Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 32, No. 3, 225-231, 2014
http://dx.doi.org/10.7848/ksgpc.2014.32.3.225

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

Selection of Spatial Regression Model Using Point Pattern Analysis

Shin, Hyun Su[1] · Lee, Sang-Kyeong[2] · Lee, Byoungkil[3]

Abstract

When a spatial regression model that uses kernel density values as a dependent variable is applied to retail business data, a unique model cannot be selected because kernel density values change following kernel bandwidths. To overcome this problem, this paper suggests how to use the point pattern analysis, especially the L-index to select a unique spatial regression model. In this study, kernel density values of retail business are computed by the bandwidth, the distance of the maximum L-index and used as the dependent variable of spatial regression model. To test this procedure, we apply it to meeting room business data in Seoul, Korea. As a result, a spatial error model (SEM) is selected between two popular spatial regression models, a spatial lag model and a spatial error model. Also, a unique SEM based on the real distribution of retail business is selected. We confirm that there is a trade-off between the goodness of fit of the SEM and the real distribution of meeting room business over the bandwidth of maximum L-index.

Keywords : Spatial Regression Model, Spatial Error Model, Kernel Density, L-index, Meeting Room Business

1. Introduction

Retail businesses, which are located on urban space, can be presented by geographical points defined in the XY coordinates. In this case, a regression model which uses point density values as a dependent variable can be introduced in order to analyse location determinants of retail businesses (Kloog *et al*., 2009). The point quadrat analysis and the kernel density analysis are generally regarded as typical methods to estimate point density values. The point quadrat analysis divides a targeted space into the same size by a grid and calculates the number of points in a grid, while the kernel density analysis is a data smoothing technique which transforms a sample of point data into a continuous surface, indicating the intensity of individual point (Bailey and Gatrell, 1995). The point quadrat analysis has been used to estimate the density of point features, however, after introducing GIS, the kernel density analysis is being widely used (Lee, 2008).

The kernel density analysis uses a distance decay function for estimating the distance from grid points as weighting and therefore kernel density values of points is affected by point distribution being able to cause the spatial autocorrelation. The spatial autocorrelation means the co-variation of observations within geographical space. In other words, high or low values of an attribute tend to cluster within the positive spatial autocorrelation, unlikely to be surrounded by neighbours with different values in negative spatial autocorrelation (Chi and Zhu, 2008). When the spatial autocorrelation occurs at the dependent variable, the ordinary least squares (OLS) regression causes violations of a basic assumption in error terms: normality, homoscedasticity and no spatial autocorrelation (Lee and Sim, 2011). In this case, a spatial regression model that is able to consider spatial autocorrelation should replace the OLS regression model (Jin *et al*., 2012). Since kernel density values of the dependent

Received 2014. 05. 07, Revised 2014. 06. 02, Accepted 2014. 06. 25

[1] Department of Urban Planning, Gachon University, Republic of Korea
[2] Corresponding Author, Member, Department of Urban Planning, Gachon University, Republic of Korea (E-mail: skylee@gachon.ac.kr)
[3] Member, Department of Civil Engineering, Kyonggi University, Republic of Korea (E-mail: basil@kgu.ac.kr)

variable follow changing in kernel bandwidths, spatial regression analysis brings about model selection problem. Despite of this problem, previous studies have undergone in the selection of a spatial regression model without clear criteria (Jin *et al*., 2012). Since the model selection problem is caused by changing in kernel bandwidths, it is important to select kernel bandwidth properly. In fact, the significance of kernel bandwidth selection during the density analysis has been confirmed in previous studies. Diggle (1985) suggested that the kernel bandwidth selection was more important than the kernel function selection. Brunsdon (1995) argued that the kernel bandwidth selection is the most important procedure and proposed the adaptive kernel algorithm. Hwang (2004), Goodwin and Unwin (2000) and Borruso and Schoier (2004) suggested the use of the L-index or 300-500 m as the kernel bandwidth in urban studies. In particular, the L-index is pretty effective in analysing point distribution pattern and therefore has received attention along with development of GIS (Lee and Lee, 2013). It has been known that point distribution pattern is clearly shown at the kernel bandwidth, which is the distance of maximum L-index (Diggle, 1985; Hwang, 2004).

