Outlier handling of Robust Geographically and Temporally Weighted Regression

G ERDA*, Indahwati and A Djuraidah
Department of Statistics, Bogor Agricultural Universisty

E-mail : erdagustriza@gmail.com

Abstract. Geographically and temporally weighted regression (GTWR) is an expansion of geographically weighted regression involving elements of time in modeling. This method produces a model that is local to each location and time so that the resulting model is more representative. In the GTWR analysis, the regression coefficient estimates are calibrated by the weighted least squares procedure, but the estimates are not robust against the outliers. In fact, if there are outliers in the data, it may create fictitious structures in the estimates of the coefficients which may mislead the result. By using a robust regression with M-estimator developed in GTWR modeling and applied on the number participants of family planning in East Java from 2009 to 2016, it can be concluded that the modeling of the robust GTWR with M-estimator can overcome the problem of outliers that occurred at the location and time studied. These are indicated by the change in the direction of the parameter estimator coefficients which are more relevant with the data plot, the fitted values that are closer to the actual value and a decrease in the MAD value in the model.

1. Introduction
Multiple linear regression is a method used to analyze the linear relationship between explanatory variables and a response variable[1]. One of the problems encountered in this regression is the presence of outliers. If there are outliers in the data, it may create fictitious structures in the estimates of the coefficients which may mislead the result[2]. This is because the resulting errors tend to be large and causes the parameter estimator to be biased and produce a false conclusion[3]. Therefore, a more robust method is needed to accommodate outlier data on the GTWR model. One of the methods that can be used is a robust regression with M-estimator. The research on robust spatial regression seems only limited to the estimates of robust geographically weighted regression (RGWR). Harris et al. in [4]examines the RGWR to measure the spatial relationships between freshwater acidification critical loads and catchment attributes. In 2014, Sari in[5], estimating the potential of paddy farming in East Java in 2012 using RGWR with M-estimator. In addition, Afifah in [6]using RGWR with Least Absolute Deviation (LAD) method to model poverty on the Java Island. The development of robust regression involving spatial and time elements can be said to have never been done. Therefore, this study aims to develop a robust geographically and temporally weighted regression (RGTWR) model using M-estimator.

2. Literature Preview
2.1 Robust Regression using M-Estimator
Robust regression is a regression method used when the distribution of errors are not normal or the presence of outliers affecting the model[7]. Robust regression is used to result model that is robust
against presence of outliers [8]. A robust regression using M-estimator is determined by minimizing the residual as follows [9]

\[ M = \sum_{i=1}^{n} \rho (\epsilon_i) \]

(1)

where \( \rho \) is an objective function. The objective function is a function that is used to find the weighting function in the robust regression. The objective function used in this research is Tukey’s [10][11] bisquare weighted function as follows:

\[ \rho(\epsilon_i) = \begin{cases} 
i \frac{c^2}{6} \left( 1 - \left( \frac{\epsilon_i}{c} \right)^2 \right)^2, \\ \frac{c^2}{6}, \\ 0 \end{cases} \quad |\epsilon_i| \leq c \]

\[ \rho'(\epsilon_i) = \begin{cases} 
i \frac{c^2}{6} \left( 1 - \left( \frac{\epsilon_i}{c} \right)^2 \right)^2, \\ \left( 1 - \left( \frac{\epsilon_i}{c} \right)^2 \right)^2, \\ 0 \end{cases} \quad |\epsilon_i| > c \]

(2)

where \( \rho'(\epsilon_i) \) is the derivative of \( \rho(\epsilon_i) \). The constant \( c \) is equal to 4.685. This value produces high efficiency and still offer protection against outliers.

