VOL. 12, NO. 15, AUGUST 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

MIXED GEOGRAPHICALLY WEIGHTED REGRESSION USING ADAPTIVE BANDWIDTH TO MODELING OF AIR POLLUTER STANDARD INDEX

Dwi Ispriyanti[1], Hasbi Yasin[1], Budi Warsito[1], Abdul Hoyyi[1] and Kukuh Winarso[2]
[1]Departement of Statistcs, Diponegoro University, Semarang, Indonesia
[2]Department of Industrial Engineering, Faculty of Engineering, Trunojoyo Madura University, Indonesia
E-Mail: dwiispriyanti@yahoo.com

ABSTRACT

Air pollution is one of the most concerned problems on earth today. It is closely related with and mostly generated from the transportation and industrialization sectors, as well as from the environmentally degrading effect of the urban physical development. Air pollution promotes the lower level of air quality, which in turn promotes the greater risk on health, especially that of the human being. This research aims to aid the government in the policy making process related to air pollution mitigation by developing a standard index model for air polluter (Air Polluter Standard Index - APSI) based on the Mixed Geographically Weighted Regression (MGWR) approach using the adaptive bandwidth. The adaptive bandwidth kernel has different bandwidth value in each observation location. Akaike Information Criterion-corrected (AICc) value is used to choose the most optimum bandwidth. The Monte Carlo Simulation is used to tests for regression coefficient non-stationarity. In this research, we also consider seven variables that are directly related to the air pollution level, which are the traffic velocity, the population density, the business center aspect, the air humidity, the wind velocity, the air temperature, and the area size of the urban forest. Based on AICc and MSE value it is know that the MGWR model with adaptive bisquare kernel is the best bandwidth to analyze this model.

Keywords: adaptive bisquare kernel, air polluter, APSI, MGWR, Monte Carlo simulation.

1. INTRODUCTION

Air pollution is a real problem that threatens the environment and even threaten human life. It is characterized by a decrease in air quality, especially in the big cities in recent years. Factors to be a major source of air pollution in large cities is transportation-engined vehicles, the exhaust gas industries, population density, shopping centers, air humidity, air temperature and wind speed, and so on. Factors which may prevent or inhibit the emergence of air pollution is the presence of many green areas and trees in city parks (Atash, 2007) and (Fahimi, *et al.*, 2012). Elements of pollutant major accordance with air pollution index (ISPU) is Carbon Monoxide (CO), particulate matter (PM), sulfur dioxide (SO2) Nitrogen Dioxide (NO2) and ozone (O3) (Fahimi, *et al.*, 2012).

The elements of these pollutants are extremely dangerous when exposed to humans for a long time and continuously. It is characterized by increased disease caused by air pollution, among others, cardiovascular, asthma, allergic, immunological disorders and cancer, both developing and developed countries (Fahimi, *et al.*, 2012). The study of air pollution showed a parallel relationship between the number of people with illnesses caused by air pollution with an increase in industrialization and urbanization in developing countries where air pollution levels are very high (Ebtekar, 2006).

The causes and effects of air pollution has been linked to the location, meaning between spatial location will be different causes and effects. Demographically, the potential impacts and causes of air pollution will differ between regions, in addition to the impact and causes of air pollution cannot use a global approach, because by using a global approach there are local variations invisible influence (Gilbert, 2011) & (Robinson, 2011). Spatial regression method frequently used is Geographically Weighted Regresssion (GWR), which is a regression method involving the effect of the location into the predictor (Fotheringham, *et al.*, 2002). In the linear regression model generated only parameter estimator that apply globally, while in the GWR models generated model parameter estimator that is local to each observation location. Mixed geographically Weighted Regression (MGWR) is a combination of global linear regression model with the GWR model. So that the model will be generated MGWR estimator parameters are global and some others are localized in accordance with the location of observations (Purhadi and Yasin, 2012).

MGWR parameter estimation is done by giving a different weighting for each location where the data is collected. In determining the value of the kernel function can be divided into two types of calculations, namely the fixed kernel and adaptive kernel bandwidth. Fixed kernel is the same bandwidth at all points of observation location, while adaptive kernel is the bandwidth that has different values for each observation location. The adaptive kernel was applied in this study in order to fix the kernel and to get better calculation accuracy.

