Malaria Morbidity Modeling in Papua Province, 2016 Using Geographically Weighted Regression (GWR) Method

Cepy Ramdhani, I. Gede Nyoman Mindra Jaya and Yusep Suparman

Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia

Abstract: Malaria is an infectious disease which is still becoming a major public health problem. Based on data from the Ministry of Health in 2016, API in Papua Province, there are 49 cases of positive malaria per 1,000 populations. Modeling the right number of malaria (Morbidity) cases will be an important instrument in eradicating malaria. Regional or spatial based analysis is a good method of describing malaria morbidity because malaria is a mosquito-borne infectious disease that affects one location and another. Geographically Weighted Regression (GWR) is a spatial-based statistical method which develops a global regression model into a local model. The GWR model aims to explore spatial diversity by forming different regression models at each observation location. In general, the results of the GWR analysis in this study indicated that the variables that influenced the model of malaria morbidity in Papua Province were the percentage of clean water usage and the number of health workers per 10,000 populations.

Key words: Spatial · Geographically Weighted Regression (GWR) · Malaria, Papua

INTRODUCTION

Based on the Millennium Development Goals (MDGs), malaria eradication is set as one of the goals that must be achieved by 2015 in the seventh goal of eradicating HIV/AIDS, malaria and tuberculosis [1-2]. Then it is proceed through the third goal in the Sustainable Development Goals (SDGs), namely ensuring a healthy life and attempting to prosper all people, with the specific goal of ending the AIDS, tuberculosis and malaria epidemic and neglected-tropical diseases by 2030 [3].

According to the 2015 World Malaria Report, malaria has attacked almost 106 countries and one third of them are recorded as malaria-high burden countries. Most cases of malaria are found in tropical countries such as in Africa, Latin America and several Asian countries including Indonesia [4]. Malaria is an infectious disease which is still becoming a major public health problem. Malaria is very pivotal on infant morbidity and mortality, children under five and mothers giving birth; besides, malaria also directly reduces work productivity by Soedarto [5]. Malaria can be prevented and controlled by cleaning environments, indoor residual spraying and mosquito nets [6], in other hand traditional malaria preventing techniques are effective for temporarily reducing the severity of the disease. Environmental management activities such as destroying mosquito breeding sites by clearing stagnant water, disposing waste either damping or burning and using toilets properly [7].

Malaria morbidity in a certain area is seen based on the Annual Parasite Incidence (API) as measured by the number of malaria positive cases per 1,000 population in one year. Although the trend of API in Indonesia tends to decline each year, API in Papua Province is still very high, based on data from the Ministry of Health in 2016, API in Papua Province is 49 cases of positive malaria per 1,000 population [8]. Malaria is an infectious disease spreading through the bite of Anopheles mosquitoes. The spread of malaria is strongly influenced by regional elements, where, if there is a malaria endemic area, the surrounding area will also have a high risk in the spread of malaria.

Modeling the right number of malaria (Morbidity) cases will be an important instrument in eradicating malaria. Regression analysis involving several explanatory variables is a statistical method that can be used to determine the factors affecting malaria morbidity.
However, this regression analysis is global which uses the average of each area to produce a parameter score that is general for all areas so that global regression becomes less able to explain the actual phenomenon [9].

According to Anselin [9], if the observation unit is in the form of area and the area of observation tends to be heterogeneous, then the proper analysis to identify the relationship between one response variable with one or more explanatory variables is by using spatial data analysis, because in this case malaria is transmitted through mosquitoes. An area with malaria endemic will definitely affect its surrounding area.

Geographically Weighted Regression (GWR) is one of the spatial-based statistical methods which is the development of a global regression model into a local model. The GWR model aims to explore spatial diversity by forming different regression models at each observation location. GWR model uses weighing matrix which the magnitude depends on the proximity between the observation areas, the closer an area is, the greater the weight will be. This method is quite effective in estimating parameters in data with spatial heterogeneity by Fotheringham [10]. Therefore, it is necessary to know the factors that influence the spread of malaria in a regional or spatial manner.

**MATERIAL AND METHODS**

**Malaria:** an infectious disease caused by parasites (Protozoa) from genus plasmodium, which can be transmitted through the bite of the Anopheles mosquito. Malaria comes from the Italian word, the word *mal* means bad and the word *area* means air. So, in literal meaning, *malaria* means a disease that often occurs in areas with bad air as the result of bad environment in agreement with Zulkoni [11].