### 2.2 Geographically and Temporally Weighted Regression

Geographically and temporally weighted regression is an extension of the GWR model [12] in order to account for variations of data [13] from both the spatial [14] and temporal sides simultaneously [15]. The GTWR model for \( p \) explanatory variables with \( y_i \) as response variable from \( (u_i, v_i, t_i) \) location or each observation can be written as follows:

\[ y_i = \beta_0(u_i, v_i, t_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i, t_i)x_{ik} + \epsilon_i \]

(4)

where \( i = 1, 2, ..., n, k = 1, 2, ..., p \) and \( (u_i, v_i, t_i) \) is the coordinate point at location \( i \) and time \( j \). The coefficients \( \beta_k(u_i, v_i, t_i) \) can be expressed as follows:

\[ \hat{\beta}(u_i, v_i, t_i) = \left( X'W(u_i, v_i, t_i)X \right)^{-1}X'W(u_i, v_i, t_i)y \]

(5)

where \( W(u_i, v_i, t_i) = diag(w_{i1}, w_{i2}, ..., w_{in}) \) and \( W(u_i, v_i, t_i) \) is matrix weighting for observation \( (u_i, v_i) \) and time \( t_i \). Spatial-temporal distance function (\( d_{ij}^{ST} \)) consists of a combination of spatial distance functions and temporal distance functions as follows:

\[ (d_{ij}^{ST})^2 = (u_i - u_j)^2 + (v_i - v_j)^2 \]

\[ (d_{ij}^s)^2 = (t_i - t_j)^2 \]

\[ (d_{ij}^t)^2 = \varphi^s[(u_i - u_j)^2 + (v_i - v_j)^2] + \varphi^t[(t_i - t_j)^2] \]

(6)

(7)

(8)

where \( \varphi^s \) and \( \varphi^t \) is the balancing parameter to influence the difference between the location and time on the spatial-temporal measurement [16]. If \( h_s^2 = \frac{h_{sT}^2}{\varphi^s} \) and \( h_T^2 = h_{sT}^2 \) then \( w_{ij}^S = \exp \left\{ -\frac{(d_{ij}^s)^2}{h_s^2} \right\} \)

\[ w_{ij}^T = \exp \left\{ -\frac{(d_{ij}^t)^2}{h_T^2} \right\} \]

(9)

where \( h_s \) is spatial bandwith, \( h_T \) is temporal spatial, and \( h_{sT} \) is spacial-temporal bandwith. If \( \tau \) is the ratio parameter of \( \tau = \frac{\varphi^t}{\varphi^s} \) with \( \varphi^s \neq 0 \) then the equation is obtained

\[ (d_{ij}^{ST})^2 = (u_i - u_j)^2 + (v_i - v_j)^2 + \tau(t_i - t_j)^2 \]

(10)

Parameter \( \tau \) is to enlarge or minimize the effect of temporal distance on spatial distance. This parameter is derived from the minimum CV criteria by initializing the initial \( \tau \) as follows:

\[ CV(\tau) = \sum_{i=1}^{n} [y_i - \hat{y}_i(\tau)]^2 \]

(11)

Then parameter estimator \( \varphi^t \) and \( \varphi^s \) can be obtained with iterative methods based on the results of estimators \( \tau \) that produce a minimum CV.
2.3 Robust GTWR with M-Estimator

The equation for location $i$ and time $t$ that contains an outlier is:

$$\rho(y_i) = \rho[\beta_0(u_i, v_i, t_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i, t_i)x_{ik} + \epsilon_i]$$

with $i = 1, 2, 3, ..., n$; $x_i = x_{i1}, x_{i2}, ..., x_{ip}$, $\beta(u_i, v_i, t_i) = \beta_0(u_i, v_i, t_i), \beta_1(u_i, v_i, t_i), ..., \beta_p(u_i, v_i, t_i)$.

The Robust regression with M-estimator is determined by minimizing the residual with equation as follows:

$$\bar{\beta}_M = \min_{\beta} \sum^{n}_{i=1} \rho(\epsilon_i) = \min_{\beta} \sum^{n}_{i=1} \rho(y_i - X_i^T \beta) = \sum^{n}_{i=1} (\rho(y_i - \beta(u_i, v_i, t_i)) = 0$$

The M-estimator is calculated using iteratively reweighted least square (IRLS). In this iteration, the value $w_i$ will change its value in each iteration, so that:

$$\hat{\beta}(u_i, v_i, t_i)^m = (X'_iW^{m-1}X_i)^{-1}X'_iW^{m-1}y_i$$

Then, for $W^m$ given weights, obtained the estimator:

$$\hat{\beta}(u_i, v_i, t_i)^{m+1} = (X'_iW^{m}X_i)^{-1}X'_iW^{m}y_i$$

where $m$ is the number of iteration. The above calculation will continue to be repeated until the convergent estimator is obtained, that is when the difference in values $\hat{\beta}(u_i, v_i, t_i)^{m+1}$ and $\hat{\beta}(u_i, v_i, t_i)^m$ approaches 0.

