Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:20

https://doi.org/10.28919/cmbn/8428

ISSN: 2052-2541

BAYESIAN SPATIAL HIERARCHICAL MIXTURE MODELS FOR EXCESS ZEROS DATA: REVIEW AND APPLICATION TO FEMALE LYMPHATIC FILARIASIS CASES

RO'FAH NUR RACHMAWATI[1,*], JULI YANDI RAHMAN[1], NOVI HIDAYAT PUSPONEGORO[2]

[1]Department of Mathematics, Indonesia Defense University, Bogor 16810, Indonesia

[2]Politeknik Statistika, STIS, Jakarta 13330, Indonesia

Abstract: Many cases of epidemiological data reported has an excessive value of zero. The number of excess zeros can be more than half, even up to 80% of all existing data. This can occur in cases of rare diseases that do not cause significant symptoms at the start of infection. Therefore, the number and the rise of reported cases becomes difficult to detect. This paper proposes several Bayesian methods which are mixture of several distributions, namely binomial, Poisson and zero-inflated Poisson, and discuss the extension of these models to spatial data with excess zeros. Spatial data is implemented with Bayesian hierarchical framework, using Besag-York-Mollié re-parameterization (BYM2) model for spatial random effects, and penalized complexity prior for latent level process in the mixture models of different types of zeros. Bayesian inference uses INLA (Integrated Nested Laplace Approximation) for more accurate and faster results for spatially based hierarchical data. We further review recent implementation of proposed Bayesian mixture models using female lymphatic filariasis cases in 2019 at 27 district city level in West Java, Indonesia, and its elevation as explanatory variable. Mixture models were compared using DIC, and the results obtained indicate that

*Corresponding author

E-mail address: rofah.nr@idu.ac.id

Received January 04, 2024

mixture distributions between Binomial-Poisson and Binomial-zero-inflated Poisson type 1 produce suitable models for characteristic of excess zeros data around 67% with high extreme observation values in certain regions.

Keywords: joint model; zero-inflated Poisson models; lymphatic filariasis disease mapping; zero-inflated data; mixture distribution.

2020 AMS Subject Classification: 62H11, 62J02.

1. INTRODUCTION

Many observed events contain a large number of zero values (excess zeros). This incident does not only occur in the field of epidemiology, but also occur in other fields such as environmental studies. Observing the number of rain events in several areas in a particular season can produce excess zeros, which allows to occur throughout the season or even the year (e.g. see [1], [2]). More cases of excess zeros are found in discrete/count data in epidemiology, including cases of rare disease. Rare diseases often do not cause specific symptoms so the number of reported cases is often undetectable and ends up with the number of cases being zero or unreported. A minimal number of cases, even zero, can reduce serious attention to rare but contagious, this is a serious event that must be avoided. Having large proportion of zeros is inevitable and cannot be omitted from the analysis. Because zeros in the data structure have an implied meaning which also contains important information. Therefore, it is very necessary to have a model that can properly handle the case of excess zeros. This is because standard single distribution models such as the binomial, Poisson, negative binomial and even zero-inflated Poisson (ZIP) cannot fit proficiently for data with excess zeros problem [3].

Spatial or spatio-temporal count data are often based on single Poisson distribution. However, the problem that is often encountered in Poisson distribution modeling is overdispersion which can result in invalid model conclusions. One way to take care overdispersion problem is with a ZIP model. ZIP can be used to overcome overdispersion which comes from excess zeros data. ZIP is a model that combines zero values with a certain probability and non-zero events with a Poisson probability distribution [4]. Hence, Feng [5] compared and highlighted differences of zero-inflated

and hurdle models for modeling zero-inflated count data and did the simulation process. From the simulation result, hurdle and zero-inflated models perform almost equivalently in the overall model fit when there are no or few zero deflations. Then [6] used zero-inflated models such as Poisson, negative binomial, ZIP, zero-inflated negative binomial and hurdle regression to analyse number of antennal care service visits during pregnancy months in Ethiopia. The model's results hurdle model was better fitted excess zeros data than any other ZIP models.

Spatial and spatio-temporal modeling for zero-inflated data makes modeling increasingly complex [7] [8]. This is because the number of observations in each regional unit is increasing and also involves spatial (and/or temporal) influences between regions. Thus, in many recent publications on spatial and spatio-temporal modeling, models for excess zeros are carried out using a single distribution method, such as [9], [10], [11]. Single distribution methods such as zero inflated models are widely used and are known considered to fit well. However, on the other hand, if the response comes from two different distributions (occurrence and number of occurrence), and each can be modelled with the same factors (namely spatial influences), then the more appropriate model to use is a mixture model [4]. Mixture models are one way to make the model fit better, control various sources of uncertainty from each process, and certainly to obtain more accurate results.

Mixture models, which are a combination of several distributions, can fit excellently on data with excess zeros rather than single distribution methods [12]. Recently, there are only a few publications that use mixture models in Bayesian spatial modeling for excess zeros data. Several mixture model publications are Lyme disease mapping in the Eastern United States [3] and male breast cancer in Ethiopia [4]. Lyme disease spatial and spatio-temporal data were implemented using INLA inferences with covariance Matérn model to describe spatial structure in each spatial coordinate. Mixture of two distributions produce logistic regression for probability of occurrence and log-linear regression for the number of occurrences that fit excess zeros proficiently. Bayesian spatial joint model also implemented to investigate rare event disease mapping such as male breast cancer in Iran [4]. Male breast cancer data are mapped using INLA inferences of mixture binomial

and zero-inflated Poisson models with BYM2 in order to define the spatial random effects in the level process. In Bayesian spatial mixture model, response from each distribution is arranged hierarchically. However, inference from the posterior distribution using the classical MCMC method can cause problems [13], especially convergence and time issues. Therefore, posterior inference from the complex spatial model is performed using INLA approach [14] [15]. In modeling with excess zero data, the binomial-Poisson mixture model is rarely used compared to ZIP model. However, in fact, the combination of these distributions is the basis of distribution for occurrence and number of occurrences, so its use can really be considered to handle cases of excess zeros data.

