International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: repository@iiasa.ac.at
Web: www.iiasa.ac.at

Working Paper

Exploratory Spatial Analysis of Regional Urbanization Patterns in the Province of Seville, Spain

Anna Shchiptsova (shchipts@iiasa.ac.at)
Richard Hewitt (richardjhewitt@hushmail.com)
Elena Rovenskaya (rovenska@iiasa.ac.at)

Approved by

Elena Rovenskaya
Program Director, Advanced Systems Analysis
September 2016

Contents

Abstract

Recent decades have seen rapid urban growth worldwide. In general, association between land use, population and economy is well-known, but our understanding of the direct economic and demographic drivers responsible for the urban land expansion is limited to the local case studies. Identification of the cause-effect links at regional and global scales is non-trivial in nature due to the complex socio-economic context of the phenomenon and the spatial character of data. This study examines the association between regional urbanization processes over space using resampling methods in regression analysis. Our main goal is to summarize the knowledge about the spatial urbanization patterns, which had been shaped in different areas of the Province of Seville, Spain in the last years of the economic growth period before the global financial crisis of 2007-08. For this purpose, we collect the available spatial cartographic and panel data on the land use, demographic and economic factors, and conduct experimental significance testing and parameter estimation using methods of permutations and bootstrapping. The model is applied at the aggregate scale of regional administrative subdivision, followed by a discussion of its approximation accuracy on a GIS lattice. Our non-temporal analysis points out that the land use variable and the spatially explicit proxies of economic activity (i.e., distance to a road and distance to a commercial center) significantly contribute to the level of population densification in different areas of the region, although the process of regional urbanization cannot be solely explained by the studied drivers.

Acknowledgments

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

About the Authors

Anna Shchiptsova is a Research Scholar with Advanced Systems Analysis Program at IIASA.

Richard Hewitt is an Associate at Observatorio para una Cultura del Territorio (OCT), Madrid, Spain.

Elena Rovenskaya is the Director of the Advanced Systems Analysis Program at IIASA; she is also a Research Scholar at the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia.

Exploratory Spatial Analysis of Regional Urbanization Patterns in the Province of Seville, Spain

Anna Shchiptsova
Richard Hewitt
Elena Rovenskaya

1 Introduction

In recent decades, we have been witnessing rapid urban growth worldwide, especially in the developing countries, but also in Europe, US and Australia. The accelerated rate of non-urban to urban land conversion drives human-induced alterations of biogeochemical cycles, climate, hydrosystems and biodiversity (Grimm et al., 2008); furthermore, it is likely to impact the global environment in the long run. By 2030, we expect cities to grow 2.5 times in their territory, consuming some 1 million km^2, or 1.1% of the total land area of the countries (Angel et al., 2005).

The conversion of land surface to urban uses is primarily caused by the original increase in the urban population. The UN estimates that the proportion of urban residents in the total population has grown from 37% in 1975 to 49% in 2005, and by 2030 this figure is expected to surpass 60% reaching 4.9 billion people (UN, 2005). Yet, we find evidence that economic constraints apparently influence the spatial urbanization patterns in different territories. For example, by 2030 the cities in developing countries are expected to increase their population by 50% and to triple their land area, with every new resident converting, on average, some 160 m^2 of non-urban to urban land during this period. At the same time, the cities in developed countries are likely to increase their population by 20% and expand 2.5 times their land area, with every new resident converting, on average, some 500 m^2 of non-urban to urban land (Angel et al., 2005). In general, association between land use, population and economy is well-known, but our understanding of the direct economic and demographic drivers responsible for the non-urban to urban land change is limited to the local case studies (Veldkamp and Lambin, 2001). At regional and global scales, we should consider urban expansion as an outcome of several types of human activities. Consequently, the complex socio-economic context of the phenomenon hampers identification of the cause-effect links between urbanization processes.

In Spain, intensive urban land expansion followed the country's entry into the European Union in 1986 and coincided with the boom in the liberalized economy before the global financial crisis of 2007-08. Artificial surfaces (i.e., surfaces with dominant human influence but without agricultural land use) grew nationally by 52% from 1987 to 2006 (OSE, 2012). The peak of this development came in the last years of the economic boom with a 21% increase in the artificial land between 2000 and 2006 years alone (OSE, 2012). This paper focuses on the urban land expansion in one of the fifty Spanish provinces, the Province of Seville. Our main goal is to summarize the knowledge about the spatial urbanization patterns, which had been shaped in this region by the end of the economic growth period. For this purpose, we collect available spatial cartographic and panel data on the land use, demographic and economic factors, and study their association over space using resampling methods in regression analysis.

Empirical literature contains regional case studies on urban land expansion originating from different fields, such as urban economics, geography, landscape ecology and climate modeling (Table 1). Remote sensing and image processing methods allow scientists to consider land use change using high-resolution GIS-based maps of several years. Typically, studies of spatially explicit data involve exploratory methods of descriptive statistics, e.g., cross-tabulation analysis (Hewitt and Escobar, 2011; Dewan and Yamaguchi, 2009) and pairwise correlation analysis (Lo and Yan, 2002). On the other hand, more advanced econometric models use aggregate spatial data to study the amount of land surface converted to urban uses in response to socio-economic changes. The data is usually available at the level of a coarse administrative subdivision (McGrath, 2005; Seto and Kaufmann, 2003; Deng et al., 2008). With few exceptions (Kumar, 2009), these studies identify population, income and agricultural land value as significant factors contributing to the size of urbanized land area in an administrative unit (Brueckner and Fansler, 1983; Paulsen, 2012).

However, conventional econometric approaches have some shortcomings in the context of urbanization processes. As a rule, researchers do not know all relevant variables, and the functional forms are chosen on the basis of convenience and familiarity (Freedman, 1997). For example, social relations and market effects are not subjected to direct measurement and are usually substituted with some proxy variables in a model (Overmars et al., 2003). Verification of statistical assumptions should include testing spatial autocorrelation in the dataset (e.g., Seto and Kaufmann, 2003). Otherwise, aspatial analysis of spatial data can cause potentially misleading results and consequent misinterpretation of the model output (Fotheringham and Rogerson, 1993; Overmars et al., 2003). In these conditions, statistical techniques are useful tools for summarizing the knowledge about spatial dataset, but cannot serve as a basis for the causal inference and subsequent prediction.

Table 1. A list of case studies on urban land expansion. "Type 1": data analysis of the GIS-based maps over several time periods, "Type 2": econometric modeling, "Type 3": spatially explicit simulation modeling.

