Spatial Random Effects Survival Models to Assess Geographical Inequalities in Dengue Fever Using Bayesian Approach: a Case Study

Sri Astuti Thamrin and Irfan Taufik

Statistics, Hasanuddin University, Makassar, Indonesia, 90245

E-mail: tuti@unhas.ac.id

Abstract. Dengue haemorrhagic fever (DHF) is an infectious disease caused by dengue virus. The increasing number of people with DHF disease correlates with the neighbourhood, for example sub-districts, and the characteristics of the sub-districts are formed from individuals who are domiciled in the sub-districts. Data containing individuals and sub-districts is a hierarchical data structure, called multilevel analysis. Frequently encountered response variable of the data is the time until an event occurs. Multilevel and spatial models are being increasingly used to obtain substantive information on area-level inequalities in DHF survival. Using a case study approach, we report on the implications of using multilevel with spatial survival models to study geographical inequalities in all cause survival.

1. Introduction

Dengue hemorrhagic fever (DHF) is an infectious disease caused by dengue viruses, which are spread by Aedes aegypti beats. In a very short period of time, this disease can lead to the death of the patient. Globally, the number of DF cases has increased significantly every year [1]. To day, DHF still becomes a health problem in Indonesia with the increasing number of patient and wider spread of occurrences. Indonesia occupies the highest position in the case of dengue disease in Southeast Asia with 10,000 cases in 2011 [2].

The increasing number of DHF patients relates to the characteristics of the surrounding of their homes such as a ward of sub district. This means that individuals are influenced by their home location and conversely, their location is formed by the individual living in it. Data containing individuals and location is hierarchical structured. Some researches have carried out analysis to such data without taking into consideration the level difference by using linier regression. This method has resulted in some problems such as the dissatisfaction in the analysis results due to the neglecting of other levels. Therefore, [3, 4] proposed a multilevel regression analysis to overcome the problem of hierarchical structured data.

The multilevel model is a regression analysis model taking into account the hierarchy of the data. In this model, the level in a hierarchical structure is defined as a level. The level 1 is the lowest level for example an individual and the level 2 is the higher one such as a location (sub district). A multi level analysis is trying to model the influence of predictor variables measured at the existing levels on response variables measured at the lowest level (level 1). The relationship between these variables with survival times is called the survival analysis.

A survival analysis is about data gathered from the records of a time, achieved by an object where this object has failed to survive. The survival data with hierarchical structured, the survival time of
observed object in a group such as the same sub district tends to have a similar characteristic or correlate each other. This might happen because of the presence of an unobserved covariate. One way to solve this problem is to include this covariate in the model as frailty or random effects. In a multilevel analysis, there are two frailty types, including the level 1 and the level 2. However, when making sub districts into a level 2 types, as this research did so, a spatial random effect needs to be taken into account.

Researches on multilevel survival analysis have been previously carried out. Modelled recurrent infection of the urinary tract by using a residual maximum likelihood (REML) approach [5]. Also used multilevel survival analysis applied on recurrent infections of chronic granulomatous disease (CGD) data by implementing a Bayesian approach [6]. Next, [7] applied the same analysis on recurrent infections of CGD data employing a hierarchical likelihood approach and [8] used a spatial correlation as frailty in survival analysis applied on Minnesota’s death birth data, but not involving frailty at multilevel analysis.

This paper discusses the implication of using multilevel with spatial survival models to study geographical inequalities in all cause DHF survival by using a Bayesian approach. In this approach, the population parameter is considered as a variable with a prior distribution. Before taking samples from a particular population, it is sometimes obtained some information with regard to parameter going to be estimated. This information is then combined with those of the sample used to estimate the population parameter. Given this, a Bayesian method is more flexibly applied due to the additional information to conclude the population’s characteristics.

