Journal of Mathematics and Statistics, 9 (2): 137-143, 2013
ISSN 1549-3644

doi:10.3844/jmssp.2013.137.143 Published Online 9 (2) 2013 (http://www.thescipub.com/jmss.toc)

GENERALIZED LINEAR MIXED MODELS WITH SPATIAL RANDOM EFFECTS FOR SPATIO-TEMPORAL DATA: AN APPLICATION TO DENGUE FEVER MAPPING

Krisada Lekdee and Lily Ingsrisawang

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand

Received 2013-03-22, Revised 2013-03-23; Accepted 2013-05-13

ABSTRACT

The Generalized Linear Mixed Models (GLMMs) with spatial random effects for spatio-temporal data are proposed. A hierarchical Bayesian method is used for parameter estimation. The random effects are assumed to be normally distributed and the spatial random effects are assumed to be proper Conditional Autoregressive (CAR) models. The proposed models are applied to Dengue fever data in Northern Thailand, including climatic covariates, rainfall and temperature. The Dengue fever maps are constructed from the posterior mean of the mortality rates.

Keywords: Generalized Linear Mixed Models, Conditional Autoregressive Models, Spatial Random Effects Spatio-Temporal Data, Dengue Fever Maps

1. INTRODUCTION

Spatio-temporal data are data collected across both time and space. Thus the data analysis has to take into account the spatial correlation across the areas and temporal correlation within each area. In Thailand, there are annual data reports of common infectious diseases from the Ministry of Public Health (MOPH) of Thailand every year. In the reports, raw data are presented and descriptive statistics, such as rates, percentages and bar charts are usually used to describe the features of those data (MOPH, 2011). It will be more informative and easier for the readers to understand if those data are analyzed thoroughly and presented in maps which are so-called disease maps (Lawson, 2008). Those data motivated us to investigate the models for the spatio-temporal data for disease mapping.

The spatio-temporal models for disease mapping found in prominent papers (Bernardinelli *et al*., 1995; Waller *et al*., 1997; Xia and Calin, 1998; Sun *et al*., 2000; Knorr-Held and Besag 1998; Nobre *et al*., 2005; Martinez-Beneito *et al*., 2008) are based on linear predictors, which may have all terms or a subset of them, expressed as Equation (1):

$$\eta_{ijk} = X_{ijk}^{T}\beta + C_k + S_i + T_j + CS_{ik} + CT_{jk} + ST_{ij} + CST_{ijk} + \varepsilon_{ijk} \quad (1)$$

where, η_{ijk}, i = 1,..,I; j = 1,…,J; k = 1,…,k are linear predictors and β is a p×1 vector of covariates X_{ijk}. C_k, S_i, T_j are additional risks of belonging to group k, living in area i and period j. CS_{ik}, CT_{ik}, ST_{ij} and CST_{ijk} are interaction terms. The ε_{ijk} are error terms. The observed data, Y_{ijk}, typically, are assumed to be Poisson distributed. For example, assuming that Y_{ijk} are Poisson distributed, Waller *et al*. (1997) propose the linear predictors as $\eta_{ijk} = C_k+S_i$ and S_i are broken to be the sum of a heterogeneity, u_i and a spatial effect, v_i, i.e., $S_i = u_i+v_i$. u_i are assumed to be normal distributed and v_i are assumed to be Conditional Autoregressive (CAR) models.

Generalized Linear Mixed Models (GLMMs) have grown in popularity due to their ability to model

Corresponding Author: Krisada Lekdee, Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand

different types of data including spatio-temporal data (Diggle *et al*., 2002; McCulloch *et al*., 2008). In GLMMs, specifically, let Y_{ij}, j = 1,…,n_i be marginally correlated observations from area i = 1,…,m. The GLMMs assume that conditional on random effects b_i, Y_{ij}, j = 1m,…,n_i, are independent and follow a distribution from the exponential family with density $f(Y_{ij}|\eta_{ij}, \varphi) = \exp\{\varphi^{-1}[Y_{ij}\eta_{ij}, -\psi(\eta_{ij})]+c(Y_{ij}, \varphi)\}$ where $\psi(.)$ and c(.) are known functions, η_{ij} are natural parameters and φ is a scale parameter, $E(Y_{ij}|b_i) = u_{ij} = \psi`(\eta_{ij})$ and $Var(Y_{ij}|b_i) = \varphi\psi``(\eta_{ij}) = \varphi\psi``(\psi^{`-1}(\mu_{ij})) = \varphi v(\mu_{ij})$; thus the variances are related to the means through the variance function $v(.) = \psi``(\psi^{`-1}(\mu_{ij}))$. The conditional mean $\mu_{ij} = E(Y_{ij}|b_i)$ can be modeled as Equation (2):