This paper aims to suggest a procedure which uses the point pattern analysis, especially L-index in order to select a unique spatial regression model. To test the procedure, we apply the L-index to meeting room business data in Seoul, find the distance of the maximum L-index, measure kernel density values of points by using the distance as the bandwidth, and input the density values as the dependent variable of spatial regression models.

2. Methodology

2.1 Spatial regression model

The OLS regression model assumes the error terms are independently, identically, and normally distributed. If the OLS regression model using kernel density values as a dependent variable can be defined as:

$$Y = X\beta + \varepsilon \qquad (1)$$

where Y is the vector of the dependent variable, the kernel density values of point features, X is the matrix of independent variables, β is the vector of parameters, and ε refers the vector of normally distributed error terms with the mean of zero and the constant variance of σ^2. Since a dependent variable is likely to show spatial autocorrelation, if Eq. (1) is estimated by the OLS regression, spatial autocorrelation can be occurred in error terms. Spatial autocorrelation can be overcome by a spatial regression model. There are two kinds of spatial regression model commonly. One is a spatial lag model (SLM) and a spatial error model (SEM). A general equation for the SLM is specified as:

$$Y = \rho WY + X\beta + \varepsilon \qquad (2)$$

where W denotes a spatial weight matrix, and ρ denotes the spatial autoregressive coefficient. For the SLM, spatial autocorrelation is modelled by a linear relation between the dependent variable Y and the associated spatially lagged variable WY (Chi and Zhu, 2008). In contrast, the SEM is specified as:

$$\begin{aligned} Y &= X\beta + u \\ u &= \lambda Wu + \varepsilon \end{aligned} \qquad (3)$$

where $\boldsymbol{\lambda}$ is the autoregressive coefficient of residuals. For the SEM, the spatial autocorrelation is modelled by an error term u and the associated spatially lagged error term Wu (Chi and Zhu, 2008). In this study, we select the method making a matrix based on spatial distance because a point feature does not have a polygon-shaped spatial boundary and only has a position. The inverse number of a distance between point features is used for actual analysis.Using spatial regression model, instead of the OLS regression, requires a test for the OLS regression errors. Non-normality, heteroscedasticity and spatial autocorrelation have to be tested. In general, non-normality is tested by Jarque-Berra statistic and heteroscedasticity by Breusch-Pagan statistic. For the spatial autocorrelation, Lagrange multiplier (LM) test is used with following null hypothesis: there is not spatial autocorrelation in dependent variables or errors. The LM-Lag static and the LM-Error statistic are used for the LM test and the model selection criteria, when the optimal model is selected between SLM and SEM. If the LM-Lag statistic is only significant, the

SLM is selected and if the LM-Error statistic is only significant, the SEM is selected (Lee and Sim, 2011).

2.2 Kernel density analysis

The point pattern is defined as a series of locations (s_1,s_2,⋯) where s_i is a vector coordinate of the *i-th* event in a specific space, *R*. An event is a standard term used for the point process to distinguish the observed location from a random location in R (Diggle, 1983). The simplest statistical model of point patterns on space is complete spatial randomness (CSR). The CSR means that an event is independently distributed, based on the same probability distribution in the target space *R*. In general, a density analysis is used to investigate the presence or absence of point pattern. The quadrat analysis and the kernel density analysis are popular in measuring density. The modifiable area unit problem (MAUP) occurs in the quadrat analysis whereas it is eased in kernel density analysis. Thus, the kernel density analysis is recently used more often. A technique estimating the kernel density is called kernel density estimation, which has a general form of the kernel estimator in Eq. (4):

$$\hat{\lambda}(s) = \sum_{i=1}^{n}\frac{1}{\tau^2}\kappa\left(\frac{s-s_i}{\tau}\right) \tag{4}$$

where $\hat{\lambda}(s)$ is the density estimator of the intensity of the spatial point pattern measurement in locations of *s*, while s_i is the *i-th* event observed. κ() is the kernel density function and τ is the bandwidth (Borruso and Schoier, 2004).

A density estimate of Eq. (4) is sensitive to the kernel bandwidth. As the bandwidth is growing, spatial variation of the density is too smoothed to show details of spatial density distribution. As the bandwidth becomes smaller, spiky results are produced, which is hard to find out a trend of changes in the density on space.

