Far East Journal of Mathematical Sciences (FJMS)

http://www.pphmj.com
http://dx.doi.org/10.17654/0972087122009
Volume 135, 2022, Pages 63-76 **P-ISSN: 0972-0871**

THE DHF MODELING IN CIAMIS REGENCY BY USING CAR-BYM, GENERALIZED POISSON, AND NEGATIVE BINOMIAL

Jajang*, Budi Pratikno and Mashuri

Department of Mathematics
Faculty of Mathematics and Natural Sciences
Jenderal Soedirman University
Indonesia
e-mail: rzjajang@yahoo.com

Abstract

Dengue hemorrhagic fever (DHF) is one of contagious diseases that may threaten human health. It is necessary to study the DHF distribution patterns in the affected area for its transmission prevention and control. This research studied disease mapping of DHF in Ciamis Regency by using generalized Poisson (GP), negative binomial (NB), and CAR-BYM models. Based on the root mean square error (RMSE), the CAR-BYM is the best model for DHF case modeling in Ciamis Regency. The highest relative risk value is that of

Received: December 6, 2021; Accepted: January 27, 2022
2020 Mathematics Subject Classification: 62J02, 62J05, 62P10.
Keywords and phrases: relative risk, disease mapping, CAR-BYM, generalized Poisson, negative binomial.
*Corresponding author

How to cite this article: Jajang, Budi Pratikno and Mashuri, The DHF modeling in Ciamis Regency by using CAR-BYM, generalized Poisson, and negative binomial, Far East Journal of Mathematical Sciences (FJMS) 135 (2022), 63-76.
http://dx.doi.org/10.17654/0972087122009

Published Online: April 9, 2022

Cijeungjing district and the lowest relative risk value is that of Lakbok district. Lakbok district is at the lowest risk of transmission and Cijeungjing district is at the highest risk of transmission.

1. Introduction

Non-overlapping data related to medical research such as disease distribution pattern in an area have high complexity and spatial heterogeneity. Research on spatial data model, such as disease mapping, requires correct parameter estimation. Disease mapping is useful to find geographic distribution of disease burden and disease incident based on risk level. The clustering of relative risk (RR) in this case is almost the same with hotspot in general spatial model case, such as in determining landscape by [1]. Disease mapping certainly cannot be separated from spatial effect. In spatial effect accommodation, this model involves spatial weights matrix. Spatial weights matrix is a non-negative matrix, scored 1 if units are close to each other and 0 otherwise. The other way to reconstruct spatial weights matrix is to combine similarity of variable attribute and proximity relationship, W-AMOEBA matrix [2, 3].

Data of the number of DHF cases are the count data. Generally, the model used to count data is Poisson, generalized Poisson, or negative binomial. However, in these models, the spatial aspect is not included. Even though in disease mapping, spatial aspect is important.

Two of the popular spatial models are spatial autoregressive (SAR) and spatial error model (SEM). The other model that can be used is conditional autoregressive-Bessag-York-Mollie (CAR-BYM). The CAR-BYM model can accommodate spatial and non-spatial aspects as the consequence of heterogeneity of cases between regions. In addition, the CAR-BYM model can detect areas with relative risk through interpolation of disease mapping resulting from information available from neighborhood. Study on spatial models for area data has been commonly carried out, especially for a model with continued response variable type [4, 5]. In the CAR-BYM model, the error distribution is dependent. Therefore, the popular parameter estimation

methods such as ordinary least square (OLS) and maximum likelihood cannot be used to estimate the parameter in the CAR-BYM model. As an alternative, Bayes method can be used to estimate the parameter of the CAR-BYM model. The advantage of this method is flexibility in assumption of the error distribution. However, to determine the estimate of parameter by using Bayes method need to find the integral for high dimensional space. Therefore, the Bayesian method through Markov Chain Monte Carlo (MCMC) is used. The MCMC method has been commonly used by researchers in medical field [4, 6-9].