$$g(\mu_{ij}) = X_{ij}^T\beta + Z_{ij}^T b_i \quad (2)$$

With the canonical link function, Equation (2) can be written as Equation (3):

$$\eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i \quad (3)$$

where, β is a p×1 vector of fixed effects, for population, related to covariates X_{ij} and b_i is a q×1 vector of random effects, for each subject, related to covariates Z_{it}. Typically, $b_i \overset{iid}{\sim} N_q(0, D)$ is assumed.

The correlation of two observations related to time is expresses as Equation (4):

$$Corr(Y_{ij}, Y_{ik}) = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij}) + Var(Y_{ik})}} \quad (4)$$

where, $Cov(Y_{ij}, Y_{ik}) = Cov(g^{-1}(X_{ij}^T\beta+Z_{ij}^T b_i), g^{-1}(X_{ik}^T\beta+Z_{ik}^T b_i))$ and $Var(Y_{ij}) = (g^{-1}(X_{ij}^T\beta+Z_{ij}^T b_i))+E(\varphi v(g^{-1}(X_{ik}^T\beta+Z_{ik}^T b_i)))$.

The most common distributions from this family are Binomial, Poisson and Normal which are associated to logit, log and identity canonical link function respectively.

The spatial correlation in GLMMs can be accounted by modifying the random effect part to include the spatial random effects, v_i, for area i, i = 1,…,m, in Equation (3). Thus the GLMMs with spatial random effects are expressed as Equation (5):

$$\eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i + v_i \quad (5)$$

The common way, in disease mapping, is to assign a Conditional Autoregressive (CAR) model first introduced by Besag (1974) to the spatial random effects which are not observed. The CAR model is defined by the conditional probability density function, $v_i \mid v_{-i}, l^1i \sim car.normal\left(\frac{1}{w_{i+}}\sum_{l^1i} w_{il}v_l, \frac{\tau_v}{w_{i+}}\right)$. Since the joint distribution of v_i is improper, a remedy is to introduce a spatial parameter. A suitable constrained, this parameter ensures a proper joint distribution for the CAR model (Gelfand and Vounatsou, 2003). The modified version is so-called a proper CAR model and expressed as $v_i \mid v_{-i}, l^1i \sim car.proper\left(\frac{\rho}{w_{i+}}\sum_{l^1i} w_{il}v_l, \frac{\tau_v}{w_{i+}}\right)$, where W is the adjacency matrix with entries $w_{i1} = 1$ if area i and 1 are neighbors and $w_{i1} = 0$ otherwise, with the diagonal entries $w_{ii} = 0$ and $w_{i+} = \sum_l w_{il}$, τ_v is a conditional variance and its magnitude determine the amount of spatial variation and ρ is the spatial parameter. The most commonly used spatial effects is based on some form of a Conditional Autoregressive (CAR) structure (Clayton and Kaldor, 1987; Cressie and Chan, 1989; Besag *et al*., 1991; Bernardinelli *et al*., 1995; Waller *et al*., 1997; Pascutto *et al*., 2000). There are several methods including a hierarchical Bayesian method used for parameter estimation in GLMMs. Moreover, a hierarchical Bayesian method has been extensively used for CAR models. Thus, in this study, the GLMMs, proper CAR models and hierarchical Bayesian method are adopted for modeling spatio-temporal data in order to construct disease maps.

As mentioned earlier, the infectious disease data motivated us to investigate the models for the spatio-temporal data for disease mapping. Dengue fever, an infectious disease caused by a family of viruses that are transmitted by mosquitoes is chosen to study. It is one of major public health problems in Thailand. In 2011, a total of 69,800 cases and 63 fatalities were reported from all provinces. The morbidity rate was 111.10 per 100,000 people. The Case Fatality Rate (CFR) was 0.0903% (MOPH, 2011). The disease maps are needed as they are communication tools to the general public about which geographic areas and people are at high risk for the Dengue fever.

In this study, we propose GLMMs with proper CAR spatial random effects to analyze Dengue data associated with rainfall and temperature in order to construct

disease maps. The proposed models are specific cases of Equation (1). The random effects are assumed to be normally distributed and the spatial random effects are assumed to be proper CAR models. The observations are assumed to be Poisson distributed. The proposed models are applied to Dengue fever data in Northern Thailand.

