ON REGULARIZATION METHODS FOR REGRESSION ANALYSIS IN THE PRESENCE OF SPATIALLY CORRELATED ERRORS: APPLICATION TO HEDONIC REGRESSION OF LAND PRICE

Morito TSUTSUMI
Assistant Professor
Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-8656 Japan
Fax: +81-3-5841-7453
E-mail:tsutsumi@planner.t.u-tokyo.ac.jp

Eihan SHIMIZU
Professor
Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-8656 Japan
Fax: +81-3-5841-7453
E-mail:shimizu@planner.t.u-tokyo.ac.jp

Hiroshi IDE
Graduate Student
Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-8656 Japan
Fax: +81-3-5841-7453
E-mail:ide@planner.t.u-tokyo.ac.jp

Jun-ya FUKUMOTO
Graduate Student
Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-8656 Japan
Fax: +81-3-5841-7453
E-mail:fukumoto@planner.t.u-tokyo.ac.jp

Abstract: Regional econometric models based on cross-sectional data are very useful for a variety of analyses of transportation projects, where the error terms are not completely independent but exhibit a spatial autocorrelation. Therefore, a number of regularization methods have been proposed and designed to improve on the ordinary least squares method in the field of applied statistics. In this paper, we focus on the spatial autocorrelation of error terms in regression model and the regularization methods for estimation of regression coefficients. First, we make a brief review of the suggested methods for spatial autocorrelation. Secondly, we apply them to a simple land price regression model, since hedonic regressions of land price have often been used to estimate various transport projects and thus their application to land price regression is one of the most interesting examples.

1. INTRODUCTION

Regional econometric models based on cross-sectional data are very useful for a variety of analyses of transportation projects, where the error terms are not completely independent but exhibit a spatial autocorrelation. This can be caused by a variety of measurement problems such as arbitrary delineation of spatial units of observation, problems of spatial aggregation, the presence of spatial externalities and spill-over effects (Anselin(1988), Griffith(1996.a)). These are often encountered in applied work and it is widely recognized that the ordinary least squares (OLS) estimator for regression models is unlikely to be a satisfactory estimator in such circumstances. Therefore a number of regularization methods have been proposed and designed to improve on OLS in the field of geography, spatial statistics, spatial econometrics and so on. However, it would be an overstatement to suggest that they have become accepted in regional science, since each approach tends to be rather self-contained, with little useful cross-reference.

In this paper, we focus on the spatial autocorrelation of error terms in regression models and the regularization methods for estimation of regression coefficients.

First, we make a brief review of the suggested regularization methods for spatial autocorrelation. Secondly, we apply them to a common simple regression model. In spite of methodological advances, the importance of their practical applications to regressions in regional analysis cannot be overemphasized. Since hedonic regressions of land values have often been used to estimate various transport projects, their application to land price regression is one of the most interesting examples. Nevertheless, few empirical studies that have employed the hedonic approach have paid attention to the analysis of spatial effects.

The remainder of this paper consists of 4 chapters. In the next chapter, following a brief statement of some problems in spatial statistical analysis, spatial autocorrelation of error terms is described in detail. Then, we make a brief review of the suggested regularization methods for spatially correlated errors in Chapter 3. Chapter 4 presents the parameter estimates and tests for spatial autocorrelation. Some concluding remarks are included in Chapter 5.

2. SPATIALLY CORRELATED ERRORS IN REGRESSION MODELS

2.1 Problems in Spatial Statistical Analysis

Spatial dependence and heterogeneity are two essential aspects of models in regional analysis, especially when cross-sectional data is used in the estimation of the models. These aspects are due to substantive nature, underlying process, misspecification, omission of essential variables, measurement errors and so on (Haining(1990)). Given a non-modifiable area and a regression model, disobedience to the assumption of error terms, that is, spatial dependence and heteroscedasticity are often encountered. Spatial dependence among the disturbances of spatial models is called spatial autocorrelation, which is serial autocorrelation in essence. Heteroscedasticity is a phenomenon where the residuals do not have a common variance.

