International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Multilevel Modelling with Eigenvector Spatial Filtering and its Application to UN Score Data in Kendari

Lemma Firari Boer[a]*, Hari Wijayanto[b] , Indahwati[c]

[a]*Graduate Student in Department Statistics, Bogor Agricultural University, Jl. Meranti Wing 22 level 4 Kampus IPB Darmaga, Bogor (16680), Indonesia*

[b,c]*Lecturer in Department Statistics, Bogor Agricultural University, Jl. Meranti Wing 22 level 4 Kampus IPB Darmaga, Bogor (16680), Indonesia*

[a]*Email: firariboer@gmail.com*

[b]*Email: hari_ipb@yahoo.com*

[c]*Email: indah.stk@gmail.com*

Abstract

Spatial dependence is a condition where locations will affect its neighborhood that tend to have the same characteristics or attributes. Eigenvector spatial filtering (ESF) is a method initially used to overcome spatial dependence in one-level linear regression by adding the eigenvector function that describes spatial effect from model. This research aims to combine ESF with multilevel modelling and applied them into the data that has both spatial dependence and hierarchy effect and compare the results with those of conventional multilevel model (without ESF). The results indicate that ESF method gave a smaller variance of level-2 random effect and AIC value. It also can be shown that the students-teachers ratio is the only significant predictor that affect UN score in Kendari at the alpha level of 5%.

Keywords: Spatial dependence; Eigenvector Spatial Filtering; Multilevel Model.

---

* Corresponding author.

1. Introduction

Multilevel modelling is an analysis used to explain the relationship between independent and dependent variables where there is a source of variability caused by hierarchical data structures. One example of hierarchical data is student data nested at school. In Goodchild [4] Spatial dependence is a condition where locations will affect its neighborhood that tend to have the same characteristics or attributes. The existence of spatial dependence indicates that the observation value from some location is related to another location, ignoring this effect will caused the violation of the assumption of residual independence.

Researches relates on the application of multilevel modelling with the existence of spatial dependence has been conducted by corado and fingleton [3] who propose the addition of spatial effect components into random effects on multilevel models. Then Pierawan and Tampubolon [8] used the specs of the SAR (Spatial Autoregressive Model) model and the SEM (Spatial Error Model) model to correct the error structure in multilevel models. Although SAR and SEM specifications are very popular there is another approach to overcome the problem of spatial dependence, this method is called Eigenvector spatial filtering (ESF).

Eigenvector spatial filtering (ESF) is a method initially used to overcome spatial dependence in one-level linear regression by adding eigenvector function that describes spatial effect from model [10], but on its development ESF can be applied into multilevel model [7]. In the application on health field, Park and Kim [7] used the ESF method to describe the spatial dependence on the data with hierarchy structure. This method is reliable because the properties of eigenvectors are mutually orthogonal so that multicollinearity will not occur.

This research aims to combine ESF with multilevel modelling and applied them into data possessing both spatial dependence and hierarchy effect and compare with conventional multilevel model (without ESF). National Exam (also called *Ujian Nasional* or UN in Indonesia) data score from elementary schools in Kendari city was used as the applied data, which is school as level-1 unit and the sub-district (*Kelurahan* in Indonesian) as level-2 unit.

2. Methodology

2.1. Multilevel Model

Multilevel modelling is used on data that has hierarchy structure, null model is the simplest model which only has the intercept without using any independent variables. Null model can be written as follows:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \tag{1}$$

index I denotes i-th individual (i=1, 2,…, N_j) and index j denotes j-th group (j=1,2,…,J), ε_{ij} denotes level-1 residual which follow $N(0, \sigma_\varepsilon^2)$. y_{ij} denotes i-th individual and j-th group of dependent variables, and β_{0j} denotes intercept which varying for each j-th group, these intercepts can be written as follows:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{2}$$

