Statistics and Computing (2023) 33:104
https://doi.org/10.1007/s11222-023-10263-x

ORIGINAL PAPER

High-dimensional order-free multivariate spatial disease mapping

Gonzalo Vicente[1] · Aritz Adin[1] · Tomás Goicoa[1] · María Dolores Ugarte[1]

Received: 1 November 2022 / Accepted: 4 June 2023 / Published online: 19 July 2023

Abstract
Despite the amount of research on disease mapping in recent years, the use of multivariate models for areal spatial data remains limited due to difficulties in implementation and computational burden. These problems are exacerbated when the number of areas is very large. In this paper, we introduce an order-free multivariate scalable Bayesian modelling approach to smooth mortality (or incidence) risks of several diseases simultaneously. The proposal partitions the spatial domain into smaller subregions, fits multivariate models in each subdivision and obtains the posterior distribution of the relative risks across the entire spatial domain. The approach also provides posterior correlations among the spatial patterns of the diseases in each partition that are combined through a consensus Monte Carlo algorithm to obtain correlations for the whole study region. We implement the proposal using integrated nested Laplace approximations (INLA) in the R package `bigDM` and use it to jointly analyse colorectal, lung, and stomach cancer mortality data in Spanish municipalities. The new proposal allows for the analysis of large datasets and yields superior results compared to fitting a single multivariate model. Additionally, it facilitates statistical inference through local homogeneous models, which may be more appropriate than a global homogeneous model when dealing with a large number of areas.

Keywords Bayesian inference · High-dimensional data · Scalable models · Spatial epidemiology

1 Introduction

Research on methodology for the spatial (and spatio-temporal) analysis of areal count data has grown tremendously in the last years, and statistical models have proven an essential tool for studying the geographic distribution of data in small areas. The main objective of these techniques is to smooth standardized mortality (incidence) ratios or crude rates to discover geographic patterns of the phenomenon under study. These models and methods have been mainly applied in epidemiology to analyse the incidence and mortality of chronic diseases such as cancer, but some recent research has demonstrated their applicability to the spatial and spatio-temporal analysis of crimes (see for example Li et al. 2014), and in particular crimes against women (see for example Vicente et al. 2018, 2020a). Although research on single disease analysis has been very fruitful and abundant since the seminal work of Besag et al. (1991), joint modelling of multiple responses provides several advantages. Firstly, it improves smoothing by borrowing strength between diseases. Secondly, and perhaps more importantly, it allows to establish relationships between different diseases, such as similar or completely different geographical distributions, i.e., correlations between spatial patterns. This is crucial, as these correlations may indicate associations with common underlying risk factors and certain (usually unknown) connections between the different diseases. The joint analysis employs multivariate spatial models that can handle both the spatial correlation within diseases and the correlation between diseases.

There is a considerable amount of research on Bayesian multivariate spatial models for count data, most of the proposals relying on Markov chain Monte Carlo (MCMC) algorithms for estimation and inference. However, their use

✉ María Dolores Ugarte
lola@unavarra.es

Gonzalo Vicente
gonzalo.vicente@unavarra.es

Aritz Adin
aritz.adin@unavarra.es

Tomás Goicoa
tomas.goicoa@unavarra.es

[1] Department of Statistics, Computer Science, and Mathematics, Institute for Advanced Materials and Mathematics (InaMat^2), Public University of Navarre, Pamplona, Spain

in practice is still limited due to a lack of " easy-to-use" implementations of the models in statistical packages and the computational burden of most of the proposals that preclude practitioners from exploiting their advantages over univariate counterparts. According to MacNab (2010), there are two approaches to multivariate modelling in disease mapping. The first one uses shared-component models (see Knorr-Held and Best 2001; Held et al. 2005), where pair-wise dependence between diseases is not a testable hypothesis, but it is assumed. In the second one, pair-wise correlation between diseases is a testable assumption and the interest is in estimating such correlations. Hereafter in the paper, multivariate models refer to this second approach. A comprehensive review of the subject can be found in the work of MacNab (2018) which discusses the three main lines in the construction of multivariate proposals based on Gaussian Markov random fields. Namely, a multivariate conditionals-based approach (Mardia 1988), a univariate conditionals-based approach (Sain et al. 2011), and a coregionalization framework (Jin et al. 2007). Regarding the latter, Martinez-Beneito (2013) derives a general theoretical setting for multivariate areal models that covers many of the existing proposals in the literature. However, this procedure is unaffordable for a moderate to large number of diseases due to the high computational cost of the MCMC algorithms. Botella-Rocamora et al. (2015) reformulate the Martínez-Beneito framework and present the so called M-models as a simpler and more computationally efficient alternative. This approach makes it possible to increase the number of diseases in the model at the expense of the identifiability of certain parameters. Recently, Vicente et al. (2020b) consider the M-models-based approach to analyse in space and time different crimes against women in India. These authors estimate the M-models using integrated nested Laplace approximations (INLA) and numerical integration for Bayesian inference (see Rue et al. 2009) and implement the procedure using the 'rgeneric' construction in R-INLA (Lindgren and Rue 2015). The result is a " ready-to-use" function for a wide audience with limited programming skills.

Several alternatives to Gaussian Markov random fields have been also proposed in the disease mapping literature. A very attractive modelling approach is the use of splines to smooth risks (Goicoa et al. 2012). Research on multivariate spline models for fitting spatio-temporal count data is not so abundant and focuses on multivariate structures to deal with the spatial and temporal dependence for one response measured in several time periods (see for example MacNab 2016; Ugarte et al. 2010, 2017). Very recently, Vicente et al. (2021) propose multivariate P-spline models to study the spatio-temporal evolution of four crimes against women. Unfortunately, inference for these multivariate proposals (and also for univariate approaches) become unfeasible when the number of areas is very large, and the scalability of the procedures is an issue.

New directions in disease mapping points towards developing new methods for Bayesian inference when the number of small areas is very large (MacNab 2022). Creating computationally efficient methods for large data sets is one of the greatest challenges in the field of univariate and multivariate spatial statistics. Several methods for massive geostatistical data (point-referenced) have been already proposed (see for example Cressie and Johannesson 2008; Lindgren et al. 2011; Nychka et al. 2015; Katzfuss 2017; Katzfuss and Guinness 2021, among others). However, in the case of areal (lattice) count data, research on the scalability of statistical models is not so abundant. Recently, Orozco-Acosta et al. (2021, 2023) propose a scalable Bayesian modelling approach for univariate high-dimensional spatial and spatio-temporal disease mapping data. They propose to divide the spatial domain into D subregions where independent models can be fitted simultaneously. To avoid the border effect in the risk estimates, k-order neighbours are added to each subregion so that some areal units will have several risk estimates. Finally, a unique posterior distribution for these risks is obtained by either computing the mixture distribution of the estimated posterior probability density functions or by selecting the posterior marginal risk estimate corresponding to the original domain to which the area belongs. This proposal reduces computational time and, in contrast to fitting a single model to the whole domain, it allows different degree of spatial smoothness over the areas within the different subdomains.

The main objective of this paper is to present a new approach to fit order-free multivariate spatial disease mapping models in domains with a very large number of small areas avoiding high RAM/CPU usage, and making it accessible to users with limited computing facilities. In particular, we combine the Orozco-Acosta et al. (2021, 2023) " divide-and-conquer" approach with a modification of the Botella-Rocamora et al. (2015) M-models to avoid overparametrization. Our approach allows for statistical inference in the subdivisions of the study domain using local homogeneous models, which seems more appropriate than a single global model when the number of small areas is large. Then, we are able to retrieve the posterior distributions of the correlations between the spatial patterns of each disease in the whole spatial domain, as well as in each of the subdivisions. We have implemented the methodology in INLA to reduce computational burden through our R package `bigDM` (Adin et al. 2023), that also implements recent high-dimensional univariate proposals.

The rest of the article has the following structure. Section 2 reviews the M-models to fit multivariate data. In Sect. 3 we present the new methodology to make the multivariate models scalable. In Sect. 4, we conduct a simulation study to compare the performance of this new modelling approach

with a single multivariate spatial M-model fitted to the whole domain. Finally, in Sect. 5, we use the new proposal to jointly analyse lung, colorectal and stomach cancer male mortality in Spanish municipalities. The paper closes with a discussion.

2 M-models for multivariate disease mapping

Let us assume that the area of interest is divided into I contiguous small areas and data are available for J diseases. Let O_{ij} and E_{ij} denote the number of observed and expected cases respectively in the i-th small area ($i = 1, \ldots, I$) and for the j-th disease ($j = 1, \ldots, J$). Conditional on the relative risks R_{ij}, the number of observed cases in the i-th area and the j-th disease is assumed to follow a Poisson distribution with mean $\mu_{ij} = E_{ij} \cdot R_{ij}$, that is,

$$O_{ij}|R_{ij} \sim Poisson(\mu_{ij} = E_{ij} \cdot R_{ij}),$$
$$\log \mu_{ij} = \log E_{ij} + \log R_{ij}.$$

Here E_{ij} is computed using indirect standardization as $E_{ij} = \sum_k n_{ijk} \cdot m_{jk}$, where k is the age-group, n_{ijk} is the population at risk in area i and age-group k for the j-th disease, and m_{jk} is the overall mortality (or incidence) rate of the j-th disease in the total area of study for the k-th age group. The log-risk is modelled as

$$\log R_{ij} = \alpha_j + \theta_{ij}, \tag{1}$$

where α_j is a disease-specific intercept and θ_{ij} is the spatial effect of the i-th area for the j-th disease. Following the work by Botella-Rocamora et al. (2015), we rearrange the spatial effects into the matrix $\boldsymbol{\Theta} = \{\theta_{ij} : i = 1, \ldots, I; j = 1, \ldots, J\}$ to better comprehend the dependence structure. The main advantage of the multivariate modelling is that dependence between the spatial patterns of the different diseases can be included in the model, so that latent associations between diseases can help to discover potential risk factors related to the phenomena under study. These unknown connections can be crucial to a better understanding of complex diseases such as cancer.

The potential association between the spatial patterns of the different diseases are included in the model considering the decomposition of $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta} = \Phi\mathbf{M}, \tag{2}$$

where Φ and $\mathbf{M}$ deal with within and between-disease dependencies, respectively. We refer to Eq. (2) as the M-model. In the following, we briefly describe the two components of the M-model.

