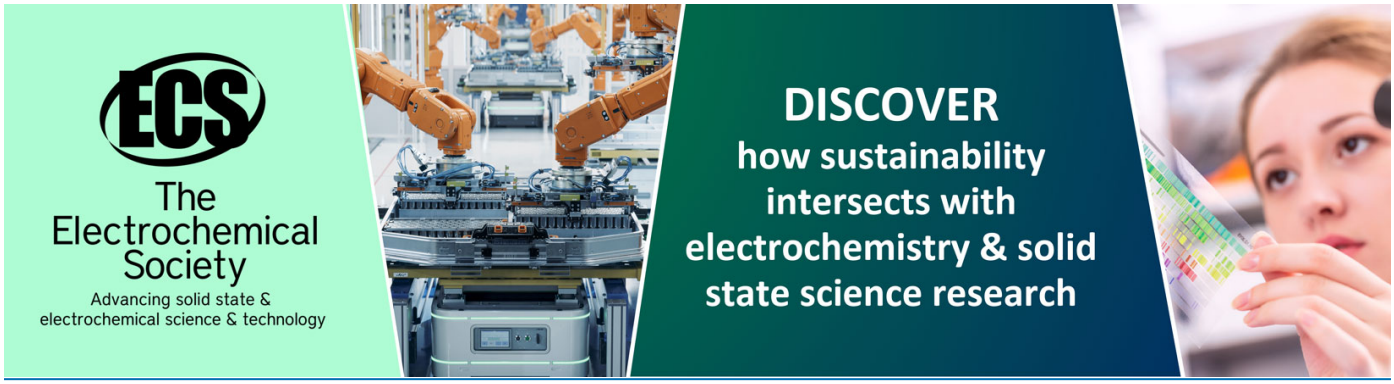
# A spatial shared component random intercept model for assessing risk of diarrhea among men and women





# A spatial shared component random intercept model for assessing risk of diarrhea among men and women

**Kismiantini[1] and I G N M Jaya[2]**

[1] Department of Statistics, Universitas Negeri Yogyakarta, Indonesia
[2] Department of Statistics, Universitas Padjadjaran, Indonesia

E-mail: kismi@uny.ac.id

**Abstract.** Diarrhea becomes a severe problem for children under five years of age. A preventive action is needed to minimize the negative effect of diarrhea. Gender risk assessment may be necessary to control diarrhea transmission as different sexes have distinct healthy behaviour. We develop a collection of candidate models of Bayesian shared component random intercept models to identify a gender group's spatial risk of diarrhea. The candidate models include Poisson-Inverse Gamma and Negative Binomial-Inverse Gamma models with different values hyperprior parameters. The results showed that the Negative-Inverse Gamma model performed better than the Poisson-Inverse Gamma with respect to the existence of overdispersion in the data. The spatial patterns of diarrhea for men and women were similar. In some sub-districts, however, it indicated that women at two different locations had a higher risk of diarrhea compared to men.

## 1. Introduction

Diarrhea is a serious public health issue, especially for children under five years of age. The World Health Organization (WHO) reported that every year approximately 525,000 children under the age of five died of diarrhea [1]. Many cases of diarrhea have been found in low-and-middle-income countries like Indonesia, Philippines, and Thailand. According to United Nations Children's Fund (UNICEF) Indonesia report, 10% of all children under the age of five died of diarrhea in 2016 [2]. In Indonesia, clean water supply and sanitation are the main factors that influence the high incidence of diarrhea in several places [3].

In developing effective strategies for preventing and controlling diarrheal diseases, it is essential to consider the characteristics of diarrhea patients. Women's health website said that women have more issues with excrement than men [4]. In this study, we assume there is a different diarrheal risk for men and women. Thus, we develop a spatial modelling of diarrheal risk by incorporating gender differences.

A spatial disease mapping study is useful to understand the spatial distribution of particular diseases ( [5]; [6]; [7]). This includes a visual representation of high/low-risk areas. It can therefore be used to create a strategy to reduce, prevent, and control infectious disease spread [8]. The spatial disease mapping allows public policy making and resource utilization [5].

