Modelling the Number of Cases of Dengue Hemorragic Fever with Mixed Geographically Negative Binomial Regression in West Java Province

Zulfanita Dien R1, Asep Saefuddin2, Anik Djuraidah3

1,2,3Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

ABSTRACT
Dengue hemorrhagic fever (DHF) is an infectious disease caused by the dengue virus of the genus Flavivirus, which is transmitted by the bite of the Aedes Aegypti. Different regional demographics cause the number of DHF cases to differ in each region following by environmental conditions in the area. The model applied is GWNBR (Geographically Weigthed Negative Binomial Regression) due to count data outcome affected by geographical effect. In some instances not all in the GWNBR model have spatial effects, sometimes the estimate parameter are constant, so the GWNBR model can be developed using a mixed model to become MGWNBR. Determination of global and local parameters using the confidence interval. This study aims to analyze the factors that influence the number of dengue cases in West Java Province in 2015 using the MGWNBR approach. Based on the comparison of AIC values, the MGWNBR model has a smaller AIC value compared to the negative binomial regression model. The variables that significant globally are population density (X1) and health worker (X2)The variables that significant locally are number of health facilities (X3) PHBS (X4) and healthy homes (X5)

Keywords: Dengue hemorrhagic fever, GWNBR, MGWNBR

I. INTRODUCTION
Dengue hemorrhagic fever (DHF) is an infectious disease caused by the dengue virus of the genus Flavivirus, family of Flaviviridae which is transmitted by Aedes Aegypti (Kemenkes 2010). DHF is one of the main health problems in Indonesia. West Java Province is the province that has the highest number of dengue cases in Indonesia, which is 21237 cases with 14 cases of death (Ministry of Health 2015). The number of these cases increased by 14.6 percent from the previous year which was as much as 18116 (Kemenkes 2014). Dengue cases in West Province are increasing every year, prevention efforts need to be done by knowing the factors that influence the development of the disease.

The number of DHF cases is one example of a count event so that the right modeling is used to determine the factors that influence the development of DHF cases namely Poisson regression. Poisson regression is one of the statistical methods that is used to find out the relationship between the variables in the form of data with one or more explanatory variables. However, sometimes in the application of Poisson regression overdispersion phenomena are found. Overdispersion occurs when the variance value is greater than the average value. If the Poisson model with overdispersion is still used in the data it will
result in the occurrence of bias parameters (Evadianti and Purhadi 2014), so that one alternative modeling solution that can be used is the negative binomial distribution. The binomial negative model can overcome the problem of overdispersion in Poisson regression because it has dispersion parameters that can explain the variability in the data (Hilbe 2011).

the spread of DHF tends to cluster (Ruliansyah et al. 2017). According to Sukendra and Kusuma (2016) infectious diseases are not influenced by administrative boundaries so it can be said that the spread of dengue disease transmitted through Aedes Aegypti can spread to areas affected by DHF. Different regional demographics cause the number of DHF cases to differ in each region following the environmental conditions in the area. The right analysis is used to determine the factors that influence the development of dengue disease by taking into account the geographical location and have a response variable in the form of count data, namely negative geographical binomial weighted regression analysis (GWNBR).

GWNBR is a weighted statistical method with variable variables in the form of count data used to estimate parameters that have spatial diversity so that each observation location has different estimator values. In some cases not all estimators in the GWNBR model have spatial effects, sometimes the explanatory variables are constant. Nakaya et al. (2005) developed a mixed model so that there are two parameters, namely local and global. So that the GWNBR model can be developed using a mixed model into mixed GWNBR. The parameter estimation in the GWNBR mixture combines parametric methods to predict global parameters and nonparametric estimates to predict local parameters. There are two methods that can be used to determine global and local parameters, namely the interval of confidence (Pongoh 2015) and the linear model of coregionalization (Mar’ah et al. 2017). The purpose of this study was to analyze the factors that influence the increase in the number of dengue cases in West Java Province in 2015 with the mixed GWNBR approach.

II. METHODS AND MATERIAL

The data used for this study are secondary data obtained from the West Java Provincial Statistics Agency and the West Java Provincial Health Office in 2015. The unit of analysis of this study is as many as 27 districts / cities in West Java Province.

