Workshop on Areal Data:

Bayesian scalable models to analyze high-dimensional areal data using the bigDM library

Aritz Adin
aritz.adin@unavarra.es
Departament of Statistics, Computer Science and Mathematics and INAMAT[2]
Public University of Navarre (UPNA)

Workshop on Spatio Temporal Modelling

25-27 June, 2024

Spatial Statistics Public University of Navarre

upna INAMAT[2] Institute for Advanced Materials and Mathematics

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Table of contents

Section 1: Introduction

Types of spatial data

Spatial data can be viewed as the result from observations of a stochastic process

$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\},$$

where $Z(\mathbf{s})$ denotes the attribute we observe at (spatial) location s.

Three types of spatial data are distinguished:

- **Geostatistical (or point-referenced) data**, where $\mathbf{s}$ varies continuously in space over a fixed domain D. Usually, we use data $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ observed at known spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ to predict the values of the variable of interest at unsampled locations.

- **Point pattern data**, where D itself is random; its index set provides the locations of random events that form the spatial point process. $Z(\mathbf{s})$ can simply take the value 1 for all $s \in D$ (indicating the occurrence of an event), or it can include additional information about some variable of interest (referred to as *marked point processes*).

- **Areal (or lattice) data**, where D is a fixed countable collection of (regular or irregular) areal units whose boundaries are clearly defined. Areal data usually arise when the number of events corresponding to some variable of interest are aggregated in areas.

Examples: geostatistical data

Remote sensing data of daytime land surface temperature (LST) and mean maximum temperature (Tmax) in Navarre during the third week of Feb-2014.

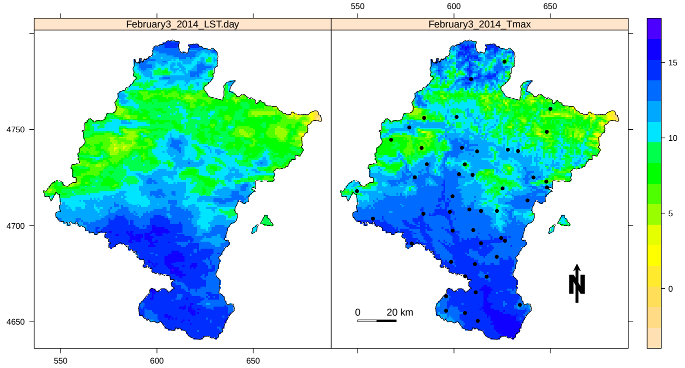


Figure 1: Militino, A., Ugarte, M., and Pérez-Goya, U. (2018). Improving the Quality of Satellite Imagery Based on Ground-Truth Data from Rain Gauge Stations. *Remote Sensing*, 10(3), 398.

Examples: point pattern data

Sky positions of 4215 galaxies within the Shapley Supercluster, recognized as the largest concentration of galaxies in the nearby universe.

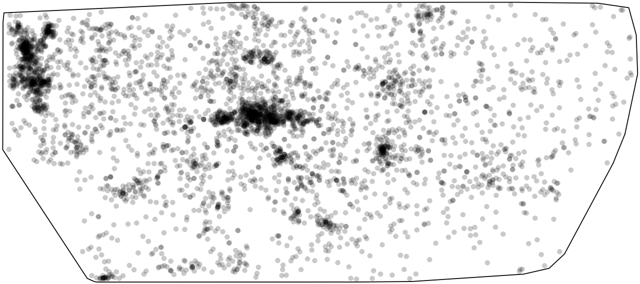


Figure 2: Baddeley, A., Turner, R. (2005). `spatstat`: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software* 12(6), pp. 1-42. Original source: M.J. Drinkwater, Department of Physics, University of Queensland.

Examples: marked point pattern data

Location of forest fires that occurred in the province of Castellón during the years 2001-2006. The circles indicate the size class for each fire: $(0, 1]$ ha, $(1, 6]$ ha, $(6, 11]$ ha, $(11, 16]$ ha, $(16, 21]$ ha, and more than 21 ha.

