Mixed geographically weighted regression (MGWR) model with weighted adaptive bi-square for case of dengue hemorrhagic fever (DHF) in Surakarta

H N Astuti1, a), D R S Saputro2, and Y Susanti3

1,2,3 Department of mathematics, FMIPA, Universitas Sebelas Maret

E-mail: a)h4rdanti@gmail.com

ABSTRACT. MGWR model is combination of linear regression model and geographically weighted regression (GWR) model, therefore, MGWR model could produce parameter estimation that had global parameter estimation, and other parameter that had local parameter in accordance with its observation location. The linkage between locations of the observations expressed in specific weighting that is adaptive bi-square. In this research, we applied MGWR model with weighted adaptive bi-square for case of DHF in Surakarta based on 10 factors (variables) that is supposed to influence the number of people with DHF. The observation unit in the research is 51 urban villages and the variables are number of inhabitants, number of houses, house index, many public places, number of healthy homes, number of Posyandu, area width, level population density, welfare of the family, and high-region. Based on this research, we obtained 51 MGWR models. The MGWR model were divided into 4 groups with significant variable is house index as a global variable, an area width as a local variable and the remaining variables vary in each. Global variables are variables that significantly affect all locations, while local variables are variables that significantly affect a specific location.

Keywords: local parameters, global parameters, MGWR model, GWR model, DHF, adaptive bi-square

1. Introduction

Tobler’s first law of Geography (Tobler in Anselin [1]), “Everything is related to everything else, but near things are more related than distant things”, has laid foundation for regional science, specifically on spatial dimension. Spatial data refer to geographically-oriented data and possess certain coordinate system as a reference framework. This implies that spatial data contain information about geographic location of an area width. Spatial analysis, in general, requires location-based data containing the location’s characteristics. It includes modeling which indicates a concept of cause and effect relationship by using a method sourced from spatial and non-spatial data sources to predict the presence of a spatial pattern.

One of spatial models is geographically weighted regression (GWR). It is a spatial regression model of which geographical location influences dependent variables (Fotheringham, et al. 2002 [9]). In other words, GWR enables to resolve spatial heterogeneity due to heterogeneous conditions of each spatial unit, whereas in fact not all of the independent variables in the GWR exert a spatial influence on dependent ones. For further references on the GWR, see Schabenberger and Gotway (2005, pp.
A Mixed Geographically Weighted Regression model (MGWR) is a combination of linear regression and the GWR. It is a regression model of which some independent variable coefficients are constant, while some other spatially vary (Brunsdon, et al.[3]). The combination is obtained after testing for spatial variability has been carried out. In the GWR, resulting parameters are only significant for certain spatial units. In several cases, the resulting parameters are significant for both certain spatial units and all of the spatial units.

The MGWR is appropriate for data which are influenced by local and global variables. The GWR and the MGWR differ on their spatial variability. All contributing factors in the GWR have spatial variability, and therefore result in local parameters which are influential for certain spatial units (Wheeler [10], Fotheringham [9], Brunsdon [8]). Meanwhile, some contributing factors in the MGWR have spatial variability, but some other do not. The factors having no spatial variability will generate a global parameter—a parameter which is influential to spatial units, while those having spatial variability will produce a local parameter. In order to explain the influence of the spatial units, weight is utilized. The present research employed bi-square weight which is determined based on the distance of the spatial units.

According to Sefuddin et al. [6], the implementation of the MGWR model with bi-square weight on poverty rates contributes to adequately high goodness of fit. Several other applied researches on the MGWR and GWR models were carried out by Pecci, F and M Sassi [7], Bai et al. [13], Huang, et al. [14], and Wrenn and Sam [15]. Both models yield dissimilar parameter estimation in every spatial unit. The GWR is considered the best model to explain the influence of low-wealth status in districts of North Sulawesi province having the lowest values of MSE and AIC and the highest value of $R^2$ in global regression and MGWR (Pongoh 2015 [11]).

