Bayesian Zero-Inflated Generalized Poisson (τ) Spatio-Temporal Modeling for Analyzing the DHF Endemic Area

Mukhsar¹, Bahriddin Apabihi¹, Sitti Wirdhana Ahmad Bakkareng², Asrul Sani¹*, Edi Cahyono¹

¹Mathematics and Statistics Department, Mathematics and Natural Sciences Faculty, Halu Oleo University-Indonesia.
²Biology Department, Mathematics and Natural Sciences Faculty, Halu Oleo University-Indonesia

Abstract

Bayesian modeling for count data is to analyze an epidemiological risk as spatially and temporally varying, e.g. Dengue Hemorrhagic Fever (DHF). The DHF data with too many zero are commonly an over dispersion. Bayesian zero-inflated generalized Poisson or BZIGP (τ) spatio-temporal. This model was developed from our previous work, called BZIP S-T. Both models is verified by using the DHF monthly data in 10 districts of Kendari city during periode 2013-2015. MCMC method is used to estimate the parameters of both models. Both models indicate the rainfall and population density are statistically significant to influence the fluctuations of DHF cases. But the BZIGP (τ) S-T is smallest deviance and the best performance model. Puwatu and Kadia districts are an endemic area.

Keywords: Bayesian spatio-temporal; DHF; Generalized Poisson; over dispersion; zero inflated

*Corresponding Author.
E-mail addresses: mukhsar@muhu@yahoo.com, mukhsar@ubo.ac.id (Mukhsar), bahridinabapihi@yahoo.com (Abapihi, B), wirdhanaaxtalora@yahoo.com (Wirdhana A. B), sanitasru@2001@yahoo.com (San A), edi_cahyono@innov-center.org (Cahyono E)
1. INTRODUCTION

DHF cases are often threatened in densely populated area of Indonesia. Some factors influence it, e.g. heterogeneity spatial as well as temporal and uncertainty factors [1]. Modeling to analyze DHF risk is adapting some factors to determine endemic locations [2]. Various studies, e.g. [3, 4, 5, 7] stated the uncertainty factor is represented as a random effect.

The DHF count data at certain times and locations are mostly zero-inflation, ignoring it will lose information [8, 9]. Several authors have analyzed such data sets by the use of zero-inflated Poisson (ZIP), see [10, 11, 12, 13, 14, 15]. The phenomenon of zero-inflated can arise as a result of clustering, distributions with clustering interpretations exhibit the feature that the proportion of observations in the zero class is greater than the estimate of the probability of zeros given by the assumed model. This model is then developed by [17] to analyze the risk of DHF monthly data in 10 districts of Kendari using Bayesian approach, called Bayesian ZIP spatio-temporal (BZIP S-T). The generalized linear model (GLM) is also integrated into BZIP S-T. Markov Chain Monte Carlo (MCMC) is used to estimate parameters of the model based on full conditional distribution (FCD).

Poisson regression was used to analyze the relationship between the response variable and one or more predictor variables, where the expected value (mean) and variance assumed to be equal (equi-dispersion). However in the discrete data analysis, sometimes the variance is greater than the mean (over-dispersion) or the variance is smaller than the mean (under-dispersion). A number of cases are often observed an over-dispersion, especially if the count data is too many zero. Several studies, e.g. [18, 19, 20, 21] are using generalized Poisson (GP) to overcome these cases. They cited examples of the data with too many zeros from various disciplines, including agriculture, econometric, patent applications, species abundance, medicine, and use of recreational facilities.

This article in developing the BZIP S-T model into a Bayesian zero-inflated generalized Poisson spatio-temporal (BZIGP S-T). The predictors used are the same, then the BZIGP S-T is expressed as BZIGP (τ) S-T. In next step is discussing description of DHF monthly data, for example, spatial and temporal detection, over-dispersion test. Furthermore, it is reviewing the BZIP S-T as the basis of BZIGP (τ) S-T developing which integrates three main components, the GP model, and FCD as well as MCMC computational techniques.