2. RESEARCH OBJECTIVE

This study aims to model Air Polluter Standard Index (APSI) with MGWR using the adaptive bandwidth based on AICc approach. The case study in this research use Air Polluter Standard Index – APSI in Surabaya City in five location as the response variable Y, while the predictor variables were the air temperature (X_1), the wind velocity (X_2), the air humidity (X_3), the traffic velocity

(X_4), the area size of the urban forest (X_5), the population density (X_6), and the business center aspect (X_7)

3. MATERIAL AND METHODS

3.1 Linear regression

Linear regression is a method that models the relationship between response variables and predictor variables. Linear regression model for p predictor variables are generally written as follows:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \quad (1)$$

where $i = 1, 2, \ldots, n$; $\beta_0, \beta_1, \ldots, \beta_p$ are the model parameters and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are error term with mean zero and common variance σ^2.

Estimation of regression parameters are done by Ordinary Least Squares (OLS) method. Testing of parameter regression model using the F distribution approach and the partial use of the t distribution approach (Rencher, 2000).

3.2 Geographically Weighted Regression (GWR)

GWR model represents the development of a global regression model where the basic idea is taken from the non-parametric regression (Mei *et al*., 2006). This model is a locally linear regression that that produces the estimator model parameters that are local to each point or location where the data is collected. GWR model can be written as follows:

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

where:

y_i : observation of respons y

(u_i, v_i) : coordinat point (*longitude*, *latitude*)

$\beta_k(u_i, v_i)$: p unknown functions of geographical locations (u_i, v_i); $k = 0,1,\ldots,p$

x_{ik} : explanatory variable at location (u_i, v_i)

ε_i :error term with mean zero and common variance σ^2

Estimation of GWR model parameter is using the Weighted Least Squares (WLS) method that gives a different weighting for each observation. So the estimator of model parameter for each location is:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{y} \quad (3)$$

The GWR method provides a feasible way for testing a global linear regression relationship for spatial data. This amounts to test the following hypotheses:

$H_0 : \beta_k(u_i, v_i) = \beta_k$ for each $k = 0,1,2,\cdots, p,$ and $i = 1,2,\cdots, n$

(there is no significant difference between the global regression model and GWR)

H_1 : at least there is a $\beta_k(u_i, v_i) \neq \beta_k, k = 0,1,2,\cdots, p$

(there is a significant difference between the globalregression model and GWR).

The test statistic given by Leung, *et al.* (2000):

$$F_{test} = \frac{\mathbf{y}^T \left[(\mathbf{I}-\mathbf{H}) - (\mathbf{I}-\mathbf{L})^T (\mathbf{I}-\mathbf{L})\right] \mathbf{y} / \tau_1}{\mathbf{y}^T (\mathbf{I}-\mathbf{H})^T (\mathbf{I}-\mathbf{H}) \mathbf{y} / (n-p-1)} \quad (4)$$

Reject H_0 if $F_{test} \geq F_{\alpha, df_1, df_2}$ where $df_1 = \tau_1^2 / \tau_2$, $df_2 = (n-p-1)$

$\tau_i = tr\left(\left[(\mathbf{I}-\mathbf{H}) - (\mathbf{I}-\mathbf{L})^T (\mathbf{I}-\mathbf{L})\right]^i\right)$, $i = 1,2$,

$$\mathbf{L} = \begin{pmatrix} \mathbf{x}_1^T \left(\mathbf{X}^T \mathbf{W}(u_1, v_1) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_1, v_1) \\ \mathbf{x}_2^T \left(\mathbf{X}^T \mathbf{W}(u_2, v_2) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_2, v_2) \\ \vdots \\ \mathbf{x}_n^T \left(\mathbf{X}^T \mathbf{W}(u_n, v_n) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_n, v_n) \end{pmatrix} \text{ and } \mathbf{H} = \mathbf{X}\left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T .$$

Testing parameters are partially done with the hypothesis as follows:

$H_0 : \beta_k(u_i, v_i) = 0$

$H_1 : \beta_k(u_i, v_i) \neq 0$ where $k = 1,2,\cdots, p$

Test statistic given by:

$$T = \frac{\hat{\beta}_k(u_i, v_i)}{\hat{\sigma}\sqrt{c_{kk}}}$$

where $\hat{\sigma}^2 = \dfrac{\mathbf{y}^T (\mathbf{I}-\mathbf{L})^T (\mathbf{I}-\mathbf{L}) \mathbf{y}}{\delta_1}$ and c_{kk} is the k-th diagonal element of the matrix $\mathbf{C}\mathbf{C}^T$ where $\mathbf{C} = \left(\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i)$.

Reject H_0 or in other words the $\beta_k(u_i, v_i)$ is a significant parameters of the model when $|T| > t_{\alpha/2, df}$, where $df = \left[\delta_1^2 / \delta_2\right]$

and $\delta_i = tr\left(\left[\left(\mathbf{I}-\mathbf{L}\right)^T\left(\mathbf{I}-\mathbf{L}\right)\right]^i\right),\ i=1,2$.

