PAPER • OPEN ACCESS

Evaluation of geographically weighted multivariate negative Binomial method using multivariate spatial infant mortality data

To cite this article: Y S Dewi *et al* 2019 *J. Phys.: Conf. Ser.* **1397** 012077

Evaluation of geographically weighted multivariate negative Binomial method using multivariate spatial infant mortality data

Y S Dewi[1,2], Purhadi[1*], Sutikno[1], and S W Purnami[1]

[1]Department of Statistics, Faculty of Mathematics, Computing and Data Sciences, Institut Teknologi Sepuluh Nopember, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Jember, Indonesia

[*]Corresponding author e-mail: purhadi@statistika.its.ac.id

Abstract. Global regression assumes that the relationships being measured are stationary over space or the model is applied equally over the whole region. If there is spatial heterogeneity on the data, then the global model is not suitable to the reality. To overcome multivariate spatial over dispersed negative binomial data, we evaluate geographically weighted multivariate negative binomial (local method) and compare it to the global method (multivariate negative binomial). The results show that the geographically weighted negative binomial performs better than the global method. The log likelihood of the local method is higher than the global method. The deviance and mean square prediction error of the local method are smaller than the global method. Moreover, the prediction of dependent variables of the local method are closer to the observed data than the global method. The estimated coefficients of the local method vary, depending on where the data are observed.

1. Introduction

Spatial data can be found in many fields, such as in transportation [1], [2], mortality [3], economic [4] and demographic [5]. The characteristic of spatial data is the existence of spatial effects on the data, the dependence between observation and location. The spatial effects can be categorized into two types [6], spatial dependence and spatial heterogeneity. In the spatial dependence, there is statistical dependence in a collection of random variables, each of which is associated with a different location. The spatial dependence leads to development of spatial area researches, Conditional Autoregressive, Spatial Autoregressive and Spatial Autoregressive Moving Average Model, among others [4], [6]. The spatial heterogeneity occurs as effects of differentiation between locations. It has the implication on the model vary between locations. It leads to development of Geographically Weighted Regression [7].

The Geographically Weighted Regression is a local regression model used to model spatially varying relationships. The spatial variation in coefficients can reveal interesting pattern contained in the spatial data. Some spatial count data researches were conducted using different techniques. For analyzing univariate spatial count data, Geographically Weighted Poisson Regression (GWPR) is described for analyzing non stationary count data and Semiparametric GWPR (SGWPR) model for mixed model is proposed, where there are global and local parameters in the model [3]. The estimation of parameters was conducted by using maximum likelihood estimation (MLE) method. The application of Generalized Linear Model (GLM) and GWPR on transportation case are described in [2]. The results show that the performance of GWPR model is better than GLM based on corrected Akaike Information Criterion (AICc). However, the GWPR assumes the equidispersion of the data. Unfortunately this assumption is often violated in the real data because data are often overdispersed. One of the methods for overcoming overdispersion is by using a geographically weighted negative binomial model [1].

On the other hand, many researchers have been developing the global regression models for Poisson data with overdispersion. For univariate dependent variable, see [8], [9], [10] and [11]. For more than one dependent variables, the multivariate negative binomial model using copula and MLE method is proposed [12]. The other researches are a robust likelihood approach for the overdispersed correlated count data analysis based on a multivariate negative binomial model using MLE via iterative Newton Raphson algorithm [13], a multivariate generalized Poisson regression model using MLE method [14] and the comparison of two bivariate negative binomial regression models that come from the different distributions derivation [15]. In a similar way to [16], a bivariate negative binomial distribution as a product of negative binomial marginals with a multiplicative factor is defined in [17]. The parameters are estimated using MLE via Newton Raphson algorithm.

