Journal of Mathematics and Statistics

Original Research Paper

Zero Inflated Poisson and Geographically Weighted Zero-Inflated Poisson Regression Model: Application to Elephantiasis (*Filariasis*) Counts Data

[1]Purhadi, [2]Yuliani Setia Dewi and [3]Luthfatul Amaliana

[1,3]Department of Statistics, Faculty of MIPA, Sepuluh Nopember Institute of Technology Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
[2]Department of Mathematics, Faculty of MIPA, University of Jember, Jl Kalimantan 37 Jember, 68121 Indonesia

Article history
Received: 21-11-2014
Revised: 16-08-2015
Accepted: 11-09-2015

Corresponding Author:
Purhadi
Department of Statistics, Faculty of MIPA, Sepuluh Nopember Institute of Technology Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
Email: purhadi@statistika.its.ac.id

Abstract: Poisson regression has been widely used for modeling counts data. Violation of equidispersion assumption can occur when there are excess of zeros of the data. For that condition we can use Zero-*Inflated Poisson* (ZIP) to analyze such data, resulting global parameter estimates. However spatial data from various locations have their own characteristics depend on their socio-cultural, geographical and economic conditions. In this paper, we first review the theoretical framework of Zero-Inflated Poisson (ZIP) and Geographically Weighted Zero Inflated Poisson (GWZIP) regression. We use Maximum Likelihood (MLE) method and EM algorithm to estimate the model parameters. The F test is used to compare the two models. Second, we fit these models to the number of *filariasis* case of East Java. In our case, there is the preponderance of zeros in the data set (65.79%). The results prove that the spatial dependence is absent, but there is weak spatial heterogeneity of the data (significance level $\alpha = 0.1$). Based on *F* test, ZIP and GWZIP regression are not significantly different.

Keywords: ZIP, GWZIP, Spatial Data, MLE, F Test, *Filariasis* Case

Introduction

Poisson regression is the well known method for modelling counts data. However this method assumes the equidispersion of the data (Bohning *et al*., 1999). Unfortunately this assumption is often violated in the observed data because data are often overdispersed. Generally, two sources of overdispersion are determined: heterogeneity of the population and excess of zeros (Khoshgoftaar *et al*., 2004; Mouatassim and Ezzahid, 2012). When the source of overdispersion is the excess zeros, the Zero Inflated Poisson Regression model fits counts data well (Lambert, 1992; Mouatassim and Ezzahid, 2012). One of method to estimate parameters of ZIP Regression is Maximum Likelihood Estimation (MLE) method. The log likelihood function can be maximized using EM algorithm.

Many research themes of ZIP Regression have been developed. Some of them are Lestari (2008) modeled counts data of commercial sex worker at Clinic for human reproduction, Putat Jaya, Surabaya; Bohning *et al*. (1999) used ZIP to analyze counts data on prevention of dental caries in children; and Mouatassim and Ezzahid (2012) fitted models to the number of claims in a private health insurance scheme. Those research themes showed that ZIP Regression fitted counts data better than Poisson Regression.

ZIP Regression estimates the parameters globally. However, data from various locations show the different conditions of them. Those are influenced by different socio-cultural, geographical and economic between them. Those conditions indicate spatial factors. Until now, the research topics of ZIP Regression have not taken into account spatial factors yet. In this study we review ZIP Regression and its development, Geographically Weighted Zero-Inflated Poisson (GWZIP) Regression to analyze excess zero counts data by considering spatial factor.

As illustration we use East Java elephantiasis (*filariasis*) poisson counts data 2012, with the proportion of zero counts data is 65.79%. *Filariasis* is an infectious tropical disease. Someone can get it from a bite by an infected mosquito, like malaria, leprous and dengue (Wulandari *et al*., 2010). The counts of *Filariasis* in East Java can be affected by spatial heterogeneity.

Some research themes about *filariasis* have been developed. For more details, one can refer to (Wulandari *et al*., 2010; Nasrin, 2008; Juriastuti *et al*., 2010). Most of them used techniques and tests to find influenced significant factors to *filariasis* case without taking into account spatial factors.

Different from ZIP Regression to *filariasis* case before, in this study we develop ZIP Regression by taking into account spatial factors.

Poisson Regression

Poisson Regression is special case of Generalized Linear Model (GLM). Standart GLM for counts data is Poisson Regression model with log link function. Poisson Regression is given by the equation below (Myers *et al*., 1990; Greene, 2003; Cameron and Trivedi, 2005):

$$\mu_i = exp\left(\mathrm{x}_i^T \beta\right) \tag{1}$$

y_i is respon variabel that follows a Poisson distribution, μ_i is the average of counts of events during a specified period.

The vector $\mathrm{x}_i^T = [x_{i,1}, x_{i,2},\ldots,x_{i,k}]$ contains the covariates and $\beta^T = [\beta_1,\beta_2,\ldots,\beta_k]$ is the vector of unknown parameters. The number k defines the dimension of the covariates vector incorporated in the model. The link function is $ln(\mu_i)$.