The matrix Φ is of order $I \times K$ and it is composed of stochastically independent columns that follow a spatially correlated distribution. Usually, $K = J$, although J and K may be different (see Corpas-Burgos et al. 2019, for a discussion). To deal with spatial dependence, different spatial priors have been considered in the literature, most of them based on the well known intrinsic conditional autoregressive (iCAR) prior (Besag 1974). Namely, the proper CAR (pCAR), a proper version of the iCAR; the Besag et al. (1991) prior (BYM), which combines iCAR and exchangeable random effects; the Leroux et al. (1999) prior (LCAR) that models spatially structured and spatially unstructured variability through a weighted sum of the iCAR precision matrix and the identity, or a modified version of the BYM model denoted as BYM2 (Dean et al. 2001; Riebler et al. 2016). In summary, the columns of $\boldsymbol{\Phi}$ follow multivariate Normal distributions with mean $\mathbf{0}$ and precision matrix $\boldsymbol{\Omega}$ whose expression depends on the spatial prior. In this paper, we consider the iCAR prior for the columns of $\boldsymbol{\Phi}$, and hence the precision matrix is $\boldsymbol{\Omega}_{\text{iCAR}} = \tau\mathbf{Q}$, where $\mathbf{Q}$ is the usual spatial neighbourhood matrix defined as $Q_{il} = 1$ if the i-th and the l-th areas are neighbours (share a common border) and 0 otherwise, $Q_{ii} = n_i$, with n_i is the number of neighbours of the i-th area, and τ is the precision parameter. We choose the iCAR prior because in the real case study all the variability is spatially structured.

On the other hand, $\mathbf{M}$ is a $K \times J$ nonsingular but arbitrary matrix and it is responsible for inducing dependence between the different columns of $\boldsymbol{\Theta}$, i.e, for inducing correlation between the spatial patterns of the diseases. In Eq. (2), the cells of $\mathbf{M}$ act as regression coefficients of the log-relative risks on the underlying patterns captured in $\boldsymbol{\Phi}$ and are treated as fixed effects with a Normal prior distribution with mean 0 and a large (and fixed) variance σ^2. Note that assigning this type of priors to the cells of $\mathbf{M}$ is equivalent to considering a Wishart prior to $\mathbf{M}'\mathbf{M}$, i.e., $\mathbf{M}'\mathbf{M} \sim Wishart(J, \sigma^2\mathbf{I}_J)$.

The multivariate approach allows the estimation of the correlation between the spatial patterns of the diseases, an interesting and useful feature, as a high positive correlation would support the hypotheses of common risk factors, and hence connections between diseases. The covariance matrix between the spatial patterns is obtained as $\mathbf{M}'\mathbf{M}$. For further details see Botella-Rocamora et al. (2015).

For notation purposes and to incorporate the dependencies between different diseases in the model, we introduce the vec$(\cdot)$ operator. Let $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_J)$ be an $I \times J$ matrix with $I \times 1$ columns $\mathbf{A}_j$, for $j = 1, \ldots, J$. The vec$(\cdot)$ operator transforms $\mathbf{A}$ into an $IJ \times 1$ vector by stacking the columns one under the other, that is, vec$(\mathbf{A}) = (\mathbf{A}_1', \ldots, \mathbf{A}_J')'$. Using this notation, the multivariate Model (1) can be expressed in matrix form as

$$\log \mathbf{R} = (\mathbf{I}_J \otimes \mathbf{1}_I)\,\boldsymbol{\alpha} + \text{vec}\,(\boldsymbol{\Theta})\,, \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$, $\mathbf{R} = (\mathbf{R}_1', \ldots, \mathbf{R}_J)'$, $\mathbf{R}_j = (R_{1j}, \ldots, R_{Ij})'$, $j = 1, \ldots, J$, and $\mathbf{I}_J$ and $\mathbf{1}_I$ are the $J \times J$ identity matrix and a column vector of ones of length I respectively. Once the between-diseases dependencies are incorporated into the model, the resulting prior distributions for vec ($\boldsymbol{\Theta}$) with Gaussian kernel has a precision matrix given by

$$\boldsymbol{\Omega}_{\text{vec}(\Theta)} = \left(\mathbf{M}^{-1} \otimes \mathbf{I}_I\right) \text{Blockdiag}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_J) \left(\mathbf{M}^{-1} \otimes \mathbf{I}_I\right)'. \tag{4}$$

Recall that this precision matrix accounts for both within and between-disease dependencies: the $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_J$ matrices control the within-diseases spatial structure and the matrix $\mathbf{M}$ deals with the between-diseases variability. Note that if $\boldsymbol{\Omega}_1 = \ldots = \boldsymbol{\Omega}_J = \boldsymbol{\Omega}_w$, the covariance structure is separable and can be expressed as $\boldsymbol{\Omega}^{-1}_{\text{vec}(\boldsymbol{\Theta})} = \boldsymbol{\Omega}_b^{-1} \otimes \boldsymbol{\Omega}_w^{-1}$, where $\boldsymbol{\Omega}_b^{-1} = \mathbf{M}'\mathbf{M}$ and $\boldsymbol{\Omega}_w^{-1}$ are the between- and within-disease covariance matrices, respectively. Note that in our case $\boldsymbol{\Omega}_w^{-1} = \boldsymbol{\Omega}^{-1}_{\text{iCAR}}$ and the precision parameter τ is set to 1 for identifiability issues. This M-model based framework includes both separable and non-separable covariance structures, and can accommodate different spatial dependency structures with different within-disease covariance matrices.

2.1 Model fitting, identifiability issues and prior distributions

Traditionally, MCMC techniques have been used for Bayesian model fitting and inference. However, they can be computationally very demanding. On the other hand, the INLA method (see Rue et al. 2009) has turned out to be very popular in recent years. It is designed for latent Gaussian fields and is based on integrated nested Laplace approximations and numerical integration. Many models used in practice are implemented in `R-INLA` (Lindgren and Rue 2015), and others can be implemented by means of generic functions with some extra-programming work. The M-model based approach is not directly available in R-INLA, but it can be implemented using the 'rgeneric' construct (see for example Vicente et al. 2020b). In this paper, we use INLA for model fitting and inference.

Spatial models usually present identifiability issues which are generally overcome using sum-to-zero constraints on the spatial random effects (see Eberly and Carlin 2000; Goicoa et al. 2018, for details). In the multivariate setting, these constraints are considered for all the diseases in the model. Additionally, the M-models bring about new identifiability issues. As pointed out by Botella-Rocamora et al. (2015), any orthogonal transformation of the columns of $\boldsymbol{\Phi}$ and of the rows of $\mathbf{M}$ in Eq. (2) causes an alternative decomposition of $\boldsymbol{\Theta}$, and therefore neither $\boldsymbol{\Phi}$ nor $\mathbf{M}$ are identifiable and inference on them should be ruled out. However, $\boldsymbol{\Theta}$ and the covariance matrix $\mathbf{M}'\mathbf{M}$ are perfectly identifiable, so inference is confined to those quantities. It is worth noting that the decomposition of the between-diseases covariance matrix as $\boldsymbol{\Omega}_b^{-1} = \mathbf{M}'\mathbf{M}$ avoids dependence on the order in which the diseases are introduced into the model, but it leads to an over-parameterization problem. In the M-model proposal, $J \times J$ parameters are used to estimate the covariance matrix even though only $J \times (J+1)/2$ parameters are required. In their paper, Botella-Rocamora et al. (2015) put independent Normal priors with mean 0 and large and fixed variance σ^2 on each entry of the matrix $\mathbf{M}$ and they show that this is equivalent to assigning a Wishart prior to the covariance matrix, i.e., $\mathbf{M}'\mathbf{M} \sim Wishart(J, \sigma^2\mathbf{I}_J)$.

To avoid the overparameterization of the covariance matrix we propose to use the Barlett decomposition of Wishart matrices (see, for example, Peña and Irie 2022). In more detail, if $\boldsymbol{\Omega}_b^{-1}$ is the $J \times J$ between-disease covariance matrix with $\boldsymbol{\Omega}_b^{-1} \sim \text{Wishart}(\upsilon, \mathbf{V})$, then the Bartlett decomposition of $\boldsymbol{\Omega}_b^{-1}$ is the factorization

$$\boldsymbol{\Omega}_b^{-1} = \mathbf{L}\mathbf{A}\mathbf{A}'\mathbf{L}'$$

where $\mathbf{L}$ is the Cholesky factor of $\mathbf{V}$, and

$$\mathbf{A} = \begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 \\ n_{21} & c_2 & 0 & \cdots & 0 \\ n_{31} & n_{32} & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{J1} & n_{J2} & n_{J3} & \cdots & c_J \end{bmatrix}, \tag{5}$$

whose diagonal elements are independently distributed as χ^2 random variables and the off-diagonal elements are independently distributed as Normal random variables. More precisely, $c_j^2 \sim \chi^2_{\upsilon - j + 1}$ and $n_{jl} \sim \text{N}(0, 1)$ for $j, l = 1, \ldots, J$ with $j > l$. Using this decomposition, only $J \times (J+1)/2$ hyperparameters (cells of $\mathbf{A}$) are needed to estimate the covariance matrix $\boldsymbol{\Omega}_b^{-1}$. Note that if $\mathbf{V} = \mathbf{I}_J$, then $\mathbf{L} = \mathbf{I}_J$. Finally, to avoid order dependence with the diseases, we introduced $\mathbf{M}$ into Eq. (4) as the eigen-decomposition of $\boldsymbol{\Omega}_b^{-1}$. Chung et al. (2015) consider a family of Wishart densities for the prior of the covariance matrix and recommend the use of $\upsilon = J + 2$ degrees of freedom to make the prior a little bit more informative. In this work we follow this recommendation. Details on how to implement this in R-INLA can be found in Appendix A.

3 Scalable Bayesian models for high-dimensional multivariate disease mapping

The M-model approach can be computationally intensive when the number of areas (I) is very large. Besides, a single homogeneous model may be questionable when the number of areas grows. These limitations highlight the need for new methods. Here, we propose to use a divide and conquer strategy partitioning the spatial domain ($\mathfrak{D}$) into D subregions, so that local multivariate spatial models can be simultaneously fitted in the different subregions. In each subregion, we consider the prior distribution with Gaussian kernel and precision matrix given in Eq. (4) to deal with within-disease spatial variation and between-disease correlations.

