Graphical interface of geographically weighted negative binomial regression (GWNBR) model using R-Shiny

H Yasin, I Suryani, P Kartikasari

Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia

Corresponding author: hasbiyasin@live.undip.ac.id

Abstract. In regression analysis, the spatial aspects need to be considered because each region has different characteristics as well as the Poisson regression model. In this method, some assumptions must be fulfilled, namely that the variance and the mean of the response variable are equal. However, counted data often have a greater variance than the mean, or what is generally called the over-dispersion phenomenon. If over-dispersion occurs, the Poisson regression is not suitable for modeling data, and that will produce biases in the parameter estimates. One method used to overcome over-dispersion in Poisson regression is negative binomial regression. The negative binomial regression model is more flexible than the Poisson regression model because it assumes that the mean and variance are not necessarily equal. Therefore, if the spatial aspect is considered, the Geographically Weighted Negative Binomial Regression (GWNBR) method is used. This study aims to compile a computational application to model spatial data using GWNBR Model using R-Shiny Web Application. And, the GWNBR model will be applied to modeling the number of dengue cases in Central Java province. The GWNBR model with Adaptive Boxcar weight is the best model because it has the smallest AIC. Using this model, two groups of districts/cities are obtained based on significant variables.

1. Introduction

Dengue Hemorrhagic Fever (DHF) is an acute disease caused by the dengue virus and transmitted through the bites of female Aedes Aegypti and Aedes Albopictus mosquitoes [1]. In Indonesia, DHF is endemic both in urban and rural areas. In urban areas the main vector is the Aedes Aegypti mosquito, while in rural areas the Aedes Albopictus mosquito is the main vector. However, it often happens that the two mosquito species occur in an area together, for example in semi-urban areas [2]. The breeding place that Aedes Aegypti often chooses is a densely populated area with inadequate sanitation, especially in standing water in the house, such as pots, flower vases, bathtubs or other water storage places such as jars, drums, or plastic buckets [3]. According to Notoatmodjo [4], environmental factors play an important role in disease transmission, especially the home environment that does not meet the requirements. The home environment is one of the factors that has a major influence on the health status of its occupants. According to the Central Java Provincial Health Office (2018) [5], the Case Fatality Rate (CFR) of DHF in Central Java Province in 2018 was 1.05 percent, a decrease compared to 2017, namely 1.24 percent. However, this figure is still higher than the national target (<1%). This shows that the handling of cases of DHF is not yet optimal. To overcome this problem, an analysis of the distribution pattern can be carried out and identify the factors that are thought to influence it.

The number of DHF sufferers in Central Java in 2018 is a discrete or count data, so the Poisson regression method can be used to analyze the factors that affect the number of DHF cases. In Poisson
regression analysis, some assumptions must be fulfilled, namely that the variance of the response variable is equal to the mean. However, often the count data has a variance greater than the mean, or so-called overdispersion phenomenon [6]. If overdispersion occurs, Poisson regression is not suitable for modeling data, and the model that will be formed produces biased parameter estimates [7]. One of the methods used to overcome overdispersion in Poisson regression is Negative Binomial regression. The negative binomial model is more flexible than the Poisson model because it assumes that the mean and variance are not necessarily the same. By paying attention to the spatial aspect (area), the Geographically Weighted Negative Binomial Regression (GWNBR) method is used [8, 9].

Spatial aspects need to be considered because each region has different geographical conditions. This condition causes a difference in the number of DHF cases from one region to another in terms of environmental conditions of the household. The DHF cases can move from one region to another. Its spread is increasingly widespread with increasing mobility and population density. Besides, the disease can be transmitted very easily and is not limited to administrative areas. So it can be said that DHF is a cross-border infectious disease. This is also following Tobler's first law in Anselin[10], which states that "Everything has a relationship with another, but something close together will have a relationship that is more than something far apart".

The problem that arises in GWNBR modeling is how to determine the optimal weighting matrix. Therefore, it is necessary to have a standard procedure for selecting the optimal kernel function to obtain a model with high accuracy. The main objective of this article is to develop a graphical user interface based GWNBR model computing software using R-Shiny Web Applications. The best model is determined by selecting the model that has the smallest AIC [11].

