Proceedings of Ahmad Dahlan International Conference on Mathematics and Mathematics Education
Universitas Ahmad Dahlan, Yogyakarta, 13-14 October 2017

Application of Geographically Weighted Regression Analysis for Modelling Pneumonia Toddlers in West Java Province

Intan Nurkartri Utami[1], Jaka Nugraha[2]

[1,2]Statistic Department of Universitas Islam Indonesia, Yogyakarta, Indonesia

Email: 13611224@students.uii.ac.id

Abstract. SDGs or Sustainable Development Goals is a joint development program until 2030 that is agreed by various countries including Indonesia in the UN resolutions forum. One of SDGs targets is to terminate infant and toddler deaths. Pneumonia is the world's leading cause of death among children, and Indonesia has become the top 10 countries with the highest toddler mortality rate due to pneumonia. West Java become the first position on the percentage estimation of pneumonia toddlers cases in java island, with different percentage of pneumonia toddlers for each district. The variation allows the influence of location to the pneumonia toddlers. this study aims to determine the factors that affect the percentage of pneumonia toddlers in every district / city in West Java by using GWR analysis. Geographically weighted Regression is the development of linear regression used to analyze spatial data. The final result shows that the modeling of pneumonia toddlers in West Java shows spatial influence. The GWR model yields R-square value of 78.08% greater than the linear regression model that is 50.49%. 8 groups are formed which each have a factor criterion that influences for each different location.

1. Introduction

SDGs or Sustainable Development Goals is a joint development program until 2030 that is agreed by various countries including Indonesia in the UN resolutions forum. There are 17 goals and 169 targets with one of the goals being a healthy and prosperous life, one of its 169 targets is to end infant and toddler deaths that can be prevented, with all countries trying to reduce neonatal mortality and under-five mortality [1]. Some of the major causes of child mortality globally are pneumonia, diarrhea and malaria that claim about 6000 children under five each day [2].

Pneumonia killed 920136 children under the age of 5 in 2015, accounting for 16% of all deaths of children under five years old. Pneumonia is an acute respiratory infection that attacks the lungs of the alveoli, when an individual has pneumonia, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake [3]. Pneumonia affects children and families everywhere, but is most prevalent in South Asia and sub-Saharan Africa. Mortality due to childhood pneumonia is strongly linked to poverty-related factors such as undernutrition, lack of safe water and sanitation, indoor air pollution and inadequate access to health care. Indonesia has become the top 10 countries with the highest toddler mortality rate due to pneumonia [4].

West Java become the first position on the percentage estimation of pneumonia toddlers cases in java island with the percentage of value 4.62% [5]. West Java Province consists of 18 regencies and 9 cities. Based on the report of West Java health profile known Number of cases of pneumonia found in toddlers in West Java in 2015 amounted to 166,888. There are 27 areas in west java with different percentage of pneumonia toddlers for each district. The variation in percentage of pneumonia in toddlers in each location / area allows the influence of location or spatial factors. Risk factors that contribute to the incidence of pneumonia also vary among others low nutrition, exclusive low breastfeeding, indoor air pollution, density, low coverage of immunization and low birth weight [6].

GWR is a development of classical regression analysis that is quite effective in performing parameter estimation on data with spatial heterogeneity. In the GWR model, the resulting regression parameters are local, so each location of observation has different regression coefficient value [7].

The previous research on the use of GWR methods such as Santoso in 2012 about external factors of pneumonia in toddlers in East Java. The data used are secondary data about pneumonia of children of East Java in 2009. Variable used in this research is percentage of malnutrition balita, percentage of toddlers get vitamin A supplementation, percentage of children under five who get immunization [8].

Subsequent research was conducted by Dzikrina in 2013 using geographically weighted regression (GWR) on leprosy data in East Java by using 8 independent variables and it was concluded that the GWR model in leprosy prevalence rate was better than modeling with multiple linear regression [9].

The research using the percentage data of children with pneumonia in West Java in 2015 using GWR method has not been done by any party before. Therefore in this research will be discussed the factors that percentage of pneumonia sufferer in balita in West Java by considering geographical factor.

