Bayesian Mixture Generalized Extreme Value Regression with Double-Exponential CAR Frailty for Dengue Haemorrhagic Fever in Pamekasan, East Java, Indonesia

D Rantini 1, N Iriawan 1*, and Irhamah 1
1 Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

*e-mail: nur_i@statistika.its.ac.id

Abstract. There is a lot of research on infectious diseases. One of the dangers is caused by dengue hemorrhagic fever (DHF) viruses. Almost all the world there are dengue viruses that spread. In health science, this research continues to increase over time. It has been widely studied by medical circles, but from a statistical point of view, patient survival can be investigated probabilistically. The survival time of DHF patients could refer to the time of arrival of DHF patients to the hospital with all the health conditions of the patient to be discharged from the hospital. It is because the patient has been declared medically cured or further improved from the initial hospital admission. In this case, survival analysis will be appropriate to be implemented, where the event of the study is DHF patients who are discharged from the hospital due to recovery. DHF patient survival data is found following the multimodal Generalized Extreme Value distribution so that a mixture model would be employed. There is a lot of transmission media for diseases transmitted by the DHF virus, but the fastest transmission medium occurs in a short time, called the Aedes Aegypti Mosquito. This transmission of the DHF virus by mosquitoes can occur between adjacent neighboring regions. This article brings together two fields of research, namely survival models with spatial random effects. The Bayesian analysis will be used as a model parameter estimation approach, whereby unmeasured autocorrelations between intersecting regions will be captured through the conditional autoregressive (CAR). Two different distributions approach for these unmeasured autocorrelations will be given, i.e. the Normal CAR and the Double Exponential CAR. Four predictor variables, namely sex, age, hematocrit level, and platelet count in each DHF patient are used to explain the variability of the survival. The Cox Proportional Hazard regression is employed to find out the significant influence of those variables on the survival spatial DHF model. The result shows that the two-component mixture Generalized Extreme Value regression model coupled with a random Double Exponential CAR effect is the best model. Based on the best model obtained, the significant predictor variables for the DHF spatial survival model are sex, hematocrit level, and platelet count.

Keywords: Bayesian, conditionally autoregressive, dengue haemorrhagic fever, frailty, generalized extreme value.
1. Introduction

Some infectious diseases that exist in the world according to the World Health Organization (WHO) include Diarrhea diseases, Malaria, HIV, TB, dengue, syphilis, and several other diseases [1]. Several infectious diseases are common in Indonesia, some of which are acute respiratory infections, diarrhea, tuberculosis, dengue fever, intestinal worms, skin diseases, malaria, and diphtheria. Among these infectious diseases, East Java province stay in the first ranked in terms of the number of cases of dengue fever [2]. Pamekasan Regency is one of the regencies in East Java Province, so this area is interesting to be studied. This study will analyze the recovery time of DHF patients calculated since a patient enters the hospital until he discharged from the hospital because of his healing.

Each recovery time of DHF data acquisition for each patient will form a data set having a certain distribution. It can have a multimodal pattern and requires to be analyzed using a survival mixture model. The recovery time of DHF patients in Pamekasan, on the other hand, supposed to have a dependency relationship between one region and another, in this case, the sub-district. Therefore, survival modeling coupled with spatial random effects is proposed as a promising method of solving this problem. This approach is referred to as a spatial effect Conditionally Autoregressive (CAR) as has been done by Besag, et.al. [3] and Cressie and Wikle [4]. In previous spatial studies, Normal CAR was frequently used. However, Double-Exponential CAR having a thicker tail is more robust and gives a better result in capturing the spatial effect pattern compared to Normal CAR [5]. This has also been proven for modeling the asthma data in Tehran in 2015 [6].

The patient's recovery rate is influenced by socio-demographic factors, some of which are gender, age, education, marital status, and type of work [7]. Recovery time can also be seen based on the laboratory test results, i.e. hemoglobin, hematocrit, leukocyte, and platelet levels [8]. These factors are believed to influence the rate of recovery of DHF patients. To be able to see the relationship of these factors on the recovery rate of DHF patients, the Cox model can be used [9].

