Robust geographically weighted regression with least absolute deviation (case study: the percentage of diarrhea occurrence in Semarang 2015)

I C Nurhayati, B Warsito, H Yasin, A Rusgiyono
Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
E-mail: ikachandra07@gmail.com

Abstract. Diarrhea is one of many health issues in the developing country like Indonesia because the sickness and the death number are still high. According to the health profile of Semarang City, the people who suffer from diarrhea from 2010-2015 are decreasing. The lowest point happened in the year 2013 with the total case of 38,001. However, there is an increasing number from 2014-2015. The distribution data of diarrhea is spatial data. The differences between environment and sanitation could cause spatial heterogeneity. The spatial heterogeneity could cause the produced variant value no longer constant, but instead, it is different in each region. Therefore, the regression model that involves the effects of spatial heterogeneity is needed, which are Geographically Weighted Regression (GWR) that is built by Weighted Least Square (WLS) adjuster. Although, GWR parameter adjuster that used WLS is very sensitive with the existence of outliers. The existence of the outlier in the data will create a vast residual. Thus, a more robust method is needed, which is the Least Absolute Deviation (LAD) methods in order to estimate the parameter on model GWR. This model is called Robust GWR (RGWR). The result shows that the model events of diarrhea on each region in Semarang City are different. Furthermore, the model events of diarrhea with the RGWR model generate MAPE 16.3396% which means the performance of RGWR is formed well.

1. Introduction
Diarrhea is one of many health issues in developing country like Indonesia, due to the high sickness and death numbers. This disease is still an endemic and also an “extraordinary” situation followed by death. According to the health profile of Semarang City, the people who suffer from diarrhea from 2010 – 2015 are decreasing. The lowest point happened in the year 2013 with the total case of 28,001. However, there is an increasing number from 2014 – 105 [1].

Diarrhea is a contagious disease. Thus a case of diarrhea in one region can affect a diarrhea case in the surrounding area [2]. One of the statistical methods that can identify the factors that affect a diarrhea case is regression analysis. Factors that affect a diarrhea case in the particular region depend on the surrounding condition, called spatial influence [3, 4]. The spatial influence that connected with the differences of environment and geographical characteristic between observed regions is spatial diversity or spatial heterogeneity. Therefore, the regression model that involves the effects of spatial heterogeneity is needed, which are Geographically Weighted Regression (GWR).

GWR model is built with Weighted Least Square (WLS) adjuster. However, parameter adjuster that is used by GWR, which is the least square, is very sensitive to the existence of outliers. Thus, a
more robust method is needed in order to accommodate the outlier in the GWR model. One of the methods that can be used is using the Least Absolute Deviation (LAD), the adjuster. LAD is an estimation method that minimalizes the total absolute number from error [5]. By utilizing LAD, the effect that caused by outlier can be solved automatically without the need to detect which data is becoming the outlier or without the need to redefine the residuals [6]. According to the spread above, in this research, the study about GWR model will be done using LAD method in the parameter adjuster, during the diarrhea case in 16 regions in Semarang City in 2015.

2. Methodology

2.1. Linear Regression

Regression is a method for measuring the influences of the independent variable to the dependent variable and predicting dependent variable using the independent variable [7]. Mathematically, the linear regression model can take the following form:

\[ y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \quad , i = 1, 2, \ldots, n \]  

where,

- \( y_i \) = Response variable on observation-i
- \( \beta_0, \beta_k \) = Intercept and regression coefficient from explanatory variable
- \( x_{ik} \) = Explanatory variable-k from observation-i
- \( \varepsilon_i \) = The remaining components from observation-i are assumed to be a Normal distributed \((\varepsilon \sim N(0, \sigma^2))\)

If stated in the form of the matrix equation, the formula will be

\[ Y = X\beta + \varepsilon \]

where:

\[ Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix} \]

One of the methods that can be used to gain those estimations is Ordinary Least Square (OLS) method. The principle of OLS methods is to minimalize the square root residual that is produced by a model with the equation of:

\[ L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left( y_i - \sum_{k=0}^{p} \hat{\beta}_k x_{ik} \right)^2 \]

Then the estimation parameter \( \hat{\beta} \) is gained with the formula \( \hat{\beta} = (X^T X)^{-1} X^T Y \).

