Contemporary Mathematics and Applications
Vol. 2, No. 2, 2020, pp. 118-128.

Modeling of Incident Status Dengue Fever in East Nusa Tenggara Using Geographically Weighted Logistic Regression Approach

A. Meylin[1], N. A. Aprilianti[1], D. Lestari[1], N. Chamidah[2,*]

[1]Student of Study Program of Statistics, Department of Mathematics, Faculty of Sciences and Technology, Airlangga University
[2]Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia

[*]Corresponding author: nur-c@fst.unair.ac.id

Abstract. Dengue fever is a disease caused by one of the four dengue viruses and this disease is an infectious disease that is spread through the bite of the Aedes Aegypti mosquito. When compared with the number of dengue cases in previous years, East Nusa Tenggara (NTT) was one of the provinces that experienced an increase in the number of dengue cases in the last three years. Previous research states that the transmission of dengue fever is caused by several factors, one of which is environmental factors of geographical location so that spatial aspects need to be involved in this study. A the statistical method that can be used to analyze spatial data in the form of a logistic regression equation that has a binary response variable is the Geographically Weighted Logistic Regression (GWLR) method. This study aims to analyze the factors that influence the high number of dengue fever cases in NTT in 2018 using GWLR approach by weighted the Gaussian kernel function. Based on the results of GWLR analysis, the number of rainy days and the number of health workers partially significantly influence the status of dengue fever events in each regency/city in NTT Province in 2018. Based on the calculation of Press's Q value, the prediction in this study was accurate with the accuracy of classification was 0.8636 or 86.36%.

Keywords: *east nusa tenggara, geographically weighted logistic regression, incident status dengue fever*.

1 Introduction

Indonesia as one of the tropical countries certainly faces many problems of tropical diseases. Starting from leprosy, malaria, chikungunya to dengue fever. In fact, some of these diseases be appointed as endemic diseases in Indonesia because they always occur in each particular period. One of them is dengue fever which always happens every year in several provinces in Indonesia. dengue fever has become an annual outbreak when entering the rainy season. Beginning in 2020 alone thousands of cases and hundreds of deaths from dengue fever have been reported in a number of provinces in Indonesia. Based on data from the Ministry of Health Republic of Indonesia, the number of dengue cases in Indonesia has reached 25693 cases from January to March 15, 2020 [1]. This figure is the highest number of dengue cases in the last three years. Compared to the number of dengue cases in previous years, NTT is one of the provinces experiencing an increase the number of dengue cases in the past three years. In 2020, the highest contributor to dengue cases in NTT was Sikka District with 1396 cases and the number of deaths reached 14 people [2]. The high number of dengue cases is caused by several factors.

The one of the factors causing the increase in dengue fever cases was environmental factors. This was also expressed by [3] in his research on the causes of dengue fever in Bandung Regency which found that environmental factors significantly influence the incidence of dengue fever in Bandung Regency. In addition, [4] analyzed the mapping of dengue fever cases in East Java Province in 2014 and concluded that there was a significant relationship between the number of dengue fever cases with the percentage of healthy homes and the ratio of health facilities. Based on the results of these studies, the authors took the initiative to analyze the causes of dengue fever in NTT Province considering that several regency/city was initially unaffected but lately there have been reported cases of dengue fever that have even become endemic in these districts / cities.

Previous research states that the transmission of dengue fever is caused by several factors, one of which is environmental factors of geographical location so that spatial aspects need to be involved in this study. Therefore, we need a statistical modeling method that takes into account aspects of the spatial approach. One method that can be used is Geographically Weighted Regression (GWR). GWR is spatial modeling with point approach. The GWR model is a global regression model that is converted into a weighted regression model. Each model parameter value is calculated at each geographical location, so each of them has different regression parameter values [5]. Regency/city in NTT Province can be categorized as dengue fever endemic area and not dengue fever endemic area. So that with the categorization of cases of dengue fever events can be analyzed using the logistic regression method by taking into account location factors or spatial factors, namely the Geographically Weighted Logistic Regression (GWLR) method. This research is expected to be a reference for related parties in determining policies related to the control of dengue outbreaks, especially in NTT Province.

