International Journal of Statistics and Applications 2015, 5(4): 150-156
DOI: 10.5923/j.statistics.20150504.03

Estimating Bias of Omitting Spatial Effect in Spatial Autoregressive (SAR) Model

Olusanya E. Olubusoye, Oluyemi A. Okunlola*, Grace O. Korter

Department of Statistics, Faculty of Science, University of Ibadan, Nigeria

Abstract Regression models commonly used to analyze cross-section and panel data assume that observations/regions are independent of one another. Relaxing this assumption of independent observations in a cross sectional setting requires that we provide a parsimonious way to specify structure for the dependence between the n observational units that make up our size n data sample. Spatial econometrics techniques allow us to account for dependence between observations which often arise when observations are collected from points or regions located in space. In this application, Monte Carlo experiment was designed using R codes to assess the performance of spatial and non spatial model. Spatial autoregressive (SAR) model was used as a typical spatial model and ordinary least squares (OLS) as non spatial model. The study showed that OLS estimate of SAR model is bias and inconsistent. Also, it is found that bias emanating from omitting spatial effect is a function of degree of spatial autocorrelation.

Keywords Spatial econometrics, Spatial autoregressive, Spatial effects, OLS, Spatial autocorrelation

1. Introduction

Sample data collected for regions or points in space are not independent, but rather spatially dependent, which means that observations from one location tend to exhibit values similar to those from nearby locations. For many social and economic processes, a better appreciation of the spatial context can potentially avoid misleading inferences and improve the strength of results and their interpretation. Knowledge about the location of a process and its interaction with processes at neighboring locations can help infer the underlying reasons and logic of the process under investigation.

Existing regression models used to analyze cross-section and panel data assume that observations/regions are independent of one another. However, in real life this is not usually the case. For instance, a conventional regression model that relates commuting times to work for region *i* to the number of persons in region *i* utilizing different commuting modes and the density of commuters in region *i*, assumes that mode choice and density of a neighboring region, say *j* does not have an influence on commuting time for region *i*. Since it seems unlikely that region *i*'s network of vehicle and public transport infrastructure is independent from that of region *j*, we would expect this assumption to be unrealistic. Ignoring this violation of independence between observations will produce estimates that are biased and inconsistent. In addition, misspecification increases the probability of wrong inferences at least as much as does the choice of a biased or inefficient estimator.

Analysis of economic data explicitly linked to location can be approached from two spatial econometrics perspectives namely, spatial heterogeneity and spatial autocorrelation. Spatial heterogeneity might arise due to a lack of structural stability across space, such as varying parameters or functional forms, and due to non homogeneity of the units of observations across space. While, spatial autocorrelation refers to the lack of independence among observations: similar to autocorrelation in time-series models (Anselin 1988).

Spatial autocorrelation among observations and the importance of relative locations is expressed in Tobler's first law of geography, which states that "everything is related to everything else, but near things are more related than distant things". Interactions among neighboring agents could, for example, induce a correlation of the variables across space, which must be accounted for in model estimation Tobler (1979).

Theoretical motivations for the observed dependence between nearby observations are countless. For instance, Ertur and Koch (2007) used a theoretical model that posits physical and human capital externalities as well as technological interdependence between regions. The study showed that this leads to a reduced form growth regression and that an average of growth rates from neighboring regions should be included.

Thomas Plümper and Eric Neumayerb (2010) identified

* Corresponding author:
yemitezoe@gmail.com (Oluyemi A. Okunlola)
Published online at http://journal.sapub.org/statistics

four specification issues in the analysis of spatial data. They argued that to avoid biased estimates of the spatial effects, researchers need to consider carefully how to model temporal dynamics, common trends and common shocks, as well as how to account for spatial clustering and unobserved spatial heterogeneity. Failure to model temporal dynamics and to control for common shocks and common trends in cross-sectional, time-series or panel data is likely to bias the estimated coefficient of the spatial effect variable, with the bias often being upward. In addition, failure to model appropriately spatial patterns in the dependent variable could lead to bias in the spatial effect estimation.