There is a relationship between observations from different locations. The near things are more related than the distant things. Accordingly, when we use global method for analyzing spatial data, there is violation to the condition, because the global method assumes that there are no relationships between observations or observations are independent in different locations. Moreover the global regression assumes that the relationship is stationary modeled throughout space, although it might be non-stationary. These conditions are not always suitable for spatial data because there are spatial effects on it. For that reason, if there is non-stationariness of the relationships between locations, the global model is not suitable with reality. Therefore it is necessary to use appropriate method to model such data. Unfortunately, the researches related to multivariate spatial data are less developed. On the other side, many data in real applications are multivariate spatial data. Therefore, the purpose of this research is to evaluate the geographically weighted multivariate negative binomial method on multivariate spatial overdispersed negative binomial data using infant mortality data with spatial effect and overdispersion. We compare it to the multivariate negative binomial (global method). We investigate five aspects in this research, the likelihood, deviance, mean square prediction error, coefficients estimate and closeness of the prediction (means of dependent variables) to the observed data.

2. Geographically weighted multivariate negative Binomial method

The model in this research is built based on multivariate negative binomial distribution , where the observation y_{ij} follow Poisson distribution with assumption $Y_{ij} \mid \varphi_i$ □ Poisson $\left(\varphi_i \mu_{ij}\right)$, i = 1,2,..., n, j = 1,2,...,m, μ_{ij} is the mean of dependent variable-j on location-i and φ_i is unobservable effect of location-i, φ_i follow gamma distribution or φ_i □ Gamma $\left(\tau^{-1},\tau^{-1}\right)$, where $E\left(\varphi\right)=1$ and $Var\left(\varphi\right)=\tau$.

2.1. The model

Let y_{ij} is the observation on location-i, dependent variable-j, where i = 1,2,…,n and j = 1,2,…,m. By using the assumption above and letting $\delta = 1/\tau$, the joint probability density function of $\left(\mathbf{y}_i, \varphi_i\right)$ is given by

$$f\left(\mathbf{y}_i,\varphi_i;\boldsymbol{\mu},\delta\right)=\frac{\left(\prod_{j=1}^{m}\mu_{ij}{}^{y_{ij}}\right)\delta^{\delta}}{\left(\prod_{j=1}^{m}y_{ij}\,!\right)\Gamma(\delta)}\exp\left(-\varphi_i\left(\sum_{j=1}^{m}\mu_{ij}+\delta\right)\right)\varphi_i^{\sum_{j=1}^{m}y_{ij}+\delta-1} \tag{1}$$

By letting $a_i=\varphi_i\left(\sum_{j=1}^{m}\mu_{ij}+\delta\right)$ and integrating the function of a_i with respect to φ_i, the marginal probability mass function of $\mathbf{y}_i$ can be written as follows

$$f\left(\mathbf{y}_i;\boldsymbol{\mu},\delta\right)=\frac{\left(\prod_{j=1}^{m}\mu_{ij}^{y_{ij}}\right)\delta^{\delta}\Gamma\left(\delta+y_{i+}\right)}{\left(\prod_{j=1}^{m}y_{ij}\,!\right)\Gamma\left(\delta\right)\left(\delta+\mu_{i+}\right)^{\delta+y_{i+}}} \tag{2}$$

for i = 1,2,...,n and j = 1,2,...,m, $y_{i+}=\sum_{j=1}^{m}y_{ij}$ and $\mu_{i+}=\sum_{j=1}^{m}\mu_{ij}$. The probability mass function (2) is known as multivariate negative binomial, where

$E\left(Y_{ij}\right)=\mu_{ij}$, $\operatorname{Var}\left(Y_{ij}\right)=\dfrac{\mu_{ij}^2}{\delta}+\mu_{ij}$ and $\operatorname{Corr}\left(Y_{ij},Y_{il}\right)=\dfrac{\sqrt{\mu_{ij}\mu_{il}}}{\sqrt{\delta+\mu_{ij}}\sqrt{\delta+\mu_{il}}}$.

