Modeling malaria incidence in Sucre state, Venezuela using a Bayesian approach [1]

Desireé Villalta
Universidad Simón Bolívar, Caracas, Venezuela, villalta@cesma.usb.ve

Lelys Guenni
Universidad Simón Bolívar, Caracas, Venezuela

Yasmin Rubio
Universidad de Carabobo,Valencia, Venezuela

Abstract: This paper presents a hierarchical Bayesian Poisson lognormal model for malaria incidence in Sucre state, Venezuela, during the period $1990-2002$. The logarithm of the relative risk of the disease for each county or municipality is expressed as an additive model that includes a multiple regression with social-economic and climatic covariates; a random effect that captures the spatial heterogeneity in the study region and a CAR (Conditionally Autoregressive) component, that recognizes the effect of nearby municipalities in the transmission of the disease each year. For most years the selected model captures well the spatial structure between the relative risks from the nearby municipalities. When a poor model fit is obtained, a t-Student model for the spatial heterogeneity parameter improves model fitting results. From the 15 municipalities in Sucre state during the study period $1990-2002$, 7 of them presented high relative risks (greater than 1) in most years. These areas are mostly agricultural areas with poor living conditions.

Keywords: hierarchical Bayesian model, Poisson lognormal model, malaria incidence, Venezuela

1 Introduction

Malaria is a parasitic infectious tropical disease that causes high mortality rates in the tropical belt. In Venezuela, Sucre state is considered the third state with the highest malaria incidence. The *Standardized Mortality Ratio* (SMR), is the ratio between the number of observed disease cases (y_i) and the expected number of cases

[1]Project funded by the National Fund for Science and Technology (FONACIT) project No. 2005-000184, Venezuela.

in the region(E_i), this is, (Banerjee, 2003)

$$SMR_i = \widehat{\Psi}_i = \frac{y_i}{E_i} \qquad i = 1, ..., k \tag{1}$$

where k is the number of subregions (in our case the number of municipalities is 15) and $E_i = p^*.n_i = \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{k} n_i}.n_i$, being p^* the total proportion of disease incidence.

This incidence rate $\widehat{\Psi}_i$ is a raw estimate of the relative risk of disease infestation in the municipality i. A value grater than 1 indicates a disease incidence greater than expected for a region; therefore this constitutes an alarm for public health authorities, (Banerjee, 2003) and (Lawson, 2003). The objective of this work is to propose a model including temporal and spatial components, to explain the dynamics of the disease and to allow simultaneously to identify the explanatory social-economic and climatic variables related with the disease incidence in Sucre state.

2 Materials and Methods

2.1 Study region and Data

The study region is located in the northeastern region of Venezuela in Sucre state. This state has 15 municipalities with an area of $11,800 km^2$. Total cases of malaria were available for 13 years during the period $1990 - 2002$. Interpolated monthly precipitation was available for the whole state using a Bayesian Kriging approach (Le and Zidek, 2007). Several social-economic variables measuring basic needs coverage, unemployment rate, housing characteristics and public services were available from the National Institute of Statistics (INE). After a dimensional reduction technique based on principal component analysis (PCA), the following covariates were used from the PCA results:X_1: Percentage of households with fair building quality and lack of public services (electricity, sewerage, drinking water); X_2: Percentage of poor households with intermediate building quality; X_3: Sewerage and drinking availability; X_4: Percentage of population in agricultural activities. Additionally, the maximum monthly precipitation during the year, X_5, was also included. Each variable was stored in a matrix of dimension of 15×13.

2.2 Spatio-temporal model

Let Y_{it} the number of malaria cases in municipality i and year t. A Poisson model is usually assumed for these quantities, where the mean rate is $\lambda_{it} = E_{it}\, \Psi_{it}$. Therefore,

$$Y_{it} \sim Poisson(\lambda_{it}) \tag{2}$$

with $t = (1, ..., T)$, being T the number of years; in this case $T = 13$.

The proposed model for Ψ_{it} is:

$$\Psi_{it} = \exp(\alpha_t + \boldsymbol{\beta}_t.\mathbf{X}_{it} + v_{it} + b_{it}) \tag{3}$$

where $v_{it} \sim N(0, \frac{1}{\tau_{ht}})$ is a parameter representing the local spatial heterogeneity of the data and $b_{it}|b_{-it} \sim N\left(\bar{b}_{it}, \frac{1}{\tau_{bt} m_{it}}\right)$ is the Conditional Auto-Regressive (CAR) component representing the spatial dependence among the neighboring counties in the transmission of the disease. For model 3, we have the vectors $\alpha_t = (\alpha_1, \alpha_2, ..., \alpha_T)$, $\beta_t = (\beta_1, \beta_2, ..., \beta_T)$, $\tau_{ht} = (\tau_{h1}, \tau_{h2}, ..., \tau_{hT})$, $\tau_{bt} = (\tau_{b1}, \tau_{b2}, ..., \tau_{bT})$, $b_{it} = (b_{1t}, b_{2t}, ..., b_{kt})$, $v_{it} = (v_{1t}, v_{2t}, ..., v_{kt})$ and $\mathbf{X}_{it}$ is the covariates matrix.

