BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1761–1772
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1761-1772

POISSON REGRESSION MODELS TO ANALYZE FACTORS THAT INFLUENCE THE NUMBER OF TUBERCULOSIS CASES IN JAVA

Yekti Widyaningsih[1*], **Zalfa Alifah Budiawan**[2]

[1,2]Statistics Study Program, Faculty of Mathematics and Natural Sciences, University of Indonesia, Kampus Baru UI Depok, Jawa Barat, 16424, Indonesia

*Corresponding author e-mail: *yekti@sci.ui.ac.id*

ABSTRACT

Article History:

Received: 30th May 2023
Revised: 16th August 2023
Accepted: 27th August 2023

Keywords:

Generalized Poisson Regression;
Geographically weighted negative binomial regression;
Geographically weighted generalized poisson regression;
Negative binomial regression.

Tuberculosis is an infectious disease and one of the world's top 10 highest causes of mortality in Indonesia. Based on this fact, it is necessary to study what factors affect number of tuberculosis cases. The number of tuberculosis cases as dependent variable is a count data that generally analyzed using Poisson regression. However, equidispersion assumption must be met, so Generalized Poisson Regression and Negative Binomial Regression are applied if the assumption is not met. Spatial aspects can be considered so Geographically Weighted Generalized Poisson Regression and Geographically Weighted Negative Binomial Regression were also conducted. Four models were built to evaluate relationship between number of tuberculosis cases and factors affecting it in Java in 2020. The explanatory variables are population density, percentage of children receiving BCG immunization, percentage of poor people, percentage of eligible drinking water facilities, percentage of family cards with access to proper sanitation, percentage of public places meet health requirements, and percentage of food management places meet hygienic requirements. This study shows that the best model for modeling the data is GWNBR with 2 groups of significant explanatory variables. Seven explanatory variables are statistically significant in 88 districts and six explanatory variables statistically significant in 12 districts.

How to cite this article:

Z. A. Budiawan and Y. Widyaningsih., "POISSON REGRESSION MODELS TO ANALYZE FACTORS THAT INFLUENCE THE NUMBER OF TUBERCULOSIS CASES IN JAVA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1671-1772, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by infection with the bacteria Mycobacterium tuberculosis and several other Mycobacterium species known as Acid-Fast Bacteria [1]. According to WHO, Tuberculosis remains as the tenth largest cause of death in the world in 2016, even though the number of deaths related to TB decreased by 22% between 2000 and 2015. The number of Tuberculosis cases in Indonesia in 2020 was 543874 with most cases happening on Java Island. The first position is occupied by West Java with 123021 cases, the second position is East Java with 65448 cases, and the third position is Central Java with 54640 cases [2]. The high TB cases are caused by several factors which are often referred to as TB risk factors. Some of the risk factors for TB include factors related to the individual, demographic factors, socioeconomic factors, and sanitation factors.

This study will focus on the number of tuberculosis cases in Java. Beside as the most tuberculosis problems in Indonesia in 2020, Java Island is also the most populous island in Indonesia. The data in this research is secondary data derived from the Provincial Health Profile, the Public Health Office, and the Central Agency of Statistics (BPS). Data is only in 2020 and does not include Banten, DKI Jakarta, and DI Yogyakarta Provinces due to incompleteness data. Therefore, this study was conducted based on 100 Districts or cities on the Java.

The number of TB cases is a count data which generally uses the Poisson Regression method in the analysis. In Poisson regression, it is necessary to assume an equidispersion condition where the variance and mean values are the same, which is rarely happen because overdispersion conditions arise, i.e., conditions where the variance is greater than the mean. Thus, the use of Poisson Regression analysis is not suitable for modeling the data. There are several methods to analyze data with overdispersion problems in Poisson Regression, including Generalized Poisson Regression (GPR) and Negative Binomial Regression (NBR) [3]. By considering the spatial effects in the data, the Geographically Weighted Negative Binomial Regression (GWNBR) and Geographically Weighted Generalized Poisson (GWGPR) methods can be used as alternatives[4]–[6]. Modeling the data by considering spatial aspects can be applied due to differences in geographical conditions, resulting in changes in the number of tuberculosis cases between regions based on regional characteristics related to environmental conditions.

Several studies discussed various factors affect the number of tuberculosis cases had been carried out such as [4], [5], [6], [7]. Based on the research that has been done, this study uses the approach of Generalized Poisson Regression (GPR), Negative Binomial Regression (NBR), Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) as a comparison in determining the factors which affects the number of TB cases in Java in 2020. The novelty of this study is the application of the Generalized Poisson Regression (GPR), Negative Binomial Regression (NBR), Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) then comparing the four models to new data, namely data on the number of tuberculosis cases in Java in 2020. The best model that meets the modeling assumptions is used to analyze and interpret modeling results.