The likelihood function of (2) is given by

$$L\left(\boldsymbol{\theta}\right)=\prod_{i=1}^{n}\frac{\left(\prod_{j=1}^{m}\mu_{ij}^{y_{ij}}\right)\delta^{\delta}\Gamma(y_{i+}+\delta)}{\prod_{j=1}^{m}y_{ij}\,!\,\Gamma(\delta)(\delta+\mu_{i+})^{(\delta+y_{i+})}} \tag{3}$$

Subsequently, by defining $\mu_{ij}\left(u_i,v_i\right)=t_{ij}e^{\mathbf{x}_i^T\boldsymbol{\beta}_j\left(u_i,v_i\right)}$, we build the geographically weighted multivariate negative binomial, where $\mu_{ij}\left(u_i,v_i\right)$ is the mean of dependent variable-j, at location-i, while $\left(u_i,v_i\right)$ is the coordinate of the location-i, t_{ij} is exposure variable of location-i and dependent variable-j, x_{ij} is independent variable at location-i and dependent variable-j, and $\boldsymbol{\beta}_j\left(u_i,v_i\right)$ is the vector coefficients of the regression at location-i and dependent variable-j. Based on [1] and equation (3), the log likelihood function of the local model can be written as

$$\ell\left(\boldsymbol{\theta}^*\right)=\sum_{i=1}^{n}\left\{\left(\sum_{j=1}^{m}y_{ij}\ln\left(t_{ij}e^{\mathbf{x}_i^T\boldsymbol{\beta}_j\left(u_l,v_l\right)}\right)\right)-\left(\delta\left(u_l,v_l\right)+y_{i+}\right)\ln\left(\delta\left(u_l,v_l\right)+\sum_{j=1}^{m}\left(t_{ij}e^{\mathbf{x}_i^T\boldsymbol{\beta}_j\left(u_l,v_l\right)}\right)\right)+A_i\right\}w_{il} \tag{4}$$

where i,l= 1,2,3,…,n, $A_i=\delta\left(u_l,v_l\right)\ln\delta\left(u_l,v_l\right)-\sum_{j=1}^{m}\ln y_{ij}\,!+\ln\left\{\dfrac{\Gamma(y_{i+}+\delta\left(u_l,v_l\right))}{\Gamma(\delta\left(u_l,v_l\right))}\right\}$,

$\boldsymbol{\theta}^*=\left[\boldsymbol{\beta}_1^T\left(u_l,v_l\right)\ \boldsymbol{\beta}_2^T\left(u_l,v_l\right)\ldots\boldsymbol{\beta}_m^T\left(u_l,v_l\right)\ \delta\left(u_l,v_l\right)\right]^T$ and w_{il} is the geographical weight.

2.2. The geographical weight and cross validation

The geographical weight is a value of weights relative to the position of (u_i, v_i) in the study area. The weights themselves are computed from a weighting scheme which is also known as kernel. In this research, we use the Fixed Bisquare Kernel weight, w_{il} as a continuous function of d_{il}. It can be defined as shown in [7],

$$w_{il} = \begin{cases} \left(1 - \left(\dfrac{d_{il}}{b}\right)^2\right)^2 ; & \text{untuk } d_{il} \leq b \\ 0 \; ; & \text{untuk } d_{il} > b \end{cases} \tag{5}$$

where b is referred to as the bandwidth, and d_{il} is the Euclidean distance between point-*i* and *l*.

The optimum bandwidth is obtained by cross-validation using the formula

$$CV(b) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} - y_{\neq ij}(b) \right)^2 \tag{6}$$

where $\hat{y}_{\neq ij}(b)$ is the fitted value of y_{ij} with the observation for location-*i* omitted from the calibration process, *n* and *m* is the number of location and dependent variables respectively. The optimum bandwidth is the bandwidth with minimum *CV*.

3. Method and algorithm

We use infant mortality data 2014 from East, Middle and West Java, Indonesia [18], [19], [20]. There are three dependent variables (Y_1, Y_2 and Y_3) which have positive correlation between them and six independent variables ($X_1 - X_6$), where they are direct and underlying causes of infant mortality. The description of the variables used in this research is presented in Table 1.

Table 1. Research variables of infant mortality data.