2. Method

The population used in this study is the percentage of children under five suffering from pneumonia in all regencies / cities located in West Java consisting of 9 cities and 18 regencies. The data used in this study is secondary data obtained from the West Java Provincial Health Office website. In this study, secondary data collection is done by downloading data from data sources previously described. The dependent variable used by the researcher for GWR analysis is the percentage of toddlers suffering from pneumonia in West Java in 2015 by Regency / City (y). Then the independent variables used by the researcher are Non-Complete Basic Immunization (x_1), percentage of infants given Exclusive breastfeeding (x_2), percentage of low birth weight (x_3), Percentage of the Poor (x_4), Percentage of BGM (x5), Percentage of Household Not PHBS (x_6).

The method of analysis used is Geographically Weighted Regression (GWR) that is, linear regression method that given the weight of the location. The software used in this analysis is GeoDa, Table and R Studio software. Here are some of the formulas used in gwr analysis.

2.1 Breusch Pagan

Breusch Pagan test is used to find the value of spatial heteroscedasticity. The following is the formula used in the breusch pagan test

$$BP = \frac{1}{2} f^T z(z^T z)^{-1} z^T f \sim x^2_{(k+1)} \quad (1)$$

with:

e_i : The error value for the i th observation.

z : Matrix size n x $(k+1)$ which contains a vector of x that has been standardized for each observation.

f : Vector (n x 1).

n : Number of observation areas

k : The number of explanatory variables.

σ^2 : Variety of remaining e_i.

Criteria for decision-making can also be done by comparing p-value with α. If p-value < α H0 is rejected, so it can be concluded that there is spatial heterogeneity.

2.2 Indeks Moran

Moran index is used to find the spatial autocorrelation value, the following is the formula used in the moran index test

$$I = \frac{n\sum_{i=1}^{n}\sum_{l=1}^{n} w_{il}(Y_i - \bar{Y})(Y_l - \bar{Y})}{\sum_{i=1}^{n}\sum_{l=1}^{n} w_{il}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (2)$$

with :

Y_i = Location variable data of i-th (i = 1, 2, ..., n)
Y_l = Location variable data of l-th (l = 1, 2, ..., n)
$\bar{Y}$ = Average observation data in all regions
w_{il} = Weight the linkage between i and l areas

2.3 GWR Equation

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1} + ... + \beta_k(u_i, v_i)x_{ik} + \varepsilon_i, i = 1,2,...,n \quad (3)$$

with:

y_i	:	Dependent variable at i-th location
(u_i,v_i)	:	Geographical location coordinates (longitude. Latitude) at i-th location
X_{ik}	:	The independent variable j-th at the i-th location
$\beta_k(u_i,v_i)$	:	The coefficient at the i-location corresponding to the j-th independent variable $x_{ji.}$
$\beta_0(u_i,v_i)$	:	Constant GWR
ε_i	:	Error at the i-location point assumed to be independent, identical and normally distributed with zero average and variance σ^2

2.4 Overall Test GWR

The significance test of the GWR model is performed to determine whether the GWR model is better used than by using a linear regression model

$H_0 : \beta_j(u_i,v_i)=\beta_j$, j=1,2,…,k and i=1,2,…,n
$H_1 : \beta_j(u_i,v_i)\neq\beta_j$, j=1,2,…,k and i=1,2,…,n
The test criteria used are, if $F_{hit} \geq F_{\alpha;(db1,db2)}$, H_0 is rejected

2.5 Partial Test GWR

This test is a test used to investigate the independent variables that significantly affect the dependent variables for each region analyzed

H0 : β_j (ui,vi)= 0, j=1,2,…,k
H1 : β_j (ui,vi)≠0
The test criteria used are, if t_{hit} = t ($\alpha_{/2}$,df) H_0 is rejected

3. Result and discussion

Figure 1 is the toddlers with pneumonia in west java in 2015. Based on figure 1, the highest percentage of toddlers with pneumonia is found in 4 areas, the largest percentage of toddlers with Pneumonia sufferers in Cirebon city. Then followed by Ciamis regency, Bandung city and Cirebon regency. The lowest percentage of toddlers with pneumonia is found in 10 areas, namely Indramayu, Bekasi, Depok City, Bekasi City, Pangandaran, Sukabumi, Tasikmalaya, Bogor, and Cianjur. Base on figure 1 it can be seen that the spread of pneumonia in toddlers form patterns and groups in certain areas.