3.3 Monte Carlo tests for regression coefficient non-stationarity

For a mixed GWR, difficulties arise when deciding whether a relationship should be fixed globally or allowed to vary locally. Here Fotheringham *et al.* (2002) adopt a stepwise procedure; where all possible combinations of global and locally-varying relationships are tested, and an optimal mixed model is chosen according to minimize AIC value. This approach is comprehensive, but computationally expensive, and is utilised in the GWR 4.0 executable software (Nakaya, *et. al.*, 2009). Alternatively, a Monte Carlo approach can be used to test for significant (spatial) variation in each regression coefficient (or relationship) from the basic GWR, where the null hypothesis is that the relationship between dependent and independent variable is constant (Fotheringham *et al.*, 2002). The procedure is analogous to that presented for the local eigen values of a GW PCA, where for the basic GWR the true variability in each local regression coefficient is compared to that found from a series of randomised data sets. If the true variance of the coefficient does not lie in the top 5% tail of the ranked results, then the null hypothesis can be accepted at the 95% level; and the corresponding relationship should be globally-fixed when specifying the mixed GWR. Observe, that if all relationships are viewed as non-stationary, then the basic GW regression should be preferred. Conversely, if all relationships are viewed as stationary, then the standard global regression should be preferred (Lu, *et al.*, 2014). Advances on the mixed GWR model, where the relationships can be allowed to vary at different rates across space can be found in Yang *et al.* (2012).

3.4 Mixed Geographically Weighted Regression (MGWR)

A key assumption for this basic GWR is that the local coefficients vary at the same scale and rate across space. However, some coefficients (and relationships) may be expected to have different degrees of variation over the study region. In particular, some coefficients (and relationships) are viewed as constant (or stationary) in nature, whilst others are not. For these situations, a Mixed GWR can be specified (Fotheringham, *et al.*, 2002). This semi-parametric model treats some coefficients as global (and stationary), whilst the rest are treated as local (and non-stationary), but with the same rate of spatial variation. In a vector-matrix notation, the MGWR model can be rewritten as (Lu, *et al.*, 2014).

$$\mathbf{y} = \mathbf{X}_a\mathbf{a} + \mathbf{X}_b\mathbf{b} + \boldsymbol{\varepsilon} \tag{5}$$

where y is the vector of dependent variables; X_a is the matrix of globally-fixed variables; a is the vector of k_a global coefficients; X_b is the matrix of locally-varying variables; and b is the matrix of local coefficients.

If we define the hat matrix for the global regression part of the model, as S_a; and that for the GW regression part, as S_b; then equation (4) can be rewritten as

$\hat{\mathbf{y}} = \hat{\mathbf{y}}_a + \hat{\mathbf{y}}_b$ where $\hat{\mathbf{y}}_a = \mathbf{S}_a\mathbf{y}$ and $\hat{\mathbf{y}}_b = \mathbf{S}_b\mathbf{y}$.

Thus the calibration procedure can be briefly described in the following six steps:

Step-1: Supply an initial value for $\hat{\mathbf{y}}_a$, say as $\hat{\mathbf{y}}_a{}^{(0)}$, practically by regressing $\mathbf{X}_a$ on $\mathbf{y}$ using ordinary least square (OLS).

Step-2: Set $i = 1$.

Step-3: Set $\hat{\mathbf{y}}_b^{(i)} = \mathbf{S}_b\left[\mathbf{y} - \hat{\mathbf{y}}_a^{(i-1)}\right]$

Step-4: Set $\hat{\mathbf{y}}_a^{(i)} = \mathbf{S}_a\left[\mathbf{y} - \hat{\mathbf{y}}_b^{(i)}\right]$

Step-5: Set $i = i + 1$

Step-6: Return to Step 3 unless $\hat{\mathbf{y}}^{(i)} = \hat{\mathbf{y}}_a^{(i)} + \hat{\mathbf{y}}_b^{(i)}$ converges to $\hat{\mathbf{y}}^{(i-1)}$.

