Spatial Prediction of Malaria Risk with Application to Bandung City, Indonesia

IGNM. Jaya, Y. Andriyana, B. Tantular

Abstract— **Identifying high-risk areas and understanding the spatial distribution of malaria is crucial for developing an early warning system for preventing, controlling, and providing targeted antimalaria interventions. Global Moran's I has been used to analyzing spatial dependence and Bayesian hierarchical models to figure out the spatial distribution and high-risk areas of malaria disease at sub-district level in Bandung city, Indonesia. We used malaria incidence data collected in 2018 for 30 sub-districts, and considered two risk factors: altitude and healthy behaviors. We evaluated 24 models by combining four likelihoods: Poisson (P), Negative Binomial (NB), Zero Inflated Poisson (ZIP), and Zero Negative Binomial (ZNB) distributions; two spatially-structured: Besag, York, and Mollié (BYM) and Leroux priors and three hyperparameter distributions: Half Cauchy (HC), Uniform (U) and Inverse Gamma (IG). Using deviance information criteria (DIC), Watanabe Akaike information criteria (WAIC) and Pseudo determination coefficient (R^2), we found that the spatial model with Poisson distribution as the likelihood, BYM for spatial structure prior, and Uniform as the hyperprior distribution had a better fit to explain the relationship between risk factors and malaria relative risk over sub-districts.**

Index Terms— **Bandung city, Bayesian, BYM, Early warning system, Leroux, Malaria**

I. Introduction

Malaria has been known as a mosquito-borne infectious disease that may cause death. It is transmitted to human through the bites of infected female Anopheles mosquitoes [1]. World Health Organization (WHO) mentioned that in 2018, there were an estimated 228 million incidences of malaria worldwide and 405,000 died. In 2008, total funding for controlling and reducing malaria incidences reached an estimated US$ 2.7 billion [1].

The geographical variation of disease burden and limited resources cause disease control and treatment to become challenging [2]. Account for the spatial heterogeneity and focus on high-risk regions is a common strategy used to overcome the limited resources. Hence, identifying spatial distribution and high-risk regions can be a crucial step for developing an early warning system for a better quality of life by preparing an effective and efficient solution to control and prevent disease transmission with a limited budget [3]. Spatial analysis is often used to identify high-risk area and the critical risk factors. Hence, disease control could be more directed.

Weather information and quality of life, such as healthy behaviors are widely used to predict disease risk variation over space [4]. However, for a small area, climate information including temperature, precipitation, and humidity may have a single value for all areas because of the limitation of weather monitoring stations. Altitude is commonly used to represent the weather variation over areas. Altitude and health behaviors information could be used effectively to improve public health awareness in a bid to prevent certain epidemics such as malaria disease. Prevention consists of detection, monitoring, and predicting disease risk that might lead to an early warning system [3].

We provide the spatial analysis to explore the spatial distribution of malaria risk in Bandung city, Indonesia and verify the risk factors which are related to the incidence risk at sub-district level. We assessed two spatial methods to address the spatial variations of malaria disease risk. The first method is descriptive approach using the standardized incidence ratio (SIR), and the second method is model-based approach using Bayesian hierarchical model. We considered using a Bayesian hierarchical Poisson model, including spatially structured and unstructured effects. Spatially structured effects is used to model spatial dependence and spatially unstructured effect for modeling spatial heterogeneity. Bayesian approach in disease mapping study gives some advantages such as its ability to account for uncertainty in the risk estimates and provide exceedance posterior probability which is useful to define high-risk clusters [3].

The rest of paper is structured as follows: In section 2, we describe spatial modelling, including discussion of global Moran's I and incidence risk modelling. Section 3 contains the empirical estimation result on spatial modelling of malaria in Bandung city, Indonesia, while section 4 consists of discussion and conclusions.

Manuscript received September 22, 2020; revised December 20, 2020. This work was supported in part by Department of Statistics Universitas Padjadjaran through research grant No: 1427/UN6.3.1/LT/2020 and supported by South Initiatives Project 2020, SI-2019-01-82

*IGNM. Jaya is a lecturer in Department of Statistics Universitas Padjadjaran, West Java, 45363, Indonesia, e-mail: (mindra@unpad.ac.id).

Y. Andriyana is lecturer in Department of Statistics Universitas Padjadjaran, West Java, 45363, Indonesia, e-mail: (y.andriyana@unpad.ac.id).

B. Tantular is lecturer in Department of Statistics Universitas Padjadjaran, West Java, 45363, Indonesia, e-mail: (bertho@unpad.ac.id)

II. Spatial modelling for developing an early warning system

A. Global Moran's I

Identify the high-risk clusters of disease risk is the crucial step in developing an early warning system. The disease clustering can be determined by evaluating the spatial dependence of disease risk across areas. Investigation of spatial dependence across sub-districts is critical in spatial modelling to provide information on whether or not the disease transmission is caused by area proximity. Disease clustering provides information on high-risk clusters. Moran's I is widely used for identifying global spatial autocorrelation in order to detecting spatial clustering of disease risk [5].

Define y_i denotes the number of incidences and N_i as the population at risk at spatial unit i $(i = 1, \dots, n)$ with n denotes the number of spatial observations (e.g., sub-districts). Moran's I is defined as follows [6]:

$$I = \frac{1}{s_y^2} \frac{\sum_{i=1}^{n} \sum_{\{j:i \neq j\}} w_{ij} \left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\sum_{i=1}^{n} \sum_{\{j:i \neq j\}} w_{ij}} \qquad (1)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, i and j are the sub-district indexes and w_{ij} is the element of spatial weight matrix **W** and defined as:

$$w_{ij} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ are adjacent neighbors} \\ 0 \ \text{ otherwise} \end{cases}$$

This study considered queen adjacency spatial weight matrix **W**. Moran's I coefficient takes values from -1 to 1. Zero value indicates that there was no global spatial autocorrelation or no spatial cluster. The strong spatial cluster is indicated by positive and large value the Moran's I coefficient.

B. Standardized incidence ratio

The disease risk is widely measured by relative risk (RR) that is the ratio of number of incidence and expected number of incidence. The expected number of incidence is a function of population at risk and/or demographic (e.g., age and gender). Let y_i denotes the number of incidences at spatial unit *i* which is count data with mean μ_i [8]. In order to evaluate the disease risk over spatial units and accommodate different structure in population at risk such as age-and sex, μ_i is commonly defined by product of relative risk (θ_i) and expected disease incidence (E_i). E_i is calculated by accounting the population at risk structure. However, the detail information of age and gender were often not available, then it is defined as $E_i = N_i(\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} N_i)$. The simple way to calculate relative risk is:

$$\hat{\theta}_i = \frac{y_i}{E_i} \qquad (2)$$

and its asymptotically standard errors (SE) is:

$$SE\left(\hat{\theta}_i\right) = \frac{\sqrt{y_i}}{E_i} \qquad (3)$$

The equation (2) is known as the crude relative risk and it is a kind of descriptive approach. The crude risk SIR is often used in disease mapping but it has many drawbacks. First, SIR is unreliable for a small area level because it tends to be high for the small expected count [9]. Second, SIR cannot be used to model the risk factors [3].