3.1 Disjoint models

A natural way to think of partitions is to consider subregions based on administrative subdivisions of the area of interest, for example provinces, states or counties. Given a partition of the spatial domain $\mathfrak{D}$, each geographic unit belongs to a single subregion, i.e. $\mathfrak{D} = \cup_{d=1}^{D} \mathfrak{D}_d$ where $\mathfrak{D}_i \cap \mathfrak{D}_j = \varnothing$ for $i \neq j$. Then, the log-risks of the models in each subregion d ($d = 1, \ldots, D$) are expressed in matrix form as

$$\begin{aligned} \log \mathbf{R}^{(d)} &= \left(\mathbf{I}_J \otimes \mathbf{1}_{I_d}\right) \boldsymbol{\alpha}^{(d)} + \operatorname{vec}\left(\boldsymbol{\Theta}^{(d)}\right), \\ \operatorname{vec}\left(\boldsymbol{\Theta}^{(d)}\right) &\sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}_{\operatorname{vec}\left(\boldsymbol{\Theta}^{(d)}\right)}\right), \\ \boldsymbol{\Omega}_{\operatorname{vec}\left(\boldsymbol{\Theta}^{(d)}\right)} &= \left[\left(\mathbf{M}^{(d)}\right)^{-1} \times \mathbf{I}_{I_d}\right] \text{Blockdiag} \\ &\quad \left(\boldsymbol{\Omega}_1^{(d)}, \ldots, \boldsymbol{\Omega}_J^{(d)}\right) \left[\left(\mathbf{M}^{(d)}\right)^{-1} \times \mathbf{I}_{I_d}\right]' \end{aligned} \tag{6}$$

where for each subregion d, $\boldsymbol{\alpha}^{(d)} = (\alpha_1^{(d)}, \ldots, \alpha_J^{(d)})'$ and $\alpha_j^{(d)}$ is a disease-specific intercept, $\mathbf{R}^{(d)} = \left(\mathbf{R}_1^{(d)'}, \cdots, \mathbf{R}_J^{(d)'}\right)'$, and each $\mathbf{R}_j^{(d)} = (R_{1j}^{(d)}, \ldots, R_{Ij}^{(d)})'$ is the vector of relative risks corresponding to disease j within the subregion d. Finally, $\mathbf{I}_{I_d}$ is the identity matrix of order I_d and $\mathbf{1}_{I_d}$ is a column vector of ones of length I_d (the number of areas within partition d), $I = \sum_{d=1}^{D} I_d$, and $\boldsymbol{\Theta}^{(d)} = \{\theta_{ij}^{(d)} : i = 1, \ldots, I_d; j = 1, \ldots, J\}$ is the matrix of spatial effects in partition d including both within and between-disease dependence structure. In more detail, this model can be expressed as

$$\begin{pmatrix} \log \mathbf{R}^{(1)} \\ \vdots \\ \log \mathbf{R}^{(d)} \\ \vdots \\ \log \mathbf{R}^{(D)} \end{pmatrix} = \mathbf{I}_J \otimes \begin{pmatrix} \mathbf{1}_{I_1} & & & \\ & \ddots & & \\ & & \mathbf{1}_{I_d} & \\ & & & \ddots \\ & & & & \mathbf{1}_{I_D} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}^{(1)} \\ \vdots \\ \boldsymbol{\alpha}^{(d)} \\ \vdots \\ \boldsymbol{\alpha}^{(D)} \end{pmatrix} + \begin{pmatrix} \operatorname{vec}\left(\boldsymbol{\Theta}^{(1)}\right) \\ \vdots \\ \operatorname{vec}\left(\boldsymbol{\Theta}^{(d)}\right) \\ \vdots \\ \operatorname{vec}\left(\boldsymbol{\Theta}^{(D)}\right) \end{pmatrix}$$

where the precision matrix of the multivariate Normal random effect vector $\left(\operatorname{vec}\boldsymbol{\Theta}^{(1)'}, \ldots, \operatorname{vec}\boldsymbol{\Theta}^{(D)'}\right)'$ is a block-diagonal matrix of dimension $IJ \times IJ$ whose blocks correspond to the precision matrices $\boldsymbol{\Omega}_{\operatorname{vec}\left(\boldsymbol{\Theta}^{(d)}\right)}$, $d = 1, \ldots, D$. The full domain log-risk is just the union of the posterior estimates of each subregion, i.e., $\log \mathbf{R} = \left(\log \mathbf{R}^{(1)'}, \cdots, \log \mathbf{R}^{(D)'}\right)'$.

3.2 Models with overlapping partitions

Disjoint partitions might suffer from border effects as areas in the boundary of a given partition would not borrow information from neigbouring areas from a contiguous subdivision. Consequently, the risk estimates in those areas may not be correct. This inconvenience can be solved by considering an alternative modelling approach in which k-order neighbours are added to each subregion of the partition, so that border areas have neighbours from other subregion of the partition. In this case, the entire spatial region $\mathfrak{D}$ is divided into a set of overlapping subregions and some small areas belong to more than one subdivision, i.e., $\mathfrak{D} = \cup_{d=1}^{D} \mathfrak{D}_d$ and $\mathfrak{D}_i \cap \mathfrak{D}_j \neq \varnothing$ for neighbouring subregions. Similar to the disjoint Model (6), D submodels will be simultaneously fitted. However, as $\sum_{d=1}^{D} I_d > I$, the final risk $\mathbf{R} = (\mathbf{R}_1', \ldots, \mathbf{R}_J')'$ with $\mathbf{R}_j' = (R_{1j}, \ldots, R_{Ij})'$, $j = 1, \ldots, J$, is no longer the union of the posterior estimates obtained for each submodel as areas located in the borders of the spatial partition would have more than one estimated posterior distribution.

Two different strategies can be considered to obtain a unique posterior estimate of the relative risk for those areas in more than one subregion. Orozco-Acosta et al. (2021) propose to calculate the mixture distribution of the estimated posterior probability density functions of the relative risks in the different subdivisions, with weights proportional to the conditional predictive ordinate (CPO) values (Pettit 1990). To compute the mixture, suppose that area i belongs to $m(i)$ subregions of the spatial domain $\mathfrak{D}$ and let $f_{ij}^{(1)}(x), \cdots, f_{ij}^{(m(i))}(x)$ be the posterior estimates of the probability density functions of the j-th disease in the i-th area. Then the mixture distribution of R_{ij} can be written as

$$f_{ij}(x) = \sum_{k=1}^{m(i)} w_k f_{ij}^{(k)}(x), \quad \text{with} \quad w_k = \frac{CPO_{ij}^k}{\sum_{k=1}^{m(i)} CPO_{ij}^k}$$

where CPO_{ij}^k is the conditional predictive ordinate of area i and disease j obtained in partition k, so that $w_k \geq 0$ and $\sum_{k=1}^{m(i)} w_k = 1$ (see for example Lindsay 1995; Frühwirth-Schnatter 2006).

More recently, Orozco-Acosta et al. (2023) consider using the posterior marginal distribution of the relative risk estimated from its original partition. Based on the results obtained from a simulation study, they show that this strategy outperforms the use of mixture distributions in terms of risk estimation accuracy and true positive/negative rates. In this paper, this is also the default strategy used to obtain unique posterior distributions for each relative risk R_{ij}.

3.3 Between-disease correlations and variance parameters

Besides increasing the effective sampling size and improving risk smoothing, one of the main advantages of multivariate disease mapping models is that they take into account correlations between the spatial patterns of the different diseases, that is, they reveal connections between them. Fitting a single multivariate model to the region of interest provides correlations between the diseases in the whole study domain thus revealing overall relationships. In addition, it also provides the diagonal elements of the between-disease covariance matrix, hereafter referred to as variance parameters. In the case of separable covariance structures (the Kronecker product of between and within disease covariance matrices) these parameters control the amount of smoothing within diseases. By dividing the spatial domain into subregions, we obtain the posterior distributions of these parameters in each of the subdivisions and we retrieve the between disease correlations and variances for the entire region. Hence, partition models provide additional information by revealing local connections between diseases in the subdivisions, which are usually based on administrative divisions.

To obtain global estimates of the parameters of interest in the overall study domain from the partition models, we adapt the consensus Monte Carlo (CMC) algorithm originally proposed by Scott et al. (2016). The idea behind consensus Monte Carlo is to divide the data into shards (in our case, the shards corresponds to different subdivisions of the spatial domain), give each shard to a worker machine which does a full Monte Carlo simulation from a posterior distribution given its own data, and then combine the posterior simulations from each worker (or submodel) to produce a set of global draws representing the consensus belief among all the workers. Here, we briefly describe how to adapt the ideas behind the CMC algorithm to our case.

Let $\boldsymbol{\psi} = (\boldsymbol{\rho}, \boldsymbol{\sigma}^2)'$ denotes the vector with the parameters of interest where $\boldsymbol{\rho} = (\rho_{12}, \ldots, \rho_{J-1,J})'$ contains the between-disease correlations and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_J^2)'$ are the diagonal elements of the between-disease covariance matrix, and let ψ_{kd} denote the local estimate of the k-th parameter of $\boldsymbol{\psi}$ in each subdomain $\mathfrak{D}_d$, $d = 1, \ldots, D$. We first extract samples of size S from the posterior marginal estimates of ψ_{kd} denoted as ψ_{kd}^s for $k = 1, \ldots, J \times (J+1)/2$, $d = 1, \ldots, D$ and $s = 1, \ldots, S$. Then, we combine the draws using weighted averages

$$\tilde{\psi}_k^s = \sum_{d=1}^{D} w_d \psi_{kd}^s, \quad \text{for } s = 1, \ldots, S$$

where w_d are normalized weights inversely proportional to the posterior marginal variances of ψ_{kd}. Finally, we approximate the posterior marginal density function of the parameter ψ_k from the combined draws $\tilde{\psi}_k^s$.

3.4 Model selection criteria

Two of the most widely used criteria to compare Bayesian models are the deviance information criterion (DIC) (Spiegelhalter et al. 2002) and the Watanabe-Akaike information criterion (WAIC) (Watanabe 2010). However, with partition models, it is not straightforward to get these quantities as we fit as many models as subdivisions. Hence, we need a procedure to estimate these quantities from the scalable models described in Sects. 3.1 and 3.2.

Extending the ideas in Orozco-Acosta et al. (2021) to the multivariate framework, we compute approximate DIC values by drawing samples from the posterior marginal distribution of the Poisson means. Denoting by $\mathbf{C}^s, s = 1, ..., S$, to the posterior simulations of $\mu_{ij} = E_{ij} \cdot R_{ij}$ (the mean of the Poisson distribution), approximate values of the mean deviance $\overline{D(\mathbf{C})}$ and the deviance of the mean $D(\overline{\mathbf{C}})$ can be respectively calculated as

$$\overline{D(\mathbf{C})} = \frac{1}{S} \sum_{s=1}^{S} -\log\left(p(\mathbf{O}|\mathbf{C}^s)\right);$$

$$D(\overline{\mathbf{C}}) = -2\log\left(p(\mathbf{O}|\overline{\mathbf{C}})\right),$$

$$\text{with } \overline{\mathbf{C}} = \frac{1}{S}\sum_{s=1}^{S}\mathbf{C}^s,$$

where $p(\mathbf{O}|\cdot)$ denotes the likelihood function of a Poisson distribution. Then, the DIC is obtained as

$$\text{DIC} = 2\,\overline{D(\mathbf{C})} - D(\overline{\mathbf{C}}).$$

Similarly, approximate WAIC values are computed as (see Gelman et al. 2014)

$$\text{WAIC} = -2\sum_{i=1}^{I}\sum_{j=1}^{J}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(O_{ij}|\mathbf{C}^s)\right) + 2\sum_{i=1}^{I}\sum_{j=1}^{J}\text{var}\left[\log\left(p(O_{ij}|\mathbf{C}^s)\right)\right].$$

4 Simulation study

We conduct a simulation study to compare the performance of the different M-models described in Sect. 2. Specifically, our interest relies on comparing the fit of a single model to the whole domain (hereafter referred to as the global model) and the partition models, in terms of parameter estimates and relative risk estimation accuracy. The $I = 7907$ municipalities of continental Spain and $J = 3$ diseases are used as the simulation template because this imitates the case study presented in Sect. 5.