In this study, survival analysis will be employed for data on the recovery time of DHF patients in Pamekasan Regency having both characteristics simultaneously, i.e. multimodal pattern and spatially regional closeness random effect. As has been stated by Susanto, et al. [10], Bayesian methods coupled with the MCMC approach, therefore, will be employed to estimate the mixture model of the Cox regression with CAR non-Gaussian spatial effect of the recovery time of DHF patients in Pamekasan Regency.

2. Spatial Survival Model

In the random effect (frailty), the spatial between strata or groups of observations can be represented by $W$, so the hazard function in the Cox model with a random (frailty) effect for each mutually independent individual can be written as follows [11]:

$$h(t_i; x_i) = h_0(t_i) \exp(\beta' x_i + W_i),$$

(1)

where $t_i$ is a vector of response variables, $x_i$ is a vector of predictor variables, $\beta$ is a regression coefficient vector, and $h_0(t_i)$ is baseline hazard.

The hazard function approach using the individual frailty model is carried out when each individual uniquely has a frailty factor that is not modeled. When each individual is grouped in a hierarchical structure, on the other hand, individuals in the same group will be treated to have the same frailty level. The modeling of this phenomenon, therefore, can employ the shared frailty modeling approach. The Cox model for each stratum (group) using shared frailty models can be seen as equation (2) [11].

$$h(t_{is}; x_s) = h_0(t_{is}) \exp(\beta' x_s + W_s),$$

(2)

where $t_{is}$ is a vector of response variables of $i$-th individual in stratum $s$, $x_s$ is a vector of predictor variables of $i$-th individual in stratum $s$, $\beta$ is regression coefficient vector, $h_0(t_{is})$ is baseline hazard of $i$-th individual in stratum $s$, and the individual frailty $W_s$ is now replaced by group frailty $W_i$.  


3. Bayesian Modeling

To be able to use spatial survival analysis with the Bayesian approach, the refining parameters are required in the prior CAR. We use two CAR priors in this paper, i.e. Normal and Double-Exponential distribution. For the recovery time of DHF patients, a mixture model will be employed for accommodating its multimodal patterns in the data analysis. So modeling the recovery time of DHF patients in this paper will combine the survival analysis modeling with a random spatial effect of CAR and estimated with Bayesian approach.

3.1. Conditionally Autoregressive (CAR) Prior

The distribution of random effects is conditional on hyperparameters $\lambda$, with priors that can be changed according to precision (i.e. as an inverse variance) of the distribution of random effects [11]. In the spatial individual frailty model, the prior of CAR $\lambda$ has a joint distribution, which is proportional to equation (3)

$$
\lambda^{(I-G)} \exp\left(-\frac{\lambda}{2} \sum_{i \text{ adj} i'} (W_i - \overline{W}_i)^2 \right) \sim \lambda^{(I-G)} \exp\left(-\frac{\lambda}{2} \sum_{i \text{ adj} i'} m_i (W_i - \overline{W}_i) \right),
$$

where $I$ is the number of units in the data, $G$ is the number of unconnected units, $i \text{ adj} i'$ indicating unit $i$ and unit $i'$ close together, $\overline{W}_i$ is the average of $W_{i,i'}$ neighboring with $W_i$, and $m_i$ is the number of neighbors [12,13]. The conditional distribution of spatial random effects for the Normal CAR prior could be shown as equation (4)

$$
W_i | W_{i,i'} \sim N\left(\overline{W}_i, \frac{1}{\lambda m_i}\right).
$$

When the conditional distribution of CAR prior follows the Double-Exponential distribution, called the Double-Exponential autoregressive model, it can be mathematically written as in equation (5) [6].

$$
T_i | t_j, j \neq i \sim \frac{1}{2d} \exp\left(-\frac{t_i - \sum_{j \neq i} b_{ij} t_j}{d}\right), i = 1, \ldots, n,
$$

where $b_{ij} = c_{ij}$, $c_{ij}$ is the element of neighborhood matrix, $c_{ij} = \sum_j c_{ij}$, $d$ is the scale parameter of the Double-Exponential distribution, and

$$
c_{ij} = \begin{cases} 1, & \text{if regions } i \text{ and } j \text{ share a common boundary, } i \neq j \\ 0, & \text{if } i = j \end{cases}
$$

3.2. Mixture Model

When the recovery time of DHF data can be collected in a sub-population structure, then mixture survival modeling can be applied [14]. Adaptively, this modeling will be able to provide ideas for dealing with multimodal data in flexible parametric modeling and statistical analysis [15].