This paper proposes mixture models and assumes responses of zero-inflated data come from two distributions: Binomial-Poisson and Binomial-ZIP with different types of zero. Each regression model for response corresponds to its link function (logit for binomial and log-linear for Poisson) which includes BYM2 as spatial random effects [16]. Parameters in mixture distribution are modelled using penalized complexity priors [17] and non-informative prior distribution [13]. As an application, excess zeros data are derived from female lymphatic filariasis cases in West Java and compare mixture models according to the deviance information criteria (DIC). The remainder of this paper is structured as follows: Section 2 provides a review of mixture models for data with excess zeros in Bayesian spatial framework. Section 3 discusses mixture model implementation in spatial count data at district city level of female lymphatic filariasis cases. The implementation divided in two cases according with and without elevation as explanatory variable. Finally, section 4 provides some conclusions, remarks and possible model developments for future publications.

2. HIERARCHICAL SPATIAL BAYESIAN MIXTURE MODELS

This section proposes and reviews some mixture models that can be implemented with excess zero response. We propose Binomial-Poisson, Binomial-ZIP mixtures with different types of zero. In zero-inflated data modelling, there are two types of zeros that can occur, i.e. the structural/true

zeros and sampling zeros. Structural zeros are conditions where true zeros/absent from an event actually occurs, while sampling zeros come from number of occurrences in area is reported zero based on change or mistake [4]. Although these three models' approaches are similar, there is a fundamental difference between three models which lies in the probability form of Poisson distribution.

2.1 Mixture Models

Under Binomial-Poisson model, it is assumed to be two-stage process that generate zero and non-zero data. Binomial-Poisson mixture model for the set of n independent and identically distributed observations of region i, $Y_i, i = 1,2,3 \dots, n$ can be described as the mixture of a point mass at zero with Binomial distribution with $n = 1$ and probability of successive zero is p, and a Poisson distribution for number of occurrences without concerning any type of zero:

$$P(Y_i = 0) = p_i, 0 \leq p_i \leq 1 \tag{1}$$

$$P(Y_i = k) = \frac{(E_i\theta_i)^k \exp(-E_i\theta_i)}{k!}, k = 0,1,2, \dots, n, \lambda > 0. \tag{2}$$

If there is only structural zero may occur in data, then mixture model used is Binomial-ZIP type0 which assume to follow probability as:

$$P(Y_i = 0) = p_i, 0 \leq p_i \leq 1 \tag{3}$$

$$P(Y_i = k) = (1 - p_i)\frac{(E_i\theta_i)^k \exp(-E_i\theta_i)}{k!\,(1 - e^{-E_i\theta_i})}, k = 1,2, \dots, n, \lambda > 0, \tag{4}$$

where Y_i denotes response of region i with $\lambda = E_i\theta_i$ as the mean of truncated Poisson distribution. If there are assumed that two types of zero may occur through processes, either by structural or sampling zero, then mixture model used is Binomial-ZIP type1 which assumed to follow probability as:

$$P(Y_i = 0) = p_i, 0 \leq p_i \leq 1 \tag{5}$$

$$P(Y_i = k) = (1 - p_i)\frac{(E_i\theta_i)^k \exp(-E_i\theta_i)}{k!}, k = 0,1, \dots, n, \lambda > 0. \tag{6}$$

This probability function is a mixture of proportion of structural zeros (p_i) and sampling zeros $(1 - p_i)$.

The interest now is in modeling the latent fields p_i and θ_i using canonical link function. This

paper uses a Bayesian framework in spatial modeling for excess zero data, so there are three hierarchies that show the levels of modeling. The three levels modelling process can be written as:

- Data Level

  Successive Event : $Y_i = 1 \sim \text{Binomial}(p_i)$

  Number of occurrences : $Y_i = k \sim \text{Poisson}(E_i\theta_i)$

- Process Level

$$\text{logit}(p_i) = \alpha_Z + \beta\gamma_i \tag{7}$$

$$\log(\theta_i) = \alpha_O + \gamma_i + \log(E_i) \tag{8}$$

with $\gamma_i = \left[\frac{1}{\sqrt{\tau_\gamma}}\left(\sqrt{\varphi}u_i + \sqrt{1-\varphi}v_i\right)\right]$. (9)

- Parameters Level

$$\begin{aligned} &\alpha_Z, \alpha_O \sim N(0, 1/\sqrt{0.0001}) \\ &\beta \sim N(0, 0.01) \\ &P\left(\left(\frac{1}{\sqrt{\tau_\gamma}}\right) > \left(\frac{0.5}{0.31}\right)\right) = 0.01 \\ &P(\varphi < 0.5) = 2/3. \end{aligned} \tag{10}$$

At data level we divided in two data/responses i.e. probability of successive event which defined as Binomial probability distribution and number of occurrences as Poisson distribution. E_i is expected case in specific region i to the whole population, and θ_i is relative risk of region i to its standard population [16]. In process level probability of successive event modelled using logistics regression and log-linear regression for number of occurrences, with α_Z, α_O are the intercept, γ_i is BYM2 used to model spatial random effects. BYM2 model is re-parameterization of BYM model with mixing parameter $\varphi \in [0,1]$ and the precision parameter τ_γ. BYM2 consist of spatially structured component u_i and an unstructured component v_i, while β represents a shared spatial random component between two regressions. In parameters level α_Z, α_O are set to have normal prior distribution with large variance, β is set to normal prior with mean 0 and $\sigma^2 = 0.01$. Mixing φ and precision parameters are set to penalized complexity prior [16], [17]. More explanation about BYM2 and penalized complexity (PC) prior we refer to [18] and [17].

