INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

**ISSN: 2454-132X**
**Impact factor: 4.295**
***(Volume 5, Issue 2)***
Available online at: www.ijariit.com

# Exploring the relative predictive efficiencies of spatial regression models

*Bhabani Shankar Das Mohapatra*
*dmbhabanishankar@gmail.com*
*Jawaharlal Nehru Technological University, Hyderabad, Telangana*

*Dr. E. G. Rajan*
*rajaneg@yahoo.co.in*
*Pentagram Research Centre Private Limited, Hyderabad, Telangana*

## ABSTRACT

***Spatial regression models are standard tools for analyzing data with spatial correlation. These models are broadly used in the social sciences for predicting the socio-economic factors. In this paper, we discuss various spatial regression models and explain the concepts based on real data to demonstrate how to obtain and interpret relevant results. We describe prediction efficiencies of various predictors relative to the efficient minimum mean square error predictor in spatial models containing spatial lags in both the dependent variable and the error term. We consider Multiple Linear Regression Model (MLRM), Spatial Autoregressive Models (SAR), Spatial Autoregressive in the Error-term Model (SEM) and Spatial Durbin Models (SDMs) to estimate the literacy progress in the districts in Odisha as a result of changing socio-economic factors over time. The goodness of fit of the different models is compared along a series of hypotheses about the performance of the specifications considering spatial relationships among the observations. The spatial analysis proved the existence of positive spatial autocorrelation and persistence of disparities in literacy attainment level across the regions during the analyzed period. The results of econometric analysis confirmed the expected positive impact of economic growth on literacy progress level as well as the necessity to incorporate the spatial dimension into the model.***

***Keywords***— *Spatial regression, Spatial analysis, Exploratory spatial data analysis, ML method, Euclidean distance, Lagrange multiplier tests*

## 1. INTRODUCTION

Conventional statistical methods usually are not practicable to be effectively used for spatial data analysis as the techniques fail to consider spatial influences among data points (neighborhood points). To fill the gap, the techniques of spatial statistics have been formalized and developed since the 1950s (Cressie, 1993) which have drawn attention and are being widely applied in spatial data modeling and analysis in the fields of social sciences.

In geo-spatial arrangement, certain patterns are highly developed whereas others are less developed regions in space. Spatial analysis and spatial data mining reflecting the geographical location of the analyzed region are the appropriate techniques for such analysis. In recent days, several geographical information systems (GIS) and other distinctive tools (R, SAS, MatLab,) have been advanced to analyze and discover the hidden useful knowledge inherent in the spatial data.

Spatial analysis solves problems in spatial data modeling and analysis using different statistical approaches to quantify two-dimensional and two-directional data dependence and heterogeneity in space. Based on the works in [1], most of the work in spatial data modeling and analysis can be referred to as either the model-driven approach or the data-driven approach [2]. The model-driven approach is employed by spatial regression analysis, starting with a model structure specification that is then fitted with the data. Methods in this category deal with model parameter estimation and model structure specification [1]. The data-driven approach in geostatistics uses variogram and kriging methods and assumes randomness in data distribution (i.e., the null hypothesis) based on a normal or randomization approach [3]. For applications, the model-driven approach mainly deals with spatial modeling related to regional and urban economics, and the data-driven approach focuses on studies of issues in geophysics, biology, and agriculture.

The purpose of this article is to provide results which exemplify the magnitude of inefficiencies of various predictors in a spatial regression model. We consider prediction issues in the context of a linear spatial regression model which contains exogenous variables, a spatially lagged dependent variable, and a spatially lagged error term. For each of our considered predictors, we provide an estimate of its predictive efficiency relative to the full information predictor. All of our results specialize in models in which one or both of these spatial lags are absent.

The study presented herein explores the model-driven approaches in spatial modeling and analysis for spatial prediction of literacy growth in the districts in Odisha. It first focuses on the assessment of the persistence of regional disparities across the districts in the area of education based on the spatial analysis of literacy level (higher secondary and

*© 2019, www.IJARIIT.com All Rights Reserved* 

secondary education of population aged 25–60, expressed in %) during the period 2011–2018, and, secondly, it investigates the impact of regional GDP growth on the literacy level in 2018 based on the estimation of corresponding non-spatial and spatial econometric models. The importance of spatial interaction and geographical proximity is also taken into account in the regression analysis testing the dependence between the attained education and regional economic growth and can be therefore considered as an interesting contribution to the discussion and to the empirical evidence in regional analyses of education attainment.

