Exploring Drivers of Urban Expansion

Anna Shchiptsova, Richard Hewitt and Elena Rovenskaya

Abstract Spatial patterns in urban land development are linked with the level and type of economic activity. Here, we develop a statistical model to explore the relationship between the spatially explicit population density and the type of land use in a region. The relationship between the type of land use (urban/non-urban) and the level of economic activity is modeled at the scale of a single cell on the geographical map. Thus, the statistical model should be tested against large samples of data points on the high-resolution maps. The challenge here is that the original socio-economic data is given at a coarser resolution than the land use (200 × 200 m cells) We present results of our spatial modeling exercise for the case study of the Seville Province, Spain.

Keywords Land use model · Urban sprawl · Multiple regression

1 Introduction

In recent decades urban systems have undergone rapid development. We have seen a transition in the population distribution from the population mostly dispersed in rural areas to a highly urbanized society, where people are concentrated in cities. Today more than 50 % of people worldwide live in a city [1] and this figure is likely to grow more in the future.

A. Shchiptsova (✉) · E. Rovenskaya
International Institute for Applied Systems Analysis, Schlossplatz 1,
2361 Laxenburg, Austria
e-mail: shchipts@iiasa.ac.at

R. Hewitt
Observatorio para una Cultura del Territorio, C/ Duque de Fernan Nunez 2,
1a planta, 28012 Madrid, Spain

E. Rovenskaya
Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow
State University, 2nd Educational Building, Leninskie Gory, 119991 Moscow, Russia

T.B. Nguyen et al. (eds.), *Advanced Computational Methods for Knowledge Engineering*, Advances in Intelligent Systems and Computing 453, DOI 10.1007/978-3-319-38884-7_12

At the regional scale we observe territory expansion of the urbanized centers. However, this process typically unfolds non-uniformly with respect to the city border. Much of this development has occurred as dispersed, low density growth outside of the major centers but within their area of economic influence. Such type of urban development is typically referred to as urban sprawl [2]. While sprawling developments are not necessarily in themselves always undesirable, they bring a range of issues such as increased energy consumption through encouragement of the use of private vehicles, causing traffic congestion and air pollution, and irreversible damage to ecosystems, caused by scattered and fragmented urban development in open lands [3].

A large body of research is dedicated to the analysis and prediction of urban expansion. Studies on land use change are based on different modeling principles including such techniques as cellular automata [4, 5], Markov chains [6] and logistic regression [7, 8]. This study focuses on the dependence between spatial patterns in land use and population distribution (without the temporal dimension). By using spatial data, we investigate whether part of the variance of the population density is explained by the land use type of the corresponding cell and the types of its immediate neighbors and if, in this way, we can capture spatial interactions.

Geographic data frequently shows spatial dependence, i.e., values at close distances are more similar than expected for independent observations. This property limits the use of the multiple linear regression model for spatial data analysis. An alternative is to incorporate a spatial lag into the model specification (e.g., spatial lag model or spatial error model). A comprehensive introduction to the econometric spatial modeling can be found in [9, 10]. But estimation of the spatial models is not easily computable. This research focuses on application of certain filtering techniques to spatial data in order to meet assumptions of standard linear regression and use conventional statistical methods to test and interpret results of spatial analysis.

We perform a case study on the Spanish Province of Seville. The choice of this region is motivated by the fact that Spain is one of Europe's urban sprawl hotspots, with problems of urban sprawl being particularly acute in the area of economic influence of major cities like Madrid and Valencia and along the Mediterranean coast.

2 Study Area

The Province of Seville is located in the Mediterranean region of Andalusia in the southwestern part of Spain. Its territory is 14,000 km^2. The terrain in this region is made up almost exclusively of river basin. The Guadalquivir river crosses the province from east to west. In the north territory includes parts of the Sierra Morena mountain range and to the very south the foothills of the Cordillera Subbetica mountain range.

The population of the region is close to two million inhabitants (2010). The province is subdivided into 105 municipalities. The large part of the population lives in the capital city Seville. The Seville municipality has the population of about

700,000 people (2011, INE), which is much larger than any other municipality; for example, the second largest municipality, Dos Hermanas, has the population of about 130,000 inhabitants (2014).

2.1 Land Use Maps

Seville has experienced notable urban development in recent years. In this region, as well as overall in Spain, urban expansion has been especially acute since the restoration of democracy in 1978, joining the EU in 1986 and skyrocketing per capita incomes in the second half of 1980s and the decade before the 2008 crisis; after the 2008 crisis, the speed of development has slowed down.

