CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 5(1)(2017), Pages 8-14
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Estimation of Geographically Weighted Regression Case Study on Wet Land Paddy Productivities in Tulungagung Regency

Danang Ariyanto, Henny Pramoedyo, Suci Astutik

Department of Statistics, Faculty of Mathematics and Natural Sciences
Brawijaya University, Malang

Email: danangariyanto75@gmail.com, pramoedyohp@yahoo.com, suci_sp@yahoo.com

ABSTRACT

Regression is a method connected independent variable and dependent variable with estimation parameter as an output. Principal problem in this method is its application in spatial data. Geographically Weighted Regression (GWR) method used to solve the problem. GWR is a regression technique that extends the traditional regression framework by allowing the estimation of local rather than global parameters. In other words, GWR runs a regression for each location, instead of a sole regression for the entire study area. The purpose of this research is to analyze the factors influencing wet land paddy productivities in Tulungagung Regency. The methods used in this research is GWR using cross validation bandwidth and weighted by adaptive Gaussian kernel function. This research using four variables which are presumed affecting the wet land paddy productivities such as: the rate of rainfall(X_1), the average cost of fertilizer per hectare(X_2), the average cost of pesticides per hectare(X_3) and Allocation of subsidized NPK fertilizer of food crops sub-sector(X_4). Based on the result, X_1, X_2, X_3 and X_4T has a different effect on each District. So, to improve the productivity of wet land paddy in Tulungagung Regency required a special policy based on the GWR model in each district.

Keywords: spatial data, geographically weighted regression, GWR, cross validation, adaptive gaussian kernel

INTRODUCTION

The conventional spatial analysis techniques use a single equation to assess the overall relationships between the dependent and independent variables across space, known as a global analytic approach. One important assumption underlying this approach is that the relationships of interest are stationary or homogeneous spatially. While the global perspective is effective in handling spatial dependence and generating less unbiased estimates (than the non-spatial modeling), it is not capable of exploring spatial non-stationarity (or heterogeneity) or identifying place-specific associations [1].

Geographically weighted regression (GWR) is a local spatial statistical technique used to analyze spatial non-stationarity, defined as when the measurement of relationships among variables differs from location to location. Unlike conventional regression, which produces a single regression equation to summarize global relationships among the explanatory and dependent variables, GWR generates spatial data that express the spatial variation in the

Submitted: 26 July 2016 Reviewed: 18 October 2017 Accepted: 2 November 2017
DOI: http://dx.doi.org/10.18860/ca.v5i1.4305

relationships among variables [2]. This approach includes locational information and smoothing techniques into regression models. In contrast to the global approach, GWR has proved to be a useful local spatial analysis tool that helps researchers to generate nuanced insights into existing literature [1]. In this research, the parameter estimation on the GWR method requires a weights matrix that calculated by adaptive gaussian kernel function. The data weighting is according to the proximity of the i-th observation location. Cross validation is used for estimating the kernel bandwidth. The case study investigated the productivity of wet land paddy in Tulungagung Regency, there are 18 districts in Tulungagung Regency and 1 district has no wet land paddy productivity. This research using four variables which are presumed affecting the wet land paddy productivities such as: the rate of rainfall(X_1), the average cost of fertilizer per hectare(X_2), the average cost of pesticides per hectare(X_3) and Allocation of subsidized NPK fertilizer of food crops sub-sector(X_4). The final result of this GWR method will be obtained productivity model of wet land paddy in Tulungagung regency. In addition, the mapping of paddy productivity per district is expected to be useful and add information especially in agriculture to increase wet land paddy productivity in Tulungagung Regency.