The subsequent sections are as follows. Firstly, for materials and methods, GLMMs with proper CAR spatial random effects are described and an application to Dengue fever data is illustrated. Then the results of data analysis are presented. Finally, the discussion and conclusion are conducted respectively.

2. MATERIALS AND METHODS

The observed data, Y_{ij}, the numbers of patients in area i, I = 1,…,m, at time j, j = 1,…,n_i, are assumed to be Poisson distributed with a natural log link function. The proposed models based GLMMs and proper CAR models are expressed as Equation (6):

$$\eta_{ij} = \log(\mu_{ij}) = X_{ij}^T\beta + Z_{ij}^T b_i + v_i \quad (6)$$

For each area, Equation (6) can be written as $\eta_I = \beta+Z_ib_i+1v_i$ where $\mu = E(Y_i|b_i, v_i)$. The sizes of vectors or matrices are: η_i ($n_i\times1$), Y_i ($n_i\times1$), X_i ($n_i\times p$), β ($p\times1$), Z_i ($n_i\times q$), b_i ($q\times1$) and 1 ($n_i\times1$). For all areas, Equation (6) can be written as $\eta = X\beta+Zb+Sv$ where $\mu = E(Y|b,v)$. Let $n = \sum_{i=1}^{m} n_i$. The sizes of vectors or matrices are: $\eta(n\times1)$, $Y(n\times1)$, $X(n\times1)$, $\beta(p\times1)$, $Z(n\times mq)$, $b(mq\times1)$, $S(n\times m)$ and $v(m\times1)$. $b\sim N_n(0,B)$ where $B_{n\times n} = diag[D,\ldots,D]$. $V\sim car.propr(\mu_v,\tau_v(D_w-\rho W)^{-1})$ where $D_W = diag[w_{i+},\ldots,w_{m+}]$. Without loss of generality, V can be reparameterized by including a sum to zero constraint on v_i. Then we get $V\sim car.proper(0,\tau_v(D_w-\rho W)^{-1})$. Under a hierarchical Bayesian framework, the specification of prior distributions, in particular for variance components, is required. In the absence of subjective prior information, a prior on each $\beta$is assumed to be normally distributed with zero mean and large variance, uniform (0,1) for ρ and inverse Wishart for D. When D is univariate, the inverse Wishart reduces to inverse gamma. A posterior inference can be easily implemented using standard Gibbs sampling Markov Chain Monte Carlo (MCMC).

For an application, the Dengue data in 2011 consisted of 68 observations from 17 provinces in Northern Thailand (MOPH, 2011). Each province contributed 4 quarterly observations over time. Let Y_{ij}, j = 1,…,4, i = 1,…,17 denote the numbers of Dengue fever patients in province i at quarter j. In stead of modeling counts, offset variables are used for modeling rates which are cases per population sizes. The model can be expressed as Equation (7):

$$\log(\mu_{ij}) = \log(pop_i) + \beta_0 + \beta_1 rain_{ij} + \beta_2 temp_{ij} + b_i + v_i \quad (7)$$

where, $\log(pop_i)$ are offset variables, pop_i are the numbers of population, $rain_{ij}$ are total rainfall, $temp_{ij}$ are temperature b_i are random intercepts capturing geographically unstructured heterogeneity among provinces and v_i are proper CAR spatial random effects capturing spatial dependence among provinces. Bayesian inference is used to fit the model by assuming a $N(0,\tau_b)$ for b_i, a proper CAR model for the joint distribution of v_i, i.e., $V\sim car.proper(0,\tau_v(D_w-\rho W)^{-1})$ and a uniform (0,1) for ρ. We use independent $N(0m,10^6)$ priors for all fixed effect parameters. $\tau_v\sim$invGamma (0.001,0.001) and$\tau_b\sim$invGamma (0.001,0.001) The MCMC Gibbs sampling was run via Open BUGS. The MCMC convergence is checked by the value of the potential scale reduction of Gelman and Rubin (1992) (Rhat) and visual examination of trace, autocorrelation, history and density plots. The Rhat values substantially closed to 1 indicate convergence.