2.3. Loglinear model

To overcome the drawback of SIR, model based approach is used. Using model based approach it is possible to account the risk factors in estimating the relative risk which is modeled through mean function. For count data, generalized linear model (GLM) i.e., log-linear model could be used. iven the Knumber of risk factors, the loglinear model of the mean μ_i is presented as [3, 8, 9]:

$$\begin{aligned} \log \mathbb{E}(y_i|\mathbf{x}) &= \log(\mu_i) \\ &= \text{offset}(\log E_i) + \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i; \\ &\qquad\qquad i = 1, \dots, n \end{aligned} \qquad (4)$$

offset$(\log E_i)$ explains the regression coefficient of $\log E_i$ is constrained to 1, $\mathbf{x}_i = (1, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iK})'$is a $(K+1) \times 1$ is vector of K number of risk factors, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)'$ is $(K+1) \times 1$ vector of regression coefficients, and the last components ϵ_i is random effect component. It is used to account the spatially structured and unstructured effects.

C. Distribution for disease count

There are two distributions were commonly used to model disease count data. First Poisson distribution and second Negative Binomial distribution. The Poisson distribution of y_i is defined as [3, 8, 9]:

$$y_i|E_i, \theta_i \sim Poisson(E_i\theta_i); \ i = 1, \dots, n \qquad (5)$$

Poisson distribution assumes the equality of the mean and variance $\mathbb{E}(y_i) = Var(y_i) = E_i\theta_i$. The Poisson regression model is similar to equation (4) [3]. However, the equality assumption may be violated because of spatial dependency or heterogeneity. When the variance is larger than the mean is known as overdispersion (i.e., $Var(y_i) > \mathbb{E}(y_i)$). The violation of the equality assumption lead to bias standard errors estimates and cause test statistics become wrong. Negative Binomial distribution was used to model the overdispersion. The idea behind the negative binomial regression model is by including second parameter ε_{it} into the Poisson distribution to model the variance of y_i with ε_t follows a Gamma distribution. The equation (5) can be extended becomes Eq. (6), [3, 10]:

$$\begin{aligned} y_i|E_i\theta_i, \varepsilon_i &\sim \text{Poisson}(E_i\theta_i\varepsilon_i) \text{ and} \\ \varepsilon_i|\varrho &\sim \text{Gamma}(\varrho, \varrho), \end{aligned} \qquad (6)$$

for $y_i = 0,1,2, \dots$ and $\varrho > 0$. The Poisson-Gamma mixture distribution of y_i is [3]:

$$\begin{aligned} p(y_i|E_i, \theta_i, \varepsilon_i) &= \text{Gamma}(\varrho, \varrho)\text{Poisson}(E_i\theta_i) \\ &= \left(\frac{\varrho^{\varrho}(\varepsilon_i)^{\varrho-1}\exp(-\varrho\varepsilon_i)}{\Gamma(\varepsilon_i)}\right)\left(\frac{\exp(-E_i\theta_i)(E_i\theta_i)^{y_i}}{y_i!}\right). \end{aligned} \qquad (7)$$

By integrating out ε_i we get the marginal probability of y_i which is a Negative Binomial (NB) distribution [10]:

$$\begin{aligned} p(y_i|E_i\theta_i, \varrho) \\ = \frac{\Gamma(y_i + \varrho)}{\Gamma(y_i + 1)\Gamma(\varrho)}\left(\frac{E_i\theta_i}{E_i\theta_i + \varrho}\right)^{y_i}\left(\frac{\varrho}{E_i\theta_i + \varrho}\right)^{\varrho}. \end{aligned} \qquad (8)$$

The mean and variance of NB distribution are $\mathbb{E}(y_i) = E_i\theta_i$ and $\text{Var}(y_i) = E_i\theta_i + (E_i\theta_i)^2/\varrho$ respectively, with ϱ is the additional Poisson variation parameter. The NB distribution has greater variance than the its mean, so that NB is appropriate distribution to deal with overdispersion. If $\varrho \longrightarrow \infty$, the NB distribution is just a Poisson distribution. Although NB is appropriate to be used to model data with overdispersion, it is not good enough to use if the model involves a lot of zero data. Large number of zeros in the data cause the Poisson and NB regression models are not appropriate to be used to predict the zero observations [11, 12]. The excess zeros are also a source of overdispersion [13]. Zero-inflated (ZI) models were used to model excess zeros [14]. It is specified as:

$$y_i \sim \begin{cases} f(y_i) & \text{with probability } 1-\pi_i \\ 0 & \text{with probability } \pi_i \end{cases} \tag{9}$$

We considered Zero Inflated Poisson (ZIPP) and Zero Inflated Negative Binomial (ZINB) for $f(y_i)$ and π_i the zero-inflation probability. By assuming that the θ_i and ε_i are the relative risk and overdispersion parameters of the NB distribution respectively, then the Eq. (9) can be defined as [12]:

$$y_i | E_i, \theta_i, \pi_i \sim ZIP(E_i\theta_i, \pi_i) \tag{10}$$

for ZIP distribution, or

$$y_i | E_i, \theta_i, \pi_i, \varepsilon_i \sim ZINB(E_i\theta_i, \pi_i, \varepsilon_i) \tag{11}$$

for ZINB distribution. Note that, $\lambda_i = E_i\theta_i$ and π_i is modeled using canonical link function [12]:

$$\begin{gathered} logit(\pi_i) = \boldsymbol{x}_i'\boldsymbol{\gamma} \\ log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \text{offset}(\log E_i) \end{gathered} \tag{12}$$

where $\boldsymbol{\gamma}$ is a vector regression coefficient of the logit model.