Table 1: Response and Explanatory Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Number of DHF Cases</td>
</tr>
<tr>
<td>X₁</td>
<td>Population density</td>
</tr>
<tr>
<td>X₂</td>
<td>Number of Health Workers</td>
</tr>
<tr>
<td>X₃</td>
<td>Number of Health Facilities</td>
</tr>
<tr>
<td>X₄</td>
<td>Percentage of Healthy and Clean Houses</td>
</tr>
<tr>
<td>X₅</td>
<td>Percentage of Healthy Houses</td>
</tr>
</tbody>
</table>

Analysis and modeling in this study were assisted by statistical programs, R (3.4.1). The steps used to analyze data on the number of dengue cases in infants in West Java Province in 2015 using the mixed GWNBR approach are as follows:

1. Describe the characteristics of the number of dengue cases in West Java Province in 2015

2. Estimating a negative binomial regression model
Binomial negative regression is one of the statistical methods used to overcome the overdispersion in the Poisson distribution. The parameter estimation is a negative binomial regression model with the following formula:

\[
\ln(\mu_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}
\]  

(1)

3. Multicollinearity Test
Multicollinearity testing can be done by looking at the VIF value (variance inflation factor) to see
whether or not there is multicollinearity. The VIF formula is as follows:

\[ VIF = \frac{1}{1-R_k^2} \quad (2) \]

with \( R_k^2 \) is the coefficient of determination of the explanatory variable for \( k = 1, 2, ..., n \). If the VIF value is > 10, the assumption of multicollinearity is not fulfilled.

4. Spatial dependencies Test

Testing spatial dependencies using Moran I. The hypothesis used to test spatial dependencies is as follows:

\[ H_0: I = 0 \]
\[ H_1: I \neq 0 \]

The Moran I test statistics are as follows:

\[ Z_{hit} = \frac{I - E(I)}{\sqrt{\text{var}(I)}} \quad (3) \]

The calculation of Moran I is based on a normal distribution, so for decision making that is if the value of Moran I is more than \( z_{\alpha/2} \) so it will reject \( H_0 \).

5. Spatial Heterogeneity Test

Testing of spatial heterogeneity can be done using the Breusch-Pagan test (Anselin 1988). The hypothesis used for testing spatial heterogeneity is as follows:

\[ H_0: \sigma_1 = \sigma_2 = \cdots = \sigma_j = 0 \]
\[ H_1: \text{at least one } \sigma_j \neq 0 \]

with statistic test as follows:

\[ BP = \left( \frac{1}{p} \right) f'Z(Z'Z)^{-1}Z'f \sim X^2(p) \]

with \( f_i = (\frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2} - 1) \)

Decision making in the Pagan Breusch test is if the value of \( BP > X^2(\alpha, p) \) so it will reject \( H_0 \).

6. Estimating the mixed GWNBR model

a. Determine the optimum smoothing parameters and spatial weight

One method that can be used to select the optimum smoothing parameter is cross validation.

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

with \( n \) adalah number of observation, \( y_i \) is the location response variable \( i \), \( \hat{y}_{x_i}(h) \) is the estimated value of observation of the location of \( i \) whose value is obtained without involving the observation of the location of the \( i \) itself.

Fotheringham et al. (2002) state that there are two spatial weighting, namely a fixed spatial kernel and an adaptive spatial kernel. The spatial weighting function aims to estimate the parameter values of each observation location.

i. Gaussian Kernel Function

\[ w_{ij} = \exp\left[ -\frac{d_{ij}^2}{b} \right] \]

with \( d_{ij} \) is the distance between locations, distance used is euclidean distance,

\[ d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \] and \( b \) is optimum smoothing parameter.

ii. Function of Kernel Bisquare

\[ w_{ij} = \begin{cases} 
1 & \text{if } d_{ij} < \frac{b}{a} \\
\left( 1 - \frac{d_{ij}}{b} \right)^2 & \text{if } d_{ij} \geq b
\end{cases} \]

b. Estimate the GWNBR model

Estimating parameters of the GWNBR model uses the following formula

\[ \beta^{(i+1)}(u_i, v_i) = (X'W(u_i, v_i)A(u_i, v_i)^{(0)}X)^{-1}X'W(u_i, v_i)A(u_i, v_i)^{(0)}z(u_i, v_i)^{(0)} \]