Where:

$$ln\left(\mu_i\right) = \mathrm{x}_i^T \beta \tag{2}$$

Maximum likelihood techniques may be used to estimate the parameters of the Poisson regression, using Iteratively Reweighted Least Square (IRLS) method (Myers *et al*., 1990) or Newton-Raphson method (Cameron and Trivedi, 2005; Greene, 2003). Given the assumption that the observations $(y_i|\mathrm{x}_i)$ are independent, the ln-likelihood function is given by:

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i \mathrm{x}_i^T \beta - \sum_{i=1}^{n} e^{\mathrm{x}_i^T \beta} - \sum_{i=1}^{n} ln\left(y_i!\right) \tag{3}$$

The null and alternative simultaneous parameters hypotheses are given below:

H_0: $\beta_1 = \cdots = \beta_k = 0$
H_1: Minimum one of $\beta_j \neq 0$, $j = 1,2,\ldots,k$

And the deviance statistic is (Myers *et al*., 1990; Greene, 2003; Cameron and Trivedi, 2005):

$$D\left(\hat{\beta}\right) = -2ln\left[\frac{L\left(\hat{\omega}\right)}{L\left(\hat{\Omega}\right)}\right] \tag{4}$$

We reject H_0 when $D\left(\hat{\beta}\right) > \chi^2_{(\alpha,n-p)}$, where $(n-p)$ is the number of degree of freedom. The deviance value declines when the number of parameters in the model increases (McCullagh and Nelder, 1989).

Test for partial parameters is written by following hypotheses:

H_0: $\beta_j = 0$; $j = 1,\ldots,k$
H_1: $\beta_j \neq 0$

The statistic test is:

$$Z_j = \frac{\hat{\beta}}{se\left(\hat{\beta}_j\right)} \tag{5}$$

We reject H_0 when $|Z_j| > Z_{\alpha/2}$, var $\left(\hat{\beta}_j\right)$ is $(j+1)^{rd}$ diagonal element of negative $\left[H\left(\hat{\beta}\right)^{-1}\right]$.

Zero-Inflated Poisson Regression

Zero-Inflated Poisson Regression (ZIPR) is proposed by Lambert (1992) to handle excess zero counts data. Observations $Y_1, Y_2,\ldots,Y_n$ are independent each other and $Y_i \sim \begin{cases} 0, \text{with probability } \pi_i \\ \text{Poisson}\left(\mu_i\right), \text{with probability } \left(1-\pi_i\right) \end{cases}$.

The probability function of Y_i is given below:

$$p\left(Y_i = y_i\right) = \begin{cases} \pi_i + \left(1-\pi_i\right)e^{-\mu_i}, & y_i = 0 \\ \dfrac{\left(1-\pi_i\right)e^{-\mu_i}\mu_i^{\,y_i}}{y_i!}, & y_i > 0 \end{cases}$$

Where:

$$\mu_i = e^{\mathrm{x}_i^T \beta} \text{ and } \pi_i = \frac{e^{\mathrm{x}_i^T \gamma}}{1+e^{\mathrm{x}_i^T \gamma}} \tag{6}$$

ZIP Regression model can be written below:

$ln\,\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$; $i = 1,\ldots,n$
$\text{logit}\,\pi_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik}$; $i = 1,\ldots,n$

Where, $\beta_{(k+1)\mathrm{x}1}$ and $\gamma_{(k+1)\mathrm{x}1}$ are the parameters of ZIP regression, $\mathrm{X}_{n\mathrm{x}(k+1)}$ is predictor variable assosiated with probability of zero state ($y_i = 0$) and mean of poisson state ($y_i > 0$).

The Y_i variable is redefined by latent variable Z_i, where Z_i~Binomial $(1, \pi_i)$. MLE method is used to estimate parameters by using Expectation–Maximization (EM) algorithm. This algorithm is iterative method to maximize likelihood function with missing data or latent variable. The function of ln likelihood is given by the equation:

$$\ln L\left(\beta,\gamma \mid y,z\right) = \sum_{i=1}^{n}\left(z_i \mathrm{x}_i^T \gamma - \ln\left(1+e^{\mathrm{x}_i^T \gamma}\right)\right)$$
$$-\sum_{i=1}^{n}\left(1-z_i\right)\ln\left(y_i!\right) + \sum_{i=1}^{n}\left(1-z_i\right)\left(y_i \mathrm{x}_i^T \beta - e^{\mathrm{x}_i^T \beta}\right)$$

Estimation of the parameters of ZIP Regression is carried out by two step, expectation and maximization in EM algorithm.

Simultaneous parameters hypotheses testing of β and γ are given by:

H_0: $\beta_1=\cdots=\beta_k=\gamma_1=\cdots=\gamma_k=0$,
H_1: Minimum one of $\beta_j \neq 0$ or $\gamma_j \neq 0$, $j = 1,2,\ldots,k$

And deviance statistic is written below:

$$G = -2ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right] \tag{7}$$

We reject H_0 when $G > \chi^2_{(\alpha,n-p)}$, where $(n\text{-}p)$ is the number of parameters under population minus the number of parameters under H_0 true. By the same way, we carry out testing of each β and γ. Then partial testing of parameters β and γ are given below:

H_0: $\beta_j = 0$ and H_0: $\gamma_j = 0$
H_1: $\beta_j \neq 0$ and H_1: $\gamma_j \neq 0$

The test statistic that is used for the hypotheses testing is deviance as written on Equation (7).

Multicollinearity

Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. Some authors have suggested a formal detection-tolerance or the Variance Inflation Factor (VIF) for multicollinearity. The formula of VIF is given below (Gujarati, 2004):

$$VIF_j = \frac{1}{(1-R_j^2)} \tag{8}$$

R_j^2 is the coefficient of determination of X_j explanatory variable on all the other X variables. VIF above 10 indicates multicollinearity problem. The multicollinearity can be handled by dropping predictor variables that are highly correlated, increasing the number of sample, ridge regression analysis or Principal Component Analysis (PCA).

