Markov Chain Monte Carlo – Based Approaches for Modeling the Spatial Survival with Conditional Autoregressive (CAR) Frailty

Nur Iriawan1† , Suci Astutik2† † , **Dedy Dwi Prastyo3†**

† Department of Statistics, Faculty of Mathematic and Natural Sciences, Institut Teknologi Sepuluh Nopember
† † PhD student of Department of Statistics, Faculty of Mathematic and Natural Sciences,
Institut Teknologi Sepuluh Nopember

Summary
Survival Model is widely used in medical field and biostatistics. This model can be used to identify the risk factors of an event and can handle the situation when risk factors change with time. Timing of an event frequently depends on the location (spatial) called as spatial survival model. In the development, survival modeling also included random effects models (frailty) to overcome the heterogeneity / sources of unexplained variance in the model. Bayesian approach couple with Markov Chain Monte Carlo (MCMC) was developed in this paper to estimate the spatial parameters of survival models with Conditional Autoregressive (CAR) frailty. The purpose of this study is to assess and implement the MCMC algorithm for modeling survival by using software WinBUGS CAR frailty that can be used to overcome the heterogeneity / sources of unexplained variance in the model because of the influence of the location.

Key words:
MCMC, Bayesian, Spatial survival, CAR frailty.

1. Introduction

Recently it has been developed a statistical method that allows for parameter estimation for models with an unknown concentration of certain opportunities forms of normalized constants and therefore it work based on full-conditional form iteratively. The method is known as Markov Chain Monte Carlo (MCMC). This method runs based on simulation techniques, which work on the scope of Bayesian statistical inference. In the development, this method can also be used to infer the spatial survival model parameters. Bayesian approach is employed for fitting hierarchical frailty model using MCMC computing method with Gibbs sampler algorithm [1]. MCMC is also very useful in determining the marginal posterior parameter that sometimes requires a very complicated integration process and quite difficult to be solved analytically. Survival Model is a model that is widely used in medical field and biostatistics [2], which can be used to identify the risk factors of an event and can handle the situation when risk factors change with time [3]. Often the timing of an event depends on the location (spatial). There are two approaches to capture the spatial factors, namely geostatistic approach using geographic location (latitude and longitude) and lattice approach which uses position of a region relative to another. Banerjee et. al. (2003) developed a hierarchal spatial survival models involving Conditional Autoregressive (CAR) distributed random effects (frailty) [4]. Inclusion of random effects or frailty term in the model can be used to address specific cases (for instance the case with spatial data) where there is diversity / variance sources that can't be explained by a vector covariate in the model. As a result, there was a bias in the estimation of survival parameters [5]. Often frailty also found that it behaves a neo-normal distribution or a mixture ([6]-[9]).
This research was conducted to assess and implement the Bayesian approaches couple with MCMC algorithm for modeling spatial survival with CAR frailty of dengue fever in the Pamekasan City Hospital, East Java, Indonesia. These coupling of methods are supposed to be able to overcome the heterogeneity/sources of unexplained variance in the model because of the influence of spatial. The algorithm will be implemented in WinBUGS.

2. Markov Chain Monte Carlo (MCMC)

MCMC is done by generating data parameter θ using Gibb's sampler. Parameter θ is expressed as a random vector with certain distribution, and the functions of estimator value, $f(\widehat{\Theta})$, is involved in joint distribution of $f(\Theta)$ [10]. The algorithms of Markov Chain Monte Carlo to obtain the posterior can be shown as follows:

i. Choose an initial value $\theta^{(0)}$.
ii. Generate samples $\theta^{(1)}$, $\theta^{(2)}$, …, $\theta^{(T)}$ from the full conditional posterior distribution of p(θ|x).
iii. Monitore convergence algorithm, if not convergent it is necessary to generate more observations.
iv. Remove the first B observations (sample burn-in)
v. Note $\{\theta^{(B+1)}, \theta^{(B+2)}, \ldots, \theta^{(T)}\}$ as a sample for posterior analysis.

