IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 6 Issue 11, November 2019
ISSN (Online) 2348 – 7968 | Impact Factor (2019) – 6.248
www.ijiset.com

Modelling and Mapping The Average of Incidence Rate Diarrhea Among Toddlers in Bandung City in 2013-2018 Using Geographically Weighted Regression

Anis Khoirunnisa[1*], I Gede Nyoman Mindra Jaya[2]

[1] Department Statistics, Padjadjaran University, Indonesia

[2] Department Statistics, Padjadjaran University, Indonesia

*email: mindra@unpad.ac.id

Abstract
Diarrhea is an infectious disease that can lead to dehydration and death, especially among toddlers. Efforts to prevent diarrhea cases can be done by controlling the factors that affect the occurrence of diarrheal diseases. Regression analysis is a statistical tool that can be used to model factors affecting high rates of diarrhea. But in this research the object of research is a spatial object, therefore the aspects of the region and time need to be considered. This research is conducted to model and map the cases of diarrhea among toddlers by considering the spatial aspects. The analysis that can be used to handle spatial heterogeneity is Geographically Weighted Regression (GWR). In estimating the parameters, the Ordinary Least Square method tends to produce inaccurate estimation. An alternative method that can be used to estimate parameters if there is spatial heterogeneity is the Weighted Least Square (WLS) method.

Keywords: *Diarrhea, Spatial Heterogeneity, GWR, WLS*

1. Introduction

Diarrhea is a contagious disease that is still a health problem and a global burden. According to the World Health Organization (WHO) in 2009 defines diarrhea as a condition of defecation with a more fluid consistency than usual with a frequency of 3 times or more in a 24 hour period. In Indonesia, based on the Diarrhea Morbidity Survey from 2013 to 2018, it showed an increasing tendency in the Case Fatality Rate (CFR) when extraordinary event of diarrhea. In 2012 the CFR when the extraordinary event of diarrhea was 1.53%, in 2013 was 1.11%, in 2014 it rose to 1.14%, in 2015 it rose to 2.47%, in 2016 it became 3.04 %, in 2017 it decreased to 1.97%, and in 2018 it increased to 4.76%. It can be seen that the CFR when the extraordinary event of diarrhea is still incompatible with the expected CFR when the extraordinary event of diarrhea is <1% [1].

The results of the Basic Health Research in 2018 showed that the province of West Java was one of the provinces that had a greater incidence of diarrhea than the national incidence of diarrhea (6.8%) at 7.4%. The toddler age group is the highest group suffering from diarrhea. In West Java the incidence of toddler diarrhea is also quite high, which is 12.8% where the incidence of diarrhea for toddlers nationally is only 11% [2].

In Bandung City, from 2013 to 2016 decreased and increased from 2016 to 2018. Incidence rate of diarrhea in toddlers in 2018 is still quite high, there are about 9,420 of 100,000 toddlers in Bandung City who have diarrhea [3].

The high incidence rate of diarrhea in toddlers in Bandung City is certainly influenced by several factors. In an effort to reduce the incidence rate of diarrhea in toddlers, it is necessary to control the factors that affect diarrhea in toddlers. One method that can overcome this problem is regression analysis. However, each district in Bandung City certainly has different characteristics. Seeing the geographical conditions, the environment, behavior and the things that against it, led to the spatial effect of spatial heterogeneity [4]. Spatial heterogeneity is a condition which cannot be explained by a global model because the characteristics between the regions of observation vary spatially. Methods that can accommodate spatial effects through Geographically Weighted Regression (GWR).