2. Materials and method

2.1. Poisson regression

Poisson regression is a nonlinear regression model with the response variable (y) following the Poisson distribution to overcome the data count. The probability function of the Poisson distribution can be stated as follows [12].

\[ f(y, \mu) = \frac{e^{-\mu} \mu^y}{y!}; y = 0, 1, 2, ..., n \]

where \( \mu \) is the mean of the response variable with a Poisson distribution where the mean and variant values of \( y \) have a value of more than 0. And the Poisson regression model equation can be written as follows.

\[ \hat{\mu}_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}) \]

where \( \hat{\mu}_i \) is the average number of events that occur in a certain time interval. The method used to estimate Poisson regression parameters is the Maximum Likelihood Estimation (MLE) method. In Poisson regression, the parameter estimated is \( \beta_k \). To get the estimated value, the steps taken are to form the likelihood function of the Poisson function and continued with the Newton Raphson iteration. Poisson regression parameter testing is to determine the effect of a parameter on the model with a certain level of significance. The Poisson regression model feasibility test is carried out using the Maximum Likelihood Ratio Test (MLRT) [13].

Overdispersion is a condition where the variance is greater than the mean in the Poisson distribution. Overdispersion can cause a conclusion to be obtained to be invalid because the standard error value becomes underestimated. This is because the resulting regression coefficient parameter is inefficient even though the regression coefficient remains consistent [14]. Overdispersion inspected by using the Pearson's Chi-square statistical test as follows:
\[ \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2} \sim \chi^2_{(n-p)}; \text{ where } p = k + 1 \]

Data is overdispersed if the Pearson’s Chi-square statistical test divided by the degrees of freedom will generate a value greater than 1.

2.2. Negative binomial regression

In negative binomial regression, the response variable is assumed to have a negative binomial distribution resulting from the Poisson-gamma mixture distribution. The Negative binomial probability mass function is as follows [15].

\[ f(y, \mu, \theta) = \frac{\Gamma(y + 1/\theta)}{\Gamma(1/\theta) \gamma^y} \left( \frac{1}{1 + \theta \mu} \right)^{\gamma/\theta} \left( \frac{\theta \mu}{1 + \theta \mu} \right)^y; y = 0, 1, 2, ..., n \]

The negative binomial regression model is expressed in the form of a linear combination between the parameter (\( \mu \)) and the regression parameter to be estimated, namely:

\[ \mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{pi}) \]

The parameter estimation of the negative binomial regression model uses the maximum likelihood method with the Newton Raphson procedure [16]. Simultaneous significance testing for the parameter of the Negative Binomial regression model uses the deviance test [17].

2.3. Geographically weighted negative binomial regression (GWNBR) model

The GWNBR model will generate local parameters, with each location having different parameters. The GWNBR model can be formulated as follows [18].

\[ y_i \sim NB \left( \exp \left( \sum_p \beta_p (u_i, v_i) x_{ip} \right), \theta(u_i, v_i) \right); i = 1, 2, ..., n \]

where \( y_i \) is the response observation value number \( i \); \( x_{ip} \) is the predictor variable in location \( (u_i, v_i) \); \( \beta_p (u_i, v_i) \) is the regression coefficient in location \( (u_i, v_i) \); \( \theta(u_i, v_i) \) is the dispersion parameters for location \( (u_i, v_i) \).

The negative binomial distribution function for each location can be written in the form of the following equation.

\[ f(y_i | \beta(u_i, v_i), \theta(u_i, v_i)) = \frac{\Gamma(\gamma + 1/\theta)}{\Gamma(1/\theta) \Gamma(\gamma + 1)} \left( \frac{1}{1 + \theta \mu_i} \right)^{\gamma/\theta} \left( \frac{\theta \mu_i}{1 + \theta \mu_i} \right)^y_i \]

where \( y_i = 1, 2, ..., \mu_i = \exp(\mathbf{x}_i^T \mathbf{\beta}_{(u_i,v_i)}) \) and \( \theta_i = \theta(u_i,v_i) \).