METHODS

Geographically weighted regression is an extension of the traditional multiple linear regression toward a local regression, in which regression coefficients are specific to a location rather than being global estimates. the specification of a basic GWR model is:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i\,; \quad i = 1,2,\dots,n \tag{1}$$

where yi is the dependent variable at location i, x_{ik} is the value of the kth explanatory variable at location i, the $\beta_k(u_i, v_i)x_{ik}$ is the local regression coefficient for the kth explanatory variable at location i, $\beta_0(u_i, v_i)$ is the intercept parameter at location i, and ε_i is the random disturbance at location i, which may follow an independent normal distribution with zero mean and homogeneous variance [3].

To facilitate the exposition, it is convenient to express the GWR model in matrix notation:

$$\boldsymbol{Y_w} = \boldsymbol{X_w\beta_w} + \boldsymbol{\varepsilon_w} \tag{2}$$

Weighted Least Square (WLS) is used for Geographically Weighted Regression parameter estimation, so:

$$\begin{aligned}
L &= \boldsymbol{\varepsilon_w}^{\boldsymbol{T}}\boldsymbol{\varepsilon_w} \\
&= (\boldsymbol{Y_w} - \boldsymbol{X_w\beta_w})^{\boldsymbol{T}}(\boldsymbol{Y_w} - \boldsymbol{X_w\beta_w}) \\
&= \boldsymbol{Y_w}^{\boldsymbol{T}}\boldsymbol{Y_w} - \boldsymbol{\beta_w}^{\boldsymbol{T}}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{Y_w} - \boldsymbol{Y_w}^{\boldsymbol{T}}\boldsymbol{X_w\beta_w} + \boldsymbol{\beta_w}^{\boldsymbol{T}}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{X_w\beta_w} \\
&= \boldsymbol{Y_w}^{\boldsymbol{T}}\boldsymbol{Y_w} - \boldsymbol{2\beta_w}^{\boldsymbol{T}}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{Y_w} + \boldsymbol{\beta_w}^{\boldsymbol{T}}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{X_w\beta_w}
\end{aligned} \tag{3}$$

As we learned in calculus, a univariate optimization involves taking the derivative and setting equal to 0. This gives us,

$$\frac{\partial L}{\partial \boldsymbol{\beta_w}} = \frac{\partial(\boldsymbol{Y_w}^{\boldsymbol{T}}\boldsymbol{Y_w} - \boldsymbol{2\beta_w}^{\boldsymbol{T}}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{Y_w} + \boldsymbol{\beta_w}^{\boldsymbol{T}}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{X_w\beta_w})}{\partial \boldsymbol{\beta_w}} = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\beta_w}} = -\boldsymbol{2X_w}^{\boldsymbol{T}}\boldsymbol{Y_w} + \boldsymbol{2X_w}^{\boldsymbol{T}}\boldsymbol{X_w}\boldsymbol{\widehat{\beta}_w} = 0$$

$$\frac{\partial L}{\partial \beta_w} = -\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{Y_w} + \boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{X_w}\boldsymbol{\widehat{\beta}_w} = 0$$

$$\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{X_w}\boldsymbol{\widehat{\beta}_w} = \boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{Y_w}$$

$$\boldsymbol{\widehat{\beta}_w} = (\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{X_w})^{-1}\boldsymbol{X_w}^{\boldsymbol{T}}\boldsymbol{Y_w} \qquad (4)$$

If $\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Y} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X\beta} + \boldsymbol{W}^{\frac{1}{2}}\varepsilon$ is equal to $\boldsymbol{Y_w} = \boldsymbol{X_w\beta_w} + \boldsymbol{\varepsilon_w}$ so,

$$= \left[(\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X})^{\boldsymbol{T}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}\right]^{-1}(\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X})^{\boldsymbol{T}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Y}$$

$$= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Y}$$

$$\boldsymbol{\widehat{\beta}_w} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{Y} \qquad (5)$$

For each i-th point, the Geographically Weighted Regression model parameter estimation is performed by matrix operation:

$$\boldsymbol{\widehat{\beta}}(u_i, v_i) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)\,\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)\,\boldsymbol{Y} \qquad i = 1,2,\dots,n \qquad (7)$$

with,

$$\boldsymbol{W_{ij}}(u_i, v_i) = \begin{bmatrix} \boldsymbol{w}_{i1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{w}_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{w}_{in} \end{bmatrix}$$