Type	Study	Region	Technique/model
1	Lo and Yang, 2002	Atlanta, Georgia, US	Correlation analysis Multiple regression
1	Dewan and Yamaguchi, 2009	Dhaka, Bangladesh	Cross-tabulation analysis Multiple regression
1	Hewitt and Escobar, 2011	Madrid region, Spain	Cross-tabulation analysis
2	Brueckner and Fansler, 1983	40 urban areas in 1970, US	Monocentric model of Alonso-Muth-Mills Multiple regression
2	Seto and Kaufmann, 2003	Pearl River Delta, China	Random coefficient regression
2	McGrath, 2005	33 largest urban areas from 1950 to 1990, US	Monocentric model of Alonso-Muth-Mills Multiple regression
2	Deng et al., 2008	coterminous China	Monocentric model of Alonso-Muth-Mills Multiple regression
2	Paulsen, 2012	329 largest urban areas in 1980,1990 and 2000, US	Monocentric model of Alonso-Muth-Mills Multiple regression
3	White et al., 1997	Cincinnati, Ohio, US	Cellular automata
3	Clarke et al., 1997	San Francisco Bay area, US	Cellular automata
3	Schneider and Pontius, 2001	Ipswich Watershed, Massachusetts, US	Binary logistic regression Multi-criteria evaluation
3	Pijanowski et al., 2002	Michigan Grand Traverse Bay Watershed, US	Artificial neural network
3	Song et al., 2015	Beijing, China	Artificial neural network
1/2	Kumar, 2009	Delhi, India	Multiple regression
1/3	Wear and Bolstad, 1998	Southern Appalachians, US	Poisson regression Binomial logit regression
1/3	Braimoh and Onishi, 2007	Lagos, Nigeria	Binary logistic regression

The coarse sampling scale does not allow econometric models to consider landscape heterogeneity and predict locations of change on a GIS-based map. Thus, spatially disaggregate effects of land use policies cannot be examined (Bocstael and Irwin, 2000). As alternative, empirical studies in geography and landscape ecology have been focused on the spatially explicit simulation modeling of land use change using local transition rules at the level of a cell or at the level of an individual land parcel of the landscape. Different approaches to model transition of a cell (or a parcel) from one land use state to another have been adopted in the literature. In general, the methods include statistical estimations (Schneider and Pontius, 2001; Wear and Bolstad, 1998; Braimoh and Onishi, 2007), machine learning techniques (Pijanowski et al., 2002; Song et al 2015) and cellular automata models, such as SLEUTH (Clarke et al., 1997) and models of the Metronamica family (White et al., 1997). Additionally, the simulation modeling approach has been extended with GIS data analysis in the "hybrid models of land use/cover change" (Irwin and Geoghagen, 2001). In particular, the exploratory analysis of the GIS-based maps over several years provides parameters for the transition rules in a model (Wear and Bolstad, 1998; Braimoh and Onishi, 2007). Overall, simulation models establish that the hypothesized interaction among cells is a possible explanation of the historic changes in the land use (Irwin and Geoghagen, 2001) by replicating these changes (White et al., 1997; Schneider and Pontius, 2001). However, such models do not include in-depth statistical testing against actual data.

We develop a regression model using the non-temporal spatial data to deduce conditions associated with the level of population densification in different areas of Seville Province at the end of the economic growth period. Thus, this work aims to complement existing models of land use change by analyzing cumulative output of urbanization processes over a single economic phase. To address limitations of previous research, we sample data at the lowest level of administrative subdivision, for which population density records were available; subsequently, we assess model approximation accuracy at the level of a GIS lattice. At the same time, we expect some relevant variables to be omitted in the dataset of available landscape attributes and socio-economic indicators of the region, and therefore, the exact distribution of regression error terms might be difficult to specify. In this connection, we conduct experimental significance testing using the method of permutations (Fisher, 1935; Fotheringham and Rogerson, 1993) and apply a non-parametric bootstrap scheme for parameter estimation (Efron, 1979).

The rest of the paper is organized as follows. In section 2, we introduce a regression model and define resampling methods for its estimation. The study area of Seville Province and available cartographic and panel data are described in section 3. Section 4 presents outcomes of spatial data analysis. We discuss results and model limitations in section 5. Some final remarks are given in the conclusion in section 6.

2 Methodology

Let us consider urbanization processes in a geographic region consisting of n administrative units. Such administrative units are, in general, heterogeneous in shape and size. We define a GIS lattice in the geographic space by putting a regular grid of cells on a study area. By assumption, administrative subdivision of a region is equivalent to partitioning its GIS lattice into disjoint polygons of cells. Suppose that statistical records per administrative unit describe demography and economy of a region. Additionally, we recognize landscape attributes to be an integral part of local economic attractiveness. We assume that natural and man-made features of relief, including type of land use, are registered on a GIS lattice using remote sensing technologies. In general, we want to relate the population density to land use and economy in an administrative unit based on the reported cartographic and panel data.

For this purpose, we put forward a multiple regression model in the following form

$$\begin{gathered} y = X\beta + \varepsilon \\ \varepsilon_1, \dots, \varepsilon_n \sim F(0, \sigma^2) \end{gathered} \tag{1}$$

where y is a $n \times 1$ vector of observable population density, X is a $n \times (p+1)$ matrix of explanatory variables and $\beta = \left(\beta_0, \beta_1, \dots, \beta_p\right)^T$ is a $(p+1) \times 1$ vector of the unknown model parameters to be estimated from the data using the ordinary least squares method. Here, each column X^i consists of observations on a single explanatory variable. By assumption, X^1 is identically 1, so that the regression equation has an intercept β_0. For a GIS-based variable, an average value over cells belonging to the same administrative unit is considered to be a single observation on this variable in X. The error term ε is a $n \times 1$ vector of independent identically distributed errors with common distribution F having mean 0 and finite constant variance σ^2. Both F and σ^2 are unknown.

