International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET1207473

Geographically Weighted Negative Binomial Regression Modeling of Tuberculosis Cases with Distribution Evaluation

Choirun Nisa[1], Muhammad Nur Aidi[2], I Made Sumertajaya[3]
[1,2,3]Department of Statistics, IPB University, Bogor, West Java, Indonesia

Article Info
Volume 7 Issue 4
Page Number: 279-285
Publication Issue :
July-August-2020

Article History
Accepted : 20 Aug 2020
Published : 28 Aug 2020

ABSTRACT

Tuberculosis (TB) is contagious disease caused by the bacteria *mycobacterium tuberculosis* and is one of the top 10 causes of death in the world. Central Java is included as one of the three provinces with highest number of TB cases in Indonesia. End the TB epidemic by 2030 is the final goal of *Sustainability Development Goal* and Indonesia has set target for TB elimination by 2035. The number of TB cases is non-negative count data. The distribution pattern of the count data needs to be noticed in order to produce valid analysis. Based on calculation of the VMR (Variance Mean Ratio) value and suitability test, the data on the number of TB cases follows negative binomial distribution. The infection of TB disease tends to be clumped and is affected by geographical factors (environmental, social, and economic). This study aims to determine factors that affecting TB cases through Geographically Weighted Negative Binomial Regression (GWNBR) model approach which considering spatial aspects. Based on the ratio of AIC and BIC value, GWNBR model with an *adaptive gaussian* kernel weighting gives the best results. The affecting factors are the number of hospitals(X_2), the percentage of population with good water access(X_4), the population density(X_5), the percentage of household with a distance of drinking water sources and feces septic tank less than 10 meters(X_7).

Keywords: TB, count data, VMR, spatial aspects, GWNBR

I. INTRODUCTION

Tuberculosis (TB) is still become global health issue. Indonesia is among the third countries with highest TB case incidences in the world (WHO 2019). The number of TB cases in Indonesia in 2018 was 566.623, where 44% of the total reported cases were in the provinces of West Java, East Java and Central Java. Pulmonary TB patients with AFB (Acid-Fast Basil) positive present a greater possibility of infection risk than pulmonary TB patients with AFB negative. A total of 49.520 new cases of AFB positive in 2018 in Central Java (Ministry of Health 2018). Each of the observation data needs to be identified its distribution's pattern so that the analysis used is suitable and the results are valid. The Variance Mean Ratio (VMR)

method is one of techniques for determining the distribution pattern of observational data by calculating the ratio of the variance and the median. If the data follows a random or Poisson distribution, variance will be proportional with the median so that the VMR value is in the range of number 1; if the value of the VMR < 1 (close to 0), it indicates uniform distribution; if the value of VMR > 1, it indicates clumped distribution or negative binomial (Krebs 2013).

The Ministry of Health Regulation number 67 of 2016 on Tuberculosis Control sets targets for national TB control program, namely TB elimination by 2035 and Indonesia TB-free by 2050. One of the steps to overcome the TB cases problem is by modeling the factors which affect TB cases. There are several methods that can be used in modeling the count data such as Poisson regression and negative binomial regression. In Poisson regression analysis, there is an assumption needs to be fulfilled, where the variance of the response variables must be equal with the median (Hilbe 2011). This kind of condition is very rare to happen because usually the count data has greater variance than the median or it is called over-dispersion condition (Cameron and Trivedi 2013). The negative binomial regression is one of solutions to overcome over-dispersion on count data (Hilbe 2011).

Tuberculosis incidence is strongly affected by spatial aspects. GWNBR model is one of methods that effective enough to estimating counted data using negative binomial distribution that has diversity. The GWNBR model generates local parameters with each location has different parameters. This study aims to evaluate the suitability of data distribution pattern through the VMR value and identify the factors that affect TB cases in Central Java in 2018 through GWNBR modeling.

