Spatial Statistics 41 (2021) 100496

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

# Scalable Bayesian modelling for smoothing disease risks in large spatial data sets using INLA

Erick Orozco-Acosta[1,2], Aritz Adin[1,2], María Dolores Ugarte[*,1,2,3]

*Universidad Pública de Navarra, Campus de Arrosadia, 31006 Pamplona, Spain*

ARTICLE INFO

*Article history:*
Received 5 October 2020
Received in revised form 21 January 2021
Accepted 21 January 2021
Available online 4 February 2021

*Keywords:*
High-dimensional data
Hierarchical models
Mixture models
Spatial epidemiology

ABSTRACT

Several methods have been proposed in the spatial statistics literature to analyse big data sets in continuous domains. However, new methods for analysing high-dimensional areal data are still scarce. Here, we propose a scalable Bayesian modelling approach for smoothing mortality (or incidence) risks in high-dimensional data, that is, when the number of small areas is very large. The method is implemented in the R add-on package bigDM and it is based on the idea of "divide and conquer". Although this proposal could possibly be implemented using any Bayesian fitting technique, we use INLA here (integrated nested Laplace approximations) as it is now a well-known technique, computationally efficient, and easy for practitioners to handle. We analyse the proposal's empirical performance in a comprehensive simulation study that considers two model-free settings. Finally, the methodology is applied to analyse male colorectal cancer mortality in Spanish municipalities showing its benefits with regard to the standard approach in terms of goodness of fit and computational time.

[*] Correspondence to: Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain.
*E-mail address:* lola@unavarra.es (M.D. Ugarte).

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain.

[2] Institute for Advanced Materials and Mathematics (InaMat$^2$), Public University of Navarre, Spain.

[3] Centro Asociado de Pamplona, UNED.

https://doi.org/10.1016/j.spasta.2021.100496

## 1. Introduction

Statistical models are an essential tool to analyse the geographical or spatial distribution of environmental and epidemiological data in small areas. Nowadays, one of the biggest challenges in the field of spatial statistics is the development of new computationally efficient methods to obtain reliable estimates of the underlying geographical patterns for large data sets. Several modern methods have been proposed to analyse massive geostatistical (point-referenced) data, where traditional estimation of Gaussian processes (GPs) becomes computationally prohibitive. Some of these approaches includes low-rank approximations to GPs such as fixed-rank kriging (Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008), stochastic partial differential equations (Lindgren et al., 2011), lattice kriging (Nychka et al., 2015), multi-resolution approximations (Katzfuss, 2017), and Vecchia approximations (Datta et al., 2016; Katzfuss and Guinness, 2021) among others, plus several parallel computation algorithmic approaches such as Gramacy and Apley (2015), Paciorek et al. (2015), Guhaniyogi and Banerjee (2018), Katzfuss and Hammerling (2017) and Lenzi et al. (2020). Sun et al. (2012) and Banerjee (2017) (and references therein) provide some background and additional recent work on massively scalable spatial processes. However, there is not much research on the scalability of statistical models for areal (lattice) count data.

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 (Lawson et al., 2016; Martínez-Beneito and Botella-Rocamora, 2019). In these studies the region of interest is divided into non-overlapping irregular areal units (administrative divisions such as states or local health areas), where epidemiological data are presented as aggregated disease counts for each geographical unit. The great variability inherent to classical risk estimation measures, such as standardized mortality/incidence ratios or crude rates, makes it necessary to use statistical models to smooth the spatial risk surface. Bayesian hierarchical models are typically used for this objective, where spatially structured random effects are included at the second level of the hierarchy.

Most research into spatial disease mapping is based on the conditional autoregressive (CAR) prior distribution (Besag, 1974), where the spatial correlation between random effects is determined by the neighbouring structure (represented as an undirected graph) of the areal units. Despite the enormous expansion of modern computers and the development of new software and estimation techniques for fully Bayesian inference, dealing with high-dimensional spatial random effects is still computationally challenging.

As far as we know, there are very few papers in the disease mapping literature proposing computationally efficient methods to analyse very large spatial data sets. Hughes and Haran (2013) give a parameterization of the areal spatial generalized linear mixed model that alleviates spatial confounding when including covariates in the model (see for example Reich et al., 2006 and Hodges and Reich, 2010) while speading computation by greatly reducing the dimension of the spatial random effect. To achieve this dimension reduction, they suggest reparameterizing the model by selecting a fixed number of eigenvectors of the Moran operator (those corresponding to the largest eigenvalues to include patterns of positive spatial dependence, i.e., attraction, or those corresponding to the smallest eigenvalues to include patterns of negative spatial dependence, i.e., repulsion). The model is implemented in the R package `ngspatial` (Hughes and Cui, 2020). Bradley et al. (2018) introduce a computationally efficient Bayesian model for predicting high-dimensional dependent count data. In particular, they propose a multivariate log-gamma distribution that leads to computationally efficient sampling of full conditional distributions within a Gibbs sampler. Very recently, Datta et al. (2019) consider a new way of constructing precision matrices for count data models using a directed acyclic graph representation derived from the original spatial neighbourhood structure of the areal units. Instead of modelling the precision matrix of the spatial random effect directly, they propose to model its (sparse) Cholesky factor using autoregressive covariance models on a sequence of local trees created from this directed acyclic graph. Although the model is order-dependent, as stated by authors of this paper, the joint density of the spatial random effect will be scalable for large data sets.

In this paper, we propose a scalable Bayesian modelling approach for smoothing mortality (or incidence) risks for high-dimensional spatial disease mapping data, that is, when the number of small areas is very large. Our method is based on the well-known "divide and conquer" approach. Instead of considering a global spatial random effect whose correlation structure is based on the whole neighbourhood graph of the areal units, the spatial domain is divided into $D$ subregions so that local spatial models can be fitted simultaneously (in parallel). Two different models are given based on the partition of the geographical units. The first model assumes that the spatial domain is divided into $D$ disjoint subregions, according to administrative subdivisions for example. Then, independent spatial models are fitted to each data subset based on the neighbourhood structure of the corresponding subgraphs. Once computations are finished, the area-specific relative risks are merged to obtain a single spatial risk surface. Clearly, assuming independence between areas corresponding to different subregions of the partition of the spatial domain could lead to border effects in risk estimates. To avoid this undesirable issue, we also propose a second modelling approach where $k$-order neighbours are added to each subregion of the spatial domain. In consequence, the main spatial domain is divided into overlapping partitions. This means that some areal units will have several risk estimates. To obtain a unique posterior distribution for these risks, we compute the mixture distribution of the estimated posterior probability density functions. In addition, approximate values for some model selection criteria are derived to perform Bayesian model comparison.

Although the methodology described in this paper could possibly be adapted to other Bayesian estimation techniques, here we use INLA (Rue et al., 2009) as it has several advantages: it is well spread and has been used in several fields (see for example Rue et al., 2017 and references therein), it is computationally efficient when fitting disease mapping models, and it is fairly easy for practitioners to handle if they are not experts in statistics.