Spatial Effects

The problems of spatial data consist of spatial dependence and spatial heterogeneity. Spatial data modeling include spatial weight matrix which its elements are the function of Euclid distance between locations. Weight matrix is built by using kernel functions, one of them is Gauss function, that is:

$$w_{il}(u_i,v_i) = exp\left(-\left(\frac{d_{il}}{h}\right)^2\right) \tag{9}$$

where, $d_{il} = \sqrt{(u_i-u_1)^2+(v_i-v_1)^2}$ is Euclid distance between location i and location l, while h is optimum bandwidth that is gotten from minimum CV.

$$CV(h) = \sum_{i=1}^{n}\left(y_i - \hat{y}_{\neq l}(h)\right)^2 \tag{10}$$

Spatial dependence indicates the dependence of observations between locations. The observation on one location influences the observation on other locations. We use Moran's index to test spatial dependence. The value of Moran's index is between -1 and 1. The formula is written below (Anselin, 1988):

$$I = \frac{n\sum_i\sum_l w_{il}(y_i-\bar{y})(y_l-\bar{y})}{\left(\sum_i\sum_l w_{il}\right)\sum_i(y_i-\bar{y})^2} \tag{11}$$

Moran's I test with the hypotheses H_0: $\lambda = 0$ (dependence spatial is present) and H_1: $\lambda \neq 0$ (There is not spatial dependence), is carried out by using Z test:

$$Z_I = \frac{I - E(I)}{\sqrt{Var(I)}} \tag{12}$$

where, we reject H_0 when $|Z_I| > Z_{\alpha 2}$.

Spatial heterogeneity effects can be identified by using *Breusch-Pagan* testing (Anselin, 1988). However *BreuschPagan* testing is sensitive to normality assumption. Because of that in this study we use *Koenker-Basset* test (Gujarati, 2004). The testing is carried out by regressing square error $(\hat{\varepsilon}_i^2)$ and square of the result of estimation. Then we test significance of the parameter. The hypotheses are H_0: represents the absence of the spatial heterogeneity Vs H_1: H_0 is not true. We test the hypothesis using Z test with criterion rejecting H_0 when $|Z_I| > Z_{\alpha/2}$ or *p-value* < α.

Geographically Weighted Zero-Inflated Poisson Model

GWZIPR Model is local model of Zero-Inflated Poisson (ZIP) Regression with local parameter estimates. Each observation of the respon variable is taken from different (u_i, v_i) location with probability $\pi_i + (1-\pi_i)\, e^{-\mu i}$ for $y_i = 0$ and $\frac{(1-\pi_i)e^{-\mu_i}\mu_i^{y_i}}{y_i!}$ for $y_i > 0$, following the equation below:

$$\mu_i = e^{x_i^T \beta(u_i,v_i)} \; and \; \pi_i = \frac{e^{x_i^T \gamma(u_i,v_i)}}{1+e^{x_i^T \gamma(u_i,v_i)}} \tag{13}$$

β (u_i, v_i) and γ (u_i, v_i) are parameters of regression on (u_i, v_i) location, X is predictor variable related to probability *of zero state* (y_i = 0) and mean of poisson state (y_i > 0).

The involvement of factor of geographical location in GWZIPR is expressed by a (u_i, v_i) coordinate. Geographical factor is the weight on GWZIPR model expressing local characteristic of parameters for each location.

Parameter Estimation of GWZIPR Model

By involving geographical factors on ZIPR and using MLE and Expectation-Maximization (EM) algorithm, the likelihood function and ln likelihood of GWZIPR model is given by the equation below:

$$L\left(\gamma(u_i,v_i),\beta(u_i,v_i)\right) = \prod_{\substack{l=1 \\ y_l=0}}^{n}\left(\frac{e^{x_i^T\gamma(u_i,v_i)}+e^{-e^{x_i^T\beta(u_i,v_i)}}}{1+e^{x_i^T\gamma(u_i,v_i)}}\right) + \prod_{\substack{l=1 \\ y_l>0}}^{n}\left(\frac{\frac{1}{1+e^{x_i^T\gamma(u_i,v_i)}}\left(exp\left(-e^{x_i^T\beta(u_i,v_i)}+y_l x_i^T\beta(u_i,v_i)\right)\right)}{y_l!}\right) \tag{14}$$

And:

$$\begin{aligned} &lnL\left(\gamma(u_i,v_i),\beta(u_i,v_i)\right) \\ &= \sum_{\substack{l=1 \\ y_l=0}}^{n} ln\left(e^{x_1^T\gamma(u_i,v_i)}+e^{-e^{x_1^T\beta(u_i,v_i)}}\right) w_{il}(u_i,v_i) \\ &- \sum_{l=1}^{n} ln\left(1+e^{x_1^T\gamma(u_i,v_i)}\right) w_{il}(u_i,v_i) \\ &+ \sum_{\substack{l=1 \\ y_l>0}}^{n}\left(-e^{x_l^T\beta(u_i,v_i)}+y_i x_l^T\beta(u_i,v_i)\right) w_{il}(u_i,v_i) \\ &- \sum_{\substack{l=1 \\ y_l>0}}^{n} ln(y_i!) w_{il}(u_i,v_i) \end{aligned} \tag{15}$$

where, w_{il} (u_i, v_i) is the weight for location i; i=1,…,n.