II. METHODS AND MATERIAL

The data used in this study is secondary data obtained from The Health Profile of Central Java Province and Central Java Province Statistics Agency in 2018 (BPS 2018). The analysis unit for this study includes 35 regencies/cities in Central Java Province.

Table 1: Response and Explanatory Variables

Variables	Explanation
Y	Number of positive TB AFB cases
X_1	Villages implementing community based total sanitation
X_2	Number of hospitals
X_3	Average length of schooling
X_4	Population with good water access
X_5	Population density
X_6	Impoverished population
X_7	Household with a distance of water drinking source and feces septic tank less than 10 meters

Analysis and modeling were carried out using *R Studio version 3.4.4* software program. The modeling of new positive TB AFB cases in Central Java Province in 2018. was carried out based on the following stages:

1. Describe the characteristics of the number of TB cases in Central Java Province in 2018.
2. Evaluate the distribution suitability of response variable with chi-square and the VMR value with the following calculation:
$$VMR = \frac{Var(Y)}{E(Y)}$$
3. 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:
$VIF = \frac{1}{1-R_k^2}$
with R_k^2 is the coefficient of determination of the explanatory variable for $k = 1,2, \dots, n$. If the

VIF value is more than 10, the assumption of multicollinearity is not fulfilled.

4. Poisson regression and negative binomial modeling to determine the best model to use.
5. Testing of spatial heterogeneity using *Breusch-Pagan* test (Anselin 1988). The hypothesis used for testing spatial heterogeneity is as follows:
   $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = 0$
   H_1: *at least one* $\sigma_i^2 \neq 0$
   with statistics test as follows:
   $$BP = \left(\frac{1}{2}\right) \boldsymbol{f}^T \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{f}$$
   with $f_i = \left(\frac{e_i^2}{\sigma^2} - 1\right)$ and decision making in the *Breusch-Pagan* test is if the value of $BP > \chi^2_{(\alpha,p)}$ so it will reject H_0.
6. One method can be used select the optimum smoothing parameter is CV (Cross Validation) criteria, with formula is as follows:
   $$CV(h) = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i}(h))^2$$
   with n is number of observations, y_i is the location response variable i, $\hat{y}_{\neq i}(h)$ is the estimated value of observation of the location of i whose value is obtained without involving the observations of the location of the i itself.
7. Fotheringham *et al.* (2002) state there are two spatial weighting, namely a fixed spatial kernel and an adaptive spatial kernel. The spatial weighting function aim to estimate the parameter values of each observation location. This study using 2 spatial weighting,
   i. Adaptive gaussian kernel
   $$w_{ij_{(u_i,v_i)}} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h_i}\right)^2\right)$$
   ii. Fixed gaussian kernel
   $$w_{ij_{(u_i,v_i)}} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right)$$
   with $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$ is the euclidean distance and h_i is optimum smoothing parameter.
8. Estimate the GWNBR model with *Newton Raphson* iteration uses the following formula:
   $$\widehat{\boldsymbol{\beta}}_{(m+1)} = \widehat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{H}_{(m)}^{-1}(\widehat{\boldsymbol{\beta}}_{(m)})\boldsymbol{g}_{(m)}(\widehat{\boldsymbol{\beta}}_{(m)})$$
9. The selection of the best model uses the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) with the following formula:
   $$AIC = 2k - 2log(L)$$
   $$BIC = -2log(L) + k\ln(n)$$
   with $log(L)$ is the maximum possible maximum estimation value, k is the number of parameters estimated, and n is the number of observations.
10. Interpret the best model and form a map of TB cases grouping based on the significantly affecting factors.

III.RESULTS AND DISCUSSION

A. Data Exploration

Central Java is one of the three provinces indicated with highest number of TB cases in Indonesia. The distribution of the number of new TB cases in Figure 1 shows with red area the total of TB cases that include in the high category is in Cilacap regencies, Banyumas regencies, Brebes regencies, Kudus regencies and Semarang City, respectively 3825, 3269, 3196, 2923 and 3400 cases. A total of 6 regencies/cities are include in medium category shown with green area and 24 regencies/cities are include in low category shown with yellow area. In addition, in Figure 1 indicates the TB cases was grouped by proximity to geographic area.