A simulation study is conducted to compare the new scalable models against the global model using the almost 8000 municipalities of continental Spain. This study reveals a competitive performance of the new models in terms of goodness of fit and computational time, that is reduced substantially. In addition, as we increase the neighbourhood ordering ($k$ parameter) in our second modelling approach, results are more similar to the global model, but this comes with a loss of computational efficiency. The new methodology is used to analyse male colorectal cancer mortality in the Spanish municipalities.

The rest of the paper is organized as follows. In Section 2 we briefly review some spatial models in disease mapping and give some details about different Bayesian inferential techniques. Section 3 introduces the new scalable models to fit high-dimensional areal count data. In Section 4 a simulation study is conducted to compare the performance of our modelling approach with the usual spatial model for areal count data. Male colorectal cancer mortality data in Spanish municipalities are analysed in Section 5. The paper concludes with a discussion and some conclusions. The methods and algorithms proposed here are implemented in the R package `bigDM` available at https://github.com/spatialstatisticsupna/bigDM, which contains a vignette to replicate the data analysis described in this paper using a simulated colorectal cancer mortality data (modified in order to preserve the confidentiality of the original data).

## 2. Spatial models for disease mapping

Let us assume that the spatial domain of interest is divided into $n$ contiguous small areas labelled as $i = 1, \ldots, n$. For a given area $i$, $O_i$ will denote the observed number of disease cases and $N_i$ the population at risk. The simplest mortality/incidence indicator is the *crude rate*, which is usually defined as the number of cases per 100,000 people, that is, $CR_i = \frac{O_i}{N_i} \times 100,000$. When the study aims to detect which areas exhibit elevated or lowered risk, the number of expected cases in each small area are usually computed. For example, if the population is divided into age-groups, the indirect standardization method is commonly used to calculate the expected number of cases as $E_i = \sum_{j=1}^{J} N_{ij} \frac{O_j}{N_j}$ for $i = 1, \ldots, n$, where $O_j = \sum_{i=1}^{n} O_{ij}$ and $N_j = \sum_{i=1}^{n} N_{ij}$ are the number of cases and the population at risk in the $j$th age-group, respectively. Note that $E_i$ represents the number of cases we expect to observe in the $i$th area if it behaves as the whole study region. 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 (areas with values higher than 1 will stand for an excess of risk, while values lower than 1 mean a lower risk for the population in that unit), these measures are extremely variable when analysing rare diseases or low-populated areas, as 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 $i$th area is assumed to be Poisson distributed with mean $\mu_i = E_i r_i$. That is,

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

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 $\alpha$ is an intercept representing the overall log-risk and $\xi_i$ is a spatial random effect. Commonly, a conditional autoregressive (CAR) prior distribution is assumed for the random effect $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$, which is a type of Gaussian Markov random field (GMRF) (Rue and Held, 2005). A GMRF, with respect to a given graph, is defined on a vector $\boldsymbol{\xi}$ by assuming a multivariate Normal distribution $\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}^{-1} = \mathbf{Q}$ is a $n \times n$ sparse precision matrix corresponding to the undirected graph of the regions under study. In what follows, we briefly review some of the most commonly used CAR priors for spatial random effects. Let $\mathbf{W} = (w_{ij})$ be a binary $n \times n$ adjacency matrix, whose $ij$th element is equal to one if areas $j$ and $k$ are defined as neighbours, usually if they share a common border (denoted as $i \sim j$), and it is zero otherwise. The joint distribution of the *intrinsic CAR* prior (iCAR) (Besag et al., 1991) is defined as

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

where $\mathbf{D}_W = diag(w_{1+}, \ldots, w_{n+})$ and $w_{i+} = \sum_j w_{ij}$ is the $i$th row sum of $\mathbf{W}$, and $\tau_{\xi} = 1/\sigma_{\xi}^2$ is the precision parameter. The symbol $^{-}$ denotes the Moore–Penrose generalized inverse of a matrix. As $\mathbf{Q}_{\xi}\mathbf{1}_n = \mathbf{0}$, where $\mathbf{1}_n$ is a vector of ones of length $n$ (i.e., $\mathbf{1}_n$ is the eigenvector associated to the null eigenvalue of $\mathbf{Q}_{\xi}$), the precision matrix of the iCAR distribution is singular and therefore, the joint distribution of $\boldsymbol{\xi}$ is improper. If the spatial graph is fully connected (matrix $\mathbf{Q}_{\xi}$ has rank-deficiency equal to 1), a sum-to-zero constraint $\sum_{i=1}^{n} \xi_i = 0$ is usually imposed to solve the identifiability issue between the spatial random effect and the intercept in Model (1).

The iCAR prior distribution only accounts for spatial correlation structures, and hence, it is not appropriate if the data variability is not only spatially structured but unstructured heterogeneity is also present. A *convolution* prior was also proposed by Besag et al. (1991) to deal with this situation (usually named as BYM prior) that combines the iCAR prior and an additional set of unstructured random effects. The model is given by

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

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The precision parameters of the spatially structured random effect ($\tau_u$) and the unstructured random effect ($\tau_v$) are not identifiable from 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 $\sum_{i=1}^{n}(u_i + v_i) = 0$ must be imposed to solve identifiability problems with the intercept.

Leroux et al. (1999) propose an alternative CAR prior (hereafter named as LCAR prior) to model both spatially structured and unstructured variation in a single set of random effects. It is given by

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

where $\tau_{\xi}$ 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 $\sum_{i=1}^{n} \xi_i = 0$ has to be considered (see Goicoa et al., 2018).

Other conditional autoregressive priors have been also given in the literature, like the *proper CAR* prior distribution described in Cressie (1993), or the reparameterization of the BYM model given by Dean et al. (2001).

### 2.1. Model fitting and inference

The fully Bayesian approach is probably the most-used technique for model fitting and inference in spatial disease mapping. Under this framework, the posterior probability distribution of the parameters of interest is obtained. Traditionally, Markov chain Monte Carlo (MCMC) techniques have been used for model inference from a fully Bayes perspective, mainly due to the development and accessibility of the well-known `WinBUGS` (Spiegelhalter et al., 2003) software. Over the last few years, other software based on MCMC methods has been popularized such as JAGS (Plummer et al., 2003) or STAN (Carpenter et al., 2017; Team, 2018), as well as other new statistical systems such as NIMBLE (de Valpine et al., 2020). An alternative to MCMC simulation methods for Bayesian inference was proposed by Rue et al. (2009). The method known as INLA is based on integrated nested Laplace approximations and numerical integration. The main goal of the INLA strategy is to approximate the marginal posterior distribution of a GMRF using numerical methods for sparse matrices to speed up computations in comparison with MCMC methods. This technique can be used easily in the free software `R` through the `R-INLA` package (http://www.r-inla.org/). The use of INLA for Bayesian inference has turned out to be very popular in applied statistics in general (see Rue et al., 2017), and in the field of spatial statistics in particular (Bakka et al., 2018).