2. RESEARCH METHODS

2.1 Materials and Data

The data in this study consisting of 100 Districts or cities on Java Island, including West Java, Central Java, and East Java Provinces obtained from the 2020 Provincial Health Office Health Profile [8], [9], [10] and data published by the Central Agency of Statistics (BPS) for 2020 [11], [12], [13]. Several provinces on Java Island such as DKI Jakarta, Banten, and DI Yogyakarta Provinces were not included in this study due to unavailability of data.

The data is secondary data from the Health Profile of the West Java Province Public Health Office, Central Java Province, and East Java Province in 2020 as well as the publication of district/city poverty data and information in the relevant provinces in 2020. This research uses district/city data in each province. The dependent variable in this study is the number of TB cases (y), while the independent variables are population density (KP), percentage of toddlers given BCG immunization (BCG), percentage of poor people (PPM), percentage of drinking water facilities according to requirements (AIR), percentage Households with access to proper sanitation facilities (JS), the percentage of TTU according to health requirements (TTU), and the percentage of food processing places (TPM) according to hygienic requirements (TPM).

2.2 Multicollinearity

Multicollinearity is a condition where in the regression model there is a mutual correlation between two or more independent variables. Multicollinearity can be checked using the VIF (Variance Inflation Factor) value criteria. The criterion for multicollinearity is if the VIF value is greater than 10. The formula for the VIF value is as follows [14]:

$$VIF=\frac{1}{1\text{-}R_p^2}; p=1,2,\ldots,k \tag{1}$$

where p : index for the p-th independent variable and $R_p^2=1\text{-}\frac{SSE}{SST}=\frac{SSR}{SST}=\frac{\sum_{i=1}^{n}(\hat{y}_i\text{-}\bar{y})^2}{\sum_{i=1}^{n}(y_i\text{-}\bar{y})^2}$ is the coefficient of determination of the regression model with the p-th independent variable as dependent variable and the other independent variables as independent variables in the model.

2.3 Overdispersion

Overdispersion is a condition in a variable (with Poisson distribution) where the variance is greater than the mean. The formula for detecting overdispersion can be written as follows [15]:

$$\hat{\phi}=\frac{\sum_{i=1}^{n}(y_i\text{-}\hat{\mu}_i)^2/\hat{\mu}_i^2}{n\text{-}k} \tag{2}$$

with y_i : the value of independent variable of i-th observation, $\hat{\mu}_i$: poisson regression coefficient of i-th observation, n : number of observations, k : number of parameters.

If $\hat{\phi}>1$, it can be said that there is overdispersion in the Poisson model [16].

2.4 Spatial Heterogeneity

The Breusch-Pagan test is used to test the homogeneity of the variance in errors. Here are the hypotheses for the Breusch-Pagan test:

$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ (There is no spatial heterogeneity)

$\mathrm{H}_1 : \sigma_i^2 \neq \sigma^2,\ i = 1,2,\ldots,n$ (There is spatial heterogeneity).

Breusch-Pagan test statistics:

$$\mathrm{BP}=\frac{1}{2}f^TZ\left(Z^TZ\right)^{-1}Z^Tf \tag{3}$$

Decision rules:

H_0 is rejected if $BP>\chi^2_{(\alpha,\ p)}$ or $p\text{-}value<\alpha$ with p is the number of independent variables.

2.5 Generalized Poisson Regression

Generalized Poisson regression (GPR) is a suitable model for count data with overdispersion cases. The GPR model is a development of the Poisson Regression model which follows the Generalized Poisson distribution. The generalized Poisson distribution function **is** [17] :

$$\mathrm{f}(y_i;\mu_i,\theta)=\left(\frac{\mu_i}{1+\theta\mu_i}\right)^{y_i}\frac{(1+\theta y)^{y_i\text{-}1}}{y_i!}exp\left(\frac{\text{-}\mu_i(1+\theta y_i)}{1+\theta\mu_i}\right),\ y_i=0,1,2,\ldots,n \tag{4}$$

where y_i is the value of random variable Y, $\mu_i=\mu_i(x_i)=exp(x_i^T\beta)$, and θ is the dispersion parameter.

The Generalized Poisson Regression model is the same as the Poisson Regression model, i.e.

$$\mu=exp(\beta_0+\beta_1x_1+\ldots+\beta_px_p) \tag{5}$$

The method used to estimate the parameters of the Generalized Poisson Regression model is Maximun Likelihood (MLE). To obtain the estimated values of the parameters of the Generalized Poisson Regression model, further numerical approaches are needed. The numerical iteration used is Newton-Raphson iteration.