3.5 Adaptive Kernel Functions

Weighting function that used to estimate the parameters in the MGWR model is the bisquare adaptive kernel functions (Fotheringham *et al.*, 2002), which can be written as follows:

$$w_j\left(u_i, v_i\right) = \begin{cases} \left(1-\left(d_{ij}/h_i\right)^2\right)^2, & \text{if } d_{ij} \le h \\ 0, & \text{if } d_{ij} > h \end{cases}$$

where d_{ij} denotes the distance between the location $\left(u_i, v_i\right)$ to location $\left(u_j, v_j\right)$ and h_i are non negative parameters are known and are usually called smoothing parameter (*bandwidth*) for location $\left(u_i, v_i\right)$. So $\mathbf{W}\left(u_i, v_i\right) = \text{diag}\left(w_1\left(u_i, v_i\right), w_2\left(u_i, v_i\right), \cdots, w_n\left(u_i, v_i\right)\right)$ and one of the methods that used to select the optimum bandwidth is the Akaike Information Criterion-corrected (AICc).

3.6 Selection of the best model

The method that is used to select the best model is Akaike Information Criterion (AIC) which is defined as follows:

$$AIC_c = 2n\ln\left(\hat{\sigma}\right) + n\ln\left(2\pi\right) + n\left\{\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right\} \tag{6}$$

where:

$\hat{\sigma}$: The estimator of standard deviation of the error

S : Matrix projection where $\hat{\mathbf{y}} = \mathbf{Sy}$

The best model selection is done by determining the model with the smallest AIC value (Nakaya, *et al.*, 2005).

3.7 Methods

The procedure to modeling Air Polluter Standard Index (APSI) with MGWR approach using the adaptive bandwidth described in the following steps:

a) Describe the data as a preliminary to determine the spread of Air Polluter Standard Index (APSI).
b) Perform the global linear regression model
c) Perform the basic GWR model
d) Monte Carlo tests for regression coefficient non-stationarity
e) Select the global regression part and the GWR part based on Monte Carlo test
f) Perform the MGWR model
g) Compare the global regression model, GWR and MGWR

4. RESULTS

Using the GWmodel R Package (Lu, *et al.*, 2016) based on AICc approach, the optimum adaptive bandwidth for each location are shown in Table-1.

Table-1. Adaptive bandwidth using Bisquare Kernel.

Location	Bandwidth
SUF 1	0.03723238
SUF 3	0.03530354
SUF 4	0.05280918
SUF 5	0.03723238
SUF 6	0.05183252

Then, using this bandwidth we estimate the GWR model. The goodness of fits for GWR model can be stated by the following hypothesis:

H_0 : $\beta_k\left(u_i, v_i\right) = \beta_k$ $k = 0,1,2,\cdots,q,$ and $i = 1,2,\cdots,n$

(GWR model is not significantly different from the Regression model)

H_1 : at least one $\beta_k\left(u_i, v_i\right) \neq \beta_k$

(GWR model is significantly different from the regression model)

Table-2. Goodness of fits of GWR Model.

Source of error	Sum of squares	Degree of freedom	F	p-value
Improvement	14886.5	12.09	4.598	0,00
GWR	122631.6	404.46		
Regression	137518.1	412.00		

Table-2 shows that the F test statistical value is 4.598 (p-value = 0.000). Using the significance value (α) of 5%, we must reject H_0, and conclude that the GWR model with adaptive bandwidth is significantly different from the regression model. Therefore, we can further conclude that the GWR model is more proper to model the Air Polluter Standard Index (APSI). This means that the location element is influential in the APSI modeling. The statistics of local parameter in GWR model with adaptive bisquare kernel shown in Table-3.

Table-3. Statistic of GWR model with Adaptive Bisquare Kernel.

Coefficient	Min	Max	Median
β_0	42.32	71.81	50.71
β_1	-3.16	7.10	2.68
β_2	-5.84	-0.66	-4.22
β_3	-4.28	7.64	-1.04
β_4	3.12	19.26	18.96
β_5	-135.70	1.52	-42.02
β_6	2.01	34.35	11.87
β_7	-45.31	29.08	-23.67

The next step is perform the monte-carlo simulation to test the regression coefficient non-stationarity. This test conducted to select the global regression part and the GW regression part. Table-4 shows that the air temperature (X_1), the wind velocity (X_2), and the air humidity (X_3) are the global regression part. Meanwhile, four other predictor variables are the GWR part because these variables have the p-value less than 0.05.

Table-4. Monte Carlo test of regression coefficient non-stationerity with adaptive bisquare kernel.

Coefficient	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
p-value	0.00*	0.09	0.30	0,09	0.00*	0.00*	0.00*	0.00*

Note: * significant at $\alpha = 5\%$

Based on the result of monte-carlo simulation to test the regression coefficient non-stationarity, the MGWR model was conducted. The model parameter of MGWR model has shown in Table-5 dan Table-6.

Table-5. Parameter of Global Regression part.