This study investigates the significance of incorporating spatial effect into regression analysis. The purpose is to build a spatial and non spatial model, design a Monte-Carlo experiment and observe the point of convergence in the two models. The specific objective is to observe the bias that could emanate from omitting spatial effect in SAR model when it exists and to examine the relationship of the bias with degree of spatial autocorrelation.

The present study is a contribution to existing work on incorporating locational aspect of sample data into the model. It is different from the existing work in that its discriminate between spatial and non spatial models, observe point of convergence in the two models, investigate bias emanating from omitting spatial effect when its exists. Also, it provides information on association between this bias and degree of spatial autocorrelation through a well-designed Monte Carlo experiment.

The rest of the paper is organised as follows: review of literature, model specification and estimation techniques, Monte Carlo design, result presentation and discussion, and concluding remark.

2. SAR Model Specification and Estimation

Spatial autoregressive (SAR) model with lagged dependent variable is given as:

$$y_i = \beta' X_i + \rho \sum_{j \neq i}^{N} w_{ij}\, y_j + u_i \quad (1)$$

Where:
$\beta = k \times 1$ vector of regression coefficients
$X_i = k \times 1$ vector of explicative variables at site i
$w_{ij}\, y_{ij} = n \times 1$ vector of lagged dependent variable
w_{ij} = the elements of a row-standardized weight matrix
ρ = Spatial effect coefficient and $u \sim N(0, \sigma^2 I)$ error term

Unlike the case of the time series analogous specification, the presence of the spatial lagged term amongst the explicative variables induces a correlation between the error and the lagged variable itself (see Anselin and Bera, 1998). Put differently, we would not want to run OLS on this model, since the presence of y on both the left and right sides means that we have a correlation between errors and the regressors and the resulting estimates will be biased and inconsistent.

Equation 1 can be written in a more compact matrix notation as

$$y = X\beta + \rho W y + u \quad (2)$$

Where:
$y = N \times 1$ *vector of dependent observations*
$X = N \times k$ matrix of observations,
W = N × N matrix of spatial weights and
$u = N \times 1$ vector of independent and identically distributed (iid) disturbances

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \dots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\begin{bmatrix} 0 & w_{12} & \dots & w_{1,n-1} & w_{1n} \\ w_{21} & 0 & \dots & w_{2,n-1} & w_{2n} \\ w_{n-1,1} & w_{n-1,2} & \dots & 0 & w_{n-1,n} \\ w_{n1} & w_{n2} & \dots & w_{n,n-1} & 0 \end{bmatrix}, \quad \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}$$

The reduced form of equation 1 is obtained as follows

$$y(I_N - \rho W) = X\beta + u$$

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} u \quad (3)$$

$$E[y/X] = (I_N - \rho W)^{-1} X\beta \quad (4)$$

Under the hypothesis of normality of the error term, the log-likelihood function of the model 1 is given by:

$$lnL\left(\frac{y}{\beta}, \rho, \sigma^2\right) = -(N/2)\ln(2\pi) - (n/2) ln\sigma^2 + ln|I - \rho W| - (1/2\sigma^2)[(y - \rho W - X\beta)'(y - \rho W - X\beta)] \quad (5)$$

Analytical solution for β *and* σ^2, conditional on ρ and given respectively as:

$$\hat{\beta}_{ML}(\rho) = (X'X)^{-1}X'(I - \rho W)y = (X'X)^{-1}X'y - \rho(X'X)^{-1}X'Wy$$

$$\hat{\beta}_{ML}(\rho) = b_0 - \rho b_L \quad (6)$$

Where $b_0 = (X'X)^{-1}X'y$ *and* $b_L = (X'X)^{-1}X'Wy$

Inspection show that b_0 is the coefficient vector from the OLS regression of y *on* X, while b_L is from OLS regression of Wy *on* X. So if ρ is known, we could compute the ML estimate of β. As a consequence, the residuals of these two OLS regressions are given as:

$$e_0 = y - Xb_0 \quad (7)$$

$$e_L = Wy - Xb_L$$

$$\hat{\sigma}^2{}_{ML}(\rho) = \frac{1}{N}\left[\left(y - \rho W - X\hat{\beta}_{ML}(\rho)\right)\right]' \left[\left(y - \rho W - X\hat{\beta}_{ML}(\rho)\right)\right]$$

$$\left.\begin{aligned}\hat{\sigma}^2{}_{ML}(\rho) &= \tfrac{1}{N}(e_0 - \rho e_L)'(e_0 - \rho e_L)\\ \hat{\sigma}^2{}_{ML}(\rho) &= \tfrac{1}{N}y'(I_N - \rho W)'M(I_N - \rho W)y\end{aligned}\right\} \quad (8)$$

$$M = I - X(X'X)^{-1}X'$$

Expression in 8 can be substituted into 5 to write a version of the log-likelihood function in terms of ρ only. This yields the concentrated log-likelihood lnL^* which is given as:

$$lnL^*\left(\frac{y}{\rho}\right) = -\frac{N}{2}\ln(2\pi) + \ln(I - \rho W) - \frac{N}{2}\ln\left[\frac{(e_0 - \rho e_L)'(e_0 - \rho e_L)}{N}\right] \quad (9)$$

$$= \frac{N}{2}\ln(2\pi) + \sum_{i=1}^{N}(I_N - \rho\omega_i) - \frac{N}{2}ln\left[\frac{1}{N}y'(I_N - \rho W)'M(I_N - \rho W)y\right]$$

Maximizing this is equivalent to minimizing

$$\min_{\{\rho\}}\left\{\frac{y'(I_N - \rho W)'M(I_N - \rho W)y}{|I_N - \rho W|^{2/N}}\right\} \quad (10)$$

This is also equivalent to

$$\min_{\{\rho\}}\left\{\frac{e_0'e_o - 2e_0'e_L + \rho^2 e_L'e_L}{\sum_j(I_N - \rho\omega_i)}\right\} \quad (11)$$

There is need to impose a constraint on the parameter ρ. Anselin and Florax (1994) point out that the parameter ρ can take on feasible values in the range $\left\{\frac{1}{\omega_{min}}, \frac{1}{\omega_{max}}\right\}$. This requires that we constrain our optimization search to values within this range. Note that ω_{min}, ω_{max} are respectively minimum and maximum Eigen value of W.

The estimator $\hat{\rho}$ is then substituted into the solution for β to yield $\hat{\beta}$:

$$\hat{\beta}_{ML} = (X'X)^{-1}X'(I_N - \hat{\rho}W)y$$

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y - \hat{\rho}(X'X)^{-1}X'Wy$$

$$\hat{\beta}_{ML} = b_0 - \hat{\rho}b_L \quad (12)$$

The steps involved in the ML estimation of the model are summarized below as:

i. Perform OLS for the models
$y = X\beta_0 + \varepsilon_o \qquad Wy = X\beta_L + \varepsilon_L$

ii. Compute residuals $e_0 = y - X\hat{\beta}_0 \;\; and \;\; e_L = Wy - X\hat{\beta}_L$

iii. Given $e_0 and\, e_L$, find ρ that maximizes the concentrated likelihood function obtained in 9

iv. Given $\hat{\rho}$ that maximizes the concentrated ML, compute $\hat{\beta} = \hat{\beta}_o - \hat{\rho}\hat{\beta}_L$ and

$$\hat{\sigma}^2{}_{\varepsilon} = \frac{1}{N}(e_0 - \hat{\rho}e_L)'(e_0 - \hat{\rho}e_L) \quad (13)$$

3. Monte Carlo Analysis

In this section, in an attempt to assess the performance of the OLS estimator when spatial autocorrelation is ignored series of Monte Carlo experiments were set up. When spatial autocorrelation consistent with a diffusion process exists in the data generating process and a spatially lagged dependent variable is omitted from the model, OLS parameter estimates for the remaining covariates will be biased and inconsistent. In the Monte Carlos, the Data Generating Process (DGP) for the case of spatial autocorrelation takes the form:

$$y = \rho Wy + \beta_0 + \beta_1 X + u \quad (14)$$

Where $u \sim N(0, \sigma^2 I)$. The independent variable, X, is normally distributed with a mean of 0 and a standard deviation of 3. Also, β_0 and β_1 are set to one (1).