Another characteristic of spatial analysis is rooted in the need to aggregate geographically referenced data. This aggregation leads to the scale effect and the zoning or aggregation effect, which are called modifiable areal unit problems (MAUP) (Openshaw *et al* (1979), Aibia(1989)). Needless to say that these problems are not independent of each other but are mutually related. Many geographically referenced data contain significant spatial autocorrelation. Consequently, as argued in the literature, spatial dependence is at the core of geographical or spatial analysis.

2.2 Spatially Correlated Errors in Regression Models

Let the standard multiple linear regression model under consideration be

$$y = X\boldsymbol{\beta} + u \tag{1}$$

where **y** is an $n \times 1$ vector of the explained variable; X is an $n \times m$ matrix of non-stochastic variables of rank m ($<n$); $\boldsymbol{\beta}$ is an $m \times 1$ parameter vector; and u is an $n \times 1$ vector of residuals:

$$\begin{array}{l} y = (y_1, \cdots, y_i, \cdots, y_n)^t, \\ \boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_m)^t, \\ u = (u_1, \cdots, u_i, \cdots, u_n)^t, \end{array} \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{pmatrix}. \tag{2}$$

In this paper, the model functionally relates the variations in land prices to structural land attributes; that is, y denotes a vector of observed land price, X denotes a matrix of land attributes, such as size, accessibility, and so on.

The standard assumptions we make about the residuals u in Eq. (1) are as follows.

$$E(u) = 0, \tag{3}$$

$$Var(u) = \sigma^2 I, \tag{4}$$

where I is a unit matrix. However, if the assumption is violated, that is, the residuals u are correlated among themselves (this correlation is called autocorrelation), then the ordinary least squares estimator is inefficient, the estimator of the residual variance is biased and the inference procedures are invalid (Cliff and Ord (1981), Anselin and Griffith(1988)). Spatial autocorrelation is a serious issue in empirical research.

Again consider the standard model below.

$$y = X\boldsymbol{\beta} \tag{5}$$

If we assume that the problem to calculate $\boldsymbol{y}$ by using the input data $\boldsymbol{\beta}$ and $\boldsymbol{X}$ is a direct problem, estimating parameter $\boldsymbol{\beta}$ from the data $\boldsymbol{X}$ and $\boldsymbol{y}$ is an inverse problem. The inverse problem is said to be *correct, correctly posed* or *well-posed* if the following two conditions hold :

(a) for each $\boldsymbol{X}$ and $\boldsymbol{y}$ the equation has a unique solution $\boldsymbol{\beta}$
(b) the solution is stable, *i.e.*, the operator X^{-1} is defined on all of the space which y belongs to and is continuous.

Otherwise, the problem is said to be *incorrectly posed* or *ill-posed* (Tikhonov *et al.*, 1990). As the number of data n is greater than that of parameter m, we are not able to solve this equation and determine $\boldsymbol{\beta}$. Thus, parameter estimation is a typical example of ill-posed problems. Least squares method and maximum likelihood are the ways to have an approximate solution which minimizes the sum of squared residuals or maximizes the likelihood, where under the standard assumption of error terms the unknown variables are the parameter $\boldsymbol{\beta}$ and the standard error σ^2. If the residuals violate the assumption, we should modify it. However, if the elements of variance-covariance matrix of error terms are all unknown, and the number of parameter exceeds the number of equation, then the problem becomes ill-posed. To improve the ill-posedness is called regularization in mathematical sciences.

3. REGULARIZATION METHODS FOR THE SPATIALLY CORRELATED ERRORS IN REGRESSION MODELS, A BRIEF OVERVIEW

A typical regularization approach is to change the solution space introducing some constrained conditions. In this chapter, we make a brief overview of the regularization methods for spatial autocorrelation of regression errors. Consider the situation where the error terms ε_i are correlated among themselves

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6}$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_i, \cdots, \varepsilon_n)^t \tag{7}$$

Among many regularization methods to consider the existence of spatial autocorrelation, the most frequently used one is a mixed regressive-spatial autoregressive model in which the spatial dependence is assumed to be generated by an autoregressive process of explained variables y_i .

$$y = \rho Wy + X\boldsymbol{\beta} + \boldsymbol{u} \tag{8}$$

where $W = \{w_{ij}\}$ is called spatial weight matrix, which denotes the effect of each zone, ρ is a parameter, and $\boldsymbol{u}$ is an error vector that satisfies the assumption (3) and (4). In order to estimate $\boldsymbol{\beta}$ with ρ , it is assumed that $\boldsymbol{u}$ obeys normal distribution $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_u{}^2 \boldsymbol{I})$, so that the maximum likelihood method can be applied.