2. DESCRIPTION OF DHF DATA

Kendari city is the capital of Southeast Sulawesi province of Indonesia. It is located geographically in the south of the equator and stretches from west to east. Kendari is
one of the cities in Indonesia as tropical country with high DHF cases. The population density of Kendari at around 1,094 people per square kilometer (km$^2$). It is situated around 3m-30m above sea level with the temperature at 23°C-32°C and the humidity at 81%-85% for the whole year. The wet season usually starts in January and ends in June. The higher rainfall (200mm-300mm) occurs during January-April and the less rainfall is around October-November (below 100 mm). Data reviews were obtained from the Meteorological, Climatological and Geophysics Agency (BMKG) and the Central Bureau of Statistics (BPS) of Kendari.

The DHF monthly data in 10 districts of Kendari, period 2013 to 2015 are showing the majority as 90% Poisson distribution and 10% binomial distribution with p-value above 5% (see Table 1). The binomial distribution is approached by the Poisson process for large number of population. It is also outlines adjacency matrix between districts. Queen Principle is used to arrange weighting matrix into a spatial contiguity.

<table>
<thead>
<tr>
<th>Code</th>
<th>Districts</th>
<th>Adjacency matrix</th>
<th>Kolmogorov-Smirnov</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mandonga</td>
<td>3,4,8,10</td>
<td>0,6</td>
<td>Poisson</td>
</tr>
<tr>
<td>2</td>
<td>Baruga</td>
<td>3,5,6,8</td>
<td>0,1</td>
<td>Poisson</td>
</tr>
<tr>
<td>3</td>
<td>Puwatu</td>
<td>1,2,4,5</td>
<td>0,2</td>
<td>Poisson</td>
</tr>
<tr>
<td>4</td>
<td>Kadia</td>
<td>1,3,5,8</td>
<td>0,1</td>
<td>Poisson</td>
</tr>
<tr>
<td>5</td>
<td>Wua-Wua</td>
<td>2,3,4,8</td>
<td>0,1</td>
<td>Poisson</td>
</tr>
<tr>
<td>6</td>
<td>Poasia</td>
<td>2,7,8</td>
<td>0,2</td>
<td>Poisson</td>
</tr>
<tr>
<td>7</td>
<td>Abeli</td>
<td>6</td>
<td>0,3</td>
<td>Binomial</td>
</tr>
<tr>
<td>8</td>
<td>Kambu</td>
<td>1,2,4,5,6</td>
<td>0,1</td>
<td>Poisson</td>
</tr>
<tr>
<td>9</td>
<td>Kendari Barat</td>
<td>1,9</td>
<td>0,1</td>
<td>Poisson</td>
</tr>
</tbody>
</table>

The spread of DHF is affected by other locations nearby. If a location is becoming DHF endemic, then the other locations are closed to it are immediately to be high risk. An aedes Aegypti population is fluctuating based on dynamical system, for example, dependency between location. Detection is required to determine the spatial dependency of DHF incident. Moran index ($\rho$) is a technique to detect spatial dependency with range -1 and 1. In January, February, March, April, and December, show positive Moran index. This means that DHF cases in adjacent locations have similar patterns. May to November, it is no founding the Moran index, because there is not DHF case.
An autocorrelation function (ACF) is a tool to detect temporal dependencies of DHF data. There are four patterns of time series data, i.e. horizontal, trend, seasonal, and cyclical [5]. The horizontal is unpredictable as well as random and tendency to go up and down. The seasonal is DHF case fluctuations occurred periodically at a certain time (quarter, quarterly, monthly, weekly, or daily) and the cyclical is DHF case fluctuations occurred in a long time. The DHF data have temporal dependencies if ACF initial value exceeds the boundary line, and then decreases gradually. The ACF showing the DHF monthly data in 10 districts of Kendari for period 2013-2015 are an initial value exceeds the boundary line on the lag-1 then decreases gradually. This means that DHF cases of Kendari is temporal dependencies.

Over-dispersed is using chi-Square, \( \theta = \frac{\chi^2}{n-k} \), where \( n \) is number of samples and \( k \) is number of parameters. The DHF monthly data in 10 districts of Kendari is over-dispersion, because \( \theta > 1 \).