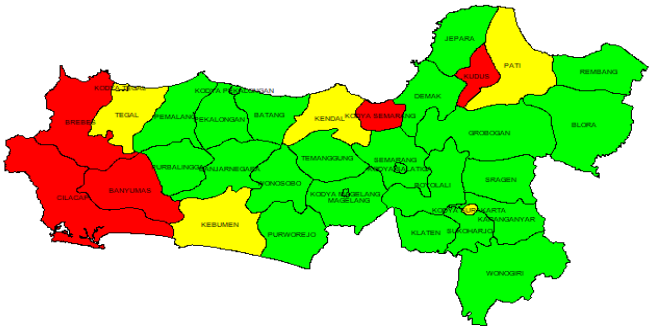

Figure 1: Number of TB cases in cities in Central Java 2018

Prior to modeling, distribution suitability evaluation is carried out through data exploration and distribution testing on response variables. The data on the number

of new TB cases has Var(Y) > E(Y) with a variance value of 859787.9 and a median value of 1414.9 which results in a VMR value of 607.685 (more than 1) and experiences over-dispersion. Based on the theory of VMR value, the data on the number of TB cases follows the negative binomial distribution. In addition, negative binomial distribution is a distribution that accommodates overdispersion violations.

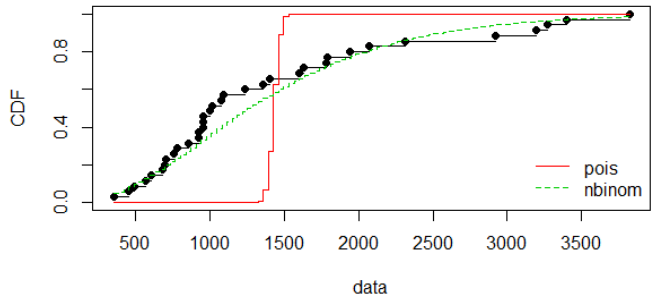

Figure 2 : Plot of empirical and theoretical distribution on applied data

Plot on Figure 2 shows the data on the number of TB cases is closer to negative binomial distribution than Poisson. Furthermore, chi-square test is conducted to find out the suitability of the data to a certain distribution opportunity. The result of chi-square test on the Poisson distribution with a value of $\chi^2(Inf) > \chi^2_{0.05,6}$ (15.5) showed that the number of TB cases did not spread as Poisson. Meanwhile, the chi-square test on negative binomial distribution with a value of $\chi^2(Inf) > \chi^2_{0.05,6}$ (15.5) showed the number of TB cases spread as negative binomial. The VMR score calculation and the distribution suitability test showed the same results that the data of TB cases follows binomial negative distribution.

Table 2: The VIF value of explanatory variables

Variables	VIF Value
X_1	1.198
X_2	1.373
X_3	4.046
X_4	2.787
X_5	3.483
X_6	1.856
X_7	2.684

Multicollinearity testing is done by calculating the VIF value in each explanatory variable. In Table 2 show 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 variable (no multicollinearity). So that 7 explanatory variables can be used for regression model formation.

B. Poisson and Negative Binomial Regression

Poisson regression modeling is conducted as a comparative evaluation of models. The AIC and BIC value of the Poisson regression model were respectively as 8122.99 and 8135.44. Negative binomial regression model is better than Poisson regression model due to the AIC and BIC values of 552.21 and 566.21. So, the selection model is negative binomial regression. Table 3 shows the results of parameter estimation of negative binomial regression model.