If $T$ is a random variable for survival time and $t_1, t_2, \ldots, t_n$ are a random sample with size $n$, then a finite mixture model can be written as equation (6)

$$
p(t) = \sum_{q=1}^Q \pi_q p_q(t),
$$
where $Q$ is the number of mixture components, and for each $q$, $p_q(t)$ is a function of the probability density function of the mixture component $q$-th, and $\pi_q$ is a non-negative proportion, where
$$\sum_{q=1}^{Q} \pi_q = 1.$$ The proportional hazard model for a mixture survival is:
$$h(t) = \sum_{q=1}^{Q} \pi_q h_q(t),$$
with $h_q(t)$ is a hazard function of the mixture $q$-th component.

For the Cox Bayesian model, the joint posterior distribution is as equation (8)
$$p(\beta, W, \lambda | t, x, \gamma) \propto L(\beta, W; t, x, \gamma) p(W | \lambda) p(\beta) p(\lambda),$$
where $p(\beta, W, \lambda | t, x, \gamma)$: joint posterior distribution; $L(\beta, W; t, x, \gamma)$: Cox likelihood; $p(W | \lambda)$: CAR distribution; $p(\beta)$: prior for $\beta$; $p(\lambda)$: hyperprior for $\lambda$; $\beta$: regression coefficient vector; $W$: weighting matrix; $\lambda$: prior unidimensional precision for the joint distribution of random effect vectors; $t$: vector of response variable; $x$: vector of predictor variable; $\gamma$: collection of event indicators, which are censored or not censored.

In Bayesian analysis, Watanabe–Akaike Information Criterion (WAIC) is an improvement over Deviance Information Criterion (DIC) standards [16], especially for comparing mixture models. The smaller the WAIC, the better the model [17].

4. Methodology

Data structure for DHF patients in Pamekasan regency which contains the patient's recovery time, the factors that influence the recovery and the patient's origin district can be seen in Table 1.

| Table 1. The Data Structure of DHF Patients Which Consists of Recovery Time, Predictor Variables and Sub-Districts Where the Patient Lives |
|---|
| No. | $T$ | t.cen | $X_1$ | $X_2$ | $X_3$ | $X_4$ | Sub-district |
|---|
| 1 | 4.787 | 1 | 2 | 6 | 45 | 125 | 1 |
| 2 | 5.725 | 1 | 1 | 2 | 57.8 | 95 | 1 |
| 3 | 4.337 | 1 | 2 | 5 | 37.3 | 66 | 1 |
| : | : | : | : | : | : | : | : |
| 146 | 5.828 | 1 | 1 | 18 | 45 | 43 | 13 |
| 147 | 4.908 | 1 | 2 | 8 | 37.9 | 36 | 13 |

where $T$: recovery time of DHF patients (days); t.cen : event indicator (1 = uncensored, 0 = censored); $X_1$: gender (1 = female, 2 = male); $X_2$: age; $X_3$: hematocrit levels; $X_4$: platelet count; sub-district: the origin of the patient (1 = Tlanakan, 2 = Pademawu, 3 = Galis, 4 = Larangan, 5 = Pamekasan, 6 = Proppo, 7 = Palenga, 8 = Peganten, 9 = Kadur, 10 = Pakong, 11 = Waru, 12 = Batumarmar, and 13 = Pasean). The steps of the analysis to determine significant parameters according to the Bayesian perspective and the model formed in this spatial survival research can be briefly seen in Figure 1.