Ciamis Regency is one of the 27 Regencies/Cities in West Java Province. In 2019, the DHF cases in Ciamis increased from previous years whereas in 2017-2018, the DHF cases in Ciamis were quite low, and in 2020, Ciamis Regency was once again a red zone for DHF case. According to the Epidemiologic Data and Surveillance Center of the Ministry of Health of the Republic of Indonesia, the causes of increase in and distribution of DHF were, among others, high mobility of the people, urban development, climate change, changes in population density, population distribution, and other epidemiologic factors. The number of regional DHF cases such as the case in Ciamis Regency is one type of area data. Interaction is possible since a DHF case is of contagious type of case. Therefore, the quantitative measure of a variable that is the attention in an area will be influenced by other areas as the consequence of interaction.

2. Research Method

2.1. The data

The research was conducted from March 2021 to October 2021. The research location was the Department of Mathematics, and the data were collected from Ciamis Regency, West Java Province.

2.2. Spatial data and spatial weights matrix

The area data was one of the spatial data besides point reference and point pattern data [10]. If D was a region consisting of non-intersecting sub-

areas, the type of area data has characteristics where fixed D is partitioned into finite number of area units [11, 12]. Spatial weights matrix (W) is a nonnegative matrix that specifies neighborhood set for each observation. Matrix W used geographic relationship (spatial contiguity, inverse distance, and k-nearest neighbors, k-NN). Matrix W with geographic distance concept is 1 for adjacent inter-areas, and 0 for between distant areas.

2.3. Count data models

2.3.1. Negative binomial and generalized Poisson model

The Poisson model is a common model that is often used for data counts. Suppose Y is random variable Poisson distributed, $Y_i \sim Poi(\mu_i)$, $i = 1, 2, ..., n$. Then the probability mass function of Y is

$$p(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ... \tag{1}$$

Expected value and variance of Y are $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i$, respectively. This condition is called *equidispersion*. Poisson's model assumes this equidispersion condition. However, in many cases, this equidispersion condition is not met, the situation is called as *overdispersion*. Common models are used in dealing with cases of overdispersion negative binomial (NB) and generalized Poisson (GP) models. Probability mass function (pmf) of NB is

$$f(y, \mu_i, m) = \exp\left[y \ln\left(\frac{m\mu}{1+m\mu}\right) + \frac{1}{m}\ln\left(\frac{1}{1+m\mu}\right) + \ln\frac{\Gamma\left(y+\frac{1}{m}\right)}{y!\Gamma\left(\frac{1}{m}\right)}\right], \tag{2}$$

where μ is the mean of Poisson random variable and m is the overdispersion parameter. In addition to the NB model, the generalized Poisson (GP) model is also commonly used to overcome the problem of overdispersion in the data count. Probability mass function (pmf) of GP is

$$f(y, \mu_i, m) = \left(\frac{\mu_i}{1+m\mu_i}\right)^{y_i} \frac{(1+my_i)^{y_i-1}}{y_i!} \cdot \exp\left[\frac{-\mu_i(1+my_i)}{1+m\mu_i}\right], \tag{3}$$

where μ is the mean of Poisson random variable, and m is the overdispersion parameter.

2.3.2. Poisson-lognormal model

Poisson-lognormal model was derived from a combination of Poisson distributions by assuming Poisson heteroscedasticity parameter. Suppose Y_i is a random variable that follows Poisson distribution, $Y_i \sim POI(\mu_i)$, $\mu_i = E_i\theta_i$ and $\theta_i = \exp(\eta_i)$. Based on the characteristics of exponential family distribution [13], the mean Poisson can be stated as

$$\log(\mu_i) = \log(E_i) + x_i'\beta + v_i, \tag{4}$$

where μ_i is the mean of Poisson distributed response variable, $\log(E_i)$ is the offset, x_i' is the free vector variable, β is the vector parameter, and v_i is the error that follows CAR.

2.3.3. Intrinsic conditional autoregressive (ICAR) model

When an area n was given to consist of non-overlapping sub-areas, each adjacent sub-area (having shared borders) was scored 1, otherwise scored 0. Spatial interaction between area pair i and j could be modeled as conditional normal $v_i \mid v_{j\neq i} \sim N\left(\sum_{i\neq j}\varphi_{ij}v_j, \tau_i^2\right)$. The adjacent spatial unit I is given by

$$E(v_i \mid v_{j\neq i}) = \mu_i + \sum_{J\in N_I}\varphi_{ij}(v_j - \mu_i) \quad \text{and} \quad Var(v_i \mid v_{j\neq i}) = \tau_i^2,\ N_i.$$