3. GENERALIZED POISSON AND ZERO-INFLATED POISSON DISTRIBUTION

Poisson regression is starting point for modeling of count data and flexible to be parameterized in the form of distribution function [22, 23, 24, 25]. Variable response has independent identically distributed (i.i.d). Suppose \( y_s, s = 1, ..., S \) is count data, where \( S \) is number of locations and predictor is \( x_s \). Then the density function of \( y_s \) is expressed

\[
Poisson(\lambda_s; y_s) = \frac{e^{-\lambda_s} \lambda_s^{y_s}}{y_s!}, \quad y_s = 0, 1, 2, ..., s = 1, ..., S.
\]

In order to adjust a to many zeros, then (1) is modified into the ZIP model. This technique was first described by [8]. The ZIP model has been discussed in various fields, e.g. epidemiology [10], health [12, 25]. Observation data in the ZIP model was divided into two process [8]. The first process, selected with probability \( \phi_s \) was generated from zero count, and the second process was selected with probability \( 1 - \phi_s \) from the Poisson process. The structure of ZIP model as i.i.d count data is written

\[
P(Y = y_s) = \begin{cases} 
\phi_s + (1-\phi_s)Poisson(\lambda_s; 0), & 0 \leq \phi_s \leq 1, y_s = 0 \\
(1-\phi_s)Poisson(\lambda_s; y_s), & y_s > 0
\end{cases}
\]  \hspace{1cm} (2)

Several researchers were used the Bayesian approach in the ZIP model, called BZIP, e.g. [11, 26, 27]. They were used MCMC method to estimate the parameters of the BZIP that generated via its FCD respectively.

The BZIP is then developed by [17] into BZIP S-T to analyze the relative risk of DHF cases in 10 districts of Kendari, Indonesia. The BZIP S-T integrates three main
Bayesian Zero-Inflated Generalized Poisson (τ) Spatio-Temporal Modeling... 1073

components, namely spatial heterogeneity, two random effects (local and global), and temporal trend. Local random effect is representing the relationship of uncertainty in location, while the global random effect is representing the uncertainty relationship between locations. Temporal trend is representing the temporal occurrence of dengue cases have the same intercept but vary temporally at each location. However, the BZIP S-T ignores over-dispersed data [17]. A fairly popular and well-studied alternative to the standard Poisson distribution is the GP distribution [20, 28, 29, 30]. They provided a guide to the current state of modeling with the GP at time, and documented many real life examples. The GP regression model is developed from (1), namely

\[
\text{GPoisson}(\lambda_j, \theta; y_j) = \left( \frac{\lambda_j}{1 + \theta \lambda_j} \right)^{y_j} \left( 1 + \theta \lambda_j \right)^{-1} \exp \left[ -\frac{\lambda_j}{1 + \theta \lambda_j} \right].
\]  

(3)

where \( \lambda_j = \lambda_j(x_j) = \exp(\sum x_j \beta_j) \) and \( \theta \) is dispersion parameter. If \( \theta = 0 \) then the model (3) is equi dispersion case, \( \theta > 0 \) is over dispersion, and \( \theta < 0 \) is under dispersion.

4. ZERO-INFLATED GENERALIZED POISSON AND ITS EXTENSION

The over-dispersed data can be overcome with zero-inflated generalized Poisson (ZIGP). Structure of the ZIGP model is organized from (2) and (3),

\[
P(Y = y_j | x_j, z_j) = \begin{cases} 
\phi_j + (1 - \phi_j)\text{GPoisson}(\lambda_j, \theta; y_j), & y_j = 0 \\
(1 - \phi_j)\text{GPoisson}(\lambda_j, \theta; y_j) & y_j > 0
\end{cases}
\]

(4)

where \( 0 < \phi_j = f(Y_j = 0) < 1 \), \( \log(\lambda_j) = \lambda_j(x_j) = \sum x_j \beta_j \), and \( \phi_j = \phi(z_j) \), \( \logit(\phi_j) = \log(\phi_j[1 - \phi_j])^{-1} = \sum z_j \rho_j \).