As an alternative model, the spatial heterogeneity parameter v_{it} can also be assumed to have a $t - Student$ distribution. The complete conditional posterior probability distributions were calculated for parameters α_t, β_t, b_{it}, v_{it}, τ_{bt},τ_{ht}.

The prior distributions for the parameters α_t, β_t, v_{it}, b_{it}, τ_{ht}, τ_{bt} of model 3, were assumed as follows: α_t and β_t are assumed Uniformly distributed; $b_{it}|b_{-it} \sim N\left(\bar{b}_{it}, \frac{1}{\tau_{bt} m_{it}}\right)$; $\tau_{ht} \sim Gamma(a_h, d_h)$ and $\tau_{bt} \sim Gamma(a_c, d_c)$, where parameters $a_h = a_c = 0.5$, $d_h = d_c = 0.0005$; b_{-it} is the parameter vector without considering the municipality i at time t; and m_{it} are the neighbors to municipality i at time t; although the number of municipalities does not change with time, we use the above notation.

3 Results

A computer code in WinBUGS was implemented for Bayesian inference using MCMC methods. Fourteen thousand samples from the parameter posterior distributions were obtained and $4,000$ samples were used for burnin. Several models were proposed by using different sets of covariates and the lognormal models with and without the CAR component (b_{it}) were also compared. The Deviance Information Criteria (DIC) (Spielgelhalter et al., 2002), and the Minimum Posterior Expected Loss Criteria (D) (Gelfand and Ghosh, 1998) were used for model selection. The DIC criteria did not show important variations among models. The D critera was more sensitive to model variations and suggested that a model with a CAR component and variables X_1, X_2, X_4 and X_5 was more appropriate, since this model presented the lowest D value. Model residuals for the selected model were tested for independence by calculating the Moran's I posterior probability interval for all years.

Posterior predictive model checks were carried out by simulating $2,000$ replicates from the posterior predictive distribution for each municipality and each year. The posterior predictive p-value $p(y_{it}^{rep} \leq y_{it}^{obs})$ was calculated to compare the observed vs. simulated values. If the p-value is close to 0 or 1, it means that the observed values are very unlike to be seen from the simulated values.

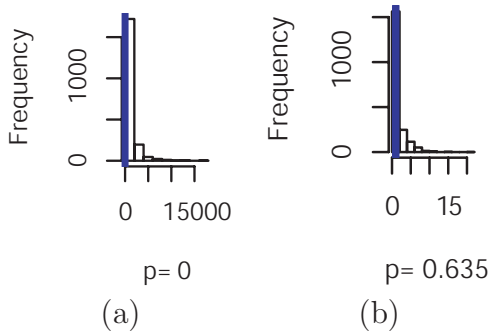

Figure 1: Posterior predictive check for the county Cruz Salmeron Acosta, year 1997 and calculated p-value, by using a normal model (a) and a $t - Student$ model (b) for the spatial heterogeneity parameter $v_{[i]}$

Model checks were satisfactory for most years and all municipalities, except for year 1997 with a good model fit only in 8 of 15 municipalities. To improve model fitting it was assumed $v[i] \sim t - Student(1, \xi, 2)$ where $\xi \sim Gamma(0.5, 0.005)$ for each municipality during year 1997. Figure 1 shows a comparison of the two posterior predictive p-values, with the normal distribution ($p - value = 0$) and the $t - Student$ distribution ($p - value = 0.635$).

From the 15 municipalities in Sucre state during the study period $1990 - 2002$, 7 of them presented relative risks greater than 1 in most years. These areas are mostly agricultural areas with poor living conditions.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. (2003) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall / CRC.

Gelfand, A.E., Ghosh, S. K. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85, pp. 1-11.

Lawson, A. B., Browne, W. J., and Rodeiro, C. L. V. (2003) *Disease Mapping with* WinBUGS *and MLwiN*, Wiley, New York.

Le, N. D., Zidek, J. (2006). Statistical Analysis of Environmental Space-Time Processes. Springer.

Spielgelhalter, D. J., Best, N., Carlin, B. P., and Van der linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J Roy. Statist.Soc., Ser. B*,64, 583-639.