Coefficient	X_1	X_2	X_3
Value	1.4435	-2.3073	-0.0247

Table-6. Parameter of GWR part.

Coefficient	Min	Max	Median
β_0	28.67	90.07	51.85
β_4	2.15	19.25	18.87
β_5	-200.30	1.97	-21.13
β_6	9.30	44.50	12.23
β_7	-51.30	17.19	-26.61

The selection of the best model is done by using the AICc criterion. Table- 6 shows the comparison of the global regression model with the GWR model and MGWR model either by using the bisquare adaptive kernel function. Table-7 shows that the MGWR model is the best model for modelling Air Polluter Standard Index - APSI in Surabaya City because it has the smallest MSE and AICc.

Table-7. Model comparisons.

Model	AICc	MSE
Regression	3,642.675	137,518.1
Basic GWR	3,586.622	112,614.3
Mixed GWR*	3,585.000	117,418.0

Note: *Best Model

5. CONCLUSSIONS

The MGWR model of the air pollution is also influenced significantly by the location factor (geographical factor). Therefore, in this study, the GWR model is suitable to model the air polluter. The local influence of GWR for observation shows that the five significant influencing predictor variables are the air temperature (X_1), the wind velocity (X_2), the air humidity (X_3), the traffic velocity (X_4), the population density (X_6). The other predictor variables are area size of the urban forest (X_5) and the business centre aspect (X_7) not significant in the model GWR.

ACKNOWLEDGEMENT

We would like to give thank to Directorate of Research and Public Services, The Ministry of Research, Technology and Higher Education Republic of Indonesia for their support. This research was funded by "PUPT" Research Grant 2017.

REFERENCES

Atash F. 2007. The Deterioration of Urban Environments in Developing Countries: Mitigating the Air Pollution Crisis in Tehran, Iran. Cities. 24(6): 399-409.

Ebtekar. 2006. Air Pollution Induced Asthma and Aleterations in Cytokine Pattern. Allergy Asthma Immunol. 5(2):47-56.

Fahimi M., Dharma B., Fetararayani D. and Baskoro 2012. Asosiasi antara polusi udara dengan IgE total serum dan tes faal paru pada polisi lalu lintas. Jurnal Penyakit Dalam. pp. 1-9.

Fotheringham A.S., Brunsdon C. and Charlton M. 2002. Geographically Weighted Regression, Jhon Wiley & Sons, Chichester, UK.

Gilbert A. and Chakraborty J. 2011. Using Geographically Weighted Regression for Environmental Justice Analysis: Cumulative Cancer Risks from Air Toxics in Florida. Social Science Research. 40: 273-286.

Leung Y., Mei C.L. and Zhang W.X. 2000. Statistical Test for Spatial Nonstationarity Based on the Geographically Weighted Regression Model, Environment and Planning. 32(5): 871-890.

Lu B., Harris P., Charlton M. and Brunsdon C. 2014. The GWmodel R package: Further Topics for Exploring Spatial Heterogeneity using Geographically Weighted Models Geo-spatial Information Science. 17(2): 85-101, http://www.tandfonline.com/doi/abs/10.1080/10095020.2014.917453.

Lu B., Harris P., Charlton M., Brunsdon C., Nakaya T. and Gollini I. 2016. Package 'GWmodel.

Mei C.L., Wang N. & Zhang W.X. 2006. Testing the importance of the explanatory variables in a mixed geographically weighted regression model. Environment and Planning. 38: 587-598.

Nakaya T., Fotheringham A.S., Brunsdon C. and Charlton M. 2005. Geographically Weighted Poisson Regression for Disease Association Mapping. Statistics in Medicine. 24(17): 2695-2717.

Nakaya T.; Charlton M.; Fotheringham A.S.; Brunsdon, C. 2009. How to use SGWRWIN (GWR4.0). National Centre for Geocomputation, National University of Ireland Maynooth.

Purhadi and Yasin, H. 2012. Mixed Geographically Weighted Regression Model (Case Study: The Percentage of Poor Households in Mojokerto 2008). European Journal of Scientific Research. 69(2): 188-196.

Rencher A C. 2000. Linear Models in Statistics, John Wiley & Sons, New York, USA.

Robinson D. and Lloyd J. M. 2013. Increasing the Accuracy of Nitrogen Dioxide (NO2) Pollution Mapping Using Geographically Weighted Regression & Geostatistic. International Journal of Applied Earth Observation and Geoinformation. 21: 374-383.

Yang W., Fotheringham A.S. and Harris P. 2012. An extension of geographically weighted regression with flexible bandwidths. Proceedings GISRUK 20th Annual Conference 2012.