Dengue Hemorrhagic Fever (DHF) is included as one of infectious diseases, which frequently occurs in Indonesia. Data of the number of its sufferers can be modeled using MGWR. The present research seeks to model the number of DHF sufferers in Surakarta by utilizing MGWR model with ‘adaptive’ (bi-square) weight.

2. Mixed Geographically Weighted Regression Model

GWR model. According to Brundson et al. [4], GWR model is influenced by such an aspect as spatial unit, in this case point in the spatial unit, and therefore the model yields estimates for every point. The model is formulated below:

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_{ki} x_{ki} + \epsilon_i$$

where $y_i, i = 1, 2, ..., n$ represents dependent variable of the $i^{th}$ datum, $\beta_{ki}, k = 1, 2, ..., m$ is regression coefficient for the $i^{th}$ location and the $k^{th}$ variable, $x_{ki}$ is the $i^{th}$ value of $x_k$, and $\epsilon_i$ is a residual which is normally distributed and independent.

2.1 MGWR Model.

The GWR was then developed to MGWR, a model which does not only consider local variable, but also the global one. It is formulated:

$$y_i = \sum_{g=1}^{q} a_{ig} x_{ig}(a) + \sum_{l=q+1}^{p} b_l(u_i, v_i)x_{il}(b) + \epsilon_i, i = 1, 2, ..., n$$

(1)

where $a_{ig}$ is a global parameter, while $b_l$ is a local parameter. Variable $x_{ig}$ is an independent variable which is related to the global parameter, while $x_{il}$ represents an independent variable which is related to local parameter, and $\epsilon_i$ is a residual in the MGWR model. In the model, an
assumption which should be met is heterogeneity. The test of heterogeneity was conducted using Breusch-pagan (BP) test with test statistic, as formulated below:

\[
BP = \frac{1}{2} r^T Z (Z^T Z)^{-1} Z^T r - n \tag{2}
\]

where \( r \) is 51x1 matrix and is formulated \( r = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} \), \( r_i = \left( \frac{\epsilon_i^2}{\sigma^2} \right) \), \( \sigma^2 = \frac{1}{n} \sum \epsilon_i^2 \), while \( Z \) is 51x4 matrix. It is decided that \( H_0 \) is rejected if \( BP \in DK \) with critical area of \( DK = \{ BP | BP > \chi^2_{(k-l)} \} \).

2.2 Parameter Estimation of the MGWR Model.
Brundson et al [3] state that since MGWR model includes both local and global parameters, the parameter estimation needs to be conducted using weight least square (WLS) for the former and ordinary least square (OLS) for the latter.

2.3 ‘Adaptative’ (Bi-square) Weight Matrix
Adaptive kernel function was utilized in the research since it enables to adapt bandwidths to in size to variations in data density. Therefore, when bandwidth is too large, only few data are involved, and vice versa. Fotheringham et al. [7] propose two weight functions, comprising gaussian and bi-square functions. The latter, as formulated below, was employed in the present research.

\[
w_{ij} = \begin{cases} 
1 - \left( \frac{d_{ij}}{h_i} \right)^2, & d_{ij} < h_i \\
0, & d_{ij} \geq h_i
\end{cases} \tag{3}
\]

where \( d_{ij} = ((x_i - x_j)^2 + (y_i - y_j)^2)^{1/2} \) represents Euclidean distance from spatial unit \((x_i, y_i)\) to \((x_j, y_j)\), while \( h_i \) is bandwidth.

3. Research Method
The present study presents an applied research which models the number of dengue hemorrhagic fever sufferers in Surakarta by applying the MGWR model and bi-square weight. Research data and procedures are described below.