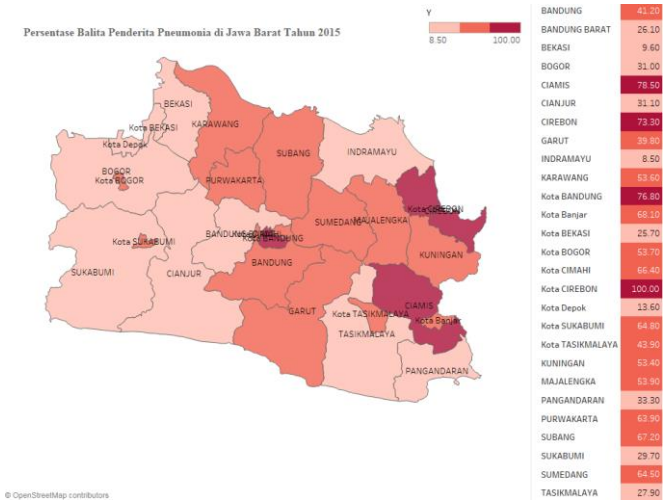

Figure 1. Percentage of pneumonia toddlers in west java

3.1. Regression linier model

From the result of linear regression analysis using software R Studi obtained linear regression model as follows

Table 1. Coefficient output

Variable	*Coefficient*	p-*value*	t_{hitug}	conclusiom
Constant (β_0)	106.49092	0.000128	4.731	Significant
Non-Complete Basic Immunization (x_1)	-0.97923	0.001125	-3.799	Significant
Exclusive breastfeeding (x_2)	0.05243	0.731249	0.348	Not Significant
low birth weight (x_3)	4.47206	0.010984	2.803	Significant
the Poor (x_4)	0.96229	0.481342	0.718	Not Significant
BGM (x_5)	-2.88024	0.088403	-1.791	Not Significant
Household Not PHBS (x_6)	-1.27229	0.018286	-2.570	Significant

Table 2. Partial test and model

Variable	*Coefficient*	p-*value*	t_{hitug}	conclusiom
Constant (β_0)	92.2972	5.11e-05	4.963	Significant
Non-Complete Basic Immunization (x_1)	-0.8419	0.00132	-3.654	Significant
low birth weight (x_3)	3.4878	0.03172	2.287	Significant
Household Not PHBS (x_6)	-0.8126	0.04934	-2.075	Significant

Table 3. Anova regression linier

R_{square}	F_{hitung}	p-*value*
0.4152	5.443	0.005611

Testing parameters simultaneously is a joint test of all parameters in the regression model. The hypothesis is

$$H0 : \beta1=\beta2=\ldots=\beta j = 0$$

H1 : There's at least one of them $\beta j \neq 0$, j = 1,2,….,k

Based on table 3, it is found that p-value = 0.005611 therefore p-value <α so that the decision is rejected H0 which means that at least one independent variable has an effect on the percentage of toddlers with pneumonia in West Java 2015.

Based on table 1 it can be seen that the p-value for the independent variable Non Complete Basic Immunization, Low Birth Weight, and Household not PHBS is smaller than α value, with a confidence level of 95%. Therefore, it can be decided that reject H0 for 3 parameters of the independent variable. So it can be concluded that the independent variables that affect the variable percentage of toddlers with pneumonia that is the variable Non complete Basic Immunization, Low Birth Weight, and Household Not PHBS. Based on table 2 can be seen a significant variable, therefore the model estimator formed from linear regression analysis is:

$$y = 92.2972 - 0.8419x_1 + 3.4878x_3 - 0.8126x_6$$

From the data that has been analyzed and the model formed in get the coefficient of determination (R^2) of 0.4152 illustrates that 41.52% variance that occurs in the dependent variable (y) can be explained by the independent variable (x). The rest of 0.6948 (69.48%) is explained by other variables not included in the model.

3.2. GWR Analysis

Prior to GWR analysis, the classical assumption is assumed, from the classical assumption results concluded that the data being analyzed does not contain multicolinearity, there is no autocorrelation, is normal, because GWR analysis will be done then the data must be heteroscedasticity.

