The Vulnerability Modeling of Dengue Hemorrhagic Fever Disease in Surabaya Based on Spatial Logistic Regression Approach

Nur Chamidah, Toha Saifuddin and Marisa Rifada

Department of Mathematics, Faculty of Science and Technology
Airlangga University, Indonesia

Abstract

In this paper, we discuss the vulnerability modeling of Dengue Hemorrhagic Fever (DHF) disease in Surabaya, Indonesia based on spatial logistic regression model approach. The model does not only include the factors that effect on the vulnerability of DHF disease but also accommodates the geographical locations simultaneously. This study uses data of the incidence of DHF disease in Surabaya that consist of response variable is specified i.e. endemic and non-endemic DHF region. While the predictor variables are factors of climate, population, and environment. The minimum of Akaike Information Criteia (AIC) value is used to determine the best model. Several weighting functions is used in this model, and we got the best model obtained by using the weighting of Fixed Gaussian function. Based on the best model, obtained prediction accuracy level of vulnerability DHF disease in Surabaya is 81.93%.

Keywords: Dengue fever, Spatial Logistic Regression, Fixed Gaussian weighting

1. Introduction

Dengue Hemorrhagic Fever (DHF) develops into a serious health problem, especially in Indonesia. Data collected last 5 years by the Department of Health show that the number of dengue cases has increased very sharply. The number of districts
affected by dengue has also increased sharply. In the 1970s, less than 10 cities or districts are infected by dengue fever, and in 2003 has become more than 200 cities or districts. The Surabaya Health Department stated that Surabaya endemic dengue and dengue fever is already spread in 31 districts in Surabaya. Ariesman [3] stated that at this time Surabaya is still an endemic DHF region. Up to 2 months of this year, dengue fever has attacked 171 people, three of them died. DHF problem in Surabaya has reached high level of vulnerability DHF disease.

Several prediction models have studied the incidence of DHF in Indonesia, including Sasmito, et al [10] and Hidayati [8] used multiple regression method, Ana, et al [1] used ARIMA models, and Ana and Widodo [2] used the ARIMAX model to model the incidence of DHF. Several models have been studied above only used weather information or climate factors but have not accommodated any aspects of the location (geospatial) simultaneously. Therefore, it is required aspects of geospatial modeling to accommodate the diversity of regional characteristics. The goal of this study is to model the vulnerability of DHF disease in Surabaya, Indonesia based on spatial logistic regression model approach with predictor variables, i.e., climate factors, population factors, and environmental factors that are assumed to be locally influence. Based on the obtained models, we can predict the vulnerability of DHF disease (endemic and non-endemic) in districts of Surabaya.

2. Preliminary notes

2.1 Geographically Weighted Regression

Geographically Weighted Regression (GWR) method is a technique used in spatial regression model that takes the framework of a simple regression model by weighted regression (Fotheringham, et al.,[6]). GWR model is expressed as follows:

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

where

\( y_i \) : observations on the \( i \)-th location \( (i = 1, 2, ..., n) \);

\( (u_i,v_i) \) : longitude latitude coordinates of the \( i \)-th point on a geographical location;

\( \beta_k(u_i,v_i) \) : the parameter model \( \beta_k(u,v) \) at the \( i \)-th point; and

\( \epsilon_i \) : error are assumed identical, independent, and Normal distribution with mean zero and variance constant \( \sigma^2 \).

Thus, each parameter value is estimated for each point of the geographical location. So, every point geographic location has different values of regression parameters. If the values of regression parameters are constant in each geographic region, the GWR models are called as global models. It means that each geographical region has the same model. This is a special case of the GWR.
2.2 Parameter Estimation of GWR Model

In the GWR model, we assume that the observational data that are close to the \(i\)-th point has more considerable influence on the estimation of \(\beta_i(u_i, v_i)\) than data that are far away from the \(i\)-th point. Local parameter \(\beta_i(u_i, v_i)\) is estimated by using weighted least squared (Leung,[8]). In GWR, an observation is weighted by the value associated with \(i\)-th point. The weights \(w_{ij}\), for \(j = 1, 2, \ldots, n\), at each location \((u_i, v_i)\) are obtained as a continuous function of the distance between the \(i\)-th point and other data points. Suppose the following matrix is a matrix of local parameters

\[
B = \begin{bmatrix}
\beta_0(u_i, v_i) & \beta_1(u_i, v_i) & \ldots & \beta_p(u_i, v_i) \\
\vdots & \vdots & \ddots & \vdots \\
\beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \ldots & \beta_p(u_n, v_n)
\end{bmatrix}
\]