The paper is organised as follows. We explained introduction in the first section, followed by material and methods. In this second section, we describe the study area, the data collection, the multilevel model, multilevel survival model, multilevel spatial survival model, and Bayesian computation. In the next section, we present the result. The conclusions and the future work are presented further in the last section.

2. Material and Methods

2.1. Study Area

Makassar is a city found in South Sulawesi, Indonesia. It is located at -5.14 latitude and 119.42 longitudes and is situated at elevation 1-25 meters above sea level. Makassar has a population of 1,747,562, making it the biggest city in South Sulawesi. The climate in Makassar City is tropical. There is significant rainfall in most months of the year. The short dry season has little effect on the overall climate.

The number of sub-districts in the Makassar city is 14, covering 143 villages. Among sub-districts, there are seven sub-districts bordering the coastal districts namely Tamalate, Mariso, Wajo, Ujung Tanah, Tallo, Tamalanrea and Biringkanaya. Makassar city residents based on the results from Socio-economic Survey National (SUSENAS) in 2013 approximately 1,408,072 people, made up of 696,086 men and 711,986 women [9].

2.2. Data Collection

We obtained data from the Dr. Wahidin Sudirohusodo hospital in Makassar City. The dataset comprises survival time’s measurements, sex, age, hematocrit, thrombocyte, leukocyte, hemoglobin, grade and location of patients with DF. A total of 270 DF patients with positive DF, which the ones admitted to hospital, were collected from period January to December 2016 in the Makassar city. We assume that each DF patient in the data set that is not linked with a death must have been alive at the end of December 2016.
The number of days they hospitalized until they recovered or were allowed to go home is treated as the response $t_{ij}$ in our models, while the remaining survivors were treated as ‘censored’, or in other words, alive at the end of the study period. There are 8.89% censored patients and 91.1% uncensored patients, including 111 (41.11%) men and 159 (46.4%) women with a mean age of 18.6 year. The minimum and maximum age of DF patients were 1 and 68 years, respectively and the average length of stay of DF patients at the hospital was 4 days. The annual dengue incidence rate (number of cases/10,000 people/year) for the January – December 2016 period for the Makassar City based on the sub-district can be seen in figure 1. The major epidemics occurred in Tamalanrea sub-district, whereas considerably fewer dengue cases were reported in Mamajang sub-district.

The response variable used in this study is the number of days they hospitalized and the predictor variables are sex, age, hematocrit, thrombocyte, leukocyte, hemoglobin, grade, and the population size. The data for the same time period for population size was obtained from the Indonesian Bureau of Statistics [9].

2.3. Multilevel Model
The multilevel model concerns populations with hierarchical structures, where samples from the population can be described as a multistage sample [3]. This model is useful for predicting relationships between variables observed at different levels in the multilevel data structure. The simplest multilevel model is a two level model where level 1 is the individual data and level 2 is the group data [10]. Generally, the regression model includes more than one predictor variable at the lowest level as well as at a higher level. The multilevel regression model for level 1 on each level 2 unit [4] is:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + e_{ij},$$  

(1)

where $Y_{ij}$ is the respond variable for level 1, $\beta_{0j}$ is an intercept for level 1, $\beta_1$ is the regression coefficient for level 1, $X_{ij}$ is the predictor variable for level 1, $e_{ij}$ is error for level 1. Index $i$ represents the unit of level 1 nested on level 2 and index $j$ is the unit of level 2. The equation (1) at the level 2, that is intercept at level 1 $\beta_{0j}$, becomes the response variable at level 2 as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j},$$  

(2)

where $\gamma_{00}$ is the intercept on level 2, $\gamma_{01}$ is the regression coefficient at the level 2, $Z_j$ is the predictor variable at the level 2, $u_{0j}$ is the error at the level 2. By substitution (2) to (1), obtained multilevel regression model as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{01} Z_j + \beta_1 X_{ij} + u_{0j} + e_{ij}$$  