A spatial shared component random intercept model is applied to accommodate the spatial autocorrelation between spatial units and share variation between gender. The model has an advantage to produce a more reliable estimation of the diarrhea risk [5].

This paper is organized as follows. Section 2 gives a brief overview of the Bayesian smoothing technique to resolve the unreliable maximum likelihood estimate of relative risk, assign conditionally autoregressive (CAR) prior to the spatial dependency and exchangeable prior to the spatial

heterogeneity, and introduce the shared component model through the Bayesian numerical method. The methodology is illustrated in Section 3 by analyzing and discussing diarrheal cases in the 30 sub-districts of Bandung, 2017. Section 4 contains a conclusion.

## 2. Material and methods

### *2.1 Material*

Data used in this study are diarrheal data obtained from Bandung's health department in 2017. Bandung is the capital city of West Java province in Indonesia. Bandung has 30 sub-districts with a total population reaching 2,497,937 in 2017. A total number of cases of diarrhea is 57,525. The number of cases diarrhea for women is slightly higher than that of men, which is around 50.33% [9].

### *2.2 Method*

To evaluate the spatial distribution of gender-based diarrhea, we develop a Bayesian shared component analysis and take into account the spatial dependency through a random intercept model ( [10], [5] , [11]).

#### *2.2.1 Risk assessment*

Standardized incidence ratio (SIR) is commonly used to evaluate the risk assessment of particular disease. First, we assume that a number of incidences $y_i$ follows a Poisson distribution with parameter $\lambda_i = E_i\theta_i$.

$$y_i \sim Poisson(E_i\theta_i)$$
$$p(y_i|E_i\theta_i) = \frac{\exp(-E_i\theta_i)\,(E_i\theta_i)^{y_i}}{y_i!} \quad (1)$$

where $E_i$ is the expected rate and $\theta_i$ is the parameter of the relative risk in region *i*, $(i = 1, 2, \dots, n)$. The expected rate $E_i$ is commonly defined as:

$$E_i = N_i \times \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} \quad (2)$$

where $N_i$ is the size of the population at risk in region *i*. If the data have a problem of over-dispersion, it is possible to use a Negative Binomial distribution [10]. The Negative Binomial (NB) distribution is defined as follows [12]:

$$p(y_i|E_i\theta_i, \varrho) = \frac{\Gamma(y_i + \varrho)}{\Gamma(y_i + 1)\Gamma(\varrho)}\left(\frac{E_i\theta_i}{E_i\theta_i + \varrho}\right)^{y_i}\left(\frac{\varrho}{E_i\theta_i + \varrho}\right)^{\varrho}. \quad (3)$$

The NB distribution has a mean of $E(Y_i) = E_i\theta_i$ and a variance of $\mathrm{Var}(Y_i) = E_i\theta_i + (E_i\theta_i)^2/\varrho$ where $\varrho$ is an additional Poisson variation parameter.

The maximum likelihood estimator of the standardized incidence ratio can be estimated as follows:

$$\hat{\theta}_i = \frac{y_i}{E_i}. \quad (4)$$

However, the maximum likelihood estimate for the relative risk is known to be unreliable due to small area issues and high variability across districts in population size ([5], [13]). To solve these problems, we use Bayesian smoothing technique based on a pure log-linear model. The pure log-linear model is a model without covariate. In this case, we assume that the relative risk is defined as $\theta_i = e^{f_{fixed}+f_{random}}$ and it can be written as:

$$\log(\theta_i) = \eta_i = f_{fixed} + f_{random}. \quad (5)$$

The $f_{fixed}$ denotes the fixed intercept and $f_{random}$ is the random component including spatial dependency and heterogeneity. The random intercept model can be written as:

$$\eta_i = \alpha + \omega_i + \upsilon_i \tag{6}$$

where $\alpha$ is an overall relative risk; $\omega_i$ and $\upsilon_i$ are spatial dependency and heterogeneity, respectively. A spatial dependency is modelled as an intrinsic Conditional Autoregressive (CAR) prior ( [10], [5], [14]). The CAR prior for the random effect $\omega$ is followed the normal distribution given by:

$$\omega_i|\omega_{j\neq i} \sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\omega_j}{\sum_{j=1}^{n} w_{ij}}, \frac{1}{\kappa_\omega \sum_{j=1}^{n} w_{ij}}\right) \tag{7}$$

where $w_{ij}$ denotes spatial weight between areas $i$ and $j$, $\sum_{j=1}^{n} w_{ij} = n_i$ and $n_i$ is number of neighbours of area *i*. We use a queen spatial contiguity matrix to present the spatial dependence. The $\kappa_\omega$ denotes a precision hyperparameter related to the conditional variance of $\omega_i$ given the values of the other elements of $\omega_j$. A spatial heterogeneity is modelled using an exchangeable prior followed a normal distribution as (i.e. a sequence of random variables that are independent and identically normally distributed (iid)):

$$\upsilon_i|\tau_\upsilon \sim N\left(0, \frac{1}{\tau_\upsilon}\right) \forall i \text{ and } i = 1, \dots, n, \tag{8}$$

where $\tau_\upsilon$ is the precision parameter of $\upsilon_i$ [15].

*2.2.2 Shared Component Analysis*

In most societies, women are more likely to report infected diarrhea than men. To prove this hypothesis, we develop a Bayesian shared component analysis because the risk factors of diarrhea for women and men are similar. The following model formulation presents a typical shared component spatiotemporal for joint groups:

$$\log(\theta_{i1}|\boldsymbol{\vartheta}) = \log(E_{i1}) + \alpha_1 + \varphi_i\gamma + \omega_{i1} + \upsilon_{i1} \tag{9}$$

$$\log(\theta_{i2}|\boldsymbol{\vartheta}) = \log(E_{i2}) + \alpha_2 + \frac{\varphi_i}{\gamma} + \omega_{i2} + \upsilon_{i2} \tag{10}$$

where $\theta_{ik}|\boldsymbol{\vartheta}$ is the expectation of $y_{ik}, (k = 1, 2)$ conditioning on the random effects $\boldsymbol{\vartheta} = (\boldsymbol{\varphi}', \boldsymbol{\omega}', \boldsymbol{\upsilon}')'$, $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_n)'$ is a random effects vector representing shared risk effects common to both groups, and the components $(\omega, \upsilon)$ are spatial dependency and heterogeneity. The elements $\varphi, \omega, \upsilon$ are assumed independent and $\gamma$ is an unknown scale parameter (i.e. a relative weight or a level importance) allowing for different shared 'risk gradients' with respect to each group outcome [16].

*2.2.3 Integrated Nested Laplace Approximation: INLA*

INLA is a Bayesian numerical method consisting three stages processes. The first stage defines the observational function $\boldsymbol{\pi}(\boldsymbol{y}|\boldsymbol{\vartheta})$, where $\boldsymbol{y}$ denotes the number of disease incidences in a vector column. The second stage defines the latent Gaussian Markov random field (GMRF) with a precision matrix $\boldsymbol{Q}$ and the third stage defines controlling hyperparameter model [17]. For the first stage, we assume that the number of disease incidences follows a Poisson distribution $y_i \sim Poisson(E_i \exp(\eta_i))$, where $\boldsymbol{y} = [y_1, \dots, y_n]'$ and $\mathbf{E} = [\mathrm{E}_1, \dots, \mathrm{E}_n]'$. The observational model is given as:

$$\pi(\boldsymbol{y}|\boldsymbol{\eta}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} p(y_i|\eta_i) \propto \sum_{i=1}^{n} [y_i \log(E_i \exp(\eta_i)) - E_i \exp(\eta_i)] \tag{11}$$