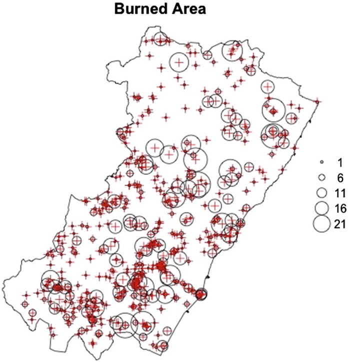
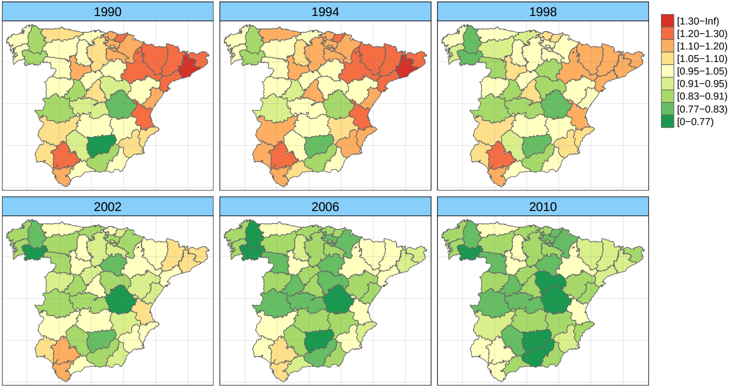


Figure 3: Díaz-Avalos, C., Juan, P., and Serra-Saurina, L. (2016). Modeling fire size of wildfires in Castellon (Spain), using spatiotemporal marked point processes. *Forest Ecology and Management*, 381, pp. 360-369.

Examples: areal data

Spatio-temporal analysis of relative risks of breast cancer mortality in the provinces of Spain during the period 1990-2010.



Introduction to disease mapping

- The development of new techniques and computational algorithms to analyse massive spatial and spatio-temporal datasets is of crucial interest in many fields such as remote sensing, geoscience, ecology, crime research and epidemiology among others.

- Disease mapping is the field of spatial epidemiology that deals with aggregated count data from non-overlapping areal units focussing on the estimation of the geographical distribution of a disease and its evolution in time.

- Three main inferential goals in disease mapping:

  1. To provide estimates of mortality/incidence risks or rates
  2. To unveil underlying spatial and spatio-temporal patterns
  3. To detect high-risk areas or hotspots

- The information acquired from these analyses is of great interest for health researchers, epidemiologists and policy makers.

Introduction to disease mapping

- Classical risk estimation measures such as the standardized mortality ratio (SMR) or crude rates, are extremely variable when analyzing rare diseases (with few cases) or low-populated areas.
- This makes necessary the use of statistical models to smooth risks (or rates) borrowing information from spatial and temporal neighbors.

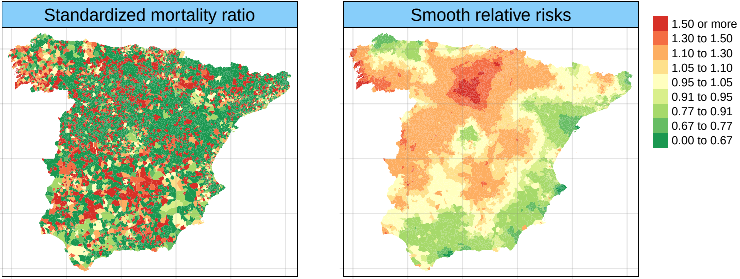


Figure 4: Maps with SMRs and smooth relative risks in the municipalities of Spain.

The joint modelling of several responses offer some advantages:

- it increases the effective sampling size and improves risk smoothing by borrowing strength between diseases
- it allows relationships between the geographical distribution of the diseases (i.e., correlations between spatial patterns)

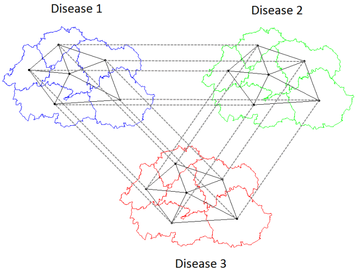


Figure 5: Toy example of a multivariate neighbourhood graph.