Variable	Description
Y_1	The number of birth deaths
Y_2	The number of neonatal deaths (after birth to 1 month)
Y_3	The number of infancy deaths (1 month to 1 year)
X_1	The percentage of handling obstetric complications
X_2	The percentage of households that have healthy behavior
X_3	The percentage of integrated health posts giving service actively
X_4	The percentage of active family planning participant
X_5	The percentage of prenatal visits to the health worker minimum four times
X_6	The percentage of prenatal getting Fe3 (90 pills)

We use the coordinates of latitude and longitude of the regencies/towns in East, Middle and West Java Indonesia as geographical factor of locations. There are 38, 35 and 27 regencies/towns in East, Middle and West Java, Indonesia respectively. Thus, there are 100 locations in total.

Geographically Weighted Negative Binomial uses MLE method via Newton Raphson algorithm for the estimation of parameters [1]. However, the algorithm has the weakness in dealing with false convergence due to improper initial value. Moreover, when analyzing more than one dependent variables, things are more complicated. In this research, we use MLE method via Nelder Mead algorithm for estimating the mean of dependent variables. Nelder Mead is an alternative algorithm for Newton Raphson. This is free derivative and robust algorithm related to the initial value for geographically weighted multivariate method [21]. The method is nonlinear optimization technique for maximizing or minimizing function on multi dimension. The Nelder Mead algorithm is a popular algorithm, because it is reliable enough even for researches in high dimension.

The algorithm proceeds through operations of the simplex to find local optimum of log likelihood function. Let $\mathbf{\Theta}$ is p dimension parameter space and $\ell\left(\boldsymbol{\theta}^*\right)$ is a function that we want to maximize (log likelihood function) or $-\ell\left(\boldsymbol{\theta}^*\right)$ minimized, where $\boldsymbol{\theta}^*\in\mathbf{\Theta}$ and $\boldsymbol{\theta}^*=[\beta_{1,0}\left(u_l,v_l\right)\ \beta_{1,1}\left(u_l,v_l\right)\ \ldots\ \beta_{1,p-1}\left(u_l,v_l\right)$ $\beta_{2,0}\left(u_i,v_i\right)\ \beta_{2,1}\left(u_l,v_l\right)\ \ldots\beta_{2,p-1}\left(u_l,v_l\right)\ \beta_{3,0}\left(u_l,v_l\right)\ \ldots\ \beta_{m,p-1}\left(u_l,v_l\right)\ \delta\left(u_l,v_l\right)]$. Each iteration is started from simplex, that is the structure formed by $p+1$ points, not in the same plane, in an p-dimension space, $p=mp+1$ where *m* is the number of dependent variables and *p* is the number of parameters for each dependent variable. Based on the evaluation of $-\ell\left(\boldsymbol{\theta}^*\right)$ on the vertex, the vertex having the worst function value is replaced by a new point with a better function value. By assuming that optimum is minimum, the vertex with highest value of the function is replaced by the new point with a lower value of the function through one of reflection, expansion or contraction operations. If all of these operations fail to find new point to replace the worst point, then the simplex will shrink to the vertex with lowest function value. See [22] and [23] for Nelder Mead algorithm in details. The cross validation is conducted using the golden section algorithm to find an optimum bandwidth. The initial estimate of geographically weighted multivariate negative binomial is taken from the coefficients and index of dispersion estimate of multivariate negative binomial for estimating the mean of dependent variables. The data analysis is conducted using R.3.4.3 software with MASS, REdas, spgwr and GW model packages.

4. Result and Discussion

The model used in this research consists three dependent variables. All correlation coefficients between the dependent variables are positive. They are 0.552, 0.449 and 0.706 for the correlation between Y_1 and Y_2, Y_1 and Y_3, and Y_2 and Y_3 respectively. Moreover the variance of dependent variables are much larger than their mean, it indicates the existence of overdispersion in the data. The variance of Y_1, Y_2 and Y_3 are 3507.36, 4444.65 and 547.00, while their means are 94, 117 and 31 respectively. Therefore, we consider multivariate negative binomial distribution for this data.

There is a high correlation between two independent variables X_5 and X_6 (Correlation = 0.909), We omit X_6 from the data analysis for overcoming multicollinearity. The selection of independent variables uses a forward stepwise method. The model with four independent variables (X_1, X_2, X_3 and X_4) has the best performance. Therefore we use those variables for further analysis.