In the second stage, the latent GMRF for uncorrelated random effects $\upsilon_i$ follows a normal distribution with a mean of 0 and a variance of $\kappa_\upsilon$, $\upsilon_i \sim N(0, \frac{1}{\kappa_\upsilon})$, where $\kappa_\upsilon$ is an precision hyperparameter of effect $\upsilon_i$. The latent model for spatial dependence component $\omega$ assumes that it follows the Besag, York, and Mollié (BYM) model which proportional to the Gaussian distribution. The probability density function of the spatial structure component $\omega$ can be written as:

$$\pi(\boldsymbol{\omega}|\kappa_\omega) \propto \kappa_\omega^{\frac{n-1}{2}} \exp\left(-\frac{\kappa_\omega}{2}\sum_{i\sim j}(\boldsymbol{\omega}_i - \boldsymbol{\omega}_j)^2\right)$$
$$= \kappa_\omega^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Q}(\omega)\boldsymbol{\omega}\right) \quad (12)$$

Two districts $i$ and $j$ are defined to be neighbour, $i\sim j$, if they are adjacent; $n$ is the number of areas; and $\boldsymbol{\omega}$ are independent zero mean normal with unknown precision parameter of $\kappa_\omega$.

Moreover, we define the inverse of hyperparameters of the hyperprior distribution instead of the hyperparameters ($\kappa_\omega, \kappa_\upsilon, \kappa_\varphi$). The inverse of hyperparameters are the variances parameters ($\sigma_\omega^2 = \frac{1}{\kappa_\omega}, \sigma_\upsilon^2 = \frac{1}{\kappa_\upsilon}, \sigma_\varphi^2 = \frac{1}{\kappa_\varphi}$). We explore two Inverse Gamma (1,1) and (1, 0.00005) distributions as the hyperprior distributions. We use an INLA package in R for analysing the data. The best hyperprior distribution is obtained by Bayesian and classical information criteria such deviance information criterion (DIC), mean absolute error (MAE), root mean square error approximation (RMSEA), and pseudo $R^2$ [10]. The DIC is given by Dbar plus pD where Dbar is the posterior of the deviance and pD is the effective number of parameters. The model with the smallest DIC, MAE, RMSEA, and highest Pseudo $R^2$ is estimated to be the best model.

The steps for the analyses were proposed based on the methods explained as follows:

a. prepare the diarrhea data for each gender in a stacked form, $y_i$,
b. calculate the expected rate $E_i$ using equation 2,
c. estimate the equations of 9 and 10 by using means INLA through R-INLA package,
d. present the relative risk estimates on the choropleth maps.

## 3. Results and discussion

In total, the study observed 57,525 individuals who had a diarrhea of whom 28,571 were male and 28,954 were female in table 1. Highest numbers of cases for both males and females diarrhea were found in the Andir sub-district, while the lowest numbers were found in the Sumurbandung sub-district in figure 1.

**Table 1**. Descriptive statistics of diarrhea by gender

| Gender | Min | Max | Median | Total |
|---|---|---|---|---|
| Men | 392 | 1786 | 912 | 28,571 |
| Women | 433 | 1727 | 909 | 28,954 |

A high number of cases of diarrhea is common in sub-districts situated in the lower left of the two spatial distribution maps in figure 1. In the lower right of the two maps, the number of cases of diarrhea in each sub-district is relatively small.