Table 3: Estimatated values of binomial regression model parameters

Parameter	Estimate	*p_value*
β_0	7.348	0.0000
β_1	0.01422	0.0684
β_2	0.07581	0.000001
β_3	-0.4584	0.000079
β_4	0.02783	0.0522
β_5	0.00007	0.1848
β_6	-0.02534	0.3473
β_7	0.02183	0.0337

The simultaneous significance test of negative binomial regression with likelihood ratio test shows that the test statistic is $D(\hat{\boldsymbol{\beta}})$ (27.52) > $x^2{}_{(0.05,7)}$ (14.06) so that H_0 is rejected which indicates that there is at least one variable affects the model. The result of the partial negative binomial regression significance test in table 7 with a significance level of 5% shows that there are 3 significant factors in the model, namely the number of hospitals (X_2), RLS (X_3), and the distance of drinking water source less than 10 meters (X_7). So that

the equation of negative binomial regression model can be written as the following:

$$\ln(\hat{\mu}) = 7.348 + 0.07581X_2 - 0.4584X_3 + 0.02183X_7$$

The spatial diversity test is done by calculating the value of Breusch-Pagan (BP), it is obtained the BP test statistic value of (18.145) > $\chi^2_{(0.05,7)}$ (14.06) and *p-value* of (0.0113) < α (0.05) which indicated there is diversity between regencies/cities (spatial heterogeneity).

C. GWNBR Model

GWNBR modeling required a partial weighting matrix. The partial weighting matrix contains kernel function that consist of interlocation distance and bandwidth. The first step is calculating euclidean distance between observatory locations. The next is determining optimum bandwidth using cross validation criteria. Table 4 shows that adaptive Gaussian weighting function is the best weight for modeling the number of new TB cases in Central Java using GWNBR approach since it has smaller CV value.

. **Table 4:** Comparison of adaptive gaussian and fixed gaussian CV value

Kernel Function	CV value
Adaptive Gaussian	24255997
Fixed Gaussian	24314049

The results of suitability test for the GWBNR model show the calculated F value of 6.315 with a level of 5% resulting a $F_{(0.05,27;27)}$ value of 1.89 indicating that there is enough evidence to state that the GWNBR model is suitable. The factors which affecting the number of new TB cases on each regency/city are quite diverse, so that they can be classified into 6 groups based on the significant factors.

The number of hospitals (X_2) has significant effect in 12 regencies/cities. Health facilities such as hospitals are involved in TB infection prevention and control program. The percentage of population with good water access (X_4) has significant effect in 10 regencies/cities. According to Girsang *et al.* (2011) showed that bad quality drinking water consumption has an effect of 23% on the incidences of TB.

The significant variable found in all regencies/cities is X_5 (population density). The Ministry of Health Regulation number 67 of 2016 states that a dense housing environment will facilitate TB infection. The percentage of households with a distance of drinking water source and feces septic tank less than 10 meters (X_7) has significant effect in 27 regencies/cities. The distance of drinking water source and feces septic tank is also a requirement of the availability of clean water. The disposal of human feces and wastes must meet the health requirements, that are must be able to prevent the waste from infiltrating and contaminating the surface of clean water sources. The water sources that meet the requirements are one of protection against disease infection (Ruswanto 2010).

A. Selection Model

Table 5 shows that a better model for data on the number of new TB cases in Central Java is the GWNBR model since it has smaller AIC and BIC values compared to the negative binomial regression model. GWNBR model gives better because it considers spatial aspects.

Table 5: Comparison of AIC and BIC value of negative binomial regression and GWNBR model

Model	AIC	BIC
Negative binomial regression	552.21	566.21
GWNBR	462.20	467.41

B. Mapping of TB cases by best model

The map grouping the regencies/cities in Central Java based on significant variables which are affecting new TB cases in Figure 2 shows that the pattern of the areas tend to clumped. Areas that are close to each other tend to have similar characteristics related to the factors that affect the incidence of TB.