The first step to estimate parameters in GWR, it is important to decide the spatial weighting matrix, which can be calculated by different methods. One method is to specify **W_{ij}** as a continuous and monotonic decreasing function of distance d_{ij} between points i and j. For adaptive kernel size, the weight of each point can be calculated by applying the Gaussian function [3]:

$$w_{ij}(u_i, v_i) = exp(-\left(\frac{d_{ij}}{h_{i(q)}}\right)^2) \qquad (8)$$

where $w_{ij}(u_i, v_i)$ is the weight of location j in the space at which data are observed for estimating the dependent variable at location i, and$h_{i(q)}$ is referred as a bandwidth. d_{ij} is eucledian distance between points i and j, $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$. Bandwidth is used to specifies how the extent of the kernel should be determined. It controls the degree of smoothing in the model. Cross-validation (CV) is an iterative process that searches for the kernel bandwidth that minimizes the prediction error of all the $y(s)$ using a subset of the data for prediction [1].

$$CV(h) = \sum_{i=1}^{n}(y_i - \hat{y}_{\neq i}(h))^2 \qquad (9)$$

where $\hat{y}_{\neq i}(h)$ is the predicted value of observation i with calibration location i left out of the estimation dataset.

The second step is testing for spatial heterogeneity. Spatial heterogeneity indicates the variation between location. So, each location has different relationship structures and parameters. Spatial data heterogeneity can be tested using the Breach-Pagan (BP) test [4]:

H₀: $\sigma^2_{(u_1,v_1)} = \sigma^2_{(u_2,v_2)} = \cdots = \sigma^2_{(u_n,v_n)} = \sigma^2$, $\sigma^2_{(u_i,v_i)} = \sigma^2$ (there is no spatial heterogeneity)

H₁: at least one $\sigma^2_{(u_i,v_i)} \neq \sigma^2$ (there is spatial heterogeneity)

$$BP = \left(\frac{1}{2}\right)\boldsymbol{f}^T\boldsymbol{Z}\,(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{f} + \left(\frac{1}{T}\right)\left[\frac{e^T W e}{\boldsymbol{\sigma^2}}\right]^{\boldsymbol{2}} \sim \chi^2_{(k+1)} \qquad (10)$$

Where $\boldsymbol{f_i} = \frac{e_i}{\sigma^2} - \mathbf{1}, \boldsymbol{e_i}$ is a vector of OLS residuals, $\boldsymbol{\sigma^2}$ is the variance based on OLS residual, T= Trace **[W$^{\mathbf{T}}$W + W$^{\mathbf{2}}$]** and $\boldsymbol{Z}$ is an N by (k+1) matrix of normal standard score (z).

The third step is testing parameters of GWR model partially using t test with hypothesis as follows [5]:

H$_{0:}$ $\beta_j(u_i, v_i) = 0$

H$_{1:}$ $\beta_j(u_i, v_i) \neq 0\ ;\ j = 1,2, \dots, k$

t test statistic can be written as follows:

$$\frac{\widehat{\beta}_k(u_i,v_i)}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{(n-k-1)} \tag{11}$$

Where c_{jj} is a diagonal element of the **CC$^{\mathbf{T}}$** matrix, with $\boldsymbol{C} = (\boldsymbol{X^T}\ \boldsymbol{W}(\boldsymbol{u_i}, \boldsymbol{v_i})\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{W}(\boldsymbol{u_i}, \boldsymbol{v_i})$**.**

The fourth step is testing parameter of GWR model simultaneously, the hypothesis is:

H$_0$: $\beta_j(u_i, v_i) = \beta_j$, where $j = 1, 2, \dots k.$

H$_1$: at least one $\beta_j(u_i, v_i)$ has a relation with location (u_i, v_i)