We search for a minimum combination of available variables, which produces statistically significant output in model (1) in terms of the method of permutations (Fisher, 1935). A permutation test computes the probability for a null hypothesis being tested, whether the original test statistic of interest is a typical element of the set of statistics derived from the given observations by an appropriate class of data reordering. The basic idea of the test is free of the assumption on the exact distribution of the test statistic; rather the reference distribution is generated from the drawn permutation sample. We run a permutation test with the test statistic R^2 to assess the overall significance of model (1). The null hypothesis of the test states that response variable has no linear relationship with the given explanatory variables, that is, $H_0\colon \beta_i = 0$ for all $i = 1 \dots p$ and the alternative hypothesis in the case of the two-tailed test is $H_1\colon \exists i\ \beta_i \neq 0\ (i = 1 \dots p)$. If the null hypothesis holds, the observations on the response variable y could have been observed in any order relative to the fixed value tuples in X. Therefore, we need to recalculate the test statistic for each of the possible permutations of y, leaving X fixed. In practice, the permutation test is usually approximated by reshuffling y k times and each time selecting a permutation randomly from the set of all possible permutations (Efron and Tibshirani, 1993). Accordingly, we calculate the approximate p-value of the test as

$$\tilde{p}\text{-value}(R^2) = \frac{\#\left\{\gamma = 1 \dots k \mid \hat{R}_\gamma^2 \geq \hat{R}^2\right\}}{k}, \tag{2}$$

where $\hat{R}_\gamma^2$ is a coefficient of determination from permutation γ of y, and $\hat{R}^2$ is a value of the test statistic for the original sample.

Further, we test significance of each individual coefficient in model (1). The null hypothesis $H_0: \beta_i = 0$ states that there is no linear relationship between y and X^i over and above any influence of the values in other explanatory variables. Thus, we need to eliminate variation caused by remaining variables and permute the obtained error terms in the reduced model without X^i. The vector of coefficients β is not known a priori, and consequently, it is impossible to carry out the exact permutation test (Anderson, 2001). Instead, we use an approximate permutation method (Freedman and Lane, 1983) as follows:

1) The test statistic $\hat{t}_i$ is computed from the coefficient t-statistic, when we regress y on X.
2) We compose a $n \times p$ matrix X', which contains all specified explanatory variables except X^i. The residuals from the regression of y on X' are subjected to permutation. For every permutation γ, we calculate a $n \times 1$ vector y' from the permuted residuals and the fixed value tuples in X'. The permutation t-statistic $(\hat{t}_i)_\gamma$ is obtained from the regression of y' on X.
3) After k replications, an approximate p-value is calculated from the generated reference distribution of t_i:

$$\tilde{p}\text{-value}(t_i) = \frac{\#\{\gamma = 1 \dots k \mid (\hat{t}_i)_\gamma \geq \hat{t}_i\}}{k} \tag{3}$$

The permutation hypothesis testing does not avoid the assumption on the observations to be independent and identically distributed. For completeness, we verify the absence of spatial autocorrelation in the residuals of model (1). In general, a pair of residuals r_i and r_j is considered to be spatially dependent, if administrative unit i is located in close proximity to administrative unit j and the covariance of r_i and r_j is non-zero. We define a row-normalized $n \times n$ matrix of spatial proximity W such that $w_{ij} = 0$, if administrative units i and j have no common boundary, i.e., they are not contiguous; otherwise, w_{ij} is inverse proportional to the total number of contiguous administrative units to administrative unit i. Here, the Moran's I (Moran, 1950) and Geary's C (Geary, 1954) coefficients are used to measure spatial autocorrelation based on the sample of residuals and the matrix of spatial proximity W. These coefficients equal

$$I = \frac{n \sum_{i,j=1 \dots n} w_{ij} r_i r_j}{\sum_{i,j=1 \dots n} w_{ij} \sum_{i=1 \dots n} r_i^2}, \tag{4}$$

$$C = \frac{(n-1) \sum_{i,j=1 \dots n} w_{ij} (r_i - r_j)^2}{2 \sum_{i,j=1 \dots n} w_{ij} \sum_{i=1 \dots n} r_i^2} \tag{5}$$

In the context of model (1), we apply the bootstrap hypothesis test (Efron and Tibshirani, 1993) with the null hypothesis of no spatial autocorrelation. That is, we test whether neighboring residuals take values that are more similar or less similar than expected if they were arranged randomly. The test statistic is calculated for original data

(y, X). After that, the data is resampled with replacement k times to get the reference test distribution. For every bootstrap replication, we change the matrix of spatial proximity W by including repetitive observations with the same weights as the original one and adjusting the spatial weights of neighbors accordingly. The approximate equal-tail p-value in the two-tailed test is defined as (Lin et al., 2011)

$$\tilde{p}\text{-value}\big(\hat{\theta}\big) = 2\min\left(\frac{\#\{\sigma = 1 \ldots k \mid \hat{\theta}_\sigma \leq \hat{\theta}\}}{k}, \frac{\#\{\sigma = 1 \ldots k \mid \hat{\theta}_\sigma > \hat{\theta}\}}{k}\right), \tag{6}$$

where $\hat{\theta}$ is a test statistic of interest and $\hat{\theta}_\sigma$ is a bootstrap statistic in run σ. In the two-tailed test, we define a probability of obtaining a result equal to or more extreme than the original test statistic $\hat{\theta}$ as a probability of obtaining a result outside of the equal-tailed interval, where one of the end points coincides with $\hat{\theta}$. An equal-tailed property means that the probability of a value to be from the left side of an interval is the same as the probability of a value to be from the right side of an interval (Efron and Tibshirani, 1993). In fact, we do both one-tailed tests and double the lowest p-value from these trials.

After that, we suppose that the specified model (1) is a correct one, and use the non-parametric bootstrap method (Efron, 1979) for parameter estimation (Freedman, 2009). Suppose that we want to estimate distribution of some statistic θ in model (1) with the specified matrix X and unknown distribution F on the basis of observed data (y, X). Consequently, we use the percentile bootstrap method (Efron and Tibshirani, 1993) to construct a confidence interval for the parameter of interest. Here, we perform the following steps to bootstrap a regression model:

1) The original data (y, X) is subjected to resampling with replacement. At first, a random sequence of indexes $(j_1, \ldots, j_n)$ is drawn from the set $\{1, \ldots, n\}$. We compose a $n \times 1$ vector y' and a $n \times (p+1)$ matrix X' by taking the selected pairs $\{(y_{j_1}, X_{j_1}), \ldots, (y_{j_n}, X_{j_n})\}$. For the sample (y', X'), we calculate bootstrap statistic $\hat{\theta}_\gamma$.
2) After k replications, we arrange a sequence $\hat{\theta}^*$ by taking bootstrap values $\{\hat{\theta}_\gamma\}_{\gamma=1\ldots k}$ in ascending order. For the given level of confidence α, we find the $\lfloor(1-\alpha)/2\,k\rfloor$ and $\lceil(1+\alpha)/2\,k\rceil$ quantiles in $\hat{\theta}^*$ and set them as the lower and upper borders of the $100\times\alpha$-% percentile confidence interval respectively. Here, $\lfloor(1-\alpha)/2\,k\rfloor$ denotes the largest integer not greater than $(1-\alpha)/2\,k$ and $\lceil(1+\alpha)/2\,k\rceil$ stands for the smallest integer not less than $(1+\alpha)/2\,k$.