(2)

Estimate of each row of equation (2) is the following equation:

\[
\hat{\beta}(i) = (X^T W(i) X)^{-1} X^T W(i) y
\]

(3)

where 
\(X\) = matrix of data from the predictor variables 
\(y\) = vector of response variables 
\(W(i) = diag [w_{i1}, w_{i2}, \ldots, w_{in}]\) = weighting matrix

In addition, to produce local parameter estimation for each geographic location, GWR produces localized versions for the entire standard regressions in all geographic locations such as the size of the goodness of fit. So, the estimation of the parameters in GWR can be obtained for all geographical locations.

2.3 Weighting GWR Model

The weighting in GWR model is very important, because it represents the weighted value of the location of observational data with each other. Therefore, we need a precisely weighting method (Chasco, et al.[5]). In the GWR weighting scheme, we have several different methods, for example, Kernel Gaussian and Bisquare weighting function (Bocci, et al., [4]). The Kernel Gaussian and Bisquare weighting function to calculate \(M\)-th point nearest, i.e.,

\[
w_j = \begin{cases} 
\frac{1}{\sqrt{2\pi}} \exp \left[ \frac{1}{2} \left( \frac{d_{ij}}{b} \right)^2 \right], & \text{if } j \text{ is one of } M - \text{th point to the nearest } i - \text{th point} \\
0, & \text{elsewhere}
\end{cases}
\]

and

\[
w_j = \begin{cases} 
\left[ 1 - \left( \frac{d_{ij}}{b} \right)^2 \right]^2, & \text{if } j \text{ is one of } M - \text{th point to the nearest } i - \text{th point} \\
0, & \text{elsewhere}
\end{cases}
\]

respectively.
In the weighting functions, \( b \) represents the distance to the \( M \)-th closest point and \( d_{ij} \) is to be the Euclidean distance:

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

For determining optimal value of \( b \), we can use least square approach based on cross-validation (CV) criterion as follows:

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

where \( i \neq i^* \) and \( \hat{y}_i(b) \) is the estimated value for \( y_i \) by leaving out the \( i \)-th observation.

### 2.4 Spatial Logistic Regression Model

Spatial logistic regression analysis is a method to estimate the regression parameters by calculating the spatial factor and an alternative approach to combine the GWR and the Generalized Linear Model (GLM) based on a logit link function to form Geographically Weighted Logistic Regression (GWLR) model. The GWLR models can be written as follows:

\[
g(x_j) = \ln \left[ \frac{\pi(x_j)}{1 - \pi(x_j)} \right]
\]

where

\[
\pi(x_j) = \frac{\exp \left( \sum_{k=0}^{p} \beta_k (u_i, v_i) x_{jk} \right)}{1 + \exp \left( \sum_{k=0}^{p} \beta_k (u_i, v_i) x_{jk} \right)}, \quad j = 1, 2, ..., n.
\]

We estimate these parameters in (5) by using maximum likelihood estimation (MLE).

There are several methods that can be used as a reference in the selection of the best model, one of them is the Akaike Information Criterion (AIC) which is defined as follows:

\[
AIC_c = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n \left\{ \frac{n + \text{tr}(S)}{n - 2 - \text{tr}(S)} \right\}
\]

where

- \( \hat{\sigma} \): The value of the standard deviation of the error estimator
- \( S \): Matrix projection where \( \hat{y} = Sy \).

Selection of the best model is done by specifying the model based on the smallest AIC value (Nakaya et al.,[9]).