3.1 Data.
Data include those on the number of DHF sufferers in 2014 which were sourced from Surakarta Department of Health and used as dependent variable \((Y)\). Factors involved in the research comprise such independent variables \((X_i)\) as the number of inhabitants, \((X_1)\), the number of houses \((X_2)\), House Index \((X_3)\), many public places \((X_4)\), the number of healthy homes \((X_5)\), the number of Posyandu \((X_6)\), area width \((X_7)\), level population density \((X_8)\), welfare of the family \((X_9)\) and high region \((X_{10})\) sourced from a public health center \((puskesmas)\) in Surakarta.

3.2 Research Procedures.
Research procedures include: 1) checking heterogeneity of data on the number of DHF sufferers by using BP test based on equation (2), 2) computing weight using bi-square in each spatial unit based on equation (3), 3) determining significant independent variable \((X_i)\) and GWR model in each spatial unit. During this procedure, both global and local
variables for MGWR model were obtained, 4) estimating parameters in MGWR model and finally MGWR model which fits to model (1) for each location was obtained.

4. Results and Discussion

Results and discussion cover data description, test of heterogeneity, ‘adaptive’ (bi-square) weight, GWR model, and MGWR model, each of which is explained further below.

4.1 Data Description

Figure 1 demonstrates that Kadipiro sub-district (A) has the most numerous DHF sufferers with total number of 40 persons. Its surrounding sub-districts include Mojosongo (B) with total number of 33 DHF sufferers, Nusukan (C) with total number of 19 DHF sufferers, and Banyuanyar (P)
with total number of 5 DHF sufferers. Other sub-districts have an average number of 1-5 DHF sufferers. This reveals the presence of a spatial indication of distribution of DHF sufferers. Furthermore, there exists a significant data inequality which signifies the presence of heterogeneity in the data. In order to find it out, the Breusch-Pagan (BP) test was carried out.

4.2 Test of Heterogeneity
The BP test with null hypothesis \((H_0)\) reveals that there is no heterogeneity in data or all spatial units having dissimilar characteristics. It results in BP value of 12.40682 with tabulated value \(\chi^2_{(3.0.05)}\) of 7.815 and therefore \(H_0\) is rejected. This implies that the DHF sufferers in Surakarta are heterogeneous and have different characteristics.

4.3 Bi-square Weight Matrix
A weight matrix quantifies spatial relationship between spatial units. The computation of the weight was conducted in each spatial unit and therefore it may differ. Table 2 presents an example of the computation in which Pasar Kliwon sub-district serves as the center of distribution.

| Sub-district       | Weight | Weight |
|--------------------|--------|--------|
| Pajang             | 0.053933 | 0.999995251 |
| Karang Asem        | 1.293993 | 0.997268126 |
| Sondakan           | 0.046209 | 0.999996514 |
| Bumi               | 0.034495 | 0.999998057 |
| Sriwedari          | 0.022023 | 0.999999208 |
| Panularan          | 0.026753 | 0.999998831 |
| Purwosari          | 0.022023 | 0.999999208 |
| Jajar              | 0.052368 | 0.999995523 |
| Kemalayan          | 0.015652 | 0.999996000 |
| Jayengan           | 0.014616 | 0.999996514 |
| Tipes              | 0.021536 | 0.999999243 |
| Serengan           | 0.019668 | 0.999999368 |
| Danakusuman        | 0.012795 | 0.999997333 |
| Joyontakan         | 0.019647 | 0.999999370 |
| Joyosuran          | 0.012628 | 0.999999740 |
| Pasar Kliwon       | 0.000000 | 1.000000000 |
| Baluwarti          | 0.006392 | 0.999999333 |
| Sangkrah           | 0.006956 | 0.999999921 |

In reference to the above weights of 51 sub-districts in which Pasar Kliwon sub-district presents as the center of distribution, weight matrix \(W\) was determined using the following formula:

\[
W = \begin{bmatrix}
    W_{11} & W_{12} & \cdots & W_{1n} \\
    W_{21} & W_{22} & \cdots & W_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    W_{n1} & W_{n2} & \cdots & W_{nn}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
    1 & 0.999999619 & \cdots & 0.999999693 \\
    0.99999619 & 1 & \cdots & 0.999996953 \\
    \vdots & \vdots & \ddots & \vdots \\
    0.999996693 & 0.999996953 & \cdots & 1
\end{bmatrix}
\]
4.4 GWR Model