5. Results and Analysis

An overview of the patient’s recovery time as the survival data which be modeled can be seen through the histogram in Figure 2. In this figure, it can be seen visually that the histogram of the survival time recovery data of DHF patients has two modes. This survival data is multimodal data with two components of the mixture. The descriptive statistics for the data in Table 1 are given in Table 2. Table 2 demonstrates that the $X_4$ variable has a greater variance than the other variables.
Table 2. Descriptive Statistics for Response Variables and Predictor Variables

| Variable | N  | Mean  | St.Dev  | Minimum | Median | Maximum |
|----------|----|-------|---------|---------|--------|---------|
| T        | 147| 4.5160| 1.6120  | 1.0980  | 4.3230 | 9.3680  |
| X1       | 147| 1.4898| 0.5016  | 1.0000  | 1.0000 | 2.0000  |
| X2       | 147| 9.9830| 7.3930  | 2.0000  | 9.0000 | 48.0000 |
| X3       | 147| 43.2850| 5.7470  | 14.3000 | 43.3000| 57.8000 |
| X4       | 147| 66.2900| 49.0300 | 10.0000 | 56.0000| 323.0000|

Suitable distribution of survival data for patient recovery time will be used as the goodness of fit distribution. Using the Kolmogorov-Smirnov and Anderson-Darling statistical test the results can be seen in Table 3.

![Survival Model Research Flowchart with Random Spatial Normal CAR and Double-Exponential CAR Effects](image)

**Figure 1.** Survival Model Research Flowchart with Random Spatial Normal CAR and Double-Exponential CAR Effects.
According to Table 3 showing does not reject the null hypothesis, the survival data on the recovery time of DHF patients follows the Generalized Extreme Value (GEV) distribution with the following probability density function

\[
f(t | \mu, \sigma, \eta) = \frac{1}{\sigma} \left(1 + \frac{\eta}{\sigma}(t - \mu)\right)^{\frac{1 - \eta}{\eta}} \exp \left(-\left(1 + \frac{\eta}{\sigma}(t - \mu)\right)^{\frac{1}{\eta}}\right),
\]  

where \( \eta, \sigma \) and \( \mu \) are shape, scale and location parameter respectively and \( \frac{\eta}{\sigma}(t - \mu) \geq -1 \). The GEV proportional hazard function with a random (frailty) effect has the following form [18]

\[
h(t_i; x_i) = h_0(t_i) \exp(\beta x_i + W_i) = \frac{1}{t_i} \left(\exp\left(\frac{1}{t_i}\right) - 1\right) \exp(\beta x_i + W_i).
\]  

To be able to continue using survival analysis with the Cox model, it can be seen in Table 4 that the proportional hazard assumption has been fulfilled.

**Table 4.** Test of Proportional-hazards Assumption

| chi-squared | degree of freedom | p-value |
|-------------|------------------|---------|
| 1.7500      | 4                | 0.7813  |

The parameter estimation results using the Bayesian MCMC method for each component of the GEV distribution can be seen in Table 5.
Table 5. The Estimation Results of the GEV Distribution Parameters for Each Component of the Mixture

| The number of observations | Parameters | \( \eta \) | \( \mu \) | \( \sigma \) |
|---------------------------|-----------|---------|---------|---------|
| 1-st component            | 126       | -0.2315 | 3.3430  | 1.2090  |
| 2-nd component            | 21        | 1.0380  | 4.4310  | 0.9079  |

Based on these results, it can be described the survival function and hazard function in Figure 3 for the recovery of DHF patients.

![Survival data of patients' recovery time using the mixture model with two components of the GEV distribution; (a) Survival function, and (b) hazard function.](image)

**Figure 3.** Survival data of patients’ recovery time using the mixture model with two components of the GEV distribution; (a) Survival function, and (b) hazard function.

The estimation of Cox regression parameters has been done by using the Bayesian MCMC approach. The estimated survival models with GEV distribution coupled with spatial random effects of Normal CAR are shown in Table 6 and Table 7 for coupling with spatial random effects of the Double-Exponential CAR.