(3)
2.3.1. Multilevel Survival

Generally, the data type has a hierarchical structure such as survival data. If the response variable is the time for the occurrence of a particular occurrence, the hierarchy of such data may appear. Kim and Dey [6] use a hazard rate with the addition of frailty or random effects of individual effects and group effects in modeling multilevel survival defined as follows:

\[ h(t, X, U_{ij}) = h_0(t) \exp(\beta_0j + \beta_iX_{ij} + U_{ij}), \]

where \( t \) is the observational time, \( X \) is the covariate variable, \( \beta \) is the covariate parameter, \( h_0(t) \) is the baseline hazard, \( u_j \) is the random effect group, \( e_{ij} \) the individual random effect, \( i = 1, 2, ..., n \), \( j = 1, 2, ..., J \), \( n \) are the numbers of individual and group, respectively.

2.3.2. Multilevel Spatial Survival

Frequently, the time data up to the occurrence of an event are grouped in strata or groups such as geographic areas. Hazard rate with the addition of frailty or spatial random effects [8] can be written as follows:

\[ h(t, X, w_j) = h_0(t) \exp(\beta_0j + \beta_iX_{ij} + w_j), \]

where \( w_j \) is the spatial random effect, \( i = 1, 2, ..., n \), \( j = 1, 2, ..., J \), \( n \) is the number of individuals, \( J \) is the number of locations. By combining a hazard model with the (4) and (5) equation we can obtain hazard model for the spatial survival multilevel as follows:

\[ h(t, X, w_j, u_j, e_{ij}) = h_0(t) \exp(\beta_0j + \beta_iX_{ij} + U_{ij}), \]

\[ U_{ij} = w_j + u_j + e_{ij} \]

the function of likelihood full for the model on the (6) equation above is:

\[ L(t_{ij}, X_{ij}, U_{ij}) = \prod_{i=1}^{n} \prod_{j=1}^{J} (\alpha t_{ij}^{-\gamma} \exp(\beta_0j + \beta_iX_{ij} + U_{ij}))^{\delta_{ij}} \exp\{-\exp(\beta X_{ij} + U_{ij}) t_{ij}^{-\alpha}\}, \]

where \( \delta_{ij} \) is a death indicator (0 if alive, 1 if dead).

2.4. Bayesian Approach

From the perspective of an estimation theory, there are two approaches namely the classic statistical approach and the Bayesian statistical approach. The former one is mainly dependent on an inferential process on sample data from a population. While the latter one apart from making use of such sample data, also takes into consideration a prior distribution called prior statistical inference with Bayesian statistical approach. The classic statistical approach views the parameter \( \theta \) as a parameter with a fixed value, while the Bayesian statistical approach regards the parameter \( \theta \) as a random variable with distribution so called prior distribution from which we can determine a posterior distribution to then acquire Bayesian estimator as the average or modus of posterior distribution [10].

The establishment of a prior from multilevel model parameter is done in order (two stage prior) for a two stage hierarchical model. Making the stage-1 prior is the establishment of a prior for the parameter \( \beta \) at the lowest level (level 1). The next stage is to determine the stage-2 prior, which is a prior for the parameter \( \gamma \) at the higher level (level 2). This prior is also called hyperprior. A prior selected for the covariate \( \beta \) parameter is an uninformative prior with \( \beta \sim N(0,100) \) and \( \gamma \sim N (0,100) \).

Suppose, we implement the lattice frailty model (6). We know assume that we observe the possibly right-censored data with \( \delta_{ij} \). The joint posterior distribution of this is given by:

\[ p(\beta, W, U | t, x, \delta) \propto \prod_{i=1}^{n} \prod_{j=1}^{J} (\alpha t_{ij}^{-\gamma} \exp(\beta_0j + \beta_iX_{ij} + U_{ij}))^{\delta_{ij}} \exp\{-\exp(\beta X_{ij} + U_{ij}) t_{ij}^{-\alpha}\}. \]

where the first term in the right-hand side is the likelihood, the second is the joint distribution of the random frailties, and the remaining terms are prior distributions. The prior and posterior distributions for each of parameters can be seen in Table 1.