Likelihood function on (15) is named incomplete likelihood because the first term is not known whether y_i = 0 comes from zero state or poisson state. Because of that Y_i is redefined by using Z_i latent variable.

$$Z_i = \begin{cases} 1, \text{from zero state} \\ 0, \text{from poisson state} \end{cases}$$

In order to obtain the solution of the equation, we use iterative method, Expectation-Maximization (EM) algorithm. Before expectation step we destinate Z_i distribution, Binomial (1, π_i) and join distribution of Z_i and Y_i, then we get new ln likelihood function below:

$$\begin{aligned} &ln\,L\left(\gamma(u_i,v_i),\beta(u_i,v_i)\,|\,y,z\right) = \\ &\sum_{l=1}^{n}(1-z_l)\left(y_l x_l^T\beta(u_i,v_i) - e^{x_l^T\beta(u_i,v_i)} - ln(y_l!)\right) w_{il}(u_i,v_i) \\ &+\sum_{l=1}^{n}\left(z_l x_l^T\gamma(u_i,v_i) - ln\left(1+e^{x_l^T\gamma(u_i,v_i)}\right)\right) w_{il}(u_i,v_i) \end{aligned} \tag{16}$$

The ln likelihood function on (16) can be rewritten as

$$\begin{aligned} &lnL\left(\gamma(u_i,v_i),\beta(u_i,v_i)|y,z\right) = lnL\left(\beta(u_i,v_i)|y,z\right) \\ &+lnL\left(\gamma(u_i,v_i)y,z\right) - \sum_{i-1}^{n}(1-z_i) ln(y_i!) w_{il}(u_i,v_i) \end{aligned}$$

Where:

$$\begin{aligned} &lnL\left(\beta(u_i,v_i)\,|\,y,z\right) \\ &= \sum_{i=1}^{n}(1-z_i)\left(y_i x_i^T\beta(u_i,v_i) - e^{x_i^T\beta(u_i,v_i)}\right) w_{il}(u_i,v_i) \end{aligned} \tag{17}$$

$$\begin{aligned} &ln\,L\left(\gamma(u_i,v_i)|y,z\right) \\ &= \sum_{i=1}^{n}\left(z_i x_i^T\gamma(u_i,v_i) - ln\left(1+-e^{x_i^T\gamma(u_i,v_i)}\right)\right) w_{il}(u_i,v_i) \end{aligned} \tag{18}$$

The last term can be ignored because it is not contained β(u_i, v_i) and γ(u_i, v_i). The expectation step is started by destinating the expectation of Z_i variable that is:

$$\begin{aligned} &E\left(Z_i\,|\,y_i,\gamma(u_i,v_i)^{(m)},\beta(u_i,v_i)^{(m)}\right) = Z_i^{(m)} \\ &Z_i^{(m)} = \begin{cases} P\left(Z_i=1\,|\,y_i=0,\gamma(u_i,v_i)^{(m)},\beta(u_i,v_i)^{(m)}\right), y_i=0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad , y_i>0 \end{cases} \\ &= \begin{cases} \dfrac{1}{1+exp\left(-x_i^T\gamma(u_i,v_i)^{(m)} - e^{x_i^T}\beta(u_i,v_i)^{(m)}\right)}, y_i=0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad , y_i>0 \end{cases} \end{aligned} \tag{19}$$

Then we substitute $Z_i^{(m)}$ from expectation step to Z_i on ln likelihood function on Equation (16). The maximization step of β(u_i, v_i) is carried out by using Newton-Raphson method as below:

$$\hat{\beta}(u_i,v_i)_{(m+1)} = \hat{\beta}(u_i,v_i)_{(m)} - H_{(m)}^{-1}\left(\hat{\beta}(u_i,v_i)_{(m)}\right) g_{(m)}\left(\hat{\beta}(u_i,v_i)_{(m)}\right)$$

Where:

$$g_{(m)}^{T} = X^{T}S^{(m)}W(u_i, v_i)(y - \mu); S^{(m)} = \text{diag}\left(1 - Z_i^{(m)}\right)$$
$$H_{(m)} = -X^{T}S^{(m)}W(u_i, v_i)TX; T = \text{diag}\left(e^{X_i^{T}\beta(u_i, v_i)}\right)$$

So we obtain $\hat{\beta}(u_i, v_i)$ below:

$$\hat{\beta}(u_i, v_i)_{(m+1)} = \left(\sum_{l=1}^{n} X_l\left(1 - Z_l^{(m)}\right)w_{il}(u_i, v_i)\hat{\mu}_{l(m)}X_l^{T}\right)^{-1}$$
$$\sum_{l=1}^{n} X_l\left(1 - Z_L^{(m)}\right)w_{il}(u_i, v_i)\hat{\mu}_{l(m)} \quad (20)$$
$$\left(X_l^{T}\hat{\beta}(u_i, v_i)_{(m)} + \left(\frac{y_l - \hat{\mu}_{l(m)}}{\hat{\mu}_{l(m)}}\right)\right)$$

Then the maximization process of $\gamma(u_i, v_i)$ is carried out by identic step like maximization of $\beta(u_i, v_i)$, using Newton-Raphson method:

$$\hat{\gamma}(u_i, v_i)_{(m+1)} = \hat{\gamma}(u_i, v_i)_{(m)} - H_{(m)}^{-1}\left(\hat{\gamma}(u_i, v_i)_{(m)}\right)\mathbf{g}_{(m)}\left(\hat{\gamma}(u_i, v_i)_{(m)}\right)$$

Where:

$$g_{(m)}^{T} = X_*^{T}R^{(m)}W(u_i, v_i)(y_* - \pi_*); R^{(m)} = \text{diag}\left(1 - Z_i^{(m)}\right)$$
$$H_{(m)} = -X_*^{T}R^{(m)}W(u_i, v_i)Q_*X_*; Q_* = \text{diag}\left(\pi_i(1 - \pi_i)\right)$$
$$X_*^{T} = \left(1, X_1^{T}, X_2^{T}, ..., X_k^{T}\right)$$
$$y_*^{T} = \left(y_1, y_2, ..., y_n, y_{n+1}, ..., y_{n+n0}\right) \text{ where } y_{n+1}, ..., y_{n0} = 0$$
$$\pi_*^{T} = \left(\pi_1, \pi_2, ..., \pi_n, \pi_{n+1}, ..., \pi_{n+n0}\right)$$