## 2. BIBLIOGRAPHIC REVIEW

Can [7] and Dubin [8] is credited as the first researchers to introduce spatial econometrics and statistical techniques to the hedonic approach, respectively. The introduction of these techniques was followed by numerous works on the spatial hedonic approach [9]. The models employed in spatial statistics and semi-parametric statistics assume a continuous spatial process in a continuous domain D [4]. This assumption is not made for the observations (realizations) but for the domain and underlying spatial process. Hence, the relationship between the observations is the same as that between the data at a new (arbitrary) site $s_m$ and the observations if the new site is in the region D. For this reason, spatial prediction with spatial statistical models and semi-parametric models is a highly intuitive technique. On the other hand, the models used in spatial econometrics are based on a discrete sub set of a domain D [4]. These models typically structuralize the mutual dependence between the observations in each discrete zone using a spatial weight matrix (SWM). The dependence is changed if another SWM is utilized; hence, the derived parameters are valid only for the particular SWM. In this study, the following spatial models have been adopted for empirical comparison: Multiple Linear Regression Model (MLRM), Spatial Autoregressive Models (SAR), Spatial Autoregressive in the Error-term Model (SEM) and Spatial Durbin Models (SDMs). For comparison, a practical method was employed for considering spatial dependence.

## 3. EXPLORATORY DATA ANALYSIS

Regression analysis often begins with exploratory data analysis. Exploratory spatial data analysis (ESDA) is an additional crucial step in spatial regression modeling, focusing on the spatial feature of data. ESDA involves visualizing spatial patterns in the data, identifying spatial clusters and spatial outliers, and diagnosing possible misspecification of spatial aspects of the statistical models, all of which can help better specify regression models. In the following, we discuss basic concepts and related issues in the context of ESDA. In particular, we review spatial autocorrelation, spatial heterogeneity, spatial weight matrix based on spatial neighborhood structures, and discuss the modifiable areal unit problem. These concepts and issues are essential in spatial regression modeling.

### 3.1 Spatial Autocorrelation

Spatial autocorrelation (also known as spatial dependence) can be defined as a similarity (or dissimilarity) measure between two values of an attribute that are nearby spatially. In other words, with positive spatial autocorrelation, high or low values of an attribute tend to cluster in space whereas, with negative spatial autocorrelation, locations tend to be surrounded by neighbors with very different values. Spatial autocorrelation can be measured by various indexes, of which the most well-known is Moran's I statistic [5]. Moran's I statistic measures the degree of linear association between an attribute (y) at a given location and the weighted average of the attribute at its neighboring locations (Wy) and can be interpreted as the slope of the regression of (y) on (Wy). Spatial autocorrelation can be visually illustrated in a Moran scatter plot, in which (Wy) on the vertical axis is plotted against (y) on the horizontal axis [6].

ESDA is usually based on spatial autocorrelation analysis both on the global and the local level. The global Moran's *I* statistic and the local Moran's *I* statistic are the main instruments for this part of our analysis. The confirmation of the spatial autocorrelation implies the presence of spatial spill-over effects, which means that the data from one region can influence the data from some other region. While the global Moran's *I* statistic provides us with a measurement of the global spatial autocorrelation—that is, how strong the spatial association is across neighbouring regions (a single value for the whole data set)—its local version enables assessing the spatial autocorrelation for one particular spatial unit (region). The LISA (Local Indicators of Spatial Association) presented by Anselin (1995) [6] can be used to determine the existence of local spatial clusters.

### 3.2 Spatial Heterogeneity

Spatial heterogeneity (also known as spatial structure, non-stationarity, or large-scale global trends of the data) refers to differences in the mean, and/or variance, and/or covariance structures including spatial autocorrelation within a spatial region. In contrast, spatial homogeneity (also known as stationarity) requires that the mean and the variance of an attribute be constant across space and that spatial autocorrelation of the attribute at any two locations depends on the lag distance between the two locations, but not the actual locations.