2.6 Negative Binomial Regression

Negative Binomial Regression (NBR) is an alternative model for dealing with the overdispersion problem based on the combined model of the Poisson and Gamma distributions which includes parameter ***θ*** [18]. The negative binomial distribution function is

$$\mathrm{f}(y_i;\mu_i,\theta) = \frac{\Gamma\left(y_i+\frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)\mathrm{y_i}!}\left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu_i}\right)^{y_i},y_i=0,1,2,\dots,n \tag{6}$$

where y_i is the value of random variable Y of i-th observation, $\mu_i = \mu_i(x_i) = exp(x_i^T\beta)$, and θ is the dispersion parameter.

The Negative Binomial Regression Model can be written as follows

$$\mu_i = exp(X_i^T\beta) \tag{7}$$

where $X_i = [1 \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ik}]^T$ is vector of independent variables for the i-th observation and ß = $[\beta_0 \quad \beta_1 \quad \beta_2 \quad \dots \quad \beta_k]^{\mathrm{T}}$ is a regression parameter vector.

The method used to estimate the parameters of the Negative Binomial Regression model is Maximun Likelihood (MLE). To obtain the estimated values of the parameters of the Negative Binomial Regression model, further numerical approaches are needed. The numerical iteration used is Newton-Raphson iteration.

2.7 Spatial Weighting Matrix

The spatial weight matrix $\mathrm{W(s_i,t_i)}$ is a matrix of size (n×n) with diagonal elements $\mathrm{w_j(s_i,t_i)}$ and $\mathrm{(s_i,t_i)}$ is the coordinate of locations. The general form of the spatial weighting matrix $\mathrm{W(s_i,t_i)}$ is

$$\mathbf{W}(\mathrm{s_i,t_i}) = \begin{bmatrix} w_1(s_i,t_i) & 0 & \cdots & 0 & 0 \\ 0 & w_2(s_i,t_i) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{n-1}(s_i,t_i) & 0 \\ 0 & 0 & \cdots & 0 & w_n(s_i,t_i) \end{bmatrix}$$

Spatial weighting values of $\boldsymbol{W(s_i,t_i)}$ is obtained based on neighborhood relations with values between 0 and 1 [19]. The weights for the jth-observation locations (s_i,t_i) at each location are a function of the Euclidean distance between locations i and j. The formation of the weight in this study using the adaptive bisquare kernel function with the following formula:

$$w_{ij} = \begin{cases} \left[1-\frac{d_{ij}^2}{b_i^2}\right]^2, & |d_{ij}| < b_i \\ 0, & \text{others} \end{cases} \tag{8}$$

where $d_{ij}=\sqrt{(s_i-s_j)^2+(t_i-t_j)^2}$ which is the Euclidean distance between locations (s_i,t_i) and b_i is the optimum bandwidth value at i-th location.

Optimum bandwidth is an optimum distance that allows the response of a region to have a strong influence on other regions. Bandwidth is determined by using cross-validation as follows

$$\mathrm{CV\ (b)} = \sum_{\mathrm{i=1}}^{\mathrm{n}}\left[y_i-\hat{y}_{\neq i}(b)\right]^2 \tag{9}$$

where $\hat{y}_{\neq i}(b)$: estimated value of y_i with observations at the i-th location removed from the estimation process and n is sample size.

2.8 Geographically Weighted Generalized Poisson Regression (GWGPR)

The GWGPR method is the development of the Geographically Weighted Regression (GWR) method for modeling spatial count data with overdispersion or underdispersion problems. The probability distribution function of the GWGPR model for each location is

$$\mathrm{f}\left(y_i\middle|\boldsymbol{x_{ip}}\beta_p(s_i,t_i),\theta(s_i,t_i)\right)=\left(\frac{\mu_i}{1+\theta\mu_i}\right)^{y_i}\frac{(1+\theta y)^{y_i\text{-}1}}{y_i!}exp\left(\frac{\text{-}\mu_i\left(1+\theta y_i\right)}{1+\theta\mu_i}\right),\ y_i=0,1,2,\ldots,n \tag{10}$$

The GWGPR models can be written as follows

$$\mu_i=exp\left(\beta_0(s_i,t_i)+\beta_1(s_i,t_i)x_{i1}+\ldots+\beta_p(s_i,t_i)x_{ip}\right),\theta(s_i,t_i) \tag{11}$$

The method used to estimate the parameters of the GWGPR model is Maximun Likelihood (MLE). To find the estimated values of the parameters of the GWGPR model, further numerical approaches are needed. The numerical iteration used is Newton-Raphson iteration.