D. Hierarchical models for relative risks

Hierarchical models of relative risk are often used to smoothen the relative risk estimate as well as overcome the overdispersion and excess zeros problems. The model is flexible for smoothing purpose by introducing random effect components. The random effect components may capture the spatial heterogeneity and spatial dependence:

$$\begin{gathered} \log(\lambda_i) = \text{offset}(\log(E_i)) + \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \\ \omega_i + v_i \, ; i = 1, \dots, n \end{gathered} \tag{13}$$

where α in an intercept and describes the overall relative risk, β_1 and β_2 denotes the regression coefficients of latitude (x1) and healthy behaviors (x2). We used intrinsic conditional autoregressive (iCAR) or BYM CAR prior for the spatially structured random effect of region i (ω_i) [15]:

$$\omega_i | \boldsymbol{\omega}_{-i}, \sigma_\omega^2, \mathbf{W} \sim \mathcal{N}\left(\frac{\sum_{j=1}^n w_{ij}\omega_j}{\sum_{i=1}^n w_{ij}}, \frac{\sigma_\omega^2}{\sum_{i=1}^n w_{ij}}\right) \tag{14}$$

where w_{ij} is similar to w_{ij} in equation (1), σ_ω^2 is the variance parameter of ω_i. We also consider Leroux CAR prior [16]:

$$\omega_i | \boldsymbol{\omega}_{-i}, \sigma_\omega^2, \boldsymbol{W} \sim \mathcal{N}\left(\frac{\rho\sum_{j=1}^n w_{ij}\omega_j}{\rho\sum_{j=1}^n w_{ij} + 1 - \rho}, \frac{\sigma_\omega^2}{(\rho\sum_{j=1}^n w_{ij} + 1 - \rho)}\right) \tag{15}$$

where ρ denotes the parameter autoregressive. Note that iCAR is the limiting case of the Leroux prior when ρ equals 1. The spatially unstructured random effect v_i is modeled by an exchangeable prior:

$$v_i | \sigma_v^2 \sim \mathcal{N}\left(0, \frac{1}{\sigma_v^2}\right) \tag{16}$$

where σ_v^2 is the variance parameter of v_i.

We assigned a vague Gaussian prior distribution for α, β_1, and β_2, that is, $\{\alpha, \beta_1, \beta_2\} \sim \mathcal{N}(0, 10^6)$ and $\log(\rho/(1-\rho)) \sim \mathcal{N}(0,100)$ [12]. For each variance, parameters are commonly assigned by inverse Gamma prior. However, it is too sensitive to the variation of the hyperprior parameters [17]. Here we use half Cauchy (HC) prior for standard deviation of each random components. Gelman (2006) [17] proposed 25 as scale parameter for the HC hyper-prior. To select the best prior and hyperprior distributions, we use model selection techniques. We consider deviance information criterion (DIC), Watanabe Akaike information criterion (WAIC), and the pseudo coefficient of determination (R^2). We fit the model in equation (4) using Integrated Nested Laplace Approximation (INLA) [18]. Bayesian exceedance probability [19, 20, 21] is computed to test the significant high-risk areas namely hotspots. The models were estimated using R-INLA packages.

INLA works in three hierarchies: the first is the likelihood model $p(\mathbf{y}|\boldsymbol{\Phi}, \boldsymbol{\psi})$, where $\mathbf{y}$ is the vector observation, $\boldsymbol{\Phi} = (\alpha, \beta_1, \beta_2, \omega, v, \rho)'$ is a vector parameters and $\boldsymbol{\psi} = (\sigma_v^2, \sigma_v^2)'$ denotes the vector of hyperparameter. The second is defining the latent Gaussian field (GMRF), $p(\boldsymbol{\Phi}|\boldsymbol{\psi}) \sim N\big(\boldsymbol{\mu}(\boldsymbol{\psi}), \boldsymbol{Q}^{-1}(\boldsymbol{\psi})\big)$. The last hierarchy is defining hyperprior distribution of the hyperparameter $\boldsymbol{\psi}$**.** The hyperprior distribution is denoted by $p(\boldsymbol{\psi})$. The posterior marginal of $\boldsymbol{\Phi}_i$ is [22]:

$$p(\boldsymbol{\Phi}_i|\mathbf{y}) = \int_{\boldsymbol{\psi}} p(\boldsymbol{\Phi}_i|\boldsymbol{\psi}, \mathbf{y}) p(\boldsymbol{\psi}|\mathbf{y}) d\boldsymbol{\psi}$$

All parameters models are approximated by INLA using the finite sum:

$$\tilde{p}(\boldsymbol{\Phi}_i|\mathbf{y}) = \sum_j \tilde{p}\big(\boldsymbol{\Phi}_i|\boldsymbol{\psi}^{(j)}, \mathbf{y}\big)\tilde{p}\big(\boldsymbol{\psi}^{(j)}|\mathbf{y}\big)\,\Delta^{(j)} \tag{18}$$

where $\tilde{p}\big(\boldsymbol{\Phi}_i|\boldsymbol{\psi}^{(j)}, \mathbf{y}\big)$ and $\tilde{p}\big(\boldsymbol{\psi}^{(j)}|\mathbf{y}\big)$ denote approximation of $p\big(\boldsymbol{\Phi}_i|\boldsymbol{\psi}^{(j)}, \mathbf{y}\big)$ and $p\big(\boldsymbol{\psi}^{(j)}|\mathbf{y}\big)$. Equation (18) is evaluated at support grid points $\boldsymbol{\psi}^{(j)}$ using suitable weights $\Delta^{(j)}$.

III. Result and Discussion

A. Data exploration

We used malaria incidences data in 2018 in 30 sub-districts of Bandung city. To estimate the relative risk (RR) and identify the high-risk clusters for developing an early warning system, a number of population is needed to adjust the population at risk variation over space. To facilitate the ecological hypothesis, two covariates altitude and healthy behaviors were used. The altitude was used as a representation of the weather variables such as temperature, humidity, and precipitation which have a high correlation with the altitude (Kazembe, 2007). The data presented in Table I was obtained from http://data.bandung.go.id/. The average of number of incidences per sub-districts for the entire city was 1 (range: 1-3), and the average of the population at risk was 83,457 (range: 24,929-13,6351) with overall mean incidence rate was 1 (range: 0-3) episodes per

100,000 persons in 2018. In 2018, the highest number of incidences of malaria were found in Buah Batu, Kiara Condong, and Sukajadi sub-districts. We observed that there were 14 (46.6%) sub-districts which have zero incidences. The excess zeros incidences may lead to overdispersion problem.

TABLE I
RESEARCH VARIABLES

id	Sub-District	Malaria Incidences	Population at Risk	Altitude	Population Density	Healthy Behaviors
1	Andir	1	106,498	733.000	20.759	56.900
2	Antapani	1	77,293	690.000	21.859	71.540
3	Arcamanik	0	73,801	680.000	8.742	62.720
4	Astanaanyar	1	76,911	695.000	23.377	58.560
5	Babakan Ciparay	0	136,254	697.000	21.700	78.840
6	Bandung Kidul	0	59,433	670.000	13.036	80.130
7	Bandung Kulon	0	136,351	709.000	21.022	56.990
8	Bandung Wetan	1	32,331	751.000	8.248	67.320
9	Batununggal	2	121,886	682.000	22.901	53.370
10	Bojongloa Kaler	0	126,477	694.000	36.286	76.730
11	Bojongloa Kidul	1	86,981	689.000	16.657	68.960
12	Buahbatu	3	100,984	670.000	13.861	67.360
13	Cibeunying Kaler	0	70,926	750.000	13.992	72.620
14	Cibeunying Kidul	0	113,885	706.000	23.612	75.730
15	Cibiru	1	73,312	706.000	12.268	81.040
16	Cicendo	1	97,903	700.000	12.791	89.150
17	Cidadap	1	54,401	848.000	7.353	69.450
18	Cinambo	0	24,929	677.000	6.578	45.440
19	Coblong	1	115,720	792.000	17.561	54.260
20	Gedebage	1	39,167	666.000	3.887	69.410
21	Kiaracondong	3	127,738	760.000	21.902	53.720
22	Lengkong	0	74,753	696.000	11.151	60.990
23	Mandalajati	0	69,283	760.000	16.039	53.270
24	Panyileukan	0	39,059	675.000	8.902	88.390
25	Rancasari	1	82,744	670.000	11.611	70.580
26	Regol	0	85,383	686.000	16.805	71.210
27	Sukajadi	3	103,390	891.000	20.254	61.480
28	Sukasari	1	75,672	856.000	12.637	38.790
29	Sumur Bandung	0	37,114	712.000	9.236	65.580
30	Ujungberung	0	83,130	698.000	13.764	63.570

Figures 1-2 presents the choropleth maps of the research variables that we used in this study.