3. RESULTS

The estimated regression coefficients and variance components are presented in **Table 1**, from MCMC runs of 30,000 iterations after discarding the initial 10,000 iterations as burn-in. The value of Rhat closed to 1 for every parameter in **Table 1** indicates that the MCMC is converged. The result shows that the morbidity risk or Relative Risk (RR) of Dengue fever will be increased when the amount of rainfall increases and also will be increased when the temperature increases. Since the spatial parameter ρ is not closed to zero, there is a spatial association among provinces. However, according to τ_v, the variation among areas is quite small. The high risk provinces and Quarters (Q) are shown in **Table 2**.

The Dengue fever maps for Q1 to Q4 of the provinces in Northern Thailand constructed from the posterior mean of the morbidity rates are shown in **Fig. 1-4**. It is clearly seen that the highest risk are in Q3.

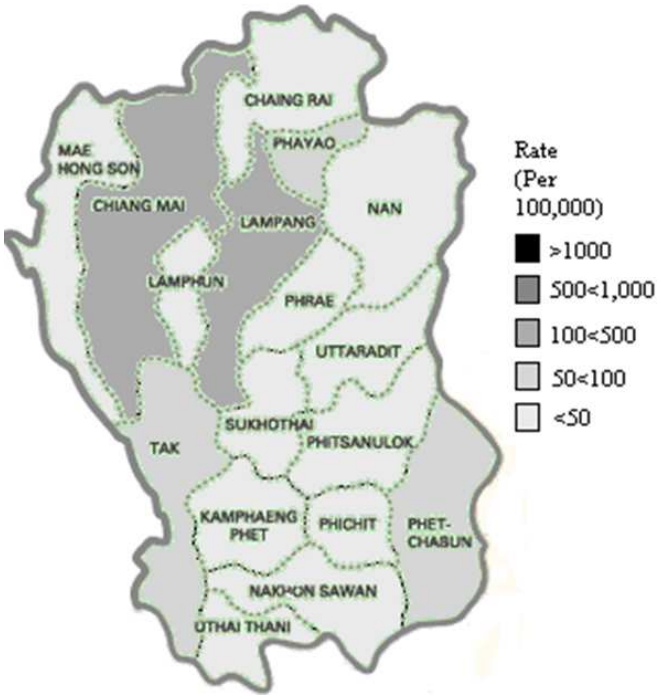

Fig. 1. Dengue fever in Q1

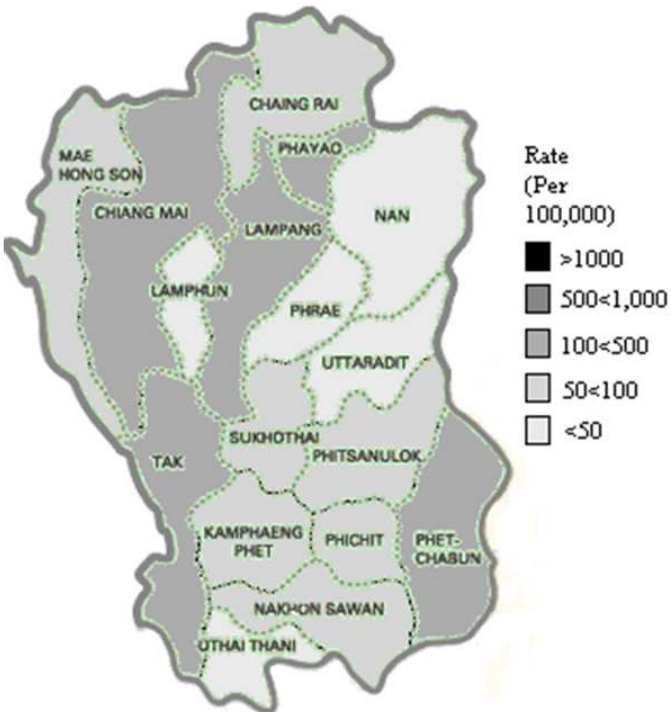

Fig. 2. Dengue fever in Q2

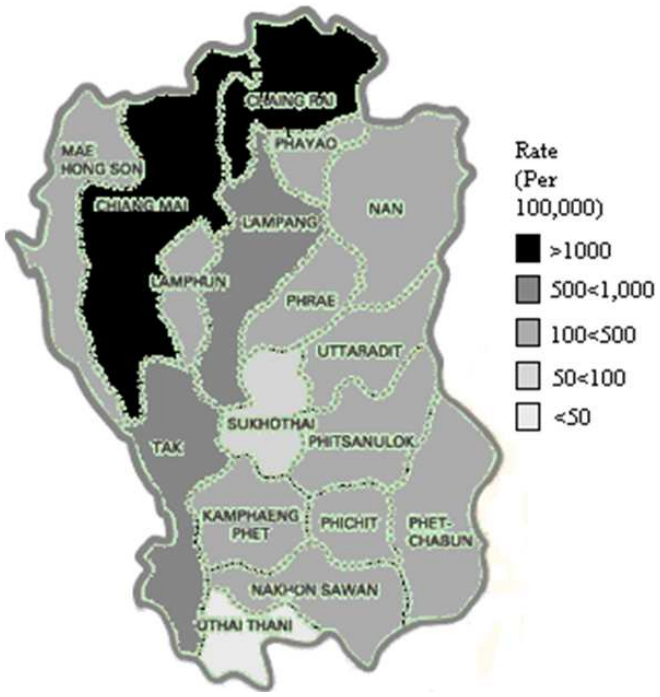

Fig. 3. Dengue fever in Q3

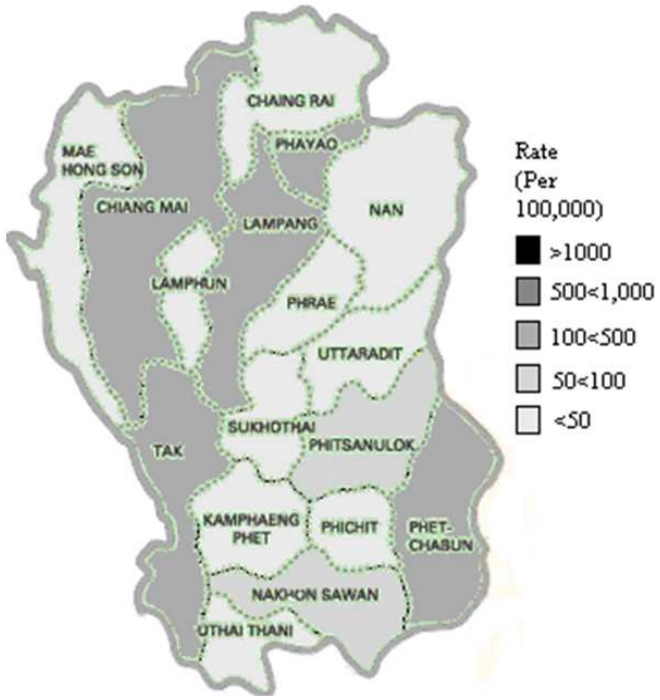

Fig. 4. Dengue fever in Q4

Table 1. Parameter estimate for dengue fever in northern Thailand

Parameter	Post Mean	95% CI***	RR**	Rhat*
Intercept	-0.8423	(-1.5130, 0.1301)	0.4307	1.0446
Rain	0.0105	(0.0063, 0.0147)	1.0106	1.0446
Temperature	0.0783	(0.0663, 0.0902)	1.0814	1.0012