### 3.3 Neighborhood Structure and Spatial Weight Matrix

To account for spatial autocorrelation in lattice data analysis, it is necessary to establish a neighborhood structure for each location by specifying those locations on the lattice that are considered as its neighbors [1]. In particular, we need to specify a spatial weight matrix **W** of dimension ( $n \times n$ ), where $n$ is the number of regions in the data sets corresponding to the neighborhood structure such that the resulting variance-covariance matrix can be expressed as a function of a small number of estimable parameters relative to the sample size (Anselin2002). Popular spatial weight matrices in spatial econometrics include the so-called ''rook's case'' and ''queen's case'' contiguity weight matrices of order one or higher, the k-nearest neighbor weight matrices, the general distance weight matrices, and the inverse distance weight matrices with different powers, the latter three of which are distance-based (Anselin1992).

In this article, we consider three different specifications of the weight matrix—the queen contiguity matrix of the first order, the 4-nearest neighbour's weight matrix and the threshold distance matrix based on Euclidean distance metric. First of all, it is necessary to define which regions are neighbours- that is, to decide which elements of matrix W will be non-zero. In case of the queen contiguity matrix, the regions are neighbours if they share any part of a common border, in the 4-nearest neighbours weight matrix each region has exactly 4 neighboring regions and in case of Euclidean distance matrix is the specification of neighboring regions based on distances between regions (the threshold value is usually chosen in order to ensure that each region has at least one neighbours). The

*© 2019, www.IJARIIT.com All Rights Reserved* 

diagonal elements of the matrix W are set to zero. In the Moran scatterplot, which is divided into four quadrants, it is possible to identify four different spatial associations: high-high (HH), low-high (LH), low-low (LL) and high-low (HL). The associations HH and LL indicate a positive spatial autocorrelation, while the associations LH and HL a negative autocorrelation. The four categories (HH, LL, LH and HL) of the spatial association correspond to the four quadrants in the Moran scatterplot.

### 3.4 Data Example

The analysis in this paper was done based on spatial data downloaded in the form of shapefile (.shp) for the district of Odisha State. The main focus was on the attained education characterized by the population aged 25–60 (in %) with secondary, higher secondary, degree and post-graduate education attainment for the villages and wards in the districts during the period 2011–2018. In order to investigate the impact of region's growth on the attained education in 2018, the GDP per capita (defined at current market prices in Purchasing Power Standard) growth rates from 2011 to 2018 were calculated based on the data.

## 4. SPATIAL REGRESSION ANALYSIS

Next, we concentrate on regression (econometric) analysis - estimation of the regression model reflecting the dependence between the attained literacy and regional economic growth.

### 4.1 Ordinary Least Square Method

The estimation of the classic linear regression model is given by the ordinary least squares (OLS) method, that is:

$$y = X\beta + \varepsilon \{1\} \quad (1)$$

Where **y** is an ($n$×1) dimensional vector of a dependent variable, **X** is a matrix of independent variables of dimension ($n$×($k$ +1)) and $k$ is a number of independent variables, **β** is (($k$ +1)×1) dimensional vector of unknown parameters and **ε** is an ($n$×1) vector of independent identically distributed (i.i.d.) error terms.

Since the presence of spatial effects can have implications on the quality of estimates, we need to test the presence of spatial dependence to choose the appropriate form of the spatial model. We can distinguish between two types of spatial dependence—spatial lag and spatial error, which also corresponds to the two forms of spatial models—SAR or Spatial Autoregressive Model and SEM or Spatial Error Model [1]

### 4.2 Spatial Autoregressive Model (SAR)

The use of the SAR model is appropriate in case the focus is on the assessment of the presence and strength of spatial interaction, and the SEM model in case of spatial dependence in the regression disturbance term [1]. The SAR model can be formulated as follows:

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

Where $\rho$ is the scalar spatial autoregressive parameter measuring the degree of dependence, **W** is a spatial weight matrix of dimension ($n$×$n$), **Wy** is an ($n$×1) dimensional vector of the spatially lagged dependent variable, **u** is an ($n$×1) the dimensional vector of error terms and all other terms were previously defined above. Since the value $\rho \neq 0$ implies the existence of spatial effects across neighbouring regions (endogenous interaction effects6), a zero value indicates no spatial dependence between observations of the considered dependent variable.