Introduction to disease mapping

- Mixed Poisson models including conditional autoregressive (CAR) priors for space and random walk priors for time including space-time interactions (Knorr-Held, 2000) are typical models in space-time disease mapping.

- Other approaches based on reduced rank multidimensional P-splines have been also proposed in this field to deal with univariate and multivariate count data (see for example Ugarte et al., 2017; Vicente et al., 2023b).

- Despite the enormous expansion of modern computers and the development of new software and estimation techniques to make fully Bayesian inference, dealing with massive data is still computationally challenging.

High-dimensional areal data

- **Question**: Are these smoothing methods 'appropriate' when analyzing very large datasets?

- **Two main problematic aspects**
  1. Computational time & resources:

     These methods are built on the idea of spatial/temporal correlation and generally use a covariance or precision matrix with dimension equal to the number of spatial locations $\times$ time points.
  2. Model assumptions:

     CAR models induces the same degree of spatial dependence through the whole adjacency graph (stationary models).

- The **R package bigDM** implements several scalable and non-stationary Bayesian models to smooth mortality or incidence risks in a high-dimensional spatial and spatio-temporal disease mapping context (Orozco-Acosta et al., 2021, 2023; Vicente et al., 2023a).

R package bigDM

- Available at CRAN (stable version) and GitHub (development version).
- The modelling approach is based on the idea of divide-and-conquer so that local models can be fitted simultaneously.
- Inference is fully Bayesian using the well-known integrated nested Laplace approximation (INLA; Rue et al., 2009) technique through the `R-INLA` package.
- Parallel or distributed computation strategies can be performed to speed up computations by using the `future` package (Bengtsson, 2020).

2. Spatial models for (high-dimensional) areal data

Statistical models in spatial disease mapping

Let us assume that the spatial domain of interest is divided into I contiguous small areas labeled as $i = 1, \ldots, I$.

- O_i and E_i denote the number of observed and expected cases, respectively, for the i-th area.
- r_{it} denotes the relative risk of mortality (incidence).

Then,

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

Depending on the specification of $\log r_i$, different models are defined.

Conditional autoregressive (CAR) models

Here we assume that

$$\log r_i = \beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta} + \xi_i \tag{1}$$

- β_0 is a global intercept (representing the overall log-risk).
- $\mathbf{x}_i^{'} = (x_{i1}, \ldots, x_{ip})$ is a p-vector of standardized covariates in the i-th area.
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{'}$ is the p-vector of fixed effect coefficients.
- $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_I)^{'}$ is a spatially structured random effect with a CAR prior distribution.

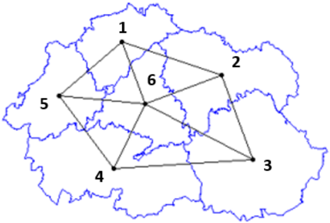
- Incorporating potential risk factors into a model is commonly referred to as ecological regression, and it confers an inferential perspective on areal data models by quantifying the relationship between a response variable and a set of covariates (see, e.g., Martínez-Beneito and Botella-Rocamora, 2019, chapter 5).

- In this type of models, both identifiability and confounding issues must be carefully taken into account.

CAR priors for random effects

- 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 $I \times I$ adjacency matrix with $w_{ij} = 1$ if $i \sim j$ (usually if they share a common border), and 0 otherwise.



$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

CAR priors for random effects

- **Intrinsic CAR (iCAR) prior distribution** (Besag et al., 1991)

$$\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{Q}_{\xi}^{-}), \quad \text{with} \quad \mathbf{Q}_{\xi} = \tau_{\xi}(\mathbf{D}_w - \mathbf{W})$$

where $\mathbf{D}_W$ is a diagonal matrix with $D_{ii} = \sum_{\{j:j\sim i\}} w_{ij}$, and $\tau_{\xi} = 1/\sigma_{\xi}^2$ is a precision parameter.