Despite the computational efficiency of INLA for Bayesian inference when fitting spatial and spatio-temporal disease mapping models for areal data, its use has not been studied in detail when the number of areas increases considerably. New parallelization strategies have been recently implemented in INLA through the integration of a special version of the PARDISO (www.pardiso-project.org) library (van Niekerk et al., 2019). However, the computational resources needed for analysing massive spatial data could be enormous, something that is not within the reach of researchers in statistics, epidemiologists or public health professionals. Thus, the main objective of this paper is to provide an alternative scalable method to perform high-dimensional spatial analysis for count data with INLA.

## 3. Scalable Bayesian models for areal count data

In this section, we propose a scalable Bayesian modelling approach for smoothing mortality (or incidence) risks for high-dimensional spatial disease mapping data. Our proposal is based on applying the "divide and conquer" approach to the spatial model described in Eq. (1), which will be named as the *Global model*. The key idea is to divide the spatial domain into $D$ subregions so that local spatial models can be simultaneously fitted in parallel reducing the computational time substantially. The LCAR prior distribution has been considered for the spatial random effect $\boldsymbol{\xi}$, but any other CAR distribution such as those described in Section 2 could be used instead in the methodology described below.

### 3.1. 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 natural choice for this partition could be the administrative subdivisions of the area of interest (such as for example, provinces or states).

Let $\mathbf{O}_d = \{O_i |$ area $i \in \mathcal{D}_d\}$ and $\mathbf{E}_d = \{E_i |$ area $i \in \mathcal{D}_d\}$ represent the observed and expected number of disease cases in each subregion, respectively. It is important to remark that the expected values are computed using all the data. Then, for $d = 1, \ldots, D$ the log-risks of the *Disjoint models* are expressed in matrix form as

$$\begin{aligned} \log \mathbf{r}_d &= \mathbf{1}_{n_d}\alpha_d + \boldsymbol{\xi}_d, \\ \boldsymbol{\xi}_d &\sim N\left(\mathbf{0}, [\tau_{\xi_d}(\lambda_{\xi_d}(\mathbf{D}_{W_d} - \mathbf{W}_d) + (1-\lambda_{\xi_d})\mathbf{I}_{n_d})]^{-1}\right) \end{aligned} \qquad (2)$$

where $\mathbf{r}_d = (r_1^d, \ldots, r_{n_d}^d)'$ is the vector of relative risks within the $d$ subregion, $\mathbf{1}_{n_d}$ is a column vector of ones of length $n_d$, $\alpha_d$ 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 $\sum_{d=1}^{D} n_d = n$. Note that this model can be also written as

$$\begin{pmatrix} \log \mathbf{r}_1 \\ \vdots \\ \log \mathbf{r}_d \\ \vdots \\ \log \mathbf{r}_D \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & & & & \\ & \ddots & & & \\ & & \mathbf{1}_{n_d} & & \\ & & & \ddots & \\ & & & & \mathbf{1}_{n_D} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \\ \vdots \\ \alpha_D \end{pmatrix} + \begin{pmatrix} \mathbf{I}_{n_1} & & & & \\ & \ddots & & & \\ & & \mathbf{I}_{n_d} & & \\ & & & \ddots & \\ & & & & \mathbf{I}_{n_D} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \vdots \\ \boldsymbol{\xi}_d \\ \vdots \\ \boldsymbol{\xi}_D \end{pmatrix}$$

where the precision matrix of the multivariate Normal random effect vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_D)'$ is a block-diagonal matrix of dimension $n \times n$ with blocks corresponding to the precision matrix of the LCAR prior within each subgraph (sub-domain). Note that the design matrix of the spatially structured random effect $\boldsymbol{\xi}$ is just the identity matrix of dimension $n \times n$. Under the formulation of Model (2), $D$ independent spatial models can be simultaneously fitted giving rise to a clear computational gain.

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. However, note that $D$ specific intercepts are estimated in Model (2). If there were interest in obtaining a single estimate of an overall log-risk $\tilde{\alpha}$ which would play the role of $\alpha$ in Model (1), we propose to extract samples from the joint posterior distribution of the linear predictors $\log \mathbf{r}_d$ for $d = 1, \ldots, D$, using the `inla.posterior.sample()` function of `R-INLA`. This function makes it possible to generate $S$ samples from the approximate joint posterior of a previously fitted *inla* object, if the argument `control.compute = list(config = TRUE)` is provided when calling the `inla()` function (see for example, Gomez-Rubio, 2020 and Martino and Riebler, 2019). After joining the samples from each partition, we could define

$$\tilde{\alpha}^s = \frac{1}{n} \sum_{i=1}^{n} \log r_i, \quad \text{for } s = 1, \ldots, S$$

and then compute the kernel density estimate of $\tilde{\alpha}$ (Sheather and Jones, 1991). We note here that this procedure to estimate the posterior distribution of a global overall log-risk is valid as independence between areas corresponding to different subregions of the partition is assumed. However, if the *k-order neighbourhood* model that will be covered in the next section is considered, only its posterior mean (or median) estimate can be easily computed.

### 3.2. k-order neighbourhood model

Assuming independence between areas belonging to different subregions could be very restrictive and it may lead to border effects in the disease risk estimates. To avoid this undesirable issue, we also propose a second modelling approach where $k$-order neighbours are added to each subregion of the spatial domain. Notice that by doing this, the main spatial domain $\mathcal{D}$ is now divided into a set of overlapping 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, multiple relative risk estimates will be obtained for some areal units. As in the disjoint Model (2), $D$ submodels will be simultaneously fitted using `R-INLA`. However, the final risk surface $\mathbf{r} = (r_1, \ldots, r_n)'$ is no longer the union of the posterior estimates obtained for each submodel, since $\sum_{d=1}^{D} n_d > n$.

To obtain a unique posterior distribution of $r_i$ for each areal unit $i$, we propose to compute a mixture distribution (see, e.g., Lindsay, 1995; Frühwirth-Schnatter, 2006) using the estimated posterior probability density function of these risks. Let us assume that area $i$ lies within $m(i)$ subregions of the spatial domain $\mathcal{D}$. That is, we have $m(i)$ estimates of the $i$th area risk. If we denote $f_1(x), \dots, f_{m(i)}(x)$ to the posterior estimates of the probability density functions, the mixture distribution of $r_i$ can be written as the weighted sum of the corresponding densities

$$f(x) = \sum_{j=1}^{m(i)} w_j f_j(x),$$

where $w_j \geq 0$ and $\sum_{j=1}^{m(i)} w_j = 1$. The approximate posterior density functions $f_j(x)$ are obtained from the corresponding submodels using the `inla.dmarginal()` function, which are evaluated at 75 equally spaced points. We propose to use the *conditional predictive ordinate* (CPO), a diagnostic measure to detect discrepant observations from a given model (Pettit, 1990), to compute the weights of the mixture distribution dividing each CPO value by the sum of the $m(i)$ different estimates. Note that giving the set of observations $\mathbf{o} = (o_1, \dots, o_n)'$, $\text{CPO}_i = Pr(O_i = o_i|\mathbf{o}_{-i})$ values denotes the cross-validated predictive probability mass at the observed count $o_i$. As described in Rue et al. (2009), the CPO quantities are computed in `R-INLA` without re-running the model by including into the `inla()` function the argument `control.compute=list(cpo=TRUE)`.

