**Spatio-temporal Bayesian model with two-level spatial structure that includes covariate effect based on mortality data in Nigeria**

U. Abdullahi 1; A. Isah2, K. O. Israel.3, and A. A. Rasheed 4

1,2,4Department of Statistics,

Federal University of Technology, Minna, Nigeria.

3Department of Animal Biology,

Federal University of Technology, Minna, Nigeria.

E-mail: u.abdullahi@futminna.edu.ng

**ABSTRACT**

*Spatio-temporal models suffer from comparability problems of relative risks (RRs) based on the removal of the covariate effect as a confounding factor on the risk estimate of the study population through distribution of standardized mortality ratios (SMRs). This paper proposed an alternative spatio-temporal Bayesian model with two-level spatial structure that included covariate effect.* *The objectives were to estimate posterior means for the model hyper-parameters and compare the performance of the two-level spatial structure model implemented with and without the covariate effect using using female breast cancer mortality data extracted from National System of Cancer Registries (NSCR) in Nigeria, from 2009-2016 through Integrated Nested Laplace Approximation (INLA) estimation procedures. The results showed that, the deviance information criterion (DIC) values for the study models were almost identical, and the posterior estimates for the parameters do not change considerably between the two models*

Keywords: Spatio-temporal modelling, Standardized mortality ratios, deviance information

 criterion, covariate effect

**I. Introduction**

Spatio-temporal models in disease mapping are useful in describing the temporal evolution of geographic patterns of mortality and disease rates. The key to applying spatio-temporal models in disease-mapping studies is to borrow strength from spatial and temporal neighbors to reduce the high variability, which is characteristic of classical disease/mortality risk estimators, such as the standardized mortality ratio (SMR), while dealing with rare diseases or areas with small populations (Ugarte *et al.*, 2014).

Research into spatial and spatio-temporal disease mapping has been carried out within a hierarchical Bayesian framework, with generalized linear mixed models (GLMM) playing a major role. A variety of spatio-temporal models dealing with single-level and two-level spatial random effects have been developed for the analysis of spatio-temporal areal data (Bernardinelli *et al*., (1995b); Knorr-Held, (2000); and Ugarte *et al.,* (2016)) respectively. The literature cited above were based on conditional autoregressive (CAR) models extending the well-known Besag–York–Mollie (BYM) model (Besag *et al*., 1991). Spatio-temporal models with a single level of spatial grouping are the parametric model with linear time trend proposed by Bernardinelli *et al*. (1995b) and the non-parametric model including different types of space-time interactions between the spatial and temporal main effects described by Knorr-Held (2000). The first study dealing with a two-level spatial random effect in spatio-temporal disease mapping was proposed by Schrödle *et al.* (2011) to analysis reported cases of bovine viral diarrhea in Switzerland. Ugarte *et al.,* (2016) proposed a new family of spatio-temporal models where the spatial effect has a two-level structure to analysis brain cancer mortality data in the municipalities of Navarre and the Basque Country.

However, these spatio-temporal models suffer from the comparability problems of relative risks (RRs) as a result of the removal of the covariate effect as a confounding factor on the risk estimate of the study population through distribution of standardized mortality ratios (SMRs). Therefore, this paper is aim to extend and develop a spatio-temporal Bayesian model with two-level spatial structure that include covariate effect. The objectives were to estimate posterior means for the model hyper-parameters and compare the performance of the two-level spatial structure model implemented with and without the covariate effect This paper is motivated by two-level spatial structure models in spatio-temporal disease mapping by Ugarte *et al.,* (2016) and also formed the basis of the proposed model.

**II. Materials and Methods**

**2.1 Data Used**

The study used secondary data on female breast cancer which contains the number of cases and the number of deaths within 60 days of surgery in each hospital in 16 hospital-based cancer registries (HBCRs) in Nigeria (see Figure 1). The data was extracted from the national system of cancer registries (NSCR) from 2009 to 2016**.** The data for each case on individual-level characteristics (that is, age, occupation, marital status and region at the time of diagnosis) were obtained at individual hospital-based cancer registry.

The study area, Nigeria has 36 states and the federal capital territory grouped into geo-political regions called north east, north south, north central, south east, south west and south south as first level areas (FLAs) (see Figure 2(a)). To evaluate the effect of increasing the number of second level areas (SLAs), states are aggregated into health areas (see Figure 2(b)).

****

Figure 1: Map of Nigeria showing areas of cancer registries



Figure 2: Map of the states of Nigeria, (a) states are grouped into 6 geo-political regions (b) states aggregated by SLAs.

**2.2 Methods**

The paper will first review spatio-temporal model with a two-level spatial random effect
by Ugarte *et al.,* (2016) and which formed the basis of the new model that include covariate effects.

**2.2.1 Spatio-temporal model with two-level spatial structure**

Suppose that the region under study is divided into n First-level Areas(FLA) labelled as that can be aggregated into Second-level Areas (SLA) labelled as where. If data are available for several time periods,

Ugarte *et al* (2016) proposed an extension of the Knorr-Held (2000) model to a two-level spatial random effectas:

 (1)

where is an overall risk level,represents the FLA for the areas,represents the SLA for theHealth areas which are in FLAsareas, represents temporally structured effect, and represents SLA space-time interactions.