For the same predictor, the (4) is called ZIGP (τ) is written into

\[
\log(\lambda_j) = \sum x_j \beta_j \quad \text{and} \quad \logit(\phi_j) = \log(\phi_j[1 - \phi_j])^{-1} = \tau \sum x_j \beta_j
\]

(5)

Model (5) is developed into spatial and temporal varying. This expansion is using Bayesian modeling technique, called Bayesian ZIGP (τ) spatio-temporal (BZIGP (τ) S-T). Assumed that the DHF case as count data, \( y_{st} \) is distributed by i.i.d Poisson distribution with parameter \( \lambda_{st} \) in district \( s^{th} \) at time \( t^{th} \). The structure of the BZIGP(τ) S-T is written
\[
\log(\lambda_s) = \log(\phi_s (1 - \phi_s))^{-1} = \sum x_{ps} \beta_p = \log(P(Ir)_s) + \beta_0 + \sum_{p=1}^P x_{ps} \beta_p + u_s + v_s + (\alpha + \delta_s) \tau_t .
\]

\[
v_s | v_s, \tau_t \sim N \left( \frac{\sum_{s=1}^S v_s}{D}, \frac{1}{h_s D} \right), \delta_s | \delta_s, \tau_t \sim N \left( \frac{\sum_{s=1}^S \delta_s}{D}, \frac{1}{h_s D} \right)
\]

where \( s = 1, 2, ..., S \) is the number of districts, \( t = 1, 2, ..., T \) is observation time, \( P \) is the number of predictors, \( P(Ir)_s \) is probability incident risk in district \( s \)th at time \( t \)th, \( x_{ps} \) is \( p \)th predictor in district \( s \)th at time \( t \)th, \( u_s \) is local random effect at district \( s \)th at time \( t \)th, \( v_s \) is trend at district \( s \)th at time \( t \)th, \( \delta_s \) is flat temporal. The \( h_s \) is precision for \( u_s \), \( \delta_s \) is precision for \( v_s \), \( \rho \) is spatial dependence parameter with \(-1 \leq \rho \leq 1\), \( D \) is total neighbors across location, \( \varepsilon(s) \) is number of neighbors of location \( s \). The \((\alpha + \delta_s) \tau_t \) is meaning the every location has the same intercept \((\alpha)\) and each location has a different contribution of DHF case. Assumed \( \beta_0 \) is flat distribution, \( \beta_p \) in normal distribution with zero mean, and \( h_\beta \) is precision parameter of \( \beta_p \) [31, 32; 33, 34, 35].

5. NUMERICAL SIMULATION AND PERFORMANCE MODEL

Likelihood for (4) is expressed

\[
L_s = -\sum \log(1 + \lambda_s^{-1}) + \sum \log \left( \frac{\lambda_s^{-1} + \exp \left( -\frac{\lambda_s}{1 + \theta_s} \right)}{1 + \theta_s} \right) + \sum \left[ y_s \log \left( \frac{\lambda_s}{1 + \theta_s} \right) + (y_s - 1) \log(1 + \theta_s) - \log(y_s) - \frac{\lambda_s (1 + \theta_s)}{1 + \theta_s} \right] .
\]

Joint prior is then defined as

\[
J_{prior} = J(\beta_0) J(\beta_p) J(u_s) J(v_s) J(\alpha) J(\delta_s) J(\beta_p | h_\beta) J(u_s | h_u) J(v_s | h_v) J(\alpha | h_u) J(\delta_s | h_v) .
\]

and \( J(\beta_p), J(u_s), J(v_s), J(\alpha), \) and \( J(\delta_s) \) are normal distribution, but \( J(\beta_0) \) is flat. For \( J(\beta_p | h_\beta), J(u_s | h_u), J(v_s | h_v), J(\alpha | h_u), \) and \( J(\delta_s | h_v) \) are gamma distribution and hyperprior of \( J(\beta_p), J(u_s), J(v_s), J(\alpha), \) and \( J(\delta_s) \) respectively.