After testing the classical assumption, spatial heterogeneity test is done by using Breush Pagan Test (BP), from this test get the value of BP equal to 0.018 which means smaller than 0.05 which means there is diversity between regions, so with the knowledge of BP value is smaller then GWR analysis can be continued. After performing the breush pagan test, the next step is to look at spatial autocorrelation values using moran's test.using moran's test.

Table 4. Value of Moran's I

Var	Moran's I
Y	0.183054
x_1	-0.0969853
x_2	0.0636082
x_3	0.0757385
x_4	0.430401
x_5	-0.0386665
x_6	0.158256

$I_0 = -\frac{1}{n-1}$

$I_0 = -0.0384615$

$H_0 : I = 0$

$H_1 : I \neq 0$

Based on Table 4, there are three variables that have Moran's I values greater than I0 = -0.0384615 indicating that there is a positive autocorrelation or clumping pattern and have similar characteristics in adjacent locations. As for the non-complete basic immunization variable (x_1) and the variable BGM (x_5) has a Moran's I value smaller than - 0.0384615. This indicates that the patterned data is spreading.

Table 5. Anova GWR

Model	SS	Df	F_{hitung}	F_{tabel}	P_{value}
GWR *Improvement*	3864.692	20.312	2.781	2.27577	0.01992
GWR *Residuals*	3068.870	16.434			

H0: βj (ui,vi)=βj , j=1,2,….,k dan i=1,2,…,27
H1: There's at least one of them βj(ui,vi)≠βj , j=1,2,….,k dan i=1,2,…,27

After GWR analysis using R studio program it is found that Fcount = 2.781 with value F0.05 (20.312,16.434) = 2.27577 so that Fcount> F0.05 (db1, db2) so the decision is reject H0 means there is influence of geographical factor on model formed

Table 6. Model regression linier and model GWR

Method	R^2	AIC
Regresi Linear	0.4152	240.9231
GWR	0.7808672	215.6791

From Table 6 we get R2 value for linear regession 0.4152 and R2 value for GWR is 0.7808672. The purpose of R2 = 0.4152 = 41.52% is to illustrate that 41.52% of the variance occurring in the dependent variable (y) can be explained by the independent (x) variable possessed. The remaining amount (69.48%) is explained by other variables not included in the model, for linear regression analysis whereas for GWR 78.08% the variance occurring in the dependent variable (y) can be explained by the independent (x) variable possessed. The remaining amount (21.2%) is explained by other variables not included in the model.

the AIC value obtained in this case is 240.9231 using linear regression and 215.6791 using GWR. It can be concluded that the GWR model is better than linear regression

Table 7. Significant variable base on region of GWR

No	Kabupaten/Kota	Significant Variable
1	Cianjur	x_2,x_3
2	Cirebon, Kota Cirebon	x_1,x_5, x_6
3	Indramayu, Kota Bandung, Kota Cimahi, Majalengka, Purwakarta, Subang, Sumedang, Bandung, Bandung Barat	x_1,x_3,x_5, x_6
4	Karawang	x_1,x_3, x_6
5	Kota Tasikmalaya, Tasikmalaya, Ciamis, Kota Banjar, Kuningan	x_1,x_6
6	Kota Sukabumi, Sukabumi, Garut	x_1,x_2, x_3
7	Kota Bogor, Kota Depok, Kota Bekasi, Bekasi, Bogor	x_1,x_2, x_3,x_6
8	Pangandaran	x_1,x_2, x_6

Table 7 is the result of grouping of variables that have a significant effect on the percentage of pneumonia in toddlers in each regency / city in West Java. One of the models formed based on significant variables such as model in Cianjur with coefficient value can be seen in the appendix..

$$y = 93.285 + 0.1169x_2 + 3.342x_3$$

The model explains that each increment of one unit of exclusive breasfedding variable will increase the percentage value of pneumonia in toddlers equal to 0.1169. and each increase of one unit of Low Birth Weight will increase the value of the percentage of pneumonia in toddlers of 3.342.