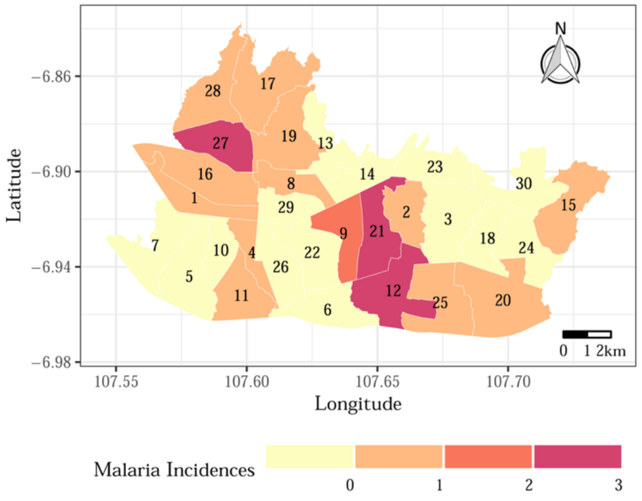
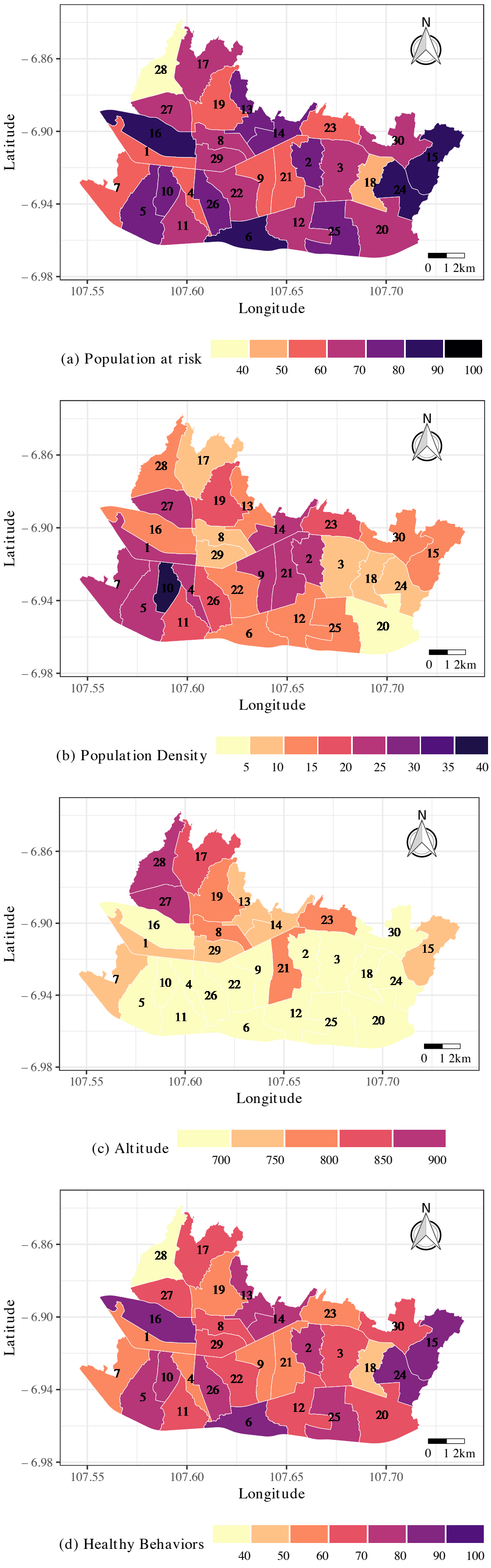


Fig 1. Spatial distribution of malaria incidences

Figure 1 presents the spatial distribution of malaria incidences in Bandung city, Indonesia. The high incidences were concentrated in northern and central Bandung which have high population density.



(a) Population at risk

(b) Population Density

(c) Altitude

(d) Healthy Behaviors

Fig. 2. (a) Population at risk, (b) Population density, (c) Altitude, and (d) Healthy behaviors

Figure 2 shows the spatial distribution of variables interest including population at risk, population density, altitude, and healthy behaviors. Population at risk is used to calculate the expected count (E_i). Two variables were considered as risk factors: altitude and healthy behaviors. The population density was not included directly in the analysis because the population variable was used to calculate E_i. It will be used to validate the spatial distribution of incidence risk or the relative risk.

B. Spatial dependence

The spatial dependence in the number of malaria incidence was evaluated by mean global Moran's I.

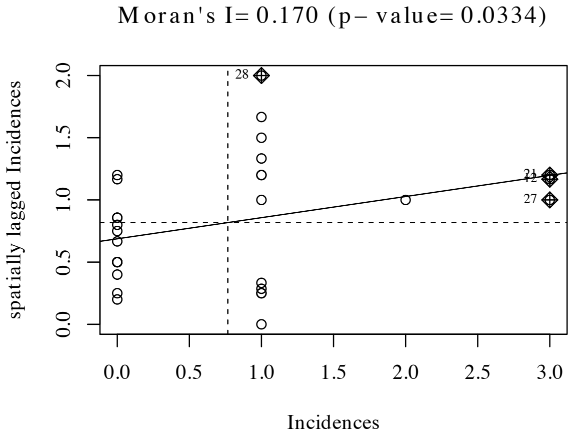


Fig. 3. Moran's Index plot

Moran's I statistics of 0.170 with (p-value<0.05) indicates there is a significant spatial dependency. Figure 3 shows the Moran's I scatter plot of malaria incidence over 30 sub-districts. Points are place in quadrant I shows that clusters with high malaria incidence was surrounded by clusters of high malaria incidences. The plots show more cluster points in quadrant I. This information can be used as an initial reference that malaria is spreading due to proximity to locations.

C. Bayesian Hierarchical Model

To estimate the relative risk of malaria, we applied Bayesian hierarchical model and compared the result with the crude SIR. A total of 24 models were estimated considering four distributions (i.e., Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial), two different spatial prior (BYM, and Leroux), and three different hyperprior distributions (half Cauchy, Uniform and Inverse Gamma). The model comparison based on DIC, WAIC, and R^2 are presented in Table 2.