Additionally, we treat model (1) as a stochastic approximation to the values on a GIS lattice. By assumption, the values of population and economic variables remain constant in the cells, which belong to the same administrative unit. Thus, we use the existing observations in a unit as a proxy for the true value at the level of a cell in this unit.

At first, let us consider approximation to the true value in an individual cell. Suppose that we have data (X', y'), where X' is a $N \times (p+1)$ matrix of explanatory variables and y' is a $N \times 1$ vector of population density values at the level of a cell; N is a total number of cells on a GIS lattice. For the given cell j, we define accuracy ρ_j of the response value from model (1) to the true value y'_j, observed in this cell, as a distance between the expected model response and the true value. That is

$$\rho_j = \left| y'_j - \hat{y}'_j \right|, \quad (7)$$

where $\hat{y}'_j = X'_j \beta$ and $j = 1, \dots, N$. $\hat{y}'_j$ is the fitted value of the response in model (1) for the observed values $X'_j = \left(x'_{j1}, \dots, x'_{jp} \right)$ of explanatory variables in cell j.

For the given subset of cells S, we measure the $(100 \times k)$-th percentile of the accuracy of the response values from model (1) to the true values $\left\{ y'_j \right\}_{j \in S}$, observed in the cells belonging to this subset, as a minimum value below which at least the k-th fraction of the cell accuracy values fall, and denote it by $\rho(k, S)$. To summarize the sample of accuracy values in subset S, we examine maximum accuracy value $\rho(1, S)$ and the sample quartiles, which coincide with the values of $\rho(0.25, S)$, $\rho(0.5, S)$ and $\rho(0.75, S)$. For the convention with standard notations, we denote the selected descriptive statistics by $\rho_{max}(S)$, $\rho(Q_1, S)$, $\rho(Q_2, S)$ and $\rho(Q_3, S)$ respectively.

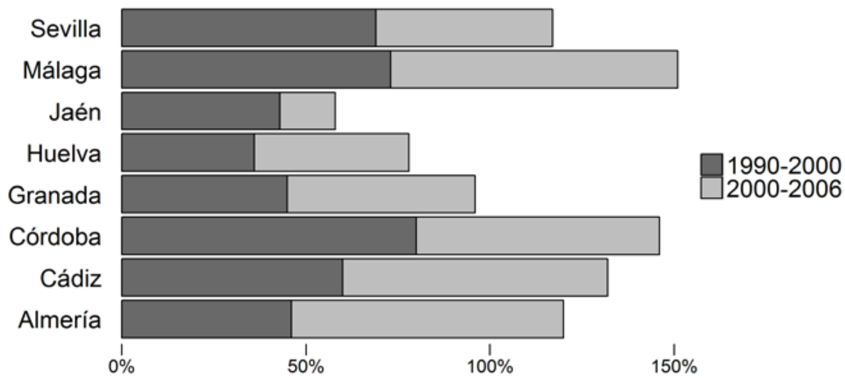
Note that, we cannot directly compute approximation accuracy in model (1). Instead, we bootstrap data (y, X) to get the reference distribution for an accuracy statistic. The values of $\rho_{max}(S)$, $\rho(Q_1, S)$, $\rho(Q_2, S)$ and $\rho(Q_3, S)$ are simply treated as model parameters, and therefore, we follow the percentile bootstrap scheme described above. Specifically, we calculate accuracy values for an estimated vector of coefficients $\hat{\beta}_\sigma$ and summarize the sample by the selected percentiles in every bootstrap replication σ. Thus, we obtain separate bootstrap samples for $\rho_{max}(S)$, $\rho(Q_1, S)$, $\rho(Q_2, S)$ and $\rho(Q_3, S)$, and find $100\times\ \alpha$-% percentile confidence intervals for these statistics.

3 Study Area and Data

The Province of Seville is located in the Mediterranean region of Andalusia in the southwestern part of Spain. It contains the region's capital, Seville and is the largest of Andalusia's eight provinces, both by surface area (14000 km^2) and by population (1.9 m. inhabitants). The Andalusia region as a whole has recently seen strong urban expansion, with more growth between 2000-2006 than in the preceding 13 years (OSE, 2012). The high proportion of the discontinuous urban fabric with respect to the total urban fabric (see Figure 1) indicates that the majority of new urban development has been of low density.

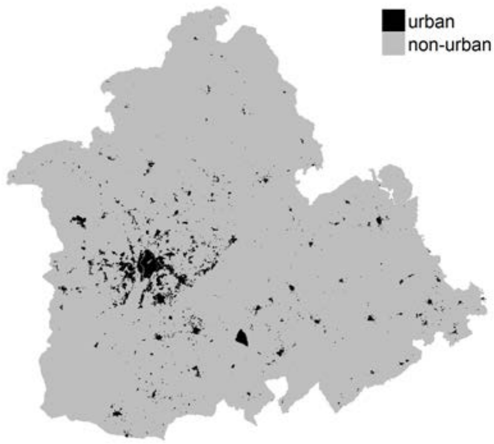
Though the province has not suffered from the sprawling suburbanization as happened in Spanish coastal areas (e.g., Malaga), notable growth of built-up areas and infrastructures has occurred around Seville city, especially to the south (Dos Hermanas, Alcala de Guadaira) and northwest (Bormujos, Tomares, Mairena de Aljarafe) of the city. On one hand, urban land has expanded near the urban cores with high population density. On the other hand, there has been a marked tendency for urban and industrial land parcels to become less concentrated and dispersed across a wider area, particularly along transportation routes.

Figure 1. Urban growth in Andalusia, years 1990-2006. New discontinuous urban fabric as a percentage of new total urban fabric (Source: EEA, 2015).



We use spatial cartographic and panel data from the regional government (REDIAM, 2015; SIMA, 2015; IECA, 2015) and national government sources (INE, 2015). In general, we rely on the available data at the end of the booming period in Spanish economy. The population density data (number of people per cell, Source: INE, 2015) is available for 2001 at the lowest level of administrative subdivision for the census data in Spain, called sections. At the same time, land use mapping (Source: REDIAM, 2015) is obtained for the year 2003. The land use data is converted to raster by the cell-center method at a resolution of 200x200 m. We reclassify the more than one hundred original land use categories into urban land class (including all artificial surfaces) and non-urban land class (including vegetation, wetlands, agricultural land and water). Figure 2 illustrates the resultant urban/non-urban land distribution in the region.