Manuscript received December 5, 2010
Manuscript revised December 20, 2010

vi. Plot the posterior distribution
vii. Get a conclusion from the posterior distribution (mean, median, etc.).

The above algorithm will provide a random sample $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots, \theta^{(T)}$ which go along with marginal distribution of $G(\theta)$ and, therefore, some of the characteristics of parameter θ can be obtained as:

1. Summary of posterior $G(\theta)$ from the sample by using a simple sample estimate. For example, the mean of the posterior can be obtained by the formula:

$$\hat{E}(G(\theta)\,|\,x) = \frac{1}{T'}\sum_{1}^{T'} G(\theta^{(t)}) \tag{1}$$

   Another measurement scale is the median and quantile (2.5% and 97.5%) gave 95% confidence interval.
2. Summary of MC error, which is a measurement scale that measures the variability of any estimate of the simulation. MC error had small value to calculate the desired parameters with increased precision.
3. Correlation of among parameters.
4. Plot of the marginal posterior distribution.

3. Spatial Survival Model

Data as the time until the occurrence of an event (time-to-event data) according to [4] are often grouped into strata / groups such as geographical area or areas. In these circumstances, hierarchical model approach using stratum-specific frailties is often suitable.

For example, if the time until recovery occurs or until an individual j ($j = 1,2,\cdots,n_i$) in the strata / groups i ($i = 1,2,\cdots,I$) experienced an event while the event is expressed as t_{ij} , a vector of covariate that affect the incidence is expressed by $\mathbf{x}_{ij}$, and by assuming that the proportional hazard $h(t_{ij};\mathbf{x}_{ij})$ follows the Weibull parametric model:

$$h(t_{ij};\mathbf{x}_{ij}) = \rho t_{ij}^{\rho-1}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij}), \tag{2}$$

then if the model capture the frailty, the proportional hazard $h(t_{ij};\mathbf{x}_{ij})$ can be expanded into:

$$h(t_{ij};\mathbf{x}_{ij}) = \rho t_{ij}^{\rho-1}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} + \mathbf{W}_i), \tag{3}$$

where ρ is the shape parameter of the baseline hazard and $\boldsymbol{\beta}$ contain the intercept for the baseline hazard. Parameter ρ represents monotonicity of hazard rate in the Weibull model. When $\rho > 1$ the hazard rate will be increasing monotone, and inversely $\rho < 1$ will be decreasing monotone, while $\rho = 0$ stating constant hazard rate (Box and Jones, 2004 in Darmofal (2008) [5]). $\mathbf{W}_i$ represents an i^{th} partition areas of region D indexed in a discrete pattern. Partitions are referred to as the 'lattice'. This model uses the method combining information about the areas adjacent to each other / its neighbors compared to metric distance information [4]. As a result, the distribution of random effect $\mathbf{W}$ is defined as,

$$\mathbf{W}\,|\,\lambda \sim CAR(\lambda), \tag{4}$$

and are called as conditionally autoregressive model which indicates the existence of spatial dependence on the composition of covariance [11], where λ is the CAR parameter distribution stating precision or variance inverse of its random effect distribution (θ).

4. Spatial Survival Simulation on WinBUGS

Data used in this research is length of stay data of patients hospitalized with dengue fever until they recovered or allowed to go home in the Pamekasan City Hospital. This length of stay data is expressed as a failure event. Spatial factors are elaborated by the neighborhood between the locations of one another (adjacent matrix). Pamekasan district map with 13 sub-district areas is presented in Figure 1.

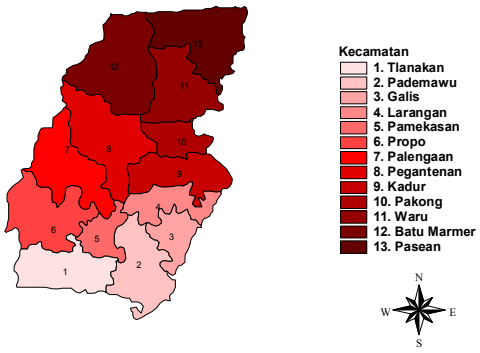


Fig. 1. Map of Pamekasan District

Response variables used in this study is length of stay (t) and the predictor variables are: sex (X1) (categorized as 1 is Female and 2 is Male), age (X2), hematocrit level (X3), and total of trombosite (X4) (categorized as 1 if X4<50,000/µl; as 2 if 50000/µl<X4<100000/µl; as 3 if 100000<X4< 150001/µl; and as 4 if X4>150,000 /µl). Some step to work with MCMC algorithm on spatial survival model is presented in Figure 2.