The parameter \( \mathbf{\beta}_{(u_i,v_i)} \) are estimated using Newton Raphson Iteration by the likelihood function as follows (Ricardo, 2013).

\[ L(\mathbf{\beta}(u_i, v_i), \theta_i | y_i, x_i) = \prod_{i=1}^{n} \left( \prod_{r=0}^{y_i-1} \left( r + \frac{1}{\theta_i} \right) \left( \frac{1}{\gamma_i} \right) \left( \frac{1}{1 + \theta_i \mu_i} \right)^{1/\theta_i} \left( \frac{1}{1 + \theta_i \mu_i} \right)^{1/y_i} \right) \]

2.4. Graphical interface of GWNBR

One of the computations of the GWNBR model can be done with R software using GWmodel and lmtest packages [19, 20]. However, not everyone is able to make the syntax of an algorithm that is quite complex. Therefore, in this study, an application was developed to analyze spatial data using the
GWNBR model based on the R-Shiny Web App [21, 22]. This application is arranged by following the algorithm in Table 2. The user interfaces display of this application can be seen in Figure 1.

**Table 1. Parameter estimation of GWNBR using Newton Raphson iteration methods**

**Algorithm**: Parameter Estimation of GWNBR using NRI Method

**Input:**
- Matrix of independent variable $X$
- Vector of dependent variable $y$
- Geographical Coordinate of observation $(u_i, v_i)$
- Type of kernel function (Gaussian, Exponential, Bisquare, Tricube, or Boxcar)
- Type of bandwidth (Fixed or Adaptive)

**Output:**
1. Initialization of parameter using coefficient of Negative Binomial Regression
   
   $\hat{\beta}_{(0)} = [\theta_0 \ \beta_{00} \ \cdots \ \beta_{p0}]$

2. Find the optimal bandwidth of GWR model using selected kernel function (Gaussian, Exponential, Bisquare, Tricube, or Boxcar) and type of bandwidth (Fixed or Adaptive).
3. Calculate the optimal Weighting Matrix based on the optimal bandwidth in step 2.
4. For each i-th location, estimate the parameter GWNBR using Newton Raphson Method

   \[
   g(\hat{\beta}_{(m)}) = \left( \frac{\partial \ln L(\beta)}{\partial \theta}, \frac{\partial \ln L(\beta)}{\partial \beta_0}, \frac{\partial \ln L(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \ln L(\beta)}{\partial \beta_p} \right)^T_{\beta=\hat{\beta}_{(m)}}
   \]

   \[
   H(\hat{\beta}_{(m)}) = \begin{bmatrix}
   \frac{\partial^2 \ln L(\beta)}{\partial \beta_0^2} & \frac{\partial^2 \ln L(\beta)}{\partial \beta_0 \partial \beta_1} & \ldots & \frac{\partial^2 \ln L(\beta)}{\partial \beta_0 \partial \beta_p} \\
   \frac{\partial^2 \ln L(\beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ln L(\beta)}{\partial \beta_1^2} & \ldots & \frac{\partial^2 \ln L(\beta)}{\partial \beta_1 \partial \beta_p} \\
   \vdots & \ddots & \ddots & \vdots \\
   \frac{\partial^2 \ln L(\beta)}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 \ln L(\beta)}{\partial \beta_p \partial \beta_1} & \ldots & \frac{\partial^2 \ln L(\beta)}{\partial \beta_p^2}
   \end{bmatrix}_{\beta=\hat{\beta}_{(m)}}
   \]

   \[
   \hat{\beta}_{(m+1)} = \hat{\beta}_{(m)} - H^{-1}(\hat{\beta}_{(m)})g(\hat{\beta}_{(m)})
   \]