Further

$$f(v_i \mid v_{j\neq i} \in S)$$

$$= \left(\frac{1}{2\pi\tau_i^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=0}^{n}\frac{\left((v_j - \mu_i) - \rho\sum_{J\in N_I}\varphi_{ij}(v_j - \mu_i)\right)^2}{\tau_i^2}\right), \tag{5}$$

$\mu_i \in R$, $\tau_i^2 \in R^+$, $|\rho| < 1$, $\varphi_{ij} = \varphi_{ji}$ and $\varphi_{ii} = 0$. In spatial autoregressive, φ_{ij} describes the element of weights matrix. The spatial matrix W is given by

$$W = (w_{ij}),\ w_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \varphi_{ij} & \text{if adjacent } i \neq j. \end{cases}$$

2.3.4. Conditional autoregressive-BYM (CAR-BYM) model

Conditional autoregressive-BYM (CAR-BYM) model is a Poisson log normal model developed for disease mapping risk. This model covers ICAR component for spatial rarefaction and ordinary random effect component for non-spatial heterogeneity. Poisson regression model has been used to estimate relative risk (RR), that is, η_i for region i, $i = 1, 2, ..., n$, given the number y_i of cases. CAR-BYM model is specified as follows:

$$\eta_i = \mu + x_i'\beta + \phi_i + \theta_i, \quad i = 1, 2, ..., n, \tag{6}$$

where x_i' = observational vector of independent variable i, β = vector of parameter, ϕ_i = ICAR component, μ = average risk level, and θ_i = random effect of non-spatial heterogeneity component.

2.3.5. Bayesian estimation framework

Suppose Y_i, $i = 1, 2, ..., n$ are random samples of probability mass functions, pmf $P(y|\theta)$, with the vector of parameter $\theta = (\theta_i, ..., \theta_p)$. Then, according to [14],

$$P(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) P(\theta). \tag{7}$$

Parameter estimation by using the Bayesian method needed information about parameter θ, is called *prior distribution*. Prior distribution is viewed as introductory knowledge of parameter θ and is determined before the given observation data. Here, without losing generality, we take one parameter θ, so we obtain prior distribution $p(\theta)$ and we then create

joint pdf/pmf $p(\theta, y)$. Based on $p(\theta)$ and $p(\theta, y)$, we obtain the posterior distribution of θ as

$$P(\theta|y) = \frac{P(\theta, y)}{p(y)} = \frac{P(y|\theta)P(\theta)}{p(y)}, \tag{8}$$

where $p(y)$ is the marginal probability and $P(y|\theta)$ is the joint pdf/pmf. Furthermore, based on this posterior distribution, the estimator for parameter θ is $\hat{\theta} = E(\theta|y)$.

The stages of parameter estimation by using Bayes are given below:

(1) Forming likelihood function:

$$l(\beta, v) = \prod_{i=1}^{n} \frac{e^{-E_i\theta_i}(E_i\theta_i)^{y_i}}{y_i!} = P(y, E, \theta|\beta, v).$$

(2) Determining prior distribution of β and v:

$$p(\beta) = \left(\frac{1}{2\pi}\right)^{\frac{P}{2}}\left(\frac{1}{\tau_\beta}\right)^{p} \exp\left(-\frac{1}{2}\sum_{h=0}^{P}\frac{\beta_h^2}{\tau_\beta^2}\right) p(v_i|v_{i\neq j}, \tau_v^2)$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=0}^{n}\left(\frac{v_i - \sum_{i\neq j}^{n}\frac{w_{ij}, v_j}{w_{ij}}}{\tau_\beta}\right)^2\right).$$

(3) Forming posterior distribution based on the equations in stage (1) and (2).

2.3.6. Markov chain Monte Carlo

Apparently, it is not easy to determine expected value with posterior pdf on (6). Therefore, Bayes parameter estimation is used through Markov Chain

Monte Carlo (MCMC). For this, let parameter θ be a vector of parameters. Then the Gibb sampler method is as follows:

For $t = 1, 2, ..., T$:

- Stage 1. Take θ_1^t from $p(\theta_1 | \theta_2^{t-1}, \theta_3^{t-1}, ..., \theta_k^{t-1}, y)$
- Stage 2. Take θ_2^t from $p(\theta_2 | \theta_1^{t-1}, \theta_3^{t-1}, ..., \theta_k^{t-1}, y)$
- …
- Stage k. Take θ_k^t from $p(\theta_k | \theta_1^{t-1}, \theta_2^{t-1}, ..., \theta_{k-1}^{t-1}, y)$.