Mixture models have two likelihoods, one for successive event for region i with Binomial distribution, $Z_i \sim \text{Binomial}(p_i, n_i = 1)$, and one for number of occurrences with (truncated) Poisson distribution, $O_i \sim (\text{Trunc})\text{Poisson}(E_i\theta_i)$. Because the linear predictors for each response are different, the response matrix must also be adjusted to combine two responses with larger matrix dimensions. Response matrix can be formed from the combination of two column vectors for each process, namely number of occurrences in region i can be zero or positive number, so the disease occurrence z_i in region i is defined as

$$z_i = \begin{cases} 1, & \text{if an event occurs} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

and the number of occurrences o_i as

$$o_i = \begin{cases} \text{NA}, & \text{if an event doesn't occur} \\ \text{occurrence number}, & \text{otherwise.} \end{cases} \tag{12}$$

Therefore, matrix response Y in mixture model can be written as

$$Y = \begin{bmatrix} z_1 & NA \\ \vdots & \vdots \\ z_n & NA \\ NA & o_1 \\ \vdots & \vdots \\ NA & o_n \end{bmatrix}. \tag{13}$$

However, literature that specifically studies the influence of demographics and geographic conditions in the context of statistical spatial Bayesian mixture modeling for excess zeros data is very rare. Therefore, in this paper the modeling will be expanded by including elevation as an explanatory variable, so equations (7) and (8) at the process level can be written as

$$\text{logit}(p_i) = \alpha_Z + \alpha_{Z,\text{elev}} X_{\text{elev}} + \beta\gamma_i \tag{14}$$

$$\log(\theta_i) = \alpha_O + \gamma_i + \alpha_{O,\text{elev}} X_{\text{elev}} + \log(E_i) \tag{15}$$

which $\alpha_{Z,\text{elev}}$ and $\alpha_{O,\text{elev}}$ state the influence of elevation in each region respectively.

2.2 INLA Inference and Code

INLA (Integrated nested Laplace approximation) makes it possible to perform approximate Bayesian inference on Gaussian latent models which are part of generalized linear mixed spatial models. Specifically, the model at the data level has the following form:

$$y_i|\boldsymbol{x}, \boldsymbol{\Theta} \sim \pi(y_i|x_i, \boldsymbol{\Theta}), i = 1,2, \dots, n, \tag{16}$$

$$\boldsymbol{x}|\boldsymbol{\Theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\Theta}), \boldsymbol{Q}(\boldsymbol{\Theta})^{-1}),$$

$$\boldsymbol{\Theta} \sim \pi(\boldsymbol{\Theta})$$

where $\boldsymbol{y}$ are the observed data, $\boldsymbol{x}$ represents a Gaussian field, and $\boldsymbol{\Theta}$ are parameters and hyperparameters. $\boldsymbol{\mu}(\boldsymbol{\Theta})$ and $\boldsymbol{Q}(\boldsymbol{\Theta})$ each represent the mean and the precision (inverse of covariance) matrix. In many situations, observation in region i, y_i are assumed to belong an exponential family with mean $\mu_i = g^{-1}(\eta_i)$. η_i is the canonical link function of response in exponential family. Linear predictor η_i can be written in additive form to accounts effects as:

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}\left(u_{ji}\right). \tag{17}$$

Here α is the intercept, $\{\beta_k\}$'s quantify the linear (fixed) effects of covariates $\{z_{ki}\}$ on the response, and $\{f^{(j)}(.)\}$'s are a set of random effects defined in terms of some covariates $\{u_{ji}\}$. Analytical approximation and numerical integration are combined to obtained approximated mean posterior distribution of the parameters. Let $\boldsymbol{x} = \left(\alpha_Z, \alpha_O, \alpha_{Z,\text{elev}}, \alpha_{O,\text{elev}}, \beta\right) \sim N(\mu, 1/\sqrt{\tau})$ denote the vector of latent Gaussian and $\boldsymbol{x}^* = (\tau_\gamma, \varphi)$ denote the vector parameter of the random components uses penalized complexity prior as stated in (10). For a more in-depth explanation on INLA inference we recommend referring to [16]. We implemented the mixture model with R-INLA package with penalized complexity prior and BYM2 model formula with the following R-code:

```
formula = Y ~  -1 + mu.z + mu.o + x_elev.z +x_elev.o +
  f(idarea, model = "bym2", graph = g, scale.model = TRUE,
    constr = TRUE,
    hyper = list(phi=list(prior="pc",
                          param = c(0.5,2/3),initial=3),
                 prec = list(prior="pc.prec",
                             param = c(1,0.01),
                             initial=1.5)))+
  f(idarea1,copy = 'idarea',fixed = FALSE)
```