ρ	0.524	(0.0280, 0.9759)	-	1.0010
τ_b	0.0041	(0.0001, 1.9716)	-	1.0023
τ_v	0.0015	(0.0002, 7.0472)	-	1.0059

Table 2. Province and Quarter (Q) at high morbidity rate (per 100,000)

Province and quarter	Post mean of rate	95% CI
Chiang Mai, Q3	2983	(2885, 3084)
Chiang Rai, Q3	1192	(1128, 1259)
Lampang, Q3	550.1	(517.1, 584.5)
Tak, Q3	505.1	(473.9, 537.1)
Chiang Mai, Q2	451.2	(425.7, 477.6)
Phetchabun, Q3	425.9	(395.6, 457.0)
Lamphun, Q3	405.2	(369.6, 442.2)
Phayao, Q3	309.1	(377.6, 431.8)
Mae Hong Son, Q3	309.1	(279.7, 339.8)
Phitsanulok, Q3	297.6	(271.7, 324.6)
Lampang, Q2	283.3	(264.1, 303.0)
Chiang Mai, Q4	281.0	(264.6, 298.1)
Phayao, Q2	270.9	(251.7, 291.2)
Tak, Q2	248.9	(231.3, 267.2)

4. DISCUSSION

The proposed models are based on GLMMs and proper CAR models. A hierarchical Bayesian method is used for parameter estimation. They are different from the model of Waller *et al*. (1997) in the way that the proper CAR models are adopted, instead of improper CAR models. Moreover, the proposed models allow random factors to be included and inverse Wishart would be applied for their variance component. The results of the analysis confirm that the Dengue fever in Northern Thailand were significantly associated with the rain and temperature. The Dengue fever maps clearly show which areas are at high risk in each quarter. They are valuable public health tools to identify the areas where an intervention is required for the Dengue fever control.