In order to determine the effects of spatial autocorrelation, we must design the spatial weight matrix **W**. The weighting system is often defined as the functions of the physical distances between zones/points such as,

$$w_{ij} = c_j \big/ d_{ij}^{\alpha} \left(i \neq j\right), \quad w_{ii} = 0 \tag{9}$$

where c_j is a constant which leads to

$$\sum_i w_{\mathrm{ij}} = 1 . \tag{10}$$

An alternative regularization method is to explain the spatial autocorrelation by adding extra variables to the initial model,

$$y = X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{u} . \tag{11}$$

This kind of method is called an expansion method (Casetti (1972)).

Another approach is to model the spatial autocorrelation through the error terms. Analogous to some approaches in time series analysis, some regularization methods have been suggested. Among them, a regression model with autoregressive model of error terms

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \qquad \boldsymbol{\varepsilon} = \lambda W\boldsymbol{\varepsilon} + \boldsymbol{u} \tag{12}$$

and with moving average model

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \qquad \boldsymbol{\varepsilon} = \lambda W\boldsymbol{u} + \boldsymbol{u} \tag{13}$$

are typical.

Kelejian and Robinson (1993) introduced another type of error components formulation to model the spatial autocorrelation among errors. The regression error term is assumed to be the sum of the two parts shown below,

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \qquad \boldsymbol{\varepsilon} = W\boldsymbol{u} + v \tag{14}$$

where

$$v \sim N(\boldsymbol{0} , \sigma_v^2 \boldsymbol{I}), E(\boldsymbol{u}v^t) = \boldsymbol{0} \ . \tag{15}$$

As shown in this section, there have been many regularization methods suggested for spatial dependence. In any case, caution is necessary in interpreting estimated parameters in applied spatial regressions.

4. AN EMPIRICAL ANALYSIS

4.1 Framework for the Empirical Comparison of Regularization Methods

Hedonic price regressions are based on the hypothesis that goods are valued for their utility-bearing attributes or characteristics and have been used to estimate the various transport projects. However, few empirical studies except Can (1992), that have employed the hedonic approach, have paid attention to the analysis of spatial effects. Some applications associated with land prices are found, such as in the papers of Takatsuka *et al* (1996.a, b) and Benirschaka and Binkley (1994), but they focus on the mechanism or formula of land price determination. Many of other applications treat the social science phenomena such as crime or disease (Hainig(1990), Getis, A.(1995)) but have little relationship to regional science. Thus, the importance of comparative study on the applications of regularization methods to hedonic land price regression cannot be overemphasized.

We simulate and compare the results by the regularization methods for spatially correlated errors. Figure 1 shows the study area, which is a part of Adachi Ward in the Tokyo Metropolis, located at about 15 to 35 minutes by Johban and Chiyoda Lines from Ueno Station in Tokyo. For estimation of the function, officially assessed land price data set by the National Land Agency in 1997 is used. Suppose that we formulate the land price function as

$$y_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ji} \tag{16}$$

where

i : point number (i =1,⋯,52)
j : category number of explanatory variables (j =1,⋯,4)
y : officially assessed land price [yen / m^2]
x_1 : lot area [m^2]
x_2 : distance to nearest station [m]
x_3 : time distance to Kita-Senju (terminal) station [minute]
x_4 : floor-area ratio based on legislation [%]

and β_0, β_j (j=1,⋯,4) are unknown parameters.