TABLE II
MODEL COMPARISON MEASURES FOR VARIOUS SPATIAL MODELS FITTED IN THE STUDY

No.	Model	Spatial Prior	DIC			WAIC			R^2		
			HC	U	IG	HC	U	IG	HC	U	IG
1	P	BYM	70.95	70.96	70.00	69.12	69.11	69.71	0.51	0.51	0.12
2	NB		71.21	71.27	70.09	69.25	69.27	69.74	0.51	0.51	0.12
3	ZIP		71.52	71.54	70.81	69.36	69.36	70.24	0.31	0.31	0.07
4	ZINB		71.81	71.81	70.98	69.45	69.45	70.35	0.33	0.33	0.06
5	P	Leroux	70.37	70.30	70.00	69.20	69.12	69.71	0.42	0.43	0.12
6	NB		70.42	70.42	70.08	69.28	69.28	69.74	0.40	0.40	0.12
7	ZIP		70.87	70.93	70.67	69.56	69.61	70.14	0.22	0.23	0.07
8	ZINB		71.22	71.22	70.68	69.71	69.71	70.12	0.23	0.23	0.07

Model: P (Poisson); NB (Negative Binomial); ZIP (Zero inflated Poisson); ZINB (Zero inflated Negative Binomial)

Table II presents the model comparison measures for the spatial models. Based on the DIC, WAIC, and R^2 values, the model (M1), a model with Poisson likelihood, BYM spatial prior, and Uniform prior had a better fit. This model explains 51% of the total variation.

TABLE III
POSTERIOR ESTIMATES FOR THE FIXED AND RANDOM EFFECTS OF THE BEST SPATIAL MODEL OF MALARIA INCIDENCE RISK AT SUB-DISTRICT LEVEL IN BANDUNG CITY, INDONESIA

Parameters	Coefficient	SE	Relative risk (RR)	95% CI
Intercept (β_0)	–3.097	4.476	0.045	0.000; 296.190
Altitude (β_1)	0.005	0.005	1.005	0.995; 1.015
Healthy Behavior (β_2)	–0.010	0.024	0.900	0.995; 1.005
σ_ω^2	0.462	0.715	0.600;1.477	1.012; 9.545
σ_v^2	0.443	0.712	0.846;1.532	1.012; 10.014
$\sigma_\omega^2/(\sigma_\omega^2+\sigma_v^2)$	0.510			

Estimates for covariates (altitude and healthy behaviors) after considering for spatially structured in the best model are given in Table III. The malaria incidence risk was positively associated with altitude (RR: 1.005, 95% CI: 0.995–1.015). Inversely, the incidence was related negatively with healthy behaviors (RR: 0.990, 95% CI: 0.995–0.1005). Both associations were not statistically significant. But, the sign of the coefficients was reasonable. It indicates that the number of malaria incidences could be reduced by increasing healthy behaviors. The proportion variance of spatially structured is about 0.510, and spatially unstructured is of 0.490, which indicates that the variation of malaria was explained similarly by spatially dependence and heterogeneity. The spatially structured and unstructured effects for each sub-district were presented in Figures 4(a)-4(b).

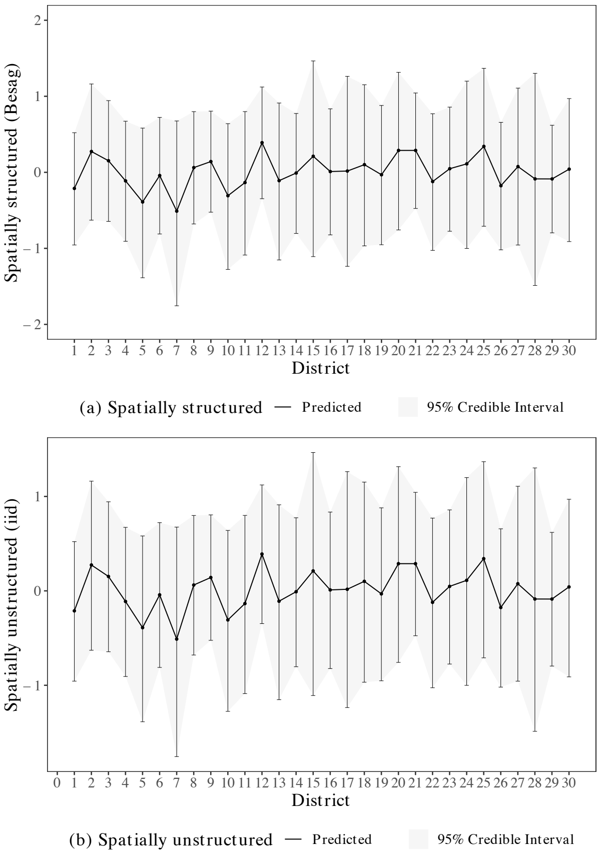
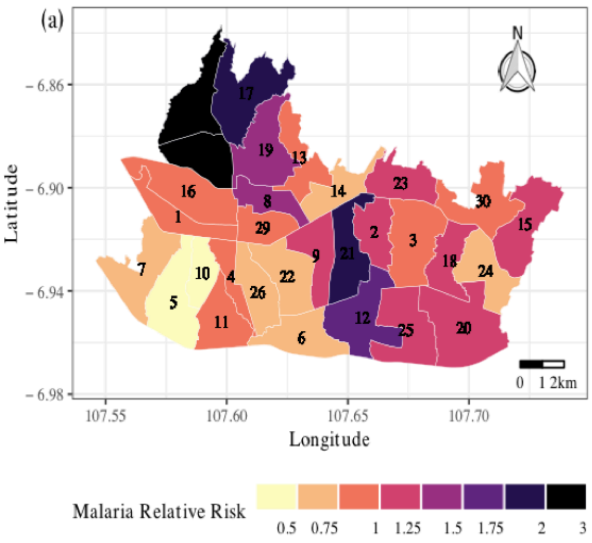