Joint posterior is multiplication of (7) and (8),

\[
J_{post} = L_s \times J(\beta_0) J(\beta_p) J(u_s) J(v_s) J(\alpha) J(\delta_s) J(\beta_p | h_\beta) J(u_s | h_u) J(v_s | h_v) J(\alpha | h_u) J(\delta_s | h_v) .
\]

The FCD for BZIGP(\( \tau \)) S-T is obtained form (9). For example, defined FCD for \( \beta_v \), named \([\beta_v]_s\), is

\[
[\beta_v]_s = L_s \times J(\beta_v) = L_s \times \text{flat}().
\]
Bayesian Zero-Inflated Generalized Poisson (τ) Spatio-Temporal Modeling...

Assuming the flat() = constant and \( \lambda_s \) is from (6) then

\[ \lambda_s = P(Ir) \exp\left( \beta_0 + \sum_{m} x_{sw} \beta_p + u_s + v + (\alpha + \delta_s) \right), \]

or \( \lambda_s = P(Ir) \exp(\beta_0) = \text{constant}, \)

is be

\[ [\beta_0] = L_s \times J(\beta_0) = \text{constant} \times \text{constant} = \text{constant}. \]  \hspace{2cm} (11)

The FCD for \( \beta_p \), named \( [\beta_p] \), is

\[ [\beta_p] = L_s \times J(\beta_p) = L_s \times \left( \beta_p | h_\beta \sim N(0, h_\beta) \right), \] \hspace{2cm} (12)

where \( \beta_p | h_\beta \sim N(0, h_\beta) \) is normal distribution with zero mean.

For the others FCD are similar to (10) and (12). The FCD is used to estimate the parameters of BZIGP (τ) S-T in open source software WinBUGS 1.4. To estimate the parameters of BZIGP(τ) S-T is allowing an Algorithm 1, whereas to estimate the parameters of BZIP S-T model have been described by [17]. DHF monthly data for period 2013-2015 in 10 districts of Kendari are response variable. Rainfall and and population density are predictors. The parameters of BZIGP(τ) S-T are achieved convergence at 10,000 iteration bur-in 10,000. There are evidenced from an ergodic mean plot that it is stable in confidence interval (Fig. 1a), density plot (Fig.1b), and autocorrelation plot (Fig.1c). For the parameters of BZIP S-T are achieved convergence at 10,000 iteration bur-in 20,000. Posterior summary of both models is described in Table 2. The rainfall and population density are statistically significant to increase of DHF cases in Kendari, where the 95% confidence interval do not contain the zero value.

Algorithm 1. MCMC Gibbs sampler for estimating the parameters of BZIGP(τ) S-T, generating for m time iterations

Start

Step 1. Define

- constant \( C = 0 \)
- zeros = 0
- Set initial number of parameters
  \[ \beta_0^{(0)}, \beta_p^{(0)}, \alpha^{(0)}, u_s^{(0)}, v^{(0)}, \delta_s^{(0)}, \tau_\beta^{(0)}, \tau_u^{(0)}, \tau_v^{(0)}, \tau_\alpha^{(0)} \]
  - zeros \( \sim \text{dpois(zeros.mean)} \)
Step 2. Define

- log-likelihood for $s$ and $t$ individual
- log-link + linear predictor

Step 3. Generate parameter for $m$ time iteration

- $\beta_0^{(m)}$ sampling from $[\beta_0]$,
- $\beta_p^{(m)}$ sampling from $[\beta_p], p = 1, ..., P$,
- $u_{st}^{(m)}$ sampling from $[u_{st}], t = 1, ..., T, s = 1, ..., S$,
- $v_{st}^{(m)}$ sampling from $[v_{st}], t = 1, ..., T, s = 1, ..., S$,
- $\alpha^{(m)}$ sampling from $[\alpha]$,
- $\delta_i^{(m)}$ sampling from $[\delta_i], s = 1, ..., S$,
- $\tau_{\beta}^{(m)}$ sampling from $[\tau_{\beta}]$,
- $\tau_{\sigma}^{(m)}$ sampling from $[\tau_{\sigma}]$,
- $\tau_{\tau}^{(m)}$ sampling from $[\tau_{\tau}]$,
- $\tau_{\delta}^{(m)}$ sampling from $[\tau_{\delta}]$,

Step 4. If the estimation of parameter is not convergent, then the next step is discard as burn-in

Note. This process is repeated to obtain convergent condition

End
Bayesian Zero-Inflated Generalized Poisson (τ) Spatio-Temporal Modeling...