2.9 Geographically Weighted Negative Binomial Regression (GWNBR)

The GWNBR method is the development of the GWR method for modeling spatial count data with overdispersion problems [20]. The probability distribution function of the GWNBR model for each location can be written as follows

$$\mathrm{f}\left(y_i\middle|x_{ip}\beta_p(s_i,t_i),\theta(s_i,t_i)\right)=\frac{\Gamma\left(y_i+\frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)y_i!}\left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu_i}\right)^{y_i},y_i=1,2,3,\ldots,n \tag{12}$$

The GWNBR models can be written as follows

$$\mu_i=exp\left(\beta_0(s_i,t_i)+\beta_1(s_i,t_i)x_{i1}+\ldots+\beta_p(s_i,t_i)x_{ip}\right),\theta(s_i,t_i) \tag{13}$$

The method used to estimate the parameters of the GWNBR model is Maximun Likelihood (MLE). To find the estimated values of the parameters of the GWNBR model, further numerical approaches are needed. The numerical iteration used is Newton-Raphson iteration.

2.10 Parameter Testing for GPR, NBR, GWGPR, GWNBR models

There are 2 types of parameters testing for the four models, namely simultaneous tests (Goodness-of-Fit) and partial tests.

1. Simultaneous test (Goodness-of-Fit)

This test uses the deviation value as a test statistic. In the GWGPR and GWNBR this test is often referred to as the Maximum Likelihood Ratio Test (MLRT).

The following is the hypothesis:

a) For GPR and NBR models

$H_0:\beta_1=\beta_2=\ldots=\beta_P=0$

H_1: there is at least one $\beta_p\neq 0,\ p=1,2\ldots,k$

b) For GWGPR and GWNBR models

$H_0:\beta_1(s_i,t_i)=\beta_2(s_i,t_i\ )=\ldots=\beta_p(s_i,t_i)=0$

H_1: there is at least one $\beta_p(s_i,t_i)\neq 0,\ p=1,2\ldots,k$

Test Statistics:

$$\mathrm{D}(\hat{\beta})=\text{-}2\,ln(\Lambda)=\text{-}2\,ln\left(\frac{\mathrm{L}(\hat{\omega})}{\mathrm{L}(\hat{\Omega})}\right) \tag{14}$$

where $L(\hat{\omega})$: Likelihood function that does not involve independent variables and $L(\hat{\Omega})$: Likelihood function involving independent variables.

Decision rules:

Reject H_0 if $D(\hat{\beta}) > \chi^2_{(a,\, p)}$ or $p\text{-}value < \alpha$ which can be interpreted that at least one independent variable in the model has a significant effect on the dependent variable.

2. Partial test

Partial test is used to find out which independent variables have a significant effect on the dependent variable.

The following is the hypothesis from the partial test:

a) For GPR and NBR models

$H_0 : \beta_p = 0$

$H_1 : \beta_p \neq 0,\ p = 1,2,\ldots,k$

b) For GWGPR and GWNBR models

$H_0 : \beta_p(s_i,t_i) = 0$

$H_1 : \beta_p(s_i,t_i) \neq 0,\ p = 1,2,\ldots,k$

Test Statistics:

a. For GPR and NBR models

$$Z_{hitung} = \frac{\hat{\beta}_p}{se\left(\hat{\beta}_p\right)} \tag{15}$$

where $\hat{\beta}_p$ is the parameter estimate β_p and se $\left(\hat{\beta}_p\right)$: standard error of $\hat{\beta}_p$

b. For GWGPR and GWNBR models

$$Z_{hitung} = \frac{\hat{\beta}_p(s_i,t_i)}{se\left(\hat{\beta}_p(s_i,t_i)\right)} \tag{16}$$

where $\left(\hat{\beta}_p(s_i,t_i)\right)$ is parameter estimate of $\beta_p(s_i,t_i)$ and se $\left(\hat{\beta}_p(s_i,t_i)\right)$ is standard error of $\left(\hat{\beta}_p(s_i,t_i)\right)$.

Decision rules:

Reject H_0 if $\left|Z_{hitung}\right| > Z_{\frac{\alpha}{2}}$ or $p\text{-}value < \alpha$ which means that the independent variables in the model affect the dependent variable.

2.11 Best Model Selection

To select the best model among the four models, the AIC value criteria is used. Akaike's Information Criterion (AIC) is a measure of the relative goodness of fit of a statistical model. Therefore, it can be concluded that the best model is the one with the smallest AIC. The formula of AIC is shown in **Equation (17)** as follows **[21]**:

$$\text{AIC} = -2\ln\left(\hat{\beta}\right) + 2k \tag{17}$$

where $ln\left(\hat{\beta}\right)$: the calculated likelihood of each model (in the case of this study are the Generalized Poisson Regression, Negative Binomial Regression, GWGPR, and GWNBR models) and k: the number of parameters in the model.