Fig. 4. (a) Spatially structured effects and (b) spatially unstructured effects



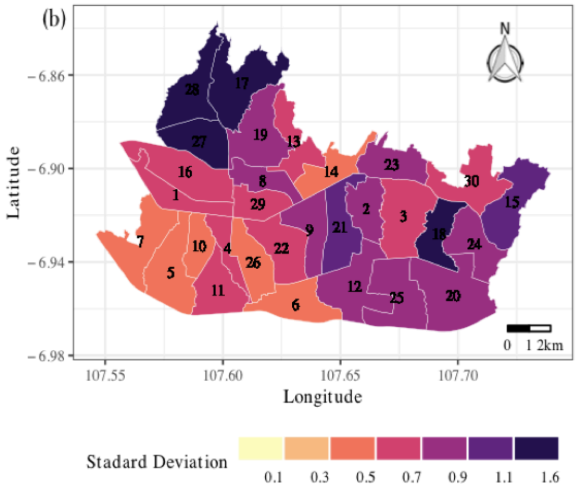
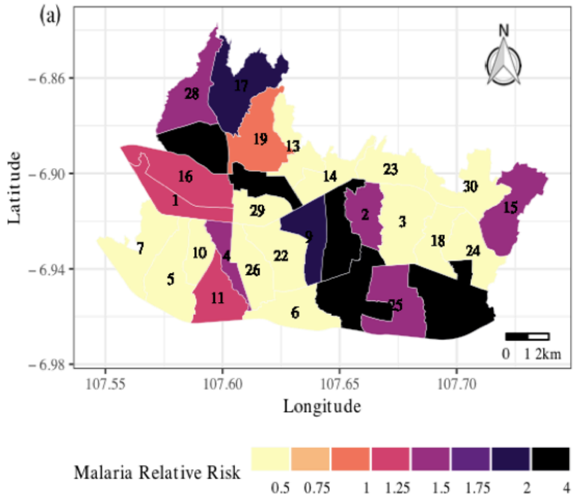


Fig. 5. (a) Malaria relative risk and (b) standard deviation based on SIR Model

Figures 5(a)-5(b) shows the relative risk and its standard deviation maps based on the SIR estimator in Eqs. 2-3. The SIR ranges from 0 to 3.370, with zero relative risks for most eastern and southwest regions.



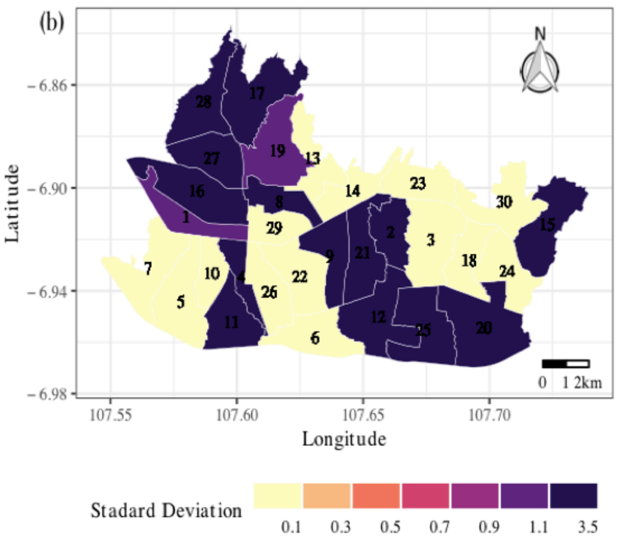


Fig. 6. (a) Malaria relative risk and (b) standard deviation based on BYM Model

Figures 6(a)-6(b) present the relative risk and its standard deviation maps based on the BYM model. The BYM ranges from 0.460 to 2.835 with relatively small relative risk most at eastern and southwest regions. The smoothed map gives homogeneous RR, which is easy to interpret. The relative risk estimates based on BYM has shrunk toward the overall means, compared to the crude SIR because of the smoothing effect of the spatial dependence and heterogeneity. The smoothing effect is also described by the plots of BYM versus crude SIR (Fig. 7).

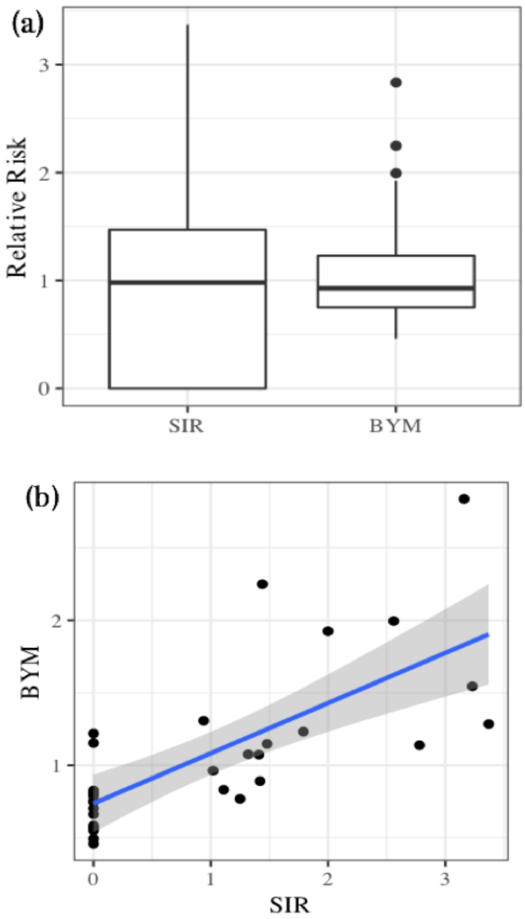


Fig. 7. (a) Boxplot SIR versus BYM and (b) scatter plot SIR versus BYM

Figure 7(a) presents the comparison between crude relative risk SMR and smoothness relative risk BYM. BYM has a smaller variation than SMR, but scatterplot (Figure 7(b)) shows that the BYM and SMR have high correlation which indicates that the smoothness does not produce underestimated estimation of the relative risk. The corresponding BYM map gives posterior probabilities of RR > 1 (Figure 8). From the map, we have three sub-districts which were categorized as hotspot areas with a posterior exceedance probability greater than 0.80. The districts are Kiaracondong, Sukajadi, and Sukasari with posterior probabilities of 0.865, 0.935, and 0.806, respectively (see Table IV). The high-risk clusters were observed mostly in the northwest and southeast of Bandung.

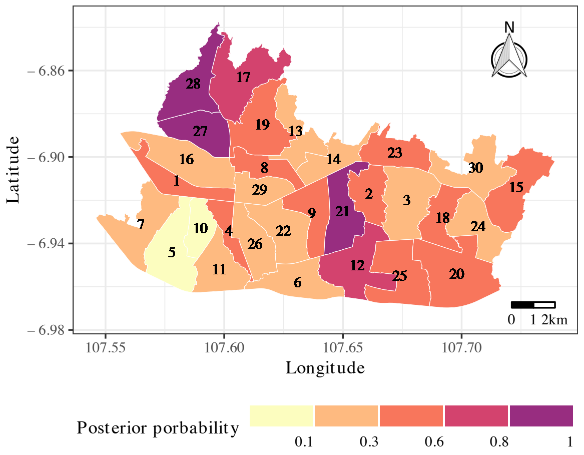


Fig. 8. Posterior probability of relative risk based on BYM

Figure 8 shows that the hotspots of malaria risk in 2018 were located at northern and central Bandung, Indonesia.