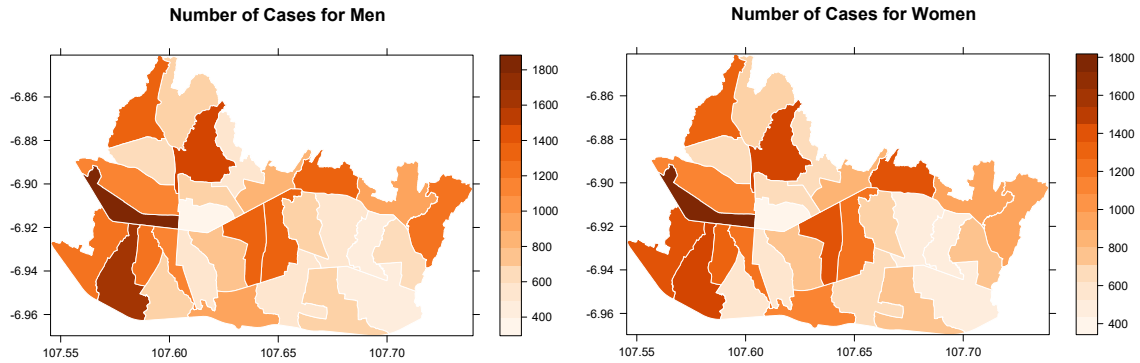



**Figure 1**. A spatial distribution of number of cases diarrhea (a) Men and (b) Women. The x-axis is longitude and y-axis is latitude. A darker color is a larger number of cases of diarrhea and a lighter is a smaller number of cases of diarrhea.

Two Bayesian shared component analyses by means Laplace approximation were developed to identify the spatial distribution of diarrhea. We evaluated Poisson and Negative Binomial distributions with two different hyperprior parameters in the Inverse Gamma distribution in table 2.

**Table 2.** Model comparison

| Model | Model | Hyperprior | Dbar | pD | DIC | MAE | RMSE | Pseudo $R^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Poisson | IG(1,0.00005) | 7037.10 | 1.34 | 7038.45 | 252.05 | 313.38 | 0.65 |
| 2 | Poisson | IG(1,0.00005) | 7037.70 | 1.94 | 7039.64 | 252.05 | 313.39 | 0.65 |
| 3 | Poisson | IG(1,0.00005) | 682.07 | 31.62 | 713.69 | 34.17 | 45.73 | 0.99 |
| 4 | Poisson | IG(1,0.00005) | 682.02 | 31.57 | 713.58 | 34.15 | 45.73 | 0.99 |
| 5 | Negative Binomial | IG(1,0.00005) | 866.09 | 2.02 | 868.11 | 283.17 | 337.58 | 0.65 |
| 6 | Negative Binomial | IG(1,0.00005) | 866.11 | 2.05 | 868.16 | 283.23 | 337.63 | 0.65 |
| 7 | Negative Binomial | IG(1,0.00005) | 866.01 | 1.98 | 868.00 | 283.09 | 337.51 | 0.65 |
| 8 | Negative Binomial | IG(1,0.00005) | 674.42 | 32.38 | 706.80 | 35.11 | 46.70 | 0.99 |
| 9 | Poisson | IG(1,1) | 7037.10 | 1.34 | 7038.45 | 252.05 | 313.38 | 0.65 |
| 10 | Poisson | IG(1,1) | 7038.10 | 2.34 | 7040.43 | 252.05 | 313.39 | 0.65 |
| 11 | Poisson | IG(1,1) | 682.53 | 32.07 | 714.61 | 34.20 | 45.72 | 0.99 |
| 12 | Poisson | IG(1,1) | 682.46 | 32.03 | 714.50 | 34.17 | 45.71 | 0.99 |
| 13 | Negative Binomial | IG(1,1) | 866.09 | 2.02 | 868.11 | 283.17 | 337.58 | 0.65 |
| 14 | Negative Binomial | IG(1,1) | 867.06 | 3.01 | 870.06 | 283.68 | 338.23 | 0.65 |
| 15 | Negative Binomial | IG(1,1) | 866.99 | 2.97 | 869.96 | 283.55 | 338.12 | 0.65 |
| 16 | Negative Binomial | IG(1,1) | 675.75 | 33.31 | 709.06 | 35.16 | 47.24 | 0.99 |

Table 2 shows that Model 8 (Negative Binomial with IG(1,0.00005)) has minimum DIC and maximum Pseudo $R^2$, while its MAE and RMSE are slightly higher than Model 4. We choose Model 8 as the best model.

**Table 3.** Posterior mean of relative risk

| Gender | Min | Max | Mean | Range |
|---|---|---|---|---|
| Men | 0.536 | 1.941 | 1.072 | 1.405 |
| Women | 0.540 | 1.959 | 1.081 | 1.419 |

Table 3 shows the posterior means of relative risk for men and women based on Model 8. The statistics show a comparable relative risk value among group of men and women. The minimum relative risk is nearly 0.5 and nearly 1.95 is the maximum value. It is supported by the relative risk correlation between men and women groups around 0.999, which was clearly presented in figure 2.