### 4.3 Spatial Error Model (SEM)

The SEM model is expressed as:

$$y = X\beta + \varepsilon,\ \varepsilon = \lambda W\varepsilon + u \quad (3)$$

Where λ is a spatial error parameter reflecting the intensity of spatial autocorrelation between regression residuals and **Wε** is an ($n$×1) dimensional vector of spatially lagged error terms. Both the SAR and the SEM model can be estimated by maximum likelihood (ML) method.

### 4.4 Spatial Durbin Model (SDM)

The formulation of the spatial Durbin model (SDM) including the spatial lags of both dependent and explanatory variables is as follows:

$$y = \rho Wy + X\beta + WX\Theta + u \quad (4)$$

Where **WX** denotes the ($n$×$k$) dimensional matrix of spatially lagged explanatory variables8, **Θ** is a ($k$ ×1) dimensional vector of parameters reflecting the exogenous interaction effects (Anselin, 2003) and all other symbols were previously explained above. Similarly, as SAR and SEM models, also the SDM model can be estimated based on the ML method. In general the SDM model plays an important role in the spatial econometrics literature [15], since it is possible to derive from it a number of other models as special cases, for instance, in case we cannot reject the null hypothesis **Θ = 0**, the SDM becomes a SAR model; if the null hypothesis **Θ** = $-\rho$**β** cannot be rejected, we will receive an SEM model from the SDM model and imposing the restrictions both ρ = 0 and **Θ = 0** yields the classic linear regression model.

## 5. EXPERIMENTAL SETUP AND RESULT

We analyze with mapping of the data via box maps for 2011 and 2018 (see figure 1 and figure 2) in order to illustrate the unequal distribution of literacy over space.

Box map is a special form of a quartile map and consists of six categories. Besides the four categories corresponding to the four quartiles, two extra categories are specified for upper and lower outliers, respectively. Therefore, as Figure 1 reveals, the first and last quartile no longer correspond to exactly one-fourth of the observations, since the lower and upper outliers, respectively, are depicted as extra categories. Three lower outliers were detected in 2011 (Baleswar and Bhadrak region) and five lower outliers in 2018 (3 out of 5 analysed Baleswar and Bhadrak region; 2 Puri and Ganjam regions.

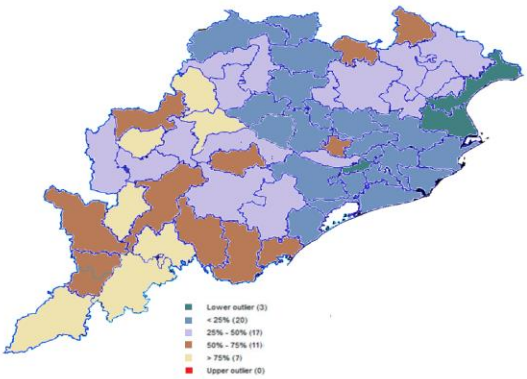


**Fig. 1: Box map for literacy level in 2011**

Figure 2 reveals that, in general, the regions with the higher (lower) percentage of the population with considered literacy level tend to be gathered together, but on the other hand, it is also possible to identify some regions with a higher percentage

*© 2019, www.IJARIIT.com All Rights Reserved* 

of "educated" population in comparison to the neighbouring regions. This is mostly the case of capital city regions or regions with higher educational institutes since these regions attract people with higher qualification.

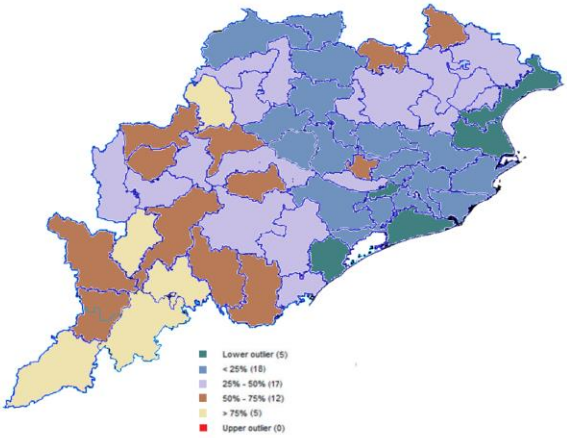


**Fig. 2: Box map for literacy level in 2018**

In order to confirm that location affects the share of people with analyzed literacy level, that is, to test for spatial autocorrelation, it is necessary to define spatial neighbourhoods and spatial weights. The spatial weight matrix W was specified in three different ways in order to show the sensitivity of its specification to the results. Firstly, a contiguity weight matrix of the queen case definition of neighbours was specified (two regions are considered as neighbours if they share any part of a common border); secondly, we used the weight matrix of 4 nearest neighbours; and, finally, W was based on Euclidean distance metric.