### 3.3. *Model selection criteria*

In this section we discuss some Bayesian model selection criteria and show how to compute them when fitting disjoint and $k$-order neighbourhood models. Given the data $\mathbf{o}$ with likelihood function $p(\mathbf{o}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the vector of unknown parameters of the model, the *Bayesian deviance* is defined as

$$D(\boldsymbol{\theta}) = -2\log(p(\mathbf{o}|\boldsymbol{\theta})) + 2\log p(\mathbf{o})$$

where $2\log p(\mathbf{o})$ denotes the deviance of the saturated model (a constant that does not depend on the model parameters). Note that under our model formulation, that is $O_i|r_i \sim Poisson(\mu_i = E_i r_i)$, the log-likelihood function is expressed as

$$\log(p(\mathbf{o}|\boldsymbol{\theta})) = \log\left(\prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{o_i}}{o_i!}\right) = \sum_{i=1}^{n} \log\left(\frac{e^{-\mu_i}\mu_i^{o_i}}{o_i!}\right).$$

Generally, the posterior mean deviance $\overline{D(\boldsymbol{\theta})}$ is considered as a measure of goodness of fit due to its robustness. However, more complex models will fit the data better, and consequently lower values of the mean deviance will be obtained. To avoid selecting models that overfit the data, several criteria that also take into account the model complexity have been proposed in the literature. The *deviance information criterion* (DIC) (Spiegelhalter et al., 2002) and *Watanabe–Akaike information criterion* (WAIC) (Watanabe, 2010), are possibly two of the best-known criteria to compare models in a fully Bayesian setting.

The DIC is computed as the sum of the posterior mean of the deviance and the number of effective parameters (a measure of model complexity)

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D,$$

where the quantity $p_D$ is defined as the posterior mean of the deviance minus the deviance computed at the posterior mean of the parameters of interest, thus,

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + (\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})) = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}).$$

Analogously to the Akaike information criterion (AIC), models with smaller DIC values provide better trade-off between model fit and complexity. To compute the DIC values in `R-INLA` for the *Global model* described in Eq. (1), the option `control.compute = list(dic = TRUE)` inside the `inla()` function is used. As described in Rue et al. (2009), instead of evaluating the deviance at the posterior mean of all parameters, INLA evaluates the deviance at the posterior mean for the latent fields and at the posterior mode for the hyperparameters (as the posterior marginal for the hyperparameters can be severely skewed).

To compare the *Global model* with the scalable models described in Sections 3.1 and 3.2, we compute approximate DIC values for the latter models by drawing samples from the posterior marginal distributions of the relative risks using the `inla.rmarginal()` function. If a total of $S$ samples are drawn, and denoting as $\boldsymbol{\theta}^s$ to the posterior simulations of $\mu_i = E_i r_i$ for $s = 1, \ldots, S$, we can compute approximate values of the mean deviance $\overline{D(\boldsymbol{\theta})}$ and the deviance of the mean $D(\bar{\boldsymbol{\theta}})$ as

$$\overline{D(\boldsymbol{\theta})} \approx \frac{1}{S}\sum_{s=1}^{S} -2\log(p(\mathbf{o}|\boldsymbol{\theta}^s)),$$

$$D(\bar{\boldsymbol{\theta}}) \approx -2\log(p(\mathbf{o}|\bar{\boldsymbol{\theta}})), \quad \text{with } \bar{\boldsymbol{\theta}} = \frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\theta}^s.$$

Similarly, to compute the WAIC values in R-INLA, the option `control.compute = list(waic=TRUE)` must be used when fitting the *Global model*. Following Gelman et al. (2014), approximate WAIC values have been also computed for the *Disjoint model* and the *k-order neighbourhood model* as

$$\text{WAIC} \approx -2\sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{s=1}^{S} p(o_i|\boldsymbol{\theta}^s)\right) + 2\sum_{i=1}^{n}\text{Var}\left[\log(p(o_i|\boldsymbol{\theta}^s))\right].$$

## 4. Simulation study

In this section, a simulation study is conducted to compare the new scalable models, i.e., the *Disjoint model* described in Eq. (2) and the *k-order neighbourhood model* described in Section 3.2, against the common spatial LCAR model described in Eq. (1), denoted as *Global model*. We base our study on the $n = 7907$ municipalities of continental Spain. To imitate the real case study that is analysed in the next section, the $D = 15$ Autonomous Regions of Spain are used as a partition of the spatial domain (see Fig. 1).

To fit the models, improper uniform prior distributions are given to all the standard deviations (square root inverse of precision parameters), and a Uniform (0, 1) distribution is considered for the spatial smoothing parameters of the LCAR prior. Finally, a vague zero mean normal distribution with a precision close to zero (0.001) is given to the intercept ($\alpha$). All the calculations are made on a twin superserver with four processors, Inter Xeon 6C and 96 GB RAM, using the full Laplace approximation strategy in R-INLA (stable) version INLA_19.09.03 of R-3.6.2.

We consider two different scenarios to compare the performance of the models. In the first scenario, a model-free true risk surface is defined by randomly assigning high and low risk values to the areas surrounding selected major cities of Spain. Considering these cities as the area centroids, the relative risks are gradually increased/decreased at different distances to get a smooth surface. Specifically, relative risks of 1.5, 1.3 and 1.2 are assigned to the municipalities that are at less than 15 km, 30 km, and 45 km respectively from the high-risk centroids. The same criterion has been used to assign reciprocal risks of 0.67, 0.77, and 0.83 to the municipalities surrounding low-risk centroids. In the second scenario, a smooth risk surface is generated by sampling from a two-dimensional isotropic P-spline model with 40 equally spaced knots for longitude and latitude. The true risk surfaces for these scenarios are displayed in Fig. 1.

In both scenarios, counts for each municipality are generated using a Poisson distribution with mean $\mu_i = E_i r_i$, where the number of expected cases $E_i$ are fixed at values equal to 1, 5, 10, and 50. A total of 100 simulations have been generated for each of the eight sub-scenarios.