According to Soedarto [5] the spread of malaria is caused by various factors including the following:

- Uncontrolled environmental changes.
- The high number of Anopheles mosquitoes.
- Relatively high population mobility towards malaria endemic areas.
- Community behavior that allows contagion.
- The widespread of malaria parasites that is immune to anti-malaria drugs.
- Limited access of health facility to reach all villages with malaria problems, due to geographical, economic and resource constraints.

**Spatial Dependencies:** Spatial dependencies are often measured by spatial autocorrelation statistics that describe the similarity between observations that is close to each other. If the distribution of data in the study area does not have patterns of relationships between that close observations, then the data is referred to as having no spatial autocorrelation by Fotheringham *et al.* [12].

**Spatial Heterogeneity:** According to Fotheringham *et al.* [10] refer to spatial heterogeneity as a condition in which a global regression model cannot explain the relationship between variables due to the characteristics of observational areas that spatially vary. Detecting spatial heterogeneity in the data can be done by the Breusch-Pagan Test by Anselin [9].

**Geographically Weighted Regression (GWR):** One of the spatial-based statistical methods which is the development of a global regression model into a local model [12]. The GWR model aims to explore spatial diversity by forming different regression models at each observation site. The global regression model is a special case of the GWR model, where the parameters in the global regression model are assumed to have constant variance. So, the GWR model shows that there are spatial variations that may be included in the model. The GWR model can be written as follows:

\[ y_i = \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \epsilon_i ; \quad i = 1, 2, ..., n \]

where:

- \( y_i \): Observation score from response variable to-i
- \( x_{ik} \): Observation score of predictor variable to-k on observation to-i
- \( u_i \): Longitude spatial coordinate for the observation to-i
- \( v_i \): Latitude spatial coordinate for the observation to-i
- \( \beta \): Regression coefficient
- \( \epsilon_i \): Error to-i

**Bandwidth Selection:** Bandwidth can be assumed as the radius of a circle, so that a point within the circle radius is still considered to have an influence. According to Fotheringham *et al.* [12] and Fotheringham *et al.* [13], who stated that the optimum bandwidth selection is one of the important things because it will affect the accuracy of the estimation.
One of the methods can be used for bandwidth selection is Cross Validation in all locations or sites, with the formula:

\[
CV = \sum_{i=1}^{n} (y_i - \hat{y}_{ii}(h))^2
\]

where: \( \hat{y}_{ii}(h) \) estimate value \( y_i \) where observation on the location \((u_i, v_i)\) excluded in the process of assessment.

**Selection of Weighting Functions:** In global regression analysis, each observation has the same weight, that is \( w_{ij} \) for each \( i, j \). Whereas the weighting process in the parameter estimation of the GWR model follows Tobler's First Law of Geography, that is the closer data to the location \( i \) will have a stronger influence in predicting the parameter on location \( i \) compared to the farther data. The weighting function used in this research is the Fixed Exponential Kernel function. For samples that have fairly regular spread in the area of research, the weighting function with fixed bandwidth is the most appropriate choice for parameter estimation. Mathematically, the Fixed Exponential Kernel weighting function can be written as follows:

\[
w_{ij}(u_i, v_i) = \exp\left( -\frac{d_{ij}}{h} \right)
\]

where:

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

is Euclidean distance between location \((u_i, v_i)\) with location \((u_j, v_j)\) and \( h \) is optimum bandwidth which is fixed or the same bandwidth being used in every location.

**Data Source:** The data used in this study are secondary data derived from the March 2017 National Socio-Economic Survey (SUSENAS) conducted by the Papua Provincial Statistics Agency [14] and malaria morbidity data from the Papua Provincial Public Health Department [15].

**Research Variables:** The response variable in this study is malaria morbidity or the number of malaria cases per 1,000 populations in Papua Province. While the predictor variables used can be seen in the table below:

<table>
<thead>
<tr>
<th>Research Variables</th>
<th>Response Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria Morbidity :</td>
<td>The number of malaria case per 10,000 populations</td>
</tr>
<tr>
<td>X(_i): Percentage of Healthy Residence</td>
<td>Predictor Variables</td>
</tr>
<tr>
<td>X(_i): Percentage of Clean Water Use</td>
<td></td>
</tr>
<tr>
<td>X(_i): The number of medic per 10,000 populations</td>
<td></td>
</tr>
<tr>
<td>X(_i): The average area height</td>
<td></td>
</tr>
</tbody>
</table>

**Research Steps**

Based on the description above, the steps of the research can be outlined as follows:

- Conducting descriptive analysis
- Modeling the data using global regression
- Identifying multicollinearity of predictor variables
- Identifying spatial dependencies and heterogeneity
- Modeling GWR
- Testing the partial significance of the estimated parameters of the GWR model
- Analysing GWR model.