By using four variables (X_1, X_2, X_3 and X_4) and Nelder-Mead algorithm~~s~~ in section 3, we find the optimum bandwidth b = 4.298 and CV = 885910.30. The geographically weighted multivariate negative binomial (local method) performs better than the multivariate negative binomial (global method) as shown in Table 2. Three criteria of the likelihood, deviance and mean square prediction error (MSPE) are used to compare the two methods. The local method has higher log likelihood and smaller deviance than the global method. Moreover, the behavior of the residual of the local method is better than the global method, indicated by a small MSPE value. The MSPE of local method is smaller than the global method.

Table 2. Goodness of fit of the local and global methods.

Method	Log likelihood	Deviance	MSPE
Local	-2049.7	1925.3	2500.8
Global	-2122.2	2073.4	2542.2

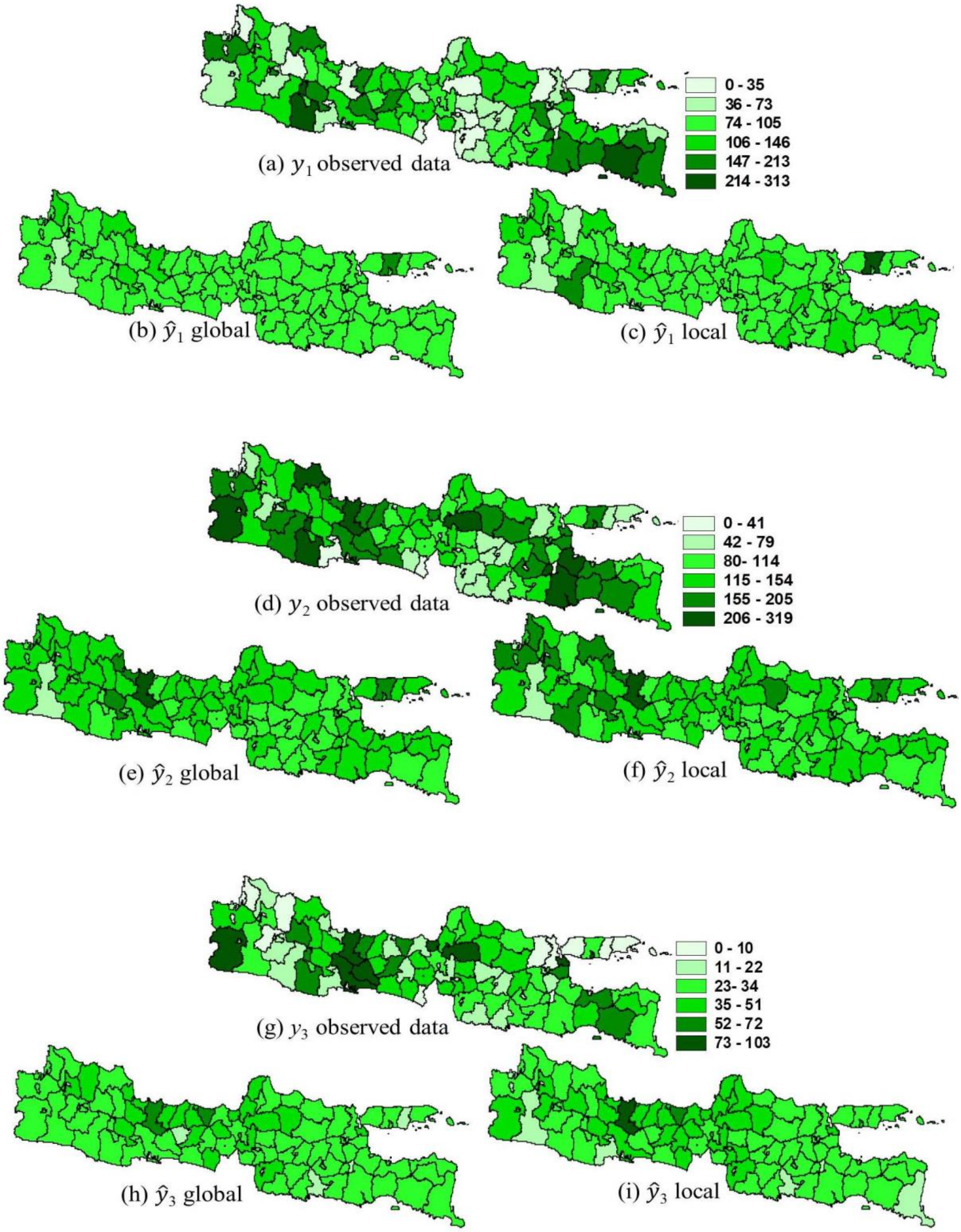

Figure 1. Comparison of observed data and their prediction using global and local methods

The comparison of observed data (**y**$_1$, **y**$_2$ and **y**$_3$) and their prediction are presented in Figure 1. The darker color represents the higher of deaths than lighter color. That figure shows that the prediction of local method tends to close to the observed data than the global method. This indicates that the local method predicts better the dependent variables than the global method. This is in accordance with

MSPE result, that the local method is better than the global method. Although the local method has better prediction than global method, the time for local method is longer than it for global method, mainly for cross validation process. Building one model in local method similar to building *n* model in global method, where *n* is the number of locations.