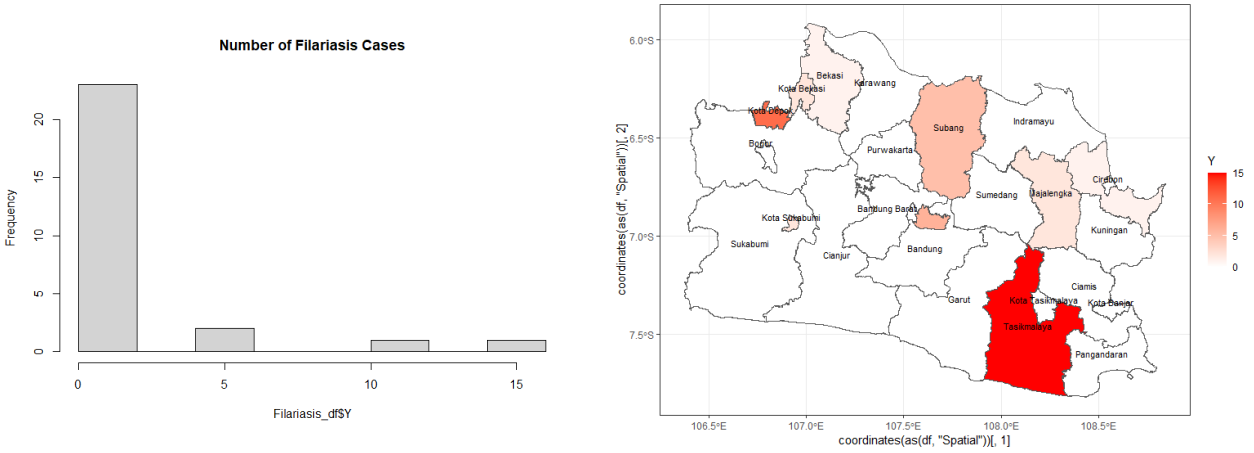
```
r.bym2 <- inla(formula,family = c('binomial','poisson'),
               data=df,E=E,verbose=F,control.predictor =
list(compute=TRUE,link=TRUE),
               control.compute = list(dic=TRUE,cpo=TRUE))
```

3. CASE STUDY: FEMALE LYMPHATIC FILARIASIS CASES IN WEST JAVA

Lymphatic filariasis still exists in Indonesia, especially in specific provinces such as Papua, East Nusa Tenggara, and West Java. According to data from the Ministry of Health of the Republic of Indonesia, almost 13,000 cases of lymphatic filariasis have been recorded [19]. Lymphatic filariasis is a type of dangerous infectious disease and can cause physical disabilities for the sufferers. Lymphatic filariasis is caused by Filaria Worms which are transmitted by various types of mosquitoes. In Indonesia, it is currently known that there are 23 species of mosquitoes from the genera Anopheles, Culex, Mansonia, Aedes and Armigeres which can act as vectors for transmitting Filariasis. This disease is chronic and if sufferer do not receive prompt and optimal treatment it can cause permanent disabilities in form of enlargement of the legs, arms and genitals in both women and men [20].

Here we consider modeling female lymphatic filariasis disease counts in West Java, Indonesia. Data was taken from West Java provincial government website [21], in 2019 at district city observation unit level in 27 regions. Many cases of zero sufferers have been reported (with 67% zeros proportion), but there are several regions that show a very high number of cases compared to other regions. Some regions do not report cases, so in this case they are assumed to have zero (sampling) cases. Figure 1(a) shows a histogram of 27 district city in West Java province, and part (b) shows map distribution of cases in each district city. The histogram in Figure 1(a) clearly states that most region have zero cases. However, there are quite real differences in region that have a very high number of cases compared to other regions, namely in Tasikmalaya with 15 cases, Depok city with 11 cases, and Bandung City with 6 cases. The histogram displays a fairly extreme distribution of data on the right. Disease distribution in Figure 1(b) shows that regions with high

cases are quite spread across various areas of West Java. The pattern of spatial dependence between regions is quite clear, areas with zero cases are in the western to central part of West Java, while areas with high cases are spread across the northern and southern regions. Several geographic and demographic influences have been studied in several publications to determine their relationship with the number of lymphatic filariasis cases [22], [23].



(a) (b)

Figure 1. The histogram (a) and the map (b) female filariasis cases in West Java at county level.

We fitted three mixture models and reported mean posterior probability for fixed and random effects, standard deviation and DIC in Table 1. We divide modeling into two large parts, namely without elevation and with elevation as an explanatory variable. In modeling without elevation, fixed effect in form of intercepts, $\hat{\alpha}_z$ and $\hat{\alpha}_O$ have significant influence (with credibility interval (CI) does not contain zero) on Binomial-Poisson mixture model with negative values. However, on contrary, intercepts have significant influence on Binomial-ZIP type 0, and only intercept in logit regression, $\hat{\alpha}_z$, has significant influence on Binomial-ZIP type 1 mixture model.