**RESULTS AND DISCUSSION**

**Descriptive Analysis of Malaria Morbidity:**

Astronomically, Papua Province is located between 2°25'-9°0' South latitude and between 130°0'-141°0' East Longitude. Papua Province is the province with the largest area in Indonesia. The regional office of national land office in Papua province shows that Papua Province has width area of 316,553.07 km\(^2\).

Based on data from the public health department in 2016, Papua Province is still the region with the highest malaria morbidity in Indonesia, which is 49 per 1,000 populations. Keerom Regency is the regency with the highest malaria morbidity with 478 per 1,000 populations and followed by Sarmi Regency with 414 per 1,000 populations. The following is a thematic map of the distribution of malaria morbidity by regency / city in Papua Province.

The explanatory variables in this study are consisted of four variables: the percentage of healthy residences (X1), the percentage of clean water use (X2), the number medic per 1,000 populations (X3) and the average area height (X4). Descriptive statistics for the four variables are shown in the table below:
Fig. 1: Thematic Map of Malaria Morbidity in Papua Province

Table 2: Descriptive Statistic of Explanatory Variables

<table>
<thead>
<tr>
<th>Var</th>
<th>Min</th>
<th>Mean</th>
<th>Med</th>
<th>Max</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_1</td>
<td>0.00</td>
<td>53.73</td>
<td>54.30</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>X_2</td>
<td>0.00</td>
<td>35.58</td>
<td>27.10</td>
<td>99.23</td>
<td>99.23</td>
</tr>
<tr>
<td>X_3</td>
<td>2.13</td>
<td>15.42</td>
<td>12.61</td>
<td>55.98</td>
<td>53.85</td>
</tr>
<tr>
<td>X_4</td>
<td>7.00</td>
<td>822.6</td>
<td>134.0</td>
<td>2513.0</td>
<td>2506.0</td>
</tr>
</tbody>
</table>

Table 3: Parameter Estimation of Global Regression Model

<table>
<thead>
<tr>
<th>Var</th>
<th>β</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-68.3734</td>
<td>-1.007</td>
<td>0.323</td>
</tr>
<tr>
<td>X_1</td>
<td>-0.0306</td>
<td>-0.042</td>
<td>0.967</td>
</tr>
<tr>
<td>X_2</td>
<td>1.4417</td>
<td>2.353</td>
<td>0.027</td>
</tr>
<tr>
<td>X_3</td>
<td>4.9365</td>
<td>3.298</td>
<td>0.003</td>
</tr>
<tr>
<td>X_4</td>
<td>0.0048</td>
<td>0.197</td>
<td>0.845</td>
</tr>
</tbody>
</table>

Based on Table 2, it can be seen that X_1 and X_2 which are the percentage numbers have a very wide range, even the X_1 range reaches 100 which is the maximum range. Likewise with X_3 which is continuous data has also a wide range. In general, because almost all variables have a wide range, there might be a significant characteristics difference among regions.

Based on Figure 2, it can be seen that the score of research variables from regencies/cities that are close to each other has similar values which are shown by the color similarity on the graph. This indicates that the research variables in a regency/city are not mutually independent, there is spatial dependence. However, to ensure that it is important to take spatial dependencies test using Moran Index and spatial heterogeneity test using the Breush-Pagan test which will be discussed in the next section.

**Global Regression Model:** Before modeling using spatial analysis, modeling will be done using global regression, in this case using Ordinary Least Square (OLS). Following is the table of results parameter estimation from the global regression model:

Fig. 2: Thematic Map the Percentage of Healthy Residences (1), the Percentage of Clean Water Use (2), the Number of Medic per 1,000 Populations (3) and the Average Area Height (4) in Papua Province
Based on simultaneous testing at the significance level of 0.05, indicates the value of F = 5.669 or p-value of 0.002 which means that the parameters are significant in the model simultaneously. Whereas based on the partial test results on the parameters of the global regression model shows that there are 2 significant explanatory variables in explaining malaria morbidity at the significant level of 0.05, namely $X_2$ (Percentage of clean water use) and $X_4$ (Number of medic per 1,000 population).