### *4.1. Results*

We evaluate the models' performance in terms of relative risk estimates by computing the mean absolute relative bias (MARB) and mean relative root mean square error (MRRMSE), defined as

$$\text{MARB} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{100}\left|\sum_{l=1}^{100}\frac{\hat{r}_i^l - r_i}{r_i}\right|,$$

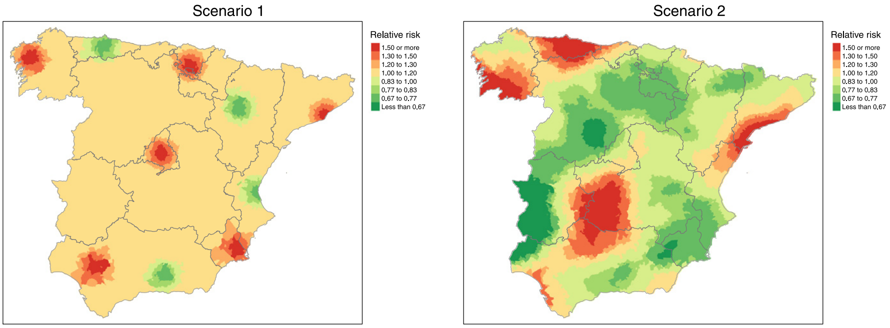

**Fig. 1.** True risk surfaces for the simulation study of Scenario 1 (left) and Scenario 2 (right).

$$\text{MRRMSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{100}\sum_{l=1}^{100}\left(\frac{\hat{r}_i^l - r_i}{r_i}\right)^2},$$

where $r_i$ is the true generated risk, and $\hat{r}_i^l$ is the posterior median estimate of the relative risk for areal unit $i$ in the $l$th simulation. In addition, coverage probabilities and 95% credible intervals' lengths have been computed.

The average values for the 100 simulated data sets in each of the sub-scenarios are computed in Table 1. The *3rd order neighbourhood model* was also considered (not shown in the table), but results did not improve those obtained with lower neighbourhood orders. Regarding computational times (in seconds), those corresponding to models simultaneously fitted in multiple machines (T1) or in a single machine (T2) are included. The maps with average values of relative risk estimates for each sub-scenario are shown in the online supplementary material.

When the number of expected cases is very low, as in sub-scenarios with $\text{E} = 1$, both model selection criteria and risk estimation accuracy measures, point to the *Global model* as the best candidate. However, small differences are observed between this model and the *kth order neighbourhood models*. As the number of expected cases increases, similar values of MARB and better values of MRRMSE are observed for our scalable model proposals in Scenario 1. The *1st order neighbourhood model* shows better or similar values in terms of model selection criteria (DIC or WAIC) for sub-scenarios $\text{E} = 5$, 10, and 50. Since in this scenario most of the high/low risk "clusters" are located inside the frontiers of the autonomous regions (see Fig. 1), the performance of the *Disjoint model* is also pretty good in terms of MRRMSE.

Scenario 2 shows a more gradual risk surface across the whole spatial domain. Then, as expected, the *Disjoint model* performs worse than the *k-order neighbourhood models*, which are able to better recover the true risk surface. In sub-scenarios $\text{E} = 1$, 5, and 10 the *second-order neighbourhood models* show slightly smaller values of DIC and WAIC than models with first order neighbourhoods. However, MARB and MRRMSE are very similar when $\text{E} = 5$, 10, and 50.

In general, we think that the new modelling proposals are a very competitive alternative to the *Global model* with a significant gain in computational time without a remarkable difference in terms of bias and variability. Empirical coverages and credible interval lengths are, in general, very similar.

## 5. Data analysis: colorectal cancer mortality in Spain

In this section, male colorectal cancer mortality data in the $n = 7907$ municipalities of continental Spain (excluding Baleares and Canary Islands and the autonomous cities of Ceuta and Melilla) are analysed using the new model proposals. According to recent studies (Ferlay et al., 2018), colorectal cancer was the second cause of cancer deaths among the male population in Europe (representing 12% of all cancers deaths) and in Spain in 2018 after lung cancer. A total of 81,934 colorectal cancer deaths (corresponding to International Classification of Diseases-10

**Table 1**
Average values of deviance information criterion (DIC), Watanabe–Akaike information criterion (WAIC), mean absolute relative bias (MARB), mean relative root mean square error (MRRMSE), empirical coverage, length of the 95% credible interval for the risks, and computational times (T1: approximate value of CPU time if all submodels are simultaneously fitted in multiple machines, T2: CPU time if all submodels are fitted in a single machine) in seconds.

| | | Model selection criteria | | Risk estimation evaluation | | | | Time | |
|---|---|---|---|---|---|---|---|---|---|
| | Model | DIC | WAIC | MARB | MRRMSE | Cov (%) | Length | T1 | T2 |
| Scenario 1 | | | | | | | | | |
| E = 1 | Global | 20800.0 | 20796.6 | 0.036 | 0.068 | 98.16 | 0.612 | 2673 | 2673 |
| | Disjoint | 20818.0 | 20801.7 | 0.043 | 0.077 | 99.24 | 0.751 | 178 | 406 |
| | 1st order neighb. | 20813.2 | 20798.2 | 0.043 | 0.073 | 99.21 | 0.743 | 292 | 546 |
| | 2nd order neighb. | 20812.2 | 20798.1 | 0.043 | 0.071 | 99.09 | 0.733 | 413 | 750 |
| E = 5 | Global | 35113.7 | 35105.4 | 0.028 | 0.058 | 98.69 | 0.423 | 1811 | 1811 |
| | Disjoint | 35135.5 | 35114.1 | 0.029 | 0.052 | 98.60 | 0.417 | 189 | 436 |
| | 1st order neighb. | 35126.4 | 35106.0 | 0.029 | 0.052 | 98.93 | 0.428 | 293 | 581 |
| | 2nd order neighb. | 35133.6 | 35114.8 | 0.029 | 0.054 | 98.82 | 0.441 | 378 | 724 |
| E = 10 | Global | 40846.5 | 40825.7 | 0.023 | 0.052 | 98.67 | 0.358 | 1799 | 1799 |
| | Disjoint | 40864.1 | 40832.0 | 0.023 | 0.044 | 98.49 | 0.328 | 182 | 417 |
| | 1st order neighb. | 40849.4 | 40817.0 | 0.023 | 0.046 | 99.00 | 0.347 | 277 | 554 |
| | 2nd order neighb. | 40861.6 | 40831.4 | 0.023 | 0.048 | 98.99 | 0.362 | 303 | 578 |
| E = 50 | Global | 54166.5 | 54050.4 | 0.014 | 0.039 | 98.29 | 0.239 | 1866 | 1866 |
| | Disjoint | 54108.6 | 54003.7 | 0.013 | 0.032 | 98.33 | 0.205 | 155 | 348 |
| | 1st order neighb. | 54083.9 | 53970.6 | 0.013 | 0.034 | 98.81 | 0.219 | 181 | 371 |
| | 2nd order neighb. | 54109.6 | 53997.3 | 0.013 | 0.035 | 98.80 | 0.228 | 244 | 458 |
| Scenario 2 | | | | | | | | | |
| E = 1 | Global | 19815.1 | 19810.3 | 0.048 | 0.109 | 99.80 | 0.811 | 1609 | 1609 |
| | Disjoint | 19894.2 | 19874.8 | 0.070 | 0.127 | 99.51 | 0.904 | 151 | 340 |
| | 1st order neighb. | 19875.1 | 19856.4 | 0.062 | 0.120 | 99.78 | 0.907 | 215 | 410 |
| | 2nd order neighb. | 19868.3 | 19850.4 | 0.058 | 0.117 | 99.89 | 0.910 | 284 | 515 |
| E = 5 | Global | 34236.2 | 34193.6 | 0.028 | 0.077 | 99.79 | 0.535 | 1922 | 1922 |
| | Disjoint | 34279.1 | 34231.5 | 0.035 | 0.080 | 99.70 | 0.527 | 146 | 327 |
| | 1st order neighb. | 34253.1 | 34201.7 | 0.031 | 0.077 | 99.85 | 0.536 | 187 | 379 |
| | 2nd order neighb. | 34250.7 | 34197.9 | 0.030 | 0.077 | 99.87 | 0.541 | 254 | 476 |
| E = 10 | Global | 40028.0 | 39942.7 | 0.022 | 0.067 | 99.77 | 0.439 | 1915 | 1915 |
| | Disjoint | 40055.3 | 39973.9 | 0.028 | 0.067 | 99.64 | 0.421 | 136 | 303 |
| | 1st order neighb. | 40025.9 | 39935.9 | 0.024 | 0.065 | 99.83 | 0.431 | 166 | 334 |
| | 2nd order neighb. | 40027.8 | 39934.5 | 0.024 | 0.065 | 99.85 | 0.436 | 231 | 425 |
| E = 50 | Global | 53403.9 | 53086.7 | 0.013 | 0.047 | 99.55 | 0.269 | 1885 | 1885 |
| | Disjoint | 53376.0 | 53105.5 | 0.015 | 0.044 | 99.53 | 0.253 | 113 | 247 |
| | 1st order neighb. | 53352.5 | 53054.5 | 0.013 | 0.044 | 99.64 | 0.260 | 152 | 302 |
| | 2nd order neighb. | 53366.9 | 53057.9 | 0.013 | 0.045 | 99.66 | 0.260 | 219 | 396 |