   Until:
   - Convergence $\hat{\beta}_{(m+1)} \approx \hat{\beta}_{(m)}$

5. Calculate the deviance model
6. Model evaluation by calculating the AIC.

2.5. Study area and data used
The data used in this study is secondary data in 2018 obtained from the Central Java Provincial Health Office [5]. The observation units used in this study were 35 districts/cities in Central Java. Considering that the data used are secondary data, it is assumed that the measuring instrument (questionnaire) used has been validated, and the officers have filled incorrectly. The dependent variable in this study is the
number of dengue cases in Central Java in 2018 (y). The independent variable consists of 5 variables, there are Percentage of Healthy Homes (X_1), Percentage of Poor People (X_2), Percentage of Clean Water Quality (X_3), Percentage of Population with Proper Drinking Water (X_4), dan Ratio of Medical Personnel (X_5).

3. Main results

3.1. Assumption check

The assumptions that must be met are the absence of multicollinearity and the occurrence of overdispersion in the Poisson regression model. A multicollinearity check can be done by looking at the Pearson correlation value. Based on Figure 3, it can be seen that all predictor variables have a Pearson
correlation coefficient of less than 0.95, which means that there are no cases of multicollinearity. Based on the overdispersion test, it can be seen that the dispersion parameter or the deviation ratio value with degrees of freedom is 81.80995 and is greater than 1, so it can be concluded that there is an overdispersion in the model. This conclusion is reinforced by the model's z-score value = 2.4728 with p-value = 0.0067. So that at the 5% significance level ($Z_{(0.05/2)}=1.96$) resulted in a decision to reject the null hypothesis (see Figure 4). Thus, to overcome overdispersion cases, negative binomial regression can be used to model the number of DHF cases in Central Java in 2018.

3.2. Model comparison
The selection of the best model based on the AIC criteria [11]. Model Comparison of the Poisson regression, negative binomial regression, and GWNBR is seen in Table 2. Based on this table, GWNBR-
Adaptive Boxcar has the smallest AIC compared to other models. This model is better at modeling the number of DHF cases in each Districts/City in Central Java in 2018.

Furthermore, based on the best model, there are three results of test: (1) based on the results of the similarity test, the $F_{\text{test}}$ value is 2.8634. At the 5% significance level, it is found that $F_{(0.05,29,29)}$ is 1.8608, so it can be concluded that there is a significant difference between the negative binomial model and the GWNBR model, (2) based on the results of simultaneous testing, the GWNBR model deviance value is 52.6051. At the 5% significance level, it is obtained $\chi^2_{(5)}$ of 11.0705, which means that at least one parameter of the GWNBR model has a significant effect, so it is necessary to continue with partial testing, (3) based on the results of testing simultaneously, it is found that the parameters are significantly different for each Districts/Cities. Value of $|t_{\text{test}}|$ parameters for each Districts/Cities are compared with the $Z_{(0.05/2)}$. If the value of $|t_{\text{test}}| > 2.045$ then reject $H_0$, which means that the variable has an influence on the model. The grouping of significant variables is presented in Table 3.

### Table 2. Models comparison based on AIC

| Model                          | Optima Bandwidth | AIC       |
|--------------------------------|------------------|-----------|
| Poisson Regression             | -                | 2568.40   |
| Negative Binomial Regression   | -                | 395.23    |
| GWNBR-Fixed Gaussian           | 1.719603         | 358.3286  |
| GWNBR-Fixed Exponential        | 3.067126         | 358.3289  |
| GWNBR-Fixed Bisquare           | 2.505283         | 358.3286  |
| GWNBR-Fixed Tricube            | 2.430764         | 358.3284  |
| GWNBR-Fixed Boxcar             | 2.122497         | 358.3281  |
| GWNBR-Adaptive Gaussian        | 32               | 358.3284  |
| GWNBR-Adaptive Exponential     | 32               | 358.3301  |
| GWNBR-Adaptive Bisquare        | 34               | 358.3294  |
| GWNBR-Adaptive Tricube         | 34               | 358.3292  |
| **GWNBR-Adaptive Boxcar**      | **33**           | **358.3264** |