Figure 1 shows the urban/non-urban land distribution in this region from the year 1956 when Spain was an autocratic country to the post-crisis year 2013. The GIS data represents the territory of the region as a regular grid of cells. We classify all the cells into two major categories: urban land and non-urban land (vegetation, wetlands, agricultural land and water). This figure illustrates the urbanized centers territory expansion unfolding over the last 60 years. Table 1 provides some basic statistics illustrating the spread of urbanization.

2.2 Economic Data

This study focuses on land use distribution in Seville region for the year 2003. As candidate drivers for the land use change in Seville, various socio-economic factors have been identified from the papers analyzing case studies of land conversion in the New Castle County, the USA [8], Wuhan City, China [7], Ecuador [11], San Francisco Bay and Sacramento areas [12] (CUF model), and San Francisco Bay area [4] (SLEUTH model). We have included those drivers of land conversion from these case studies, which are relevant for the Seville province. These potential land use drivers include socio-economic factors defined on the GIS-based maps (with 200 × 200 m cells) and those obtained from the census data. Where the data for the year 2003 was not available, the closest year, for which the data was available, has been chosen.

The GIS-based maps in the collected dataset include data on transportation networks (i.e., the distance to the nearest road (2005) and the distance to the nearest airport (2006)), biophysical factors (i.e., the distance to the nearest waterfront (2005) and the distance to the nearest area of forest (2006)), data on physical proximity to different infrastructure objects (i.e., the distance to the nearest area of commercial or industrial land use (2006), the proximity to a city center with more than 10,000 inhabitants (2011) and the proximity to a city center with more than 50,000 inhabitants (2011)).

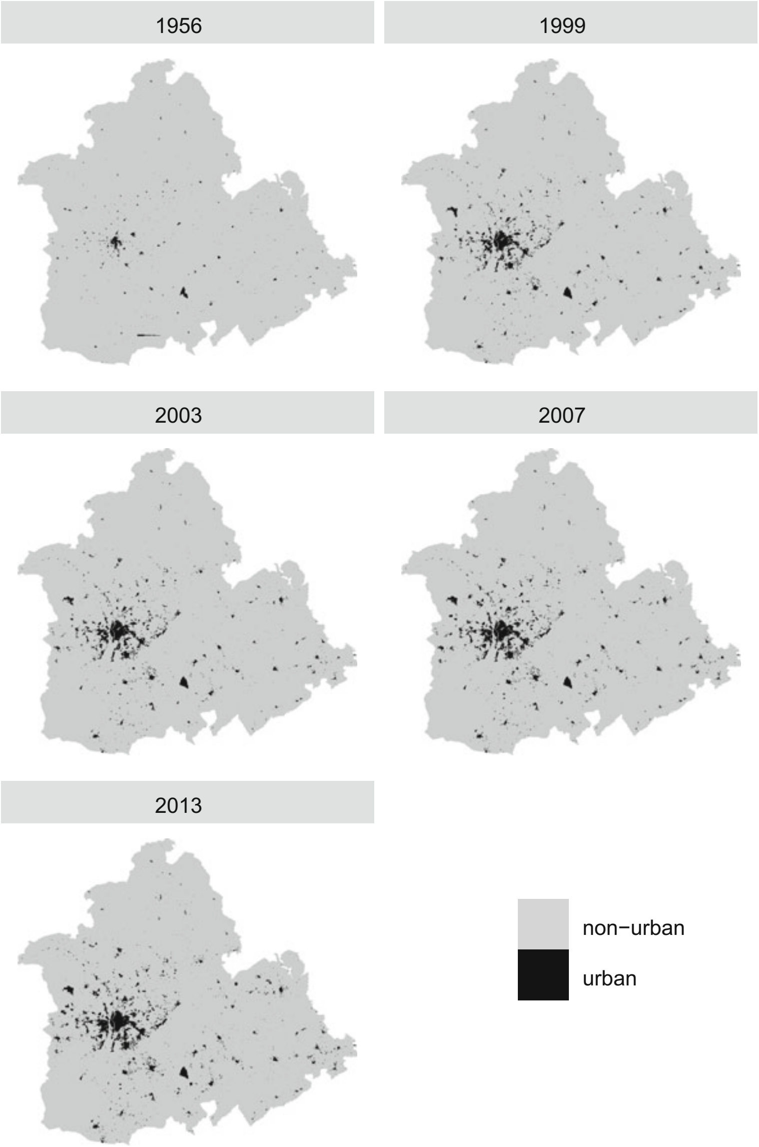

Fig. 1 Historical transformation of land use in the Province of Seville, Spain

Table 1 Spread of urbanization in the Province of Seville estimated from the GIS data