Then we obtain $\hat{\gamma}(u_i, v_i)$:

$$\hat{\gamma}(u_i, v_i)_{(m+1)} = \left(\sum_{l=1}^{n+n_0} \mathbf{X}_{*l}^{T} r_{*l}^{(m)} w_{il}(u_i, v_i)\hat{\pi}_{l(m)}\left(1 - \hat{\pi}_{l(m)}\right)\mathbf{X}_{*l}\right)^{-1}$$
$$\sum_{l=1}^{n+n_0} \mathbf{x}_{*l}^{T} r_{*l}^{(m)} w_{il}(u_i, v_i)\hat{\pi}_{l(m)}\left(\mathbf{X}_{*l}\hat{\gamma}(u_i, v_i)_{(m)} + \left(\frac{y_l - \hat{\pi}_{l(m)}}{\hat{\pi}_{l(m)}}\right)\right) \quad (21)$$

Then $\beta(u_i, v_i)$ and $\gamma(u_i, v_i)$ are replaced by $\hat{\beta}(u_i, v_i)$ and $\hat{\gamma}(u_i, v_i)$ on the Equation (20) and (21). After that we repeat from expectation step. We continue doing those steps (Expectation-Maximization) until $\hat{\beta}(u_i, v_i)$ and $\hat{\gamma}(u_i, v_i)$ are convergen.

Hypotheses Testing of GWZIPR Model

We use *F* test to compare GWZIPR and ZIPR model, to know significance of geographical factors. The hypotheses are written below for $i = 1, 2, \ldots, n$:

H_0: $\beta_j(u_i, v_i) = \beta_j$ and $\gamma_j(u_i, v_i) = \gamma_j; j = 1, 2, \ldots, k$
H_1: Minimum there is one $\beta_j(u_i, v_i) \neq \beta_j$ or $\gamma_j(u_i, v_i) \neq \gamma_j$

F test for the hypothesis is given by:

$$F = \frac{G_1 / df_1}{G_2 / df_2} \quad (22)$$

Statistic *F* follows *F* distribution with degree of freedom (df_1, df_2). G_1 and G_2 are devian values of ZIP and GWZIP regression model with degree of freedom df_1 and df_2 respectively. We reject H_0 when $F > F_{(\alpha, df1, df2)}$.

Simultaneous parameter hypotheses testing of GWZIPR model is given below:

The parameters testing of $\beta(u_i, v_i)$ and $\gamma(u_i, v_i)$:

H_0: $\beta_1(u_i,v_i) = \ldots = \beta_k(u_i, v_i) = \gamma_1(u_i,v_i) = \ldots = \gamma_k(u_i,v_i) = 0$
H_1: Minimum there is one $\beta_j(u_i, v_i) \neq 0$ or $\gamma_j(u_i, v_i) \neq 0$
$i = 1, 2, \ldots, n; j = 1, 2, \ldots, k$

We use *G* test:

$$G = -2\left[lnL(\hat{\omega}) - lnL(\hat{\Omega})\right] \quad (23)$$

We reject H_0 when $G > \chi^2_{(\alpha, p-q)}$, where p is the number of parameters under population and q is the number of parameters under H_0 true.

The parameters testing of $\beta(u_i, v_i)$:

H_0: $\beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \ldots = \beta_k(u_i, v_i) = 0$
H_1: Minimum there is one $\beta_j(u_i, v_i) \neq 0; j = 1, 2, \ldots, k$

We use G test, Equation (23) with criterion we reject H_0 when $G > \chi^2_{(\alpha, k)}$.

We carry out the same step as simultaneous parameter hypothesis testing of $\beta(u_i, v_i)$ for parameters testing of $\gamma(u_i, v_i)$. In order to know which parameters are significant in each study area, we carry out the partial testing of the parameters. The hypotheses are:

H_0: $\beta_j(u_i, v_i) = 0$ and H_0: $\gamma_j(u_i, v_i) = 0$
H_1: $\beta_j(u_i, v_i) \neq 0$ H_1: $\gamma_j(u_i, v_i) \neq 0$

By using statistic test in Equation (23), we reject H_0 when $G > \chi^2_{(\alpha, 1)}$.

Application

Filariasis or *elephantiasis* is an infectious disease caused by filaria worm and transmited by mosquitos bites. There are three species of worm can cause *filariasis* disease, those are *Wuchereria bancroft*, *Brugiatimori* and *Brugiamalayi*. *Filariasis* can be infected by all of species of mosquito like *Anopheles*, *Aedes*, etc.

The infection of *filariasis* occurs when the sources of infection are available, those are human with *mikrofilaria,* vector (mosquito) and other vulnerable human to *filariasis* (Nasrin, 2008). Some factors which

can trigger the emergence of *filariasis* case are presence of filaria worm (*Brugia malayi, Brugia timori, Wuchereria bancrofti*) (Pratiknya, 2000), human (Mardesni, 2006), mosquito (DKRI, 2007), environment factor (Soedarto, 1990) including interior and exterior environment, behaviour factor (Juriastuti *et al*., 2010) and the knowledge about *filariasis* (Nasrin, 2008).

Preventive efforts of *filariasis* case are conducted by health counseling activities, physical construction of a healthy house (DKRI, 2006), spraying, using wire netting, mosquito nets, mosquito coils and profilaksis.

Moreover some handling efforts of *filariasis* are reporting to the local health department, protecting the patients from mosquito bites, finding the sources of infection, special treatment and controlling vector (mosquito) in the endemic area (Mardesni, 2006).

