Bayesian Modeling Approach in Big Data Contexts: an Application in Spatial Epidemiology

Erick Orozco-Acosta, Aritz Adin and María Dolores Ugarte
Department of Statistics, Computer Science and Mathematics and InaMat^2. Public University of Navarre, Spain
Email: erick.orozco@unavarra.es, aritz.adin@unavarra.es, lola@unavarra.es

***Abstract*—In this work we propose a novel scalable Bayesian modeling approach to smooth mortality risks borrowing information from neighbouring regions in high-dimensional spatial disease mapping contexts. The method is based on the well-known "divide and conquer" approach, so that the spatial domain is divided into D subregions where local spatial models can be fitted simultaneously. Model fitting and inference has been carried out using the integrated nested Laplace approximation (INLA) technique. Male colorectal cancer mortality data in the municipalities of continental Spain have been analyzed using the new model proposals. Results show that the new modeling approach is very competitive in terms of model fitting criteria when compared with a global spatial model, and it is computationally much more efficient.**

***Index Terms*—Disease mapping, High-dimensional data, INLA, Parallel computing**

I. Introduction

Disease mapping is the field of spatial epidemiology that studies the link between geographic locations and the occurrence of diseases, focusing on the estimation of the spatial and/or spatio-temporal distribution of disease incidence or mortality patterns. The great variability inherent to classical risk estimation measures, such as standardized mortality/incidence ratios or crude rates, makes necessary the use of statistical models to estimate smooth spatial risk surfaces borrowing information from neighbouring regions. The information adquired from these analyses is invaluable for health researchers and policy-makers as it helps to formulate hypothesis about the etiology of a disease, to look for risk factors and and also to allocate funds efficiently in hot spot areas. However, scalability of the models, i. e. the usefulness of the models when the number of small areas increases considerably, is an aspect that has not been studied much.

II. Spatial models in Disease mapping

Let us assume that the spatial domain of interest is divided into n contiguous small areas labeled as $i = 1, \ldots, n$. For a given area i, let O_i and E_i denote the observed and expected number of disease cases, respectively. Using these quantities, the *standardized mortality/incidence ratio* (SMR or SIR) is defined as the ratio of observed and expected cases for the corresponding areal unit. Although its interpretation is very simple, these measures are extremely variable when analyzing rare diseases or low-populated areas, as it is the case of high-dimensional data. To cope with this situation, it is necessary to use statistical models that stabilize the risks (rates) borrowing information from neighbouring regions.

Generalized linear mixed models (GLMM) are typically used for the analysis of count data within a hierarchical Bayesian framework. Conditional to the relative risk r_i, the number of observed cases in the ith area is assumed to be Poisson distributed with mean $\mu_i = E_i r_i$. That is,

$$\begin{aligned} O_i|r_i &\sim Poisson(\mu_i = E_i r_i),\ i = 1, \ldots, n \\ \log \mu_i &= \log E_i + \log r_i, \end{aligned}$$

where $\log E_i$ is an offset. Depending on the specification of the log-risks different models are defined. Here we assume that

$$\log r_i = \alpha + \xi_i, \tag{1}$$

where α is an intercept representing the overall log-risk and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$ is a spatial random effect for which a conditional autoregressive (CAR) prior is usually assumed. The spatial correlation between CAR random effects is determined by the neighbouring structure (represented as an undirected graph) of the areal units. Let $\mathbf{W} = (w_{ij})$ be a binary $n \times n$ adjacency matrix, whose ijth element is equal to one if areas j and k are defined as neighbours (usually if they share a common border), and it is zero otherwise. Here, the prior distribution proposed by Leroux et al. [1] has been considered, which is given by

$$\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{Q}_\xi^{-1}), \quad \text{where} \quad \mathbf{Q}_\xi = \tau_\xi[\lambda_\xi(\mathbf{D}_W - \mathbf{W}) + (1-\lambda_\xi)\mathbf{I}_n]$$

and $\tau_\xi = 1/\sigma_\xi^2$ is the precision parameter, $\lambda_\xi \in [0, 1)$ is a spatial smoothing parameter, $\mathbf{D}_W = diag(w_{1+}, \ldots, w_{n+})$ and $w_{i+} = \sum_j w_{ij}$ is the ith row sum of $\mathbf{W}$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. We will refer to this model as the *Global model*.

III. Scalable Bayesian model proposal

Instead of considering a global spatial random effect whose correlation structure is based on the whole neighbourhood graph of the areal units, we propose to divide the main spatial domain into D subregions so that local spatial models can be simultaneously fitted in parallel reducing the computational time substantially. Two different models are proposed based on the partition of the geographical units.