Year	Number of urban cells	% of urban cells in the map (%)	Relative growth (%)
1956	2461	0.71	–
1999	9218	2.63	274.56
2003	10,105	2.88	9.62
2007	12,289	3.50	21.61
2013	13,674	3.89	11.27

The census data on potential factors includes several types of records characterizing the social and economic activity in the region. Namely, the group of general economic factors includes data on the income per capita (2003, Euros, Source: SIMA), people employed (2001, number of people, Source: SIMA), and economically active population (2001, number of people, Source: SIMA). Population factors are composed from the percentage of population younger than 20 years old, the percentage of population between 20 and 64 years old, and the percentage of population older than 65 years (2001, number of people, Source: Instituto de Estadistica y Cartografia de Andalucia, Consejeria de Economia, Innovacion, Ciencia y Empleo). Land economic factors include data on the real estate transactions (2004, number of transactions, Source: Diputacion de Sevilla, Anuario Estadistico de la Provincia de Sevilla) and the number of dwellings built (2001, number of houses, Source: Instituto Nacional de Estadistica). Finally, social factors are represented by the number of secondary schools (2005, number of centers, Source: SIMA).

3 Modeling

Here, we put forward a multiple regression model that relates the expansion of urban territories with the spatial population growth in the following form

$$\begin{aligned} y &= X\alpha + Z_1\beta^1 + Z_2\beta^2 + \gamma + \epsilon \\ \epsilon &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{1}$$

In (1) y contains a $n \times 1$ vector of the section-based dependent variable, X is a $n \times p$ matrix of the GIS-based explanatory variables describing the land use types of the given cell and its neighboring cells, Z_1 represents a $n \times k_1$ matrix of the socio-economic GIS-based explanatory variables and Z_2 is a $n \times k_2$ matrix of the municipality-based explanatory variables. α, β^1 and β^2 are $p \times 1$, $k_1 \times 1$ and $k_2 \times 1$ vectors of the corresponding coefficients respectively and γ is an intercept. The error term ϵ is a $n \times 1$ vector of independent identically normally distributed variables with mean equal zero and variance equal σ^2. I_n denotes a $n \times n$ identity matrix. Observations are accounted at the level of a single cell. Thus, the number of cells defines the sample size n.

We take the population density as dependent variable in the regression equation (1). Note that population density is defined at the lowest level of administrative division for the census data in Spain (census tracts called sections). As the section level is coarser than the cell level, we assign the value of the population density in a given section to every cell that belongs to this particular section. In the same way, we also extend the municipality-level socio-economic data to the level of a single cell on a GIS-based map.

Note also, that in general, X includes the type of land use in a focal cell and the types of land use in its Moore neighborhood, which comprises the eight cells surrounding a central cell on a two-dimensional square lattice. Alternatively, the information about the cell neighborhood can be aggregated and represented just by a number of urban cells in it (including the type of a focal cell). In what follows, we employ the latter aggregation for defining X.

3.1 Implementation

The computations are done in the Clojure programming language (version 1.6.0). Regression estimates are obtained using Incanter 1.9.0, a Clojure-based, R-like statistical computing and graphics environment for the JVM. For principal component analysis, we use R version 3.0.3.

The input GIS-based maps are stored in an ESRI ArcInfo ASCII raster file format. The census data is given as tabular records in a csv file format.

3.2 Pre-processing of the GIS-based Explanatory Variables

Before performing the regression analysis, we rescale and clean the source GIS-based maps. We use the rule that if either any of neighboring land use types of a cell are undefined or socio-economic data is not set for the corresponding section or municipality, we exclude the cell from the sample. The algorithm of GIS data cleaning is done sequentially:

Step 1. Exclusion of cells with neighborhoods containing undefined values—we exclude cells, which either belong to the border of the studied area or have undefined values in their neighborhood in any of the given maps in the dataset.

Step 2. Normalization of the GIS-based explanatory variables—we bring all the values of these variables into the range [0, 1].

Step 3. Exclusion of cells, which fall into protected natural areas in the Seville province, or cells whose neighborhood contains cells belonging to these areas.