2. Method

The data used in this study is the average incidence rate of diarrhea among toddlers in Bandung City in 2013-2018 obtained from the Bandung City Health Office. Observed variables include:

Y : the average incidence rate of diarrhea in toddlers (case number/100,000 toddlers)
X_1: percentage of healthy homes (%)
X_2: percentage of healthy latrines (%)
X_3: percentage of clean water (%)
X_4: percentage of decent drinking water (%)
X_5: percentage of clean and healthy life behavior (%)
X_6: percentage of vitamin A (%)
X_7: percentage of poor nutrition status (%)
X_8: population density (person/Km^2)

2.1. Multiple Linear Regression Analysis

Regression analysis is a method generally used to explain the relationship between response variables and predictor variables. The term regression was first introduced by Francis Galton in 1886 [5]. Regression models with several predictor variables can be formulated as in the following:

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

i = 1,…, N and k = 1,…, K
y_i : the i-th observation value of response variable
β_0 : intercept
$\beta_1, \ldots \beta_K$: slop coefficient
X_{ik} : the k-th independent variables for observation i
ε_i : error at location i, assumed ε_i ~ IID (0, σ^2)

These basic assumptions are known as classic assumptions which consist of:

1. $E(\varepsilon_i) = 0$, for i = 1,…, N
2. $Var\ (\varepsilon_i) = \sigma^2$, for i = 1,…, N
3. $Cov\ (\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$

Estimating the parameters of the linear regression model uses the Ordinary Least Square (OLS) method by minimizing the sum of squares error. Estimating parameters of the model is obtained from the following equation:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \quad (2)$$

2.2. Spatial Heterogeneity

Before doing modeling using GWR, first detect the effects of spatial heterogeneity through the Breusch-Pagan test with the following formulation:

$$BP = \frac{1}{2}[\boldsymbol{g'X(X'X)^{-1}X'g}] \sim \chi^2_{(K)} \quad (3)$$

where:

$\boldsymbol{g'} = (g_1, g_1, \ldots, g_N)^T$ with $g_i = \frac{e_i^2}{\left(\frac{e^t e}{n}\right)} - 1$

e_i : observation value at location i of residuals
$\boldsymbol{e}$: residual vector of observation
$\mathbf{X}$: N × (K + 1) matrix that containing vector of the predictor variable

2.3. Non-Local Multicollinearity

The local multicollinearity problem is detected if $\text{VIF}_k(u_i, v_i)$ value of each k-th of predictor variable at location i > 10. Following is the formula $\text{VIF}_k(u_i, v_i)$ value:

$$\text{VIF}_k(u_i, v_i) = \frac{1}{(1 - R_k^2(u_i, v_i))} \quad (4)$$

with $R_k^2(u_i, v_i)$ as coefficient of determination between X_k with other explanatory variables for each location.

2.4. Geographically Weighted Regression

Regression analysis uses the assumption that each observation is independent of each other and homogeneous. This assumption is not fulfilled when using spatial data because each observation point is related to other observation points [6]. Geographically Weighted Regression is a method used to explore spatial effects [7]. The Geographically Weighted Regression model can be written as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{K} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (5)$$

$i = 1, \ldots, N$
with:
y_i : observation value at location i of response variable
$\beta_0(u_i, v_i)$: intercepts at location i
$\beta_k(u_i, v_i)$: local parameter of the k-th predictor variable at location i
$x_{ik} = x_{i1}\ x_{i2}\ \ldots\ x_{iK}$: observation value at location i and of the k-th predictor variable
(u_i, v_i) : coordinate point (latitude, longitude) location i
ε_i : error at location i, assumed ε_i ~ IID (0, σ^2)

The estimated parameters of the GWR model are carried out by the Weighted Least Square (WLS) method, namely by providing a different weight at each observation location. The following is a form of parameter estimation from the GWR model for each location:

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = (\mathbf{X}^T \boldsymbol{W}(u_i, v_i)\mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{W}(u_i, v_i)\mathbf{y} \quad (6)$$

The weighting used in estimating GWR model parameters is the Gaussian fixed kernel function expressed in the following equation:

$$w_{ij} = exp\left(-\frac{1}{2}\left({}^{d_{ij}}/_{b}\right)^2\right) \quad (7)$$

In forming the weighting function, the optimum bandwidth value is needed. The method often used in determining the optimum bandwidth is Cross Validation (CV) with the following formula:

$$CV(b) = \sum_{i=1}^{N} [y_i - \bar{y}_{\neq i}(b)]^2 \quad (8)$$

To see whether the modeling using GWR produces a better model, the suitability of the model is tested using the F test statistic. Furthermore, the significance of the parameter testing is done using the t-test statistic to determine which parameters significantly affect the response variable at each observation location. The coefficient determination (R^2) needs to be calculated because it is the most commonly used quantity to measure how far the ability of

the model in explaining the variation of response variables in research.