In general, the computation method for a Bayesian inference uses Markov Chain Monte Carlo (MCMC). In MCMC, there are some algorithm types including Gibbs Sampling which is used to withdraw samples form a complex distribution with high dimensions [11]. The main concept of Gibbs
Sampling is how to find the form of conditioned univariate distributions containing all random variables but only one variable is going to be taken its values.

\[ f(\beta) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{\beta^2}{200}\right) \]

\[ f(\alpha) = \frac{1}{\Gamma(1)} \exp(-\alpha) \]

\[ f(e) = \frac{1}{\Gamma(1)} \exp(-e) \]

\[ f(\gamma) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{\gamma^2}{200}\right) \]

\[ f(w_j) = \frac{1}{\sqrt{2\pi/\lambda_j}} \exp\left(-\frac{(w_j - \bar{w}_j)^2}{2/(\lambda \gamma_j)}\right) \]

\[ f(u) = \frac{1}{\Gamma(1)} \exp(-u) \]

Table 1. Prior Distribution and hyperprior for each of parameters

| Parameter     | Prior distribution form |
|---------------|-------------------------|
| \( \beta \sim N(0,100) \) | \[ f(\beta) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{\beta^2}{200}\right) \] |
| \( \alpha \sim \Gamma(1,1) \) | \[ f(\alpha) = \frac{1}{\Gamma(1)} \exp(-\alpha) \] |
| \( e \sim \Gamma(1,1) \) | \[ f(e) = \frac{1}{\Gamma(1)} \exp(-e) \] |
| \( \gamma \sim N(0,100) \) | \[ f(\gamma) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{\gamma^2}{200}\right) \] |
| \( w_j | \lambda \sim N(\bar{w}_j, 1/(\lambda \gamma_j)) \) | \[ f(w_j) = \frac{1}{\sqrt{2\pi/\lambda_j}} \exp\left(-\frac{(w_j - \bar{w}_j)^2}{2/(\lambda \gamma_j)}\right) \] |
| \( \lambda \sim \Gamma(1,1) \) | \[ f(u) = \frac{1}{\Gamma(1)} \exp(-u) \] |

We choose the normal prior for \( \beta, \rho \sim Gamma\left(\alpha, \frac{1}{\alpha}\right) \) and \( \lambda \sim Gamma(a, b) \). Diffuse priors are represented by large positive values for \( \alpha, a \) and \( b \). The joint posterior density of \( (\beta, W, \rho) \) in (4) is analytically intractable because the integration of the joint posterior density is not easy to perform. Hence, the Gibbs sampler is used to update the parameter in the model. The complexity of the likelihood in (8) excludes closed-form full conditionals; the parameters may be updated conveniently using Metropolis-Hastings algorithm [12]. In this study, MCMC computations were implemented using the OpenBUGS system [13].

3. Results

The first stage is examining the Weibull distribution on the data using the Mann test, which is special for one distribution. This test shows that the DF survival time data is a weibull distribution. In this paper, spatial correlation was involved as a frailty term at the level 2 and using Moran’I with the queen contiguity neighborhood tested it. The Moran’I test shows that there is the autocorrelation spatial among location. Figure 2 illustrated queen contiguity neighborhood. Next step is to estimate the posterior parameters of the proportional hazard models for spatial survival multilevel use Bayesian approach via Gibbs Sampling algorithm. The posterior parameters that were estimated namely \( \beta, \alpha, \sigma_{\epsilon_{ij}}, \gamma, w, \sigma_{u_{ij}}, \lambda \).