Step 4. Exclusion of cells, which do not have a specified value of the population density in the section they belong to.

Note that protected natural areas include UNESCO World Heritage sites, Ramsar wetland sites, Nature network 2000 sites, biosphere reserves, protected areas and European Diploma sites. Since urban development is not allowed at all of these sites, they have been excluded from the sample.

3.3 Principal Component Analysis

First, we find out that the municipality-based variables exhibit strong pairwise correlation as shown in Table 2. Note that in case of moderate- and big-size samples (i.e., those containing more than 80 points), the critical Pearson correlation coefficient that ensures the statistical significance at 0.05 level is near 0.25.

Because of a large number of highly correlated variables, the principal component analysis is applied to reveal interrelationships and remove multicollinearity in the set of the municipality-based economic factors. This kind of transformation of the original variables allows obtaining orthogonal factors, which certainly do not correlate with each other.

In this case, we are able to reduce the number of factors to the first two principal components, which explain 99 % of the total sample variance. The first component represents the average yield of 9 out of 10 variables. The second principal component correlates with the remaining original explanatory variable—income per capita. The intuition behind the revealed two first principal components is the following. The first principal component separates municipalities with a high number of inhabitants (by assigning bigger values) from the underpopulated territories. The second principal component separates observations from the municipalities with medium population (in terms of the Seville Province) from other points in the sample.

3.4 Data Compilation

After the procedure described in Sect. 3.2, we have a dataset with 235,678 observations, which fall into 99 different municipalities with 1,219 sections. In contrast to the GIS data, municipalities and sections do not divide territory uniformly and vary in size significantly. In highly populated areas section size coincides with one cell, while the upper value of the size can exceed 20,000 cells in other territories.

Table 3 contains a list of variables, which serve as an input to the multiple regression analysis.

Table 2 Correlations of the municipality-based variables ($n = 99$)

	Income per capita	Pop employed	Pop active	Pop under 20	Pop from 20–65	Pop over 65	Real estate	Dwellings built	Schools
Income per capita	1	−0.06	−0.06	−0.07	−0.06	−0.05	−0.06	−0.12	−0.09
Pop employed	−0.06	1	1	1	1	1	0.99	0.96	0.99
Pop active	−0.06	1	1	1	1	1	0.99	0.96	0.99
Pop under 20	−0.07	1	1	1	1	0.99	0.99	0.97	0.99
Pop from 20–65	−0.06	1	1	1	1	1	0.99	0.96	0.99
Pop over 65	−0.05	1	1	0.99	1	1	0.99	0.94	0.99
Real estate	−0.06	0.99	0.99	0.99	0.99	0.99	1	0.96	0.99
Dwellings built	−0.12	0.96	0.96	0.97	0.96	0.94	0.96	1	0.97
Schools	−0.09	0.99	0.99	0.99	0.99	0.99	0.99	0.97	1

Table 3 Variables included in the multiple regression analysis

Variable	Range
GIS-based	
Number of cells with urban land use in the neighborhood	$0 \dots 9$
Distance to the nearest road	$[0, 1]$
Distance to the nearest area of commercial or industrial land use	$[0, 1]$
Distance to the nearest airport	$[0, 1]$
Distance to the nearest waterfront	$[0, 1]$
Distance to the nearest area of forest	$[0, 1]$
Distance to the nearest city with more than 10,000 inhabitants	$[0, 1]$
Distance to the nearest city with more than 50,000 inhabitants	$[0, 1]$
Municipality-based	
PC1 (first principal component)	$[-28.44, 0.89]$
PC2 (second principal component)	$[-4, 1.45]$
Section-based	
Population density (people per cell, 2001)	$[0, 2704]$

4 Results

In what follows, we take the log transformed population density as the dependent variable in the model.

The advantage of the classical multiple linear regression is that we can easily interpret the estimated coefficients if the standard assumptions of the model are fulfilled. However, the latter is rarely a case for the spatial data, and this study is not an exception. Figure 2c shows that residuals are not statistically independent and substantial heteroscedasticity is present when we apply regression (1) to all cells from the GIS-based maps which remain in the dataset after the pre-processing step described in Sect. 3.2.

A non-uniform administrative division of the territory is one of the reasons that may cause the violation of the error independence assumption. Sections necessarily represent the entire province, including areas where human economic activity is very low due to difficult or undeveloped terrain. As a rule, these sections contain many cells (more than 100) and are characterized by a low population density and monotonic change in the distance-related explanatory variables.