Figure 1: Plot of parameter estimation using WinBUGS 1.4 of BZIGP(τ) S-T, convergence at 10,000 iterations burn-in 10,000, (a) ergodic mean, (b) density, and (c) autocorrelation.

The BZIGP(τ) S-T is smaller deviance, 653.9 if it compared to BZIP S-T, 1550, (see Table 2). The smallest deviance is the best performance model (Congdon, 2010). The BZIGP(τ) S-T model results show the Puwatu and Kadia District consistent as highest DHF case (black color) in Kendari (see Fig.2). Both districts are endemic dengue location for intervention to prevent the spread of dengue to other locations.

Table 2. Posterior summary, parameter estimation of BZIP S-T 10,000 iterations burn-in 20,000 and BZIGP(τ) S-T 10,000 iterations burn-in 10,000

<table>
<thead>
<tr>
<th>Node</th>
<th>Mean</th>
<th>SD</th>
<th>MC error</th>
<th>2.50%</th>
<th>Median</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>BZIP S-T (see [17])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta0</td>
<td>0.28</td>
<td>0.091</td>
<td>0.003</td>
<td>0.132</td>
<td>0.271</td>
<td>0.49</td>
</tr>
<tr>
<td>beta1</td>
<td>0.25</td>
<td>0.072</td>
<td>0.0005</td>
<td>0.126</td>
<td>0.242</td>
<td>0.42</td>
</tr>
<tr>
<td>beta2</td>
<td>0.93</td>
<td>0.523</td>
<td>0.006</td>
<td>0.281</td>
<td>0.78</td>
<td>2.28</td>
</tr>
<tr>
<td>deviance</td>
<td>1550</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BZIGP(τ) S-T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta0</td>
<td>1.4</td>
<td>0.3</td>
<td>0.02</td>
<td>0.7</td>
<td>1.4</td>
<td>1.9</td>
</tr>
<tr>
<td>beta1</td>
<td>0.06</td>
<td>0.03</td>
<td>0.001</td>
<td>0.1</td>
<td>0.06</td>
<td>0.008</td>
</tr>
<tr>
<td>beta2</td>
<td>0.0012</td>
<td>0.6</td>
<td>0.002</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>deviance</td>
<td>653.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. CONCLUSION AND FUTURE WORK

This study develop the BZIGP(τ) S-T model based on our previous work, called BZIP S-T. We are using DHF monthly data in 10 districts of Kendari, for period 2013-2015 as response. The rainfall and population density are predictors. The MCMC Gibbs sampler is computational techniques to estimate the parameter of both models through its FCD respectively. Parameters estimation of BZIGP(τ) S-T attained convergence at 10,000 iterations and 10,000 burn-in. The BZIP S-T attained convergence at 10,000 iterations and 20,000 burn-in. Both models show the rainfall and population density are statistically significant influencing of DHF case in Kendari city. The BZIGP(τ) S-T is the better model because it smallest deviance, 653.9. The Puwatu and Kadia district are consistent as highest DHF location in Kendari city. Both locations are an endemic DHF. Future study will apply this model using DHF daily data because the dengue cases are very quickly fluctuate.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support of Kementerian Riset, Teknologi, dan Pendidikan Tinggi (KEMENRISTEK DIKTI) of Indonesia. We thank also to the Department of Health, BPS, and BMKG of Kendari for their permission to use their observation data in this study.
Bayesian Zero-Inflated Generalized Poisson (τ) Spatio-Temporal Modeling...

REFERENCES


Bayesian Zero-Inflated Generalized Poisson (τ) Spatio-Temporal Modeling...