5. Simulation Result and Discussion

The first step for doing the MCMC simulation in spatial survival model is defining a matrix of spatial weights. This matrix is used as parameters of the CAR prior distribution in the formation of spatial survival model. Queen Contiguity criteria can be used for determining the spatial weight matrix by employing the neighborhood of their respective areas of the so-called contiguity / adjacent. Contiguity/adjacent matrix for Pamekasan district is shown in Figure 3.

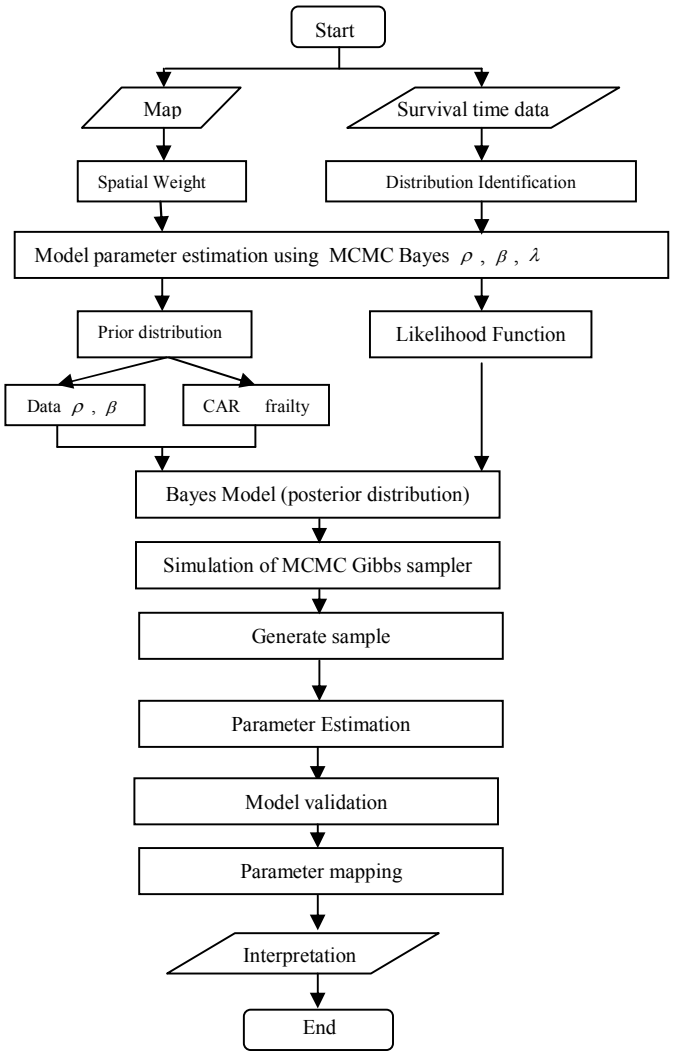


Fig. 2. MCMC Algorithm on Spatial Survival Model

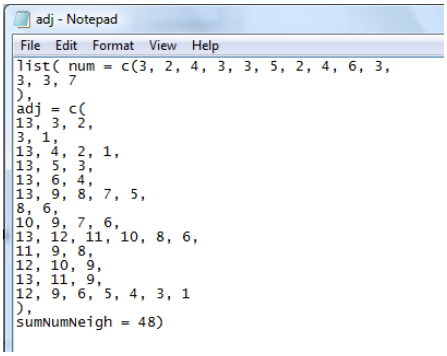


Fig. 3. Adjacent Matrix of Pamekasan Districts

Secondly, modeling assumptions that must be met in the proportional hazard modeling is that the hazard function of the categorical predictor variables have to be proportional at all times. Proportional testing can be done by using plot –ln[-ln $S(t)$] descriptively. Figure 4 shows lines between categories (male and female) are parallel, and the proportional hazard assumption, therefore, can be fulfilled. It's meant that the predictor variables were independent of time and the relationship between the cumulative hazard is proportional / constant every time.