Two different scenarios have been considered to recover the possible underlying generating process of spatially correlated disease risks. In the first scenario, samples are generated from a fixed covariance structure based on the spatial neighbourhood graph of the whole area under study, that is, the global model is used as the generating model. In contrast, in the second scenario, independent samples for each partition (Spanish Autonomous Regions, see Fig. 5 in Appendix B) are generated using the covariance structures of the partition, that is, the Disjoint model is used as the data generating mechanism. Further details are given below.

4.1 Data generation

One advantage of multivariate models is their ability to reveal relationships between different diseases in terms of correlations between their underlying spatial patterns. To evaluate how well these correlation parameters are estimated, we start by sampling from a multivariate Normal distribution with precision matrix $\boldsymbol{\Omega}_{\text{vec}(\boldsymbol{\Theta})} = \boldsymbol{\Omega}_b \otimes \boldsymbol{\Omega}_{iCAR}$. Here, the elements of the between-disesease covariance matrix are fixed, that is,

$$\Omega_b^{-1} = \begin{pmatrix}\sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3\end{pmatrix}\begin{pmatrix}1 & \rho_{12} & \rho_{13}\\ \rho_{21} & 1 & \rho_{23}\\ \rho_{31} & \rho_{32} & 1\end{pmatrix}\begin{pmatrix}\sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3\end{pmatrix} = \begin{pmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_2^2 & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_3^2\end{pmatrix}$$

where σ_j^2 are variance parameters, and $\rho_{kl} = \rho_{lk}$ are between-disease correlation coefficients. Note that σ_{kl} denotes the covariances between each pair of diseases. Then, for each sample of $\text{vec}(\boldsymbol{\Theta}^r)$, $r = 1, \ldots, 100$ we compute the relative risks R_{ij}^r following Eq. (3). Finally, we generate O_{ij} counts for area i and disease j using a Poisson distribution with mean $\mu_{ij}^r = E_{ij} \cdot R_{ij}^r$, where E_{ij} are the expected number of cases of our case study data (lung, colorectal and stomach cancer mortality in Spanish males).

In Scenario 1, the neighbourhood graph of all the 7907 municipalities is used to define the spatial precision matrix $\boldsymbol{\Omega}_{iCAR}$ (the global model is used to generate the data). In addition, we fix the parameters of the between-disease covariance matrix as $\sigma_1^2 = 0.25$, $\sigma_2^2 = 0.16$, $\sigma_3^2 = 0.09$, $\rho_{12} = 0.7$, $\rho_{13} = 0.5$ and $\rho_{23} = 0.1$. In Scenario 2, $D = 15$ independent samples are generated from multivariate Normal distributions with precision matrices equal to $\boldsymbol{\Omega}_{\text{vec}(\boldsymbol{\Theta}^d)} = \boldsymbol{\Omega}_b^{(d)} \otimes \boldsymbol{\Omega}_{iCAR}^{(d)}$, where $\boldsymbol{\Omega}_{iCAR}^{(d)}$ is the spatial precision matrix of the areas within subdomain $d = 1, \ldots, D$, and different between-disease covariance matrices $\boldsymbol{\Omega}_b^{(d)}$ are considered en each subdivision (the disjoint model, k = 0, is used to generate the data). Here, the variance parameters are fixed to $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.4$ and $\sigma_3^2 = 0.3$, while similar values to the ones estimated with the partition models in the case study presented in the next section are used as correlation coefficients (see Table 6 in Appendix B). We increase the variance parameters in Scenario 2 to get stronger smoothing effects in each subdivision. Note that the variance parameters are the same in all the subdivisions, but they cannot be considered as global variance parameters because the covariance structures, based on the neighbourhood matrices, are different. Hence, in this scenario we do not have true parameter values for the global model.

4.2 Results: Scenario 1

Table 1 compares the true values of model parameters in Scenario 1 (variance parameters and correlation coefficients) against average values of posterior mean estimates over the 100 simulated data sets. In addition, estimated standard errors, simulated standard errors (derived from the sample variance of the parameter estimates) and empirical coverages of the 95% credible intervals are also displayed. Note that

for the partition models, these posterior marginal distributions are obtained by using the CMC algorithm described in Sect. 3.3. In terms of model parameters, multivariate models give very accurate estimates of the real values, both in terms of posterior mean and posterior standard deviation estimates (note that nearly identical values are obtained from estimated and simulated standard errors). As expected, slightly better results are obtained when fitting the global model, as this is the true generating model in Scenario 1. Regarding partition models, the higher the neighbourhood order, the more similar the CMC estimates of the correlation coefficients are to those of the global model.

Table 2 displays average values of model selection criteria (posterior mean deviance $\overline{D(\boldsymbol{\theta})}$, effective number of parameters p_D, DIC and WAIC) for the global and the partition models, as well as the accuracy of the relative risk estimates quantified by the mean absolute relative bias (MARB), the mean relative root mean square errors (MRRMSE) and empirical coverages of the 95% credible intervals for the risks. Note that the MARB and MMRMSE are defined for each small area i and disease j as

$$\text{MARB}_{ij} = \left| \frac{1}{100} \sum_{r=1}^{100} \frac{\hat{R}_{ij}^{(r)} - R_{ij}^{(r)}}{R_{ij}^{(r)}} \right| \quad \text{and}$$

$$\text{MMRMSE}_{ij} = \sqrt{\frac{1}{100} \sum_{r=1}^{100} \left(\frac{\hat{R}_{ij}^{(r)} - R_{ij}^{(r)}}{R_{ij}^{(r)}} \right)^2}$$

where $R_{ij}^{(r)}$ and $\hat{R}_{ij}^{(r)}$ denote the true value and the posterior median estimate of the relative risks for the r-th data set ($r = 1, \ldots, 100$). Model selection criteria point towards partition models, though differences are mild. Regarding MARB, MMRMSE and 95% coverage values, differences between the global and the partition models are practically negligible.

4.3 Results: Scenario 2

In contrast to the previous scenario, it should be noted that in Scenario 2 we cannot compare the global estimates of the model parameters against the true values of the variance parameters and between-disease correlations, since different values have been used to generate the risk surfaces in each subdomain and we do not have true global values. However, we can compare the model's performance in terms of model selection criteria and risk estimation accuracy (see Table 3). As expected, the Disjoint model ($k = 0$) shows the best performance according to these measures, as this is the true generating model in Scenario 2. In terms of MARB and MRRMSE, partition models also outperform the Global model.

We are also interested in analyzing if the partition models are able to recover the local between-disease covariance structures of the true generating process. In Table 6 (Appendix B) we compare these values against the average values of posterior mean estimates of local parameters in each subdivision over the 100 simulated data sets for the Disjoint model. For almost every subdivision, very accurate estimates are obtained for both variance parameters and correlation coefficients. For the latter, the median value of the empirical coverage of the 95% credible intervals is 0.95 (with $Q_1 = 0.93$ and $Q_3 = 0.97$). As expected, these estimates get worse as the neighbourhood order of the models increases, since the estimated local correlations correspond to enlarged subdivision rather than the subdivisions themselves. Even so, the median values of the empirical coverage of the 95% credible intervals for the between-disease correlations are 0.89 (with $Q_1 = 0.84$ and $Q_3 = 0.92$) and 0.86 (with $Q_1 = 0.79$ and $Q_3 = 0.90$) for 1st-order and 2nd-order neighbourhood models, respectively. All the results are shown in Tables 7 and 8 in Appendix B.

5 Case study

In this section we jointly analyse mortality data for lung, colorectal, and stomach cancer in men in the 7907 municipalities of mainland Spain (excluding Baleares and Canary Islands and the autonomous cities of Ceuta and Melilla) during the period 2006-2015 using the new proposal. During the ten years of the study, a total of 162,602 deaths from lung cancer (corresponding to codes C33-C34 of the International Classification of Diseases-10), 82,967 from colorectal cancer (C17-C21) and 33,170 from stomach cancer (C16) were registered for male population of mainland Spain, which correspond to global rates of 76.48, 39.02 and 15.60 deaths per 100,000 male inhabitants, respectively.

5.1 Model fitting and model selection

We fit the disjoint model ($k = 0$) and the k-order neighbourhood model for $k = 1, 2, 3$ in R-INLA using $D = 15$ subdivisions of the spatial domain. These subdivisions are also of interest as they correspond to Autonomous Regions of Spain (NUTS2 level from the European nomenclature of territorial units for statistics, shown in Fig. 5 in Appendix B). In these partitions, the highest value of I_d (number of municipalities) is 2245 and corresponds to the Autonomous Region of Castilla y León, a rather vast territory from central to northwestern Spain with about 5% of the total Spanish population. Although this subregion is large, we maintain this subdivision as it represents the administrative division of Spain into Autonomous Regions. We also fit the multivariate spatial M-models over the entire spatial domain (global

Table 1 Average values of posterior mean, posterior standard deviation (SD), simulated standard errors (sim) and empirical coverage of the 95% credible intervals (EC) for model parameters based on 100 simulated data sets for Scenario 1

	True value	Mean	SD	Sim	EC	Mean	SD	Sim	EC
		Global model				Disjoint model			
σ_1^2	0.25	0.250	0.011	0.011	0.95	0.240	0.012	0.012	0.83
σ_2^2	0.16	0.160	0.010	0.010	0.95	0.158	0.011	0.011	0.96
σ_3^2	0.09	0.092	0.009	0.010	0.92	0.101	0.010	0.011	0.74
ρ_{12}	0.70	0.700	0.025	0.026	0.95	0.690	0.026	0.029	0.89
ρ_{13}	0.50	0.487	0.044	0.046	0.95	0.452	0.045	0.048	0.80
ρ_{23}	0.10	0.089	0.059	0.057	0.96	0.077	0.057	0.065	0.95
		1st-order nb model				2nd-order nb model			
σ_1^2	0.25	0.241	0.011	0.012	0.84	0.239	0.010	0.012	0.75
σ_2^2	0.16	0.159	0.010	0.010	0.94	0.155	0.009	0.010	0.89
σ_3^2	0.09	0.100	0.010	0.010	0.79	0.097	0.009	0.010	0.83
ρ_{12}	0.70	0.691	0.025	0.029	0.92	0.695	0.023	0.032	0.82
ρ_{13}	0.50	0.461	0.043	0.051	0.81	0.468	0.040	0.048	0.83
ρ_{23}	0.10	0.079	0.055	0.058	0.91	0.082	0.053	0.060	0.95