In this study, we fit ZIP and GWZIP regression, the method that we develop to filariasis counts data, by taking into account spatial factors. The data contains information about filariasis counts and factors which are thought likely to influence the *filariasis* case.

The response variable is the number of filariasis per regency in East Java.

The covariate matrix contains the variables assosiated with health (including the age). Those variables are written at Table 1.

Notice

Y: The case of *filariasis.* X_1: The percentage of households having healthy lifestyle behavior (Uloli *et al*., 2008)*,* X_2: The percentage of households having healthy outhouse (Rahayu, 2005), X_3: The percentage of households having healthy trash can (Rahayu, 2005), X_4: The percentage of households having healthy wastewater management (Soedarto, 1990). X_5: The percentage of the residents 20-39 years of age (Wulandari *et al*., 2010; Juriastuti *et al*., 2010), X_6: The percentage of healthy counseling activities (Nasrin, 2008; Juriastuti *et al*., 2010). u_i and v_i represent altitude and longitude

Table 1. Structure of data

Y	X_1	X_1	…	X_6	u_i	v_i
y_1	x_{11}	x_{21}	…	x_{61}	u_1	v_1
y_2	x_{12}	x_{22}	…	x_{62}	u_2	v_2
y_3	x_{13}	x_{23}	…	x_{63}	u_3	v_3
⋮	⋮	⋮	⋱	⋮	⋮	⋮
y_{38}	$x_{1;38}$	$x_{2;38}$	…	$x_{6;38}$	u_{38}	v_{38}

Table 2. Descriptive statistics for variables

Variable	Mean (%)	StdDev (%)	Minimum (%)	Maximum (%)
Y	0.579	0.976	0.00	4.00
X_1	43.720	14.790	8.50	65.74
X_2	77.290	16.040	25.32	97.46
X_3	60.220	25.620	0.00	88.62
X_4	58.970	26.100	0.00	100.00
X_5	0.305	0.026	0.26	0.37
X_6	1.217	1.070	0.00	4.05

Table 3. VIF value of predictor variables

Variable	X_1	X_2	X_3	X_4	X_5	X_6
VIF	1.063	1.497	2.488	2.455	1.322	1.267

Table 4. Parameter estimates of zip regression

Parameter	Estimates	StdError	Z	p-value
β_0	2.874	2.824	1.0180	0.3090
β_1	0.042	0.026	1.6310	0.1030
β_2	-0.011	0.013	-0.8320	0.4050
β_3	0.037	0.021	1.7000	0.0890
β_4	0.012	0.020	0.5980	0.5500
β_5	-17.025	10.805	-1.5760	0.1150
β_6	-1.917	0.744	-2.5770	0.0099*
γ_0	23.920	70.624	0.3039	0.7350
γ_1	4.181	5.502	0.7600	0.4470
γ_2	-0.967	1.430	-0.6770	0.4990
γ_3	0.890	1.166	0.7630	0.4460
γ_4	-0.892	1.321	-0.6750	0.4990
γ_5	-57.270	261.965	-0.2190	0.8270
γ_6	171.665	218.010	-0.7870	0.4310

Descriptive Analysis of Filariasis Counts Data

Based on descriptive analysis as shown in Table 2, elephantiasis is rare case at regencies in East Java, minimum zero case and maximum only 4 cases. There are more than 50% in average of households in study area have healthy life facilities (outhouse, trash can, wastewater management). There is a tiny percentage (average 0.305%) of residents 20-39 years of age, minimum 0.26% and maximum 0.37%. The healthy counseling activities are rarely conducted, only 1.217% in average, minimum 0% and maximum 4.05%. It means there are regencies which not conduct healthy counseling activities at all.

Identification of multicollinearity is presented in Table 3. From the table, all VIF values are less than 10. This means there is no multicollinearity between predictor variables.

ZIP Model to Filariasis Counts Data

We estimate the parameters of *filariasis* counts data using R program. The results are presented in Table 4.

Table 4 shows that by using significance level $\alpha = 0.05$, there is one variable which is significant to ln model (parameterβ) and there is not variable significant to logit model (parameterγ). So we can state that the ZIP model is appropriate to model *filariasis* counts data.

GWZIP Model to Filariasis Counts Data

Spatial factors on the GWZIPR model are identified by testing spatial effect, dependence and heterogeneity. We use Moran's I to test spatial dependence. By using the significance level ($\alpha = 0.1$), the result shows that there is not spatial dependence effect (*p-value* 0.7405 >α). The testing by using *Koenker-Basse test* indicates the existence of weak spatial heterogeneity, *p-value* 0.0933 < α. Then we involve spatial factor in the model by carrying out GWZIPR model.

In this study, the coordinate of altitude and longitude represent geographical factor of regencies/towns. We use euclid distance to measure the distance between regencies/towns. We get optimum bandwidth by choosing minimum CV. Then Euclid distance and optimum bandwidth are used into kernel function to obtain spatial weight matrix. We use Gauss Kernel function in this step. Table 5 is an example containing euclid distances and the weights on GWZIPR model of Pacitan Regency (Regency 1).

We use R program to estimate parameters of GWZIPR model using EM algorithm. The summary of parameter estimates of all regencies/towns is given in the Table 6.

Table 6 shows that parameter estimates $\hat{\beta}_5$ and $\hat{\gamma}_5$ have high standard error. They are different from the standard error of other parameters, as the effect, interval confidence of those parameters are very width. This is associated with the result of testing of the parameter of each regency/town.

The result of the hypothesis testing of equality between GWZIPR and ZIP regression model shows that there is no significant difference between two models, $F = 0.2935 < F_{(0.1;14,24)} = 1.7974$, we cannot reject H_0.