Figure 2. Land use distribution in the Province of Seville (Spain), year 2003 (Source: REDIAM, 2015).



Additionally, we take the number of cells with urban land use in a cell neighborhood, including the cell itself, as a cell value on the GIS lattice in the sample. Here, we consider the Moore neighborhood, which comprises the eight cells surrounding a central cell on a two-dimensional square lattice.

Statistical records on social and economic activity in the region are available at the level of a municipality, which is a coarser scale of administrative subdivision than the scale of sections. In this connection, we extend data to the level of a section by assigning the value in a given municipality to every section that belongs to this particular municipality. The group of general economic variables includes the data on income per capita (2003, Euros, Source: SIMA, 2015), people employed (2001, number of people, Source: SIMA, 2015), and economically active population (2001, number of people, Source: SIMA, 2015). Population factors consist of 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: IECA, 2015). Land economic variables include data on real estate transactions (2004, number of transactions, Source: AEPS, 2015) and number of dwellings built (2001, number of houses, Source: INE, 2015). Finally, social factors are represented by the number of secondary schools (2005, number of centers, Source: SIMA, 2015). To avoid multicollinearity in the data, we transform these factors using the principal component analysis (Jolliffe, 2002).

We collect GIS-based data on physical proximity of a cell to different infrastructure objects, landscape features and geographic locations. These factors represent spatially explicit proxy measurements of economic activity in the region (Veldkamp and Lambin, 2001). The first part of topographic data comprises the distance to the nearest road (2005, km, Source: CNIG, 2015), the distance to the nearest area of commercial or industrial land use (2006, km, Source: EEA, 2015) and the distance to the nearest airport (2006, km, Source: EEA, 2015). Landscape data describes distance to the nearest waterfront (2005, km, Source: CNIG, 2015) and distance to the nearest area of forest (2006, km, Source: EEA, 2015). Lastly, we collect data on proximity to a city center with more than 10,000 inhabitants (2011, km, Source: INE, 2015) and on proximity to a city center with more than 50,000 inhabitants (2011, km, Source: INE, 2015).

Before statistical modeling, we clean and rescale the source GIS-based maps of the land use and topographic variables. At first, we exclude cells from the sample, which either contain an undefined value or have undefined values in their Moore neighborhood in any of the given maps in the dataset. After that, we apply normalization by bringing cell values of an individual variable into the range $[0,1]$:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{8}$$

where x is an original cell value, x_{min} and x_{max} denote the minimum and maximum values among all cell values of the variable on a regional GIS lattice. Finally, we drop cells, which fall into protected natural areas in Seville Province (REDIAM, 2015), or cells whose Moore neighborhood contains cells belonging to these areas, as urban development is not possible in these territories. Note that protected natural areas include UNESCO World Heritage sites, Ramsar wetland sites, Nature network 2000 sites, biosphere reserves and European Diploma sites.

4 Results

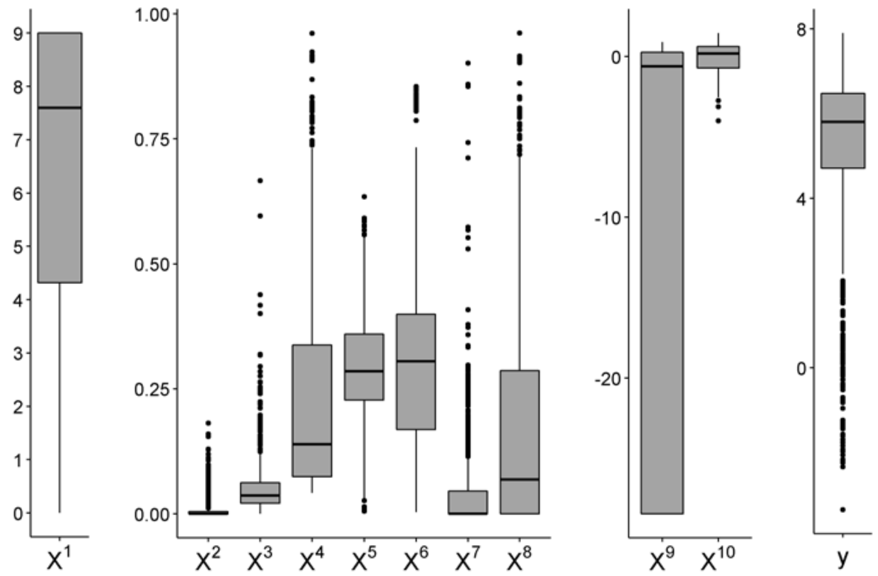
After carrying out principal component analysis, we are able to reduce the number of economic factors gathered from statistical records to two variables. The first principal component represents the average yield of 9 out of 10 collected economic variables. The remaining 10th factor correlates with the second component. Overall, both principal components explain 99% of the total variance of the aggregate economic data in municipalities. After data preprocessing, we compose a list of regression variables, which is shown in Table 2.

Table 2. Regression variables.

Variable	Name	Type of original data
X^1	Number of cells with urban land use in the Moore neighborhood	GIS-based map
X^2	Distance to a road	GIS-based map
X^3	Distance to a commercial center	GIS-based map
X^4	Distance to an airport	GIS-based map
X^5	Distance to waterfront	GIS-based map
X^6	Distance to forest	GIS-based map
X^7	Distance to a city (> 10 ths inh.)	GIS-based map
X^8	Distance to a city (> 50 ths inh.)	GIS-based map
X^9	First principal component of economic data	Statistical records per municipality
X^{10}	Second principal component of economic data	Statistical records per municipality
y	Logarithmic population density	Statistical records per section

The sample contains 836 points (sections), which fall into 99 municipalities. In total, the section's territory comprises 235,678 cells on the GIS lattice. As a useful summary, Figure 3 illustrates the range of sample observations.

Figure 3. Box plots of the sample observations at the level of a section.
Scale on y-axis: X^1 - 1:1 [cell], X^2 - 1: 19.36 [km], X^3 - 1:22.56 [km], X^4 - 1:109.05 [km], X^5 - 1:55.42 [km], X^6 - 1:29.01 [km], X^7 - 1:62.55 [km], X^8 - 1:103.25 [km], X^9, X^{10} - 1:1 [-], y - ln(people per cell).