We ran initially parallel MCMC for each of the models. Then, we monitored them using measurements of sample autocorrelations within the chains, cross-correlations between the parameters, and plots of the sample traces. For the spatial models, however, convergence was much slower. Here our diagnostic tools suggested discarding the first 10 000 iterations from each chain as pre-convergence burn-in. Then, we retained the remaining 50 000 iterations yielded a final sample of posterior analysis. We also note that the lattice model took shorter to run. This can be attributed to the computations involving matrix inversions and determinant evaluations of dimension 14 x 14 within each iteration of the sampler (there are 14 counties in Makassar). The CAR model, on the other hand, avoids this problem since it directly models the weight (inverse dispersion) matrix. We run our models in OpenBUGS, while R with map tool package was used for mapping the results.
Figure 2. Map of Makassar City with queen contiguity.

Table 2. Posterior summaries for the spatial survival multilevel model.

| Model   | Parameter                                | Mean     | 95% Credible Interval (CI)                        |
|---------|------------------------------------------|----------|--------------------------------------------------|
| Level 1 | $\beta_{\text{age}}$                     | -0.019   | (-0.033, -0.006)                                 |
|         | $\beta_{\text{sex}}$                     | 0.139    | (-0.132, 0.409)                                  |
|         | $\sigma_{\epsilon ij}^2$                 | 2.866    | (0.8817, 6.422)                                  |
|         | $\alpha$                                 | 2.081    | (1.888, 2.274)                                   |
|         | $\gamma_{00}$                            | -2.969   | (-4.257, -1.637)                                 |
| Level 2 | $\gamma_{\text{population density}}$     | 7.372 x 10^{-7} | (-9.842 x 10^{-6}, 1.093 x 10^{-5}) |
|         | $\sigma_{\mu ij}^2$                      | 2.878    | (0.845, 6.354)                                   |

Table 2 provides 2.5, 50, and 97.5 posterior percentiles for the main effects in our hazard proportional multilevel models. In all two models, all of the predictors are significant at the 0.05 levels, except sex and density population. In the CAR model, the age covariate increases the posterior median hazard rate by a factor of $e^{-0.019} = 0.981$.

The mapping summaries of our results can be seen in Figure 3. We use the Maptool R package to map resulting fitted hazards rates. Figure 3 describes the DF disease mapping in Makassar City. This disease mapping presents the estimated result of the spatial effect for each sub-district to survival of DF patients in Makassar City. From Figure 3, we can see that the lowest spatial effect to survival of DF patients occur in Makassar sub-district (-0.101) and the highest spatial effect to survival of DF patients occur in Biringkanaya sub-district (0.099). The fitted model indicates excess mortality in the north, which is accentuated and extended to a generally increasing pattern from south to north by the survival CAR model.

We now consider several additional plots for our Weibull CAR frailty model with covariates to the observed survival times. Figure 4 illustrates the estimated survival probability curve for a selected set of covariate based on a sample in the DF dataset. The dashed curve is similar, except that it is under the CAR frailty model; for this model, 95% equal-tail point wise credible intervals are also shown as dotted curves. The estimated survival falls quite rapidly towards 0.10 when the survival time of this DF patient is about 7 days but then decays very slowly.

4. Conclusion
In this paper, we already presented the frailty modelling for spatially correlated and multilevel survival data. The Bayesian method coupled with MCMC computational technique was used as an approach to multilevel spatial survival models with CAR frailty. The case study involved DF survival period January
to December 2016 at Dr. Wahidin Sudirohusodo hospital in Makassar, with covariates given by sex, age, and population density. Here, we proposed ‘lattice’ models with CAR.

**Figure 3.** Posterior mean frailties with covariate, multilevel spatial survival, Makassar DF data.

**Figure 4.** Kaplan-Meier curve for multilevel spatial survival model for DF dataset in Biringkanaya sub-district, Makassar, Indonesia.

Overall, this result shows that only the age covariate describe significantly the DF patients’ survival in Makassar City. The lowest and the highest spatial effect to DF patients’ survival occur in Makassar
and Biringkanaya sub-districts, respectively. This trend, combined with the clear emergence of the Makassar urban area, strongly suggests the need for fitting covariates in our model, most of which vary spatially.