Table 1. Mean posterior probability, standard deviation and DIC

Coefficients	Binomial-Poisson			Binomial-Poisson with Elevation		
	Mean (95% CI)	$\hat{\sigma}$	**DIC**	**Mean (95% CI)**	$\hat{\sigma}$	**DIC**
Fixed Effects						
$\hat{\alpha}_z$	-1.622 (-3.332, -0.075)	0.825		-0.967 (-3.332, 1.304)	1.173	
$\hat{\alpha}_O$	-1.679 (-3.661, -0.200)	0.871		-1.491 (-3.933, 0.640)	1.162	
$\hat{\alpha}_{Z,\text{elev}}$	-	-		-0.306 (-1.031, 0.334)	0.344	
$\hat{\alpha}_{O,\text{elev}}$	-	-	68.18	-0.191 (-1.484, 0.393)	0.400	68.36
Random Effects						
$1/\sqrt{\hat{\tau}_u}$	0.717 (0.306, 1.431)	0.291		0.653 (0.287, 1.292)	0.260	
$\hat{\varphi}$	0.133 (0.032, 0.343)	0.081		0.155 (0.026, 0.443)	0.109	
$\hat{\beta}$	1.220 (0.680, 1.766)	0.279		1.227 (0.735, 1.712)	0.248	
Coefficients	**Binomial-ZIP type 0**			**Binomial-ZIP type 0 with Elevation**		
	Mean (95% CI)	$\hat{\sigma}$	**DIC**	**Mean (95% CI)**	$\hat{\sigma}$	**DIC**
Fixed Effects						
$\hat{\alpha}_z$	-0.886 (-1.863, 0.044)	0.484		-0.578 (-1.973, 0.779)	0.699	
$\hat{\alpha}_O$	1.207 (-0.251, 2.150)	0.614		0.594 (-1.362, 1.911)	0.832	
$\hat{\alpha}_{Z,\text{elev}}$	-	-		-0.121 (-0.485, 0.235)	0.183	
$\hat{\alpha}_{O,\text{elev}}$	-	-	111.85	0.204 (-0.134, 0.555)	0.171	111.99
Random Effects						
$1/\sqrt{\hat{\tau}_u}$	4.866 (1.014, 14.580)	3.681		5.014 (0.903, 16.489)	4.321	
$\hat{\varphi}$	0.176 (0.010, 0.606)	0.159		0.187 (0.012, 0.633)	0.166	
$\hat{\beta}$	1.062 (0.466, 1.655)	0.302		1.081 (0.482, 1.700)	0.309	
$\hat{p}$	0.653 (0.473, 0.808)	0.086		0.653 (0.473, 0.808)	0.086	
Coefficients	**Binomial-ZIP type 1**			**Binomial-ZIP type 1 with Elevation**		
	Mean (95% CI)	$\hat{\sigma}$	**DIC**	**Mean (95% CI)**	$\hat{\sigma}$	**DIC**
Fixed Effects						
$\hat{\alpha}_z$	-1.566 (-3.251, -0.039)	0.815		-0.995 (-3.363, 1.248)	1.168	
$\hat{\alpha}_O$	-1.436 (-3.266, 0.030)	0.847		-1.468 (-3.962, 0.632)	1.165	
$\hat{\alpha}_{Z,\text{elev}}$	-	-		-0.261 (-0.905, 0.364)	0.322	
$\hat{\alpha}_{O,\text{elev}}$	-	-	71.66	-0.073 (-0.073, 0.467)	0.277	71.22
Random Effects						
$1/\sqrt{\hat{\tau}_u}$	0.803 (0.243, 2.012)	0.467		0.645 (0.264, 1.334)	0.278	
$\hat{\varphi}$	0.136 (0.035, 0.331)	0.077		0.125 (0.035, 0.295)	0.068	
$\hat{\beta}$	1.218 (0.649, 1.794)	0.291		1.209 (0.688, 1.710)	0.260	
$\hat{p}$	0.066 (0.012, 0.182)	0.045		0.062 (0.015, 0.155)	0.037	

In models with elevation, three mixture models do not have significant intercepts for either logit or log-linear regressions. Likewise with elevation, three models state that there is not strong enough evidence to state that elevation has significant influence. From a geographical background, West Java has very diverse regional altitudes. Regions with high number of extreme cases stand at very different altitudes, such as Tasikmalaya (15 cases) at 411.66 meters above sea level (masl), Depok city (11 cases) at 87.8 masl, while Cimahi city (zero cases) at 794.36 masl. This supports evidence that in this case study no significant relationship was found between regional altitudes and number of female lymphatic filariasis cases. The standard deviation, $\hat{\sigma}$, for fixed effects has the smallest value in Binomial-ZIP type 0, while other two mixture models have almost the same standard deviation.

The posterior mean for random effects shows that Binomial-ZIP type 0 model has the highest mixing parameter value, $\hat{\varphi}$. This shows that influence of spatial structure dependence in this case is around 17.6% and 18.7%, while spatial structure independence is around 82.4% and 81.3% for modeling with and without elevation respectively. $\hat{\beta}$ near to 1 shows the spatial pattern of occurrence and number for occurrences are similar in mixture model. Both with and without elevation modeling, $\hat{\beta}$ is close to 1 with the most similar random effect values for both regressions is being Binomial-ZIP type 0. Meanwhile, $1/\sqrt{\hat{\tau}_u}$ shows the marginal deviation from regression intercept (initial risk) α, independent in the graph. The highest marginal deviation is in Binomial-ZIP type 0 with $\frac{1}{\sqrt{\hat{\tau}_u}} = 4.866$, while the Binomial-Poisson and Binomial-ZIP type 1 models have almost the same small marginal deviation or in other word the models have large precision values. The mean posterior of zero probability for logistic regression are estimated the same $\hat{p} = 0.653$ for Binomial-ZIP type 0 for both with and without elevation, and 0.066 and 0.062 for Binomial-ZIP type 1 for with and without elevation modeling respectively. Binomial-ZIP type 1 gives lower probability owing to some zeros covered by Poisson distribution. According to DIC, Binomial-Poisson and Binomial-ZIP type 1 performs better compared to Binomial-ZIP type 0.

Low DIC values are shown by Binomial-Poisson and Binomial-ZIP type 1 models. To investigate in more detail the estimator values for the two models, Table 2 (without elevation

modeling) represents estimator values for spatial random component, $\hat{\gamma}_i$, the mean posterior of relative risk for occurrence probability and mean posterior of relative risk for number of occurrences in five highest case districts cities in West Java. In Table 2, both models show the same order for five regions with the highest number of lymphatic filariasis cases affecting women. The Binomial-Poisson model shows a higher probability of occurrence and number of occurrences compared to the Binomial-ZIP model. The region with the greatest probability and number of occurrences is Tasikmalaya. Table 3 shows modeling with elevation. Both models still show the same order, but there is a change in order of the relative risk of occurrence in logistic regression model. In this model, the greatest occurrence probability is in Depok city.