3. RESULTS AND DISCUSSION

3.1 Multicollinearity

Multicollinearity is examined to determine whether there are correlations between two independent variables. Multicollinearity was checked through VIF scores. **Table 1** shows the VIF scores of each independent variable.

Table 1. VIF score of each independent variable

Variable	VIF Score
X_1 (KP)	1,519220
X_2 (BCG)	1,115338
X_3 (PPM)	1,541758
X_4 (AIR)	1,290480
X_5 (JS)	1,285694
X_6 (TTU)	1,837326
X_7 (TPM)	1,460800

Based on the results in **Table 1**, all VIF scores are smaller than 10 for all independent variables. It shows that all independent variables meet the assumption of non-multicollinearity. In other words, there is no correlation between any two independent variables.

3.2 Overdispersion

Overdispersion of samples have been checked by calculating the Pearson chi-square value divided by the degree of freedom. The result is $\widehat{\phi} = 614.341 > 1$ which defined that there is overdispersion in the Poisson Regression model.

3.3 Spatial Heterogeneity

Spatial heterogeneity test carried out using the Breusch-pagan test. The following is a hypothesis from the Breusch-Pagan test, namely

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ (There is no spatial heterogeneity)

$H_1 : \sigma_i^2 \neq \sigma^2,\ i = 1,2,\ldots,n$ (There is spatial heterogeneity).

Based on the analysis, it was obtained that $BP = 22.573 > \chi^2_{(0.05;7)} = 14.0671$ concluded that the data contained spatial heterogeneity or variance between different observation locations are different.

3.4 Generalized Poisson Regression

Generalized Poisson Regression (GPR) used as an alternative when overdispersion conditions occur in the Poisson Regression model. It recognized that there is overdispersion in the Poisson model used to analyze the data using Generalized Poisson Regression. **Table 2** is the result of the estimations parameter of GPR.

Table 2. GPR Parameter Estimation

Variable	Parameter	Estimation	*p-value*
Intercept	$\widehat{\beta}_0$	5.210572	2×10^{-16}
KP	$\widehat{\beta}_1$	5.885873×10^{-5}	**0.001002**
BCG	$\widehat{\beta}_2$	-1.890808×10^{-3}	0.649958
PPM	$\widehat{\beta}_3$	9.979145×10^{-3}	0.557173
AIR	$\widehat{\beta}_4$	5.935316×10^{-3}	0.062009
JS	$\widehat{\beta}_5$	-1.471216×10^{-2}	**0.000286**

TTU	$\hat{\beta}_6$	-5.154265×10^{-3}	0.240764
TPM	$\hat{\beta}_7$	-2.058292×10^{-3}	0.551502
Deviance		1417954	
AIC		1646.237	

Note: the p-value in **bold** is a significant independent variable.

Simultaneous testing is carried out with the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_7 = 0$$

H_1: there is at least one $\beta_k \neq 0,\ k = 1,2\ldots,7$

Based on Table 2, it can be concluded that H_0 is rejected because $deviance = 1417954 > \chi^2_{(0.05;7)} = 14.0671$ which means that there is at least one independent variable affect number of TB cases.

Partial testing for parameters is carried out with the following hypotheses:

$$H_0 : \beta_p = 0$$
$$H_1 : \beta_p \neq 0,\ p = 1,2,\ldots,7$$

Based on Table 2, only the KP and JS variables have $p\text{-}value < \alpha = 0.05$. Therefore, only the KP (population density) and JS (access to proper sanitation facilities) variables affect the number of TB cases.

3.5 Negative Binomial Regression

Negative Binomial Regression (NBR) used as an alternative when overdispersion conditions occur in the Poisson Regression model. It has known that there is overdispersion in the Poisson model that the data is analyzed using Negative Binomial Regression. Table 3 is the result of estimations parameter of NBR..

Table 3. NBR Parameter Estimation

Variabel	**Parameter**	**Estimation**	***p-value***
Intercept	$\hat{\beta}_0$	1.01	$<2\times10^{-16}$
KP	$\hat{\beta}_1$	4.122×10^{-5}	0.074150
BCG	$\hat{\beta}_2$	-7.202×10^{-3}	0.141668
PPM	$\hat{\beta}_3$	-2.119×10^{-2}	0.284725
AIR	$\hat{\beta}_4$	6.77×10^{-3}	0.065442
JS	$\hat{\beta}_5$	-2.036×10^{-2}	$\mathbf{0.000108}$
TTU	$\hat{\beta}_6$	-9.76×10^{-3}	0.059438
TPM	$\hat{\beta}_7$	1.995×10^{-6}	0.99609
Deviance		2458651	
AIC		1636.4	

Note: the p-value in **bold** is a significant independent variable.