The coefficients estimates of global and local methods and their standard error (in parentheses) are presented in Table 3. They show that there are differences between the coefficients estimate resulted by both methods. This is because the coefficients estimates of spatial data model have the local character, those depend on the location where the data are observed.

Table 3. Estimated coefficients for the three locations using global and local methods.

Coefficient	Global		Local					
			Madiun		Cilacap		Karawang	
β_{10}	4.505	(0.459)	4.678	(0.933)	4.440	(0.664)	4.451	(1.877)
β_{11}	0.001	(0.003)	0.004	(0.008)	0.002	(0.004)	-0.002	(0.005)
β_{12}	-0.002	(0.003)	-0.003	(0.005)	-0.002	(0.004)	0.015	(0.007)
β_{13}	-0.003	(0.003)	-0.011	(0.006)	-0.003	(0.004)	-0.008	(0.005)
β_{14}	0.003	(0.003)	0.004	(0.003)	0.003	(0.005)	-0.003	(0.023)
β_{20}	4.657	(0.458)	4.622	(0.931)	4.700	(0.663)	5.622	(1.870)
β_{21}	0.005	(0.003)	0.009	(0.008)	0.005	(0.004)	0.003	(0.005)
β_{22}	0.000	(0.003)	0.002	(0.005)	0.000	(0.004)	0.010	(0.007)
β_{23}	-0.007	(0.003)	-0.015	(0.006)	-0.009	(0.004)	-0.012	(0.005)
β_{24}	0.002	(0.003)	0.002	(0.003)	0.001	(0.005)	-0.015	(0.023)
β_{30}	2.387	(0.471)	2.757	(0.956)	2.466	(0.680)	2.438	(1.930)
β_{31}	0.008	(0.003)	0.009	(0.008)	0.009	(0.004)	0.007	(0.005)
β_{32}	0.007	(0.003)	0.011	(0.005)	0.006	(0.005)	0.008	(0.007)
β_{33}	-0.004	(0.003)	-0.015	(0.006)	-0.005	(0.004)	-0.004	(0.005)
β_{34}	0.001	(0.003)	0.001	(0.003)	0.001	(0.005)	-0.001	(0.024)
τ	0.339	(0.046)	0.396	(0.087)	0.322	(0.062)	0.307	(0.092)
Loglikelihood	-2122.244		-2049.741					
Deviance	2073.387		1925.304					
MSPE	2542.177		2500.843					

5. Conclusion

Geographically weighted multivariate negative binomial is a development method of geographically weighted negative binomial. This method is used when there are multivariate spatial count data having positive correlation between them. The dependent variables are predicted by independent variables, where each location has the local character coefficients based on where the data are observed.