Table 2. Top 5 counties with the highest relative risk, without elevation

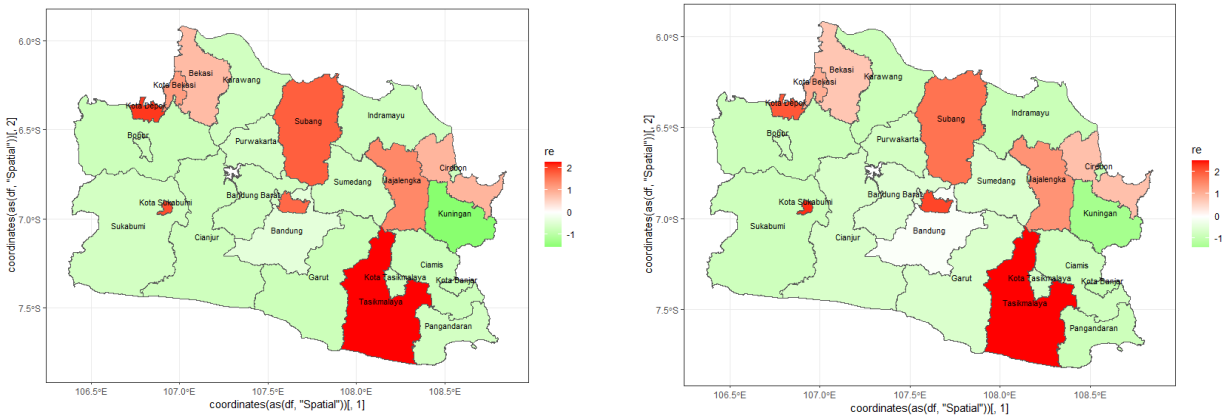
County Name	**Estimated Relative Risk (RR) for Z (occurrence probability)**		**Estimated Relative Risk (RR) for O (number of occurrences)**	
	Binomial-Poisson	**Binomial-ZIP Type 1**	**Binomial-Poisson**	**Binomial-ZIP Type 1**
Tasikmalaya	0.955	0.947	27.721	27.776
Depok City	0.934	0.922	18.532	18.573
Sukabumi City	0.900	0.886	16.912	17.194
Subang	0.881	0.864	10.272	10.355
Bandung City	0.863	0.843	8.303	8.363

Table 3. Top 5 counties with the highest relative risk, with elevation

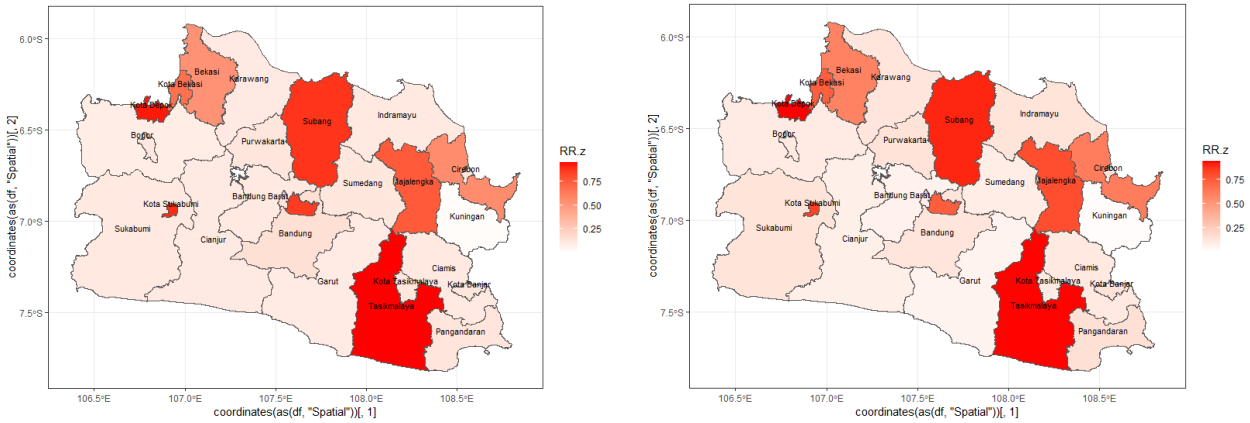
Estimated Relative Risk (RR) for Z (occurrence probability)		**Relative Risk (RR) for O (number of occurrences)**	
Binomial-Poisson	**Binomial-ZIP Type 1**	**Binomial-Poisson**	**Binomial-ZIP Type 1**
0.948	0.943	27.772	27.926
Depok City	Depok City	Tasikmalaya	Tasikmalaya
0.946	0.934	18.616	18.646
Tasikmalaya	Tasikmalaya	Depok City	Depok City
0.906	0.898	17.226	18.063
Subang	Subang	Sukabumi City	Sukabumi City
0.840	0.808	10.355	10.438
Sukabumi City	Sukabumi City	Subang	Subang
0.804	0.789	8.367	8.588
Majalengka	Majalengka	Bandung City	Bandung City

Disease mapping is based on the smallest DIC values in Binomial-Poisson model. Figure 2(a) presents a mapping of spatial random effects estimator which is the BYM2 component in regression equation, (b) presents estimated relative risk for occurrence probability and (c) estimated relative risk for number of occurrences. Left and right side of Figure 2 state for modeling without and with elevation respectively. In Figure 2(a), spatial random component is very capable of representing the spatial pattern of lymphatic filariasis cases very well. The model can detect region with a high number of extreme cases such as Tasikmalaya, Depok city and Subang which are marked with a strong red color. Likewise, the model very well detects regions with a low number of cases colored in white and green. The red color indicates disease potential is above average, whereas the green color indicates disease potential of region is below average with an average value of 0, and the average is indicated with white color. White region is clearly visible in model with elevation, namely Bandung which is in the central region of West Java.