Table 2 Average values of model selection criteria (mean deviance, effective number of parameters, DIC and WAIC) and risk estimation accuracy (MARB, MRRMSE and empirical coverage -EC- of the 95% credible intervals) based on 100 simulated data sets for Scenario 1

	Model selection criteria				Risk estimation accuracy		
	$\overline{D(\boldsymbol{\theta})}$	p_D	DIC	WAIC	MARB	MRRMSE	EC
Global	78521.9	3046.9	81568.8	81504.7	0.024	0.191	0.950
Disjoint ($k = 0$)	78299.9	3329.3	81629.1	81529.2	0.023	0.196	0.957
1st-order nb ($k = 1$)	78407.9	3154.7	81562.6	81499.4	0.024	0.193	0.953
2nd-order nb ($k = 2$)	78454.5	3091.5	81546.0	81496.5	0.024	0.192	0.950

Table 3 Average values of model selection criteria (mean deviance, effective number of parameters, DIC and WAIC) and risk estimation accuracy (MARB, MRRMSE and empirical coverage -EC- of the 95% credible intervals) based on 100 simulated data sets for Scenario 2

	Model selection criteria				Risk estimation accuracy		
	$\overline{D(\boldsymbol{\theta})}$	p_D	DIC	WAIC	MARB	MRRMSE	EC
Global	78766.9	5385.6	84152.5	83894.7	0.062	0.322	0.947
Disjoint ($k = 0$)	78505.8	5132.6	83638.4	83451.3	0.051	0.299	0.954
1st-order nb ($k = 1$)	78420.2	5465.5	83885.7	83650.9	0.055	0.314	0.957
2nd-order nb ($k = 2$)	78457.1	5460.2	83917.3	83694.3	0.057	0.317	0.955

model), and compare the results with those obtained with the new proposal.

Previously, univariate models were also fitted to each disease using a BYM2 spatial prior. The covariance matrix of this prior copes with both spatial structured variability and unstructured variability. Results (not shown here to conserve space) show that most of the variability is spatially structured. Since the computational cost of this prior makes it difficult its use in a multivariate setting, and most of the variability is spatially structured, we fit the joint multivariate proposal given in Eq. (6) by considering an iCAR prior for the spatial random effects.

For the partition models, we distribute the submodels over 2 machines with four processors Intel Xeon Silver 4108 and 192GB RAM on each machine (Ubuntu 20.04.4 LTS operative system), using the simplified Laplace approximation strategy in R-INLA (Lindgren and Rue 2015) (stable version INLA_22.05.07, R version R−4.1.2) and simultaneously running 3 models in parallel on each machine using the `bigDM` package (Adin et al. 2023).

Table 4 displays the posterior mean deviance $\overline{D(\boldsymbol{\theta})}$, the effective number of parameters p_D, the DIC, and the WAIC for the global and the scalable models together with the computing time (in minutes). The total time for the scalable models is obtained by adding the running time and the merging time. The running time refers to the elapsed time for all the submodels fitted with R-INLA, and the merging time refers to the combination (when necessary) of the posterior distributions of the risks, the approximation of the DIC/WAIC values, and the computation of global estimates of the between-diseases correlation coefficients using the proposed CMC algorithm. As expected, the computational

Table 4 Model selection criteria and computational time, in minutes, for multivariate models with iCAR spatial prior using the simplified Laplace approximation strategy if INLA

Model	Model selection criteria				Time (in min)		
	$\overline{D(\boldsymbol{\theta})}$	p_D	DIC	WAIC	Run	Merge	Total
k = 0	76779.7	2471.9	79252.6	79204.8	5.4	0.7	6.1
k = 1	76894.6	2327.4	79222.0	79187.3	6.5	1.1	7.6
k = 2	76942.0	2289.4	79231.4	79211.9	7.7	1.1	8.8
k = 3	77007.0	2231.8	79238.8	79220.0	8.2	1.1	9.3
Global	77186.8	2164.2	79351.0	79283.9	33.2	–	33.2

cost raises as the neighbourhood order (k) increases, though the scalable proposal is faster than the global model for all values of k. The greatest reduction in time in comparison with the global model is obtained for $k = 0$, being the global model about 5.5 times slower. When the neighbourhood order increases, the difference in computing time is less pronounced. The global model is about 4.3, 3.8, and 3.6 times slower than the scalable models with k =1, 2, and 3, respectively. Regarding model selection criteria, scalable Bayesian models outperform the global model. The greater reduction in DIC and WAIC is obtained for the 1st-order neighbourhood model. However, increasing the neighbourhood order may improve the between-disease correlation estimates.

Table 5 Descriptive statistics of the estimated between-disease correlations with the global, and $k = 0, 1, 2$-order neighbourhood models, using an iCAR prior for spatial random effects

ρ	Model	Mean	SD	$q_{.025}$	$q_{.5}$	$q_{.975}$	Mode
$\rho_{1.2}$	Global	0.70	0.04	0.63	0.70	0.77	0.70
	k = 0	0.66	0.04	0.58	0.66	0.73	0.66
	k = 1	0.68	0.04	0.60	0.68	0.74	0.68
	k = 2	0.71	0.03	0.65	0.71	0.77	0.71
$\rho_{1.3}$	Global	0.46	0.05	0.36	0.46	0.55	0.46
	k = 0	0.55	0.05	0.46	0.55	0.63	0.55
	k = 1	0.55	0.04	0.47	0.56	0.63	0.56
	k = 2	0.50	0.04	0.42	0.50	0.57	0.50
$\rho_{2.3}$	Global	0.57	0.05	0.46	0.57	0.67	0.57
	k = 0	0.54	0.05	0.43	0.54	0.64	0.54
	k = 1	0.56	0.05	0.45	0.56	0.65	0.56
	k = 2	0.59	0.04	0.50	0.59	0.68	0.60

5.2 Joint analysis of male mortality from three types of cancer in Spain

In this subsection, the spatial patterns of lung, colorectal, and stomach cancer mortality risks in men are examined in the municipalities of continental Spain using the scalable multivariate proposal presented in Sect. 3.

We begin with a comparison of the estimated risks obtained with the global model, the disjoint model ($k = 0$) and the k-order neighbourhood models ($k = 1$, 2 and 3). Figure 1 displays dispersion plots of the posterior median estimates of the relative risks obtained with the partitioned models versus those obtained with the global model. The left, central and right columns correspond to lung, colorectal and stomach cancer, respectively. The neighbourhood order in the partition models are represented in the different rows. The largest differences are observed between the global and the disjoint models. This is expected because areas in the border of a subdivision do not borrow strength from neighbouring areas located in a contiguous subdivision. As the neighbourhood order k increases, the risk estimates are more similar to the global model. Figure 2 displays the spatial patterns of lung cancer mortality risks (top) and the posterior probabilities of risk exceedance (bottom), $P(R_{ij} > 1|\mathbf{O})$, obtained with the global, the disjoint (k = 0) and the partition models (k = 1, 2, 3). To save space, maps for colorectal and stomach cancer are provided in Figs. 6 and 7 (Appendix B). Though differences in risks estimates are observed in the dispersion plots, it is harder to appreciate them on the maps.

Multivariate models borrow information from nearby areas and the different diseases. Additionally, they present other advantages over univariate counterparts, such as the possibility of estimating correlations between the spatial patterns of the diseases. Moderate to high correlations may suggest the existence of underlying risk factors affecting the diseases under study, which in turns implies connection between them. This information may be crucial to better understand diseases such as cancer in which known risk factors only explain a small percentage of the cases. Spatial patterns may be associated to factors like access to treatment or life style that might have an impact on mortality.

Posterior distributions of the between-disease correlations obtained with the disjoint (k = 0) and the partition models (k = 1, 2) are displayed in Fig. 3 together with correlations for whole Spain obtained with the CMC algorithm and with the global model. Here, $\rho_{1.2}$, $\rho_{1.3}$, and $\rho_{2.3}$ denote the correlation parameters between lung and colorectal, lung and stomach, and colorectal and stomach cancer, respectively. Summary statistics (mean, median, mode, standard deviation, 2.5 and 97.5 percentiles) of the between-disease posterior correla-

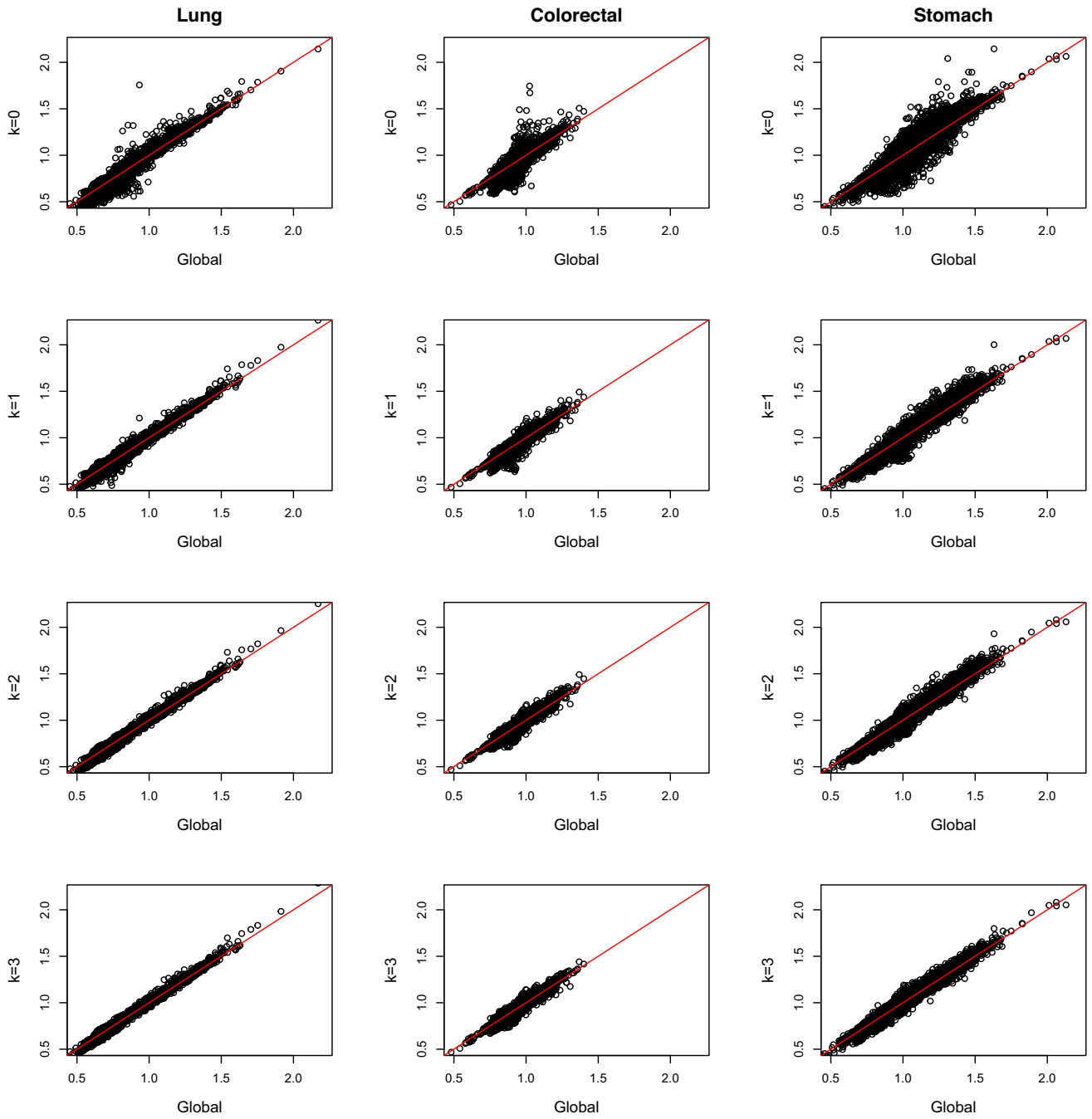

Fig. 1 Dispersion plots of the posterior median estimates of relative risks for lung (left column), colorectal (central column) and stomach (right column) cancer mortality data obtained with the partitioned model ($k = 0, 1, 2, 3$ from top to bottom) versus the global model