If the condition is stable in case of iteration τ_0, then the estimation of the parameter θ_i is given by

$$\hat{\theta}_k = \hat{E}(\theta_k | y) = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \theta_k^t.$$

3. Result and Discussion

3.1. Description analysis

The correlational analysis between response and predictor variables is the initial stage of regression model. Correlation between variables is initial description of the relationship between predictor variable and response variable. However, here we used only three predictor variables. These are the number of health workers, population density, and altitude. Meanwhile, the response variable used is the number of dengue hemorrhagic fever (DHF).

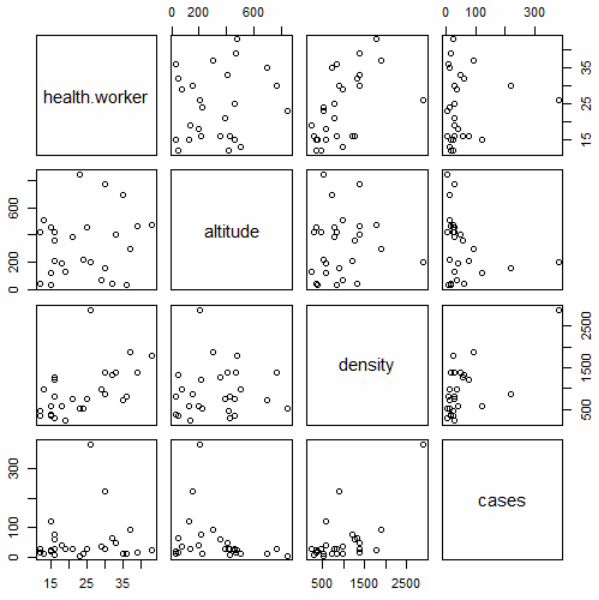

Figure 1. The scatterplot among variables.

The number of health workers consisted of the number of midwives and the number of community health center personnel. The scatterplot of the variables is presented in Figure 1.

Based on Figure 1, we cannot see relationship clearly, and hence it is necessary to create a model for interpretation of the relationship between predictor variables and response variable. From the model, we also can know contribution of each of the predictor variables to the response variable.

3.2. Model

In disease mapping, relative risk (RR) is often used to measure the risk of a region with others. Relative risk values in a model data count depends on observations and the expected values. Here, the expected values are obtained from the best model. The best model is selected from generalized Poisson, negative binomial, and CAR-BYM models. The criterion used for model selection is the root mean square error (RMSE).

The results of analysis of variance (ANOVA) of the three models that are used for this data are listed in Tables 1-3.

Table 1. ANOVA of the generalized Poisson model

Coefficients	Estimate	Std. error	z-value	Pr(> | z |)
(Intercept):1	3.7305	0.3514	10.616	< 2e-16***
(Intercept): 2	1.1196	0.1208	9.265	< 2e-16***
density	0.0011	0.0001	8.560	< 2e-16***
health.worker	-0.0275	0.0131	-2.103	0.0355*
altitude	-0.0013	0.0006	-2.340	0.0193*

Table 2. ANOVA of the negative binomial model

Coefficients	Estimate	Std. error	z-value	Pr(> | z |)
(Intercept)	3.8239	0.3995	9.572	< 2e-16***
density	0.0010	0.0003	3.942	8.08e-05***
health.worker	-0.0138	0.0172	-0.801	0.423
altitude	-0.0025	0.0006	-4.042	5.30e-05***

Signif. codes: '***' for $\alpha = 0.001$ and '*' for $\alpha = 0.05$

Based on Tables 1 and 2, we see that the effects of density and altitude to the number of DHF are significant. However, health.worker effect is not significant.