### 3. Methods

#### 3.1 Parameter Estimation of Spatial Logistic Regression Model

Parameter estimation of spatial logistic regression models are calculated by using MLE method. The initial step of the method is to establish the following likelihood function for each location \( i (i=1,2,\ldots,n) \):
Vulnerability modeling of dengue hemorrhagic fever disease

\[
L(\beta(u, v_i)) = \prod_{j=1}^{n} P(Y = y_j) = \prod_{j=1}^{n} \pi(x_j)^{y_j}(1 - \pi(x_j))^{1-y_j} = \left\{ \prod_{j=1}^{n} \left[ 1 + \exp \left( \sum_{k=0}^{p} \beta_k(u, v_i) x_{jk} \right) \right]^{-1} \right\} \exp \left[ \sum_{k=0}^{p} \left( \sum_{j=1}^{n} y_{jk} \beta_k(u, v_i) \right) \right].
\]

Next, logarithm of likelihood function is

\[
\ln L(\beta(u, v_i)) = \sum_{k=0}^{p} \left( \sum_{j=1}^{n} y_{jk} \beta_k(u, v_i) \right) - \sum_{j=1}^{n} \ln \left[ 1 + \exp \left( \sum_{k=0}^{p} \beta_k(u, v_i) x_{jk} \right) \right].
\]

Geographic factors are to be the weighting factor of GWLR models. This factor has different values for each location that shows the local nature of the GWLR models. Therefore, we give weight \( w(u, v_i) \) to logarithm of likelihood function in (7). So, we have:

\[
\ln L'(\beta(u, v_i)) = \sum_{j=1}^{n} w_j(u, v_i) y_{j} x_{jk} - \sum_{j=1}^{n} x_{jk} w_j(u, v_i) \ln \left[ 1 + \exp \left( \sum_{k=0}^{p} \beta_k(u, v_i) x_{jk} \right) \right].
\]

Estimation parameter of \( \beta(u, v_i) \) is obtained by derivating equation (8) with respect to \( \beta_k(u, v_i) \) and the result is to be equal to zero,

\[
\frac{\partial \ln L'(\beta(u, v_i))}{\partial \beta_k(u, v_i)} = 0
\]

\[
\Leftrightarrow \sum_{j=1}^{n} w_j(u, v_i) y_{j} x_{jk} - \sum_{j=1}^{n} x_{jk} w_j(u, v_i) \frac{\exp \left( \sum_{k=0}^{p} \beta_k(u, v_i) x_{jk} \right)}{1 + \exp \left( \sum_{k=0}^{p} \beta_k(u, v_i) x_{jk} \right)} = 0
\]

\[
\Leftrightarrow \sum_{j=1}^{n} w_j(u, v_i) y_{j} x_{jk} - \sum_{j=1}^{n} x_{jk} \pi(x_j) w_j(u, v_i) = 0.
\]

Because of function in (9) is implicit equation, we use numerical iterative procedure, i.e., Newton Raphson method Iteratively Reweighted Least Squares (IRLS) to solve (9).

In general, the equation for Newton Raphson iteration is given by:

\[
\beta^{(i+1)}(u, v_i) = \beta^{(i)}(u, v_i) - H^{(i)}^{-1}(\beta^{(i)}(u, v_i)) \ g^{(i)}(\beta^{(i)}(u, v_i))
\]

where
\[ g^{(i)}(\beta^{(i)}(u, v)) = \begin{bmatrix} \frac{\partial \ln L'(\beta(u, v))}{\partial \beta_0(u, v)} \\ \frac{\partial \ln L'(\beta(u, v))}{\partial \beta_1(u, v)} \\ \vdots \\ \frac{\partial \ln L'(\beta(u, v))}{\partial \beta_p(u, v)} \end{bmatrix}, \quad H^{(i)}(\beta^{(i)}(u, v)) = \begin{bmatrix} h_{00} & h_{01} & \cdots & h_{0p} \\ h_{10} & h_{11} & \cdots & h_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p0} & h_{p1} & \cdots & h_{pp} \end{bmatrix}, \] and

\[ H^{(i)}(\beta^{(i)}(u, v)) \] is the Hessian matrix which elements

\[ h_{kk} = \frac{\partial^2 \ln L'(\beta(u, v))}{\partial \beta_k(u, v) \partial \beta_k(u, v)}. \]

Thus, we have:

\[ \frac{\partial^2 \ln L'(\beta(u, v))}{\partial \beta_k(u, v) \partial \beta_k(u, v)} = \sum_{j=1}^{n} x_{jk} x_{jk} w_{j}(u, v) \pi(x_j)(1-\pi(x_j)) \]