5. CONCLUSION

The GLMMs with proper CAR spatial random effects under a hierarchical Bayesian framework for spatio-temporal data are proposed. The proposed models are applied to Dengue fever data in Northern Thailand. The covariates considered are rain and temperature. The Dengue fever maps which are important tools for Dengue fever control are produced using the posterior mean of the mortality rates.

6. ACKNOWLEDGEMENT

We gratefully thank Assoc. Prof. Dr. Yisheng Li for his valuable advice and Mr. Allen Doyle for his kind proofreading. We also thank the Department of Statistics and the Faculty of Science of Kasetsart University for technical support.

7. REFERENCES

Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli and M. Ghislandi *et al*., 1995. Bayesian analysis of space-time variation in disease risk. Stat. Med., 14: 2433-2443. PMID: 8711279

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Royal Stat. Soc. Series B-Statistical Methodol., 36: 192-236.

Besag, J., J. York and A. Mollie, 1991. Bayesian image restoration, with two applications in spatial statistics. Annals Instit. Stat. Math., 43: 1-20. DOI: 10.1007/BF00116466

Clayton, D. and J. Kaldor, 1987. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics, 43: 671-681. PMID: 3663823

Cressie, N. and N.H. Chan, 1989. Spatial modeling of regional variables. J. Am. Stat. Assoc., 84: 393-401. DOI: 10.1080/01621459.1989.10478783

Diggle, P.J., P.J. Heagerty, K.Y. Liang and S.L. Zeger, 2002. Analysis of Longitudinal Data. 2nd Edn., Oxford University Press, New York, ISBN-10: 0198524846, pp: 379.

Gelfand, A.E. and P. Vounatsou, 2003. Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics, 4: 11-15. DOI: 10.1093/biostatistics/4.1.11

Gelman, A. and D.B. Rubin, 1992. Inference from iterative simulation using multiple sequences. Stat. Sci., 7: 457-511.

Knorr-Held, L. and J. Besag, 1998. Modelling risk from a disease in time and space. Stat. Med., 17: 2045-2060. PMID: 9789913

Lawson, A.B., 2008. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. 1st Edn., Taylor and Francis, Boca Raton, Fla., ISBN-10: 1584888415, pp: 368.

Martinez-Beneito, M.A., A. Lopez-Quilez and P. Botella-Rocamora, 2008. An autoregressive approach to spatio-temporal disease mapping. Stat. Med., 27: 2874-2889. PMID: 17979141

McCulloch, C.E., S.R. Searle and J.M. Neuhaus, 2008. Generalized, Linear and Mixed Models. 2nd Edn., John Wiley and Sons, Inc., New Jersey, ISBN-10: 0470073713, pp: 384.

MOPH, 2011. AESR Annual Epidemiological Surveillance Report 2011.

Nobre, A.A., A.M. Schmidt and H.F. Lopes, 2005. Spatio-temporal models for mapping the incidence of malaria in Para. Environmetrics, 16: 291-304. DOI: 10.1002/env.704

Pascutto, C., J.C. Wakefield, N.G. Best, S. Richardson and L. Bernardinelli *et al*., 2000. Statistical issues in the analysis of disease mapping data. Stat. Med., 19: 2493-2519. DOI: 10.1002/1097-0258(20000915/30)19:17/18<2493::AID-SIM584>3.0.CO;2-D

Sun, D., R.K. Tsutakawa, H. Kim and Z. He, 2000. Spatio-temporal interaction with disease mapping. Stat. Med., 19: 2015-2035. DOI: 10.1002/1097-0258(20000815)19:15<2015::AID-SIM422>3.0.CO;2-E

Waller, L.A., B.P. Carlin, H. Xia and A.E. Gelfand, 1997. Hierarchical spatio-temporal mapping of disease rates. J. Am. Stat. Assoc., 92: 607-617. DOI: 10.1080/01621459.1997.10474012

Xia, H. and B.P. Calin, 1998. Spatial-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. Stat. Med., 17: 2025-2043. DOI: 10.1002/(SICI)1097-0258(19980930)17:18<2025::AID-SIM865>3.0.CO;2-M