### Table 3. District/City grouping based on significant variables in the GWNBR model

| Significance Variables | District/City                                                                 | Total |
|-----------------------|-------------------------------------------------------------------------------|-------|
| $X_1, X_2, X_3, X_4, X_5$ | Purbalingga and Brebes, Cilacap, Banyumas, Banjarnegara, Kebumen, Purworejo, Wonosobo, Grobogan, Jepara, Demak, Semarang, Kendal, Batang, Pekalongan, Pemalang, Tegal, Magelang City, Semarang City, Pekalongan City, Tegal City, Magelang, Boyolali, Klaten, Sukoharjo, Wonogiri, Karanganyar, Sragen, Blora, Rembang, Pati, Kudus, Temanggung, Surakarta City, Salatiga City | 2     |

### 4. Conclusion

Empirical results show that the GWNBR model with Adaptive Boxcar weight is the best model because it has the smallest AIC. Using this model, two groups of districts/cities are obtained based on significant variables. The variables affecting the number of DHF cases in all districts/cities in Central Java Province are the percentage of healthy houses, the percentage of clean water quality, and the ratio of medical personnel. Further study is needed to optimize the weights matrix using the Neural Network approaches.
Acknowledgments
This research was fully supported and funded by “Riset Madya” Research Grant 2020 under the Research Assignment Letter No: 2024/UN7.5.8/PP/2020.

References
[1] Achmadi U F 2010 *Buletin Jendela Epidemiologi* 2 1 15-20
[2] Soedarto 2012 *Demam Berdarah Dengue* (Surabaya: Sagung Seto)
[3] Tobing T, Aunudin, Rahman L A 2011 *Dengue Fever (DF) Case Modelling in East Java with Poisson and Negative Binomial Models* (Bogor: Institut Pertanian Bogor)
[4] Notoatmodjo S 2003 *Ilmu Kesehatan Masyarakat, Prinsip-Prinsip Dasar* (Jakarta: Rineka Cipta)
[5] Official Health of Central Java Province 2018 *Health Profile of Central Java Province 2018* (Semarang: Indonesia)
[6] Cameron A, Trivedi P 1998 *Regression Analysis of Count Data* (Cambridge: Cambridge University Press)
[7] Hardin J, Hilbe J 2007 *Generalized Linear Models and Extensions Second Edition* (Texas: Stata Press)
[8] Fotheringham A S, Brunsdon C, Charlton M 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (USA: Wiley)
[9] Hidayanti U, Latra I N, Purnadi 2015 *Jurnal Sains dan Seni* 4 2 120
[10] Anselin L 1988 *Spatial Econometrics: Methods and Models* (Dordrecht: Kluwer Academic Publishers)
[11] Nakaya T, Fotheringham A S, Brunsdon C, Charlton M 2005 *Stat. Med.* 24 7 2695-2717
[12] Myers R H 1990 *Classical and Modern Regression with Application* (Boston: PWS-KENT Publishing Company)
[13] McCullagh P, Nelder J A 1989 *Generalized Linear Models* (London: Chapman and Hall)
[14] Cameron A, Trivedi P 1990 *J. Econom.* 46 1 347-346
[15] Greene W 2008 *Functional Forms for the Negative Binomial Model for Count Data, Foundation, and Trends in Econometrics* (New York: New York University)
[16] Hilbe J 2011 *Negative Binomial Regression, Second Edition* (New York: Cambridge University Press)
[17] Hosmer D and Lemeshow S 1995 *Applied Logistic Regression* (New York: John Wiley and Sons Inc.)
[18] da Silva A R and Rodrigues T C V 2014 *Stat. Comput.* 24 5 769–783
[19] Gollini I, Lu B, Charlton M, Brunsdon C and Harris P 2015 *J. Stat. Soft.* 63 17 1-50
[20] Lu B, Harris P, Charlton M and Brunsdon C 2014 *Geo Spat. Inf. Sci.* 17 2 85-101
[21] Chang W 2020 Package ‘shiny’: *Web Application Framework for R*
[22] Fawcett L 2018 *J. Stat. Educ.* 26 1 2–16