We fit model (1) to the combinations of explanatory variables described in Table 2. In permutation hypothesis tests, we compute p-values for R^2 and $t_1, \ldots, t_{10}$ statistics using formulas (2) and (3) respectively. We calculate a 95%-normal approximation interval around the estimated p-value, which equals $\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/k}$ and defines approximation precision at the given level of confidence. Estimations suggest that the land use variable (X^1), the distance to a road (X^2) and the distance to a commercial center (X^3) represent the minimum combination of explanatory variables in model (1), for which the null hypothesis of no relationship is rejected in the permutation tests of overall model significance and individual coefficients significance at the level 0.05 (Table 3).

Table 3. Approximate p-values with 95% normal approximation confidence intervals. "Model 1": model (1) specified with all explanatory variables, "Model 2": model (1) specified with X^1, X^2 and X^3 explanatory variables. The ** symbol indicates significance at the level 0.05.
(number of replications 10,000)

	Model 1	Model 2
$\tilde{p}$-value (t_1)	0.000 (0.000, 0.000)**	0.000 (0.000, 0.000)**
$\tilde{p}$-value (t_2)	0.000 (0.000, 0.000)**	0.000 (0.000, 0.000)**
$\tilde{p}$-value (t_3)	0.004 (0.003, 0.005)**	0.008 (0.006, 0.010)**
$\tilde{p}$-value (t_4)	0.726 (0.717, 0.735)	–
$\tilde{p}$-value (t_5)	0.296 (0.287, 0.305)	–
$\tilde{p}$-value (t_6)	0.959 (0.955, 0.963)	–
$\tilde{p}$-value (t_7)	0.262 (0.254, 0.271)	–
$\tilde{p}$-value (t_8)	0.551 (0.541, 0.560)	–
$\tilde{p}$-value (t_9)	0.161 (0.154, 0.168)	–
$\tilde{p}$-value (t_{10})	0.375 (0.366, 0.385)	–
$\tilde{p}$-value (R^2)	0.000 (0.000, 0.000)**	0.000 (0.000, 0.000)**

Results of bootstrap tests show that model (1) specified with X^1, X^2 and X^3 variables is consistent with the null hypothesis of no spatial autocorrelation in the residuals at the significance level 0.05. Table 4 shows a mean estimate with a 95% percentile confidence interval for the Moran's I coefficient specified by (4) and the Geary's C statistics given in (5) used as measures of spatial autocorrelation. The p-value in hypothesis testing is computed from formula (6).

Table 4. Bootstrap estimates and approximate p-values for the measures of spatial autocorrelation in model (1) specified with X^1, X^2 and X^3 explanatory variables.
(number of replications 10,000)

Statistic	Mean	95% percentile confidence interval	p-value
Moran's I	0.005	(-0.074, 0.085)	0.689
Geary's C	0.993	(0.850, 1.144)	0.972

After permutation testing, we specify model (1) with the statistically significant combination of X^1, X^2 and X^3 variables, and bootstrap regression parameters (Table 5).

The bootstrap estimates of model coefficients 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 a commercial center, the higher the population density is, but this variable has a lesser impact. The R^2 estimate in Table 5 suggests that about 85% of the total variance in the logarithmic population density is explained by X_1, X_2 and X_3 variables.

Table 5. Bootstrap estimates in model (1) specified with X^1, X^2 and X^3 explanatory variables. (number of replications 10,000)

Coefficient	Mean	95% percentile confidence interval	Statistic	Mean	95% percentile confidence interval
β_1	0.414	(0.370, 0.454)	R^2	0.851	(0.829, 0.872)
β_2	-34.415	(-45.704, -25.389)	MSE (mean square error)	0.754	(0.654, 0.859)
β_3	-3.700	(-5.889, -1.416)			
β_0	2.966	(2.656, 3.301)			

In the end, we derive how accurate model (1) with X_1, X_2 and X_3 variables approximates population density values at the level of a cell on the GIS lattice. Table 6 demonstrates results for the cells in the entire GIS-based map and for the cells grouped by the number of cells with urban land use in the cell Moore neighborhood including the cell itself. We consider that model (1) is weakly accurate, if the accuracy estimate is less than the interquartile range (IQR) of the population density sample, which equals 1.76 people per cell. For the entire GIS-based map, the expected accuracy drops from 185% of IQR for Q_3 to 119% for Q_2, and then, to 60% of IQR for Q_1. The same tendency (178%-115%-59%) is detected, when we examine expected approximation accuracy for the group of cells with no urban cells in the Moore neighborhood. However, this result is expected as the cells in the group compose 92% of the total cell amount in the sample. Overall, model (1) is weakly accurate only for 25% of cells in each group. Note that approximation precision measures the degree of variation in the model response under the fixed values of explanatory variables and coincides with the variance in the error terms in model (1). Its estimate is given by the mean square error in Table 5.

Table 6. Bootstrap mean with 95% percentile confidence intervals for the accuracy percentiles (people per cell) in the cells grouped by the number of cells with urban land use in a cell Moore neighborhood in model (1) specified with X^1, X^2 and X^3 explanatory variables. "All": the sample of the entire GIS-based map, "non-urban": the sample of cells with no urban cells in the Moore neighborhood, "urban": the sample of cells with at least one urban cell in the Moore neighborhood (including the cell itself) (number of replications 10,000).

S	number of points	$\rho(Q_1,S)$	$\rho(Q_2,S)$	$\rho(Q_3,S)$	$\rho_{max}(S)$
Group "non-urban"	216,729	1.04 (0.92, 1.15)	2.02 (1.81, 2.29)	3.14 (2.85, 3.50)	15.95 (11.92, 20.97)
Group "urban"	18,949	1.74 (1.65, 1.84)	3.34 (3.20, 3.48)	4.56 (4.41, 4.71)	10.72 (9.96, 14.18)
Group "all"	235,678	1.06 (0.95, 1.18)	2.10 (1.88, 2.34)	3.26 (3.00, 3.61)	15.95 (11.92, 20.96)

5 Discussion

In this paper, we have performed a non-temporal analysis of spatial land use, population and economic data to analyze cumulative output of these urbanization processes in different areas of Seville Province. The outlined algorithm of statistical modeling and estimation focuses on several inherent aspects of the spatial phenomenon under study. First and foremost, we recognize that some relevant variables can be omitted in the collected dataset. Consequently, we have conducted experimental significance testing using permutations and bootstrapping to summarize the partial knowledge about spatial dependencies. In the second place, we have chosen to sample data at the lowest level of administrative subdivision, which is simultaneously the most accurate spatial measure available for the population density data in Seville Province. Furthermore, the methodology is not directly applicable at the level of the GIS lattice, where the necessary assumption of independent observations does not hold due to the presence of spatial autocorrelation.