Then the result of the simultaneous parameter hypothesis testing of GWZIPR model shows that $G = 278.5349 > \chi^2_{(12)} = 18.549$, so we reject H_0. It means GWZIPR model is appropriate to model East Java *filariasis* counts data. Base on those results, we can use ZIP or GWZIPR to build the model of East Java *filariasis* counts data.

Table 5. Euclid distance and the weights of Pacitan regency

Regency/ town	Euclid distance	Weight	Regency/ town	Euclid distance	Weight
1	0.000	1.000	20	1.406	0.757
2	0.295	0.988	21	0.601	0.950
3	1.638	0.685	22	0.670	0.939
4	2.012	0.565	23	1.769	0.643
5	0.669	0.939	24	2.780	0.336
6	1.961	0.581	25	0.512	0.964
7	1.215	0.812	26	0.885	0.895
8	1.968	0.579	27	0.967	0.876
9	0.871	0.899	28	0.271	0.990
10	0.165	0.996	29	0.971	0.875
11	0.818	0.910	30	0.878	0.897
12	0.760	0.922	31	0.881	0.896
13	0.270	0.990	32	1.010	0.866
14	0.341	0.984	33	0.936	0.884
15	0.767	0.920	34	0.230	0.993
16	0.895	0.893	35	0.724	0.929
17	1.755	0.648	36	1.593	0.699
18	0.751	0.924	37	0.794	0.915
19	2.031	0.559	38	0.488	0.967

Table 6. Summary of parameter estimates of GWZIPR of all regencies/town

Parameter	Average	StdError	90% confident power	Interval upper
β_0	1.4016	11.28984	-17.16851	19.97177
β_1	0.0017	0.00039	0.00101	0.00229
β_2	-0.0089	0.00032	-0.00946	-0.00840
β_3	0.0168	0.00037	0.01616	0.01737
β_4	0.0037	0.00032	0.00321	0.00425
β_5	-5.6389	135.35929	-228.28516	217.00737
β_6	-0.2622	0.16117	-0.52726	0.00294
γ_0	7.6065	44.85090	-66.16667	81.37971
γ_1	0.0066	0.00156	0.00404	0.00917
γ_2	-0.0357	0.00128	-0.03784	-0.03362
γ_3	0.0671	0.00147	0.06465	0.06950
γ_4	0.0149	0.00127	0.01282	0.01699
γ_5	-22.5556	541.43716	-913.14065	868.02950
γ_6	-1.0487	0.64468	-2.10905	0.01175

Table 7. Parameter estimates of GWZIPR Model of Pacitan Regency

Parameter	Estimate	StdError	Z_{value}
β_0	1.329	9.70073	0.137
β_1	0.003	0.00034	8.926*
β_2	-0.007	0.00023	-29.973*
β_3	0.016	0.00035	46.317*
β_4	0.003	0.00033	9.225*
β_5	-5.701	111.78431	-0.051
β_6	-0.295	0.11119	-2.653*
γ_0	7.314	38.69841	0.189
γ_1	0.012	0.00134	8.926*
γ_2	-0.029	0.00097	-29.973*
γ_3	0.063	0.00136	46.317*
γ_4	0.011	0.00119	9.225*
γ_5	-22.804	447.13725	-0.051
γ_6	-1.180	0.44478	-2.653*

We carry out partial testing of the GWZIPR model to know which variables influence model significantly. The result shows that all variables except the percentage of the residents 20-39 years of age (X_5) are significant in all regencies/towns. This is in accordance with the summary Table 6 showing standard error of parameter estimates $\hat{\beta}_5$ and $\hat{\gamma}_5$ are large on all regencies/towns. An example of local parameter estimates of GWNBR model, Pacitan Regency is given in the Table 7.

By comparing $Z_{(0.05)}$ =1.64486 and $|Z_{value}|$ from Table 7 we can conclude that five variables (X_1, X_2, X_3, X_4 and X_6) are significant and the model for Pacitan regency can be written below:

$$ln(\hat{\mu}_1) = 1.329 + 0.003x_{11} - 0.007x_{21} + 0.016x_{31} + 0.003x_{41} - 5.701x_{51} - 0.295x_{61} \quad (24)$$

$$\hat{\mu}_1 = e^{(1.329+0.003x_{11}-0.007x_{21}+0.016x_{31}+0.003x_{41}-5.701x_{51}-0.295x_{61})}$$

$$logit(\hat{\pi}_1) = 7.314 + 0.012x_{11} - 0.029x_{21} + 0.063x_{31} + 0.011x_{41} - 22.804x_{51} - 1.180x_{61} \quad (25)$$

Equation (24) describes that if percentage of households having healthy lifestyle behavior (X_1) increase 1%, then it will increase the average *filariasis* counts 0.0045 and vice versa. By the same way we can interpret other variables. Unfortunately those are contrary to reality that the increasing of percentage of households having healthy lifestyle behavior (X_1), trash can (X_3) and wastewater management (X_4) should reduce *filariasis* case counts, not conversely. Those cases are probably caused by small observation (only 38 observations) and there is weak spatial heterogeneity of the data (significance level $\alpha = 0.1$).

Logit model (parameter γ on the Table 7) shows that probability of no *filariasis* case ($y_i = 0$) in each regency/town is influenced by X_1, X_2, X_3, X_4 and X_6 variables. Based on both models in Equation (24) and (25) and Table 7, predictor variables which influence *poisson state* and *zero state* are the same.