For each iteration \( i^{th} \), we apply

\[ g_k^{(i)} = \frac{\partial \ln L'(\beta(u, v))}{\partial \beta_k(u, v)} = \sum_{j=1}^{n} w_{j}(u, v) y_j x_{jk} - \sum_{j=1}^{n} x_{jk} \pi(x_j)^{(i)} w_{j}(u, v) \]

\[ h_{kk}^{(i)} = \frac{\partial^2 \ln L'(\beta(u, v))}{\partial \beta_k(u, v) \partial \beta_k(u, v)} = -\sum_{j=1}^{n} x_{jk} x_{jk} w_{j}(u, v) \pi(x_j)^{(i)} (1-\pi(x_j)^{(i)}) \]

where \( \pi(x_j)^{(i)} = \frac{\exp\left(\sum_{k=0}^{p} \beta_k^{(i)}(u, v) x_{jk}\right)}{1 + \exp\left(\sum_{k=0}^{p} \beta_k^{(i)}(u, v) x_{jk}\right)} \)

By repeating the procedure for each iteration of the \( i^{th} \) point regression, the local parameter estimator will be obtained. The iteration will stop when circumstances converge, that is when \( \| \beta^{(i+1)}(u, v) - \beta^{(i)}(u, v) \| \leq \epsilon \), where \( \epsilon \) is a very small positive number.

### 3.2 Description of Data and Steps of Estimating Parameters

The data used in this study includes data the incidence of DHF disease in 83 districts in Surabaya city that consist of 78 endemic and 5 non-endemic district. The predictor variables in this model are climate (rainfall, temperature, and humidity), population and environmental factors. Description and sources of data are showed in Table 1.
Table 1. Description and Sources of Data

<table>
<thead>
<tr>
<th>Description of Data</th>
<th>Sources of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>The incidence of DHF disease in the last 3 years in each district (endemic and non-endemic)</td>
<td>Health Department of Surabaya</td>
</tr>
<tr>
<td>Climate factors (rainfall, temperature, and humidity)</td>
<td>Climate Department of Surabaya.</td>
</tr>
<tr>
<td>Population factors (population density, the population aged &lt;15 years, and the population of at least high school educated)</td>
<td>Central Bureau of Statistics of Surabaya.</td>
</tr>
<tr>
<td>Environmental factors (Percentage of no larvae of household, healthy life behavior and the number of larvae observer)</td>
<td>Health center in each district in Surabaya</td>
</tr>
</tbody>
</table>

The response variable in this study is the level of vulnerability to DHF disease in every district in Surabaya consists of endemic (category 1) and non-endemic (category 0). The description of variables in this study are given in Table 2.

**Table 2** The summary of variables.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Response variable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variables</strong></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>The vulnerability of DHF in each district</td>
</tr>
<tr>
<td><strong>Climate factors</strong></td>
<td></td>
</tr>
<tr>
<td>X&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Rainfall</td>
</tr>
<tr>
<td>X&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Temperature</td>
</tr>
<tr>
<td>X&lt;sub&gt;3&lt;/sub&gt;</td>
<td>Humidity</td>
</tr>
<tr>
<td><strong>Population Factors</strong></td>
<td></td>
</tr>
<tr>
<td>X&lt;sub&gt;4&lt;/sub&gt;</td>
<td>Population density</td>
</tr>
<tr>
<td>X&lt;sub&gt;5&lt;/sub&gt;</td>
<td>The number of people aged least than 15 years old</td>
</tr>
<tr>
<td>X&lt;sub&gt;6&lt;/sub&gt;</td>
<td>The number of people educated at least high school</td>
</tr>
<tr>
<td><strong>Environmental Factors</strong></td>
<td></td>
</tr>
<tr>
<td>X&lt;sub&gt;7&lt;/sub&gt;</td>
<td>Percentage of no larvae of household</td>
</tr>
<tr>
<td>X&lt;sub&gt;8&lt;/sub&gt;</td>
<td>Healthy life behavior</td>
</tr>
<tr>
<td>X&lt;sub&gt;9&lt;/sub&gt;</td>
<td>The number of larvae observer</td>
</tr>
</tbody>
</table>

The steps for estimating parameters of GWLR model are as follows:
1. Determine latitude \((u_i)\) and longitude \((v_i)\) in every district in Surabaya.
2. Calculate the euclidian distance between the observation by geographic location. Euclidian distance in \((\ldots)\) between the location at coordinates \((u_i, v_i)\) and the location at coordinates \((u_j, v_j)\). This calculation is run for the entire observation locations.
3. Determine the optimal bandwidth \((b)\) based on CV method in (4).
4. Calculate the weighting matrix by using kernel functions, i.e., *Fixed Gaussian, Fixed Bisquare, Adaptive Bisquare* and *Adaptive Gaussian*. The weighting matrix calculations are run for $i = 1, 2, ..., 83$.