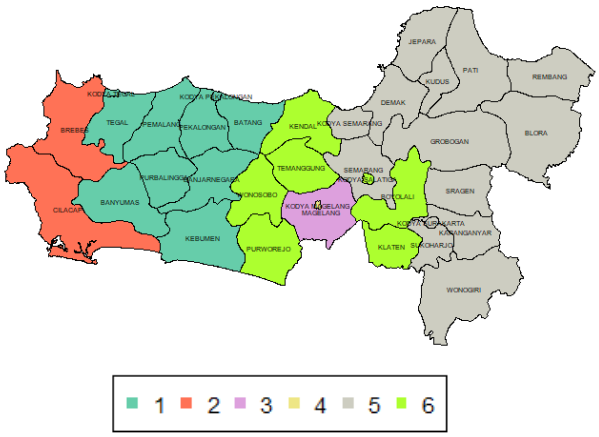

Figure 2: Classification of regencies/cities in Central Java based on factors which affect the number of TB cases

IV.CONCLUSION

The distribution suitability evaluation of count data through the calculation of the VMR value and chi-square test shows that the data on the number of TB cases follows negative binomial distribution. Based on distribution evaluation results, the GWNBR model gives the best results with smaller AIC and BIC values than the Poisson regression model and negative binomial regression. The GWNBR model with the adaptive Gaussian spatial kernel weighting classifies regencies/cities in Central Java into 6 groups based on factors which is significantly affecting the number of new TB cases. Overall, the affecting factors are the number of hospitals (X_2), the percentage of population with good water access (X_4), the population density (X_5), the percentage of household with the distance of drinking water source and feces septic tank less than 10 meters (X_7). The significant variable which is found in all regencies/cities is population density(X_5).

V. REFERENCES

[1]. [BPS] Badan Pusat Statistik Provinsi Jawa Tengah. 2019. Provinsi Jawa Tengah Dalam Angka. Pemerintah Provinsi Jawa Tengah

[2]. [Dinkes] Dinas Kesehatan Provinsi Jawa Tengah. 2018. Profil Kesehatan Provinsi Jawa Tengah Tahun 2018. Pemerintah Provinsi Jawa Tengah.

[3]. [Kemenkes] Kementrian Kesehatan Indonesia. 2016. Peraturan Menteri Kesehatan Republik Indonesia No. 67 Tahun 2016 Tentang Penanggulangan Tuberkulosis. Jakarta (ID): Kemenkes.

[4]. [WHO] World Health Organization. 2019. Global Tuberculosis Report 2019. France (FR): WHO.

[5]. Anselin L. 1988. Spatial Econometrics: Methods and Models. Netherlands (NL): Kluwer Academic Publisher.

[6]. Cameron AC, Trivedi PK. 2013. Regression Analysis of Count Data. Cambridge (UK): Cambridge University Press.

[7]. Fotheringham AS, Brunsdon C, Charlton M. 2002. Geographically Weighted Regression the Analysis of Spatially Varying Relationships. Chichester (UK): John Wiley and Sons.

[8]. Girsang M, Tobing K, Rafrizal. 2011. Faktor Penyebab Kejadian Tuberkulosis Serta Hubungannya Dengan Lingkungan Tempat Tinggal di Provinsi Jawa Tengah. Buletin Peneliti. Kesehatan. 39(1): 34 - 41.

[9]. Hilbe MJ. 2011. Negative Binomial Regression: 2nd. New York (USA): Cambridge University Press.

[10]. Krebs CJ. 2013. Ecological Methodology. Ed ke-4. New York (US): Harper Collins Publisher.

[11]. Ruswanto B. 2010. Analisis Spasial Sebaran Kasus Tuberkulosis Paru Ditinjau dari Faktor Lingkungan Dalam dan Luar Rumah di Kabupaten Pekalongan [Disertation]. Semarang (ID): Universitas Diponegoro.

Cite this article as :

Choirun Nisa, Muhammad Nur Aidi, I Made Sumertajaya, "Geographically Weighted Negative Binomial Regression Modeling of Tuberculosis Cases with Distribution Evaluation", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 4, pp. 279-285, July-August 2020. Available at doi : https://doi.org/10.32628/IJSRSET1207473
Journal URL : http://ijsrset.com/IJSRSET1207473