GWR model refers to a model developed from multiple linear regression by utilizing weight as a basis for spatial analysis. Bi-square function employs bandwidth to form weight matrix for each spatial unit. The optimal bandwidth for the bi-square weight is 33 km. The GWR results in 51 models, one of which is that of Pasar Kliwon sub-district which is formulated:

\[ y_{P_k} = 3.829165 + 0.643972X_3 + 3.579917X_7 - 0.000345X_8. \]

4.5 MGWR Model

MGWR model is yielded from global and local variables having significant influence on the number DHF sufferers. Such global variables as house index \( X_3 \) and level population density \( X_8 \) were obtained in the model. The result of the analysis of each sub-district from GWR model was classified in Table 3.

| Group | Sub-district(s) | Significant variable(s) | The number of the sub-districts | Percentage |
|-------|----------------|-------------------------|-------------------------------|------------|
| 1     | Kadipiro, Mojosongo, Nusukan, Manahan, Jebres, Mangkubumen, Gilingan, Jajar, Tipis, Srewedari, Jayengan, Banyuanyar, Sumber, Kemlayan, Sondakan, Baluwarti, Serengan, Purwosari, Punggawan, Panularan, Pajang, Kestalan, Kepatihan Kulon, Karang Asem, Bumi | \( X_7 \) | 25 | 49.02% |
| 2     | Pucang sawit, Jagalan, Tegal Harjo | \( X_7, X_8 \) | 3 | 5.88% |
| 3     | Pasar Kliwon, Semanggi, Sangkrah, Kampung Sewu Joyontakan, Danukusuman, Joyosuran, Kauman, Timuran, Sudiroprajan, Setabelan, Purwodiningratran, Penumping, Laweyan, Kratonan, Ketelan, Kerten, Keprabon, Kepatihan Wetan, Kedung Lumbu, Gajahan, Gandekan | \( X_3, X_7, X_8 \) | 4 | 7.84% |
| 4     | No significant variable | No | 19 | 37.26% |

Considering the GWR model and the classification shown by Table 3, local variable for the MGWR model includes area width \( (X_7) \), while global variable covers House Index \( (X_3) \) and level population density \( (X_8) \). After significant test for global variable, House Index was the significant global variable. The House Index, which refers to the number of houses positively infected with larvae and/or pupae, exerts a significant influence on the number of DHF sufferers in Group 3. Meanwhile, area width gives a significant influence on the number of DHF sufferers in all of the groups. Figure 2 illustrates the mapping of significant variables demonstrated in Table 3.
Figure 2. The classification of sub-districts based on significant variables

Figure 2 presents that 1) numerous DHF sufferers in Group 1 influenced by area width \( (X_7) \) are found in 25 sub-districts (49.02%) and spread over Surakarta, except areas bordered with Sukoharjo, 2) numerous DHF sufferers in Group 2 influenced by area width \( (X_7) \) and level population density \( (X_6) \) are found in 3 sub-districts (5.88%), 3) numerous DHF sufferers in Group 3 (7.84%) influenced by area width \( (X_7) \), level population density \( (X_6) \) and House Index \( (X_3) \) are found in 4 sub-districts, and 4) numerous DHF sufferers in Group 4 (37.26%) are not influenced by 10 variables predicted to give influence on the number of the DHF sufferers. The parameter estimation is demonstrated by Table 3.