The study examines the bias of the OLS estimates of $\beta 1$ when spatial autocorrelation in the DGP are omitted from the OLS specification. In addition, the study investigate the performance of OLS varying both the number of observations (and the corresponding spatial weights matrices) and the degree of spatial autocorrelation, as reflected in the autoregressive parameters ρ.

For each set of experiments, the observations are arrayed in regular square lattices. Monte Carlos is performed for four different sizes of square lattice structures: a 25×25 $lattice$ $(N = 25), a$ 100×100 $lattice$ $(N = 100)$, a 400 400 $lattice$ $(N = 400), 900 \times 900$ $lattice$ $(N = 900.$ In each case a queen contiguity definition of neighbours is employed. The performance of OLS is examined for five values of ρ : $0, 0.2, 0.5, 0.7, 0.9$. For each combination of lattice size and ρ, 1000 replications were performed.

In order to implement the estimation techniques discussed in the previous section and to observe the distribution of β_1 in the simulation, R statistical software code was developed.

4. Monte Carlo Result

This section of the study presents the result of the Monte Carlo design discussed earlier. To allow for simplicity, the section is divided into two. The first focus on the comparison of OLS and SAR model and contain information on the point of convergence of the two models. The second subdivision furnishes information on the relationship between biases of omitting spatial in SAR and the degree of spatial autocorrelation. These are discussed sequentially below.

4.1. OLS Versus SAR Model

Table 1. OLS versus SAR model

	N=25		N=100	
	OLS	SAR	OLS	SAR
b_1	0.97234	1.000338	0.98654	0.985023
s.e(b_1)	0.10502	0.032889	0.01190	0.011227
$\hat{\sigma}_\varepsilon$	0.521	0.4268	0.3269	0.3082
$\hat{\rho}$		0.2489		0.0917
LM-pvalue		0.0018		0.0009
LR-p-value		0.0059		0.0019
W-p-value		0.0028		0.0012
	N=400		N=900	
	OLS	SAR	OLS	SAR
b_1	0.984737	0.984877	0.994513	0.994832
s.e(b_1)	0.006667	0.0066593	0.003877	0.0038842
$\hat{\sigma}_\varepsilon$	0.4086	0.4074	0.3606	0.3602
$\hat{\rho}$		-0.0103		
LM-pvalue		0.5412		0.5820
LR-pvalue		0.5719		0.6465
W-pvalue		0.5597		0.6414

Source: Author's computation from the simulation

Table 1 shows the result for OLS and SAR model respectively for various numbers of observations considered in the simulation. From the table, (especially when N= 25 and 100) the consequences of taking spatial effect into account are quite clear. The residual standard error of spatial models is much smaller than that of the least squares regression counterpart. In essence, spatial model performs better than the OLS especially where spatial effect parameter is highly significant as indicated by LM, LR and Wald diagnostic test result for spatial dependence.

From table 1, the result of the three approaches considered to examine the presence of spatial effect in the data set was found to have the same results. For N=25 and 100 spatial effect parameter is found to be significant and for N= 400 and 900 the null hypothesis of absence of spatial effect are accepted. The results obtained in these two cases are not significantly different from OLS result. Hence, it is of no use to go for the spatial model when it has been confirmed by a diagnostic test-statistic that spatial effect does not exist in the dataset.

4.2. Relationship between Bias of Omitting Spatial Effect and Degree of Spatial Autocorrelation

This segment of the paper examines the relationship between bias of omitting spatial effect and the degree of spatial autocorrelation. In the simulation, the values of spatial effect parameter (ρ) are fixed, varying the number of observation and the corresponding lattice size replicated 1000 times. In each case, the density of b_1 is shown so as to clearly know how the bias emanating from omitted spatial is related to the degree of spatial autocorrelation.