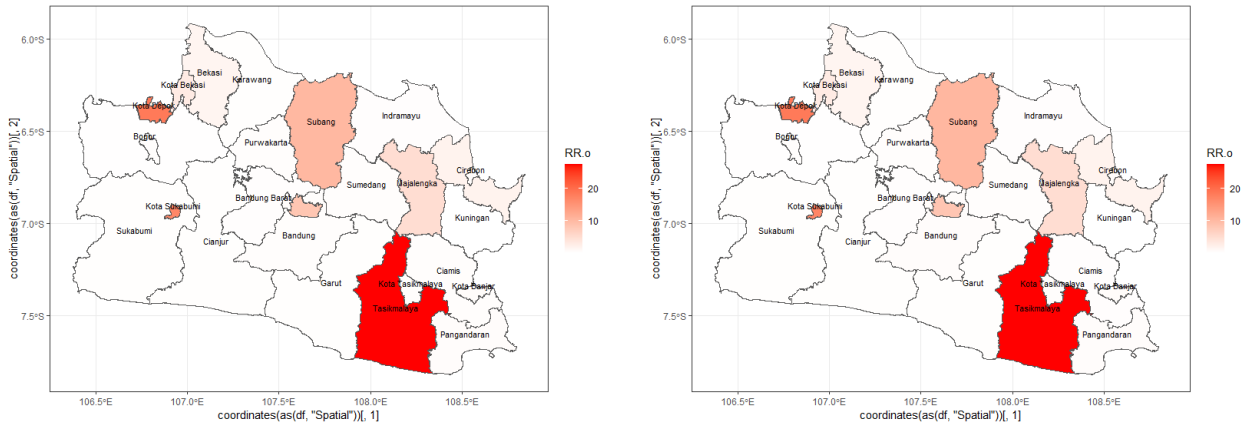
Figure 2(b) shows the same spatial pattern. However, if we see in more detail according to Tables 2 and 3, the chance of occurrence probability in model without elevation is Tasikmalaya, namely 0.955, while in model with elevation region with the largest occurrence probability is Depok city with a chance of 0.943. Figure 2(c) also shows the same spatial pattern with Tasikmalaya as the region with the largest number of occurrences, but in general model with elevation has a higher number of occurrences in all areas.



(a) Estimated spatial random effects $\hat{\gamma}_i$



(b) Estimated relative risk for occurrence probability



(c) Estimated relative risk for number of occurrences

Figure 2. Binomial-Poisson without elevation (left), with elevation (right)

4. CONCLUSIONS

This paper establishes a spatial mapping of number of lymphatic filariasis cases affecting women in West Java. The modeling used proposes a mixture distribution method for spatial data with excessive zero. The proposed mixture distributions are Binomial-Poisson, Binomial-ZIP type 0 and Binomial-ZIP type 1. The matrix structure is a combination of two responses, each of which has a regression equation, logistic regression for occurrence probability, and log-linear regression for number of occurrences. The spatial random component in both regressions uses BYM2 which is a combination of spatially structured and spatially unstructured random effects. Using the BYM2 model in the proposed mixture models provide very good results. This is because BYM2 prior parameters are reparametrized using a scaled precision matrix. Therefore, the interpretation of

parameters is very clear, precision parameter indicates marginal deviation from the initial risk (intercept), and mixing parameter expresses the variation between spatial dependence and spatial independence of the data. In addition, BYM2 uses a PC prior which provides the advantage of a simpler model so that the spatial random component scaling technique becomes more efficient.

Because there are two types of zeros in data, namely structural zero and sampling zero, we propose a combined Binomial-ZIP model of type 0 and type 1. In many case studies regarding data with excess zeros, the number of non-zero cases that occur has a value that is not very high or around zero. However, in this case study of lymphatic filariasis, the number of extreme events occurred in several regions, there were even some regions with 11 to 15 cases in 1 year for rare types of disease. For this reason, this paper also proposes a combined Binomial-Poisson distribution model. The Poisson distribution in this mixture model is expected to be able to overcome the phenomenon of regions that have a number of extreme cases. Inference uses INLA to avoid convergence issue of parameters posterior distribution, and selection of the best model uses smallest DIC.

Finally, Bayesian spatial mixture model was applied to female lymphatic filariasis data in 27 districts cities in West Java. The three proposed models can map lymphatic filariasis cases very well. Binomial-ZIP type 0 and Binomial-ZIP type 1 models have similar spatial estimation patterns but type 0 has a larger DIC value. This indicates that filariasis data contains not only structural zero but also sampling zero which must be taken into account in the analysis. The Binomial-Poisson model has the smallest DIC value, this indicates that extreme case events can still be handled very well by Binomial-Possion model. For this reason, the Binomial-Poisson distribution to model zero-inflated data with the number of extreme cases in several regions can be considered for use in modeling. Tasikmalaya has the highest spatial risk compared to other counties. The probability of a lymphatic filariasis case occurring in this region is around 95% and the relative risk of lymphatic filariasis cases in Tasikmalaya can reach 27 times compared to its standard population. Furthermore, in this case there was no significant relationship between elevation and the number of female lymphatic filariasis cases in West Java.

Even though Binomial-Poisson mixture model has the best performance in this case study, exploration using comprehensive simulations still needs to be done. This is to further investigate the characteristics of zero-inflated data that are suitable for this model. In this case the proportion of zero data is around 67% with extreme event values in several regions which are very suitable for the Binomial-Poisson model. The mixture model used in this paper can be developed using other zero-inflated models, including the negative binomial distribution [24], or can even use the beta distribution [25]. Recently, studies on lymphatic filariasis using a differential equation approach have been carried out as in [26]. The development of modeling based on differential equations can also be developed into spatial modeling with regionalization based on stochastic differential equations on point spatial data. Moreover, the development of spatio-temporal and multivariate modeling for zero-inflated data with mixture models is also very interesting to study in future research.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] R.N. Rachmawati, A. Djuraidah, A.H. Wigena, et al. Spatio-temporal Bayes regression with INLA in statistical downscaling modeling for estimating West Java rainfall, in: Proceedings of the 1st International Conference on Statistics and Analytics, ICSA, Bogor, 2019.

[2] A. Djuraidah, R.N. Rachmawati, A.H. Wigena, et al. Extreme data analysis using spatio-temporal Bayes regression with INLA in statistical downscaling model, Int. J. Innov. Comput. Inf. Control. 17 (2021), 259-273.