5. Estimate the parameters model.

6. Test significant of the parameters model.

7. Determine the best model by using the weighting matrix that has the smallest value of AIC in (6) models.

### 4. Results and Discussion

In this study, we used the correlation coefficient (Pearson Correlation) criterion to determine whether there are multicollinearity among the predictor variables. The correlation coefficients between the predictor variables are presented in Table 3.

<table>
<thead>
<tr>
<th>Table 3 Coefficients of Correlation between predictor variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_3$</td>
</tr>
<tr>
<td>$X_4$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$X_5$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$X_6$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$X_7$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$X_8$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$X_9$</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows that all $p$-values of correlation coefficients among the predictor variables of climate factors ($X_1$, $X_2$, and $X_3$) is 0.000. So, the predictor variables of climate factors are not modeled together in one model. In addition, there are significant correlation coefficients between the variable $X_5$ and $X_4$, $X_6$. Then, $X_5$ is not included in the model. Thus the predictor variables modeled in this study are $X_1$, $X_4$, $X_6$, $X_7$, $X_8$, and $X_9$.

Further, to build predictive model of DHF disease in Surabaya based on spatial logistic regression model approach, we use predictor variables $X_1$, $X_4$, $X_6$, $X_7$, $X_8$, and $X_9$ by using a weighted kernel function. The goodness of fit criterion used in this study is the AIC value. The best model is the model that has the smallest AIC value. The results of bandwidth value ($b$) and the optimal AIC values obtained are showed in Table 5.
Table 5. Optimal bandwidth value and AIC of spatial logistic regression models based on several kernels weighting function

<table>
<thead>
<tr>
<th>Weighting Function</th>
<th>bandwidth (b)</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Gaussian</td>
<td>0.074</td>
<td>31.460</td>
</tr>
<tr>
<td>Fixed Bisquare</td>
<td>0.170</td>
<td>31.535</td>
</tr>
<tr>
<td>Adaptive Gaussian</td>
<td>45.000</td>
<td>31.660</td>
</tr>
<tr>
<td>Adaptive Bisquare</td>
<td>83.000</td>
<td>32.86</td>
</tr>
</tbody>
</table>

Table 5 shows that the smallest AIC value is 31.460. It means that the best model in this study uses Fixed Gaussian weighting function.

Based on hypothesis testing with the significance level ($\alpha$), i.e., 10%, we obtain that spatial logistic regression model by using Fixed Gaussian weighting function that rainfall variable ($X_1$) is significant in the 30 districts, the population of at least high school educated variable ($X_6$) is significant in 62 districts, whereas for population density variable ($X_4$), percentage of no larvae of household variable ($X_7$), healthy life behavior variable ($X_8$) and the number of larvae observer variable ($X_9$) are not significant in the entire districts. Then, we get the accuracy of prediction of the level of vulnerability of DHF disease at 83 districts in Surabaya is 81.93%. From the estimation of vulnerability of DHF disease model at 83 districts in Surabaya, we divide them into 3 levels areas, i.e., low (0-0.3), moderate (0.31-0.7) and high level (0.71-1) of vulnerability of DHF disease. The thematic map of vulnerability of DHF disease at 83 districts in Surabaya is showed in Figure 1 as follows:

Figure 1 The thematic map of vulnerability of DHF disease at 83 districts in Surabaya.

Figure 1 shows that almost all districts in Surabaya (95%) have high vulnerability of DHF disease (endemic areas).
5. Conclusions

Based on spatial logistic regression approach, we obtain the best model for the vulnerability of DHF disease in Surabaya by using weighted Gaussian fixed which has the smallest AIC value. Because of there are predictor variables that are not significant in all districts, so that these variables influence globally. To increase the accuracy of prediction, in the further research, we can use the semiparametric spatial logistic regression model that has predictor variables locally and globally influence.

Acknowledgement. This Research was funded by the Directorate General of Higher Education of Indonesia through Featured Research University Grant 2013 with contract number: 91/AUPT-Baru/2013.

References


Received: December 15, 2013