TABLE IV
ESTIMATION OF BYM AND POSTERIOR PROBABILITY OF MALARIA DISEASES IN BANDUNG CITY, INDONESIA, 2018

Id	Sub-District	BYM			Posterior Probability
		Mean	SE	95% CI	
1	Andir	0.903	0.560	0.217; 2.331	0.328
2	Antapani	1.074	0.714	0.233; 2.912	0.430
3	Arcamanik	0.808	0.543	0.177; 2.211	0.261
4	Astanaanyar	0.891	0.607	0.189; 2.461	0.315
5	Babakan Ciparay	0.460	0.349	0.079; 1.372	0.070
6	Bandung Kidul	0.586	0.456	0.104; 1.781	0.133
7	Bandung Kulon	0.550	0.438	0.084; 1.698	0.119
8	Bandung Wetan	1.284	0.850	0.303; 3.480	0.540
9	Batununggal	1.232	0.741	0.295; 3.101	0.541
10	Bojongloa Kaler	0.493	0.364	0.090; 1.442	0.082
11	Bojongloa Kidul	0.769	0.541	0.154; 2.171	0.238
12	Buahbatu	1.545	0.854	0.425; 3.679	0.709
13	Cibeunying Kaler	0.827	0.619	0.150; 2.444	0.272
14	Cibeunying Kidul	0.663	0.445	0.142; 1.810	0.168
15	Cibiru	1.147	0.935	0.158; 3.601	0.439
16	Cicendo	0.832	0.656	0.124; 2.546	0.281
17	Cidadap	1.924	1.390	0.337; 5.520	0.730
18	Cinambo	1.220	1.254	0.125; 4.534	0.407
19	Coblong	1.307	0.753	0.339; 3.199	0.590
20	Gedebage	1.138	0.887	0.197; 3.464	0.434
21	Kiaracondong	1.995	1.000	0.608; 4.441	0.865
22	Lengkong	0.704	0.509	0.139; 2.030	0.197
23	Mandalajati	1.153	0.782	0.243; 3.171	0.470
24	Panyileukan	0.748	0.726	0.084; 2.671	0.222
25	Rancasari	1.076	0.746	0.212; 3.004	0.426
26	Regol	0.570	0.407	0.115; 1.630	0.118
27	Sukajadi	2.835	1.536	0.742; 6.603	0.935
28	Sukasari	2.248	1.530	0.410; 6.156	0.806
29	Sumur Bandung	0.755	0.522	0.170; 2.114	0.224
30	Ujungberung	0.788	0.551	0.160; 2.217	0.250

IV. Discussion and Conclusion

We have analyzed and mapped the relative risk of malaria by considering altitude and healthy behaviors as the risk factors.. This study is an application of and spatial regression for epidemiological data at small area [3, 9]. Using a Bayesian spatial hierarchical model, we obtained more homogenous relative risk estimates than crude standardized incidence ratios by smoothing the data through BYM CAR model with Uniform hyperprior distribution [16]. The smoothing model is more comfortable to interpret where the posterior estimate has high specificity and low sensitivity [21]. This is an important property to avoid false-positive, thereby predicting true clusters in the maps. The BYM CAR model of malaria risk gives a more reliable estimate of the relative risk of disease than the crude estimate SIR [3, 21, 23]. A nonparametric model, such as P-spline may also have a similar advantage. However, it can give a computational challenge [24].

Over smoothing issue becomes an exciting topic in disease mapping study. Several alternatives to CAR models have been proposed to be able to distinguish the high-risk and low-risk clusters better. Allowing the spatial autocorrelation has different values from 1 using BYM model and selecting the appropriate interval of the relative risk on the choropleth maps may also be possible to minimize the over smoothing problem realized by the CAR model. As far as we know, this is the first study to evaluate the spatial distribution of malaria risk in Bandung city, Indonesia. However, there is a limitation on data access. The newest data that we can access is Malaria data in 2018 due to the government regulation that published health profile information after one until two years later. We also realized that under-reported malaria incidences might occur in the community because most people resort to home or community-based care [24]. They usually visit the health facility or modern biomedical care at a health facility if the disease is perceived to be severe or near-fatal [25, 26]. Therefore, the relative risk pattern described in this study describes the risk of severe malaria [24]. The under-reported effect is minimized in this study by defining the sub-district level as units of spatial analysis and using the aggregate data over sub-districts [24]. Additionally, we believe that the availability of health facilities in each sub-district improves the data reported quality.

In this study, we evaluate the effects of altitude and healthy behaviors on malaria risk. Because the study includes a small sample size (30 sub-districts), we only focus on the direction of the effect and avoid the discussion of statistical signification. High malaria risk was generally associated with high altitude and low healthy behaviors. In our study, naturally, high altitude areas are of high malaria transmission increased malaria incidence, while at low altitude, malaria risk is decreased. This result confirmed the research study by Kazembe (2007) [24], and it was similar to the dengue disease incidence in Bandung city, Indonesia. The high-risk dengue was found in high altitude areas which are in northern regions of Bandung city [3]. The consistency of these results with dengue disease is possible because of the similar kinds of vectors. Both diseases are transmitted by the same vector, i.e., mosquitoes. Mosquitoes benefit from rainfall and high-altitude regions which commonly have a high frequency of rainfall [3]. Rainfall provides aquatic environments for mosquitoes breeding sites [27].

The high index of healthy behaviors is corresponding to low malaria risk. Health behaviors are important risk factors that have to get more attention in disease control and prevention [28]. The random effect component of spatial autocorrelation was found to influence the malaria risk. This component provides a characterization of spatial patterns in the data. The Spatial autocorrelation accounts for the relevant unobserved variables [3, 24, 29]. At the same time, spatial regression modelling is essential to consider malaria modelling [29]. This study provides risk maps that could be used for developing an early warning system as guidance for the government to define priority areas for the focusing of limited resources. This is an essential point and future model aimed at a sub-district level for the effectiveness and efficiency of malaria disease prevention and control [2, 24]. Based on the result of our study, we suggest that various risk factors influence the spatial variation in malaria risk in the Bandung city regions. Give more attention to the risk factors that have a high impact on malaria risk could help the government to develop an effective and efficient strategy in controlling malaria spread and reduce the negative effect on society.

The Bayesian hierarchical model that has been used can be extended to identify spatiotemporal clusters by taking into account the space time variation [30]. It is important to explore the spatial evolution of disease transmission.

REFERENCES

[1] WHO, "Malaria," WHO, 13 January 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/malaria. [Accessed 22 June 2020].

[2] L. Kazembe, K, Immo, T. Holtz, and B. Sharp, "Spatial analysis and mapping of malaria risk in Malawi using point referenced prevalence of infection data," International Journal of Health Geographics, vol. 5, no. 41, pp.1-9, 2006.

[3] I. G. N. M. Jaya and H. Folmer, "Bayesian spatiotemporal mapping of relative dengue disease risk in Bandung, Indonesia," Journal of Geographical Systems, vol. 22, no. 1, pp. 105-142, 2020.

[4] X. Wu, Y. Lu, S. Zhou, L. Chen, B. Xu, "Impact of climate change on human infectious diseases: Empirical evidence and human adaptation, " Environment International, vol. 86, pp. 14-23, 2016.