The finding from this study shows that the density of b_1 is peakest at a high level of the spatial effect parameter. Therefore the bias becomes larger as the spatial effect parameter increases. This implies that the bias resulting from omitting spatial effect in SAR model is a function of degree or magnitude of spatial effect in the data. If the spatial effect parameter estimate is negligible, so also will the bias and if the spatial effect parameter estimate is strong so also the bias. See figures 1-4 for details.

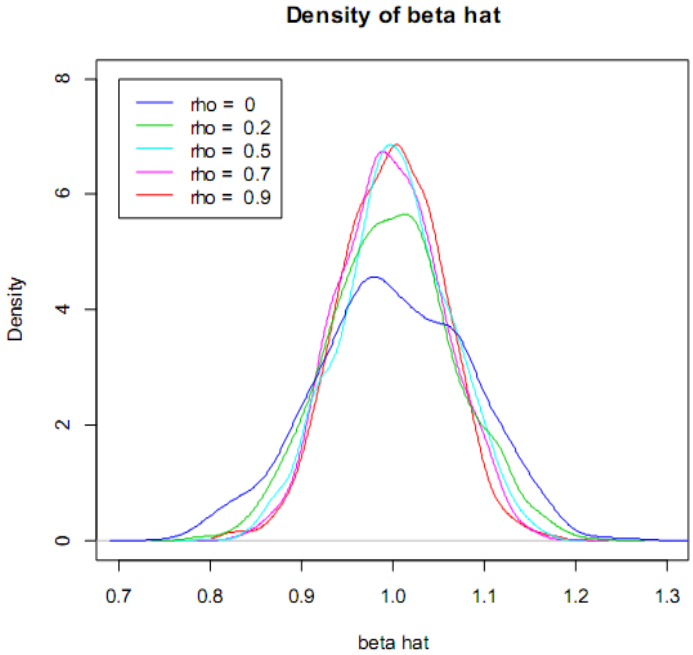

Figure 1. Distributions of b1 for N=25

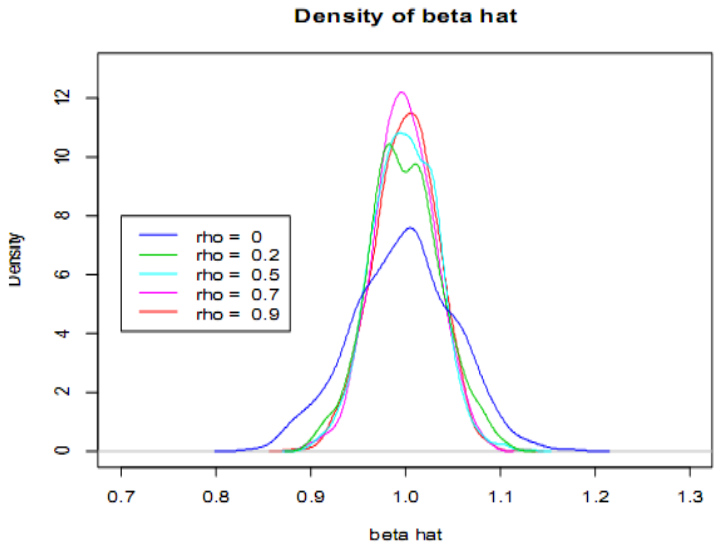

Figure 2. Distributions of b1 for N=100

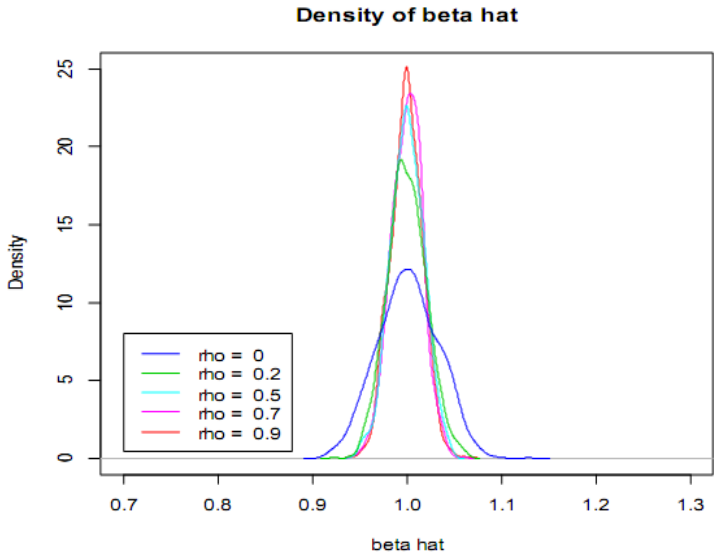

Figure 3. Distributions of b1 for N=400

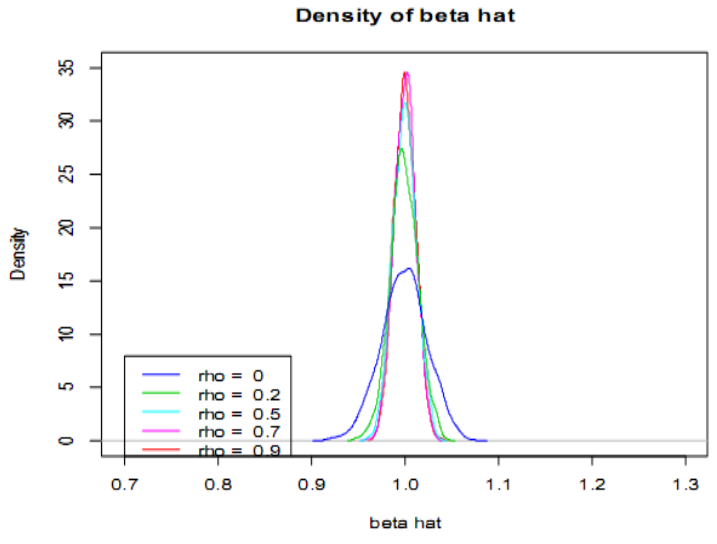

Figure 4. Distributions of b1 for N=900

5. Concluding Remarks

This study shows the significance of using the spatial autoregressive model (SAR) when spatial autocorrelation exist in a data set. Spatial econometrics methods allow us to account for dependence between observations, which often arise when observations are collected from points or regions located in space.

The present study shows that the use of OLS to estimate the SAR model is not appropriate because the spatial lag is correlated with the error term and the included explanatory variable. Thus, OLS will produce biased and inconsistent estimate. Findings show that the spatial model performs better than non-spatial model because a lesser error variance was produced in the spatial model in comparison to the non-spatial model.