If the spatial graph is fully connected (matrix $\mathbf{Q}_{\xi}$ has rank-deficiency equal to 1, or equivalently, $\mathbf{Q}_{\xi}\mathbf{1}_I = \mathbf{0}$), a sum-to-zero constraint $\sum_{i=1}^{I} \xi_i = 0$ is usually imposed to solve the identifiability issue between the spatial random effect and the intercept in Model (1).

- **Convolution CAR (or BYM) prior distribution** (Besag et al., 1991)

$$\boldsymbol{\xi} = \mathbf{u} + \mathbf{v}, \quad \text{with} \quad \begin{array}{l} \mathbf{u} \sim N(\mathbf{0}, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-}), \\ \mathbf{v} \sim N(\mathbf{0}, \tau_v^{-1}\mathbf{I}_I) \end{array}$$

The precision parameters τ_u and τ_v are not identifiable form the data (MacNab, 2011), just the sum $\xi_i = u_i + v_i$ is identifiable. Hence, similar to the iCAR prior distribution, the sum-to-zero constraint $\sum_{i=1}^{I}(u_i + v_i) = 0$ must be imposed to solve identifiability problems with the intercept.

CAR priors for random effects

- **Leroux CAR (LCAR) prior distribution** (Leroux et al., 1999)

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

  where τ_{ξ} is the precision parameter and $\lambda_{\xi} \in [0,1]$ is a spatial smoothing parameter.

  Even the precision matrix $\mathbf{Q}_{\xi}$ is of full rank whenever $0 \leq \lambda_{\xi} < 1$, a confounding problem still remains and consequently, a sum-to-zero constraint $\sum_{i=1}^{I} \xi_i = 0$ has to be considered (Goicoa et al., 2018).

- **BYM2 prior distribution** (Riebler et al., 2016)

$$\boldsymbol{\xi} = \frac{1}{\sqrt{\tau_{\xi}}}\left(\sqrt{\lambda_{\xi}}\mathbf{u}_{\star} + \sqrt{1-\lambda_{\xi}}\mathbf{v}\right),$$

  where $\mathbf{u}_{\star}$ is the scaled intrinsic CAR model with generalized variance equal to one (Sørbye and Rue, 2014) and $\mathbf{v}$ is the vector of unstructured random effects.

  Unlike the LCAR model, the variance of $\boldsymbol{\xi}$ is expressed as a weighted average of the covariance matrices of the structured and unstructured spatial components

$$\text{Var}(\boldsymbol{\xi}|\tau_{\xi}) = \frac{1}{\tau_{\xi}}(\lambda_{\xi}\mathbf{R}_{\star}^{-} + (1-\lambda_{\xi})\mathbf{I}_I),$$

  where $\mathbf{R}_{\star}^{-}$ indicates the generalised inverse of the scaled spatial precision matrix.

Scalable models for spatial areal data

References

Bengtsson, H. (2020). A Unifying Framework for Parallel and Distributed Processing in R using Futures.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Goicoa, T., Adin, A., Ugarte, M. D., and Hodges, J. S. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, 32(3):749–770.

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

Leroux, B. G., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Halloran, M. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191. Springer-Verlag: New York.

MacNab, Y. C. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research*, 20(1):49–68.

Martínez-Beneito, M. A. and Botella-Rocamora, P. (2019). *Disease Mapping: From Foundations to Multidimensional Modeling*. CRC Press, Boca Raton.

Orozco-Acosta, E., Adin, A., and Ugarte, M. D. (2021). Scalable Bayesian modeling for smoothing disease mapping risks in large spatial data sets using INLA. *Spatial Statistics*, 41:100496.

Orozco-Acosta, E., Adin, A., and Ugarte, M. D. (2023). Big problems in spatio-temporal disease mapping: methods and software. *Computer Methods and Programs in Biomedicine*, 231:107403.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

Ugarte, M. D., Adin, A., and Goicoa, T. (2017). One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: model fitting