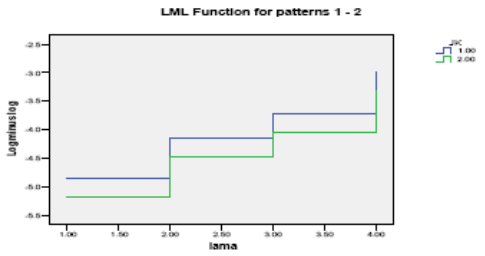


Fig. 4. Proportional Hazard Assumption for

The next step is determining the distribution of the length of stay as its survival time (t). From the goodness of fit test, all of 13 districts in Pamekasan, statistically can be modeled as Weibull distributions, even with different significant level which are all greater than 5%. This study, therefore, will employ the Weibull proportional hazard to estimate the spatial survival models with CAR frailty.

Summary of MCMC simulation results on parameter estimation spatial survival model with frailty CAR are shown in Table 1. Simulation was done 51,000 iterations which has burn-in on iteration 1,000. Time execution for this process is 440 seconds or 7.3 minutes.

Based on Table 1, there is one area within 13 areas in that region with no parameter significant. This area is Pamekasan sub-district. This is because the prior

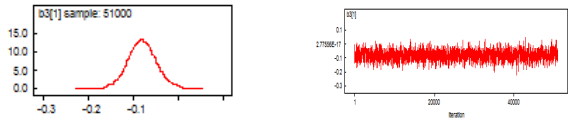
distribution parameters chosen for this sub-district is uni-modal, whereas from the plot of the data showed a tendency to be multi-modal or mixture distribution. This issue will become a significant research in the future.

Posterior distribution and iteration history plots of the MCMC output of the significant parameters are presented in Figure 5 to Figure 16.

Table 1. MCMC Simulation Results on Spatial Survival Models with CAR Frailty

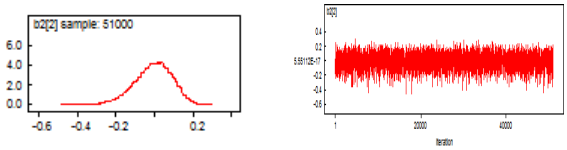
Sub-district	Mean Parameter model survival spatial dengan frailty CAR					
	b1	**b2**	**b3**	**b4_1**	**b4_2**	**b4_3**
Tlanakan	-0.540	-0.094	**-0.082**	-0.728	-0.452	-0.097
Pademawu	-0.429	-0.004	**-0.105**	-0.103	1.028	-0.326
Galis	-0.583	**-0.574**	0.038	-0.043	**3.549**	-0.014
Larangan	-0.818	**-0.168**	**-0.132**	-0.124	**-2.819**	-0.600
Pamekasan*	-0.603	-0.020	-0.040	-0.674	-0.689	0.023
Propo	**-1.052**	0.039	-0.078	-0.188	-0.712	-0.013
Palengaan	**-1.251**	-0.187	**-0.096**	-0.100	-0.537	-0.012
Pegantenan	-0.968	-0.212	**-0.143**	0.008	-1.831	-0.016
Kadur	-0.712	**-0.739**	**-1.843**	-0.009	**79.790**	-0.014
Pakong	-0.832	**-0.958**	0.039	-0.012	7.270	-0.003
Waru	**-1.051**	-0.130	-0.060	0.004	-0.002	-0.172
Batu Marmar	-0.890	**-0.434**	**-0.187**	0.091	3.350	0.153
Pasean	**-1.053**	-0.286	-0.081	0.117	-1.276	-0.022

Note : : significance on 95% confidence interval
* : Pamekasan beside as the district name, its is also as a sub-district name.