2 Geographically Weighted Logistic Regression (GWLR)

GWLR is an extension of the logit model and enables the spatial variation in the log-odds to be determined. The dependent variable is still categorical and the independent variables can be either continuous or categorical [6]. The geographical location is entered into the model through the weighting function. Weight (w_{ij}) is given at each observation. The model of GWLR is as follows:

$$\pi(\mathbf{X}_\mathrm{j}) = \frac{\exp(g(\pi(\mathbf{X}_\mathbf{j})))}{1+\exp(g(\pi(\mathbf{X}_\mathbf{j})))} \tag{1}$$

where $g\left(\pi(\mathbf{X}_\mathbf{j})\right)$ is the logit form for GWLR model as follows:

$$\begin{aligned} g\left(\pi(\mathbf{X}_\mathbf{j})\right) &= \ln\left[\frac{\pi(\mathbf{X}_\mathbf{j})}{1-\pi(\mathbf{X}_\mathbf{j})}\right], \qquad 0 < \pi(\mathbf{X}_\mathbf{j}) < 1 \\ &= \mathbf{X}_\mathbf{j}\, \boldsymbol{\beta}_\mathbf{j}(\mathrm{u_i}, \mathrm{v_i}), \qquad j = 1,2,\dots,p \end{aligned} \tag{2}$$

Weighting is used to give different emphasis to different observations in producing parameter estimators. Spatial weighting function that is commonly used is fixed kernel weighting, here are two examples of fixed kernel weighting:

a) Fixed Gaussian Kernel

$$\mathrm{w_{ij}} = \exp\left[-\left(\frac{\mathrm{d_{ij}}}{\mathrm{h}}\right)^2\right] \tag{3}$$

b) Fixed Bi Square Kernel

$$\mathrm{w_{ij}} = \begin{cases} \left(1-\left(\frac{\mathrm{d_{ij}}}{\mathrm{h}}\right)^2\right)^2, & \text{if } \mathrm{d_{ij}} > \mathrm{h} \\ 0, & \text{if } \mathrm{d_{ij}} > \mathrm{h} \end{cases} \tag{4}$$

with $\mathrm{d_{ij}} = \sqrt{(\mathrm{u_i}, \mathrm{v_i})^2 + (\mathrm{u_j}, \mathrm{v_j})^2}$.

In the formation of a GWLR model, bandwidth a very important role because it will affect the accuracy of the model of the data, namely adjusting the variance and bias of the model. Optimal bandwidth can be chosen based on the smallest Akaike Information Criterion (AIC) value. The AIC formula for the GWLR method is as follows:

$$\mathrm{AIC} = 2\mathrm{n}\log(\hat{\sigma}) + \mathrm{n}\log(2\pi) + \mathrm{n} + \mathrm{tr}(\hat{\mathbf{y}}\mathbf{y}^\mathbf{T}) \tag{5}$$

Testing the similarity between the GWLR model and the logistic regression model aims to determine the significance of geographical factors. The hypothesis used is as follows:

$\mathrm{H_0}$: there is no significant difference between the GWLR model and logistic regression

$\mathrm{H_1}$: there is significant difference between the GWLR model and logistic regression

The test statistics used are as follows:

$$\mathrm{F_{hit}} = \frac{\mathrm{D}(\hat{\beta})/\mathrm{db_1}}{\mathrm{D}(\hat{\beta}(\mathrm{u_i}, \mathrm{v_i}))/\mathrm{db_2}} \tag{6}$$

where $\mathrm{D}\hat{\beta} = -2\sum_{\mathrm{i}=1}^{\mathrm{n}} \pi(\mathrm{x_i})g(\pi(\mathrm{x_i})) + \log(1-\pi(\mathrm{x_i}))$ and

$D(\hat{\beta}(u_i, v_i)) = 2\{\sum_{i=1}^{n}[y_{1i} \ln(\hat{\pi}(x_i)) + y_{0i} \ln(1 - \pi(x_i))]\} - 2\{\sum_{i=1}^{n} n_{1i} \ln(n_{1i}) + n_{0i} \ln(n_{0i}) + n\ln(n)\}$

The test criterion is to reject H_0 if it is of value $F_{hit} > F_{(\alpha,db_1,db_2)}$.

Testing the parameters of the GWLR model partially, with the following hypothesis:

$H_0: \beta_k(u_i, v_i) = 0$

$H_1: \beta_k(u_i, v_i) \neq 0;\ i = 1, 2, \dots, n;\ k = 1, 2, \dots, p$

The test statistics used for this test can be obtained by the Wald test as follows:

$$Z_{hit} = \frac{\hat{\beta}_k(u_i,v_i)}{Se(\hat{\beta}_k(u_i,v_i))}\ ;\ k = 1, 2, \dots, p \qquad (7)$$

Test statistics in equation (7) approach the standard normal distribution. The test criterion is reject H_0 if the value$|Z_{hit}| > Z_{\alpha/2}$ .Value $Z_{\alpha/2}$ can be obtained from the standard normal distribution table.