2.3 Kernel bandwidth selection

Diggle(1985) suggested that the K-index determined by secondary characteristic analysis of point features should be used for a proper kernel bandwidth. The K-index is computed from Eq. (5) and decides whether point distribution is random with comparing the number of points existed in a certain distance from a specific point and the number of points theoretically expected.

$$\widehat{K}(\tau) = \frac{R}{n^2}\sum\sum\frac{I_h(d_{ij})}{w_{ij}} = \frac{1}{\lambda^2 R}\sum\sum\frac{I_h(d_{ij})}{w_{ij}} \tag{5}$$

where *R* is the area of a target space, *n* is the number of events, $\lambda(=\frac{n}{R})$ is the average density of events, d_{ij} is a distance between s_i and s_j, I_h () is an indicator function (1 in case of $d_{ij} < \tau$, and 0 for the rest cases), and w_{ij} is a weighting to remove a boundary line effect; locations within the distance of τ from boundary line are supposed to have smaller events than expected value.

In a random pattern, the probability of points existing in all locations is same and independent. Therefore, $\lambda\pi\tau^2$ is the number of average points expected to be found in a certain distance, from a specific point. In other words, $K(\tau) = \pi\tau^2$ means the isotropic condition without spatial interaction and $K(\tau) > \pi\tau^2$ means cluster distribution and $K(\tau) < \pi\tau^2$ means regular distribution. Therefore, we can find a distance, where a cluster appears, by drawing a graph of the K-index. However, the L-index which is square root of the K-index, is practically used instead of K-index because the graph of K(τ) is typically increasing exponentially. L(τ) is defined as:

$$L(\tau) = \sqrt{\frac{K(\tau)}{\pi}} - \tau \tag{6}$$

Eq. (6) which originally suggested by Ripley (1976), redeemed by Cressie (1991). The L-index on CSR distribution has the merit of coinciding with the x-axis of a graph due to a deducted distance. Diggle (1985) and Hwang (2004) reported that a distance of the maximum point of L(τ) is suitable for the bandwidth of kernel density.

3. Application and Analysis

3.1 Analysis data

Meeting room business in Seoul is used as the analysis data, which are composed of 110 observations (Fig. 1). Most businesses were found using the internet NAVER Map and DAUM Map from May 6 to 10, 2013. Both physical and locational characteristics data of meeting room business was compiled by visiting sites, investigating building resisters, collecting data of KOSIS, the small enterprise development agency and Seoul institute, and using ArcGIS 10.0.

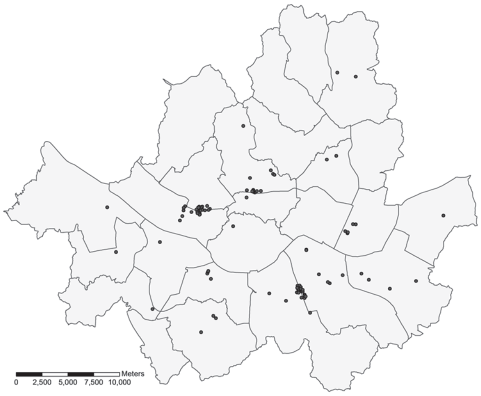


Fig.1. Meeting room business in Seoul, Korea

3.2 Selection of kernel bandwidth using the L-index

The kernel density values of meeting room business was measured by bandwidth increasing from 100 m to 1500 m and then, Moran's I was measured to find the spatial autocorrelation. Moran's I statistic summarizes the global clustering of similar values into one index and the range is -1 to +1. As it is getting close to +1, positive spatial autocorrelation is shown, but as it is getting close to -1, negative spatial autocorrelation is shown. In general, a case of >+0.3 or <-0.3 is decided to show relatively strong autocorrelation (Yim and Lee, 2012). As shown in Fig. 2, Moran's I increases proportional to increasing of kernel bandwidth. Especially, kernel bandwidths from 600 m to 1500 m show strong spatial autocorrelation. The goodness of fit of the spatial regression model usually tends to increase as global spatial autocorrelation grows. Therefore, if the kernel bandwidth is selected only based on goodness of fit, the distance of maximum point, 1500 m has to be selected in this interval. However, this selection is meaningless because there can be larger values anytime.