**2.2.2 Model Formulation**

Suppose that the region under study is divided into first-level areas (FLAs) labeled as
 that can be aggregated into second-level areas (SLAs) labeled as , where . For each area , data are available for different time periods ., Let denote the number of counts of the individual in the FLA region at the time interval, with individual characteristics of interest, where, and. Here, is the number of counts that follows a Poisson distribution. In this model, the mean is defined in terms of a rate and the patient age as for area and time, which can be written as:

 (2)

 (3)

The log-risk is modelled as:

 .. (4)

where is an overall risk level, is the coefficients of covariates on the response, is the value of the covariate for the areas, represents the FLA for the areas, represents the SLA for the Heath Areas which are in FLAsareas, denotes temporal effects and is the SLA space-time interaction effect.

In modelling the covariate effect, univariate normal priors with mean zero and precision are assignedto each of the regression effects*.* The precision parameter is assigned a gamma distribution *.*

For the spatial random effects, gaussian exchangeable distributions are assumed for the spatial random effects and , whereas for the temporal effect , a dynamic neighboring structure using a random walk of order 1, was used. That means that in space, each HA may have its own risk, but all HA within a State region share a common spatial effect, whereas in time, each year has two neighbors, the previous point and the following one, except for the first and last year, which only depends on one. For the interaction term , we assume that the effects and interact assuming an exchangeable distribution, , with the precision parameter of the random effect.

**2.2.3 Assumption(s) of the Proposed Model**

In addition to assumptions of the Ugarte *et al.,* (2016) model, the proposed model assumed that:

1. The mean is defined in terms of a rate and the patient age for area and time.

**III. Results**

The first step in the analyses is by fitting regression model of the covariate effect (patient’s age) categorized into five age-groups, that is, (0-14 years) as Puberty age group and (15 *–* 29 years), (30*−*49 years), (50 *–* 69 years) and (70 + years) as adult’s age groups, on the response variable (female breast cancer mortality) to know the female age-group with higher risk on breast cancer. The spatial and temporal effects were not included in these preliminary analyses.

We examined the *p-value* of each regression age-group, the most significant age groups were (50 *–* 69 years) with the smallest *p-value* (5.00e-16 < 0.05); followed by (30 *–* 49 years) with *p-value* (5.81e-15< 0.05) and (70 + years) with *p-value* (1.60e-11) and the significant age group was (15-29 years) with *p-value* (0.000113). and not significant age groups was (0-14 years) with the smallest *p-value* (0.678686>0.05). The multiple R-squared ( = 0:9999), implies that 99.99% of variation in patients age was well explained by the fitted model and that the regression model fit the sample data. Adjusted R-squared (:9999), implies 99.99% of goodness of fit of the model. To evaluate the impact of adult’s age groups on the risk of mortality, the model in Equation (4) can be reformulated as:

 (5)

**3.1 Model Selection**

To select the best model, the study considered the deviance information criterion (DIC) among several other quantities for model choice. The DIC is the sum of the posterior mean of the deviance, (a measure of goodness of fit) and the number of effective parameters (a measure of model complexity). The DIC is a well-known Bayesian model choice criterion to decide which model provides the best trade-off between model fit and complexity (Spiegelhalter *et al.*, 2002). Models with the smallest DIC value provide the best trade-off between model fit and complexity.

Table 1. Deviance information criterion (DIC) for the models using random walks of first order

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model |  |  |  |  |
| Two -level model without covariateTwo-level model with Covariate | 1506.6461509.407 | 10.340411.1017 | 1516.5071516.493 | 0.16400.1637 |

 **: mean deviance; : effective number of parameters,: deviance information criterion; and : Mean Logarithmic Score.**

Table 1 shows summary of the posterior mean of the deviance, the number of effective parameters and the deviance information criterion (DIC) as a measure of trade-off between model fit and complexity for the wo models. The result obtained in Table 1 suggested that two-level model with covariate effect is the best model and displayed the best fit for the data.

**3.2 Posterior estimate**

The estimated posterior means and quantiles intervals for the model parameters implemented with and without covariate effect and the DIC as a model selection criterion is reported in Table 2. The fixed effects *(,* , and } estimated as relative risks: an increase of age in the adult’s age groups (15 *–* 29 years), (30*−*49 years), (50 *–* 69 years) and decrease in the age group (70 + years) with an increase of around 2.64% , 7.41%, 12.86%, and decrease around 9.02% respectively in the risk of female breast cancer mortality..