codes C18–C21) were registered for male population in the municipalities of continental Spain during the 2006–2015 period, which represents an overall crude rate of 38.54 deaths per 100,000 male inhabitants. The indirect age-standardization method has been used to compute the number of expected cases using 5-years age groups (internal standardization). This method allows us to compare the relative risk of each municipality with the whole of Spain during the study period. The expected number of cases ranges from 0 to 6129 (with mean and median values of 1.8 and 10.4, respectively), while the number of observed cases varies from 0 to 5814 (with mean and median values of 2.0 and 10.4, respectively).

As in the simulation study, the *Global model*, the *Disjoint model*, and $k = 1, 2, 3$ *order neighbourhood models* have been fitted with `R-INLA` using the $D = 15$ Autonomous Regions of Spain as a partition of the spatial domain. The same hyperprior distributions described in Section 4 have been also considered here. Results are shown in Table 2. 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*,

**Table 2**
Model selection criteria ($\overline{D(\boldsymbol{\theta})}$: mean deviance, $p_D$: effective number of parameters, DIC: deviance information criterion, WAIC: Watanaba–Akaike information criterion), computational time (T.run: running time, T.merge: merging time, T.tot: Total time) in seconds and data size ($n = \sum_{d=1}^{D} n_d$).

| Model | $\overline{D(\boldsymbol{\theta})}$ | $p_D$ | DIC | WAIC | T.run | T.merge | T.total | $n$ |
|---|---|---|---|---|---|---|---|---|
| Global | 26667.6 | 548.5 | 27216.1 | 27237.9 | 1929 | – | 1929 | 7907 |
| Disjoint | 26510.7 | 656.8 | 27167.5 | 27166.7 | 110 | 26 | 136 | 7907 |
| 1st order neighbourhood | 26533.5 | 634.2 | 27167.6 | 27170.5 | 132 | 63 | 195 | 8979 |
| 2nd order neighbourhood | 26557.9 | 616.5 | 27174.3 | 27183.3 | 166 | 83 | 249 | 10646 |
| 3rd order neighbourhood | 26586.0 | 583.0 | 27169.0 | 27175.4 | 219 | 107 | 326 | 12553 |

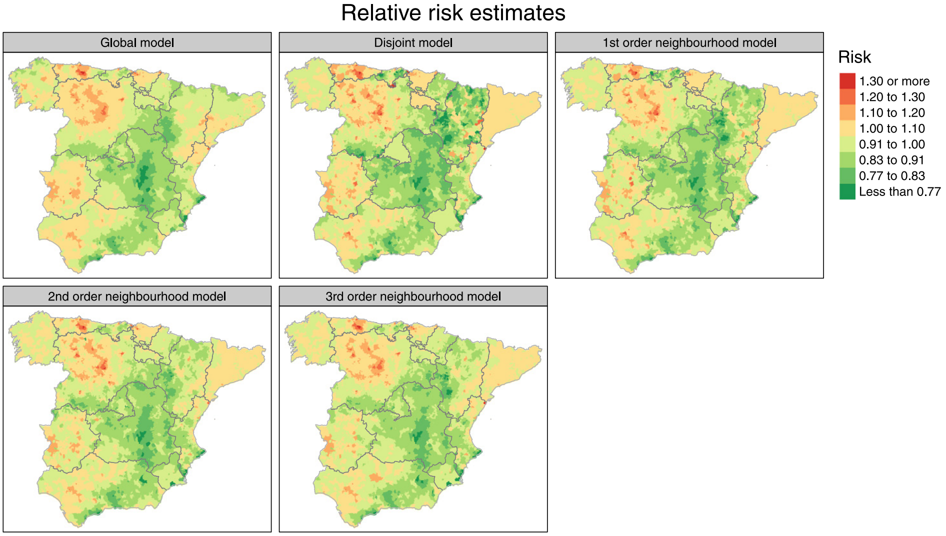

**Fig. 2.** Maps of posterior median estimates for $r_i$ of male colorectal cancer mortality data in Spanish municipalities during the period 2006–2015.

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 increase as higher values of neighbourhood order are considered. The largest values of $n_d$ (number of areas for each subdivision) correspond to the autonomous region of Castilla and León, located in north-west Spain, with a total of 2245, 2451, 2744 and 3047 municipalities for neighbourhood models with $k = 0$ (Disjoint model), 1, 2, and 3 respectively.

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 $r_i$, and posterior exceedance probabilities $P(r_i > 1|\mathbf{O})$ of male colorectal cancer mortality risks are shown in Figs. 2 and 3 respectively. In general, very similar spatial patterns are observed for all the models, but *2nd* and *3rd order neighbourhood models* seem to show a spatial risk surface more similar to the *Global model*. Even though small differences are observed in DIC and WAIC values between the scalable model proposals, a greater variability in the degree of spatial smoothness among autonomous regions is observed for the *Disjoint model*, leading to not very reasonable relative risk estimates in some regions as Madrid or Aragón. As expected, this effect seems to be corrected when including neighbouring areas in the spatial sub-domains in the *k-order neighbourhood models*.