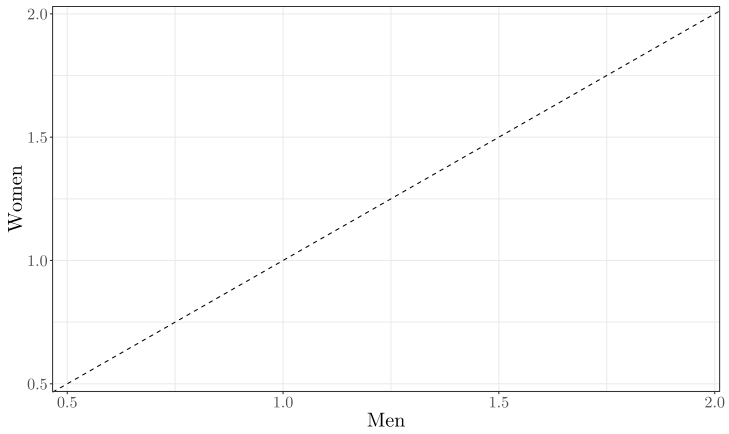



**Figure 2.** Relative risk for men versus women

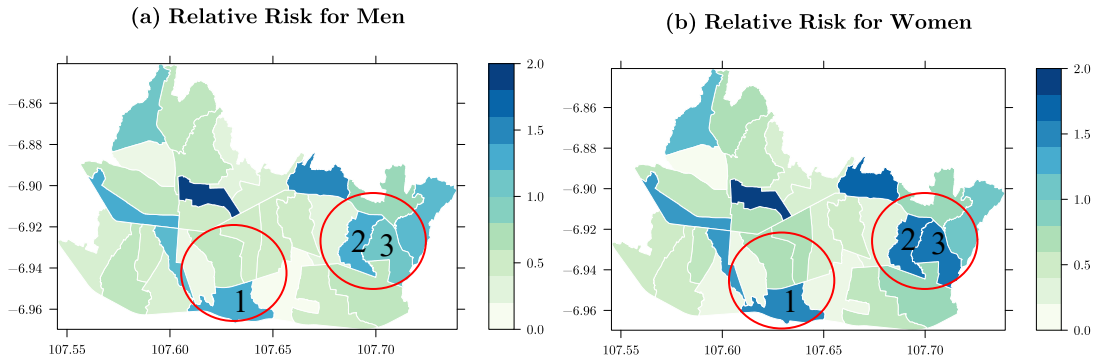



**Figure 3.** Spatial distribution of the relative risk (a) Men and (b) Women. The x-axis is longitude and y-axis is latitude. A darker color is a larger relative risk and a lighter is a smaller relative risk.

Figure 3 shows the relative risk spatial distribution among men and women groups. Highest relative risks for both males and females diarrhea were found in the Bandung Wetan sub-district, while the lowest risks were found in the Sukajadi sub-district. Overall, groups of men and women have comparable spatial relative risk patterns. However, some sub-districts have distinct relative risk values that are indicated by the red circle. Women have a greater relative risk than men in those three sub-districts. The sub-districts are (1) Bandung Kidul, (2) Cinambo, and (3) Panyileukan.

This study conducted spatial and statistical analyses to asses high-low risk diarrhea among men and women. The Bayesian shared component analyses by means Laplace approximation using Poisson and Negative Binomial with Inverse Gamma hyperprior distribution were develop as a new measurement tool to summarize the relative risk of diarrhea in each geographic area. The Negative Binomial model is a better model than the Poisson model because of overcoming overdispersion case. Our findings show that women have a higher risk diarrhea in several locations than men. These results are in line with [18] and [19] findings. The exact reason for a high relative risk of diarrhea in these locations, however, remain unclear. Further study is needed to define the most important diarrhea variables at these locations and decrease the burden of diarrheal diseases.