Table.2: Posterior estimates (means standard deviation (SD)and quantile s) and DIC for the

 Two-level model with and without patients age as covariate

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | SD | 2.5% | 50% | 97.5% |
| Without Patients age as covariate DIC= 1516.507 |
|  | 0.037112.39240.966769.35042.129877.0241106.4027 | 0.35974.41370.364110.09960.362520.119574.5801 | 0.03243.84600.262049.80921.428535.106338.1419 | 0. 05871.2.32520.961069.18802.124073.7112105.3460 | 0.073421.32641.704489.83242.8647114.7685257.0629 |
|  With Patients age as covariate DIC= 1516.493  |
|  | 0.03830.0264  0.07410.12860.090212.24610.998869.35992.130776.8130106.2911 | 0.2177-0.38320.36590.37980.37204.26740.396210.10910.363419.908474.1643 | 0.3084- 0.81130.6029-0.6387-0.11463.70030.294149.81871.429435.257437.1219 |  0.0597-0.07091.30540.09380.60081.2.17890.993169.19752.124974.0669104.5275 | 0.07660.70432.05220.86371.357921.19011.736589.84192.8656116.0051256.0429 |

This means that the variables on patient’s age explain part of the variability in the risk of breast cancer mortality in Nigeria. Using the DIC values presented in Table 2 as a tool for evaluating the fit of a model, suggests that the model with covariate effect is better suited for the data (with a DIC of 1516.507 against 1516.493 obtained). Though the two DIC values are almost identical, and their posterior estimates for the parameters do not change considerably between the two models 

Figure 3. (a) Map of the spatial pattern of mortality risks and (b) Map of posterior probabilities of the spatial pattern of mortality risk obtained from Equation (1)



Figure 4. (a) Map of the spatial pattern of mortality risks; (b) Map of posterior probabilities of the spatial pattern of mortality risk obtained from Equation (5)

The maps of the spatial pattern of mortality risks and their posterior probability of exceeding 1 are shown in Figures (3) and (4) receptively. Figures (4) is interpreted as the residual mortality risk for each health area after the risk factors (15 *–* 29 years), (30*−*49 years), (50 *–* 69 years) and (70 + years) are taken into account



Figure.5. Map of posterior means of the spatial pattern of mortality risk (a) without covariate effect (b) with covariate effect

Figure 5 shows the map of posterior means of the spatial pattern of mortality risks of the two-level model implemented with covariate effect (as residual relative risks).and without covariate effect Their posterior estimates did not change substantially and the two maps are generally similar in appearance.

**IV. Discussion and Conclusion**

The statistical models used in spatio-temporal disease mapping have, in general, a single level of spatial dependence of random effects. But in some cases, it makes sense to consider natural groupings of the small areas. For example, districts or municipalities can be grouped into provinces or health areas, or counties can be aggregated in states affected by similar health policies. Also, in spatio-temporal disease mapping models, the first step is the removal of the effect of the confounding factors on the risk estimate in the study population through distribution standardization. Standardization of mortality rates (SMRs) or disease incidence is a basic tool in both demography (Rothman *et al*., 2008) and epidemiology (Rothman *et al*., 2008; and Woodward, 2013). This study compared the results of two possibilities of spatio-temporal models. An, additional model was fitted based on covariate effect (patients age) as an alternative to the traditional use of SMRs in these types of models.

The result showed from Table 1, the posterior mean of the deviance, the number of effective parameters and the deviance information criterion (DIC) as a measure of trade-off between model fit and complexity for the two models suggested that two-level model with covariate effect is the best model and displayed the best fit for the data. The result when using the additional model based on covariate effect, model with patients age as covariate performed better than without patients age as covariate with respect to the trade-off between model fit and complexity based on DIC Figure 4 showed that the two methods are generally similar in appearance.

From the spatio-temporal areal data analyses, there is no substantial difference between the two models If the covariate variables are significant then, the model with covariate effect can be implemented and becomes an alternative to the usual use of standardized mortality ratios (SMRs). As a result of this, the following conclusions are recommended:

1. To implement model with covariate effect, the covariate variable(s) must be significant.
2. As the covariate variable(s) is/are not significant, the usual use of standardized mortality ratios (SMRs) should be implemented

**References**

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995b). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443.

Besag, J., York, J., & Molliė, A. (1991). Bayesian image restoration with application in spatial statistics. Annals of the Institute of Statistical Mathematics, 43 (1): 1–20.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17–18), 2555-2567.

Rothman K. J, Greenland S, Lash T. L, (2008). Modern epidemiology (third editions). Lippincott Williams & Wilkins, 758 pp.

Schrödle, B., Held, L., & Riebler, A. (2011). *Using INLA for the Evaluation of Veterinary Surveillance Data* from Switzerland: a case study. J Roy Stat Soc C; 60: 261-279.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4):583–639.

Ugarte, M. D., Adin, A., & Goicoa, T. (2016). Two-level spatially structured models in spatio-temporal disease mapping. Statistical Methods in Medical Research, 25(4):1080–1100.

Ugarte, M. D., Adin, A., Goicoa, T., and Militino, A. F. (2014). On fitting spatiotemporal disease mapping models using approximate Bayesian inference. *Statistical Methods in Medical Research*, 23(6):507–530.

Woodward, M. (2013). *Epidemiology: study design and data analysis*. Chapman and
Hall/CRC, New York.