3. Result and Discussion

3.1. Data description

The following is a map of the average incidence rate of diarrhea for toddlers in Bandung from 2013 to 2018:

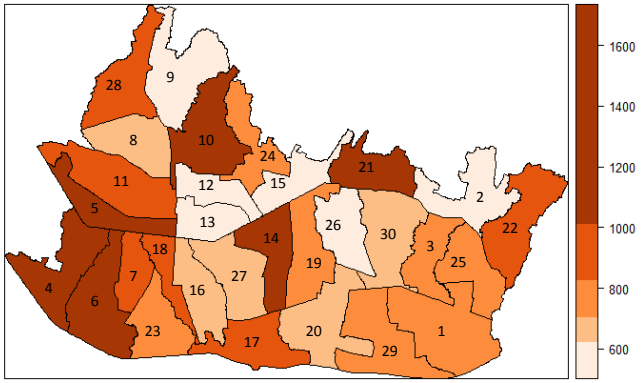


Figure 2 Map of Average Incidence Rate of Diarrhea Among Toddlers in Bandung City

Based on Figure 2 it can be seen that the locations that have the darkest colors (has a high average of incidence rate) are Bandung Kulon District (Number 4), Andir (Number 5), Babakan Ciparay (Number 6), Coblong (Number 10), Batununggal (Number 14), and Mandalajati (Number 21)). Whereas Ujung Berung District (Number 2), Cidadap (Number 9), Bandung Wetan (Number 12), Sumur Bandung (Number 13), Cibeunying Kidul (Number 15), and Antapani (Number 26) has the brightest colors (has a high average of incidence rate).

3.2. Ordinary Least Square (OLS)

The following are the results of modeling using the Ordinary Least Square (OLS) method:

Table 1 Estimates Parameters of Global Model

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4137	1044	3.9620	0.0007
β_1	10.3500	6.8290	1.5160	0.1444
β_2	-20.7100	6.5720	-3.1520	0.0048
β_3	0.7843	6.4810	0.1210	0.9048
β_4	-12.5000	4.7660	-2.6230	0.0159
β_5	-23.5600	7.6630	-3.0740	0.0058
β_6	-15.8100	7.2110	-2.1930	0.0397
β_7	-77.6600	51.7800	-1.5000	0.1486
β_8	0.0316	0.0065	4.8960	0.0001

So the regression model obtained is as follows:

$$\hat{y} = 4137 + 10.35X_1 - 20.71X_2 + 0.78X_3 - 12.5X_4 - 23.56X_5 - 15.81X_6 - 77.66X_7 + 0.0316X_8$$

In this section, modeling the average incidence rate of diarrhea in toddlers in Bandung City using the Geographically Weighted Regression (GWR) method. The model assumption test is first performed to answer whether a regression model is valid or not is used as an explanation for the influence between predictor variables on the response variable. Testing of residuals using Kolmogorov-Smirnov concluded that residuals were normally distributed, testing using Durbin Watson concluded that there were no autocorrelations, both positive and negative, and based on the VIF value obtained, each variable had a VIF value of less than 10, which means that there was no multicollinearity problem.

Testing spatial heterogeneity with the Pagan Breusch test gives a BP value of 20,803 while $\chi^2_{0.05;8}$ of 15,503 so that from the test results it can be concluded that there is spatial heterogeneity. In addition, based on local VIF values obtained, each variable at each location has a local VIF value of less than 10, which means there is no local multicollinearity problem. Therefore, analysis using GWR is appropriate.