Simultaneous testing is carried out with the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_P = 0$$

H_1: there is at least one $\beta_p \neq 0,\ p = 1,2\ldots,7$

Based on Table 3, it can be concluded that H_0 is rejected because $deviance = 2458651 > \chi^2_{(0.05;7)} = 14.0671$ which means that there is at least one independent variable affect the number of TB cases.

Partial testing for parameters is carried out with the following hypotheses:

$$H_0 : \beta_p = 0$$
$$H_1 : \beta_p \neq 0,\ p = 1,2,\ldots,7$$

Based on Table 3, only the JS variable has $p\text{-}value < \alpha = 0.05$ Therefore, only JS (access to proper sanitation facilities) variables affect the number of TB cases

3.6 Geographically Weighted Generalized Poisson Regression (GWGPR)

GWGPR was a development of GPR, where the development lies in the spatial aspects that are considered in the estimation of model parameters. The spatial aspect is indicated by the spatial weighting matrix.

Hypothesis testing for model suitability is:

$H_0 : \beta_1(s_i,t_i) = \beta_2(s_i,t_i) = \ldots = \beta_p(s_i,t_i) = 0$

H_1: there is at least one $\beta_p(s_i,t_i) \neq 0,\ p = 1,2\ldots,7$

After processing the data using R studio, the deviance value is 37171.71. If the significance level is $a = 0.05$, then $\chi^2_{(0.05;7)} = 14.0671$. Therefore, $deviance = 37171.71 > \chi^2_{(0.05;7)} = 14.0671$. It was concluded that H_0 being rejected means that there is at least 1 independent variable affects the number of TB cases.

Testing the hypothesis for the significance of the parameters partially is:

$H_0 : \beta_p(s_i,t_i) = 0$

$H_1 : \beta_p(s_i,t_i) \neq 0,\ p = 1,2,\ldots,7$

The partial test results produce different parameters for each district/city. Two groups were obtained based on the partial test results for the GWGPR model. The grouping results showed six variables have a significant effect on all cities or Districts in Java Island, namely population density (KP), percentage of poor people (PPM), percentage of drinking water facilities that meet the requirements (AIR), percentage of households with access to proper sanitation facilities (JS), percentage of public places that meet health requirements (TTU), and percentage of places eating meets hygienic requirements (TPM). Meanwhile, the percentage of children under five who were given BCG immunization (BCG) influenced the number of TB cases only in some districts or cities in group 1.

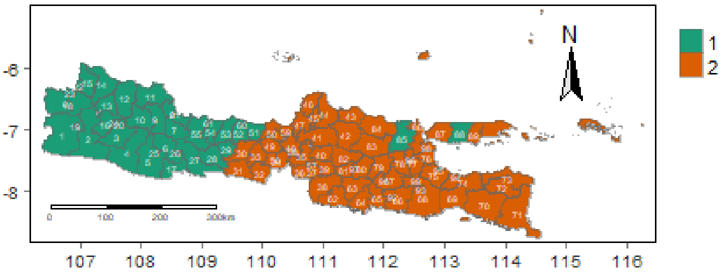

Figure 1. GWGPR Model Grouping Map in Java Based on Significant Independent Variables

Based on **Figure 1**, a map with horizontal axis is the degree of longitude and vertical axis is the degree of latitude, there were 40 districts/Cities in Java which are included in group 1, the group with all significant independent variables. It can be seen that the Districts or cities in group 1 tend to be in the west of the island. Whereas for group 2, the group with the variable percentage of children under five being given BCG immunization (BCG) was not significant, totaling 60 districts/cities, tending to be in the east of Java Island.

3.7 Geographically Weighted Negative Binomial Regression (GWNBR)

GWNBR was a development of Negative Binomial Regression, its development lies in the presence of spatial aspects that are considered in the estimation of model parameters. The spatial aspect is indicated by the spatial weighting matrix.

Hypothesis testing for model suitability is:

$H_0 : \beta_1(s_i,t_i) = \beta_2(s_i,t_i) = \ldots = \beta_p(s_i,t_i) = 0$

H_1: there is at least one $\beta_p(s_i,t_i) \neq 0,\ p = 1,2\ldots,7$

After processing the data using R studio, the deviance value is 76946.6 If the significance level is $\alpha = 0.05$, then $\chi^2_{(0.05;7)} = 14.0671$. Therefore, $deviance = 76946.6 > \chi^2_{(0.05;7)} = 14.0671$. It concluded that H_0 being rejected means there is at least one independent variable affects the number of TB cases.