Concerning the different specifications of, the global Moran's *I* statistics were computed with respect to different specifications the **W** matrix during the period 2011–2018 for the three above-mentioned specifications of weight matrices. (Figure 3)

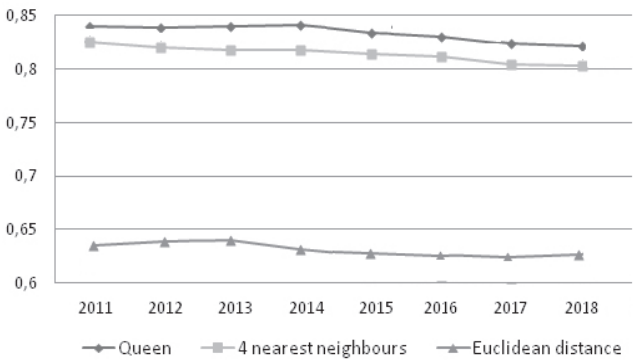


**Fig. 3: Global Moran's I during 2011–2018 for different specification of weight matrices**

Since the global Moran's *I* values based on queen case and 4 nearest neighbours of weight matrix were similar, the global Moran's *I* values based on Euclidean distance weight matrix were substantially lower. It is important to mention that these weight matrices are used in a row-standardized form—that is, the elements of each row sum to one and each neighbour of the concrete region is given equal weight. The values of global Moran's *I* statistics are larger than the expected value $E(I) = -1/(n-1) = -0.00398$ which indicates the positive spatial autocorrelation, which means that it is much more likely that regions with high (low) percentage of population with considered literacy level will have neighbours with also high (low) share of population with at least upper secondary education than in case of pure randomness.

The results of the spatial analysis revealed the persistence of disparities in considered literacy level Odisha during the analyzed period. Moreover, as we can see in figure 2, in 2018 even a greater number of outliers were identified.

We present the results of regression (econometric) analysis—estimation of the regression model reflecting the dependence between the attained literacy and regional economic growth. The results are presented separately for queen case weights (Table 1) and Euclidean distance weights (table 2).

In the first step, the classic linear regression model (1) was estimated based on OLS. Estimation results are gathered in Tables 1 and 2. Both parameters $\beta 0$ and $\beta 1$ were statistically significant, the expected positive relationship between the attained education and regional economic growth was confirmed. However, the $R^2$ was very low and the diagnostic statistics—Moran's *I* applied on regression residuals and the Lagrange Multiplier tests—indicated that we can clearly reject the null hypothesis of non-spatial dependence, which means that the spatial aspect should be taken into account.

The estimation results by ML for the SAR and the SEM model are also given in Tables 1 and 2. All the estimated parameters were statistically significant; the positive value of regression parameter $\beta 1$ confirms the positive impact of growth on the attained education. The statistical significance of spatial parameters ρ and $\lambda$ confirms the strong positive and significant spatial autocorrelation, that is, the presence of spatial effects across neighbouring regions. The higher percentage of the literate population in a specific region will tend to push up the rate in the neighbouring regions. The low values of Moran's *I* statistic applied on corresponding spatial residuals in SAR and SEM indicates no further evidence of spatial autocorrelation.

**Table 1: Estimation results of MLRM, SAR, SEM and SDM models; Weight: Queen**

| Estimation | MLRM | SAR | SEM | SDM |
|---|---|---|---|---|
| | Least Square | Maximum Likelihood | Maximum Likelihood | Maximum Likelihood |
| $\beta 0$ | 71. 957 | 11.734 | 70.341 | 10.780 |
| $\beta 1$ | 47.312 | 12.401 | 27.553 | 23.135 |
| $\lambda$ | - | - | 0.675 | - |
| $\rho$ | - | 0.598 | - | 0.603 |
| $\theta$ | - | - | - | |
| $R^2$ | 0. 198 | 0.604 | 0.689 | 0.694 |
| Moran's I (error) | 13.896 | | | |
| Moran's I Spatial Residue | | -0.049 | -0.0785 | -0.0624 |