The statistic test is:

$$\frac{SSE(H_1) / \left[\frac{\delta_1^2}{\delta_2}\right]}{SSE(H_0) / n-k-1} \square\ F^*_{(\frac{\delta_1^2}{\delta_2}, n-k-1)} \tag{12}$$

where:

$SSE(H_0) = \mathbf{y^T(I - L)^T(I - L)y}$

$SSE(H_1) \ = \mathbf{y^T(I - S)y}$

$\delta_1 = trace\left\{(\mathbf{I - L})^{\mathrm{T}}(\mathbf{I - L})\right\}$

$\delta_2 = trace\left\{(\mathbf{I - L})^{\mathrm{T}}(\mathbf{I - L})\right\}^2$

$$\mathbf{L}_{(n\times n)} = \begin{pmatrix} x_1^T\left[X^TW(u_1,v_1)X\right]^{-1}X^TW(u_1,v_1) \\ x_2^T\left[X^TW(u_2,v_2)X\right]^{-1}X^TW(u_2,v_2) \\ \vdots \\ x_n^T\left[X^TW(u_n,v_n)X\right]^{-1}X^TW(u_n,v_n) \end{pmatrix}$$

$\mathbf{S} \ = \mathbf{X(X^TX)^{-1}X^T}$

$\mathbf{I}$ = identity matrix ordo n

RESULTS AND DISCUSSION

The results of Breusch-Pagan (BP) test in wet land paddy productivities in Tulungagung Regency is:

$$BP = \left(\frac{1}{2}\right)\boldsymbol{f}^{\,T}\boldsymbol{Z}\,(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{f} + \left(\frac{\mathbf{1}}{T}\right)\left[\frac{e^T\boldsymbol{W}e}{\boldsymbol{\sigma^2}}\right]^{\mathbf{2}} = 7438{,}636$$

Because BP test > $\chi^2_{(0,05;\ 5)} = 11{,}070$, the decision is Reject H_0. There is spatial heterogeneity in wet land paddy productivities in Tulungagung Regency, so we can use geographically weighted regression to estimate parameter model wet land paddy productivities in Tulungagung Regency.

The next step is to determine the parameter estimation of the GWR model based on the equation 7. Table 1 show the parameters estimation GWR model.

Table 1. Parameters estimation GWR model

No	district	β_0	β_1	β_2	β_3	β_4
1	Besuki	26,553	-0,02175	0,0000165	0,0000363	-0,0015
2	Bandung	40,055	0,01395	0,0000014	0,0000374	-0,00065
3	Pakel	27,372	-0,02315	0,0000179	0,0000311	-0,0016
4	Campurdarat	35,107	-0,01342	0,0000077	0,0000396	-0,00154
5	Kalidawir	42,804	0,01075	-0,0000009	0,0000370	-0,00036
6	Pucanglaban	59,237	0,12682	-0,0000260	0,0000264	0,007461
7	Rejotangan	55,276	0,08255	-0,0000189	0,0000297	0,005735
8	Ngunut	40,515	0,03892	-0,0000185	0,0000672	0,010068
9	Sumbergempol	41,269	0,05628	-0,0000239	0,0000751	0,011815
10	Boyolangu	40,946	0,06148	-0,0000224	0,0000706	0,0117
11	Tulungagung	41,456	0,05585	-0,0000207	0,0000664	0,011093
12	Kedungwaru	41,122	0,05185	-0,0000186	0,0000621	0,01075
13	Ngantru	42,624	0,03504	-0,0000128	0,0000508	0,007392
14	Karangrejo	45,889	0,04003	-0,0000111	0,0000406	0,005076
15	Kauman	59,895	0,10582	-0,0000251	0,0000278	0,0076
16	Gondang	59,386	0,17687	-0,0000308	0,0000234	0,008816
17	Pagerwojo	60,541	0,16031	-0,0000305	0,0000244	0,009065
18	Sendang	41,348	0,02341	-0,0000006	0,0000365	-0,000099