2.2. Spatial Heterogeneity Test

Breusch-Pagan test can be used to detect the effect of spatial heterogeneity [8]. Hypothesis test that is used in the Breusch-Pagan test is:

- \( \text{H}_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma^2 \)
- \( \text{H}_1: \text{there should be at least one } i \text{ therefore } \sigma_i^2 \neq \sigma^2 \quad (i = 1, 2, \ldots, n) \)

Level of significance = \( \alpha \)

Statistics test

\[ \text{BP} = \frac{1}{2} f^T Z (Z^T Z)^{-1} Z^T f \]

\( f = (f_1, f_2, \ldots, f_n)^T \) with \( f_i = \left( \frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \). \( Z \)-Matrix is a matrix of size \( n \times (p+1) \) that consist of standardized explanatory variables. The value of BP is a scalar value that follows noncentral Chi-Square distribution p.
Decision Criteria
Dismiss $H_0$ if $BP > \chi^2_{(p,\alpha)}$. With $p$ is the number of independent variables or if $p$-value < $\alpha$. If $H_0$ is dismissed, then there will be characteristic differences in one region with another region. Therefore, it needs to be done using Geographically Weighted Regression (GWR).

2.3. Spatial Outlier Detection
Based on [9], a spatial outlier is defined as local instability or spatial object that picture non-spatial attribute which is relatively extreme or significantly different from another spatial object in the neighborhood. One of the methods that can be used to detect the outlier is Spatial Z Statistic Test with the hypothesis test as follows:

$H_0 : \mu_i = \mu_{S(x)}$ (The location observation value $i$ is not a spatial outlier)

$H_1 : \mu_i \neq \mu_{S(x)}$ (The location observation value $i$ is a spatial outlier)

Level of significance = $\alpha$

Statistics test

$$Z_{stat} = \frac{S(x) - \mu_{S(x)}}{\sigma_{S(x)}} \quad (4)$$

with $S(x)$ as a difference between observation value $i$ with an average value of observed location which is close with location $i$

Decision Criteria
Observation value from location $i$ is detected as a spatial outlier if the value $Z_{stat} > Z_{\alpha}$. The value of $Z_{\alpha}$ is a value of normal distribution $Z$ table for significance value of $\alpha$.

2.4. Geographically Weighted Regression
Geographically Weighted Regression (GWR) is the development from the regression model counted in every location of observation. GWR model according to [10] could be written as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n \quad (5)$$

Parameter assessment using WLS is the same as linear regression, with minimalizing the number of square residuals using the following equation:

$$\sum_{j=1}^{n} w_j(u_i, v_i) \varepsilon_j^2 = \sum_{j=1}^{n} w_j(u_i, v_i) \left[ y_j - \beta_0(u_i, v_i) - \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} \right]^2 \quad (6)$$

If written in the matrix, it will be

$$\mathbf{e}^T \mathbf{W}(u_i, v_i) \mathbf{e} = [\mathbf{Y} - \mathbf{X}\mathbf{\beta}(u_i, v_i)]^T \mathbf{W}(u_i, v_i) [\mathbf{Y} - \mathbf{X}\mathbf{\beta}(u_i, v_i)] = \mathbf{Y}^T \mathbf{W}(u_i, v_i) \mathbf{Y} - 2\mathbf{\beta}^T(u_i, v_i)\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{Y} + \mathbf{\beta}^T(u_i, v_i)\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{X} \mathbf{\beta}(u_i, v_i)$$