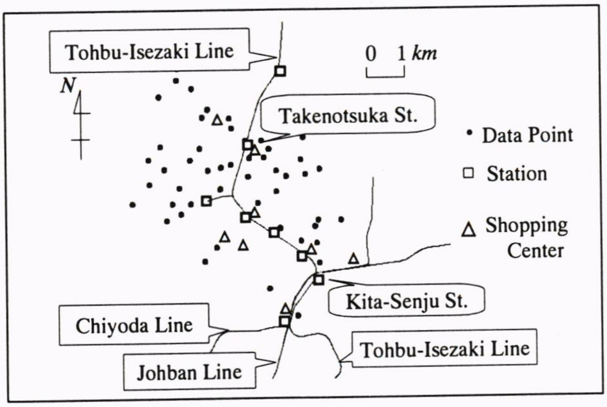

Figure 1. Study Area

One standard regression model and five alternative models for regularization shown below are applied to the same data set .

Model 1	$y = X\boldsymbol{\beta} + u$		
Model 2	$y = \rho Wy + X\boldsymbol{\beta} + u$		(1)'
Model 3	$y = X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + u$		(7)'
Model 4	$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,	$\boldsymbol{\varepsilon} = \lambda W\boldsymbol{\varepsilon} + u$	(10)'
Model 5	$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,	$\boldsymbol{\varepsilon} = \lambda Wu + u$	(11)'
Model 6	$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,	$\boldsymbol{\varepsilon} = Wu + v$	(12)'

where

$$u \sim N(0, \sigma_u^2 I), v \sim N(0, \sigma_v^2 I), E(uv^t) = 0 \ . \tag{17}$$

4.2 Detection of Spatial Correlation in the Residuals of Regression Model

The estimation results of Model 1 by OLS are shown in Table 1. The correlation coefficient is 0.890. Incidentally, there is no multicollinearity in this formulation. Figure 2 illustrates the residuals in Model 1. It is implied that the residuals are spatially correlated among themselves.

There are many approaches to assessing spatial autocorrelation. One statistic that is often used is the Moran's statistic (Moran (1950)) defined as

Table 1. Estimation Result of model 1

parameter	coefficinet	t-value	standard deviation
β_0	3.12×10^5	18.5	1.69×10^4
β_1	2.03×10^2	2.79	72.8
β_2	-25.8	-6.27	4.12
β_3	-2.51×10^3	-2.09	1.20×10^3
β_4	2.31×10^2	5.16	44.7

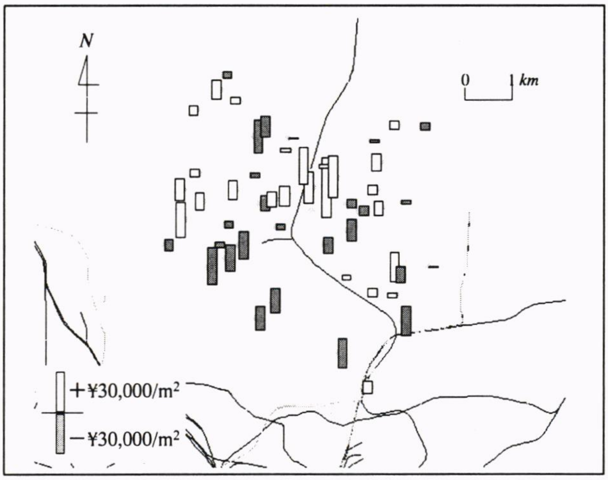

Figure 2. Visualization of Spatial Autocorrelation

$$I = \frac{n}{s}\frac{e^t We}{e^t e} \tag{18}$$

where e is a vector of OLS residuals,

$$s = \sum_j \sum_i w_{ij} \,. \tag{19}$$

Under the standardization by Eq.(9), we have

$$n = s \,. \tag{20}$$

Moran's statistic will be employed in this study because of its relative simplicity compared to other statistics such as Lagrange multiplier statistic. The generalization of Moran's statistic shown by Cliff and Ord (1973) is formulated as

$$Z = \frac{I - E[I]}{\sqrt{Var[I]}} \tag{21}$$

$$E[I] = \frac{1}{n-k} tr(\boldsymbol{MW}) \tag{22}$$

$$Var[I] = \frac{\{tr(\boldsymbol{MWMW}^t) + tr(\boldsymbol{MW})^2 + [tr(\boldsymbol{MW})]^2\}}{(n-k)(n-k+2) - \{E[I]\}^2} \tag{23}$$

where

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t \,. \tag{24}$$

The distribution of the standardized Moran's statistic is shown to be asymptotically normal. However, it should be noted that Moran's statistic cannot test the significance of spatial autoregressive coefficient ρ or λ directly (Anselin and Rey (1991)). Tests for the presence of spatial autocorrelation by Moran's statistic are carried out, in which a null hypothesis of no spatial dependence is tested against the hypothesis of dependence as reflected in particular structure.

