makes it difficult currently to use models to predict the land spatial pattern with an accuracy of more than 85% [12, 23, 25, 27]. For example, Li and Yeh [12], who integrated cellular automata (CA) and neural networks to simulation of land development, obtained an accuracy of about 83%, whereas Verburg et al. [25], who used the CLUES (Conversion of Land Use and its Effects) model to simulate land use change scenarios in Ecuador, managed to obtain an accuracy range of 71–90%. The reason could be because land use change is driven by both biophysical and social–economic factors, and few models could synthetically incorporate these driven factors to make a precise prediction.

Secondly, Maotiao River Basin is located in the middle of Guizhou province, which is located in the Karst mountain regions of southwest China. In this region land degradation is most serious, the land surface is uneven and the topographic condition is complicated [8, 26]. All these increase the uncertainty of prediction of land spatial pattern. Even the Chinese Academy of Science (CAS) uses the TM/ETM image to interpret the land spatial pattern; after manual adjustment, the accuracy is just about 75%.

6. Conclusion

Using GIS and MLR, this study constructed a model to predict the 2000 farmland spatial pattern based on 2000 spatial data. When the cut-point is 85%, the accuracy of the predicted farmland is 71%, and the predicted non-farmland's is about 73%. The result is quite satisfactory.

Based on the results, the factor affecting the farmland spatial pattern most was slope, followed by population density, elevation and GDP per capita. Two variables, distance to roads and distance to villages, do not show significance in the analysis of maximum likelihood estimates and Somers' *D* index. It proves that accessibility has little impact on farmland spatial pattern. These findings may help the government to utilize the province's agricultural resources effectively.

In comparison with other spatial models, our model (combining MLR with GIS) is much easier to handle than CA and CLUE. The model also requires few parameters. There are also several factors that result in an incorrect prediction. First, it is important to recall that farmland spatial pattern data from which the model was developed correspond to a single year, and may be insufficient for long-term prediction. Second, we cannot model some government policies. Yet these policies are quite important for farmland use change and they have a substantial effect on farmland spatial pattern.

Acknowledgements This research was supported by The Key Project (40335046) of National Natural Science Foundation of China and The Research Fund for the Doctoral Program of Higher Education (20040001038). The authors are also grateful to the anonymous reviewers for their insightful comments and helpful suggestions. However, any errors or shortcomings in the paper are the responsibility of the authors.

References

1. Bian, L., & West, E. (1997). GIS modeling of elk calving habitat in a prairie environment with statistics. *Photogrammetric Engineering & Remote Sensing*, *63*, 161–167.
2. Cai, Y. (1990). *The territorial structure and resources development in Guizhou Province*. Beijing: Ocean Press.
3. Cai, Y. (1997). *Geography and environment– system analytical methods*. Beijing: The Commercial Press.
4. Cai, Y. (2001). A study on land use/cover change: The need for a new integrated approach. *Geographical Research, 20*, 645–652.
5. Christensen, R. (1996). *Analysis of variance, design and regression applied statistical methods*. New York: Chapman & Hall.
6. He, T., & Xie, D. (2002). Status of soil and water loss of slope-dry-cultivated land and countermeasures of realignment in Guizhou mountainous region. *Journal of Guizhou University(Agricultural and Biological Science)*, *21*, 280–286.
7. http://www.stats.gov.cn.
8. Huang, Q., & Cai, Y. (2006). Assessment of karst rocky desertification using the radial basis function network model and GIS technique: A case study of Guizhou Province, China. *Environmental Geology*, *49*, 1173–1179.
9. Kleinbaum, D. G. (1994). *Logistic regression: A self-learning text*. New York: Springer.
10. Kleinbaum, D. G., & Pacific, G. (1998). *Applied regression analysis and other multivariable methods*. Duxbury Press.
11. Lambin, E. F. (1999). Land-use and land-cover change: Implementation strategy, IGBP Report No. 48/IHDP Report No. 10, IGBP Stockholm.
12. Li, X., & Yeh, A. G. (2002). Neural-network-based cellular automata for simulating multiple landuse changes using GIS. *International Journal of Geographical Information Science*, *16*, 323–343.
13. Ludeke, A. K., Maggio, R. C., & Reid, L. M. (1990). An analysis of anthropogenic deforestation using logistic regression and GIS. *Journal of Environmental Management*, *31*, 247–259.
14. Maddala, G. S. (1988). *Introduction to econometrics*. New York: Macmillan.
15. Mertens, B., & Lambin, E. F. (2000). A spatial model of land-cover change trajectories in a frontier region in southern Cameroon. *Annals of Association of American Geographer*, *90*, 467–494.
16. Narumalani, S., Jensen, J. R., Althausen, J. D., Burkhalter, S., & Mackey, Jr. (1997). Aquatic macrophyte modeling

using GIS and multiple logistic regression. *Photogrammetric Engineering & Remote Sensing*, *63*, 41–49.
17. Pereira, J. M. C., & Itami, R. M. (1991). GIS-based habitat modeling using logistic multiple regression: A study of the Mountain Graham red squirrel. *Photogrammetric Engineering & Remote Sensing*, *57*, 1475–1486.
18. Rawlings, J. O. (1998). *Applied regression analysis: A research tool*. New York: Springer.
19. SAS Institute. (1997). *SAS/STAT software: Changes and enhancements through release 6.12*. Cary, NC: SAS Institute, Inc.
20. Thornton, P. K., & Jones, P. G. (1998). A conceptual approach to dynamic agricultural land-use modeling. *Agricultural System*, *57*, 505–521.
21. Turner, M. G. (1987). Land use changes and net primary production in the Georgia landscape. *Environment Management*, *11*, 237–247.
22. Vega, G., Woodard, P. M., Titus, S. J., Adamowicz, W. L., & Lee, B. S. (1995). A logit model for predicting the daily occurrence of human caused forest fires. *International Journal of Wild Fire*, *5*, 101–111.
23. Veldkamp, A., & Fresco, L. O. (1996). CLUE-CR: An integrated multi-scale model to simulate land use change scenarios in Costa Rica. *Ecological Modeling*, *91*, 231–248.
24. Veldkamp, A., & Lambin, E. F. (2001). Predicting land-use change. *Agriculture, Ecosystems & Environment*, *85*, 1–3.
25. Verburg, P. H., Veldkamp, A., de, K., Kok, K., & Bouma, J. (1999). A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. *Ecological Modeling*, *116*, 45–61.
26. Wang, S., & Liu, Q. (2004). Karst rocky desertification in southwestern China: Geomorphology, land use, impact and rehabilitation. *Land Degradation & Development*, *15*, 115–121.
27. Wu, F., & Webster, C. J. (1998). Simulation of land development through the integration of cellular automata and multicriteria evaluation. *Environmental & Planning B*, *25*, 103–126.
28. Xiong, K. (2002). *The study of Karst rocky desertification using the GIS & RS tech, a case study of Guizhou Province*. Beijing: Geology Press.