3. Simulation and Results

The data used in this study is secondary data about factors that affect number of family planning participants in East Java Province. The data were obtained from the book "Health Profile of East Java" published by East Java Provincial Health Office from 2009 to 2016. The response variable in this study was the number of family planning participants ($Y$), while the explanatory variables used were the number of fertile couples ($X_1$), per capita expenditure ($X_2$), average school length ($X_3$), and number of health facilities ($X_4$). The analysis steps are as follows: (1) Exploring the number family planning participants. (2) Examining the multicollinearity and heterogeneity of spatial and time of data. (3) Model analysis using GTWR by determining spatial-temporal parameter ($\tau$) spatial parameter ($\phi^T$), temporal parameter ($\phi^T$) and optimum spatial-temporal bandwidth ($h_{opt}$) using exponential kernel function. (4) Detecting outlier data using boxplot. (5) Analyzing RGTRW with M-estimator for each location and time by calculating the value of $\hat{y}_i = x_i^T \hat{\beta}(u_i, v_i, t_i)^0$ at each location, with $\hat{\beta}(u_i, v_i, t_i)^0$ is obtained from GTWR modeling. Calculating weighted value $w_i$ using Tukey’s bisquare weighted function by:

$$w_i = \left\{ \begin{array}{ll}
1 - \left( \frac{\epsilon_i}{c} \right)^{2} & \text{if } |\epsilon_i| \leq c \\
0 & \text{if } |\epsilon_i| > b
\end{array} \right.$$  

with $c = 4.685$, $\epsilon_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}}$, $\hat{\sigma} = \frac{\text{median} |\epsilon_i - \text{median}(\epsilon)|}{0.6745}$ and $\epsilon_i = y_i - \hat{y}_i$. At the same time calculating $\hat{\beta}(u_i, v_i, t_i)^m = (X'_iW^{m-1}X_i)^{-1}X'_iW^{m-1}y_i$. Also, Selecting the best model between GTWR and RGTRW.

Figure 1. Map Distribution Number of Family Planning Participants in East Java (in Ten Thousand Souls)
The distribution of number of family planning participants in each regency of East Java from 2009 to 2016 is shown in Figure 1. All districts in the province are grouped into five groups based on the lowest and the highest number of family planning participants. Regency with the highest number of family planning participants for 2009 to 2015 were in the same area that are in Malang, Jember, and Surabaya. In 2010, 2014 and 2015, Sidoarjo regency managed to increase the number of its family planning participants into the highest number of groups but in the next year, Sidoarjo regency could not maintain its position and become a regency in the group with the number of family planning participants second highest. The boxplot of the number of family planning participants in East Java is presented in Figure 2. It can be seen that the pattern of data distribution in 2009 to 2012 tends to be similar, but in the following year, the pattern looks down. In 2014, although the number of family planning participants increased, but again declined until 2015 and then increased again in 2016. The pattern of increase and decrease that occurred indicate the heterogeneity of time on the data. In addition, it can be said that the distribution of number family planning participants in East Java is skewed right with almost all the years studied had outliers. There are three regencies that tend to be outliers, namely Jember, Malang, and Surabaya. The number of family planning participants in those districts is quite different with the number of other districts, thus indicating outliers. Multicollinearity checks among variables were based on VIF (Variance Inflation Factor) values. VIF values greater than five (VIF> 5) indicate the occurrence of multicollinearity. Based on Table 1, it is known that all VIF values in explanatory variables are less than five so it can be said that the explanatory variables are independent and there is no multicollinearity problem,

![Figure 2. Boxplot of Number Family Planning Participants in East Java in 2009-2016](image)

Table 1: VIF values of each explanatory variable

| Variable | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|----------|-------|-------|-------|-------|
| VIF      | 3.401 | 1.349 | 1.496 | 3.580 |

Meanwhile, homogeneity testing was determined partially and simultaneously using Breusch-Pagan test. The results of the test in Table 2 show that observations from 2009 to 2011 show no spatial heterogeneity, but for the next year there was spatial heterogeneity among the data. When simultaneous spatial heterogeneity is tested, it is found that there is spatial variation in the data on the number of family planning participants in East Java. Spatial heterogeneity causes a difference in the characteristics of each location. Therefore, it can be concluded that the application of spatial regression using geographical weighted regression appropriately applied to model number family planning participants in East Java to overcome the heterogeneity that occurred.