3.3. Geographically Weighted Regression

GWR modeling is done by entering the spatial weighting using the weighted least square method. In estimating the parameters of the GWR model a spatial weighting matrix is needed based on the Euclidean distance matrix and the optimum bandwidth value. Based on the calculation results, the optimum bandwidth value is 0.1838733. The weighting matrix obtained for each location is then used to form the model so that each location has a different model. The summary of the estimated parameters of the GWR model can be seen in the table as follows:

Table 2 Summary of Estimated Parameters Values of GWR Model

Parameter	β(ui,vi) Value		
	Min	Median	Max
(Intercept)	4046.1	4151.083	4251.668
β_1	10.06081	10.39174	10.7653
β_2	-21.2169	-20.9213	-20.5626
β_3	0.387725	0.81123	1.542027
β_4	-13.1151	-12.5297	-11.7093
β_5	-24.712	-23.8354	-22.46
β_6	-16.5275	-15.8015	-15.3368
β_7	-80.1883	-77.8096	-74.6733
β_8	0.0312	0.031719	0.032052

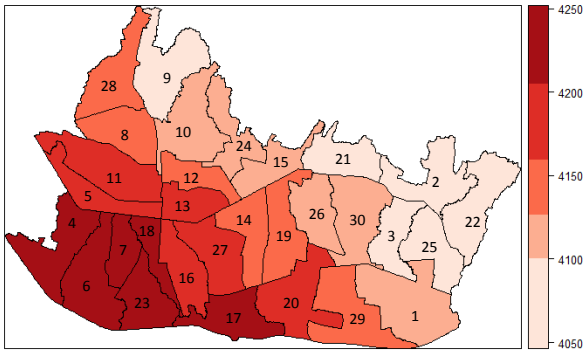
From Table 2 it can be seen that the variables that have a positive influence are the variables of healthy houses, clean water, and population density, while the other

variables have negative influences. The estimated parameter values obtained are different for each location, the following models are formed based on the estimated parameters in the Kiaracondong District:

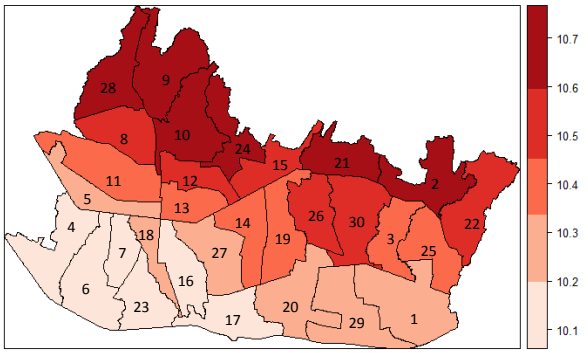
$$\hat{y}_{Kiaracondong} = 4138.654 + 10.383X_1 - 20.875X_2 + 0.771X_3 - 12.328X_4 - 23.458X_5 - 16.006X_6 - 76.965X_7 + 0.032X_8$$

Mapping of the estimated β parameter can be seen in the following figure:

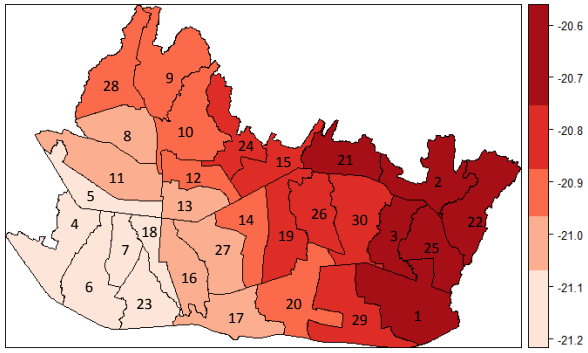
(a) Intercept (β_0)