Equation (2) is called level-2 model where γ_{00} denotes overall average and $u_{0\mathrm{j}}$ is the deviation from overall average or level-2 residual which $u_{0\mathrm{j}} \sim \mathrm{N}(0, \sigma_{u0}^2)$. Variance of y_{ij} can be written as sum of level-1 and level-2 variance:

$$var(y_{ij}) = var(u_{0\mathrm{j}}) + var(\varepsilon_{ij}) = \sigma_{u0}^2 + \sigma_{\varepsilon}^2 \tag{3}$$

whereas covariance from two individuals within group are equal to level-2 variance:

$$Cov(Y_{ij}, Y_{i'j}) = var(u_{0\mathrm{j}}) = \sigma_{u0}^2. \tag{4}$$

Therefore, correlation between two individuals or intraclass correlation (ρ) is:

$$\rho = \sigma_{u0}^2/(\sigma_{u0}^2 + \sigma_{\varepsilon}^2) \tag{5}$$

We can write random intercept model without interaction where each level has only one independent variable:

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij} \tag{6}$$

Equation (6) is level-2 model, for level-2 model is given by:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + u_{0\mathrm{j}} \tag{7}$$

By substituting equation (7) into equation (6), random intercept model can be written as follows:

$$y_{ij} = \gamma_{00} + \gamma_{01}Z_{1j} + \beta_{1j}X_{ij} + u_{0\mathrm{j}} + \varepsilon_{ij} \tag{8}$$

Where $y_{ij}$is dependent variable, γ_{00} is the overall average of dependent variable, γ_{01} is level-2 slope of variable Z_{1j}, β_{1j} is level-1 slope of variable X_{ij}, X_{ij} level-1 independent variable, Z_{1j} is level-2 independent variable, $u_{0\mathrm{j}}$ is level-2 residual which $u_{0\mathrm{j}} \sim \mathrm{N}(0, \sigma_{u0}^2)$, ε_{ij} is level-1 residual which $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_{\varepsilon}^2)$. $\gamma_{00} + \gamma_{01}Z_{1j} + \beta_{1j}X_{ij}$ is called fixed effect and $u_{0\mathrm{j}} + \varepsilon_{ij}$ is called random effect.

Generally, maximum likelihood (ML) is used as the estimating parameter. There are two types of ML in multilevel regression analysis such as Full Maximum Likelihood (FML) and Restricted Maximum Likelihood (RML). The difference between those methods are in the likelihood function where FML has both regression and variance component while RML only has the variance component within its likelihood function.

2.2. Spatial Weight Matrix

Spatial weight matrix is a matrix that describe relation between locations. This relation can be based on the distance or contiguity. There are three types of contiguity such as rook, bishop, and queen (Figure 1) [6].

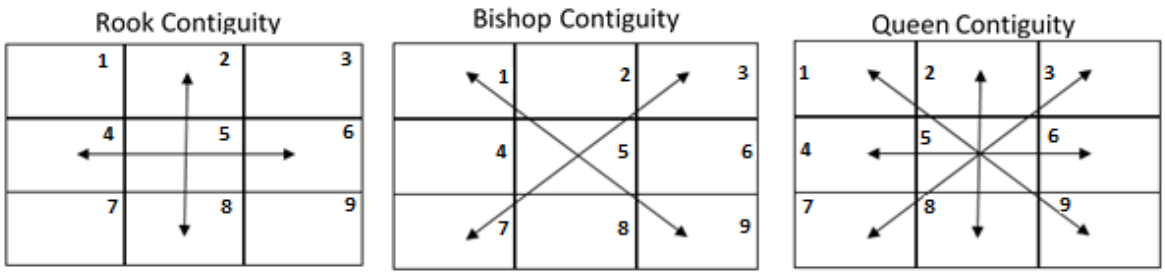

Figure 1: Contiguity spatial relationship

Constructing spatial weight matrix can be done by giving 1 if between two locations share their border and 0 if the two locations do not share their border. This can be written as:

$$w_{ij} = \begin{cases} 1;\ i \text{ and } j \text{ share their border} \\ 0; i \text{ and } j \text{ doesn't share their border} \end{cases} \quad (9)$$

To make sure that there is no scaling effect, usually row normalization was applied. This method is the most popular among the standardization methods. Row normalization can be written as follows:

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \quad (10)$$

2.3. Spatial Dependence Test

Moran's I is the modification of Pearson's correlation coefficient. Within the concept of spatial observation, variables will tend to be similar to its nearby observation [1]. To accommodate this, x_i and x_j are weighted by w_{ij}.