tions are also shown in Table 5. In general, the posterior distributions estimated with the CMC algorithm for the partition models are very similar to those obtained with the global model. Similar to the posterior estimates of the relative risks, closer values to the global model are observed as the neighbourhood order k increases.

Finally, Fig. 4 displays a map with the posterior medians and standard deviations of the between-diseases correlations $\rho_{1,2}$ (left), $\rho_{1,3}$ (center), and $\rho_{2,3}$ (right), for the different subdivisions (Autonomous Regions) obtained with the 1st-order neighbourhood partition model. Partition models can provide the correlations over the whole study domain, but also the correlations for the different subdivisions. This is an advantage over the global models as we add information at different administrative divisions. Moreover, the variability in the posterior medians of the correlations across the subdivisions may indicate a lack of stationarity that the global model cannot cope with, and hence the advantages of the

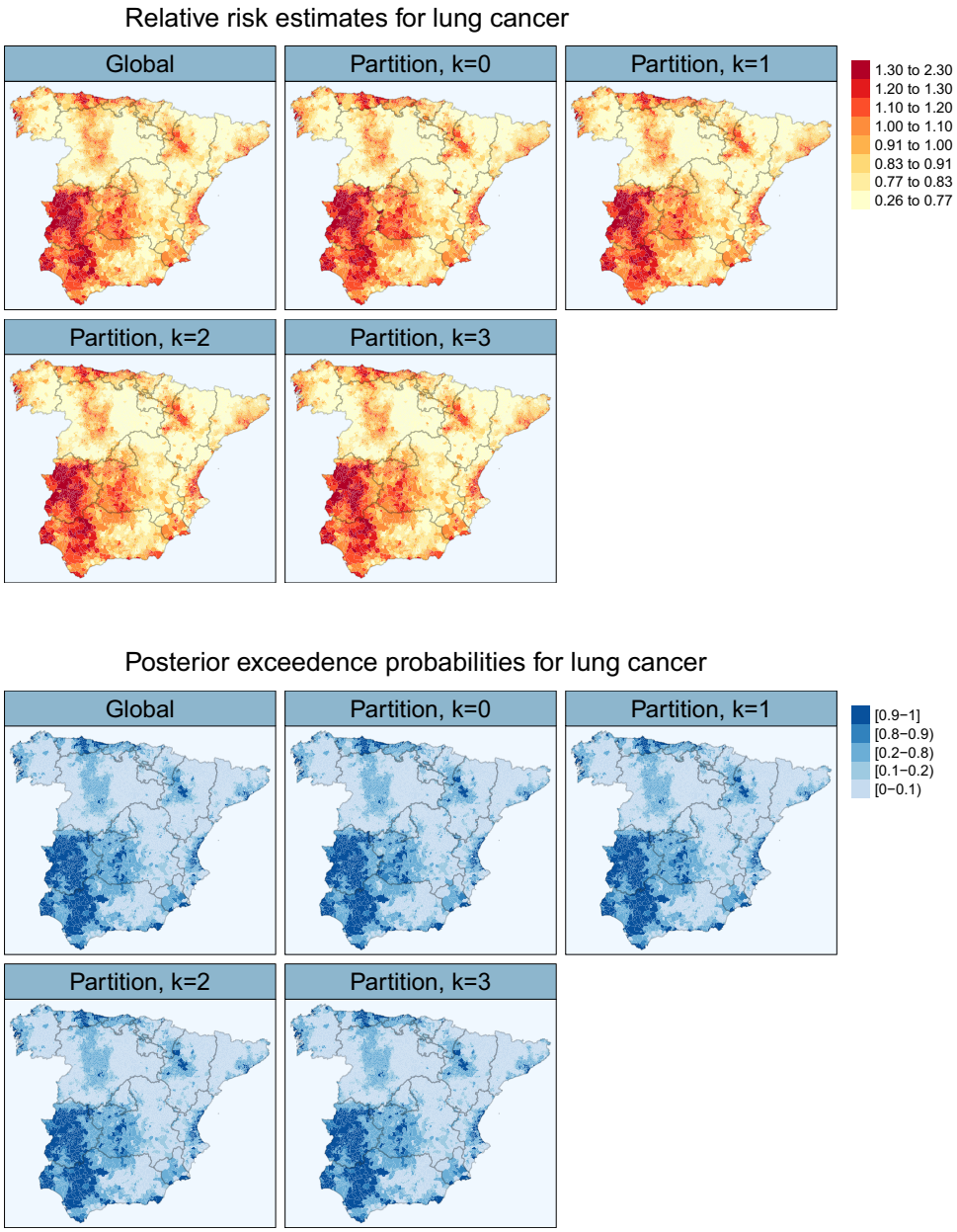

Fig. 2 Maps of posterior median estimates of mortality relative risk for lung cancer (top) and posterior exceedance probabilities $P(R_{ij} > 1|\mathbf{O})$ (bottom) in continental Spain

partition models. When the number of small areas is large, the use of a global model with one single precision (smoothing) parameter may be questionable while local models add more flexibility to deal with the spatial heterogeneity across the map.

6 Discussion

Spatial areal models have a long tradition in epidemiology to study the geographical pattern of a disease. While initially focused on modelling a single disease, spatial models have evolved into a multivariate framework with two notable objectives: to improve estimates by borrowing strength from other diseases and neighbouring areas, and to estimate latent correlations between the spatial patterns of the diseases under study to address the connections between them and to hypothesize common risk factors. Research on spatial multivariate models has received considerable attention in recent years, although their use is not yet widespread in epidemiology mainly because (i) the implementation of multivariate models in available software requires advanced computing skills and (ii) computational issues are accentuated when the number of small areas is large as computing time may become prohibitive. Vicente et al. (2020b, 2021) provide an implementation of multivariate CAR and P-splines in R-INLA that

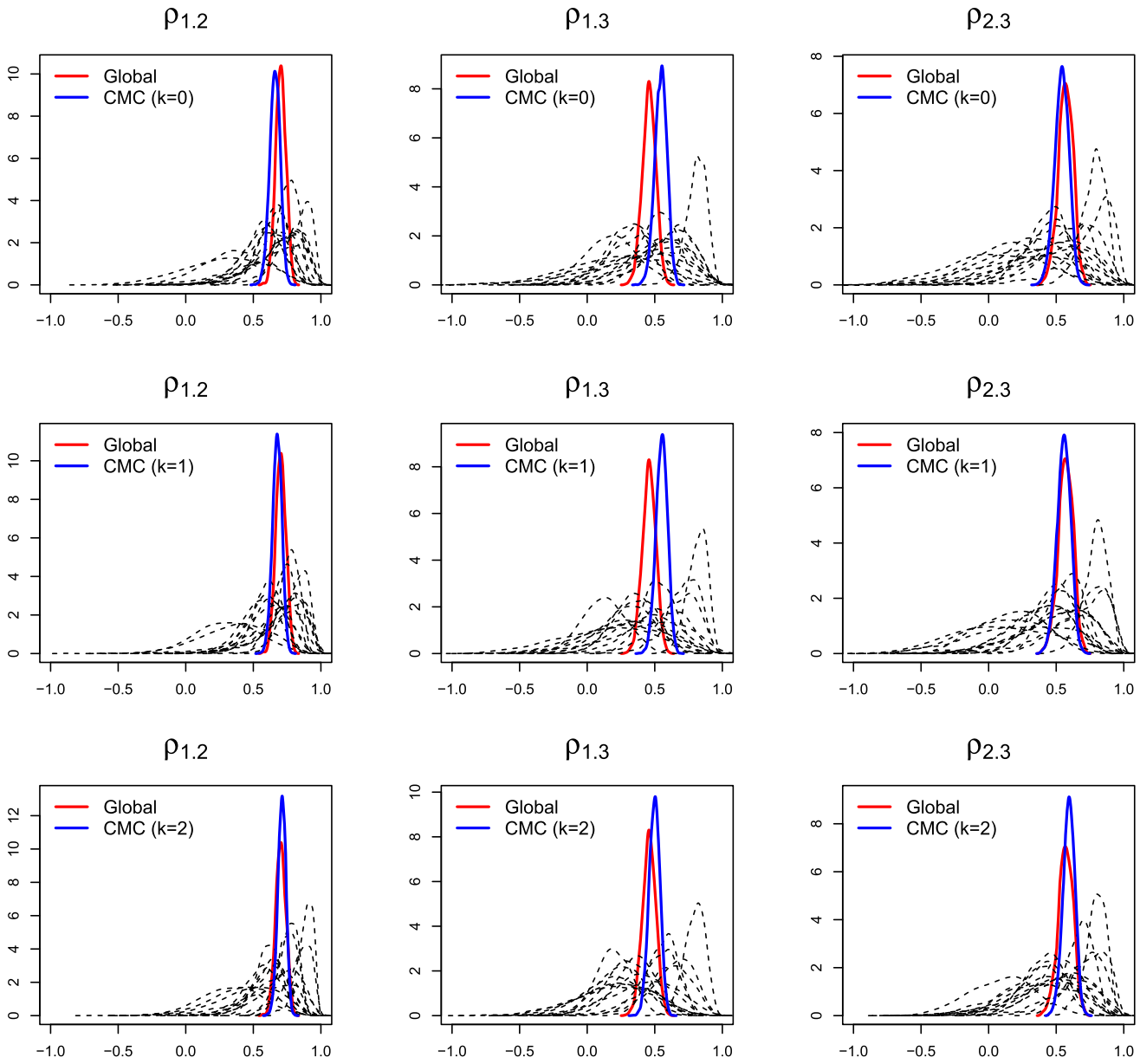

Fig. 3 Posterior distributions of the estimated between-disease correlations with the global, and $k = 0, 1, 2$-order neighbourhood models, using an iCAR prior for spatial random effects

can be used by a wide audience without advanced computer skills.