The definition of odd is the comparison between the probability of success and the probability of failure. The equation for calculating the odd ratio value is as follows (Harlan, 2018):

$$\text{Odd} = \frac{\pi(x)}{1-\pi(x)} = \exp(\mathbf{X}_j\boldsymbol{\beta})\ ; j = 1, 2, \dots, n \text{ dan } 0 < \pi(x) < 1 \qquad (8)$$

While Apparent Error Rate (APPER) is a value that is used to see the chance of error in classifying an object. The APPER value is as follows [7]:

$$\text{APPER} = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \times 100 \qquad (9)$$

where

n_{11} is the number of failed event is correctly classified as a failed event

n_{12} is the number of failed event is incorrectly classified as a successful event

n_{21} is the number of successful event is incorrectly classified as a failed event

n_{22} is the number of successful event is correctly classified as a successful event

Based on the calculation of the APPER value, the accuracy of classification values obtained is equal to:

Accuracy of Classification = 1 – APPER (10)

The final step to assessing the model as a whole is to determine level of prediction accuracy. This determination is carried out using a statistical test called Press's Q with the following hypothesis:

H_0 : inaccurate classification

H_1 : accurate classification

The test statistics used for this test can be obtained by the Press's Q value as follows [8]:

$$\text{Press's Q} = \frac{[N-(qk)]^2}{N(k-1)} \qquad (11)$$

where

N is number of samples

q is many events are correctly classified or $\mathrm{q} = \mathrm{n}_{11} + \mathrm{n}_{22}$
k is number of groups
The test criterion is reject H_0 if the value of $\mathrm{Press's\ Q} > \chi_{(\mathrm{db},\alpha)}$.

3 Data And Step Analysis

The data used in this study were the number of dengue fever events in 2016 – 2018 as response variable and the number of rainy days, the number of health facilities, the number of health workers, the distance to the provincial capital, the amount of rainfall, the percentage of illiterates as predictor variables. This data sourced from the NTT Book in Figures 2019 – 2017 issued by the Badan Pusat Statistik (BPS) or Central Bureau of Statistics of NTT Province. The unit observations of this study were all regency/city in NTT Province which consisted of 22 regencies/cities. Response variables were classified into:

0: regencies/cities that had dengue cases in the period 2016 - 2018 but not every year (non-endemic)
1: regencies/cities that had dengue cases in the period 2016 – 2018 (endemic).

Table 1 Variable Definition

Variable	Variable Name	Variable Captions
Y	Type of dengue fever events	The number of dengue fever events in 2016 – 2018
X1	The number of rainy days	The number of rainy days during 2018
X2	The number of health facilities	The number of puskesmas and hospitals in each district/city in 2018
X3	The number of health workers	The number of doctors in each district/city in 2018
X4	The distance to the provincial capital	Distance of district/city capital to provincial capital in kilometers
X5	The amount of rainfall	The amount of rainfall in millimeters per district/city in 2018
X6	The percentage of illiterates	Percentage of illiterate population in each district/city in 2018

The steps taken in data analysis to achieve the research objectives were:

1. Describe the factors that influence the incidence of dengue fever in NTT based on a thematic map
2. Conduct heterogeneity tests and spatial dependency tests as assumptions that must be met used the Breusch Pagan test and Moran’s I test
3. Modeling and estimating dengue fever incidence data in regency/city in NTT Province based on the Geographically Weighted Logistic Regression (GWLR) approach with the following steps:

   a. Determining latitude (u_{ij}) and longitude (v_{ij}) in each district / city in NTT Province
   b. Calculating the Euclidean distance (d_{ij}) between observations based on geographical location. Euclidian distances were calculated between locations at coordinates (u_i, v_i) and locations with coordinates (u_j, v_j)
   c. Determining the best bandwidth (h) based on the minimum AIC value according to equation (5)
   d. Calculating the weighting matrix using the Kernel function, namely Fixed Gaussian (3) and Fixed Bi Square (4)
   e. Estimating the parameters of the GWLR model by entering all the predictor variables according to equation
   f. Testing the similarity of the model between the GWLR model and the global logistic regression model using equation (6)
   g. Partially testing the significance of parameters (7)
   h. Calculating the odd ratio value according to equation (8)
   i. Calculating the accuracy classification of the GWLR model according to equation (10)
   j. Calculating the Press's Q value according to equation (11)
4. Interpreting factors that significantly influence the incidence of dengue fever in each regency/city in NTT Province based on thematic maps.