Unlike in, e.g., [9, 11], we do not use more advanced statistical models that can incorporate spatial autocorrelation in these areas, but apply a certain filtering to deal with non-independence and heteroscedasticity of the residuals. For this purpose, two filters have been constructed:

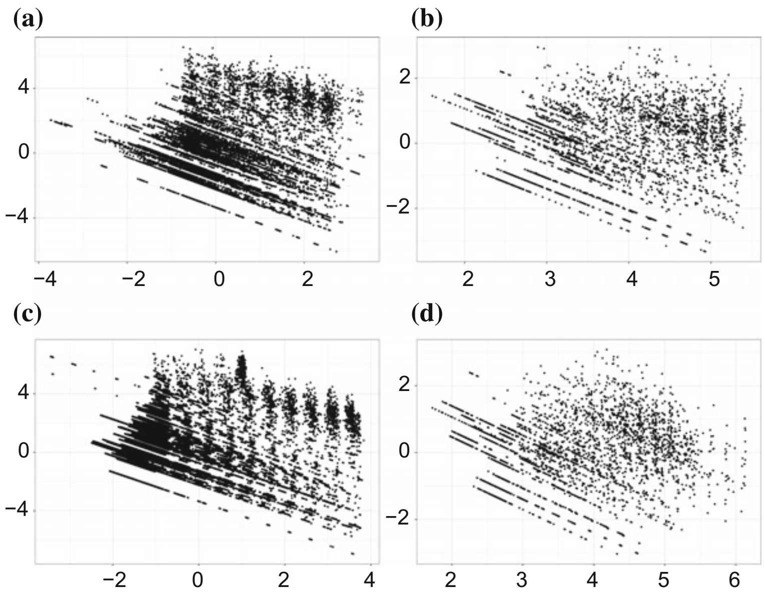

Fig. 2 Residuals (y-axis) versus predicted values (x-axis). **a** Filter 1—Model 1. **b** Filter 2—Model 1. **c** All cells—Model 1. **d** Filter 2—Model 2

Filter 1. Exclude all cells, which have either all non-urban or all urban cells in their Moore neighborhoods (including the focal cell itself).

Filter 2. Exclude all cells, which have all urban neighbors around them (including the focal cell itself) as in Filter 1, and additionally, exclude all sections, where the share of non-urban cells in the total number of cells is more than 70 % and the number of non-urban cells in this section exceeds 100.

Note that in both filters we exclude the urbanized cells with entirely urbanized neighborhoods, because the population density in these cells is likely to be highly dependent on some other (missing in this study) independent variables (for all urbanized cells with entirely urbanized neighborhoods, the distance to roads remains zero and the distance to commercial centers changes insignificantly to capture the population density variance).

4.1 Results of Multiple Regression

In the current exploratory research, we perform the regression analysis applying either Filter 1 or Filter 2, and also varying the number of the GIS-based explanatory variables. Figure 2a, b shows the residual plots against predicted values after applying Filter 1 and Filter 2 correspondingly. In case of the original GIS-based maps,

Table 4 Multiple regression results (Filter 2)

Variable	Coefficient
Model 1	
Number of urban neighbors	0.23 (0.22, 0.24)
Distance to roads	−8.42 (−10.86, −5.97)
Distance to the commercial centers	−1.46 (−2.44, −0.49)
PC1	0.02 (0.01, 0.02)
PC2	0.29 (0.26, 0.32)
R^2 (adjusted R^2)	0.42 (0.42)
Model 2	
Number of urban neighbors	0.22 (0.21, 0.23)
Distance to roads	−6.65 (−9.06, −4.25)
Distance to the commercial centers*	−0.07 (−1.02, 0.88)
Distance to airport	−1.45 (−2.28, −0.61)
Distance to waterfront	−0.35 (−0.7, 0.0)
Distance to forest	0.52 (0.31, 0.73)
Distance to the city with >10 ths people	−1.59 (−2.06, −1.13)
Distance to the city with >50 ths people	3.05 (2.27, 3.82)
PC1	0.01 (0.0, 0.02)
PC2	0.28 (0.25, 0.31)
R^2 (adjusted R^2)	0.47 (0.46)

*Not significant (p-value > 0.05)

the sample size n is 235,678 cells (case c in Fig. 2). Filter 1 reduces this number to 17,351 cells, while Filter 2 leaves 4,138 points for the regression analysis.