Crimestat III specialized in point process analysis is used to measure the L-index. Fig. 3 shows a graph of L(τ) against the distance of meeting room business. As can be seen, L(τ) increases up to a distance of 924.6 m whereupon it decreases again. It means that meeting room business shows the strongest clustering at a distance of 924.6 m.

The kernel density is measured at a bandwidth of 924.6 m and the result is shown in Fig. 5 (a). The kernel density is higher at the downtown Gangnam and Sinchon. Moran's I statistic is 0.3491 and therefore strong spatial autocorrelation is confirmed.

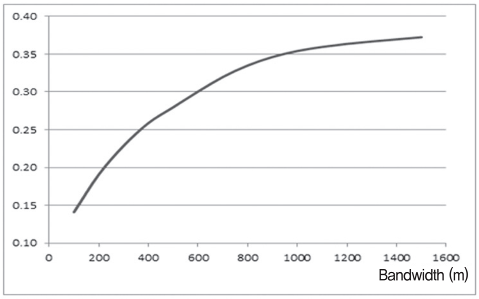


Fig. 2. Moran's I statistic against kernel bandwidth

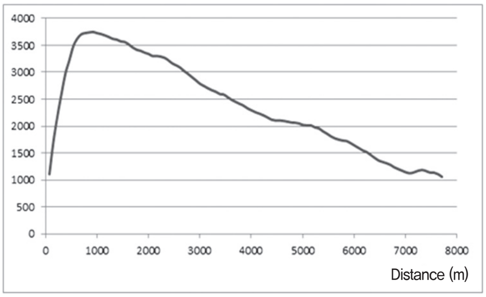


Fig. 3. L-index change against distance

3.3 Spatial autocorrelation for OLS regression errors

The dependent variable is the kernel density of meeting room business which is measured at a bandwidth of 924.6 m, and independent variables for regression analysis are summarized in the Table 1. Both non-normality and heteroscedasticity of the OLS regression errors have to be identified to apply a spatial regression model. As shown in Table 2, the result of Jarque-Bera test does not show the normal distribution of errors at the 1% significance level. In addition, the result of Breusch-Pagan test shows the heteroscedasticity of errors at the 1% significance level. The LM-Lag test and The LM-Error test have to be performed to select one among spatial regression models, SLM and SEM. The p-value of the LM-Lag test is 0.7649, and therefore the result is not significant whereas the p-value of LM-Error

Table 1. Basic statistics of independent variables of OLS and SEM

Independent variables	Variable details	Mean	Std. Dev.
Gangnam station distance	A distance from Gangnam station	6479.4	4767.6
Jonggak station distance	A distance from Jonggak station	6831.9	3411.1
Sinchon station distance	A distance from Sinchon station	7049.3	4785.9
Commercial facility density	Commercial facility floor area of dong / dong area	0.5	0.6
Office building density	Office building floor area of dong / dong area	0.3	0.5
Private institute density	The number of private educational institutes within 300 m search-radius	17.1	14.5
Subway distance	A distance from a close subway station	335.9	401.5
Frontal road width	The width of a frontal road	15.7	12.9

test is 0.049, and therefore the result is significant at the 5% significance level. The SEM is selected as a final model because the spatial autocorrelation is only confirmed in the LM-Error test.

Table 2. Spatial autocorrelation test for OLS errors

Test	Values	Probability
Jarque-Bera Test	10.4212	0.0055
Breusch-Pagan Test	22.1725	0.0046
LM-Lag Test	0.0894	0.7649
LM-Error Test	3.8760	0.0490

3.4 SEM analysis result

The result of the OLS and the SEM application is presented in Table 3. A spatial regression coefficient, λ of the SEM is strong positive and significant and therefore, spatial autocorrelation is confirmed. Pseudo R² instead of R² is computed because the SEM is estimated by Maxim Likelihood (ML) estimation. Therefore, it is impossible to directly compare the goodness of fit of the OLS and the SEM by using R². Instead, Log-likelihood, Akaike Information Criterion (AIC), and Schwartz Criterion (SC) are used to compare two models. In general, improvement of model goodness of fit is decided when Log-likelihood increases and AIC and SC decrease (Anselin, 2005). By comparing Log-likelihood, AIC and SC, it was confirmed that goodness of fit has been improved by using the SEM.