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (11)$$

With I is Moran's I, w_{ij} is spatial weight matrix, n is number of observations, $\bar{x}$ mean of X, and $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$

Moran's I hypothesis and test statistics can be written as follows:

H0: I=0 (there is spatial autocorrelation)

H1: I≠0 (there is no spatial autocorrelation)

$$Z_{hit} = \frac{I-E(I)}{\sqrt{Var(I)}} \quad (12)$$

With $E(I) = -\frac{1}{n-1}$, rejech H0 if $|Z_{hit}| > Z\alpha_{/2}$ or $|Z_{hit}| > 1.96$ if α level at 5%.

2.4. Eigenvector Spatial Filtering (ESF)

Filtering is one of the methods to account the presence of spatial autocorrelation by filtering the variables within the model into two types of variables, i.e. nonspatial and spatial variables. Filtering uses spatial weight matrix to capture the covariance from one or more of random variables which describe characteristics from areal units [5]. The examples of filtering process are the extraction of $(I_n - \rho W)^{-1}(X\beta + \varepsilon)$ from spatial autoregressive model (SAR) and $X\beta + (I_n - \theta W)^{-1}\varepsilon$ from spatial error model (SEM). Eigenvector Spatial Filtering (ESF) is a method that utilize the Eigen decomposition from transformation of spatial weight matrix [2]:

$$\left(I - 11^T/_n\right) W \left(I - 11^T/_n\right) = MWM \tag{13}$$

With *I* identity Matrix ($n \times n$), 1 is vector of ones ($n \times 1$), *W* is spatial weight matrix. The Eigen decomposition from MWM will obtain *k* eigenvalues $(\lambda_1, \lambda_2, \dots, \lambda_k)$ and *k* eigenvectors $(v_1, v_2, \dots, v_k)$.

$$det\left[\left(I - 11^T/_n\right) W \left(I - 11^T/_n\right) - \lambda I\right]=0 \tag{14}$$

$M = \left(I - 11^T/_n\right)$ at equation (14) is called projection matrix. This projection matrix, together with W, are used to generate spatial proxy variables which is a set of eigenvectors from MWM. Performance of projection matrix depends on the utilized model. For example, suitable projection matrix for OLS is $M = I - X(X'X)^{-1}X'$ [10].

2.5. ESF Multilevel Model

Park and Kim [7] stated that the integration of random intercept model and ESF can be written as the sum of fixed effect and random effect with linear combinations of eigenvectors.

$$y_{ij} = \gamma_{00} + \gamma_{01}Z_{1j} + \beta_{1j}X_{ij} + u_{0\mathrm{j}} + \varepsilon_{ij}$$

$$y_{ij} = \left(\gamma_{00} + \gamma_{01}Z_{1j} + \beta_{1j}X_{ij}\right) + \left(\gamma_0 + \gamma_1 v_{1j} + \cdots + \gamma_n v_{nj}\right) + \left(u_{0j}^* + \varepsilon_{ij}\right) \tag{15}$$

It can be seen from equation above that filtering process are conduct by separating the spatial autocorrelation effect which is describe by residual component.

2.6. Data

Applied data used in this research are the UN score of elementary school in Kendari city 2016, Primary Education Data (*Data Pokok Pendidikan* or DAPODIK in Indonesian) provided by Kendari Board of Education, and the Village Potential Data (*Potensi Desa* or PODES) which is provided by the Central Bureau of Statistics. Below are the variables that were used:

- Dependent variable:

Y: UN Score (overall average score)

- Level-1 independent variables (elementary schools):

  Students-groups ratio (x_1), students-classrooms ratio (x_2), students-teachers ratio (x_3), percentage of teachers with qualification (x_4), percentage of certified teachers (x_5), percentage of civil servant teachers (x_6), percentage of proper classrooms (x_7).

- Level-2 independent variables (sub-district):

  Number of people with social insurance administration organization (Z)

2.7. Analysis

- Explore data in order to capture general description about data.
- Perform Moran's *I* test for spatial dependency.
- Perform stepwise for selecting level-1 and level-2 independent variables.
- Perform Multilevel model and ESF multilevel model.
- Calculate AIC and interpreted the model.