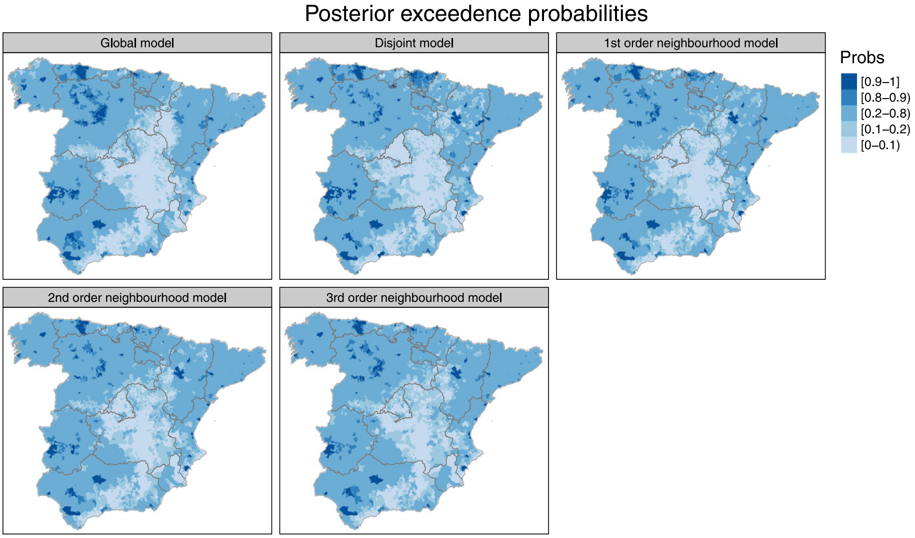
Posterior exceedence probabilities
Global model
Disjoint model
1st order neighbourhood model
2nd order neighbourhood model
3rd order neighbourhood model
Probs
[0.9–1]
[0.8–0.9)
[0.2–0.8)
[0.1–0.2)
[0–0.1)

**Fig. 3.** Maps of posterior exceedance probabilities $P(r_i > 1|\mathbf{O})$ of male colorectal cancer mortality data in Spanish municipalities during the 2006–2015 period.

## 6. Discussion

The "divide and conquer" strategy has been extensively used to analyse big data in other contexts such as machine learning, commonly using a Bayesian approach to compute tractable posterior distributions (posterior samples if MCMC methods are considered). One of the key questions is how to combine the estimation of the parameters of interest from each subsample to obtain robust final estimates. Many combination methods have been proposed in the literature, such as the kernel density product estimation proposed by Neiswanger et al. (2013), the consensus Monte Carlo algorithm (Scott et al., 2016) where a weighted average of the posterior distributions obtained from the subsample data are defined, the mixture-based approach proposed by Scott et al. (2017), or the recently proposed global consensus Monte Carlo algorithm (Rendell et al., 2020), among others.

In this paper we avoid MCMC methods and rely on the INLA technique, as it is relatively simple for practitioners to use. In particular, we develop scalable Bayesian models for smoothing mortality or incidence risks in spatial disease mapping when the number of small areas is very large. We propose to divide the main spatial domain into subregions so that local spatial models can be simultaneously fitted reducing the computational time substantially. Although the methodology described in this paper uses the INLA estimation strategy, it could also be adapted to other Bayesian fitting techniques.

As stated, the new proposals must define a partition of the spatial domain as a first step. The administrative divisions of the area of interest (such as provinces, states or local health areas) are a natural choice for this partition. However, if the user has no idea on how to define this initial partition, a random partition can be also considered by defining a grid over the associated cartography with a certain number of rows and columns (see the vignette accompanying the `bigDM` package for further details). We note here that in the real data analysis performed in Section 5 very similar results are obtained when using a random partition over the municipalities of Spain instead of the spatial partition defined by the Autonomous Regions, in particular when fitting the *k-order neighbourhood models*. The second stage of our proposal is to fit independent hierarchical Bayesian models including spatially structured and unstructured random effects to smooth the risks in each subregion. Here, two different modelling approaches are defined: a *Disjoint model* where each geographical unit is contained into a single subregion, and a *k-order neighbourhood model* where an overlapping set of regions are defined by adding neighbouring areas to those regions located in the border of the partition. This second approach allows us to eliminate the independence assumption between areas belonging to different subregions, avoiding border effects. Finally, the results of the models are merged to obtain a unique risk estimate for each areal unit. For the *k-order neighbourhood model*, we compute a mixture distribution of the estimated posterior probability density functions using the CPO's to calculate the mixture weights. In addition, approximations to model selection criteria such as DIC and WAIC are also derived.

Both the simulation study and the real data analysis indicate that the new methodology provides reliable risk estimates with a substantial reduction in computational time. Moreover, the new scalable models avoid the high RAM/CPU memory usage when analysing massive spatial data. In cases where small differences in model selection criteria are observed between the *Disjoint* and *k-order neighbourhood model*, we recommend using the *k-order neighbourhood model* to avoid overfitting and border effects.

We would like to highlight that the main objective of this paper is to propose a new Bayesian computational strategy to estimate the posterior distribution of relative risks when dealing with high-dimensional data as this is the main purpose of disease mapping models (Ugarte et al., 2006). The scope of this article does not include estimating associations between the response variable and certain covariates (via ecological spatial regression). If this were the objective, two important problems would have to be tackled: one, the potential spatial confounding between fixed and random effects and the other, the need (or not) to estimate a single regression coefficient for each variable in all partitions.

Finally, we think that the great potential of this methodology is its extension to the spatio-temporal setting. The complexity inherent to spatio-temporal interaction models and the even higher dimensionality associated to this type of data, makes it necessary the use of scalable techniques for Bayesian inference in small area data. We are currently investigating this issue.

## Supplementary material

The maps with average values of relative risk estimates for each sub-scenario of the simulation study presented in Section 4 are available at https://emi-sstcdapp.unavarra.es/bigDM/SupplementaryMaterial.pdf.

## Acknowledgments

This research has been supported by the Spanish Ministry of Science and Innovation (project MTM 2017-82553-R (AEI/FEDER, UE)). It has also been partially funded by la Caixa Foundation, Spain (ID 1000010434), Caja Navarra Foundation, Spain, and UNED Pamplona, Spain, under agreement LCF/PR/PR15/51100007 (project REF P/13/20). We would like to thank the Spanish Statistical Office (INE) for providing the data.

## References

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., Lindgren, F., 2018. Spatial modeling with R-INLA: A review. Wiley Interdiscip. Rev. Comput. Stat. 10 (6), e1443.

Banerjee, S., 2017. High-dimensional Bayesian geostatistics. Bayesian Anal. 12 (2), 583.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848. http://dx.doi.org/10.1111/j.1467-9868.2008.00663.x.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2), 192–225. http://dx.doi.org/10.1111/j.2517-6161.1974.tb00999.x.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43 (1), 1–20. http://dx.doi.org/10.1007/bf00116466.

Bradley, J.R., Holan, S.H., Wikle, C.K., et al., 2018. Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). Bayesian Anal. 13 (1), 253–310.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Softw. 76 (1).

Cressie, N., 1993. Statistics for Spatial Data. Revised edition. John Wiley & Sons.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 209–226. http://dx.doi.org/10.1111/j.1467-9868.2007.00633.x.

Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111 (514), 800–812. http://dx.doi.org/10.1080/01621459.2015.1044091.