Table 2 shows the result of Moran's tests for the residuals of Model 1. As the values of Moran's statistic depend on the assumed structure of $\boldsymbol{W}$, we tested with five types of parameter α in Eq.(9). In all cases, a null hypothesis of no spatial dependence is rejected at the significance level of 1 %.

Table 2. Detection of Spatial Autocorrelation in Model 1

Parameter α	0.5	1	2	5	∞
Moran's I	0.013	0.062	0.195	0.433	0.428
Z	5.21	4.77	4.01	3.74	2.85
Probabilities for normal distribution to exceed the value of Z	0	0	6.0×10^{-5}	1.9×10^{-4}	4.4×10^{-3}

4.3 Comparison among the Applications of Regularization Methods

As the parameter for spatial weight matrix $\boldsymbol{W}$, α =2 is often adopted analogous to gravity law. Thus, from now on, we will continue to analyze under the condition that α =2. A summary overview of the results of Models 1 to 6 is given in Table 3. All the parameters were estimated by maximum likelihood.

Table 3. Comparison among the Results of the Models

	Model					
	1	2	3	4	5	6
Correlation Coefficient	0.890	0.86	0.917	0.843	0.857	0.876
AIC	1152	1149	1141	1146	1148	1149
ρ	—	0.48	—	—	—	—
λ	—	—	—	0.68	0.56	—
$\beta_0\ (\times10^5)$	3.12	1.45	2.24	3.17	3.21	3.03
β_1	203	167	234	122	136	136
β_2	-25.8	-18.5	-24.8	-28.2	-28.2	-26.1
$\beta_3\ (\times10^3)$	-2.51	-1.76	-0.040	-1.49	-1.80	-1.04
β_4	231	218	236	229	224	261
γ_1			909			
γ_2			32.2			

It is not easy to compare Model 1 with Model 2 or 3, for the structures of the latter are somewhat modified. However, Models 4, 5 and 6 are models where only the structure of error terms is modified. With respect to the parameter β_3 directly associated with evaluation of transport projects, it should be remarked that its value in model 6 is less than half of that in model 1. Hedonic regressions of property values, especially land prices, have been used to estimate the benefits of various infrastructure projects. However, compared with other statistical problems such as multicollinearity of explainable variables, correlation and heterogeneity of the errors are largely ignored in the applications. Our empirical results imply that lack of attention to spatial autocorrelation may lead to serious mistakes in project evaluation. In addition, the regression coefficients, which affect the estimation of infrastructure projects directly, depended on the Model we choose and their differences cannot be ignored in measuring the impact of infrastructure projects.

Incidentally, Table 4 shows the standardized Moran's statistics for the residual ***u*** in Models 1 to 5. As there are two dependent error terms in Model 6, Moran's test is not employed for it. Null hypothesis of no spatial dependence is not rejected at the significance level of 5% in Models 2, 4, 5 while it is rejected in Model 3. In this case, Model 3 is not effective in getting rid of spatial autocorrelation of the residuals.

Table 4. Detection of Spatial Autocorrelation in Model 1 to 5

Model	1	2	3	4	5
Z	4.01	2.04	3.00	0.79	1.42
Probability for normal distribution to exceed the value of Z	0	0.041	0.003	0.428	0.157

5. CONCLUSIONS

Regional econometric models based on cross-sectional data have been indispensable in regional and transport analysis. However, not enough attention had been paid to its estimation in empirical applications except to the multicollinearity of explanatory variables. Spatial correlation and heteroscedasticity of error terms that violate the independence assumption on which the statistical analysis is based are often encountered and affect the parameter estimation, but little attention has been paid to them.