Table 3. ANOVA of the CAR-BYM model

	Median	2.5%	97.5%	n.effective	Geweke.diag
(Intercept)	0.2514	-0.4553	1.0594	16.7	0.8
Density	0.0012	0.0007	0.0018	7.9	-1.3
health.worker	-0.0675	-0.0959	-0.0191	7.5	-0.7
altitude	-0.0011	-0.0030	0.0004	7.7	1.7
tau2	0.0207	0.0036	0.1590	86.2	0.8
sigma2	0.5295	0.2891	1.0312	128.2	1.3

The value of Geweke.diag on Table 3 is used to test significance of the predictor variables. The Geweke's diagnostic is used to determine the burn-in period of the smallest early portion of the chain that passes the diagnostic. From the Geweke.diag value of Table 3, we see that the altitude and density are significant.

Based on Tables 1-3, we can see that the coefficient of the population density is positive and significantly correlated with the number of DHF cases. This means that if the population density in a district increases, then the number of DHF cases also increases. Meanwhile, coefficients of the number of health workers and altitude are negative. This means that if the health workers and altitude increase, then the number of DHF cases decreases.

Furthermore, to calculate the relative risk of DHF for each district in Ciamis regency, we selected the best model by using the root mean square error (RMSE). Figure 2 shows accuracy model by using plot between DHF actual and DHF prediction.

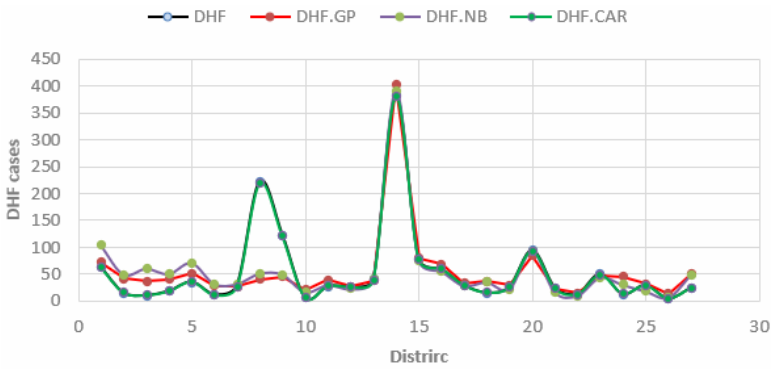

Figure 2. Plot of DHF actual and DHF predictions.

Based on Figure 2, we can see that accuracy of the CAR-BYM prediction for DHF is good (green color line close to black color line). Based on generalized Poisson, negative binomial, and CAR-BYM model, we then computed their root mean square errors. The RMSEs found are 40.95, 40.64 and 1.29, respectively. Due to this, we conclude that the CAR-BYM model is the best model for DHF modeling. Therefore, here we used CAR-BYM model to determine relative risk (RR). The relative risk (RR) values for each district resulting from CAR-BYM model are presented in Table 4. Table 4 gives the detail of RR value in each district. Figure 1 is the presentation of RR value in map.

Table 4. The relative risk (RR) values of each district

No	Districts	Relative risk	No	Districts	Relative risk
1	Banjarsari	0.76	14	Ciamis	3.23
2	Banjaranyar	0.33	15	Baregbeg	1.47
3	Lakbok	0.16	16	Cikoneng	0.91
4	Purwadadi	0.39	17	Sindangkasih	0.48
5	Pamarican	0.44	18	Cihaurbeuti	0.26
6	Cidolog	0.47	19	Sadananya	0.57
7	Cimaragas	1.38	20	Cipaku	1.18
8	Cijeungjing	3.43	21	Jatinagara	0.72
9	Cisaga	2.70	22	Panawangan	0.21
10	Tambaksari	0.26	23	Kawali	0.97
11	Rancah	0.41	24	Lumbung	0.38
12	Rajadesa	0.43	25	Panjalu	0.52
13	Sukadana	1.37	26	Sukamantri	0.18
			27	Panumbangan	0.34

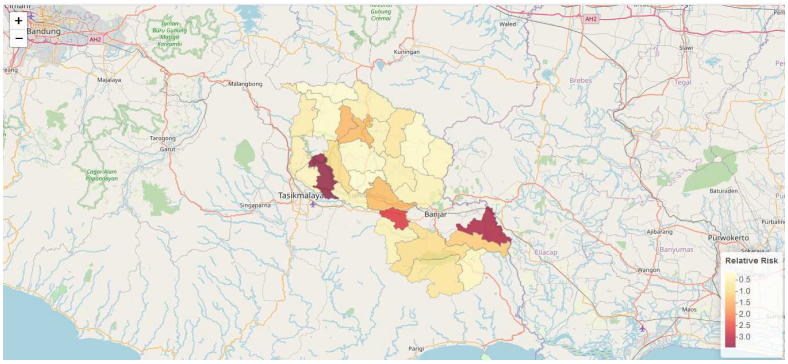

Figure 3. The DHF cases distribution mapping based on RR value in Ciamis Regency.