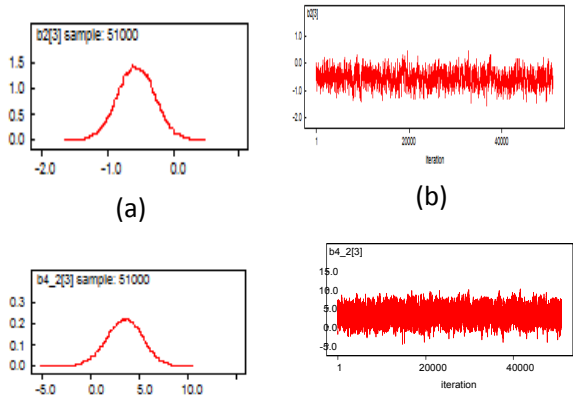
(a) (b)

Fig. 5. (a) Posterior distribution and (b) history iteration plots of the hematocrit levels parameters in the Tlanakan sub-district



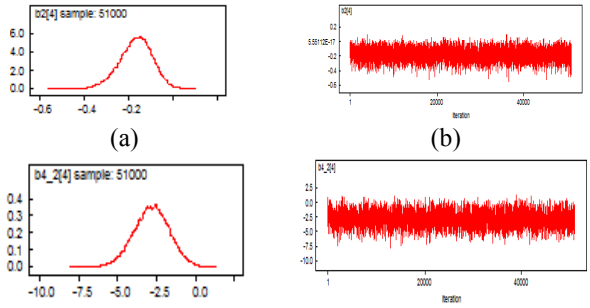
(a) (b)

Fig. 6. (a) Posterior distribution and (b) history iteration plots of the hematocrit levels parameters in the Pademawu sub-district



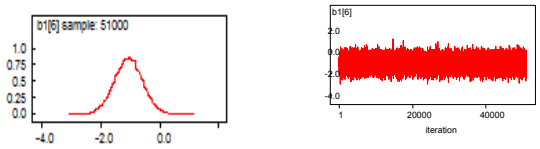
(a) (b) (c) (d)

Fig. 7. (a) Posterior distribution and (b) history iteration plots of the age parameters, (c) Posterior distribution and (d) history iteration plots of the trombosite parameters (total from 50,000 to 100,000/µL), in the Galis sub-district.



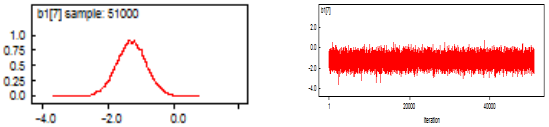
(a) (b) (c) (d)

Fig. 8 (a) Posterior distribution and (b) history iteration plots of the parameters of age, (c) Posterior distribution and (d) history iteration plots of the trombosite parameters (total from 50,000 to 100,000/µL), in the Larangan sub-district.



(a) (b)

Fig. 9 (a) Posterior distribution and (b) history iteration plots of the gender parameters in the Propo sub-district.



(a) (b)

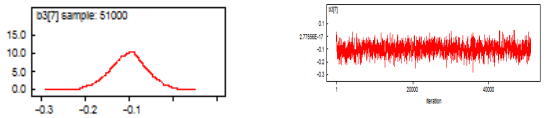


Fig. 10. (a) Posterior distribution and (b) history iteration plots of the gender parameters, (c) Posterior distribution and (d) history iteration of the age parameters, in the Palengaan sub-district.

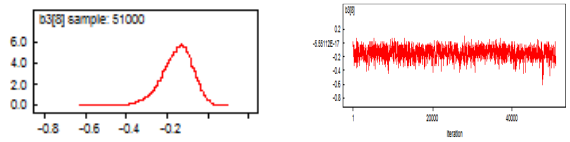


Fig. 11 (a) Posterior distribution and (b) history iteration plots of the hematocrit level parameters in the Pegantenan sub-district.