In this paper, we have focused on the spatially correlated errors in regression models. A brief overview of regularization methods for spatial autocorrelation of regression residuals was made. It was shown that there have been many regularization methods suggested for spatial dependence. Then, some empirical results associated with hedonic regression of land prices were presented, since despite the increasing use of such methods, there has been relatively little emprical analysis of land price regressions. Several regularization methods were applied in the presence of spatially correlated errors. Our empirical results implied that lack of attention to spatial autocorrelation might lead to serious mistakes in project evaluation. And it was demonstrated that regression coefficients, which affect the estimation of infrastructure projects directly, depended on the regularization method that we chose and their differences could not be ignored in measuring the impact of infrastructure projects. We should note that several alternatives which are effective in getting rid of spatial autocorrelation of the residuals may lead to quite different results in project evaluation.

For future reference it should be noted that specification of the spatial weight matrix is also an important issue since the misspecification causes serious mistakes in parameter estimation (Griffith(1996.b)).

We believe that our empirical analysis provides some useful implications for spatially correlated errors in hedonic regression models.

ACKNOWLEDGEMENTS

This work was partially supported by the National Science Foundation under Grants 09875121 and 10750399 from the Japanese Ministry of Education.

REFERENCES

Anselin, L.(1988) **Spatial Econometrics: Methods and Models.** Kluwer Academic, Dordrecht.

Anselin, L. and Griffith, D.A.(1988) Do spatial effects really matter in regression analysis?. **Papers of the Regional Science Association 65**, 11-34.

Anselin, L. and Rey, S.(1991) Properties of tests for spatial dependence in linear regression models. **Geographical Analysis 23,** 112-131.

Arbia, G. (1989) **Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems**. Kluwer Academic, Dordrecht.

Benirschaka, M. and Binkley, J. K.(1994) Land price volatility in a geographically dispersed market. **American Journal of Agricultural Economics 76**, 185-195.

Can, A. （1992） Specific and estimation of hedonic housing price models. **Regional Science and Urban Economics 22**, 453-474.

Casetti, E. （1972） Generating model by the expansion method: applications to geographical research. **Geographical Analysis 4**, 81-91.

Cliff, A.D. and Ord, J.K. (1973) **Spatial Autocorrelation.** Pion, London.

Cliff, A.D. and Ord, J.K. (1981) **Spatial Processes: Models and Applications.** Pion, London.

Getis, A.(1995) Spatial filtering in a regression framework : examples using data on urban crime, regional inequality, and government expenditures, In Anselin, L. and Florax, R.J.G.M.(eds.), **New Directions in Spatial Econometrics**. Springer, Heidelberg.

Griffith, D.A.(1996.a) The Need for Spatial Statistics. In Arlinghaus, S. L. (ed.), **Practical Handbook of Spatial Statistics.** CRC Press, Boca Raton.

Griffith, D.A.(1996.b) Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In Arlinghaus, S. L. (ed.), **Practical Handbook of Spatial Statistics.** CRC Press, Boca Raton.

Haining, R.(1990) **Spatial Data Analysis in the Social and Environmental Sciences.** Cambridge University Press, Cambridge.

Kelejian, H. and Robinson, D.(1993) A Suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. **Papers in Regional Science 72**, 297-312.

Moran, P. A. P. (1950) Note on continuous stochastic phenomena. **Biometrika 37**, 17-23

Openshaw, S. and Taylor, P.J.(1979) A million or so correlated coefficients: three experiment on the modifiable areal unit problem. In Wrigley, N. and Bennett, R.J. (eds.), **Statistical Applications in the Spatial Sciences.** Pion, London.

Takatsuka, H. and Higuchi, Y.(1996.a) A Statistical Study of Spatial Relationship of Land Prices with Spatial Autocorrelation Analysis. **Studies in Regional Science 26,** 139-153 (in Japanese).

Takatsuka, H. and Higuchi, Y.(1996.b) Land price models considering spatial dependence of expectation: specification and estimation methods. **Journal of Applied Regional Science 2,** 53-63 (in Japanese).

Tikhonov, A. N., Goncharsky, A. V., Stepanov, V. V. and Yagola, A. G. (1990) **Numerical Methods for the Solution of Ill-Posed Problems.** Kluwer Academic, Dordrecht.