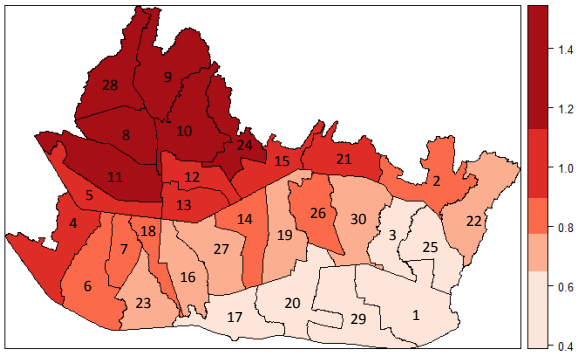
(b) Healthy Home (β_1)



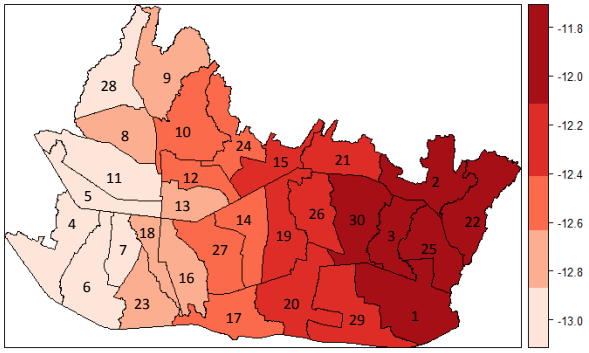
(c) Healthy Latrines (β_2)



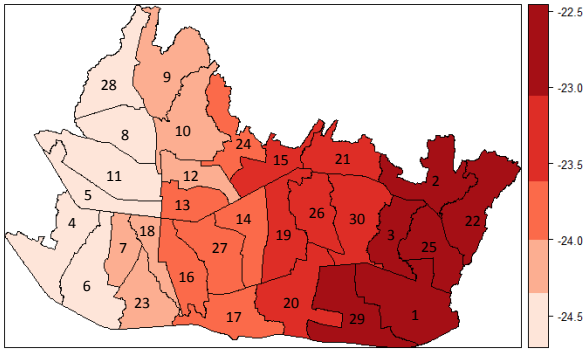
(d) Clean water (β_3)



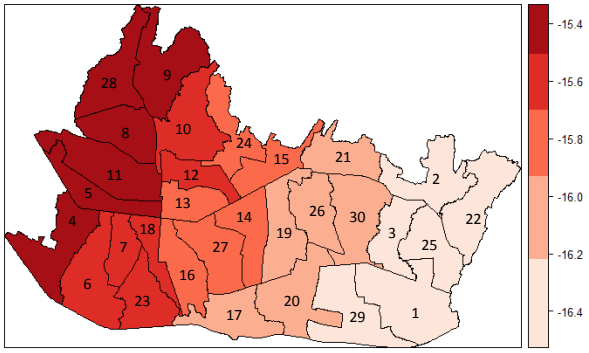
(e) Decent Drinking Water (β_4)



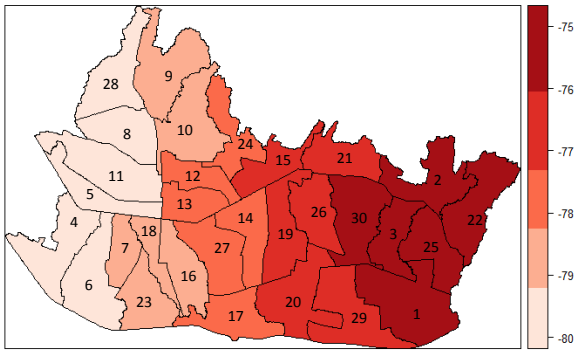
(f) Clean and Healthy Behavior (β_5)



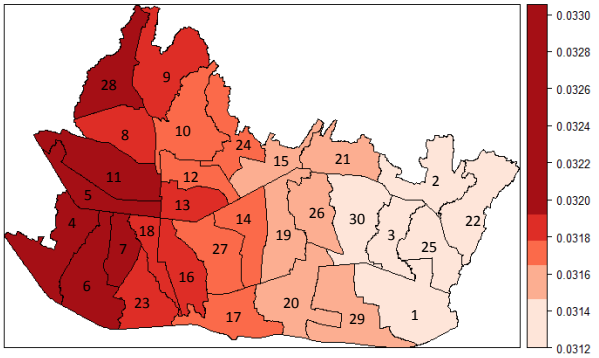
(g) Provision of Vitamin A (β_6)