The study discovered that spatial autocorrelation exists in cases where N=25 and 1000 and the SAR results were significantly different from their OLS counterpart. The smaller error variance in these cases signifies that the SAR model outperforms the OLS (non-spatial) model. The test-statistic values of LM, LR and Wald diagnostic tests for presence spatial dependence in the dataset are not significant hence the parameter estimate of spatial model are not different from their OLS counterparts. In addition, the study established the existence of relationship between bias resulting from omitting spatial effect in SAR model and the magnitude of spatial autocorrelation.

Thus, the SAR model is recommended when spatial diagnostics show the presence of spatial autocorrelation in a dataset. The use of a spatial model if spatial autocorrelation is absent in a dataset results in waste of time and resources.

REFERENCES

[1] Anselin L. and Rey S., 1991. Properties of tests for spatial dependence in linear regression models, Geographical Analysis, 23: 110-131.

[2] Anselin, L. 2002. Under the hood: Issues in the specification and interpretation of spatial regression models. Agricultural Economics 27, 247267.

[3] Baltagi B.H., 2001. Econometric Analysis of Panel Data, (second edition), John Wiley and Sons, Chichester, England.

[4] Cliff A.D. and Ord J.K., 1981. Spatial Processes: Models and Applications, Pion, London.

[5] Cressie, N. 1993. Statistics for Spatial Data, Revised edition, New York: John Wiley.