Acknowledgements
The authors acknowledge the Directorate of Research and Community Service, Directorate General Strengthening Research and Development Ministry of Research, Technology and Higher Education Subject Contract Number: 005/ ADD/ SP2H/ LT/ DPRM/ VIII/ 2017, for funding this research.

Appendix

model{
    for(i in 1:Nsubj) {
        obs.t[i] ~ dweib(alpha, mu[i])I(t.cen[i],)  
        log(mu[i]) <- v[Id[i]] + beta1*Usia[i] + beta2*JenisKelamin[i] + e[Id[i]]
    }
    for(j in 1:14) { v[j] ~ dnorm(my[j], sigmau)  
        my[j] <- gamma0 + gamma1*KepPend[j] + W[j]
        e[j] ~ dnorm(0.0, sigmase)}
    for (i in 1:nsum) {weights[i] <- 1
        W[1:regions] ~ car.normal(adj[], weights[], num[], tau)  
        W.mean <- mean(W[])
        gamma0 ~ dnorm(0.0,0.001)
        gamma1 ~ dnorm(0.0,0.001)
        beta1 ~ dnorm(0.0, 0.001)
        beta2 ~ dnorm(0.0, 0.001)
        alpha~ dgamma(1, 1)
        tau ~ dgamma(1, 1)
        sigmase ~ dgamma(1, 1)
        sigmau ~ dgamma(1, 1)
    }
}

For full ODC files contact the author directly.

References
[1] World Health Organisation 2012 Global Strategy for Dengue Prevention and Control (2012-2020) WHO Library Cataloguing in Publication Data
[2] National Geographic Indonesia Berita Kesehatan [Online; accessed on 15 May 2016]
[3] Hox J J 2002 Multilevel Analysis: Techniques and Applications (New Jersey: Lawrence Erlbaum Associated Inc)
[4] Hox J J and Maas C J M 2005 Multilevel Analysis:Encyclopedia of Social Measurement 2(1) 785-793
[5] Wang K, Yau K K, Lee A H and McLachlan G J 2007 Multilevel Survival Modeling of Recurrent Urinary Tract Infections Computer Methods and Programs in Biomedicine 87(1) 225-229
[6] Kim S and Dey D K 2008 Modeling Multilevel Survival Data Using Frailty Models Communications in Statistics-Theory and Methods 37(1) 1734-1741
[7] Mustafafi K D 2015 Analisis Data Multilevel Survival dengan menggunakan Metode Hierarchical Likelihood (Universitas Gadjah Mada)
[8] Banerjee S, Wall M M, and Carlin B P 2003 Frailty Modeling for Spatially Correlated Survival Data, with application to infant mortality in Minnesota Biostatistics 4(1) 123-142
[9] Biro Pusat Statistik Kota Makassar 2014 Indeks Pembangunan Manusia Kota Makassar 2014.
BPS Catalog  https://makassarkota.bps.go.id/websiteb/pdf_publikasi/Indeks-Pembangunan-Manusia-Kota-Makassa-2014.pdf [Online; accessed on 31 October 2016]

[10] Thamrin S A, Amran, Kresna A K, Rahmi S, and Ansariadi 2017 Bayesian inference for spatial parametric proportional hazards model using spatsurv R. *AIP Conference Proceedings* 1827 020015-1–020015-16

[11] Gilks W R dan Wild P 1992 Adaptive Rejection Sampling for Gibbs Sampling *Applied Statistics* 41(2) 337-348

[12] Carlin, B P and Louis T A 2000 *Bayes and Empirical Bayes Methods for Data Analysis* ed Boca R (FL: Chapman and Hall/CRC Press)

[13] Lunn D, Spiegelhalter D, Thomas A and Best N 2009 The BUGS project: Evolution, critique and future directions (with discussion) *Statistics in Medicine* 28 3049-3082