Table 4. Parameter estimation of local variables \( \hat{\beta}_0, \hat{\beta}_7, \) and \( R^2 \) for 11 sub-districts

| Sub-district  | \( \hat{\beta}_0 \)  | calculated t value | \( \hat{\beta}_7 \)  | calculated d t value | Local \( R^2 \)  |
|---------------|----------------------|--------------------|----------------------|----------------------|-----------------|
| Pajang        | 4.623182             | 6.206654           | 6.113438             | 6.286946             | 0.643961        |
| Laweyan       | 4.611724             | 6.048076           | 4.777195             | 2.726309             | 0.363126        |
| Karang Asem   | -1.453233            | -0.462694          | 6.939781             | 8.599443             | 0.884828        |
| Sondakan      | 4.706330             | 6.481070           | 5.853682             | 5.282221             | 0.581010        |
| Bumi          | 4.892310             | 6.574185           | 5.631906             | 3.421980             | 0.429166        |
| Sriwedari     | 5.142577             | 6.702331           | 6.289033             | 3.749680             | 0.456800        |
| Panularan     | 5.602855             | 6.082733           | 7.522728             | 3.718553             | 0.460204        |
| Purwosari     | 5.963288             | 6.117586           | 8.430034             | 4.128316             | 0.520631        |
| Serengan      | 3.967844             | 3.731161           | 3.950396             | 1.726267             | 0.347405        |
| Kerten        | 4.872237             | 7.177317           | 7.020797             | 9.456836             | 0.775779        |
| Pasar Kliwon  | 2.901746             | 3.412516           | 2.365120             | 1.404927             | 0.347395        |

Parameter estimation for the MGWR model was carried out using global parameter estimation, \( \hat{\beta}_3 \), of 1.153763 and local parameters, \( \hat{\beta}_0 \) and \( \hat{\beta}_7 \), as figured out by Table 4 for 11 sub-districts.
Table 5. Significant local parameter estimation of $\beta_0$ and $\beta_7$ for the MGWR model

| Sub-district   | $\beta_0$   | $\beta_7$   |
|---------------|-------------|-------------|
| Pajang        | 4.623182    | 6.113438    |
| Laweyan       | 4.611724    | 4.777195    |
| Karang Asem   | Not Significant | 6.939781    |
| Sondakan      | 4.706330    | 5.853682    |
| Bumi          | 4.892310    | 5.631906    |
| Sriwedari     | 5.142577    | 6.289033    |
| Serengan      | 3.967844    | Not Significant    |
| Kerten        | 4.872237    | 7.020797    |
| Pasar Kliwon  | 2.901746    | Not Significant    |

By referring to Table 4, 51 MGWR models for each sub-district are gained. The general formula of MGWR model for predicting the number of DHF sufferers is constructed below:

$$\hat{Y}_i = \hat{\beta}_{i0} + 1.361575X_{i3} + \hat{\beta}_{i7}X_{i7}$$

The model results in $R^2$ value of 0.78, indicating that 78% of total variability is explained by such variables of House Index and area width. An example of MGWR model for Pasar Kliwon sub-district is formulated:

$$\hat{y}_{pk} = 2.901746 + 1.153763X_3.$$  

It means that an increase in a unit of House Index leads to an increase in a number in a number one or two DHF sufferers. Provided that the House Index is constant, a number of one or two sufferer(s) exist(s). The model has R-squared value ($R^2$) of 34.7%, indicating that 34.7% of total variability is explained by House Index, while 65.3% by other variables excluded in the observation. The other example of MGWR model was MGWR model for Pajang sub-district is formulated:

$$\hat{y}_{pj} = 4.623182 + 1.153763X_3 + 6.113438X_7.$$  

It means that an increase in a unit of House Index lead to an increase in a number one or two DHF sufferers. Provides that the House Index is constant, a number of one or two sufferer(s) exist(s). An increase in one km$^2$ of area width leads to an increase in number six or seven DHF sufferers. Provided that area width is constant, a number of six or seven sufferers exists. The model has R-squared value ($R^2$) of 64.4%, indicating that 64.4% of total variability is explained by House Index and area width, while 35.6% by other variables excluded in the observation.

In reference to the two aforementioned MGWR models with global variable of House Index, it is clear that an increase in a unit of House Index causes an increase in a number of one or two DHF sufferer(s) in each sub-district in Surakarta.