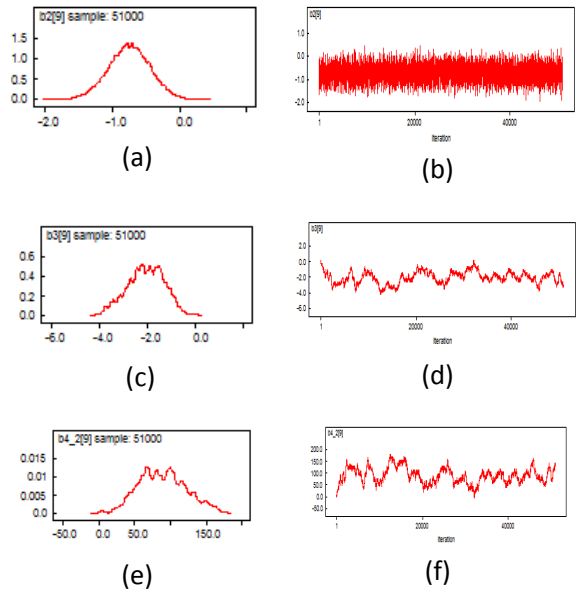


Fig. 12. (a) Posterior distribution and (b) history iteration plots of the age parameters, (c) posterior distribution and (d) history iteration plots of the hematrocrit level parameters, (e) posterior distribution and (f) history iteration of the trombosite parameters (total from 50,000 to 100,000/µL), in the Larangan sub-district.

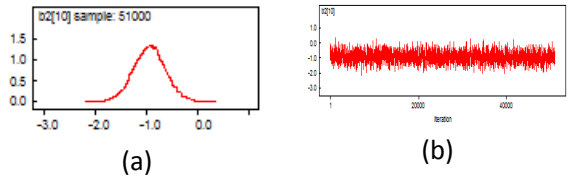


Fig. 13 (a) Posterior distribution and (b) history iteration plots of the age parameters in the Pakong sub-district.

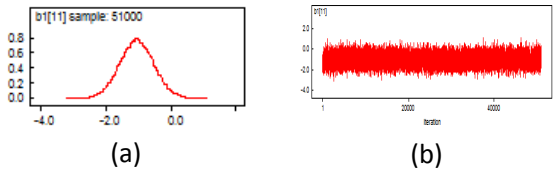


Fig. 14 (a) Posterior distribution and (b) history iteration plots of the gender parameters in the Waru sub-district.

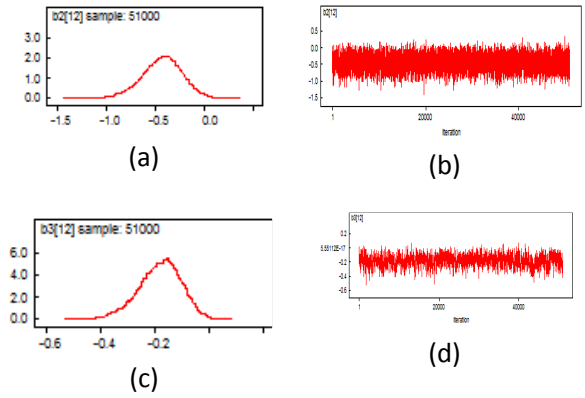


Fig. 15. (a) Posterior distribution and (b) history iteration plots of the age parameters, (c) posterior distribution and (d) history iteration of the hemotocrit level parameters, in the Batu Marmer sub-district.

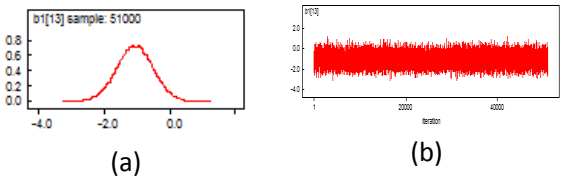


Fig. 16. (a) Posterior distribution and (b) history iteration plots of the gender parameters in the Pasean sub-district.

Figure 5 to Figure 16, except posterior in Figure 12, show that all the posterior distribution of parameters perform a normal distribution and their history iteration shows such fastly mixing MCMC processes. One reason of Figure 12 is not to be ideally performs because the amount of data (patient) in that area is too small, only 2 people.