Our results suggest that some important drivers of land use change are indeed missing from the dataset. The following thought experiment may help to illustrate this outcome. Let's assume that conversion of land is being caused only by some changes in the selected demographic and economic variables from the very start of urbanization process. Then it is reasonable to expect association between population, economic indicators and land use distribution in the resultant cumulative land use map after several years of conversion. However, our regression model shows low precision in comparison with the spread of population density values. The expected mean square error makes 43% of the IQR in the density sample with the 95% confidence interval from 37% to 49%. Results indicate that distribution of the error terms contains substantial information about population growth. For that reason, we cannot argue that cumulative change can be solely explained by the studied drivers.

The scale dependence of the reported results is examined using approximation accuracy of the model at the higher resolution of a GIS lattice. In general, the model is accurate (relative to the IQR of the population density sample) in 25% of cells in the entire GIS-based map and in any of the cell groups studied. Moreover, accuracy estimates exceed model precision, which points out significant variation in the values of explanatory variables from the section average. On the other hand, we do not consider landscape heterogeneity by excluding slope and elevation from the dataset. We expect these variables to correlate with the remaining ones in the regression model. However, application of the both landscape features can further refine in what spatial conditions aggregate results are biased at the fine scale. In general, we anticipate that smaller sections have higher urbanization rate and better approximation accuracy.

In the case study of Seville Province, we have come to the conclusion that economic variables from statistical records provide little or no information about the population density. Possibly, non-relevance of these factors is caused by the scale at which observations were sampled. The sampling has been done in 99 municipalities, which is approximately 10 times less than the number of sections. Even so, economic activity is represented by the proxy GIS-based variables in the model. Inclusion of these variables agrees with many studies on land use drivers (Veldkamp and Lambin, 2001). We found it interesting that significant proxies (i.e., distance to a road and distance to a commercial center) are the same location variables which often used in the investigation of land use change on a GIS lattice (e.g., Wear and Bolstad, 1998). In contrast, our research has examined a static snapshot of urbanization processes and was conducted at the scale of administrative subdivision.

Note that, the urban land variable (i.e., number of cells with urban land use in the Moore neighborhood) has a limited range of values and proxy variables enter into the regression equation with negative signs. Immediately, we can deduce an upper bound on the predictions from model (1). Formally, the expected logarithmic density in a section is less or equal than $9 \times 0.454 + 3.301 = 7.387$ people per cell with the right border of the 95% percentile confidence interval around this expected value equal to 8.246, as results in Table 5 suggest. Besides, we draw attention to the limited applicability of the proxy measurements. In the areas with high population density, the distance proxies take zero values, but evidently, cannot become negative. On the other hand, raw economic variables, e.g., income per capita, do not have such constraint on their value range. This can be an argument for a land use modeler to introduce a new condition into a model of land use change, such that sections in which population density exceeds the calculated upper bound should have a totally urbanized territory.

6 Conclusion

The outcomes of our study are somewhat discouraging for the goals of land use modeling. On one hand, statistical analysis at the fine aggregate scale points out that the economic panel data should be substituted with the spatially explicit proxy measurements. At the same time, the model with proxy variables is imprecise and simultaneously inaccurate (even in the rough estimation) at the cell level. Therefore, we have concluded that non-temporal spatial analysis of land use, population and economy provides partial knowledge about drivers standing behind urbanization phenomenon. Besides, if we assume that cumulative land use change can be inferred from a set of measurable variables, then not all of these urbanization driving forces are reflected in the collected dataset.

To our understanding, this conclusion shows methodological difficulties at modeling socio-economic processes with high spatial resolution. After all, no conventional guidelines exist on how to measure these variables at the level of a GIS lattice. We come out with the recommendation for a land use modeler to necessarily incorporate uncertainty associated with economic drivers in a model of land use change. In this case, the quantified interdependence between population densification and urban land distribution can help to refine the probability of change in different areas of the land use map. We view this direction as a promising one for future research.

Software

All computations were done using Clojure v.1.8.0 and Incanter v.1.5.7, a Clojure-based, R-like statistical computing and graphics environment for the JVM. The software components with data can be accessed through the link http://www.iiasa.ac.at/web/home/research/researchPrograms/AdvancedSystemsAnalysis/land-use-spatial-analysis.html at the International Institute For Applied Systems Analysis (IIASA). 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.

References

[1] Anderson, M. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences, 58(3): 626-639. DOI: 10.1139/f01-004

[2] Angel S., Sheppard S. C., Civco D. L., Buckley R., Chabaeva A., Gitlin L., Kraley A., Parent J., & Perlin M. (2005). The Dynamics of Global Urban Expansion. The World Bank, Washington D.C : Transport and Urban Development Department.

[3] Bockstael, N. E., & Irwin, E. G. (2000). Economics and the land use-environment link. In H. Folmer, & T. Tietenberg (Eds.), The international yearbook of environmental and resource economics 1999/2000 (pp. 1-54) Edward Elgar Publishing, Northampton, MA.

[4] Braimoh, A. K., & Onishi, T. (2007). Spatial determinants of urban land use change in Lagos, Nigeria. Land Use Policy, 24(2): 502-515. DOI:10.1016/j.landusepol.2006.09.001

[5] Brueckner, J. K., & Fansler, D. A. (1983). The economics of urban sprawl: Theory and evidence on the spatial sizes of cities. The Review of Economics and Statistics, 65(3): 479-482. DOI: 10.2307/1924193

[6] Clarke, K. C., Hoppen, S., & Gaydos, L. (1997). A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. Environment and planning B: Planning and design, 24(2): 247-261. DOI: 10.1068/b240247

[7] de Almeida, C. M., Batty, M., Monteiro, A. M. V., Camara, G., Soares-Filho, B. S., Cerqueira, G. C., & Pennachin, C. L. (2003). Stochastic cellular automata modeling of urban land use dynamics: empirical development and estimation. Computers, Environment and Urban Systems, 27(5): 481-509. DOI:10.1016/S0198-9715(02)00042-X

[8] Deng, X., Huang, J., Rozelle, S., & Uchida, E. (2008). Growth, population and industrialization, and urban land expansion of China. Journal of Urban Economics, 63(1): 96-115. DOI:10.1016/j.jue.2006.12.006

[9] Dewan, A. M., & Yamaguchi, Y. (2009). Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. Applied Geography, 29(3): 390-401. DOI:10.1016/j.apgeog.2008.12.005

[10] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7(1): 1-26. DOI:10.1214/aos/1176344552

[11] Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

[12] Fisher, R. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.