4 Result And Discussion

Before estimating the parameters of the model with spatial regression analysis, the data to be used must be tested with a spatial assumption test. Spatial assumption test aims to determine whether the data to be analyzed has fulfilled the basic assumptions of spatial regression. There were two tests of spatial assumptions in this study, which are as follows [9]:

a. Breusch Pagan Test

   This test is conducted to determine the effect of spatial heterogeneity. This study used the Breusch Pagan test method. The hypothesis used was as follows:

   H_0 : there was no spatial heterogeneity
   H_1 : there was spatial heterogeneity

   The result for Breusch Pagan test, where probability value calculation was 0.01538, which fewer than 10% as significance level. Therefore, H_0 was rejected. It concluded that there is spatial heterogeneity on data. Thus, the data has fulfilled the GWLR spatial assumption test, i.e. there is spatial heterogeneity on data.

b. Moran's I Test

   This test is conducted to determine the effect of spatial dependence. This study used the Moran's I test method. The hypothesis used was as follows:

   H_0 : There was no spatial dependencies
   H_1 : There was spatial dependencies

The result for Moran’s I test, where probability value calculation was 0.00256, which fewer than 10% as significance level. Therefore, H_0 was rejected. It concluded that there is spatial dependence on data. Thus, the data has fulfilled the GWLR spatial assumption test, i.e. there is spatial dependence on data.

The next step, we chose the optimum bandwidth based on the minimum AIC value. The results for each weighting given in Table 2 as follows:

Table 2 Comparison of AIC Values of each Weighting Kernel Function

Weighting Kernel Function	**Bandwidth**	**AIC**
Fixed bi-Square	1.375	-248.485
Fixed Gaussian	2.223	23.229

Table 1 showed that the Fixed bi – Square Kernel function had the smallest AIC value compared to other weighting functions. The smallest AIC was -248.485, with the bandwidth value 23.229. Thus, Kernel Fixed bi-square function weighting was used to estimate the best model in this study.

After selected the best weighting functions, the goodness of model test was performed to determine whether the resulting GWR model was suitable for this data. The hypothesis used was as follows:

H_0 : states the GWLR model was not appropriate

H_1 : states the GWLR model was appropriate

With an α of 0.10, the test criteria H_0 was rejected if $F > F_{(\alpha;df_1;df_2)}$ or $F > F_{(0.1,11.936,29)}$ or $F > 1.77996$. The results given in Table 3 as follows:

Table 3 The Conformity Test of The GWLR Model

Model	***Deviance***	**DOF**	**F**
Global logistic regression	15.765	15.000	13.653
GWLR	11.814	153.476	

The result for statistic test, where F value calculation was 13.653, which more than F statistic in table with 10% as significance level, degree of freedom 15 and 153.476, equals to 1.532. Therefore, H_0 was rejected. It concluded GWLR model is better than the global logistic regression model.

Partial testing of GWLR model parameters was performed at each of the i locations. The test statistic used t test with rejected testing criteria if t value calculation more than t statistics in table with 10% as significance level, degree of freedom 16, equals to 1.746. For example, it will be analyzed estimation model in West Manggarai Regency as endemic area. The result partial testing of the parameters for West Manggarai Regency using GWLR model given in Table 4 as follows:

Table 4 Partial Test of West Manggarai Regency Parameters

Parameters	Estimation	t count
β_0	1.102685	0.230808
β_1	-4.698108	-0.072446
β_2	9.723685	0.162948
β_3	-8.123778	-1.751253
β_4	-0.295799	-0.014913
β_5	3.025027	0.073405
β_6	-3.405644	-0.056230

Table 4 showed that β_3 that had a significant effect on West Manggarai Regency. The GWR model for West Manggarai Regency was as follows:

$$\pi(x) = \frac{\exp(f(x))}{1+\exp(f(x))} \tag{12}$$

where

$$f(x) = 1.102685 - 8.123778x_3 \tag{13}$$

thus

$$\pi(x) = \frac{\exp(1.102685-8.123778x_3)}{1+\exp(1.102685-8.123778x_3)} \tag{14}$$

Odd ratio values for West Manggarai Regency was as follows:

$$\text{odd ratio} = \exp(\beta_3) = \exp(-8.123778) = 0.0003 \tag{15}$$

Based on the odd ratio value in equation (15), it was known that the chance of West Manggarai Regency was classified as dengue fever endemic region by 0.0003 times the chance of West Manggarai Regency was classified as non-endemic dengue fever region if the number of health workers increased by one unit. This was the same as the chance that West Manggarai Regency was classified as a non-endemic region more than the chance that West Manggarai Regency was classified as an endemic region if the number of health workers increased.