Here, we present two models, which differ in the number of explanatory variables. Model 1 includes two GIS-based explanatory variables, which have a visually identifiable correlation trend with the population density on the respective scatterplots. Model 2 takes all available factors as explanatory variables. Figure 2d presents the residual plot for Model 2 after Filter 2 has been applied.

Table 4 presents the estimates obtained from Model 1 and Model 2. In both cases we use the sample obtained from Filter 2.

4.2 Discussion

Figure 2 illustrates that filtering of the original observations facilitates the fulfillment of the standard multiple regression assumptions regarding the error independence and homoscedasticity. Note that a well-recognized problem in the analysis of spatial data is the presence of spatial autocorrelation in the dataset. The effects of spatial dependence on the conventional statistical methods are various, including a likely overestimation of R^2 and the unreliability of the significance tests. Filtering helps to deal with the spatial autocorrelation in such a way, that we select a sample of (presumably) independent observations (cells) from the original GIS-based land use map consequently removing spatial correlations in the sample.

Coefficient estimates in Table 4 indicate that the population density is higher in the cells with more urbanized neighborhoods and is lower in the cells which are far away from the transportation routes. The closer a cell is to commercial centers, the higher the population density is, but this variable has a lesser impact. All estimates in Table 4 are statistically significant (p-value < 0.05).

The R^2 coefficient suggests that about 40 % of the total variance in the population density is explained by Model 1 using Filter 2. This value almost doubles compared to the case when Filter 1 is used. Despite the fact that we cannot explain most of the variance of the population density using the collected set of socio-economic indicators inside the urban areas (and for the whole territory in general), the results obtained so far suggest, that the land use neighborhood partly captures the spatial pattern of the population distribution, which is caused by unknown drivers not included in this study.

5 Conclusion

We have showed that filtering procedures can be used to deal with a non-independence of observations in case of spatial data. Heteroscedasticity and correlation have been detected in the residuals of the standard multiple regression model in the case study of the Seville province. We have constructed special filters to isolate the territory at the borders of urban areas, where spatial correlation is not present. The obtained results have showed a more random pattern in the residuals of the regression model.

The model suggests that for the remaining part of the spatial data (after filtering), a decent part of the variance of the population density can be explained by the abundant land use types in the neighborhood of the focal cell. This finding can help obtain new insights related to the phenomenon of urban sprawl, which occurs at the fringe of urban areas outside of the city centers with high population densification.

Acknowledgments The authors would like to acknowledge DG research for funding through the FP7-funded COMPLEX project #308601, http://www.complex.ac.uk.

References

1. United Nations Population Division: World Population Prospects: The 2004 Revision and World Urbanization Prospects: The 2005 Revision. United Nations, New York (2005)
2. Angel, S., Sheppard, S.C., Civco, D.L.: The Dynamics of Global Urban Expansion. Transport and Urban Development Department, The World Bank, Washington, DC (2005)

3. Frenkel, A., Ashkenazi, M.: Measuring urban sprawl: how can we deal with it? Environ. Plann. B Plann. Des. **35**(1), 56–79 (2008)
4. Clarke, K.C., Hoppen, S., Gaydos, L.J.: A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay Area. Environ. Plann. B Plann. Des. **24**, 247–261 (1997)
5. Batty, M., Xie, Y., Sun, Z.: Modeling urban dynamics through GIS-based cellular automata. Comput. Environ. Urban Syst. **23**, 205–233 (1999)
6. Lopez, E., Bocco, G., Mendoza, M., Duhau, E.: Predicting land-cover and land-use change in the urban fringe: a case in Morelia City Mexico. Landscape Urban Plann. **55**, 271–285 (2001)
7. Cheng, J., Masser, I.: Urban growth pattern modeling: a case study of Wuhan City PR China. Landscape Urban Plann. **6**, 199–217 (2003)
8. Huang, B., Zhang, L., Wu, B.: Spatiotemporal analysis of rural-urban land conversion. Int. J. Geograph. Inf. Sci. **23**(3), 379–398 (2009)
9. Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht (1988)
10. LeSage, J., Pace,R.K.: Introduction to Spatial Econometrics. CRC Press (2009)
11. Overmars, K.P., de Koning, G.H.J., Veldkamp, A.: Spatial autocorrelation in multi-scale land use models. Ecol. Modell. **164**, 257–270 (2003)
12. Landis, G.D.: Imagining land use futures: applying the California urban futures model. J. Am. Plann. Assoc. **61**(4), 438–457 (1995)