[3] A. Arab, Spatial and spatio-temporal models for modeling epidemiological data with excess zeros, Int. J. Environ. Res. Public Health. 12 (2015) 10536–10548. https://doi.org/10.3390/ijerph120910536.

[4] N. Asmarian, S.M.T. Ayatollahi, Z. Sharafi, et al. Bayesian spatial joint model for disease mapping of zero-inflated data with R-INLA: A simulation study and an application to male breast cancer in Iran, Int. J. Environ. Res. Public Health. 16 (2019), 4460. https://doi.org/10.3390/ijerph16224460.

[5] C.X. Feng, A comparison of zero-inflated and hurdle models for modeling zero-inflated count data, J. Stat. Distrib. Appl. 8 (2021), 8. https://doi.org/10.1186/s40488-021-00121-4.

[6] D.B. Bekalo, D.T. Kebede, Zero-inflated models for count data: an application to number of antenatal care service visits, Ann. Data. Sci. 8 (2021), 683-708. https://doi.org/10.1007/s40745-021-00328-x.

[7] D. Kang, J. Choi, Bayesian zero-inflated spatio-temporal modelling of scrub typhus data in Korea, 2010-2014, Geospat. Health. 13 (2018), 215-223. https://doi.org/10.4081/gh.2018.665.

[8] B. Neelon, Bayesian zero-inflated negative binomial regression based on pólya-gamma mixtures, Bayesian Anal. 14 (2019), 829-855. https://doi.org/10.1214/18-ba1132.

[9] C.E. Lee, S. Kim, Applicability of zero-inflated models to fit the torrential rainfall count data with extra zeros in South Korea, Water. 9 (2017), 123. https://doi.org/10.3390/w9020123.

[10] Y. Lee, M.M. Alam, M. Noh, et al. Spatial modeling of data with excessive zeros applied to reindeer pellet‐group counts, Ecol. Evol. 6 (2016), 7047-7056. https://doi.org/10.1002/ece3.2449.

[11] O. Lyashevska, D.J. Brus, J. van der Meer, Mapping species abundance by a spatial zero-inflated Poisson model: a case study in the Wadden Sea, the Netherlands, Ecol. Evol. 6 (2016), 532-543. https://doi.org/10.1002/ece3.1880.

[12] H. Zhu, S.M. DeSantis, S. Luo, Joint modeling of longitudinal zero-inflated count and time-to-event data: A Bayesian perspective, Stat. Meth. Med. Res. 27 (2016), 1258-1270. https://doi.org/10.1177/0962280216659312.

[13] S. Martino, A. Riebler, Integrated nested Laplace approximations (INLA), preprint, (2019). http://arxiv.org/abs/1907.01248.

[14] J.B. Illian, S. Martino, S.H. Sørbye, et al. Fitting complex ecological point process models with integrated nested Laplace approximation, Methods Ecol. Evol. 4 (2013), 305-315. https://doi.org/10.1111/2041-210x.12017.

[15] P. Moraga, Geospatial health data: Modeling and visualization with R-INLA and shiny, CRC Press, Boca Raton, 2019.

[16] D. Simpson, H. Rue, A. Riebler, et al. Penalising model component complexity: a principled, practical approach to constructing priors, Stat. Sci. 32 (2017), 1-28. https://doi.org/10.1214/16-sts576.

[17] A. Riebler, S.H. Sørbye, D. Simpson, et al. An intuitive Bayesian spatial model for disease mapping that accounts for scaling, Stat. Methods Med. Res. 25 (2016), 1145-1165. https://doi.org/10.1177/0962280216660421.

[18] E.L. Davis, J. Prada, L.J. Reimer, et al. Modelling the impact of vector control on lymphatic filariasis programs: current approaches and limitations, Clinic. Infect. Dis. 72 (2021), S152-S157. https://doi.org/10.1093/cid/ciab191.

[19] Y. Cheng, X. Wang, Q. Pan, et al. Modeling the parasitic filariasis spread by mosquito in periodic environment, Comput. Math. Methods Med. 2017 (2017), 4567452. https://doi.org/10.1155/2017/4567452.

[20] K.H. Lee, C. Pedroza, E.B.C. Avritscher, et al. Evaluation of negative binomial and zero-inflated negative binomial models for the analysis of zero-inflated count data: application to the telemedicine for children with medical complexity trial, Trials 24 (2023), 613. https://doi.org/10.1186/s13063-023-07648-8.

[21] B. Tang, H.A. Frye, A.E. Gelfand, et al. Zero-inflated beta distribution regression modeling, J. Agric. Biol. Environ. Stat. 28 (2022), 117-137. https://doi.org/10.1007/s13253-022-00516-z.

[22] World Health Organization, Lymphatic filariasis, World Health Organization, 1 June 2023. https://www.who.int/news-room/fact-sheets/detail/lymphatic-filariasis. [Accessed 15 December 2023].

[23] Ministry of Health of Republic Indonesia, Health Services, Directorate General of Health Services. https://yankes.kemkes.go.id/view_artikel/70/penyakit-kaki-gajah. [Accessed 15 December 2023].

[24] West Java Province of Indonesia, Open Data Jabar, Public Health Office of West Java, 1 December 2019. https://opendata.jabarprov.go.id/id/dataset/jumlah-penderita-kronis-filariasis-berdasarkan-kategori-kasus-dan-jenis-kelamin-di-jawa-barat. [Accessed October 2023].

[25] M. Blangiardo, M. Cameletti, Spatial and spatio-temporal Bayesian models with R-INLA, John Wiley & Sons, 2015.

[26] I.H. Febiriana, V. Adisaputri, P.Z. Kamalia, et al. Impact of screening, treatment, and misdiagnose on lymphatic filariasis transmission: a mathematical model, Commun. Math. Biol. Neurosci. 2023 (2023), 67. https://doi.org/10.28919/cmbn/7983.