The GWR model for district Besuki based on Table 1 can be written as follows:

$$\hat{y}_1 = 26{,}553 - 0{,}0217x_1 + 0{,}0000165x_2 + 0{,}0000363x_3 - 0{,}00150x_4 \qquad (13)$$

Other GWR models for the other districts has the same way as in equation 8 based on Table 1. Testing parameters of GWR model simultaneously conducted to determine the effect of weighting in the process of parameter estimation on the case of paddy productivity in Tulungagung Regency. Results of simultan parameter test based on equation 12. F test is 8,904. Value of F test is greater than $F_{(0,05;10,3)} = 8{,}785$. This shows that simultaneously the predictor variables X1, X2, X3 and X4 have significant effect spatially on response variable. The GWR model of each district is formed on the basis of influential parameters, so partial parameter testing is performed by Table 2 using t test.

Table 2. t test statistics for parameter of GWR model

No	District	t test statistic β_0	t test statistic β_1	t test statistic β_2	t test statistic β_3	t test statistic β_4
1	Besuki	5,429	-1,23768	2,974011	7,616609	-1,81654
2	Bandung	12,447	0,907966	0,456556	11,73849	-0,89729
3	Pakel	5,814	-1,27889	2,910577	4,486399	-1,91835
4	Campurdarat	10,604	-0,82625	2,479856	11,72245	-1,91886
5	Kalidawir	18,321	0,886951	-0,41056	13,04989	-0,58983
6	Pucanglaban	13,264	5,188448	-5,04069	7,326105	4,716729
7	Rejotangan	14,778	3,647647	-4,34952	8,891485	3,844718
8	Ngunut	19,811	2,362104	-8,07605	15,72467	11,62082
9	Sumbergempol	14,223	3,046957	-5,58308	10,28998	10,56913
10	Boyolangu	14,431	3,073174	-6,91394	12,4301	11,11775
11	Tulungagung	19,018	3,071231	-7,73936	13,62258	11,04025
12	Kedungwaru	19,701	2,941217	-8,31189	14,96989	11,01064
13	Ngantru	18,828	2,111468	-7,95199	18,32336	9,755659
14	Karangrejo	17,363	2,105942	-6,30701	13,46622	5,196659
15	Kauman	13,516	3,402	-4,47746	6,19037	3,55978
16	Gondang	5,404	1,296335	-4,49924	2,7131	3,188156
17	Pagerwojo	9,567	2,655057	-4,6735	5,152779	3,895126
18	Sendang	12,836	1,525182	-0,21392	11,48837	-0,13862

Significant if t test > $t_{(0,05;13)}$ = 2,16

Based on Table 2, The Farmers need to add the average cost of pesticides per hectare(X_3) to each district because the effect is positive significant to the wet land paddy productivity. In the districts of Ngantru, Karangrejo and Gondang need to add the average cost of fertilizer per hectare(X_2) because it has a significant positive effect on paddy productivity. The addition of X_2 and X_3 is not continuously. There are 12 districts that need to be added by the allocation of subsidized NPK fertilizer of food crops sub-sector(X_4) because it proved to have a significant positive effect on wet land paddy productivity. The district is grouping according to the variables that significantly affect the GWR model, the result is listed at Tabel 3.

Table 3. The Grouping according to the significant variables of GWR model

district	Group	Significant variables
Pucanglaban, Rejotangan, Ngunut, Sumbergempol, Boyolangu, Tulungagung, Kedungwaru, Kauman, Pagerwojo	Yellow	X_1, X_2, X_3 and X_4
Ngantru, Karangrejo, Gondang	Red	X_2, X_3 and X_4
Besuki, Pakel, Campurdarat	Purple	X_2 and X_3
Bandung, Kalidawir, Sendang	Green	X_3

Based on Table 3, a spread pattern of variables that has a significant effect on paddy productivity by districts in Tulungagung Regency presented in Figure 1.