Concluding Remarks

In this study, we have introduced two regression models for counts data: Zero-Inflated Poisson Regression (ZIPR) and Geographically Weighted Zero-Inflated Poisson Regression (GWZIPR). Maximum likelihood techniques are used to estimate the parameters of both models. The EM algorithm is used to maximize the likelihood by using Newton Raphson method. Moran's I and *Koenker-Basset* test are used to know the presence of spatial dependence and heterogeneity. Euclid distance and Gauss Kernel Function is used to obtain spatial weight matrix. Comparing ZIPR and GWZIPR are tested by using F Test.

The tests have proved the spatial independence of the number of East Java *filariasis* cases 2012. However, there is weak spatial heterogeneity of the data (significance level $\alpha = 0.1$). In such data we have shown that ZIP and GWZIP regression models are not significantly different (based on F test). Nevertheless, the two models have different significant variables. The

significant variable of ZIP regression models is X_6. Then X_1, X_2, X_3, X_4 and X_6 variables are significant on GWZIP Regression Model. Which one should we use? In such situation, probably it is a good idea if we choose the model that can give logical interpretation. In this condition the model of ZIP Regression gives more logical meaning than GWZIP Regression.

Acknowledgement

We thank East Java Provincial Department for providing access to Elephantiasis (*Filariasis*) Data.

Author's Contributions

Purhadi: Participated in all experiments, designed the research plan and coordinated the research.

Yuliani Setia Dewi: Participated in the data-analysis interpretation of data and contributed to the writing of the manuscript.

Luthfatul Amaliana: Participated in organized the study, data processing, data-analysis and interpreted of data.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. 1st Edn., Springer Science and Business Media, Dordrecht, ISBN-10: 9024737354, pp: 284.

Bohning, D., E. Dietz, P. Schlattmann, L. Mendonca and U. Kirchner, 1999. The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. J. Royal Statistical A, 162: 195-209. DOI: 10.1111/1467-985X.00130

Cameron, A.C. and K.P. Trivedi, 2005. Micro econometrics: Methods and Applications. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521848059, pp: 1034.

DKRI, 2006. Profil kesehatan Indonesia. Departemen Kesehatan RI., Pusat Data dan Informasi, Depkes RI, Jakarta.

DKRI, 2007. Profil kesehatan Indonesia. Departemen Kesehatan RI, Pusat Data dan Informasi, Depkes RI, Jakarta.

Greene, W.H., 2003. Econometric Analysis. 5th Edn., Pearson Education India, Upper Saddle River, N.J., ISBN-10: 817758684X, pp: 1026.

Gujarati, D.N., 2004. Basic Econometrics. 4th Edn., Tata McGraw Hill, ISBN-10: 0070597936, pp: 1032.

Juriastuti, P., M. Kartika, I.M. Djaja and D. Susanna, 2010. Faktor risiko kejadian *filariasis* di kelurahan jati sampurna. J. Makara UI Kesehatan, 14: 31-36.

Khoshgoftaar, T.M., K. Gao and R.M. Szabo, 2004. Comparing software fault predictions of pure and zero-inflated poisson regression models. Int. J. Syst. Sci., 36: 705-715. DOI: 10.1080/00207720500159995

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. ASA Am. Society Quality Control Technometr., 34: DOI: 10.2307/1269547 1-14

Lestari, A., 2008. Pemodelan zero-inflated Poisson regression (Studi kasus: Data pekerja seks komersil di klinik putat jaya Surabaya). MSc. Tesis, Institut Teknologi Sepuluh Nopember, Surabaya.

Mardesni, F., 2006. Hubungan lingkungan rumah, perilaku dan pekerjaan dengan kejadian *filariasis* Di Kabupaten Muaro Jambi Tahun. Msc. Tesis, FKM, ilmu kesehatan masyarakat fakutas, Kesehatan Masyatakat Universitas Indonesia.

McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2th Edn., CRC Press, ISBN-10: 0412317605, pp: 532.

Mouatassim, Y. and E.H. Ezzahid, 2012. Poisson regression and zero-inflated Poisson regression: Application to private health insurance data. Eur. Actuary J., 2: 187-204. DOI: 10.1007/s13385-012-0056-2

Myers, R.H., D.C. Montgomery, G.G. Vining and T.J. Robinson, 1990. Generalized Linear Model with Applications in Engineering and Sciences. 2th Edn., John Wiley and Sons, Inc., New Jersey.

Nasrin, 2008. Faktor-faktor lingkungan dan perilaku yang berhubungan dengan kejadian *filariasis* di kabupaten Bangka Barat. MSc. Tesis, Jurusan Kesehatan Lingkungan, Universitas Diponegoro, Semarang.

Pratiknya, A.W., 2000. Dasar-Dasar Metodologi Penelitian Kedokteran dan Kesehatan. 1st Edn., Raja Grafindo Persada, Jakarta, ISBN-10: 9794213667, pp: 236.

Rahayu, A., 2005. Kejadian *filariasis* dan hubungannya dengan faktor risiko lingkungan fisik rumah di Kab. Subang th. 2005. MSc. Tesis, Fakultas Kesehatan Masyarakat, Universitas Indonesia, Depok.

Soedarto, 1990. Penyakit-Penyakit Infeksi di Indonesia. 1st Edn., Widya Medika, ISBN-10: 9795190083, pp: 183.

Uloli, R., S. Soeyoko and Sumarni, 2008. Analisis faktor-faktor risiko kejadian *filariasis*. Berita Kedokteran Masyarakat, 24: 44-50.

Wulandari, S.P., B.S.S. Ulama and I. Rahmawati, 2010. Pemodelan resiko penyakit kaki Gajah (*Filariasis*) di provinsi papua dengan regresi zero-inflated Poisson. Forum Statistika dan Komputasi, 15: 8-16.