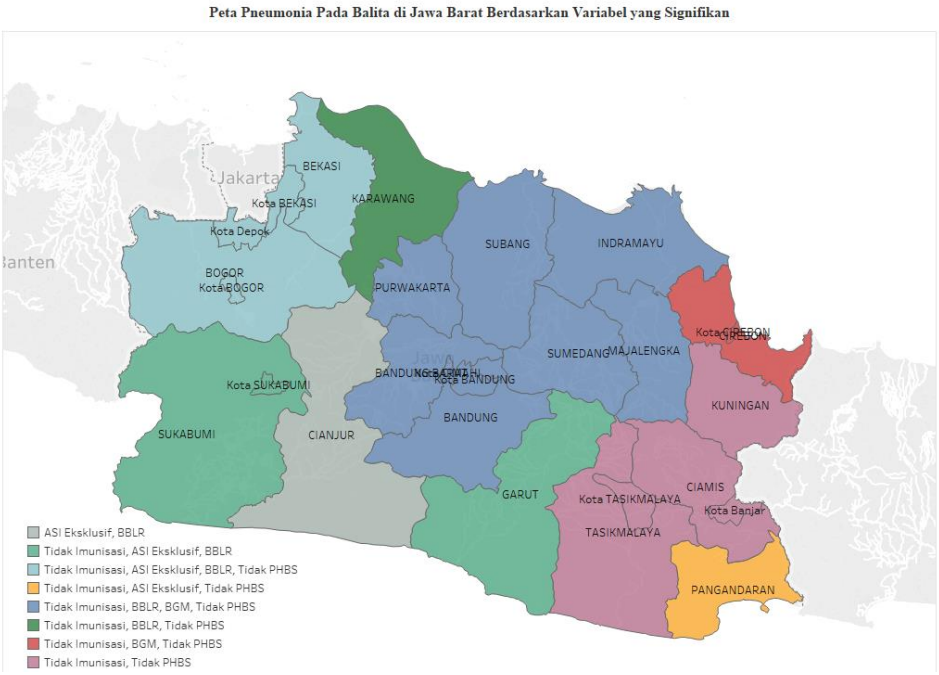

Figure 2. Pneumonia toddlers in West Java by significant variable

4. Conclusion

A significant variable affecting Pneumonia in Toddlers are Non complete Basic Immunization (x_1), Low Birth Weight (x_3) and Household not PHBS (x_6). The predicted model of pneumonia in toddlers globally is

$$y = 92.2972 - 0.8419\,x_1 + 3.4878\,x_3 - 0.8126x_6$$

Significant variables in each District / City are Non complete Basic Immunization (x_1), Exclusive Breastfedding (x_2), Low Birth Weight (x_3), BGM (x_5), Household not PHBS (x_6). There are 8 groups formed based on the significant variables in each District / City. They are 1) Kabupaten Cianjur with significant variable are x_2 dan x_3, 2) Kabupaten Cirebon dan Kota Cirebon with significant variable are x_1, x_5, x_6, 3) Kabupaten Indramayu, Kota Bandung, Kota Cimahi, Kabupaten Majalengka, Kabupaten Purwakarta, Kabupaten Subang, Kabupaten Sumedang, Kabupaten Bandung dan Kabupaten Bandung Barat with significant variable are x_1, x_3, x_5, x_6, 4) Kabupaten Karawang with significant variable are x_1, x_3, x_6, 5) Kota Tasikmalaya, Kabupaten Tasikmalaya, Kabupaten Ciamis, Kota Banjar, Kabupaten Kuningan with significant variable are x_1 dan x_6, 6) Kota Sukabumi, Kabupaten Sukabumi dan Kabupaten Garut with significant variable are x_1, x_2, x_3, 7) Kota Bogor, Kota Depok, Kota Bekasi, Kabupaten Bekasi dan Kabupaten Bogor with significant variable are x_1, x_2, x_3, x_6, 8) Kabupaten Pangandaran with significant variable are x_1, x_2, x_5, x_6

Geographically weighted regression (GWR) is better than Linear Regression model seen based on R2 and AIC values, where R2 value 0.4152 and AIC 240.9231 for Linear Regression and R2 0.7808672 and AIC 215.6791 for GWR