Datta, A., Banerjee, S., Hodges, J.S., Gao, L., 2019. Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. Bayesian Anal. 14 (4), 1221–1244. http://dx.doi.org/10.1214/19-BA1177.

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodríguez, A., Temple Lang, D., Paganin, S., 2020. NIMBLE User manual. R package manual version 0.9.1. http://dx.doi.org/10.5281/zenodo.1211190, URL https://r-nimble.org.

Dean, C., Ugarte, M., Militino, A., 2001. Detecting interaction between random region and fixed age effects in disease mapping. Biometrics 57 (1), 197–202. http://dx.doi.org/10.1111/j.0006-341x.2001.00197.x.

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O., Bray, F., 2018. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Eur. J. Cancer 103, 356–387. http://dx.doi.org/10.1016/j.ejca.2018.07.005.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Science & Business Media.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24 (6), 997–1016. http://dx.doi.org/10.1007/s11222-013-9416-2.

Goicoa, T., Adin, A., Ugarte, M., Hodges, J., 2018. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. Stoch. Environ. Res. Risk Assess. 32 (3), 749–770. http://dx.doi.org/10.1007/s00477-017-1405-0.

Gomez-Rubio, V., 2020. Bayesian Inference with INLA. CRC Press.

Gramacy, R.B., Apley, D.W., 2015. Local Gaussian process approximation for large computer experiments. J. Comput. Graph. Statist. 24 (2), 561–578. http://dx.doi.org/10.1080/10618600.2014.914442.

Guhaniyogi, R., Banerjee, S., 2018. Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. Technometrics 60 (4), 430–444. http://dx.doi.org/10.1080/00401706.2018.1437474.

Hodges, J.S., Reich, B.J., 2010. Adding spatially-correlated errors can mess up the fixed effect you love. Amer. Statist. 64 (4), 325–334. http://dx.doi.org/10.1198/tast.2010.10052.

Hughes, J., Cui, X., 2020. Ngspatial: Fitting the centered autologistic and sparse spatial generalized linear mixed models for areal data. R package version 1.2-2.

Hughes, J., Haran, M., 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (1), 139–159. http://dx.doi.org/10.1111/j.1467-9868.2012.01041.x.

Katzfuss, M., 2017. A multi-resolution approximation for massive spatial datasets. J. Amer. Statist. Assoc. 112 (517), 201–214. http://dx.doi.org/10.1080/01621459.2015.1123632.

Katzfuss, M., Guinness, J., 2021. A general framework for vecchia approximations of Gaussian processes. Statist. Sci. 36 (1), 124–141. http://dx.doi.org/10.1214/19-STS755.

Katzfuss, M., Hammerling, D., 2017. Parallel inference for massive distributed spatial data using low-rank models. Stat. Comput. 27 (2), 363–375. http://dx.doi.org/10.1007/s11222-016-9627-4.

Lawson, A.B., Banerjee, S., Haining, R.P., Ugarte, M.D., 2016. Handbook of Spatial Epidemiology. CRC Press.

Lenzi, A., Castruccio, S., Rue, H., Genton, M.G., 2020. Improving Bayesian local spatial models in large data sets. J. Comput. Graph. Statist. 1–28. http://dx.doi.org/10.1080/10618600.2020.1814789.

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

Lindgren, F., Rue, H., Lindstr̀om, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498. http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x.

Lindsay, B.G., 1995. Mixture models: Theory, geometry and applications. In: NSF-CBMS Regional Conference Series in Probability and Statistics. JSTOR, pp. i–163.

MacNab, Y.C., 2011. On Gaussian Markov random fields and Bayesian disease mapping. Stat. Methods Med. Res. 20 (1), 49–68. http://dx.doi.org/10.1177/0962280210371561.

Martínez-Beneito, M.A., Botella-Rocamora, P., 2019. Disease Mapping: From Foundations to Multidimensional Modeling. CRC Press.

Martino, S., Riebler, A., 2019. Integrated nested Laplace approximations (INLA). arXiv:1907.01248.

Neiswanger, W., Wang, C., Xing, E., 2013. Asymptotically exact, embarrassingly parallel MCMC. arXiv preprint arXiv:1311.4780.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2015. A multiresolution Gaussian process model for the analysis of large spatial datasets. J. Comput. Graph. Statist. 24 (2), 579–599. http://dx.doi.org/10.1080/10618600.2014.914946.

Paciorek, C.J., Lipshitz, B., Zhuo, W., Kaufman, C.G., Thomas, R.C., et al., 2015. Parallelizing Gaussian process calculations in R. J. Stat. Softw. 63 (10), 1–23. http://dx.doi.org/10.18637/jss.v063.i10.

Pettit, L., 1990. The conditional predictive ordinate for the normal distribution. J. R. Stat. Soc. Ser. B Stat. Methodol. 52 (1), 175–184. http://dx.doi.org/10.1111/j.2517-6161.1990.tb01780.x.

Plummer, M., et al., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vol. 124. Vienna, Austria., pp. 1–10.

Reich, B.J., Hodges, J.S., Zadnik, V., 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics 62 (4), 1197–1206. http://dx.doi.org/10.1111/j.1541-0420.2006.00617.x.

Rendell, L.J., Johansen, A.M., Lee, A., Whiteley, N., 2020. Global consensus Monte Carlo. J. Comput. Graph. Statist. 1–11. http://dx.doi.org/10.1080/10618600.2020.1811105.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. CRC Press, Boca Raton.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392. http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x.

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K., 2017. Bayesian Computing with INLA: A review. Annu. Rev. Stat. Appl. 4, 395–421. http://dx.doi.org/10.1146/annurev-statistics-060116-054045.

Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., McCulloch, R.E., 2016. Bayes And big data: The consensus Monte Carlo algorithm. Int. J. Manag. Sci. Eng. Manag. 11 (2), 78–88.

Scott, S.L., et al., 2017. Comparing consensus Monte Carlo strategies for distributed Bayesian computation. Braz. J. Probab. Stat. 31 (4), 668–685.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B Stat. Methodol. 53 (3), 683–690. http://dx.doi.org/10.1111/j.2517-6161.1991.tb01857.x.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (4), 583–639. http://dx.doi.org/10.1111/1467-9868.00353.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2003. WinBUGS User Manual. Citeseer.

Sun, Y., Li, B., Genton, M.G., 2012. Geostatistics for large datasets. In: Advances and Challenges in Space-Time Modelling of Natural Events. Springer, pp. 55–77.

Team, S.D., 2018. Stan modeling language users guide and reference manual, version 2.18. 0. URL https://mc-stan.org/.

Ugarte, M.D., Ibáñez, B., Militino, A.F., 2006. Modelling risks in disease mapping. Stat. Methods Med. Res. 15 (1), 21–35.

van Niekerk, J., Bakka, H., Rue, H., Schenk, L., 2019. New frontiers in Bayesian modeling using the INLA package in R. J. Stat. Softw. in press. arXiv:1907.10426.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11 (Dec), 3571–3594.