(h) Malnutrition Status (β_7)



(i) Population density (β_8)



3.4. Model Conformity Test

Testing the suitability of the GWR model obtained from the test with the F-statistic gives a $F_{calculate}$ of 2.2408 while the F_{table} of 2.0102 so that the decision is to reject H_0 means that it can be concluded that the GWR model is better than the global model (multiple linear regression analysis).

3.5. Test the Significance of Parameters

After obtaining the estimated value of the parameters, testing the significance of the parameters needs to be done to find out the predictor variables that have a significant effect on the average incidence rate of diarrhea in toddlers in each district. If the value of $|T_{calculate}| \geq t_{(\alpha/2;n-p-1)} = t_{(0.025;21)} = 2.0796$, the k-th parameter at location i significantly influences the average incidence rate of diarrhea in toddlers. The following is a grouping table for districts based on significant predictor variables:

Table 3 Classification Based on Factors Affecting the Average Incidence Rate of Diarrhea in Toddlers in Bandung City

Local Significant Variables	Districts
X_2, X_4, X_5, X_6, X_8	Ujung Berung, Mandalajati, Rancasari, Gede Bage, Cidadap, Panyileukan, Bandung Kidul, Cibiru, Antapani, Buah Batu, Cinambo, Bandung Wetan, Arcamanik, Cibeunying Kaler, Sukajadi, Coblong, Batununggal, Regol, Lengkong, Cibeunying Kidul, Bojongloa Kidul, Bandung Kulon, Babakan Ciparay, Andir, Bojongloa Kaler, Cicendo, Sumur Bandung, Astanaanyar, Kiaracondong

The coefficient of determination generated by the global model and the GWR model can be seen in the following table:

Table 4 Comparison of the Coefficient of Determination

Global Model	Local Model
52.21%	66.76%

Based on Table 4, it can be seen that in the local model (GWR) of 66.76% the predictor variables are able to explain the response variable (average incidence rate of diarrheain toddlers in Bandung City) and the remaining 33.24% is explained by other variables outside the model. Whereas the global model (multiple linear regression) is only 52.21%, the predictor variables are able to explain the response variable (average incidence rate of diarrheafor toddlers in Bandung City) and the remaining 47.79% is explained by other variables outside the model.

IV. Conclution

Based on the results and discussion, it can be concluded that each district has different estimated parameter values, so the model formed is different for each district. Modeling using local regression (GWR) gives better results than the global model (multiple linear regression), this is supported by the R^2 value of the GWR model (66.76%) which is greater than multiple linear regression (52.21%).

Acknowledgments

Based on the results of this study, the government is expected to give more attention to districts that have a high contributing factor to cause diarrhea in toddlers and the need for socialization to the community related to significant factors that affect toddlers diarrhea in order to increase public awareness so that the community can participate active in efforts to prevent diarrhea in toddlers and is expected to be able to reduce the incidence rate of diarrheafor toddlers in Bandung City.

Reference

[1] Kementrian Kesehatan Republik Indonesia. (2019). *Profil Kesehatan Indonesia 2018*. Jakarta: Kementrian Kesehatan Republik Indonesia.

[2] -----. (2019). *Laporan Nasional Riset Kesehatan Dasar 2018*. Jakarta: Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan.

[3] Dinas Kesehatan Kota Bandung. (2018). *Profil Kesehatan Kota Bandung 2017*. Bandung: Dinas Kesehatan Kota Bandung.

[4] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

[5] Gujarati, D. N. (2004). *Basic Econometrics* (4th ed.). New York: McGraw-Hill Companies, Inc.

[6] Charlton, M., & Fotheringham, A. S. (2009). Geographically Weighted Regression. *White Paper*, National Centre for Geocomputation.

[7] Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. West Sussex: Wiley.