The values of RR for each district are presented in Figure 3. Based on Figure 3, we found and identified which areas have high, medium, and low RR.

The DHF cases of Cijeungjing and Ciamis districts are 222 and 382, respectively. However, the RR of Cijeungjing district (3.43) is higher than the RR value of Ciamis district (3.22). These are two districts having exceedingly high number of DHF cases compared to average RR of the Ciamis Regency. Table 4 has also shown that the most of its RR values are below 1. This means that the effect of high DHF in Ciamis Regency occurs in certain districts having high RR values.

Figure 3 shows the distribution of RR values for each district. The dark color indicates a high risk area of DHF and light color indicates a low risk area of DHF.

4. Conclusion

The CAR-BYM model is the best model that can be used for DHF case modeling in Ciamis Regency. Based on factors studied, it is noted that if both the number of health workers and altitude are increased, then the DHF cases decrease. Also, it is obtained that if the population density is increased, then the DHF cases increase.

The best model for modeling the DHF cases in Ciamis Regency is the CAR-BYM model. The relative risk for each area is determined by using the CAR-BYM model. The relative risk is directly proportional to the number of DHF cases.

The highest relative risk value is that of Cijeungjing district and the lowest relative risk value is that of Lakbok district. Lakbok district is at the lowest risk of transmission and Cijeungjing district is at the highest risk of transmission.

Acknowledgement

The authors would like to express their gratitude to the Ministry of Research and Technology and Higher Education for their funding for this research through Research Grant BLU UNSOED year 2021.

References

[1] T. A. Nelson and B. Boots, Detecting spatial hot spots in landscape ecology, Ecography 31(5) (2008), 556-566.

[2] J. Aldstadt and A. Getis, Using AMOEBA to create a spatial weights matrix and identify spatial clusters, Geographical Analysis 38(4) (2006), 327-343.

[3] Jajang, A. Saefuddin, I. W. Mangku and H. Siregar, Comparing performances of WG, WGnew and WC on dynamic spatial panel model by Monte Carlo simulations, Far East J. Math. Sci. (FJMS) 90(1) (2014), 17-34.

[4] Y.-J. Cheng et al., Geographical information systems-based spatial analysis and implications for syphilis interventions in Jiangsu province, People's Republic of China, Geospatial Health 7(1) (2012), 63-72.

[5] P. Zhi-Hang et al., Spatial distribution of HIV/AIDS in Yunnan province, People's Republic of China, Geospatial Health 5(2) (2011), 177-182.

[6] X. Han and L. Lee, Bayesian estimation and model selection for spatial Durbin error model with finite distributed lags, Regional Science and Urban Economics 43(5) (2013), 816-837.

[7] J. Hendricks and C. Neumann, A Bayesian approach for the analysis of error rate studies in forensic science, Forensic Science International 306 (2020), 110047.

[8] R. Srinivasan and P. Venkatesan, Bayesian random effects model for disease mapping of relative risks, Ann. Biol. Res. 5(1) (2014), 23-31.

[9] H. S. Stern and N. Cressie, Posterior predictive model checks for disease mapping models, Stat. Med. 19(17-18) (2000), 2377-2397.

[10] C. Gaetan and X. Guyon, Spatial Statistics and Modeling, Springer, Vol. 90, 2010.

[11] S. Banerjee, B. P. Carlin and A. E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, CRC Press, 2014.

[12] N. A. Cressie, Statistics for Spatial Data, John Wiley and Sons, Inc., New York, 1993.

[13] P. McCullagh and J. A. Nelder, Generalized Linear Models, CRC Press, Vol. 37, 1989.

[14] J. Ibrahim, M. Chen and D. Sinha, Bayesian survival analysis, Springer Series in Statistics, Springer, NY, New York, 2001.