5. Appendices

Kabupaten/Kota	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Cianjur	93.285	-0.913	0.117	3.342	-0.461	-1.908	-0.761
Cirebon	158.330	-0.898	-0.145	2.727	0.070	-2.955	-1.757
Indramayu	136.488	-0.981	-0.046	3.793	0.801	-3.264	-1.628
Karawang	90.295	-1.081	0.240	4.772	1.139	-2.849	-1.114
Kota Bandung	107.624	-0.980	0.048	3.687	0.454	-2.620	-1.111
Kota Cimahi	103.721	-0.990	0.073	3.809	0.492	-2.608	-1.073
Kota Cirebon	165.985	-0.881	-0.176	2.426	-0.128	-2.911	-1.811
Kota Tasikmalaya	159.664	-0.820	-0.186	1.712	-0.611	-2.380	-1.558
Kuningan	175.636	-0.852	-0.231	1.786	-0.467	-2.722	-1.832
Majalengka	144.902	-0.922	-0.103	3.005	0.252	-2.898	-1.596
Purwakarta	95.822	-1.044	0.154	4.337	0.855	-2.778	-1.077
Subang	104.930	-1.058	0.095	4.317	1.099	-3.100	-1.221
Sumedang	123.188	-0.965	-0.023	3.470	0.483	-2.809	-1.328
Tasikmalaya	162.444	-0.811	-0.198	1.571	-0.688	-2.349	-1.575
Kota Banjar	183.871	-0.826	-0.281	1.287	-0.634	-2.577	-1.881
Kota Sukabumi	82.777	-1.040	0.319	4.674	0.153	-1.996	-0.920
Sukabumi	82.846	-1.038	0.316	4.653	0.141	-1.994	-0.915
Kota Bogor	81.608	-1.129	0.463	5.561	0.705	-2.266	-1.147
Kota Depok	82.181	-1.160	0.501	5.865	1.047	-2.551	-1.247
Bandung	111.500	-0.939	0.016	3.315	0.149	-2.434	-1.088
Bandung Barat	101.438	-0.996	0.090	3.881	0.509	-2.596	-1.052
Ciamis	161.587	-0.816	-0.193	1.641	-0.646	-2.374	-1.575
Pangandaran	199.250	-0.875	-0.342	1.387	-0.377	-2.843	-2.171
Kota Bekasi	81.757	-1.128	0.437	5.608	1.268	-2.366	-1.210
Bekasi	84.775	-1.104	0.347	5.205	1.243	-2.615	-1.153
Bogor	83.097	-1.197	0.568	6.171	0.835	-2.878	-1.296
Garut	78.302	-1.153	0.528	6.677	0.824	-2.716	-1.212
Ciamis	93.285	-0.913	0.117	3.342	-0.461	-1.908	-0.761

6. References

[1] BAPPENAS 2015 Kesehatan Dalam Kerangka Sustainable Development Goals retrieved from http://www.sdgsindonesia.or.id/index.php/dokumen?start=5

[2] Unicef 2013 Sekitar 150.000 anak Indonesia meninggal pada tahun 2012 retrieved from https://www.unicef.org/indonesia/id/media_21393.html

[3] WHO 2016 Pneumonia retrieved from http://www.who.int/mediacentre/factsheets/fs331/en/

[4] Unicef 2016 Pneumonia retrieved from https://www.unicef.org/health/index_91917.html

[5] Kementrian Kesehatan RI 2016 Profil Kesehatan Indonesia tahun 2015 retrieved from www.depkes.go.id/.../profil-kesehatan-indonesia/profil-kesehatan-Indonesia-2015.pdf

[6] Dinas Kesehatan Jawa Barat 2016 Profil Kesehatan Provinsi Jawa Barat 2015 retrieved from www.diskes.jabarprov.go.id/index.php/arsip/categories/MTEz/profile-kesehatan

[7] A S Fotheringham, and C Brunsdon, M Chartlon 2002 *Geographically Weighted Regresion, the analysis of spatially varying relationships* (John Wiley and Sons, LTD)

[8] F P Santoso 2012 *Faktor Eksternal Pneumonia Pada Balita di Jawa Timur dengan Pendekatan Geographically Weighted Regression* Thesis Program Studi Statistika Institut Teknologi Sepuluh November

[9] A M Dzikrina dan W P Santi 2013 Pemodelan Angka Prevalensi Kusta dan Faktor-Faktor yang Mempengaruhi di Jawa Timur dengan Pendekatan Geographically Weighted Regression *E-jurnal Sains dan Seni* **2**