Table 2. Validation of the Spatial Survival model with CAR Frailty

Model	MAE
Survival without frailty	4.115
Survival with CAR frailty	4.073

Validation of MCMC Simulation on spatial survival models with CAR frailty are carried out by comparing this result with spatial survival model without frailty [12]. Mean Absolute Error (MAE) is employed here to choose the best model. The smallest MAE value of the model indicates the best model. The comparison is showed in Table 2, which demonstrates that MCMC for spatial survival models with CAR frailty can reduce the error about 3.8% compared with no frailty survival model.

6. Conclusion

This paper has presented the Bayesian couple with MCMC computational methods as an approach to spatial survival models with CAR frailty using WinBUGS software. These methods not only have succeeded to demonstrate the accuracy in estimating parameters of spatial survival models with frailty CAR and have shown lower model error than without frailty, but also can overcome the heterogeneity/sources of unexplained variance in the model because of the spatial effect.

Acknowledgments

We would like to express our sincere gratitude to Research Institute of Institut Teknologi Sepuluh Nopember (ITS) funded this research as part of the Professor Research.

References

[1] Carlin, B. P., & Louis, T. A., (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2 ed.). Boca Raton: FL: Chapman and Hall/CRC Press.

[2] Cox, D. R., & Oakes, D., (1984). *Analysis of Survival Data.* London: Chapman and Hall.

[3] Dohoo, I. R., (2008). Quantitative epidemiology: Progress and Challenges. *Preventive Veterinary Medicine, 86*, p. 260-269.

[4] Banerjee, S., Wall, M. M., & Carlin, B. P., (2003). Frailty Modeling for Spatially Correlated survival data, with application to infant mortality ini Minnesota. *Biostatistics*, p. 123-142.

[5] Darmofal, D., (2008). *Bayesian Spatial Survival Models for Political Event Processes*. Department of Political Science, University of South Carolina. 350 Gambrell Hall. Columbia.

[6] Iriawan, N., (1999). *On Stable and Adaptive Neo-Normal Distributions*., (p. 384 - 389). Yogyakarta.

[7] Iriawan, N., (2000). *Computationally Intensive Approaches to Inference in Neo-Normal Linear Models*. CUT-Australia.

[8] Iriawan, N., (2001a). *Penaksiran Model Mixture Normal Univariabel: Suatu Pendekatan Metode Bayesian dengan MCMC*., (p. 105-110). Yogyakarta.

[9] Iriawan, N., (2001b). *Studi tentang 'Bayesian Mixture Normal' dengan menggunakan metode Markov Chain Monte Carlo (MCMC)*, Laporan Penelitian, LEMLIT-ITS.

[10] Box, G. E. P., & Tiao., (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-wesley.

[11] Wall, M. M., (2004). A Close Look at the Spatial Structure Implied by the CAR and SAR Models. *Journal of Statistical Planning and Inference, 121*, p. 311-324.

[12] Amalia, S., Iriawan, N. & Prastyo, D.,2010, Implementation of Bayesian Mixture Survival : *Case Study of the Dengue Hemorragic Fever Patients Recovery*, Proceeding of International Conference of Mathematics and Natural Science 2010, Bandung.

Nur Iriawan, received the B.Science in Statistics from ITS, Indonesia, M.Science in Computer Science from Indonesia University, Indoneisa and PhD in Mathematics and Statistics from Curtin University of Technology, Western Australia in 1986, 1990 and 2000 respectively. He has been a Lecturer Staff of ITS since 1988. His research interest in computational statistics: Bayesian (MCMC), Stochastic modeling and simulation.

Suci Astutik, received the B.Science from Brawijaya University and M.Science degrees from ITB in 1993 and 2000, respectively. She has been a PhD student of ITS since 2009. She has also been a lecturer staff of Brawijaya University since 1999. Her current research interest in computational statistics: Bayesian (MCMC), Stochastic modeling

Dedy Dwi Prastyo, received the B.Science and M.Science degrees from ITS in 2006 and 2008, respectively. At present, he is a lecturer staff in Department of Statistics, ITS, Indonesia. His current research interest is statistical computation and modeling.