5. Conclusions
The research concludes that House Index ($X_3$) serves as an independent variable which has global influence, while area width ($X_7$) exerts local influence. The MGWR model results in R-squared value ($R^2$) of 0.78. The general formula of MGWR model for predicting the number of DHF sufferers using bi-square weight is shown below:

$$\hat{Y}_i = \hat{\beta}_{i0} + 1.361575X_{i3} + \hat{\beta}_{i7}X_{i7}.$$
The formula indicates that when other variables are constant, an increase in one unit of House Index leads to an increase in a number of one or two DHF sufferer(s) in each sub-district in Surakarta.

References

[1] Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography (46, 234-240).
[2] Anselin, L. (1998). Spatial econometrics : methods and models. Dordrecht: Kluwer Academic Publishers.
[3] Brunsdon, C., S. A. Fotheringham, and M. Charlton. (1996). Geographically weighted regression: a method for exploring spatial nonstationary. Geographical Analysis (Vol. 28 No. 4, 281-298).
[4] Brunsdon C., S. A. Fotheringham, and M. Charlton. (1999). Some notes on parametric significance tests for geographically weighted regression. Journal of Regional Science (39, 497-524).
[5] Profil kesehatan provinsi Jawa Tengah. (2013). Semarang: Dinas Kesehatan Provinsi Jawa Tengah (Health Department of Central Java province).
[6] Saefuddin, A., N. A. Setiabudi, and N. A. Achsani. (2011). On comparisson between ordinary linear regression and Geographically Weighted Regression: with application to Indonesian poverty data. European Journal of Scientific Research (Vol. 57 No. 2, 275-285).
[7] Pecci, F and M Sassi. (2008). A mixed Geographically Weighted approach to decoupling and rural development in the EU-15. Paper prepared for presentation at the 107th EAAE Seminar "Modelling of Agricultural and Rural Development Policies". Sevilla, Spain, January 29th - February 1st.
[8] Brunsdon, C., Fotheringham, A.S. and Charlton M.E. (1999). Some notes on parametric significance tests for geographically weighted regression. Journal of Regional Science (39, 497-524).
[9] Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). Geographically weighted regression: The analysis of spatially varying relationships. West Sussex: Wiley.
[10] Wheeler, D.C. and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical Systems (7, 161-187).
[11] Pongoh, F., I M Sumertajaya, M N Aidi. (2015). Geographicchal Weighted Regression and Mix Geographicchal Weighted Regression. International Journal of Statistics and Applications, 5(1), 1-4. doi: 10.5923/j.statistics.20150501.01.
[12] Wei, C.H.; Qi, F. (2012). On the estimation and testing of mixed geographically weighted regression models. Econ. Modell (29, 2615–2620).
[13] Bai, Y.; Wu, L.; Qin, K.; Zhang, Y.; Shen, Y.; Zhou, Y. (2016). A geographically and temporally weighted regression model for ground-level PM2.5 estimation from satellite-derived 500 m resolution AOD. Remote Sens (8, 262).
[14] Huang, B.; Wu, B.; Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. Int. J. Geogr. Inf. Sci (24, 383–401).
[15] Wrenn, D.H.; Sam, A.G. (2014). Geographically and temporally weighted likelihood regression: Exploring the spatiotemporal determinants of land use change. Reg. Sci. Urban Econ (44, 60–74).
[16] Wheeler, D.; Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J. Geogr. Sys. (7, 161–187).
[17] Cho, S.; Lambert, D.M.; Kim, S.G.; Jung, S. (2009). Extreme coefficients in geographically weighted regression and their effects on mapping. Glsci. Remote Sens. (46, 273–288).
[18] Lloyd, C. D. (2007). Local models for spatial analysis. Boca Raton: CRC.
[19] Schabenberger, O. and Gotway, C. A. (2005). Statistical methods for spatial data analysis. London: Chapman & Hall.
[20] Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*. NJ: Wiley, Hoboken.