[6] Diniz-Filho, J. A. F. et al. 2003. Spatial autocorrelation and red herrings in geographical ecology. Global Ecol. Biogeogr. 12: 53-64.

[7] Dormann, C. F. 2007a. Assessing the validity of autologistic regression. Ecol. Modell. 207: 234-242.

[8] Dormann, C. F. 2007b. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecol. Biogeogr. 16: 129-138.

[9] Dormann, C. F. 2007c. Promising the future. Global change predictions of species distributions. Basic Appl. Ecol. 8: 387-397.

[10] Ertur and Koch 2007. Dual gravity: Using spatial econometrics to control for multilateral resistance, Center for Operations Research and Econometrics (CORE), Universit´e catholique de Louvain, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.

[11] Getis A. and Ord J.K., 1992. The analysis of spatial association by distance statistics, Geographical Analysis, 24: 189-206.

[12] Greene W.H., 2003. Econometric Analysis (fifth edition), New York, Macmillan.

[13] Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in ecology: The flexibility of Eigen function spatial analyses in exploiting relative location information. Ecology 87: 2603-2613.

[14] Gujarati D., 2003. Basic Econometrics, (fourth edition), McGraw-Hill.

[15] Hawkins, B. A. et al. 2007. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. Ecography 30: 375-384.

[16] Jetz, W. et al. 2005. Local and global approaches to spatial data analysis in ecology. Global Ecol. Biogeogr. 17: 97-98.

[17] Johnston J., 1991. Econometric Methods, McGraw Hill, New York.

[18] Kennedy P., 2003. A Guide to Econometrics, (fifth edition), Blackwell Publishers.

[19] Kiihn, I. 2007. Incorporating spatial autocorrelation may invert observed patterns.Div. Distrib. 13: 66-69.

[20] Kissling, W. D. and Carl, G. 2007. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecol. Biogeography.

[21] Kmenta J., 1997. Elements of Econometrics, (second edition), Macmillan, New York.

[22] Lennon, J. J. 2000. Red-shifts and red herrings in geographical ecology. Ecography 23: 101-113.

[23] LeSage, J. P. and K. R. Pace 2009. Introduction to Spatial Econometrics, Boca Raton: CRC Press/Taylor & Francis.

[24] LeSage, J.P 2005, 2010. Spatial econometrics toolbox for Matlab, http://www.spatial-econometrics.com.

[25] Magrini, S. 2004. Regional (di) Convergence in Handbook of Regional and Economics Volume 4 (eds. Vernon Henderson and Jacques-Francois Thisse), North Holland, Amsterdam.

[26] Mankiw N.G., 1995. The growth of nations, Brooking Papers on Economic Activity, 1: 275-326.

[27] Mankiw N.G., Romer D. and Weil D., 1992. A contribution to the empirics of economic growth, Quarterly Journal of

Economics, 107: 407-437.

[28] Pace, R. Kelley and James P. LeSage 2009. Biases of OLS and Spatial Lag Models in the Presence of an Omitted Variable and Spatially Dependent Variables, Progress in Spatial Analysis: Methods and Applications, eds, Antonio Pa'ez, Julie Gallo, Ron N. Buliung, and Sandy Dall'erba. Berlin: Springer-Verlag.

[29] Pace, R. Kelley, James P. LeSage and Shuang Zhu 2010. Spatial Dependence in Regressors and its Effect on Estimator Performance.

[30] Palma, L. et al. 1999. The use of sighting data to analyse Iberian lynx habitat and distribution. J. Appl. Ecol. 36: 812-824.

[31] Robert J. Franzese, Jr. and Jude C. Hays 2007. Spatial Econometric Models of Cross-Sectional Interdependence in Political Science Panel and Time-Series-Cross-Section Data, Oxford University Press.

[32] Thomas Plümpera and Eric Neumayerb 2010. Model Specification in the Analysis of Spatial Dependence.

[33] Tobler W. 1979. "Cellular Geography," Philosophy in Geography, pp. 379-386, Dordrecht.

[34] Reidel.Woolridge J.M., 2002. Econometrics: A Modern Approach (second edition).