*Note: Table 1 refers to the statistical significance at 1% level of significance*

Also, the SDM model was estimated and its parameters $\beta 0$, $\beta 1$ and ρ were statistically significant at a 1% significance level. A different situation occurs in the case of parameter **θ** which was in the queen case definition of the weight matrix (Table 1) statistically significant at 5% significance level, but under Euclidean weights (Table 2) no statistical significance was proved. The negative sign of this parameter indicates that higher growth in neighbouring regions connected *inter alia* with better career opportunities and better economic well-being will attract a certain share of the better literate population to move to such regions and thus the percentage of the population with considered education attainment level in analyses region will go down.

*© 2019, www.IJARIIT.com All Rights Reserved* 

**Table 2: Estimation results of MLRM, SAR, SEM and SDM models; Weight: Euclidean distance**

| Estimation | MLRM | SAR | SEM | SDM |
|---|---|---|---|---|
| | Least Square | Maximum Likelihood | Maximum Likelihood | Maximum Likelihood |
| $\beta 0$ | 71. 957 | 3.562 | 65.450 | 3.451 |
| $\beta 1$ | 47.312 | 13.521 | 25.70 | 20.254 |
| $\lambda$ | - | - | 0.754 | - |
| $\rho$ | - | 0.798 | - | 0.803 |
| $\theta$ | - | - | - | |
| $R^2$ | 0. 198 | 0.625 | 0.652 | 0.624 |
| Moran's I (error) | 10.563 | - | - | - |
| Moran's I Spatial Residue | | -0.031 | -0.0223 | -0.0314 |

*Note: Table 2 refers to the statistical significance at 1% level of significance*

Regarding the statistical significance of parameters and other model characteristics, we prefer the SDM model for queen case weights and SEM model for Euclidean distance weights. We can also conclude that in both cases the consideration of the space in econometric modelling is an unavoidable part of the estimation procedure in order to receive econometrically correct results.

## 6. CONCLUSION

In this research we performed the spatial analysis of literacy level of the population aged 25–60 with at least secondary education across the state of Odisha during the period 2011–2018. We also investigated the impact of regional growth on the share of people with specified literacy level.

## 7. REFERENCES


[1] Luc Anselin Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity (*references*) Geographical Analysis, Vol. 20, No. 1 (January 1988) Ohio State University Press

[2] Luc Anselin S*patial Dependence and Spatial Structural Instability in Applied Regression Analysis*, Geographical Analysis, (May 1990) Ohio State University Press

[3] Schabenberger O, Pierce FJ (2002) Contemporary statistical models for the plant and soil sciences. Boca Raton, FL: CRC Press.

[4] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[5] Moran, P. (1948) The Interpretation of Statistical Maps. Journal of the Royal Statistical Society, 10, 243-251.

[6] Luc Anselin, Local Indicators of Spatial Association-LISA, Geographical Analysis. Vol.27, No 2 (April 1995, Ohio State University Press

[7] Can A. The measurement of neighborhood dynamics in urban house prices, Economic Geography 1990; 66:254–72.

[8] Dubin RA. Estimation of regression coefficients in the presence of spatially autocorrelated error terms, The Review of Economics and Statistics 1998; 70: 466–74.

[9] Anselin L, Le Gallo J. Interpolation of air quality measures in hedonic house price models, spatial aspects, Spatial Economic Analysis 2006; 1(1): 31–52.

[10] Schabenberger O, Gotway CA. Statistical methods for spatial data analysis. Boca Raton, Chapman&Hall/CRC; 2005.

[11] Kelejian HH, Prucha IR. The relative efficiencies of various predictors in spatial models containing spatial lags, Regional Science and Urban Economics 2007; 37: 283–432.

[12] Kato T. A further exploration into the robustness of spatial autocorrelation specifications, Journal of Regional Science 2008; 48( 3): 615–39.

[13] Rubin DB. Inference with missing data, Biometrika 1976; 63: 581–92.

[14] LeSage JP, Pace RK. Models for spatially dependent missing data, The Journal of Real Estate Finance and Economics 2004;29(2): 233–54.

[15] LeSage, J. and Pace, K.R. (2009) Introduction to Spatial Econometrics. Chapman and Hall/CRC.


*© 2019, www.IJARIIT.com All Rights Reserved* 