Therefore the estimation of GWR parameter model could be obtained from the equation below:

$$\hat{\mathbf{\beta}}(u_i, v_i) = [\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{Y}$$

Weighted matrix on the GWR model is a very important aspect. The role of weight represents the location of one observation data with another. Weighted matrix is a diagonal matrix from weight in each observed location, which pictured as:

$$\mathbf{W}(u_i, v_i) = \begin{bmatrix}
    w_{11} & 0 & \cdots & 0 \\
    0 & w_{12} & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & w_{1n}
\end{bmatrix}$$

where $w_{ij} = w_j(u_i, v_i)$ with $j = 1, 2, \ldots, n$ is the weighted location $j$ in order to estimate the parameter on location-$i$. 


One of the commonly used methods in weighting GWR model is Kernel Function. This research uses Adaptive Gaussian Kernel Function with the formula:

$$ w_j(u_i, v_i) = \exp \left( -\frac{1}{2} \left( \frac{d_{ij}}{h_{(q)}} \right)^2 \right) $$

with,

$$ d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} $$ as a Euclidean distance between location-\(i\) and location-\(j\)

$$ h_{(q)} \text{ = bandwidth adaptive which determine q as the closest distance from location-}i $$

Bandwidth selection has a significant impact on the GWR outcome. Therefore, looking for optimum bandwidth value needs to be done in order to generate a good model. One of the methods that can be used to choose optimum bandwidth value is Cross-Validation (CV), which formulated as:

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

Where \(\hat{y}_{\neq i}(h)\) is the estimation value \(y_i\) where the observation in the location-\(i\) is dismissed from the estimation process.

**2.5. Robust Geographically Weighted Regression**

The spatial outlier is sometimes hard to detect in GWR model. Therefore, a more robust method is needed. The term robust could be explained as the insensitive or the rigidity to small changes on the assumption [11]. One of the methods that can be applied to the GWR model is LAD [12]. The model that used on RGWR is the same as the model that used on GWR model. As well as the used weighted matrix, the differences are on the used criteria and the chosen optimum bandwidth [13].

The choosing for optimum bandwidth criteria on RGWR can be done using Absolute Cross Validation (ACV) criteria procedure. According to [6], the outlier does not affect the ACV score, which causes the ACV score to be more robust than the CV score. ACV criteria using the absolute value from the difference of the response variable \(y_i\) moreover, the estimation value of \(\hat{y}_{\neq i}(h)\), which formulated as:

$$ ACV(h) = \sum_{i=1}^{n} |y_i - \hat{y}_{\neq i}(h)| $$

The optimal value from bandwidth \(h\) can be chosen in the same way as the criteria for a CV. Which by choosing the value of bandwidth \(h\) that produce the smallest ACV(h).

Estimated parameter from the RGWR model can be done by adopting the following equation:

$$ L = \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - \sum_{k=0}^{p} \hat{\beta}_k x_{ik}| $$

The solution to produce the regression parameter \((\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)\) can not be done with the differentiation process as shown in the OLS method. The completion of the solution from a regression using LAD is similar to the completion using a linear program [14]. The following is the estimation of the RGWR model parameter in the linear program:

minimizing \(\sum_{i=1}^{n}(\varepsilon_i^+ - \varepsilon_i^-)w_j(u_i, v_i)\),

with constraints

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

and \(\varepsilon_i^+ , \varepsilon_i^- \geq 0\)

$$ \varepsilon_i^+ = \varepsilon_i \text{ I}(\varepsilon_i > 0), \varepsilon_i^- = -\varepsilon_i \text{ I}(\varepsilon_i < 0) $$

The problem of the linear program above can be solved using the Simplex algorithm.
3. Research Variables
The variables used to consist of one response variable and six independent variables that are thought to affect the incidence of diarrhea. Those independent variables are the percentage of healthy household ($X_1$), the percentage of household with unprotected wellspring ($X_2$), the percentage of population density ($X_3$), the percentage of the resident who has access to clean drinking water ($X_4$), the percentage of household without toilet ($X_5$), and percentage of household with private toilet ($X_6$).