**Identification of Multicollinearity:** To detect multicollinearity can be done by looking at the value of variance inflation factors (VIF). VIF value can be used to detect multicollinearity because this value represents correlation coefficient between predictor variables in the regression model. Here are the VIF values of global regression with explanatory variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>1.0133</td>
<td>1.5817</td>
<td>1.2013</td>
<td>1.8135</td>
</tr>
</tbody>
</table>

Based on Table 4, it can be seen that the VIF value for all variables is less than 10, therefore it can be concluded that there is no multicollinearity among explanatory variables.

**Identification of Spatial Dependencies and Heterogeneity**

The main requirement in Geographically Weighted Regression (GWR) modeling is that there must be spatial dependencies and heterogeneity. To detect it, Moran Index test for dependencies and Breusch-Pagan test for heterogeneity were carried. Following are the results of the two tests:

The Table 5 enumerates that the p-value for the Moran Index is 1.47-e7 so it can be concluded that there are spatial dependencies at the significance level of 0.05. Whereas Breusch-Pagan's p-value is 0.097 which it can be concluded that there is spatial heterogeneity at significance level 0.10. The existence of spatial dependencies and heterogeneity shows that modeling of malaria morbidity in Papua Province cannot use global regression, because it will cause parameter estimation become inefficient. Thus local modeling is needed to take into account to be able to explain the spatial effects on the data.

**GWR Modeling:** In spatial data-based modeling, the assumption is that each observation location relates or influences other observation locations, however the closer range proximity of one location to another location will affect the strength of the relationship so that a weighting is needed. Before determining the weighing matrix, first we need to determine how much bandwidth is appropriate, in order not to produce oversmooth or undersmooth models.

Based on the results of the calculation using fixed exponential kernel function with cross validation technique (CV) optimum bandwidth for this research is obtained and that is 1.531 with CV score 20.6704. The bandwidth score of 1.531 indicates that the observation location which is in the radius of 1.531 degrees from the observation location still has an influence on malaria morbidity at the observation location. Furthermore, the bandwidth is used as a weight of an observation location to other regions so the weighing matrix is obtained for each location.

After obtaining the weighing matrix for each observation location, then modeling using GWR was done using the R program and the result of GWR model parameter estimation on each location which is summarized as follows:

Table 6 shows the minimum, median and maximum score of the GWR model parameters estimation from each observation location. It can be seen that the variables $X_2$ and $X_4$ have a positive relationship with the response variable in all observation locations, it is seen that the minimum and maximum score of parameter estimation from the two variables are positive. Variable $X_2$ shows a negative relationship on response variables in all area. On the other side variable $X_4$ has negative relation in several areas and negative in other locations on response variable. As in global regression, to detect the occurrence of multicollinearity in local regression can also be seen from the local VIF score, with summary as follows:

---

### Table 4: VIF Explanatory Variables Score

<table>
<thead>
<tr>
<th>Variable</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>1.0133</td>
<td>1.5817</td>
<td>1.2013</td>
<td>1.8135</td>
</tr>
</tbody>
</table>

### Table 5: Moran’s I dan Breusch-Pagan Test

<table>
<thead>
<tr>
<th>Test</th>
<th>Score</th>
<th>p-value</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moran’s I</td>
<td>0.0934</td>
<td>1.47-e7</td>
<td>Reject $H_0$</td>
</tr>
<tr>
<td>Breusch-Pagan</td>
<td>7.8578</td>
<td>0.097</td>
<td>Reject $H_0$</td>
</tr>
</tbody>
</table>

Source: R Processing Result

---

### Table 6: Summary of GWR Model Parameter Estimation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Med</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.1896</td>
<td>0.0509</td>
<td>0.1932</td>
</tr>
<tr>
<td>$X_1$</td>
<td>-0.1766</td>
<td>0.0181</td>
<td>0.1402</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.2828</td>
<td>0.4245</td>
<td>0.6720</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.2687</td>
<td>0.4359</td>
<td>0.8911</td>
</tr>
<tr>
<td>$X_4$</td>
<td>-0.0748</td>
<td>-0.0124</td>
<td>0.1301</td>
</tr>
</tbody>
</table>