Testing the hypothesis for the significance of the parameters partially is:

$$H_0 : \beta_p(s_i,t_i) = 0$$

$$H_1 : \beta_p(s_i,t_i) \neq 0,\ p = 1,2,\ldots,7$$

The partial test results produce different parameters for each district/city. Two groups were obtained based on the partial test results for the GWNBR model. The grouping results showed six variables have a significant effect on all cities or Districts in Java Island, namely population density (KP), percentage of poor people (PPM), percentage of drinking water facilities that meet the requirements (AIR), percentage of households with access to proper sanitation facilities (JS), percentage of public places that meet health requirements (TTU), and percentage of places eating meets hygienic requirements (TPM). Meanwhile, the percentage of children under five who were given BCG immunization (BCG) influenced the number of TB cases only in some districts or cities in group 1.

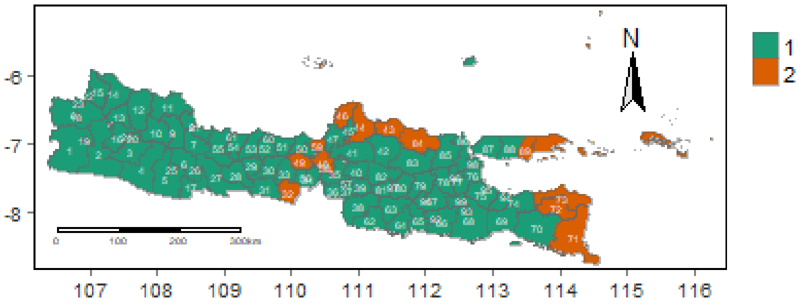

Figure 2. GWNBR Model Grouping Map in Java According to Significant Independent Variables

Based on Figure 2, a map with horizontal axis is the degree of longitude and vertical axis is the degree of latitude, there were 87 Districts or Cities in Java were included in group 1 or groups with all significant independent variables. It can be seen that the Districts or Cities in group 1 dominates Java Island both on the west and east sides. Whereas for group 2, the group with the variable percentage of children under five receiving BCG immunization (BCG) was not significant and the other independent variables were significant, totaling 13 districts or cities spread over a small part of Central Java and East Java.

3.8 Best Selection Model

The best model is determined using the AIC score of each model. The model with the smallest AIC score is the best model. Table 4 shows AIC Scores for every Models.

Table 4. AIC Scores for every Models

Model	AIC
Generalized Poisson Regression	1646.23
Negative Binomial Regression	1636
Geographically Weighted Generalized Poisson Regression	1562.579
Geographically Weighted Negative Binomial Regression	1562.449

Based on Table 4, the smallest AIC score is the AIC score for the GWNBR model. Thus, the best model for modeling the number of TB cases in Java in 2020 is the GWNBR model.

These results indicate that, according to the best model, the GWNBR model, the number of tuberculosis cases in Java in 2020 that occurred in 87 Districts or cities in Java Island was affected by all independent variables in which these Districts or cities dominated Java Island both in the west and east side while 13 other Districts or cities in Java Island are not affected by variable percentage of children under five receiving BCG immunization (BCG) in which the Districts or cities are spread over a small part of Central Java and East Java.

4. CONCLUSIONS

Based on the results of the analysis that has been carried out on data modeling on the number of tuberculosis cases in Java Island in 2020, several conclusions can be drawn, namely before paying attention to the spatial aspect, the Negative Binomial Regression model is a better model in explaining data on the number of tuberculosis cases compared to the Generalized Poisson Regression model, then based on the analysis of Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) performed with the kernel adaptive bisquare weighting function, 2 groups are produced with group 1 being the locations with all independent variables are significant and group 2 being the locations with all independent variables are significant except for the percentage under five given BCG immunization (BCG). In the GWGPR method, there are 40 Districts/Cities included in group 1 and 60 Districts or Cities included in group 2. Meanwhile in GWNBR, there are 87 Districts or Cities included in group 1 and 13 Districts or Cities included in group 2. The GWNBR model was chosen as the best model from four models that have been analyzed for the GWNBR model has the smallest AIC score among the other four models. This model can overcome the problem of overdispersion in spatial count data.

ACKNOWLEDGMENT

The Directorate of Research and Development of Universitas Indonesia (DRPM UI) funded this study as an additional paper through a grant of Hibah Publikasi Terindeks Internasional (PUTI) Q2 2022—2023 No.: NKB-668/UN2.RST/HKP.05.00/2022.

REFERENCES

[1] Kemenkes RI, “InfoDATIN Tuberkulosis 2018,” *Pusat Data dan Informasi Kementerian Kesehatan RI*, Jakarta, 2018. [Online]. Available: https://pusdatin.kemkes.go.id/article/view/18101500001/infodatin-tuberkulosis-2018.html

[2] Kompaspedia, “Sebaran Kasus Tuberkulosis di Indonesia,” 2021. https://kompaspedia.kompas.id/baca/infografik/peta-tematik/sebaran-kasus-tuberkulosis-di-indonesia

[3] B. Irawati, “Perbandingan Analisis Generalized Poisson Regression (GPR) dan Regresi Binomial Negatif untuk Mengatasi Overdispersi Studi Kasus: Pemodelan Jumlah Kasus Kanker Serviks di Jawa Timur,” *Jurnal Matematika*, vol. 2, no. 2, pp. 13–24, 2012.