## 4. Conclusion

Increasing cases of diarrhea over the past few years require public attention. Diarrhea becomes a severe problem especially for children under five years of age. Accurate information on the spatial distribution of diarrheal risk is essential to the development of an early warning scheme for diarrhea. Detailed information based on gender may be required to control the transmission of diarrhea because different gender have different healthy behaviour. We develop Bayesian shared component analysis by means of Laplace approximation and consider the random intercept model. The random intercept model accommodates the spatial dependencies of the data. We observed that men and women have a similar spatial pattern of diarrhea. However, some sub-districts in the south and east have a different value of the relative risk. Women in these sub-districts are at a higher risk. The limitation of this study is that it does not quantify the risk factors due to the limited data available. Future research will be carried out in order to obtain a more accurate and precise model by considering the effects of several risk factors.

## Acknowledgments

We are very grateful for the comments made by the editor and the reviewers as they have contributed to improve an earlier version of this article. We also thank the health office, city of Bandung, for supplying the aggregated dengue and population data per sub-district. The authors thank to the Rector Padjadjaran University (Unpad). This paper is funded by ALG Unpad contract: 2296/UN6.D/KS/2018 and RFU Unpad contract: 1732 d/UN6.RKT/LT/2018.

## References

[1] WHO 2017 Diarrhoeal disease. https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease (Accessed 4 July 2019)
[2] UNICEF 2018 Diarrhoeal disease. https://data.unicef.org/topic/child-health/diarrhoeal-disease/. (Accessed 28 May 2018)
[3] Komarulzaman A, Smith J and de Jong E 2016 *Global Public Health* **12** pp 1141-55
[4] Middleton T 2016 Womens Health. https://www.womenshealthmag.com/health/a19940393/digestive-issues-in-women/. (Accessed 11 July 2019)
[5] Jaya I G N M, Folmer H, Ruchjana B N, Kristiani F and Yudhie A 2017 *Modeling of Infectious Diseases: A Core Research Topic for the Next Hundred Years ed* R Jackson and P Schaeffer (Cham: Springer) pp 239-254
[6] Lawson A B 2013 *Bayesian Disease Mapping, Hierarchical Modelling in Spatial Epidemiology, 2nd ed*. (Boca Raton: Chapman and Hall/CRC)
[7] Kumar S V, Devika S, George S and Jeyaseelan L 2017 *Clin. Epidemiol.Glob. Health* **5** pp 87-96
[8] Christaki E 2015 *Virulence* **6** pp 558-65
[9] Bandung 2018 Bandung Health Profile 2018 (Bandung: Bandung Health Department)
[10] Jaya I G N M and Folmer H (2019) Bayesian Spatiotemporal Mapping of Relative Dengue Disease Risk in Bandung, Indonesia *Manuscript of International Journal of Geographical Analysis* pp. 1-30
[11] Knorr-Held L and Best N G 2001 *J. R. Stat. Soc. Ser A* **164** pp 73-85
[12] Mohebbi M, Wolfe R and Forbes A 2014 *Int. J. Environ. Res. Public Health* **11** pp 883-902
[13] Clayton D and Kaldor J 1987 *Biometrics* **43** pp 671-81
[14] Besag J, York J and Mollié A 1991 *Ann. Inst. Stat. Math* **43** pp 1-20
[15] Rue H and Held L 2005 *Gaussian Markov Random Fields: Theory and Applications* (London: Chapman & Hall)
[16] Schrodle B and Held L 2011 *Environmetrics* **22** pp 725-734
[17] Blangiardo M and Cameletti M 2015 *Spatial and Spatio-Temporal Bayesian Model with R-INLA* (London: John Wiley)
[18] Kandala N B, Ji C, Stallard N, Stranges S and Cappuccio FP 2008 *Ann. Trop. Med. Parasit* **102** pp 427-45
[19] Hirai M, Roess A, Huang C and Graham J 2016 *Global Health Action* **9** pp 32833