Table 3. Comparing SEM and OLS estimation

Variables	OLS		SEM	
	Coefficient	t-value	Coefficient	t-value
Intercept	16.4716	7.46***	24.1498	3.46***
Gangnam station distance	-0.0008	-4.40***	-0.0002	-0.44
Jonggak station distance	0.0003	1.08	0.0005	1.42
Sinchon station distance	-0.0012	-6.11***	-0.0026	-9.00***
Commercial facility density	3.2349	2.83***	4.7046	4.73***
Office building density	3.4548	2.40**	4.7584	4.10***
Private institute density	0.2464	5.39***	0.2325	5.95***
Subway distance	-0.0003	-0.24	-0.0001	-0.12
Frontal road width	-0.0727	-1.65	-0.0447	-1.21
λ			0.915***	
R²	0.671		0.750	
AIC	698.473		680.205	
SC	722.777		704.510	
Log-likelihood	-340.236		-331.103	

Notes: *** and ** are significant at $p \leq 0.01$ and $p \leq 0.05$, respectively.

Coefficients of significant independent variables of the SEM can be explained as follows. Gangnam station distance and Sinchon station distance have negative effects on the kernel density of meeting room businesses. On the contrary, commercial facility density, office building density, and private institute density have positive effects. In other words, kernel density values of meeting room business are in inverse proportion to Gangnam station distance and Sinchon station distance, but proportionally increase with growing up in the commercial facility floor area ratio of a dong of the lowest administrative unit in Korea, the office building floor area ratio of a dong, and the number of private educational institutes. The coefficient of subway distance variable and frontal road width variable are not significant. This means that the distance of subway station and the width of a frontal road have no effects on the kernel density of meeting room businesses.

3.5 Evaluation of the methodology

Changes of R^2 are analysed by applying SEM to intervals more than 600m where spatial autocorrelation is observed by Moran's I. The result is shown in Fig. 4. Like Moran's I in Fig. 2, R^2 increases as a bandwidth grows. A bandwidth of maximum R^2 cannot be identified because R^2 is monotonically increasing. The distance of maximum L-index, 924.6 m, is not the point where R^2 is the highest, as shown in Fig. 4, however, this bandwidth reveals the clustering pattern most clearly. By comparing Fig. 3 and Fig. 4, we can notice that there is a trade-off between goodness of fit of the SEM and the real distribution pattern of meeting room business over a distance with the maximum L-index. The goodness of fit of the SEM increases proportionally to the growth of kernel bandwidths whereas explanation power on distribution pattern of meeting room business decreases.

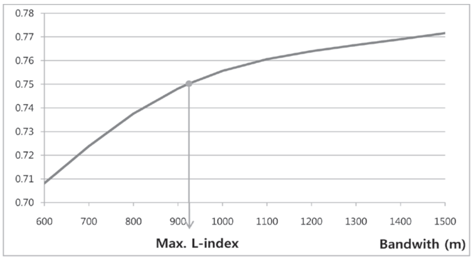
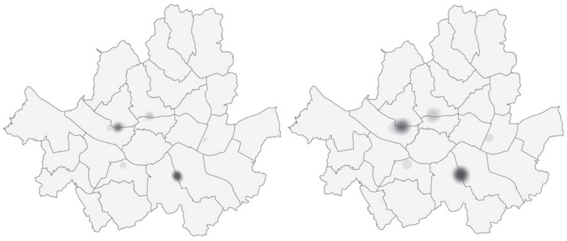


Fig. 4. Changes of SEM R^2 against kernel bandwidth



(a) **(b)**

Fig. 5. Kernel Density (a) bandwidth 924.6m (b) bandwidth 1500m

On the contrary, both the L-index and the goodness of fit of the SEM increase in the smaller bandwidth ranges. This informs us that the kernel density of bandwidth 1500 m of Fig. 5 (b) covers the larger area than bandwidth 924.6 m of Fig. 5 (a), but its clustering is weaker.