After getting the GWNBR model then look for the confidence interval for each parameter using the following formula:

\[ \beta_j \pm t_{\alpha/2} S_j \sqrt{\hat{\epsilon}_{jj}} \]

Criteria is that if the confidence interval has a percentage above 65 percent then the variable can be used as a global parameter and if the percentage is less than 65 percent then the variable can be used as a local parameter.
c. Predict the mixed GWNBR model parameters

Algorithm for estimating the mixed GWNBR model as follows:

i. Initiation γ (which comes from a negative binomial regression model)

ii. Using GWNBR for errors adjusted as estimating local parameters and calculating smoother $S$

iii. Use the formula as follows:

$$\gamma^{(l+1)} = (X_{par}A^{(l)}X_{par})^{-1}X_{par}A^{(l)}(I - S^{(l)})z^{(l)}$$

to get global parameters.

iv. Iteration is carried out until the convergence is reached.

v. Choose the best model with the AIC criteria

The selection of the best model uses the Akaike Information Criteria with the following formula:

$$AIC = -2\log_e\left(L(\hat{\theta} | data)\right) + 2K$$

with $\log_e\left(L(\hat{\theta} | data)\right)$ is the maximum possible maximum estimation value, $\theta$ is an unknown parameter, $K$ is the number of parameters estimated, and $n$ is the number of observations.

7 Interpret the results of the best models

III. RESULTS AND DISCUSSION

West Java Province is the province that has the highest number of dengue cases in Indonesia. Districts that have the highest number of dengue cases are Bandung Regency with a total of 3640 cases, while Districts which have the least number are Pangandaran Regency with 29 cases. The map of the number of dengue cases in West Java Province is shown in Figure 1.

Based on the estimation results, the negative binomial regression model is obtained by estimating the parameters presented in Table 2.

**Table 2**: Estimator parameters for negative binomial regression models

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>5.063</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$-2.05 \times 10^{-6}$</td>
<td>0.953</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.18</td>
<td>$1.83 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.031</td>
<td>2.41e-9</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.011</td>
<td>0.31</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.018</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Deviance = 28.277  
Df = 21  
AIC = 397.5

Based on the estimation results for negative binomial regression in Table 2 above, simultaneous and partial testing can be carried out to model the number of dengue cases in West Java Province. Simultaneous testing of the parameter estimators modeled the number of dengue cases in West Java Province using 5% level, so that the values of $\chi^2(4, 0.05) = 9.48$. Value of $\chi^2$ is less than $D(\hat{\beta})$ which is 28.277, so
that there is enough proof to state that there is at least
one explanatory variable that significant 5% level.
Partial testing of parameter estimators modeled the
number of DHF cases in West Java Province with 5%
level obtained values $z_{(0.05/2)}$ is 1.96. The value of $z$
Table is less than $z_{hitung}$ in each explanatory variable,
so that there is proof evidence to state that the
explanatory variable significant at 5% level. The
explanatory variables that have a significant effect are
health workers, the number of health facilities, and
healthy homes.

Multicolinearity testing is done by calculating the VIF
value (variance inflation factor) in each explanatory
variable. If the VIF value is > greater than 10, it can be
assumed that there is a correlation between the
explanatory variables. In Table 3 it can be seen that
the VIF value in the explanatory variable is less than
10 so it can be concluded that between explanatory
variables are not correlated for each variables.

Table 3: VIF values for explanatory variables

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>VIF Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>2.93</td>
</tr>
<tr>
<td>X2</td>
<td>2.98</td>
</tr>
<tr>
<td>X3</td>
<td>1.33</td>
</tr>
<tr>
<td>X4</td>
<td>1.46</td>
</tr>
<tr>
<td>X5</td>
<td>1.49</td>
</tr>
</tbody>
</table>

Spatial dependency is indicated by the existence of
interrelationships between regions. Testing of spatial
dependencies is done by calculating the value of the
Moran Index. Testing of spatial heterogeneity (spatial
diversity) is done by calculating the value of Breusch-
Pagan. The results of testing the spatial assumptions
are presented in Table 4.