Table 2: Test of Breusch Pagan Statistics

| Year   | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2009-2016 |
|--------|------|------|------|------|------|------|------|------|-----------|
| p-value | 0.9918 | 0.255 | 0.6965 | 0.0673 | 0.0080* | 0.0296* | 0.8791 | 0.0759** | 0.0499*   |

Description: * significant at $\alpha = 0.10%$; ** significant at $\alpha = 5\%$
The data used in this study involves geographic and time elements that have heterogeneity in each location and time, so the precise estimation is determined using GTWR modeling. In GTWR modeling, the weights are established based on the Euclidean distance matrix using the best selected kernel function based on the smallest Cross Validation (CV). The smallest CV value is generated by Exponential kernel weighted function that is $4.19E11$ with optimum bandwidth of GTWR of $0.265$. In addition, the spatial distance parameter ($\phi_s$) is $1.230$, the temporal distance parameter ($\phi_T$) is $0.0008$, and the optimum ratio parameter ($\tau$) is $0.00072$. The residual of the GTWR modeling are explained with the boxplot in Figure 3. It can be seen that for each year there are errors which have very large/small value compared with the general data. The errors is detected to be outliers, either upper outlier or lower outlier. The existence of the outlier observation may create fictitious structures in the estimates of the coefficients which may mislead the result. Therefore, a more robust modeling is needed to accommodate outliers.

![Figure 3. Boxplot of Residual for GTWR Modelling year 2009-2016](image)

Geographically and temporally weighted regression robust (RGTWR) is a method to overcome the outlier for the estimation generated in fit with the actual data. The RGTWR estimation is determined through the iteration process by utilizing the GTWR parameter as the initial value. The iteration process uses M-estimator until a convergent value is obtained. Each location and time produces a convergent estimators at different iterations, so that to ease the process, the iteration is determined by a maximum of 45 iteration processes with the level of accuracy used 0.0 01.

![Figure 4. Boxplot of Parameter Estimations for (a) GTWR and (b) RGTWR Year 2009-2016](image)

Based on the modeling, it is found that the modeling on RGTWR is considered able to improve the predicted result of GTWR parameter. This is indicated by changes in the direction of the relationship with the coefficient value of the predictor of number of fertile couples and the number of health facilities. In GTWR modeling, the coefficients of these parameters show negative results in some cities, but after the modeling with RGTWR, it is found that the parameters change direction and give positive value more appropriate to the plot of the. In addition, although the errors generated by GTWR and RGTWR tend to be symmetrical, but the residual RGTWR modeling results in a much smaller range and range.
Moreover, the resulting outliers also tend to stay apart and separate from the general distribution of data. Distribution of parameter estimations for GTWR and RGTWR model is described in Figure 4. It is found that the estimated range of parameters $\beta_1$ and $\beta_4$ generated by the RGTWR model is greater than the estimated values of $\beta_1$ and $\beta_4$ parameters in GTWR. In contrast, the estimates of $\beta_2$ and $\beta_3$ generated from the RGTWR method tend to provide a lower estimated range than the GTWR modeling results. Meanwhile, comparison of actual and prediction of the number of family planning participants in East Java year 2012 using GTWR and RGTWR model is illustrated in Figure 5. The RGTWR modeling results estimators are closer to the actual values meaning the RGTWR modeling is able to improve the predictive value. Meanwhile, Figure 7 shows the comparison of residual line boxes from GTWR and RGTWR modeling.

Figure 5. Plot of Actual Data, Result of Prediction of GTWR and GTWR from Number of Family Planning Participants of East Java 2012

The comparison of model goodness based on MAD and $R^{2}_{corrected}$ is shown in Table 3. RGTWR modeling can reduce the value of MAD meaning that the variance error of RGTWR is smaller, resulting in the closest approximation to the actual data. Meanwhile, the $R^{2}_{corrected}$ of RGTWR results in a lower value of about 5% compared to $R^{2}_{corrected}$ of GWRT, which is 0.848. Although the value of $R^{2}_{corrected}$ in GTWR modeling is lower, the RGTWR modeling is still selected as the best model in modeling the number of family planning participants in East Java from 2009 to 2016 because RGTWR is able to decrease the MAD value and the difference of $R^{2}_{corrected}$ is small.