4. Conclusion

When a spatial regression model has a dependent variable using kernel density values, we cannot select a unique model because kernel density values of a dependent variable change as kernel bandwidth changes. For this reason, this paper suggests a procedure using the L-index for selecting a unique spatial regression model. To test the procedure, we apply spatial regression model to meeting room business data in Seoul. As a result, the distance of maximum L-index is found and used to measure kernel density as the bandwidth. The measured density values are used as a dependent variable of spatial regression models. By testing the spatial autocorrelation, the spatial error model is turned out to be more suitable than the spatial lag model. The selected SEM is not a model with the best goodness of fit but with the best real distribution pattern.

This study have meaning in that the proposed procedure can contribute to finding location determinants of retail businesses which have only point features without other information. In this study, meeting room business is established as point feature data, nonetheless, we can find that the density of meeting room businesses is in inverse proportion to Gangnam station distance and the Sinchon station distance, but proportionally increases with growing up in commercial facility density, office building density, or private institute density. We expect that the proposed procedure will be applied to other commercial businesses.

This study has a limitation in that we test the procedure for meeting room business with small observations. Though these data include total meeting room business in Seoul, it is not sufficient to evaluate the methodology. A follow-up study, which analyses large sample data such as coffee shops or chain stores, should be needed.

References

Anselin, L. (2005), *Exploring Spatial Data with GeoDa: a*

Workbook, Center for Spatially Integrated Social Science

Bailey, T. C., and Gatrell, A. C. (1995), *Interactive Spatial Data Analysis*, Longman, London

Borruso, G. and Schoier, G. (2004), Density analysis on large geographical databases, Search for an Index of Centrality of Services at Urban Scale*, ICCSA 2004, LNCS 3044*, pp. 1009-1015

Brunsdon, C. (1995), Estimating probability surfaces for geographical point data: an adaptive kernel algorithm, *Computers and Geosciences*, Vol. 21, No. 7, pp. 877-894

Chi, G. and Zhu, J. (2008), Spatial regression models for demographic analysis, *Population Research and Policy Review*, Vol. 27, pp. 17-42

Cressie, N. (1991), *Statistics for Spatial Data*, John Wiley and Sons, New York

Diggle, P. J. (1983), *Statistical Analysis of Spatial Point Patterns*, Academic Press, London

Diggle, P. J. (1985), A kernel method for smoothing point process data, *Applied Statistics*, Vol. 34, No. 2, pp. 138-147

Goodwin, M.T. and Unwin, D. (2000), Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations, *Transactions in GIS*, Vol. 4, No. 4, pp. 305-317

Hwang, S. (2004), Temporal extensions of K function, *UCGIS Fall* 2004, pp. 1-18

Jin, C.-J., Park, H.-S., and Kang, J.-M. (2012), An empirical analysis of locational tendency of coffee shops around Hongik university, *Journal of the Urban Design Institute of Korea*, Vol. 13, No. 5, pp. 71-82. (in Korean with English abstract)

Kloog, I., Haim A., and Portnov, B. (2009), Using kernel density function as an urban analysis tool: investigating the association between nightlight exposure and the incidence of breast cancer in Haifa, Israel, *Computers, Environment and Urban Systems*, Vol. 33, pp. 55–63.

Lee, B. (2008), Applying the L-index for analyzing the density of point features, *The Journal of GIS Association of Korea*, Vol. 16, No. 2, pp. 237-247. (in Korean with English abstract)

Lee, H. Y. and Sim, J. H. (2011), *GIS Geomatics*, Bobmunsa, Paju. (in Korean)

Lee, S.-K. and Lee, B. (2013), Assessing the appropriateness of the spatial distribution of Standard lots using the L-index, *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, Vol. 31, No. 6-2, pp. 601-609. (in Korean with English abstract)

Ripley, B. D. (1976), The second-order analysis of stationary point processes, J*ournal of Applied Probability,* Vol. 13, pp. 255-66.

Yim, P. and Lee, S. (2013), Estimation of prices and rents of knowledge industrial centers in Seoul metropolitan area considering spatial autocorrelation, *Journal of the Korea Real Estate Analysts Association*, Vol. 19, No. 2, pp. 5-20. (in Korean with English abstract)