[4] E. U. L. Fitri, “Pemodelan Faktor-Faktor yang Mempengaruhi Jumlah Kasus Tuberkulosis di Jawa Timur Menggunakan Metode Geographically Weighted Generalized Poisson Regression dan Geographically Weighted Negative Binomial Regression,” *Institut Teknologi Sepuluh Nopember, Surabaya*, 2017.

[5] S. Indahwati and M. Salamah, “Analisis Faktor-Faktor yang Memengaruhi Jumlah Kasus Tuberculosis di Surabaya Tahun 2014 Menggunakan Geographically Weighted Negative Binomial Regression,” *Jurnal Sains dan Seni ITS*, vol. 5, no. 2, 2016.

[6] A. S. N. Zaina, R. S. Pontoh, and B. Tantular, “Pemodelan Dan Pemetaan Penyakit TB Paru di Kota Bandung Menggunakan Geographically Weighted Negative Binomial Regression: Studi Kasus Dinas Kesehatan Kota Bandung,” in *E-Prosiding Seminar Nasional Statistika| Departemen Statistika FMIPA Universitas Padjadjaran*, 2021, p. 8.

[7] S. Noorcintanami, Y. Widyaningsih, and S. Abdullah, “Geographically weighted models for modelling the prevalence of tuberculosis in Java,” *J Phys Conf Ser*, vol. 1722, no. 1, p. 012089, Jan. 2021, doi: 10.1088/1742-6596/1722/1/012089.

[8] Dinas Kesehatan Provinsi Jawa Barat, *Profil Kesehatan Provinsi Jawa Barat*. Bandung, 2020. [Online]. Available: https://diskes.jabarprov.go.id/assets/unduhan/Profil%20Kesehatan%20Jawa%20Barat%20Tahun%202020.pdf

[9] Dinas Kesehatan Provinsi Jawa Tengah, *Profil Kesehatan Provinsi Jawa Tengah*. Semarang, 2020. [Online]. Available: https://dinkesjatengprov.go.id/v2018/dokumen/Profil2020/mobile/index.html

[10] Dinas Kesehatan Provinsi Jawa Timur, *Profil Kesehatan Provinsi Jawa Timur*. Surabaya, 2020. [Online]. Available: https://dinkes.jatimprov.go.id/userfile/dokumen/PROFIL%20KESEHATAN%202020.pdf

[11] BPS Provinsi Jawa Barat, “Persentase Penduduk Miskin (Persen), 2019-2021,” *Badan Pusat Statistik Provinsi Jawa Barat*, 2021. https://jabar.bps.go.id/indicator/23/51/1/persentase-penduduk-miskin.html

[12] BPS Provinsi Jawa Tengah, “Kemiskinan, 2019-2021,” *Badan Pusat Statistika Provinsi Jawa Tengah*, 2021. https://jateng.bps.go.id/indicator/23/34/1/kemiskinan.html

[13] BPS Provinsi Jawa Timur, “Jumlah dan Persentase Penduduk Miskin di Provinsi Jawa Timur Menurut Kabupaten/Kota, 2017-2021,” *Badan Pusat Statistika Provinsi Jawa Timur*, 2021. https://jatim.bps.go.id/indicator/23/421/1/jumlah-penduduk-miskin-menurut-kabupaten-Kota-di-jawa-timur.html

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[15] P. McCullagh, *Generalized linear models*. Routledge, 2019.

[16] F. Famoye, J. T. Wulu, and K. P. Singh, “On the generalized Poisson regression model with an application to accident data,” *Journal of Data Science*, vol. 2, no. 3, pp. 287–295, 2004.

[17] F. Famoye, “Restricted generalized poisson regression model,” *Commun Stat Theory Methods*, vol. 22, no. 5, pp. 1335–1354, Jan. 1993, doi: 10.1080/03610929308831089.

[18] J. M. Hilbe, *Negative binomial regression*. Cambridge University Press, 2011.

[19] L. Anselin, *Spatial econometrics: methods and models*, vol. 4. Springer Science & Business Media, 1988.

[20] A. R. da Silva and T. C. V. Rodrigues, “Geographically weighted negative binomial regression—incorporating overdispersion,” *Stat Comput*, vol. 24, pp. 769–783, 2014.

[21] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” *Selected papers of hirotugu akaike*, pp. 199–213, 1998.