Table 6. Estimation of Cox Regression Parameters with Spatial Random Effects of Normal CAR

| parameter | Normal Conditionally Autoregressive (CAR) | mean | std.dev | MC Error | 2.50% | median | 97.50% |
|-----------|------------------------------------------|------|---------|----------|-------|--------|--------|
| \( b_0 \) |                                           | -0.05057 | 0.16230 | 0.01082 | -0.35110 | -0.06967 | 0.28520 |
| \( b_{21} \) |                                         | 0.05385 | 0.37860 | 0.02536 | -0.76390 | 0.19740 | 0.57590 |
| \( b_{11} \) |                                         | 0.48320 | 0.61110 | 0.04094 | -0.25690 | 0.33000 | 1.61200 |
| \( b_{13} \) |                                         | -0.02891 | 0.25740 | 0.01720 | -0.39630 | -0.11520 | 0.46230 |
| \( b_{22} \) |                                         | 0.12730 | 0.10850 | 0.00718 | -0.06042 | 0.11690 | 0.34500 |
| \( b_{24} \) |                                         | 0.15150 | 0.21910 | 0.01464 | -0.19680 | 0.14080 | 0.61910 |
| \( b_{33} \) |                                         | -0.98670 | 0.03838 | 0.00241 | -1.05200 | -0.98960 | -0.90650 |
| \( b_{41} \) |                                         | -0.48700 | 0.14530 | 0.00966 | -0.72800 | -0.47990 | -0.21050 |
| \( b_{42} \) |                                         | -0.01848 | 0.01406 | 7.93E-04 | -0.05153 | -0.01608 | -0.00184 |
| \( b_{43} \) |                                         | -0.04523 | 0.09176 | 0.00604 | -0.22280 | -0.04961 | 0.11480 |
Based on Table 6, it can be concluded that the significant parameters are $b_3$, $b_2$, and $b_4$, where $b$ is the estimator of $\beta$. The hazard function for models with Normal CAR spatial random effects can be seen in Equation (11) for the first component

$$h_1(t; x_i) = \frac{1}{t_i^2 \left( \exp \left( \frac{1}{t_i} \right) - 1 \right)} \exp \left( -0.05057 - 0.9867 x_{i1} - 0.01848 x_i + W_i \right),$$

and Equation (12) for the second component

$$h_2(t; x_i) = \frac{1}{t_i^2 \left( \exp \left( \frac{1}{t_i} \right) - 1 \right)} \exp \left( 0.05385 - 0.487 x_{i3} + W_i \right).$$

### Table 7. Estimation of Cox Regression Parameters with Double-Exponential CAR Spatial Random Effects

| Parameter | Mean | Std. Dev | MC Error | 2.50% | Median | 97.50% |
|-----------|------|----------|----------|-------|--------|--------|
| $b_{0i}$  | -0.23120 | 0.20560 | 0.01372 | -0.59820 | -0.16120 | 0.06598 |
| $b_{02}$  | -1.22100 | 0.35470 | 0.02375 | -1.82300 | -1.28400 | -0.63290 |
| $b_{1i}$  | -0.42490 | 0.22140 | 0.01479 | -0.87090 | -0.35900 | -0.05491 |
| $b_{3i}$  | -0.61870 | 0.14120 | 0.00938 | -0.90030 | -0.61850 | -0.34750 |
| $b_{21}$  | 0.10980 | 0.11000 | 0.00728 | -0.06427 | 0.09992 | 0.30820 |
| $b_{22}$  | 0.28860 | 0.26790 | 0.01792 | -0.06958 | 0.21940 | 0.76120 |
| $b_{3i}$  | -0.94450 | 0.02646 | 0.001510 | -0.99480 | -0.94460 | -0.89440 |
| $b_{32}$  | -0.51680 | 0.14280 | 0.009504 | -0.76870 | -0.51940 | -0.22740 |
| $b_{4i}$  | -0.02108 | 0.01284 | 0.32E-04 | -0.05043 | -0.01869 | -0.00228 |
| $b_{42}$  | -0.01426 | 0.08955 | 0.005881 | -0.21480 | 0.00907 | 0.11070 |

Based on Table 7, it can be concluded that the significant parameters are $b_{3i}$, $b_{32}$, $b_{3i}$, and $b_{4i}$. The hazard function for models with Double-Exponential CAR spatial random effects can be seen in Equation (13) for the first component

$$h_1(t; x_i) = \frac{1}{t_i^2 \left( \exp \left( \frac{1}{t_i} \right) - 1 \right)} \exp \left( -0.23120 - 0.4249 x_{i1} - 0.9445 x_{i3} - 0.02108 x_{i4} + W_i \right),$$

and Equation (14) for the second component

$$h_2(t; x_i) = \frac{1}{t_i^2 \left( \exp \left( \frac{1}{t_i} \right) - 1 \right)} \exp \left( -1.22100 - 0.6187 x_{i1} - 0.5168 x_{i3} + W_i \right).$$