The geographically weighted multivariate negative binomial performs better than the global method (multivariate negative binomial) in predicting the means of multivariate spatial overdispersed data. This is indicated by higher log likelihood and smaller deviance and MSPE. Moreover, the local method predicts the dependent variables better than the global method. The prediction of the local method tends to be closer to the observed data than the global method. However the computational time for local method is longer than the global method.

Acknowledgement

The authors thank East, Middle and West Java Provincial Health Department for providing access to the infant mortality data.

References

[1] da Silva A R and Rodrigues T C V 2014 Geographically Weighted Negative Binomial Regression-incorporating overdispersion *Stat. Comput.* **24** 769–83

[2] Hadayeghi A, Shalaby A S and Persaud B N 2010 Development of planning level transportation safety tools using Geographically Weighted Poisson Regression *Accid. Anal. Prev. J.* **42** 676–88

[3] Nakaya T, Fotheringham A S, Brunsdon C and Charlton M 2005 Geographically weighted Poisson regression for disease association mapping *Stat. Med.* **24** 2695–717

[4] LeSage J and Kelley R 2009 *Introduction to Spatial Econometrics* (Chapman and Hall/CRC)

[5] Mullen W F, Jackson S P, Croitoru A, Crooks A, Stefanidis A and Agouris P 2015 Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features *GeoJournal* **80** 587–605

[6] Anselin L 2001 Spatial econometrics *A companion to Theor. Econom.* 310–30

[7] Fotheringham A S, Brunsdon C and Charlton M 2002 Geographically Weighted Regression-The Analysis of Spatially Varying Relationships

[8] Dean C and Lawless J F 1989 Tests for detecting overdispersion in poisson regression models *J. Am. Stat. Assoc.* **84** 467–72

[9] Consul P C and Famoye F 1992 Generalized poisson regression model *Commun. Stat. - Theory Methods* **21** 89–109

[10] Ridout M, Demetrio C G . and Hinde J 1998 Models for count data with many zeros *Int. Biometric Conf.* 1–13

[11] Famoye F and Singh K P 2006 Zero-Inflated Generalized Poisson Regression Model with an Application to Accident Data *J. Data Sci.* **4** 117–30

[12] Shi P and Valdez E A 2014 Multivariate negative binomial models for insurance claim counts *Insur. Math. Econ.* **55** 18–29

[13] Solis-Trapala I L and Farewell V T 2005 Regression analysis of overdispersed correlated count data with subject specific covariates *Stat. Med.* **24** 2557–75

[14] Famoye F 2015 A Multivariate Generalized Poisson Regression Model *Commun. Stat. - Theory Methods* **44** 497–511

[15] Dewi Y S, Purhadi, Sutikno and Purnami S W 2017 Comparison of Bivariate Negative Binomial Regression Models for Handling Over dispersion *Int. J. Appl. Math. Stat.* **56** 53–62

[16] Lakshminarayana J, Pandit S N N and Srinivasa Rao K 1999 On a bivariate poisson distribution *Commun. Stat. - Theory Methods* **28** 267–76

[17] Famoye F 2010 On the bivariate negative binomial regression model *J. Appl. Stat.* **37** 969–81

[18] Dinas Kesehatan Provinsi Jawa Timur 2015 *Profil Kesehatan Provinsi Jawa Timur 2014*

[19] Dinkes Jawa Tengah 2014 *Profil Kesehatan Provinsi Jateng Tahun 2014*

[20] Dinas Kesehatan Provinsi Jawa Barat 2015 *Profil Kesehatan Provinsi Jawa Barat Tahun 2014*

[21] Dewi Y S, Purhadi, Sutikno and Purnami S W 2019 Comparison of Nelder Mead and BFGS Algorithms on Geographically Weighted Multivariate Negative Binomial *Int. J. Adv. Sci. Enginering Inf. Technol.* **9** 979–87

[22] Nash J C 2014 On best practice optimization methods in R *J. Stat. Softw.* **60** 1–14

[23] Small C G and Wang J 2003 *Numerical methods for nonlinear estimating equations* (Oxford University Press on Demand)