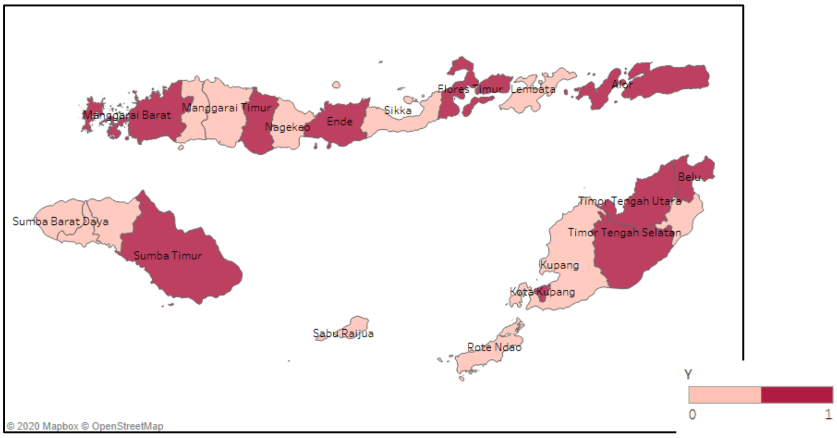

Figure 1 Thematic Map of Incident Status of dengue fever in NTT Province

Figure 1 showed thematic map of incident status of dengue fever in NTT. Code 0 was regency/city non endemic dengue fever, code 1 was regency/city endemic dengue fever. Based on Figure 1, the number of regency/city classified as dengue fever endemic areas was 10, while the number of regencies/cities classified as non-endemic dengue fever was 12.

Table 5 Prediction

Observation	π	$\hat{y}$	**Y**	**Information (Accurate or Inaccurate**
1	0.8808	0	0	**Accurate**
2	0.9999	1	1	**Accurate**
3	4.9697 x 10^{-27}	1	0	**Inaccurate**
4	3.3214 x 10^{-5}	1	1	**Accurate**
5	0.1736	1	1	**Accurate**
6	0.1446	1	1	**Accurate**
7	0.0049	1	1	**Accurate**
8	0.5	0	0	**Inaccurate**
9	1	1	1	**Accurate**
10	0.0024	0	0	**Inaccurate**
11	2.1564 x 10^{-6}	1	1	**Accurate**
12	0.0004	1	1	**Accurate**
13	0	0	0	**Accurate**
14	0	0	0	**Accurate**
15	0	1	1	**Accurate**
16	0.3989	0	0	**Accurate**
17	0.9999	0	0	**Accurate**
18	0.0004	0	0	**Accurate**
19	0.0388	0	0	**Accurate**
20	1.1678 x 10^{-60}	0	0	**Accurate**
21	0.1597	0	0	**Accurate**
22	0.0024	0	1	**Inaccurate**

Table 5 was a table grouping variables that have a significant effect on the response variable in each district/city presented in Table 6.

Table 6 Variables that have a Significant Effect on the Status of dengue fever in NTT Province

Significant Variables	**The province**
X_1	Rote Ndao Regency
X_3	Manggarai and West Manggarai Regency

Based on Table 6 can be known the factors that had a significant effect on each regency/city in NTT Province were presented in the thematic map as follows:

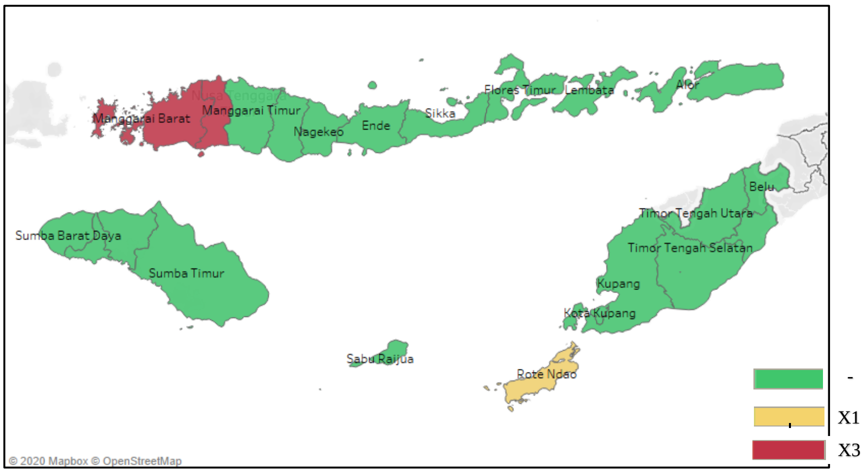

Figure 2 Thematic Map of Factors that significantly influence the Status of dengue fever in NTT Province

Based on factors that significantly influence the GWLR model was obtained to determine the estimated value of the status of dengue fever events in each regency/city in NTT Province. The alleged value obtained from the GWLR model was an opportunity so that a cut point was needed to classify the status category of dengue fever events. The cut point value used in this study was 0.5 where the estimated results of more than 0.5 will be categorized as dengue fever endemic areas and the estimated results less than 0.5 will be categorized as non-endemic areas. Based on the estimation results, a table can be formed to calculate the accuracy of classification as follows:

Table 7 Classification Results of Incident Status Dengue Fever Based on GWLR Method

Observation	**Estimate**	
	Endemic	**Non Endemic**
Endemic	8	2
Non Endemic	0	12

Based on Table 7, APPER value for this estimate was 0.0909. Therefore, the accuracy of classification was 0.9091 or 90.91%. This value more than the accuracy of classification on global logistic regression model, which was 86.36%. It concluded GWLR model is better than the global logistic regression model.

Then, we assessed the model as a whole is to determine level of prediction accuracy use Press's Q value. The hypothesis used was as follows:

H_0 : inaccurate classification

H_1 : accurate classification

The result for statistic test, where χ^2 value calculation was 11.636, which more than $\chi^2_{(0.1;1)}$ equals to 2.705. Therefore, H_0 was rejected. It concluded the prediction was accurate.

5 Conclusion

In 2018 almost 50% regency/city were still classified as dengue endemic areas. There were two variables that significantly influence the status of dengue fever events in NTT Province, namely the number of rainy days and the number of health workers. The variable number of rainy days had significant effect in Rote Ndao Regency while the number of health workers had significant effect in the Manggarai and West Manggarai Regency. Then, based on Press's Q value, the prediction in this study was accurate with the accuracy of classification was 0.8636 or 86.36%.

The East Nusa Tenggara provincial government should pay more attention to regencies/cities that were included in the dengue endemic area by increasing the number and quality of available health workers. The active role of the community in supporting programs made by the government and maintaining environmental cleanliness was also very necessary to eradicate dengue fever given that the weather factors that cannot be controlled.

6 References

[1] Ministry of Health Republic of Indonesia, News and Information Page, < https://www.kemkes.go.id/article/view/20062200001/penambahan-kasus-dbd-masih-tinggi.html>

[2] Ministry of Health Republic of Indonesia, News and Information Page, < https://www.kemkes.go.id/article/view/20030900009/atasi-klb-dbd-di-sikka-menkes-bawa-tim-dokter.html>

[3] Respati, T., Raksanagara, A., Djuhaeni, H., Sofyan, A., Faridah, L., Agustian D., and Sukandar, H., 2017, Journal of Aspirator 9, 91 – 96.

[4] Arniva, N. S., and Purhadi P., 2016, Journal of Sains & Seni 5, 277 – 282.

[5] Fischer M. M., and Getis, A., 2009, Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications (Springer Science & Business Media, London, 2009), pp. 461 – 462.

[6] Kobayashi, A., 2019, International Encyclopedia of Human Geography, (Elsevier, Amsterdam, 2019), pp. 163 – 164.

[7] Harlan, J., 2018, Analisis Regresi Logistik, (Gunadarma, Depok, 2018).

[8] Lomax, R. G., Hahs-Vaughn, D. L., 2013, An Introduction to Statistical Concepts: Third Edition, (Taylor and Francis Group, New York, 2013), chapter 19.

[9] Anselin, L., 1988, Spatial Econometrics: Method and Models (Kluwer Academic Publisher, The Netherlands, 1988).