3. Result and Discussion

3.1. General Description

Kendari is a city in Southeast Sulawesi province in Indonesia with four districts and 64 sub-districts. The total number of elementary school in Kendari is 114. Figure 2 shows the distribution map of elementary schools in Kendari.

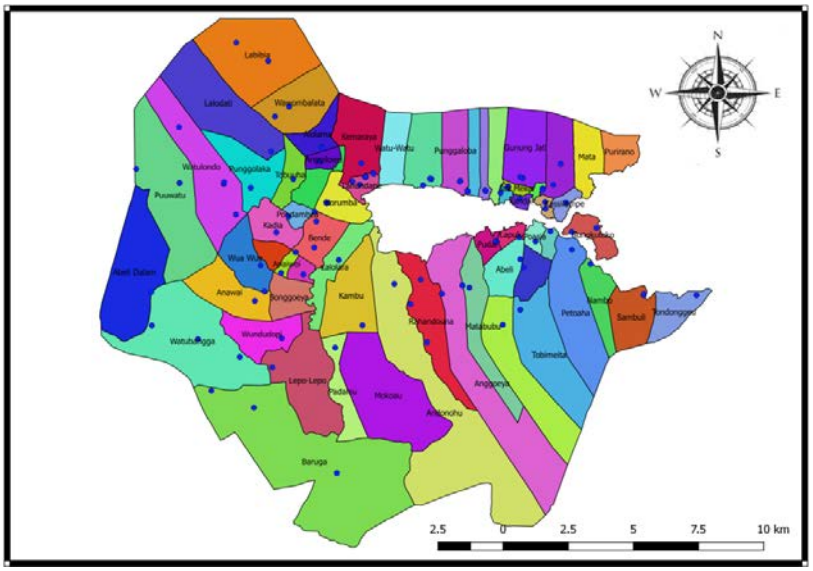

Figure 2: Contiguity spatial relationship

According to the Permendikbud (Ministry of Education and Culture regulation) no. 17 year 2017 article 26, the number of groups should be between 6 and 24 for every elementary schools, then in Permendikbud No. 23 years of 2013 article 23 section, the students-groups ratio for every elementary schools are should not be greater than 32 people, also in Permendikbud No. 23 years of 2013 article section 5 and 8 stated that for every elementary schools there is one teacher for 32 students and there are 70% among the total number of teachers with the academic qualifications of bachelor degree or diploma and 35% of teachers are certified.

Whereas in DAPODIK data, students-groups ratio and students-classroom ratio for every sub-district are no more than 32 and almost 50% of sub-district have 32 of students-classroom ratio and 18% of sub-district have students-classroom ratio no more than 60.

From the data, it also can be known that all sub-districts have students-teachers ratio below 32 people and percentage of proper classrooms above 50. In 85% of sub-districts, 70% of their teachers are qualified; in 73% of sub-district, more than 35% of its teachers are certified. Finally, in 83% sub-district, the percentage of civil servant teachers reach above 50%.

For UN score data, it can be shown that all sub-district have the average score between 70 to 90, it means that all district have a good achievement in UN score. Below is chart of UN score of data by sub-district:

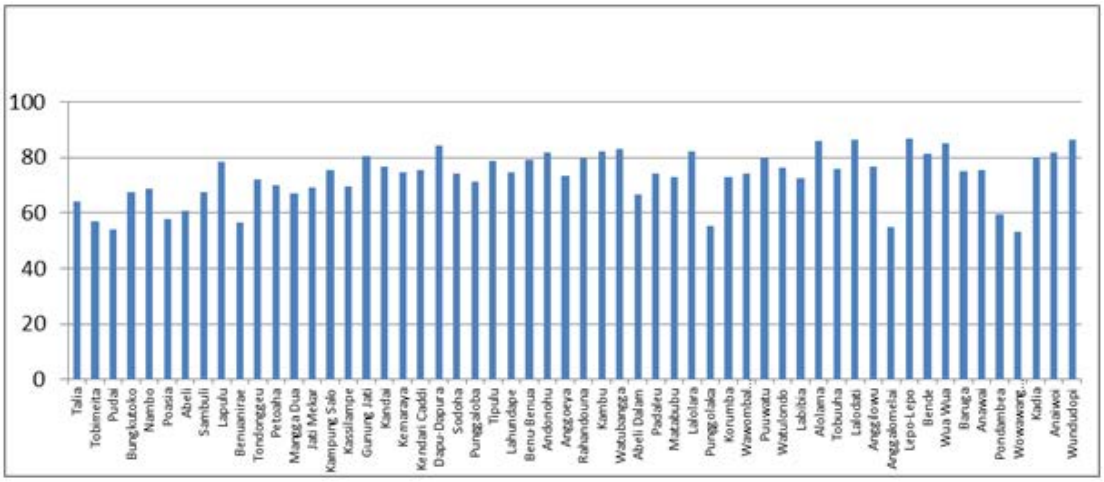

Figure 3: UN score in Kendari

3.2. Modelling

Before conducting the modelling, it is necessary to check the presence of spatial dependence by Moran's *I* test. The test show that there is an existence of spatial correlation (0.255) at significant level of 5%.

Table 1: Stepwise Multilevel Regression

Variable	SumSq	MeanSq	NumDF	DenDF	F.value	elim.num	Pr(>F)
x_1	0.7978	0.7978	1	100.71	0.0151	1	0.9023
x_2	3.2344	3.2344	1	77.28	0.0622	2	0.8036
x_7	12.0086	12.0086	1	92.03	0.2353	3	0.6288
x_6	42.6139	42.6139	1	92.68	0.8437	4	0.3607
x_4	47.8711	47.8711	1	77.07	0.9652	5	0.3289
x_3	254.3366	254.3366	1	107.3	5.0662	Kept	0.0264
x_5	327.6141	327.6141	1	101.41	6.5259	Kept	0.0121
Z	206.2941	206.2941	1	47.28	4.1092	Kept	0.0483

From the Table 1, it can be seen that level-2 variable (Z) are included into the model as well as for x_3 and x_5 for level-1 variables because *p-value* is less than 0.05. There is a large difference of scale between Z and other variables in order to reduce the scaling effect logarithmic transformation is applied on Z.

After that, we checked the multicollinearity assumption by calculating VIF from the selected variables obtained through the stepwise process. Referring VIF value that is less than 2 from Table 2, it can be said that multicollinearity assumption is not violated. Normality assumption was also checked, and the QQ plot (Figure 4) shows that the points are following the straight line indicating that the normality assumption is not violated.

Table 2: VIF

Variable	x_3	x_5	Z
VIF	1.1826	1.2185	1.0361

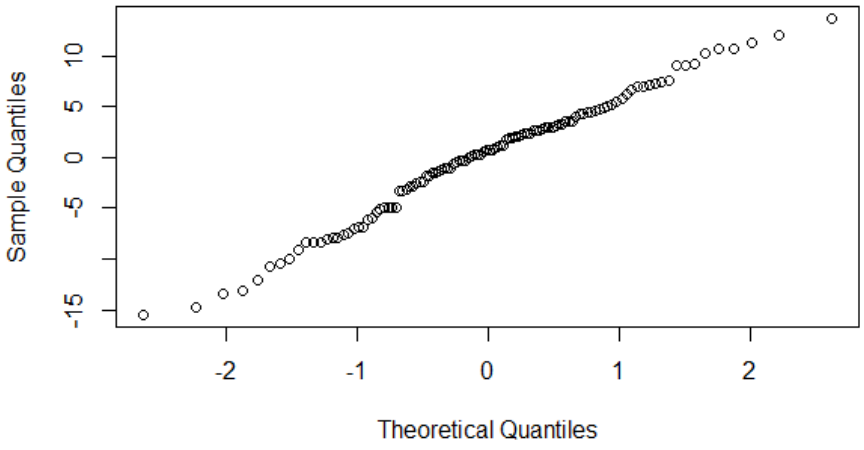

Figure 4: QQ-plot

It can be seen from Table 3 that we built three models: null model, multilevel model without ESF, and ESF multilevel model. Null model shows that data have a hierarchy effect because level-2 variance are high (40.8704) resulting to high intraclass correlation value (0.4185). Intraclass correlation also indicates that 41.85% of total variance are coming from level-2 effect of sub-district. For multilevel model without ESF, all level-1 and level-2 variables are at the significant level of 5%, these can be interpreted as students-teacher ratio (x_3), percentage of teachers with certification (x_5), and the number of people with social insurance administration organization (Z) that affected UN score in Kendari at the alpha level of 5%.