To compare the two models, WAIC values can be seen in Table 8.
Table 8. Comparison of WAIC values for Cox survival models with spatial effects of Normal CAR compared to Double-Exponential CAR

| Spatial CAR Effect       | WAIC     |
|--------------------------|----------|
| Normal                   | 626.880  |
| Double Exponential       | 726.191  |

6. Conclusion
Based on the analysis and model results obtained in Equations (11), (12), (13) and (14), it is found that the age variable is not significant. This means that the patient's age is not too significant in influencing the patient's recovery rate. Gender, hematocrit levels, and platelet counts, on the other hand, were significant in the two models. But the number given is negative, for hematocrit and platelet variables it means that the greater the hematocrit and platelet, the faster the patient's recovery rate. The alleged existence of leptokurtic patterns in spatial survival of DHF patients was apparently not significantly influential in the Cox regression model in Pamekasan, although Double Exponential CAR was able to elaborate variables that influenced DHF survival better than Normal CAR.

References
[1] Mabey D, Peeling RW, Ustianowski A, Perkins MD. Tropical infectious diseases: diagnostics for the developing world. Nat Rev Microbiol. 2004;2(3):231.
[2] Arsin AA. EPIDEMIOLOGI DEMAM BERDARAH DENGUE (DBD) DI INDONESIA. Masagena Press; 2013.
[3] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991;43(1):1–20.
[4] Cressie N, Wikle CK. Statistics for spatio-temporal data. John Wiley & Sons; 2015.
[5] Besag J, Kooperberg C. On conditional and intrinsic autoregressions. Biometrika. 1995;82(4):733–46.
[6] Motarjem K, Mohammadzadeh M, Abyar A. Bayesian Analysis of Spatial Survival Model with Non-Gaussian Random Effect. J Math Sci. 2019;237(5):692–701.
[7] Ramzan M, Ansar A, Nadeem S. Dengue epidemics: knowledge perhaps is the only key to success. J Ayub Med Coll Abbottabad. 2015;27(2):402–6.
[8] Leong AS-Y, Wong KT, Leong TY-M, Tan PH, Wannakrairat P. The pathology of dengue hemorrhagic fever. In: Seminars in diagnostic pathology. Elsevier; 2007. p. 227–36.
[9] Lee ET, Wang J. Statistical methods for survival data analysis. Vol. 476. John Wiley & Sons; 2003.
[10] Susanto I, Iriawan N, Kuswanto H, Fithriasari K, Ulama BSS, Suryaningtyas W, Anindya Apriliyanti Pravitasari. On The Markov Chain Monte Carlo Convergence Diagnostic of Bayesian Finite Mixture Model for Income Distribution. In: Journal of Physics: Conference Series. IOP Publishing; 2018. p. 12014.
[11] Darmofal D. Bayesian spatial survival models for political event processes. Am J Pol Sci. 2009;53(1):241–57.
[12] Banerjee S, Wall MM, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics. 2003;4(1):123–42.
[13] Bernardinelli L, Montomoli C. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Stat Med. 1992;11(8):983–1007.
[14] Astuti AB, Iriawan N, Irhamah, Kuswanto H. Bayesian mixture model averaging for identifying the different gene expressions of chickpea (Cicer arietinum) plant tissue. Commun Stat Methods. 2017;46(21):10564–81.
[15] Rufo MJ, Martín J, Pérez CJ. New approaches to compute Bayes factor in finite mixture models. Comput Stat Data Anal. 2010;54(12):3324–35.
[16] Zhao L, Shi J, Shearon TH, Li Y. A Dirichlet process mixture model for survival outcome data: assessing nationwide kidney transplant centers. Stat Med. 2015;34(8):1404–16.
[17] Watanabe S. A widely applicable Bayesian information criterion. J Mach Learn Res. 2013;14(Mar):867–97.
[18] Diamoutene A, Barro D, Somda SMA, Noureddine F, Kamsu-Foguem B. Survival analysis in living and engineering sciences. 2016;