Table 3: Comparison of Model Goodness

| Model   | MAD          | $R^{2}_{corrected}$ |
|---------|--------------|---------------------|
| GTWR    | 18217.240    | 0.897               |
| RGTWR   | 17909.512    | 0.848               |

4. Conclusion

Modelling of RGTWR with M-estimator is quite effective to overcome the problem of outliers that occurred in the data. The results of this study show that RGTWR with M-estimator can change the direction of the parameter estimator coefficient which is more appropriate with the data plot, result the approximate value that is closer to the actual value, create residual that has range tending to be smaller and decrease the MAD value in the model.

Acknowledgments

The author thank to the Department of Statistics, Bogor Agricultural University and Indonesia Endowment Fund for Education (LPDP) for their support.

References

[1] Jamilatuzzahro Caraka R. and Herliansyah R, 2018 *Aplikasi Generalized Linear Model dengan R* 1st ed. Yogyakarta: INNOSAIN Graha Ilmu Yogyakarta.

[2] Zhang H and Mei C, 2011 Local Least Absolute Deviation Estimation of Spatially Varying Coefficient
Models: Robust Geographically Weighted Regression Approaches Int. J. Geogr. Inf. Sci. 25, 9 p. 1467–1489.

[3] Caraka R E et al., 2018 Ecological Show Cave and Wild Cave: Negative Binomial Gllvm’ s Arthropod Community Modelling Procedia Comput. Sci. p. 1–8.

[4] Harris P Fotheringham A S and Juggins S, 2010 Robust Geographically Weighted Regression: A Technique for Quantifying Spatial Relationships between Freshwater Acidification Critical Loads and Catchment Attributes Ann. Assoc. Am. Geogr. 100, 2 p. 286–306.

[5] Sari M S., 2014, Estimasi Parameter Robust Geographically Weighted Regression dengan Metode Robust-M: Studi Kasus: Pemetaan Potensi Pertanian Padi Jawa Timur Tahun, Indonesia.

[6] Afifah R, 2017, Robust Geographically Weighted Regression dengan Metode Least Absolute Deviation pada Kasus Kemiskinan di Pulau Jawa, Universitas Padjadjaran.

[7] Olive D., 2005 Applied Robust Statistics Carbondale: Southern Illinois University.

[8] Draper N R and Smith H, 2002 Applied Regression Analysis 3rd ed Wiley Interscience Publications.

[9] Fox J and Weisberg S, 2002 Robust Regression New York.

[10] Sutikno D U Gio P U Caraka R E and Ahmar A S, 2018 Brief Overview of STATCAL Statistical Application Program J. Phys. Conf. Ser. 1028.

[11] Kurniawan I D Rahmadi C Caraka R E and Ardi T A, 2018 Short Communication: Cave-dwelling Arthropod community of Semedi Show Cave in Gunungsewu Karst Area, Pacitan, East Java, Indonesia Biodiversitas 19, 3 p. 857–866.

[12] Caraka R E and Yasin H, 2017 Geographically Weighted Regression (GWR) Sebuah Pendekatan Regresi Geografis 1st ed. MOBIUS GRAHA ILMU.

[13] Caraka R E and Yasin H, 2018 Spatial Data Panel 1st ed. Wade Group.

[14] Kurniawan I D Soesilohadi R C H Rahmadi C Caraka R E and Pardamean B, 2018 The difference on Arthropod communities’ structure within show caves and wild caves in Gunungsewu Karst area, Indonesia Ecol. Environ. Conserv. 24, 1.

[15] Wang P, 2006, Exploring Spatial Effects on Housing Price: The Case Study of The City of Calgary, University of Calgary.

[16] Huang B Wu B and Barry M, 2010 Geographically and Temporally Weighted Regression for Modeling Spatio-Temporal Variation in House Prices Int. J. Geogr. Inf. Sci. 24, 3 p. 383–401.