[13] Fotheringham, S., & Rogerson, P. (1993). GIS and Spatial Analytical Problems. International Journal of Geographical Information Systems, 7(1): 3-19. DOI: 10.1080/02693799308901936.

[14] Freedman, D., & Lane, D. (1983). A Nonstochastic Interpretation of Reported Significance Levels. Journal of Business & Economic Statistics, 1(4): 292-298. DOI: 10.2307/1391660

[15] Freedman, D. (1997). From Association to Causation via Regression. Advances in Applied Mathematics, 18(1): 59-110. DOI:10.1006/aama.1996.0501

[16] Freedman, D. (2009). Statistical models: Theory and Practice. Cambridge University Press.

[17] Geary, R. (1954). The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, 5(3): 115-145. DOI: 10.2307/2986645

[18] Grimm N. B., Faeth S. H., Golubiewski N. E., Redman C. L., Wu J., Bai. X., & Briggs J. M. (2008). Global change and the ecology of cities. Science, 319: 756-760. DOI: 10.1126/science.1150195

[19] Hewitt, R., & Escobar, F. (2011). The territorial dynamics of fast-growing regions: Unsustainable land use change and future policy challenges in Madrid, Spain. Applied Geography, 31(2): 650-667. DOI:10.1016/j.apgeog.2010.11.002

[20] Irwin, E. G., & Geoghegan, J. (2001). Theory, data, methods: developing spatially explicit economic models of land use change. Agriculture, Ecosystems & Environment, 85(1): 7-24. DOI:10.1016/S0167-8809(01)00200-6

[21] Kumar, P. (2009). Assessment of economic drivers of land use change in urban ecosystems of Delhi, India. AMBIO: A journal of the human environment, 38(1): 35-39. DOI: 10.1579/0044-7447-38.1.35

[22] Lin, K.-P., Long, Z.-H., & Ou, B. (2011). The Size and Power of Bootstrap Tests for Spatial Dependence in a Linear Regression Model. Computational Economics, 38(2): 153-171. DOI: 10.1007/s10614-010-9224-0

[23] Lo, C. P., & Yang, X. (2002). Drivers of land-use/land-cover changes and dynamic modeling for the Atlanta, Georgia metropolitan area. Photogrammetric Engineering and Remote Sensing, 68(10): 1073-1082.

[24] McGrath, D. T. (2005). More evidence on the spatial scale of cities. Journal of Urban Economics, 58(1): 1-10. DOI:10.1016/j.jue.2005.01.003

[25] Moran, P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37(1-2): 17-23. DOI: 10.2307/2332142

[26] Overmars, K., de Koning, G., & Veldkamp, A. (2003). Spatial Autocorrelation in Multi-scale Land Use Models. Ecological Modelling, 164: 257-270. DOI: 10.1016/S0304-3800(03)00070-X

[27] Paulsen, K. (2012). Yet even more evidence on the spatial size of cities: Urban spatial expansion in the US, 1980–2000. Regional Science and Urban Economics, 42(4): 561-568. DOI:10.1016/j.regsciurbeco.2012.02.002

[28] Pijanowski, B. C., Brown, D. G., Shellito, B. A., & Manik, G. A. (2002). Using neural networks and GIS to forecast land use changes: a land transformation model. Computers, environment and urban systems, 26(6): 553-575. DOI:10.1016/S0198-9715(01)00015-1

[29] Schneider, L. C., & Pontius, R. G. (2001). Modeling land-use change in the Ipswich watershed, Massachusetts, USA. Agriculture, Ecosystems & Environment, 85(1): 83-94. DOI:10.1016/S0167-8809(01)00189-X

[30] Seto, K. C., & Kaufmann, R. K. (2003). Modeling the drivers of urban land use change in the Pearl River Delta, China: Integrating remote sensing with socioeconomic data. Land Economics, 79(1): 106-121. DOI: 10.2307/3147108

[31] Song, W., Pijanowski, B. C., & Tayyebi, A. (2015). Urban expansion and its consumption of high-quality farmland in Beijing, China. Ecological Indicators, 54: 60-70. DOI:10.1016/j.ecolind.2015.02.015

[32] UN (2005). United Nations Population Division - World Population Prospects: The 2004 Revision and World Urbanization Prospects: The 2005 Revision. New York: United Nations.

[33] Veldkamp, A., & Lambin, E. (2001). Editorial: Predicting Land-use Change. Agriculture, Ecosystems and Environment, 85(1-3): 1-6. DOI: 10.1016/S0167-8809(01)00199-2
[34] Wear, D. N., & Bolstad, P. (1998). Land-use changes in southern Appalachian landscapes: spatial analysis and forecast evaluation. Ecosystems, 1(6): 575-594. DOI: 10.1007/s100219900052
[35] White, R., Engelen, G., & Uljee, I. (1997). The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. Environment and Planning B: Planning and Design, 24(3): 323-343. DOI: 10.1068/b240323

Web References

[1] AEPS (2015). Anuario Estadístico de la Provincia de Sevilla, Diputación de Sevilla. http://www.dipusevilla.es/portalestadistico/anuario/. Accessed 1.08.16.
[2] CNIG (2015). Download Center of the National Center for Geographic Information. http://centrodedescargas.cnig.es/CentroDescargas/inicio.do/. Accessed 1.08.16.
[3] EEA (2000). The revised and supplemented Corine land cover nomenclature. EEA technical report no 38. http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-clc1990-250-m-version-06-1999/corine-land-cover-technical-guide-volume-2/corine-land-cover-technical-guide-volume-2/. Accessed 1.08.2016.
[4] EEA (2015). European Environment Agency, Copernicus Land Monitoring service CORINE land cover. http://land.copernicus.eu/pan-european/corine-land-cover/. Accessed 1.08.16.
[5] IECA (2015). Instituto de Estadística y Cartografía de Andalucía. http://www.juntadeandalucia.es/institutodeestadisticaycartografia/. Accessed 1.08.16.
[6] INE (2015) Instituto Nacional de Estadística. http://www.ine.es/. Accessed 1.08.16.
[7] OSE (2012) Informe de sostenibilidad en España 2012. http://www.upv.es/contenidos/CAMUNISO/info/U0637061.pdf/. Accessed 1.08.16.
[8] REDIAM (2015) Andalusian Government environmental information service. http://www.juntadeandalucia.es/medioambiente/site/rediam/. Accessed 1.08.16.
[9] SIMA (2015) Sistema de Información Multiterritorial de Andalucía. http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2-en.htm/. Accessed 1.08.16.