In this paper, we present a new approach to analyse multivariate areal count data when the number of small areas is very large. In particular, we combine the methodology proposed by Orozco-Acosta et al. (2021) for high-dimensional disease mapping with a modification of the multivariate approach given by Botella-Rocamora et al. (2015) to avoid overparameterization, obtaining a scalable Bayesian modelling approach to multivariate disease mapping. Our proposal begins with the partitioning of the spatial domain into subregions with substantially fewer small areas. The multivariate models can then be fitted simultaneously (using both parallel or distributed computation strategies) in each of these regions, reducing computational time and avoiding memory and storage problems. Dividing the whole spatial domain into disjoint regions may induce border effects as the areas in the limits of a given subdivision do not borrow information from neighbouring areas located in a different subregion. To overcome this issue, we consider k-order neighbourhood models that incorporate neighbouring areas to those regions located on the partition boundary. Finally, variance parameters and between-disease correlations for the whole area are obtained by means of an adaptation of a consensus Monte Carlo algorithm. The correlation coefficients indicate potential geographic factors related (or not) to the

different diseases. If the covariance structure is separable, the variance parameters measure the amount of smoothing for each disease. In addition to the CMC algorithm, we have also considered the Weierstrass rejection sampler (WRS) proposed by Wang and Dunson (2013) to recover the parameters of interest for the whole study region (results not shown to save space). In this algorithm, the posterior of the target distribution in the whole area is approximated by combining posterior samples of the subdivisions using rejection sampling. Though it was originally proposed to combine posterior draws from independent MCMC subset chains, it can be adapted to other Bayesian estimation techniques such as INLA through the R package `weierstrass` (available at https://github.com/wwrechard/weierstrass). In general, very similar posterior marginal estimates are obtained with both algorithms.

One of the key issues with partition models is to choose the neighbourhood order. Here we use model selection criteria such as DIC and WAIC. Our conclusions are that, in general, the larger the neighbourhood order, the more similar the partition model is to the global model. However, increasing too much the neighbourhood order, the benefits of our proposal in terms of computational time vanish. Overall, first or second order neighbourhood models are appropriate. From the simulation study, we conclude that even when the underlying generating process is the Global model, the partition models are very competitive in terms of risk estimation accuracy. Moreover, the global between-disease correlation coefficients are well recovered with the partition models. If the geographical distribution and correlation structure of the underlying process varies across the whole map (which seems very realistic in practice), better results are obtained with our modelling proposals than with the usual global model.

Very recently, a new hybrid approximate procedure that uses the Laplace method with a low-rank variational Bayes correction has been proposed as part of the R-INLA project (Van Niekerk and Rue 2021; Van Niekerk et al. 2023). The latest versions of the R-INLA package allow to run the models using this new approximation strategy (named as "compact" mode) resulting in a substantial reduction in computational time. This new approximation method appears to be very promising. However, further research is necessary to explore its accuracy in estimating hyperparameters, such as between-disease correlations.

Moreover, when there is a large number of areas, the suitability of a global homogeneous model (with a single precision/smoothing parameter) for the entire study region may be doubtful. Instead, implementing various local homogeneous models can provide increased flexibility in capturing the spatial heterogeneity present across the map.

In conclusion, it can be argued that partition models offer several advantages over a global model. Firstly, they accelerate computations through the classical integrated nested Laplace approximations and alleviate storage and memory problems. Secondly, they offer a dual benefit. Even if the global model is appropriate, we can provide both a global spatial pattern for the entire region and local patterns for the subdivisions, which is particularly beneficial for our case. Lastly, it's worth noting that as the number of diseases grows, so does the number of hyperparameters in the covariance matrix, resulting in a greater computational burden. This issue warrants further research.

In our case study, we use an administrative division of the municipalities of continental Spain corresponding to $D = 15$ Autonomous Regions. This partition is a natural choice as Autonomous Regions in Spain are responsible for developing and implementing health policies, and life style may change from region to region. By utilizing subdivisions,

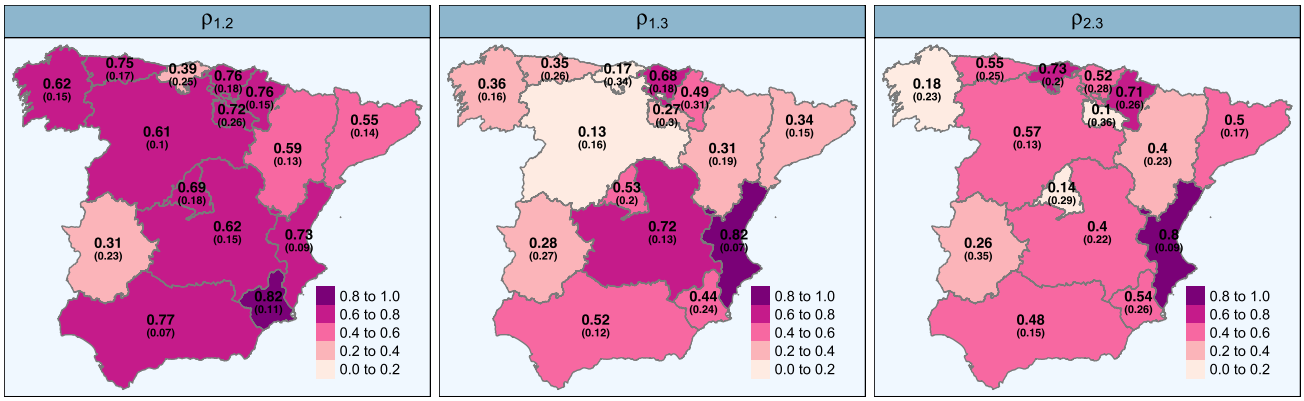

Fig. 4 Maps of posterior medians of between-disease correlations and standard deviation (in brackets) for the different subdivisions obtained with the 1st-order neighbourhood partition model. Correlations between lung and colorectal cancer are displayed on the left ($\rho_{1,2}$), the central map displays the correlations between lung and stomach cancer ($\rho_{1,3}$), and the map on the right displays the correlation between colorectal and stomach cancer ($\rho_{2,3}$)

we can obtain estimates that reveal associations between diseases which may be linked to specific policies, different lifestyles, or other geographical factors that have a local impact. This could potentially explain the observed differences in between-disease correlations across subdivisions. However, this partition may have some disadvantages. For instance, the Region of Castilla and León comprises 2245 municipalities, which is still a large number. To address this issue, we have also employed a finer partition based on 47 provinces rather than Autonomous Regions. Although the overall results are similar, the partition based on Autonomous Regions yields better recovery of the global between-disease correlations.

The M-models for multivariate disease mapping described in this paper are implemented in the `R` package `bigDM`, which also includes several scalable spatial and spatio-temporal Poisson mixed models for areal count data in a fully Bayesian setting using INLA. The package also contains a vignette to replicate the data analysis described in Sect. 5 using simulated data to preserve the confidentiality of the original data.

Acknowledgements This work has been supported by the project PID2020-113125RB-I00/MCIN/AEI/10.13039/501100011033. It has also been partially funded by the Public University of Navarra (project PJUPNA2001). We would like to thank the valuable comments made by two anonymous reviewers and the editor that have contributed to clarify some aspects of an earlier version of this paper.

Funding Open Access funding provided by Universidad Pública de Navarra.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

A Appendix

In this Appendix we briefly explain how to implement the Bartlett decomposition in R-INLA. This requires that the hyperparameters have support on $\mathbb{R}$. So, we will reparameterise the elements c_j described in Eq. (5) as

$$\theta_j = \log(c_j), \;\; j = 1, \ldots, J,$$

and the log priors for c_j are given as the corresponding log priors for $\theta_j, \forall j = 1, \ldots, J$.
For each $c_j^2, \forall j = 1, \ldots, J$, we assign a chi-square distribution with $J + 2 - j + 1$ degrees of freedom, so the log prior for θ_j is

$$\log \pi(\theta_j) = \log(2) + 2 \cdot \theta_j + \log f_j \left[\exp(2\theta_j)\right]$$

where $f_j(\cdot)$ is the probability density function (pdf) of c_j^2. This expression is obtained as follows.

$$\begin{aligned}
\theta_j &= \log(c_j) = \frac{1}{2}\log\left(c_j^2\right) = \frac{1}{2}\log(x_j) \Rightarrow x_j \\
&= g^{-1}(\theta_j) = \exp(2\theta_j) \\
\frac{dx_j}{d\theta_j} &= 2\exp(2\theta_j) \Rightarrow \left|\frac{dx_j}{d\theta_j}\right| = 2\exp(2\theta_j) \\
\pi(\theta_j) &= f_j\left[g^{-1}(\theta_j)\right]\left|\frac{dx_j}{d\theta_j}\right| = f_j\left[\exp(2\theta_j)\right]\, 2\exp(2\theta_j) \\
\log \pi(\theta_j) &= \log f_j\left[\exp(2\theta_j)\right] + \log(2) + 2 \cdot \theta_j
\end{aligned}$$

Note that non-diagonal elements in $\mathbf{A}$ (see Eq. (5)) have support on $\mathbb{R}$, so there is no need to reparameterize them, i.e.,

$$\theta_j = n_{il}, \;\; j = J + 1, \ldots, J(J+1)/2.$$

Finally, let us denote $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J, \theta_{J+1}, \ldots, \theta_{J(J+1)/2})'$. Then,

$$\pi(\boldsymbol{\theta}) = \prod_{j=1}^{J(J+1)/2} \pi(\theta_j) = \prod_{j=1}^{J} \pi(\theta_j) \times \prod_{j=J+1}^{J(J+1)/2} \pi(\theta_j),$$

and taking logarithms

$$\begin{aligned}
\log \pi(\boldsymbol{\theta}) &= \sum_{j=1}^{J} \log \pi(\theta_j) + \sum_{j=J+1}^{J(J+1)/2} \log \pi(\theta_j) \\
&= \sum_{j=1}^{J} \left\{\log(2) + 2 \cdot \theta_j + \log f_j\left[\exp(2\theta_j)\right]\right\} \\
&\quad + \sum_{j=J+1}^{J(J+1)/2} \log \phi(\theta_j) \\
&= J\log(2) + 2\sum_{j=1}^{J} \theta_j + \sum_{j=1}^{J} \log f_j\left[\exp(2\theta_j)\right] \\
&\quad + \sum_{j=J+1}^{J(J+1)/2} \log \phi(\theta_j)
\end{aligned}$$

where $f_j(\cdot)$ are the pdf of the chi-squared distribution with $J + 2 - j + 1$ degrees of freedom, $j = 1, \ldots, J$, and $\phi(\cdot)$ is the pdf of the standard Normal distribution.