4. Results and Discussion
Referring to Figure 1, the district in Semarang City that has the highest and the lowest number in diarrhea case are shown. Tugu District is the highest district that has the highest percentage of diarrhea case in Semarang City, the numbers are 3.3862%, while the lowest case is in Ngaliyan District with the number of 0.7875%. The mean percentage of diarrhea case in Semarang City in the year 2015 is 1.4795%.

![Figure 1. Graphic of Percentage of Diarrhea Occurrence in Semarang City During the Year of 2015](image)

4.1. Multiple Linear Regression
The next step is to shape the linear regression model with Ordinary Least Square (OLS). The variable percentages of diarrhea case in Semarang City during 2015 ($Y$) is connected with six independent variables that are allegedly affecting them. Thus, the regression model is obtained as:

$$\hat{y} = 19.995 - 0.01502X_4 - 0.4057X_2 - 6 \times 10^{-3}X_3 - 0.1539X_4 + 0.2564X_5 - 0.02059X_6$$

However, from the six independent variables only two independent variables that are significant towards the model, those are $X_4$ (the percentage of the resident who has access to clean drinking water) with p-value = 0.0449 and $X_5$ (the percentage of household without toilet) with p-value = 0.0350. The following is the obtained significant model:

$$\hat{y} = 14.629537 - 0.135580X_4 + 0.172463X_5$$
4.2. Spatial Heterogeneity Test and Spatial Outlier Detection

Because of the value of $BP = 11.089 > \chi^2_{0.05;2} = 5.9915$ and the p-value $= 0.0000309$ is smaller than $\alpha = 0.05$, it is shown that there is spatial heterogeneity or there is a different characteristic on each diarrhea case’s data in every district in Semarang City during 2015.

| Parameter  | Minimum | Maximum | Mean  | Median | Standard Deviation |
|------------|---------|---------|-------|--------|-------------------|
| $\hat{\beta}_0$ | 9.8417  | 10.8169 | 10.1706 | 10.1054 | 0.2530 |
| $\beta_3$  | -0.1017 | -0.0916 | -0.0950 | -0.0942 | 0.0027 |
| $\hat{\beta}_5$ | 0.1881  | 0.2122 | 0.2040 | 0.2054 | 0.0071 |

4.3. GWR Model

Before modeling the diarrhea case in Semarang using RGWR, GWR will be modeled first. The following is the summary of the GWR parameter model from affecting variable:

4.4. RGWR Model

According to ACV criteria, bandwidth Adaptive Gaussian Kernel is made for each location as follow:
Table 3. Adaptive Bandwidth Value Every Location RGWR Model

| Location     | Bandwidth | Location     | Bandwidth |
|--------------|-----------|--------------|-----------|
| Mijen        | 0.2256    | Genuk        | 0.2256    |
| Gunung Pati  | 0.1903    | Gayamsari    | 0.1803    |
| Banyumanik   | 0.1368    | Semarang Timur | 0.1782 |
| Gajah Mungkur| 0.1387    | Semarang Utara | 0.1668 |
| Semarang Selatan | 0.1669 | Semarang Tengah | 0.1625 |
| Candisari    | 0.1450    | Semarang Barat | 0.1392 |
| Tembalang    | 0.1567    | Tugu         | 0.1429    |
| Pedurungan   | 0.1908    | Ngaliyan     | 0.1442    |

Then count weighted matrix for each location. For example, a weighted matrix of Tugu District:

\[
W(u_i, v_j) = \begin{bmatrix}
0.9101 & 0 & \ldots & 0 \\
0 & 0.8073 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0.9968
\end{bmatrix}
\]

The summary of estimated value parameter for the RGWR model with the least absolute deviation obtained by using R software is as follows:

Table 4. Summary of Estimated Parameter RGWR Model

| Parameter | Minimum | Maximum | Mean  | Median | Standard Deviation |
|-----------|---------|---------|-------|--------|--------------------|
| \( \hat{\beta}_0 \) | 1.0168  | 12.9642 | 6.0966 | 2.9351 | 5.5497             |
| \( \hat{\beta}_4 \) | -0.1205 | 0.0048  | -0.0498 | -0.0194 | 0.0575             |
| \( \hat{\beta}_5 \) | -0.1314 | 0.2839  | 0.0747  | 0.0296 | 0.1798             |

According to Table 4, it can be seen that explanatory variables in every district have a different effect on diarrhea case percentage. The explanatory variable used is a significant explanatory variable in the GWR model. For example, the RGWR model made for Tugu district is

\[
\hat{y} = 12.9642 - 0.1205X_4 + 0.2839X_5
\]

On outline, the explanatory variable which has the most significant regression coefficient is the variable X₅ (percentage of household without toilet). If other explanatory variable coefficients are considered fixed, then the diarrhea case percentage in Tugu district will be increased by 0.2839 if the percentage of household without toilet increased by 1%. Then, in variable X₄ (percentage of a resident who has access to available water), if the percentages of a resident who has access to available water increased by 1%, then the diarrhea case in Tugu district will be decreased by -0.1205 assuming fixed variable.

According to the result of obtained estimation, RGWR model with MAPE is 16.3396%. Which means the model performance obtained from estimated parameter by using RGWR with least absolute deviation is good because it is in the range 10% - 20%. While the model performance obtained from estimated parameter by using GWR is 35.8665% which means that the model performance is still in the good limit because it is in the range 20% - 50%.

5. Conclusion

The diarrhea case modeling resulted by RGWR are varied in every location. On outline, the explanatory variable which has the most significant regression coefficient is in variable X₅ (percentage of household without toilet). The RGWR model can generate a better-estimated parameter than the GWR model. It is shown by MAPE value that generated by RGWR model of 16.3396%, while GWR model is 35.8665%. It shows that the use of the robust technique for spatial data that contain outlier has better performance.
Acknowledgments
We would like to give thanks 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 “PTUPT” Research Grant 2018.

References
[1] Kementerian Kesehatan Republik Indonesia 2015 Buku Profil Kesehatan Kota Semarang 2015 (Jakarta: Kementerian Kesehatan RI).
[2] Hariyanti L 2010 Spatial Autoregressive Structural Equational Model pada Prevalensi Diare (Studi Kasus di Pulau Jawa dan Madura) (Tesis) (Bandung: Universitas Padjajaran).
[3] Jensen R, Gatrell, Jay D, McLean, Daniel D 2005 Geo-Spatial Technologies in Urban Environment (New York: Springer).
[4] Bekti, RD 2014 AIP Conference Proceeding 1621 (1)
[5] Zhang H, Zao H 2008 Biometrika, 95 107.
[6] Wang FT, Scott DW 1994 Journal of the American Statistical Association 89 65.
[7] Gujarati D 2006 Ekonometrika Dasar (Jakarta: Erlangga).
[8] Anselin, L 1988 Spatial Econometrics: Method and Models (Kluwer Academic Publisher. The Netherlands)
[9] Maimon, O, Rokach, L 2010 Data Mining and Knowledge Discovery Handbook Second Edition (New York: Springer).
[10] Fotheringham AS, Charlton M, Brunson, C 2002 Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chicester. (UK: John Wiley and Sons).
[11] Warsito B, Yasin H, Ispriyanti D 2018 Journal of Physics Conference Series 2018.
[12] Zhang H, Mei C 2011 International Journal of Geographical Information Science 25 1467.
[13] Afifah R, Andriyana Y, Jaya IGN 2017 AIP Conference Proceedingm 1827 (1).
[14] Wagner HM 1959 Journal of the American Statistical Association 54 206.