Table 4: Testing of spatial assumptions

<table>
<thead>
<tr>
<th>Dependency Test</th>
<th>Heterogeneity Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moran Index</td>
<td>-0.63</td>
</tr>
<tr>
<td>$z_{statistic}$</td>
<td>-6.67</td>
</tr>
<tr>
<td>$z_{0.05/2}$</td>
<td>1.96</td>
</tr>
</tbody>
</table>

| Moran Index | -0.63 |
| $z_{statistic}$ | -6.67 |
| $z_{0.05/2}$ | 1.96 |

GWNBR produces different regression parameter
values at each observation location. The first step
taken in GWNBR modeling is to determine the kernel
weighting function used to produce the best model
with the model selection criteria using the AIC value.

Table 5: Selection of the best kernel weighting
functions

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWNBR Fix Bisquare</td>
<td>309.95</td>
</tr>
<tr>
<td>GWNBR Fix Gauss</td>
<td>350.70</td>
</tr>
<tr>
<td>GWNBR Adaptive Bisquare</td>
<td>255.40</td>
</tr>
<tr>
<td>GWNBR Adaptive Gauss</td>
<td>339.82</td>
</tr>
</tbody>
</table>

Based on Table 5, it can be seen that GWNBR
Adaptive Bisquare model has the smallest AIC value.
The results of the model suitability test using the F
test produce a calculated F value of 16.28 with a level
of 5% resulting in a value $F_{(0.05;21;21)}$ amounting to
2.08 resulting in the conclusion that there is enough
proof to state that there is a difference between the
negative binomial regression model and the GWNBR
model.

The GWNBR model has explanatory variables that
affect each district / city. The map of explanatory
variables for each Regency / City is presented in
Figure 2. The closeness between regions between
districts / cities affects the explanatory variables that
influence the region.
The selection of global and local parameters can be done using the confidence interval. The selection criteria using a confidence interval, if the percentage of the confidence interval is more than 65%, can be grouped into global parameters. Variable $X_1, X_2$ each has a percentage of more than 65%, 85.18% and 74.07%, so that these variables can be grouped into global variables. Variables $X_3, X_4, X_5$ each has a percentage of 29.62%, 62.96% and 62.96% so that it can be grouped into local variables. The division of global and local variables is used to estimate the parameters in the Mixed Geographical Binomial Negative Regression model (GWNBR).

Mixed GWNBR is a model that combines global and local parameters in one model. The mixed GWNBR model has different parameter estimation values in each field. Comparison of the AIC values of the negative binomial regression model and the mixed GWNBR model based on the AIC values are presented in Table 6.

Table 6: Comparison of the AIC values of the negative binomial regression model and the mixed GWNBR model

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative binomial regression</td>
<td>397.50</td>
</tr>
<tr>
<td>Mixed GWBNR</td>
<td>286.57</td>
</tr>
</tbody>
</table>

The mixed GWNBR model has an AIC value smaller than the negative binomial regression model so that the spatial model is better used. In addition, the mixed GWNBR model has more variables that influence each Regency / City than the GWNBR model. The map of explanatory variables that influence the mixed GWNBR model is presented in Figure 3.

One of the factors that influence the incidence of DHF (Kusuma and Sukendra 2016; Muliansyah and Baskoro 2016) is population density. Uncontrolled aspects of mobility and urbanization are factors that increase population, especially in urban areas and industrial estates. This condition has resulted in the creation of a conducive environment for the breeding of dengue vectors so that it will cause an increase in the number of dengue cases (Muliansyah and Baskoro 2016). This is because urban areas and industrial estates have a settlement layout that coincides with each other so that it will facilitate the transmission process (Ashlilihah et al. 2015). Another aspect that affects the incidence of DHF is clean and healthy living behavior and healthy homes (Candra 2010; Ashlilihah et al. 2016). According to Respati et al. (2017) said that basic sanitation can be used to determine risk factors for dengue cases.

Based on the variables that influence the GWNBR and mixed GWBNR models, it can be explained that the variables in the mixed GWNBR model provide more extensive information than the GWNBR model. The influential variables in the mixed GWNBR model not only pay attention to the variables in each region but also consider variables that influence globally.
IV. CONCLUSION

Based on the results of the study it can be seen that the mixed GWNBR model provides a good model of the number of dengue cases in West Java Province in 2015 compared to the negative binomial regression model. The variables that influence the mixed GWNBR model are population density ($X_1$) and health workers ($X_2$) influential globally. Variable number of health facilities ($X_3$), PHBS ($X_4$), and healthy houses ($X_5$) influential local.

V. REFERENCES