A. Disjoint models

Let consider a partition of the spatial domain $\mathcal{D}$ into D subregions, that is $\mathcal{D} = \bigcup_{d=1}^{D} \mathcal{D}_d$ where $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$ for all $i \neq j$. In our disease mapping context, this means that each geographical unit belongs to a single subregion. A

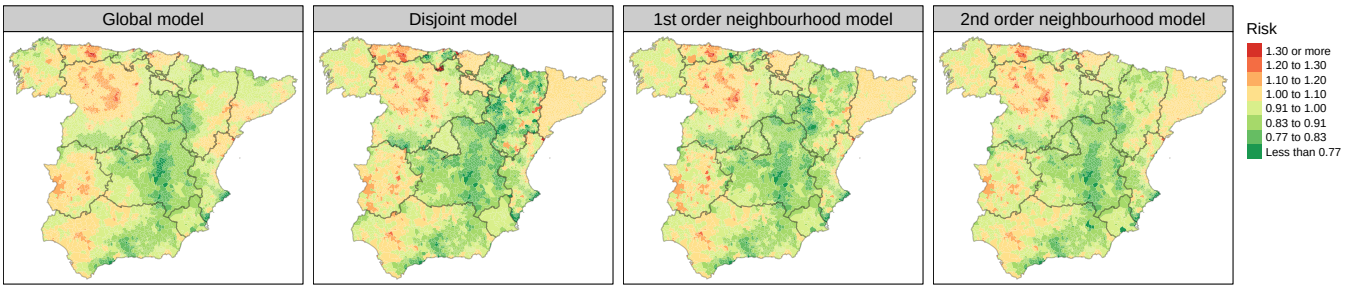

Fig. 1. Maps of posterior median estimates for r_i of male colorectal cancer mortality data in Spanish municipalities during the period 2006-2015.

natural choice for this partition could be the administrative subdivisions of the area of interest (such as for example, provinces or states). Then, for $d = 1, \ldots, D$ the log-risks of the *Disjoint models* are expressed in matrix form as

$$\begin{aligned} \log \mathbf{r}_d &= \alpha_d + \boldsymbol{\xi}_d, \\ \boldsymbol{\xi}_d &\sim N\left(\mathbf{0}, [\tau_{\xi_d}(\lambda_{\xi_d}(\mathbf{D}_{W_d} - \mathbf{W}_d) + (1 - \lambda_{\xi_d})\mathbf{I}_{n_d})]^{-1}\right) \end{aligned}$$

where α_d is an intercept, $\boldsymbol{\xi}_d = (\xi_1^d, \ldots, \xi_{n_d}^d)'$ is the vector of spatial random effects within each subregion with a LCAR prior distribution, $\mathbf{W}_d$ is the neighbourhood subgraph of the areas belonging to $\mathcal{D}_d$, and $\mathbf{I}_{n_d}$ is the identity matrix of dimension n_d, with $\sum_{d=1}^{D} n_d = n$. Since we have defined a partition of the spatial domain $\mathcal{D}$, the log-risk surface $\log \mathbf{r} = (\log \mathbf{r_1}, \ldots, \log \mathbf{r_D})'$ is just the union of the posterior estimates of each submodel.

B. K-order neighborhood models

Assuming independence between areas belonging to different subregions could be very restrictive and may lead to border effects in the disease risk estimates. To avoid this undesirable issue, we also propose a second modeling approach where k-order neighbours are added to each subregion of the spatial domain. Notice that doing this, the main spatial domain $\mathcal{D}$ is now divided into overlapping set of regions, that is, $\mathcal{D} = \bigcup_{d=1}^{D} \mathcal{D}_d$ but $\mathcal{D}_i \cap \mathcal{D}_j \neq \emptyset$ for neighbouring subregions. In consequence, for some areal units multiple relative risk estimates will be obtained. To obtain a unique posterior distribution of r_i for each areal unit i, we propose to compute a mixture distribution of the estimated posterior probability density functions obtained from different models.

IV. Data analysis: Colorectal cancer in Spain

We illustrate the models's behaviour by estimating colorectal cancer mortality risks in Spanish municipalities during the period 2006-2015, where the $D = 15$ Autonomous Regions of Spain are used as a partition of the spatial domain. Model fitting has been carried out using the well-known integrated nested Laplace approximation (INLA) [2] technique for Bayesian inference through the `R-INLA` package. The results are shown in Table I. The computational time for the scalable model proposals are divided into: 1) *running time*, which corresponds to the maximum time of the $D = 15$ submodels (that is, assuming that all models have been simultaneously fitted), and 2) *merging time*, corresponding to the computation of the mixture distribution of the risks and the approximate DIC and WAIC values. As expected, the complexity and computational time of the models increases as higher values of neighbourhood order are considered. Besides the significant reduction in the computational time required to fit the models in INLA, the model selection criteria suggest that the new model proposals outperform the *Global model* in this real data analysis. The maps with posterior median estimates of r_i are shown in Fig. 1.

TABLE I
Model selection criteria and computational time.

Model	DIC	WAIC	T.run	T.merge	T.total
Global	27216.1	27237.9	1929	–	1929
Disjoint	27167.5	27166.7	110	26	136
1st order neighbourhood	27167.6	27170.5	132	63	195
2nd order neighbourhood	27174.3	27183.3	166	83	249

V. Conclusions

Similar spatial patterns are observed for all the models, but *2nd order neighbourhood models* seem to be the most similar to the *Global model*.

The scalable model proposal described in this work has been implemented in the `bigDM` package. The main potential of this methodology is its extension to the spatio-temporal domain. The complexity inherent to spatio-temporal interaction models and the even higher dimensionality associated to this type of data, makes necessary the use of scalable techniques for Bayesian inference. We are currently investigating this issue.

Acknowledgment

This work has been supported by Project MTM2017-82553-R (AEI/FEDER, UE)

References

[1] B. G. Leroux, X. Lei, and N. Breslow, "Estimation of disease rates in small areas: A new mixed model for spatial dependence," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer New York, 2000, pp. 179–191.

[2] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 2, pp. 319–392, 2009.