[5] M. Jackson, L. Huang, Q. Xie, and R. Tiwari, "Research a modified version of Moran's I," International Journal of Health Geographics, vol. 9, no. 33, pp. 1-10, 2010.

[6] P. Moran, "Notes on Continuous Stochastic Phenomena." Biometrika, vol. 37, pp. 17-23, 1950.

[7] D. Kang, H. Choi, J. H. Kim, and J. Choi, "Spatial epidemic dynamics of the COVID-19 outbreak in China. International Journal of Infectious Diseases," vol. 94, pp. 96-102, 2020.

[8] I. G. N. M. Jaya, H. Folmer, B. N. Ruchjana, F. Kristiani and A. Yudhie, "Modeling of infectious diseases: a core research topic for the next hundred years," in Regional Research Frontiers - Vol. 2 Methodological Advances, Regional Systems Modeling and Open Sciences, USA, Springer International Publishing, 2017, pp. 239-254.

[9] D. Clayton and J. Kaldor, "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," Biometrics, vol. 43, no. 3, pp. 671-681, 1987.

[10] M. Mohebbi, R. Wolfe, A. Forbes, "Disease Mapping and Regression with Count Data in the Presence of Overdispersion and Spatial Autocorrelation: A Bayesian Model Averaging Approach," International Journal of Environmental Research and Public Health, vol. 11 no. 1, pp. 883-902, 2014

[11] L. Bermúdez, D. Karlis, and I Morillo, "Modelling Unobserved Heterogeneity in Claim Counts Using Finite Mixture Models," risks, vo. 8, no. 10, pp. 1-13, 2020.

[12] M. Blangiardo and M. Cameletti, Spatial and Spatio-temporal Bayesian Models with R – INLA, Chichester: John Wiley & Sons, 2015.

[13] H. Yang, R. Li, R. Zucker, and A. Buu, "Two-stage model for time varying effects of zero-inflated count longitudinal covariates with applications in health behaviour research." Journal of the Royal Statistical Society: Series C, vol. 65, no. 3, pp. 431–444, 2015.

[14] O. Loquiha, N. Hens, L. Chavane, M. Temmerman, N. Osman, C. Faes, and M. Aerts, "Mapping maternal mortality rate via spatial zero-inflated models for count data: A case study of facility-based maternal deaths from Mozambique, " PLoSE ONE, vol. 13, no. 11, pp. 1-21, 2018.

[15] J. Besag, J. York and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," Annals of the Institute of Statistical Mathematics, vol. 43, no. 1, pp. 1-20, 1991.

[16] B. G. Leroux, X. Lei and N. Breslow, "Estimation of Disease Rates in Small Areas: A New Mixed Model For Spatial Dependence," in Statistical Models in Epidemiology, the Environment, and Clinical Trials, New York, Springer, 2000, pp. 179-191.

[17] A. Gelman, "Prior Distributions for Variance Parameters in Hierarchical Models," Bayesian Analysis, vol. 1, no. 3, pp. 515–534, 2006.

[18] H. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." Journal of the Royal Statistical Society, vol.7, no. 2, pp. 319–392, 2009.

[19] A. Lawson, "Hotspot Detection and Clustering, Ways and Means." Environmental and Ecological Statistics, vol. 17, no. 2, pp. 231–245, 2010.

[20] A. Lawson and C. Rotejanaprasert, "Childhood Brain Cancer in Florida, a Bayesian Clustering Approach." Statistics and Public Policy, vol. 1, no. 1, pp. 99–107, 2014.

[21] S. Richardson, A. Thomson, N. Best and P. Elliott, "Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies," Environmental Health Perspectives, vol. 112, no. 9, pp. 106-1025, 2004.

[22] B. Schrodle and L. Held, "Spatio-temporal disease mapping using INLA," Environmetrics, vol. 22, no. 6, pp. 725-734, 2011.

[23] M. Ugarte, B. Ibanez and A. Militino, "Modelling risks in disease mapping," Stat. Meth. Med. Res, vol. 15, pp. 21-35, 2006.

[24] L. Kazembe, "Spatial modelling and risk factors of malaria incidence in northern Malawi. Acta Tropica, vol. 102, pp. 126-137, 2007.

[25] L. Held, M. Hohle and M. Hofmann, "A statistical framework for the analysis of multivariate infectious disease surveillance data," Statistics in Medicine, vol. 5, p. 187–199, 2005.

[26] D. De Savigny, C. Mayombana, E. Mwageni, H. Masanja, "Care-seeking patterns for fatal malaria in Tanzania. Malaria Journal, vol. 3, no. 27, pp. 1-15, 2004.

[27] S. Lindsay, L. Parson and C. J. Thomas, "Mapping the ranges and relative abundance of the two principal African malaria vectors, Anopheles gambiae sensu stricto and An. arabiensis, using climate data.," Proc Biol Sci, vol. 265, no. 1399, p. 847–854, 1988.

[28] E. Kupcewicz, A. Szypulska and A. Doboszyńska, "Positive Orientation as a Predictor of Health Behavior during Chronic Diseases," Int J Environ Res Public Health, vol. 16, no. 18, p. 3408, 2019

[29] Y. Susanti, S. Sulistijowati, H. Pratiwi, T. Liana, "Paddy Availability Modeling in Indonesia Using Spatial Regression." IAENG International Journal of Applied Mathematics, vlol. 45, no. 4, pp. 398-403, 2015.

[30] I. G. N. M. Jaya and H. Folmer, "Identifying Spatiotemporal Clusters by Means of Agglomerative Hierarchical Clustering and Bayesian Regression Analysis with Spatiotemporally Varying Coefficients: Methodology and Application to Dengue Disease in Bandung, Indonesia," Geographical Analysis, pp. 1-51, 2020.

I Gede Nyoman Mindra Jaya received the B.S. degree (S.Si) from Department of Statistics Universitas Padjadjaran in 2003 and Master of Science from Institut Pertanian Bogor in 2009. He starts Doctor of Philosophy, Faculty Spatial Science in Groningen University, Netherlands. He is active in Spatial and Bayesian researches.

Yudhie Andriyana received his Bachelor of Statistics at Universitas Padjadjaran in 2002. He obtained his Master of Mathematics with Specialization in Statistis in 2009 at Tehcnische Universiteit Kaiserslautern, Germany. In the early 2015, obtained his PhD in Statistics at Mathematics Department, KU Leuven, Belgium. For the past five years, he has conducted intensive researches in the area of flexible modellings.

Bertho Tantular received the B.S. degree (S.Si) from Department of Statistics Universitas Padjadjaran in 1998 and Master of Science from Institut Pertanian Bogor in 2009. He starts Doctor of Mathematics at Faculty of Mathematics and Natural Science, Universitas Padjadjaran. He is active in Longitudinal Data and Multilevel Analysis researches