Source: R Processing Result
Table 7: Summary of VIF Local Score

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Average</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>1.008</td>
<td>1.033</td>
<td>1.061</td>
</tr>
<tr>
<td>X₂</td>
<td>1.243</td>
<td>1.526</td>
<td>1.913</td>
</tr>
<tr>
<td>X₃</td>
<td>1.168</td>
<td>1.236</td>
<td>1.341</td>
</tr>
<tr>
<td>X₄</td>
<td>1.478</td>
<td>1.775</td>
<td>2.140</td>
</tr>
</tbody>
</table>

Source: R Processing Result

Table 8: Variables that Affect Malaria Morbidity in each Regencies

<table>
<thead>
<tr>
<th>No</th>
<th>Regency</th>
<th>Significant Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jayawijaya, Kepulauan Yapen, Jayapura, Biak Numfor, Paniai, Puncak Jaya, Mimika, Boven Digoel, Mappi, Asmat, Yahukimo, Pegunungan Bintang, Tolikara, Sarmi, Keerom, Waropen, Mamberamo Raya, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Intan Jaya, Deiyai, Jayapura City</td>
<td>X₂ and X₃</td>
</tr>
<tr>
<td>2</td>
<td>Merauke, Nabire</td>
<td>X₁</td>
</tr>
<tr>
<td>3</td>
<td>Supiori</td>
<td>X₂</td>
</tr>
</tbody>
</table>

Based on Table 7, the VIF score for all variables in all observation locations is less than 10, this shows that there is no multicollinearity between explanatory variables in all observation locations. Therefore modeling using GWR can be applied to this study.

**GWR Parameter Estimation:** Furthermore, to determine the factors that influence the model of malaria morbidity in each regency, partial testing was carried out as follows:

- \( H₀ : β_i(u, v) = 0 \) (There is no explanatory variable influence of on response variable)
- \( H₁ : β_i(u, v) ≠ 0 \) (There is explanatory variable influence on response variable);

\( k = 1,2, \ldots, p \)

With score \( α = 0.05 \), the score of t-table was obtained that is \( t_{0.025,24} = 2.064 \). After that by comparing the computed t score with ± table t score, the factors that influence the model of malaria morbidity in each location will be obtained. Based on the results of partial parameter estimation processing using R 3.4.4 software, the conclusions are as follows:

- From the four explanatory variables included in the GWR model, there are no locations where all of the explanatory variables are significant to malaria morbidity.
- Most of the locations (26 regencies from 29 regencies) have two explanatory variables which are significant in estimating malaria morbidity in Papua Province, namely the percentage of clean water use (\( X₃ \)) and the number of medic per 10,000 populations (\( X₄ \)).

Whereas for the other 3 locations, there is only one variable that is significant in estimating malaria morbidity in Papua Province, namely the variable of medic number per 10,000 population (\( X₁ \)) in Merauke and Nabire Regencies and the percentage of clean water use (\( X₃ \)) in Supiori regency.

Fig. 3: Distributions of Variables that Affect Malaria Morbidity in Papua

In general, the variables that influence the model of malaria morbidity in Papua Province are the percentage of clean water use and the number of medic per 10,000 populations, while the percentage of healthy residences and the average height of the area does not affect...
spatially on malaria morbidity in Papua Province. The distribution of variables that influence malaria morbidity is mapped as follows:

**CONCLUSIONS**

Based on the previous discussion, several conclusions can be drawn as follow:

- The global regression model formed from four independent variables shows it simultaneous influence toward malaria morbidity, but there are only two variables that are partially significant, namely the percentage of clean water use and the number of medic per 10,000 populations.
- Analysis of the GWR model with the fixed exponential kernel weighting method to get optimal bandwidth provides different results with global regression, so that the GWR model is suitable to be used for modeling malaria morbidity in Papua Province.
- In general, the variables that influence the model of malaria morbidity in Papua Province are the percentage of clean water use and the number of medic per 10,000 populations. These variables affect 26 regencies / cities in Papua Province, while the remaining 3 regencies, each have only one influential variable.

**ACKNOWLEDGEMENTS**

We would like to thanks to the Central Bureau of Statistics Indonesia and Master Program of Applied Statistics Universitas Padjadjaran for their supports.

**REFERENCES**