Multilevel modeling with ESF are applied by adding eigenvectors function from transformed spatial weight matrix into model. Three eigenvectors are selected by minimizing the effect of spatial dependence into multilevel model [11]. The model shows that only students-teacher ratio (x_3) that affected UN score in Kendari at the alpha level of 5% referring to p-value of x_3 (0.0491) is less than 0.05. All independent variables have positive coefficient, meaning that positive relation with the response variable. The result also shows that ESF method gave a smaller variance of level-2 random effect and AIC value indicating that ESF method are giving the best model compare to others.

Table 3: Modeling

Variables	Null Model	Multilevel Model	ESF Multilevel Model
Level-1			
x3		0.3905 (0.0283)	0.3293 (0.0491)
x5		0.1174 (0.0134)	0.0623 (0.1825)
Level-2			
Z		0.8984 (0.0476)	0.5304 (0.1784)
Random Effect			
Level-2 variance	40.8704	28.5946	9.1500
Level-1 variance	56.7762	50.2025	54.2783
Intercept	73.3 (0.0000)	55.3398 (0.0000)	62.2390 (0.0000)
Eigenvector			3 eigenvectors
AIC	834.3032	822.6230	792.7634
Log-likelihood	-414.1516	-405.3115	-387.3817

4. Conclusion

Modelling with ESF produces a smaller variance of level-2 random effect, indicating that ESF method are capable of dealing with spatial dependence effects. It also can be shown that the students-teachers ratio (x_3) is the only significant predictor that affect UN score in Kendari at the alpha level of 5%.

5. Recommendation

Using sub-district as level-2 unit gives its own limitations in this research, because sub-district is the smallest unit of PODES data then the availability of data becomes very limited. In Podes data for kendari city many variables are not available so the candidates for level-2 variables becomes less. To get a better model, further research can focus on the type of spatial weight matrix and another type of multilevel model. Since ESF is method that used to overcome spatial dependence problem, it might be possible to apply this method on another type of modelling besides the multilevel model.

Acknowledgements

The authors would like to acknowledge to the Educational Board of Kendari, the Central Bureau of Statistics (BPS), and the Department of Statistics, Bogor Agricultural University, for their support on this paper.

References

[1] Anselin L, “Spatial Econometrics: Methods and Models”, Kluwer Academic Publisher, Dordrecht, 1988, pp. 101-103.

[2] Chun Y, Griffith DA, “A quality assessment of eigenvector spatial filtering based parameter estimates for the normal probability model”, Spatial Statistics, Vol. 10, pp. 1-11, Apr. 2014.

[3] Corrado L, Fingleton B, “Multilevel Modelling with Spatial Effect”, University of Strathclyde press, Glasgow, 2011, pp. 1-21.

[4] Goodchild MF, “Geographical Information Science”, Geographical Information System, Vol. 6, pp. 31-45, Feb 1992.

[5] Griffith AD, “Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization”, Springer, New York, 2003, pp 31-130.

[6] LeSage JP, “The Theory and Practice of Spatial Econometrics”, Department of Economics University of Toledo, Toledo, 1999, pp. 1-20 .

[7] Park YM, KIM Y, “A Spatially Filtered Multilevel Model to Account for Spatial Dependency: Application to Self-Rated Health Status in South Korea”, International Journal of Health Geographics, Vol. 13, pp. 1-6, Feb 2014.

[8] Pierawan AC, Tampubolon G, “Spatial Dependence Multilevel Model of Well-Being Across Region in Europe”, Applied Geography, Vol. 47, pp. 168-176, Feb 2014.

[9] Snijders TAB, Bosker RJ, “Multilevel Analysis: An introduction to basic and advance multilevel modeling”, Sage Publication, London, 1999, pp 38-83.

[10] Tiefelsdorf M, Grifith DA, “Semiparametric filtering of spatial autocorrelation: the eigenvector approach”. Environment and Planning, Vol. 39, pp. 1193-1221, Mar 2007.

[11] Xu H, “Compare Spatial and Multilevel Regression Models for Binary Outcome in Neighborhood Study”,Sociological Methodology, Vol 44, pp. 229-272, Aug 2014.