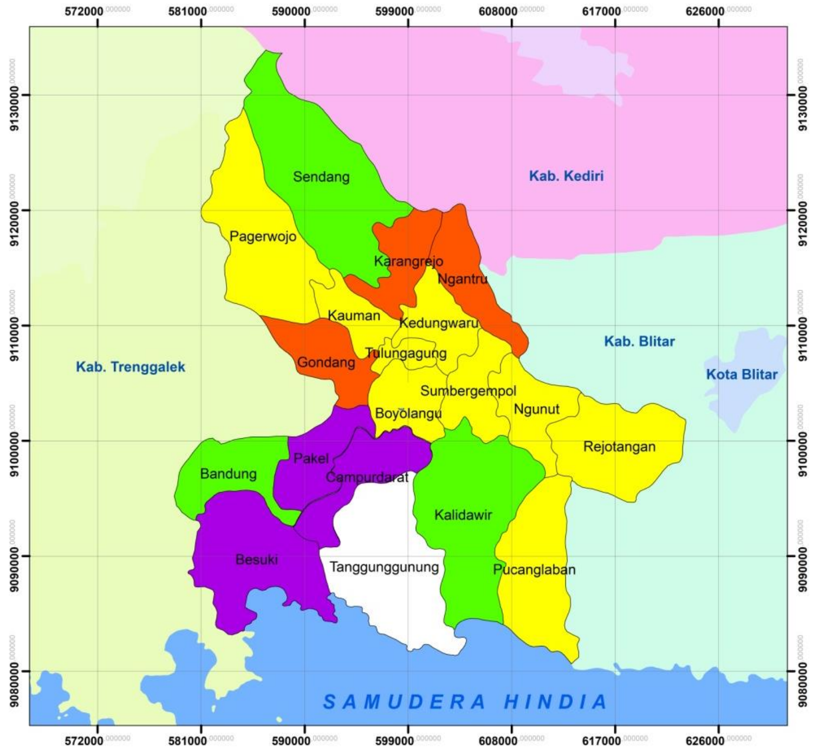

Figure 1 . The spread pattern of variables that significantly affected by district on wet land paddy productivities in Tulungagung regency

The yellow group is the group where the four variables, X_1, X_2, X_3 dan X_4 is significant to the productivity of wet land paddy. The yellow group consisted of 9 districts. Only Variable X_2, X_3 and X_4 had a significant effect on the red group, there were 3 districts in the red group. Purple group is a group with 2 variables that have significant effect that is X_2 and X_3. There are 3 districts that enter the purple group. The last group is green group where only X_3 has significant effect. Green group consists of 3 districts. The District of Tanggunggunung has no wet land paddy productivity so the colour is white.

CONCLUSION

Based on the result, X_1, X_2, X_3 and X_4 has a different effect on each district. So, to improve the productivity of wet land paddy in Tulungagung Regency required a special policy based on the GWR model in each district.

REFERENCES

[1] A. S. Fotheringham, C. Brundson and M. Charlthon, Geographically Weighted Regression : The Analysis of Spatially Varying Relationships, United Kingdom: John Wiley and Sons, Ltd, 2002.

[2] C. L. Mei, Geographycally Weighted regression Technique for Spatial Data Analysis, China: Jiaotong University, 2005.

[3] M. M. F. a. A. Getis, Handbook of Applied Spatial Analysis, Berlin: Heidelberg and New York, 2010, pp. 255-278.

[4] L. Anselin, Spatial Econometrics : Methods and Models, Netherlands: Kluwer Academic Publishers, 1988.

[5] Y. Leung, C. L. Mei and W. X. Zhang, "Statistical Test for Spatial Non-Stationary Based on the Geographically Weighted Regression Model," *Environment and Planning A,* vol. 32, pp. 9-32, 2000.