Code

The R-INLA code to assign log prior distributions to the hyperparameters of the M-models (elements of the **A** matrix) can be checked in the `Mmodel_icar()` function of the `bigDM` package.

B Appendix

In this Appendix we include additional tables and figures regarding the simulation study (Sect. 4) and the results of the joint analysis of mortality data for lung, colorectal and stomach cancer (case study of Sect. 5).

Figure 5 displays the map of the administrative division of Spain into Autonomous Regions.

Tables 6, 7 and 8 compares the true values of model parameters (local correlation coefficients in each subdivision) against average values of posterior mean estimates over the 100 simulated data sets for Scenario 2.

Figure 6 displays the spatial patterns of colorectal cancer mortality risks (top) and the posterior probabilities of risk exceedance (bottom), $P(R_{ij} > 1|\mathbf{O})$, obtained with the global and the disjoint models. Similarly, Fig. 7 displays the spatial patterns of stomach cancer mortality risks (top) and the posterior probabilities of risk exceedance (bottom), $P(R_{ij} > 1|\mathbf{O})$, obtained with the global and the disjoint models.

Fig. 5 Map of the administrative division of Spain into Autonomous Regions

Table 6 Disjoint model

Parameter	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov
	Andalucía					Aragón					Asturias				
α_1	−0.20	−0.20	0.01	0.01	0.97	−0.20	−0.21	0.03	0.03	0.91	−0.20	−0.20	0.03	0.04	0.95
α_2	−0.10	−0.10	0.02	0.01	0.95	−0.10	−0.10	0.03	0.03	0.98	−0.10	−0.10	0.04	0.04	0.96
α_3	0.10	0.10	0.02	0.02	0.97	0.10	0.09	0.05	0.04	0.97	0.10	0.09	0.05	0.05	0.94
σ_1^2	0.50	0.51	0.05	0.04	0.96	0.50	0.53	0.09	0.08	0.93	0.50	0.58	0.15	0.14	0.94
σ_2^2	0.40	0.41	0.04	0.05	0.95	0.40	0.45	0.10	0.09	0.95	0.40	0.49	0.14	0.13	0.94
σ_3^2	0.30	0.32	0.05	0.05	0.93	0.30	0.39	0.12	0.12	0.87	0.30	0.41	0.15	0.11	0.94
ρ_{12}	0.76	0.75	0.04	0.04	0.94	0.58	0.52	0.12	0.11	0.94	0.71	0.64	0.12	0.11	0.95
ρ_{13}	0.52	0.52	0.07	0.08	0.94	0.30	0.24	0.17	0.20	0.90	0.34	0.28	0.20	0.19	0.96
ρ_{23}	0.47	0.46	0.08	0.08	0.96	0.37	0.29	0.19	0.18	0.94	0.51	0.44	0.19	0.18	0.97
	Castilla y León					Cataluña					Comunidad Valenciana				
α_1	−0.20	−0.20	0.02	0.02	0.96	−0.20	−0.20	0.02	0.02	0.94	−0.20	−0.20	0.02	0.02	0.96
α_2	−0.10	−0.11	0.03	0.03	0.93	−0.10	−0.10	0.02	0.02	0.96	−0.10	−0.10	0.02	0.02	0.94
α_3	0.10	0.10	0.04	0.03	0.97	0.10	0.10	0.03	0.03	0.93	0.10	0.10	0.03	0.03	0.92
σ_1^2	0.50	0.51	0.06	0.06	0.97	0.50	0.51	0.05	0.05	0.93	0.50	0.51	0.06	0.06	0.92
σ_2^2	0.40	0.42	0.06	0.07	0.92	0.40	0.42	0.05	0.05	0.95	0.40	0.42	0.06	0.06	0.97
σ_3^2	0.30	0.34	0.08	0.08	0.91	0.30	0.31	0.05	0.05	0.95	0.30	0.32	0.06	0.06	0.93
ρ_{12}	0.60	0.58	0.08	0.07	0.97	0.54	0.52	0.06	0.07	0.88	0.72	0.71	0.05	0.05	0.98
ρ_{13}	0.12	0.13	0.13	0.13	0.93	0.34	0.34	0.09	0.09	0.94	0.81	0.79	0.06	0.06	0.96
ρ_{23}	0.56	0.53	0.11	0.11	0.96	0.48	0.46	0.09	0.09	0.96	0.79	0.76	0.07	0.07	0.93
	La Rioja					Madrid					Murcia				
α_1	−0.20	−0.22	0.06	0.06	0.95	−0.20	−0.20	0.04	0.04	0.97	−0.20	−0.20	0.03	0.03	0.97
α_2	−0.10	−0.13	0.08	0.08	0.92	−0.10	−0.11	0.04	0.05	0.94	−0.10	−0.11	0.04	0.04	0.96
α_3	0.10	0.07	0.10	0.09	0.97	0.10	0.09	0.06	0.06	0.96	0.10	0.09	0.05	0.05	0.96
σ_1^2	0.50	0.70	0.22	0.16	0.90	0.50	0.55	0.10	0.09	0.96	0.50	0.64	0.18	0.17	0.91
σ_2^2	0.40	0.64	0.25	0.22	0.86	0.40	0.44	0.10	0.09	0.96	0.40	0.53	0.16	0.14	0.90
σ_3^2	0.30	0.66	0.32	0.26	0.84	0.30	0.37	0.10	0.10	0.93	0.30	0.41	0.15	0.13	0.96
ρ_{12}	0.65	0.51	0.21	0.16	0.98	0.66	0.61	0.10	0.11	0.90	0.80	0.74	0.10	0.11	0.94
ρ_{13}	0.26	0.20	0.28	0.22	0.99	0.52	0.50	0.13	0.14	0.94	0.42	0.38	0.20	0.16	0.98
ρ_{23}	0.11	0.07	0.31	0.25	1.00	0.12	0.08	0.17	0.17	0.95	0.49	0.42	0.20	0.20	0.93
	Cantabria					Castilla - La Mancha									
α_1	−0.20	−0.20	0.04	0.04	0.96	−0.20	−0.20	0.02	0.02	0.88					
α_2	−0.10	−0.11	0.05	0.05	0.93	−0.10	−0.10	0.02	0.02	0.96					
α_3	0.10	0.08	0.07	0.07	0.94	0.10	0.09	0.03	0.03	0.88					
σ_1^2	0.50	0.59	0.16	0.14	0.95	0.50	0.50	0.05	0.06	0.91					
σ_2^2	0.40	0.51	0.16	0.14	0.93	0.40	0.42	0.06	0.06	0.94					
σ_3^2	0.30	0.49	0.19	0.17	0.91	0.30	0.33	0.07	0.07	0.95					
ρ_{12}	0.37	0.29	0.19	0.18	0.92	0.60	0.58	0.07	0.08	0.93					
ρ_{13}	0.14	0.11	0.23	0.20	0.96	0.71	0.67	0.08	0.07	0.99					
ρ_{23}	0.69	0.58	0.19	0.15	0.99	0.38	0.37	0.11	0.11	0.93					
	Extremadura					Galicia									
α_1	−0.20	−0.20	0.02	0.02	0.96	−0.20	−0.20	0.01	0.02	0.95					
α_2	−0.10	−0.10	0.03	0.03	0.96	−0.10	−0.10	0.02	0.02	0.96					
α_3	0.10	0.09	0.04	0.04	0.95	0.10	0.10	0.03	0.03	0.93					
σ_1^2	0.50	0.54	0.09	0.07	0.94	0.50	0.52	0.06	0.06	0.96					
σ_2^2	0.40	0.44	0.09	0.09	0.95	0.40	0.41	0.06	0.06	0.94					

Table 6 continued

Parameter	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov
σ_3^2	0.30	0.36	0.10	0.10	0.93	0.30	0.32	0.06	0.07	0.89					
ρ_{12}	0.30	0.28	0.12	0.11	0.97	0.61	0.61	0.07	0.07	0.93					
ρ_{13}	0.27	0.25	0.16	0.15	0.98	0.36	0.34	0.11	0.11	0.93					
ρ_{23}	0.24	0.22	0.18	0.17	0.92	0.17	0.16	0.12	0.12	0.96					
	Navarra					País Vasco									
α_1	−0.20	−0.21	0.04	0.04	0.93	−0.20	−0.20	0.03	0.03	0.94					
α_2	−0.10	−0.11	0.05	0.05	0.98	−0.10	−0.10	0.03	0.03	0.94					
α_3	0.10	0.07	0.07	0.06	0.94	0.10	0.09	0.04	0.04	0.98					
σ_1^2	0.50	0.57	0.13	0.14	0.89	0.50	0.53	0.09	0.09	0.94					
σ_2^2	0.40	0.47	0.13	0.10	0.98	0.40	0.44	0.09	0.09	0.93					
σ_3^2	0.30	0.47	0.17	0.15	0.90	0.30	0.36	0.09	0.10	0.92					
ρ_{12}	0.73	0.67	0.12	0.10	0.96	0.73	0.69	0.08	0.08	0.92					
ρ_{13}	0.44	0.37	0.20	0.16	0.98	0.65	0.61	0.11	0.11	0.96					
ρ_{23}	0.65	0.52	0.19	0.19	0.94	0.47	0.42	0.15	0.13	0.94					

Average values of posterior mean, posterior standard deviation (SD), simulated standard errors (sim) and empirical coverage of the 95% credible intervals (Cov) for local estimates model parameters based on 100 simulated data sets for Scenario 2

Table 7 1st-order neighbourhood model

Parameter	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov
	Andalucía					Aragón					Asturias				
α_1	−0.20	−0.20	0.01	0.02	0.86	−0.20	−0.20	0.03	0.04	0.79	−0.20	−0.19	0.03	0.07	0.67
α_2	−0.10	−0.09	0.02	0.02	0.90	−0.10	−0.10	0.03	0.04	0.91	−0.10	−0.09	0.04	0.07	0.70
α_3	0.10	0.10	0.02	0.02	0.93	0.10	0.10	0.04	0.04	0.97	0.10	0.10	0.05	0.06	0.93
σ_1^2	0.50	0.57	0.05	0.05	0.59	0.50	0.74	0.10	0.12	0.28	0.50	0.84	0.18	0.26	0.47
σ_2^2	0.40	0.46	0.05	0.06	0.78	0.40	0.60	0.10	0.12	0.49	0.40	0.67	0.17	0.20	0.58
σ_3^2	0.30	0.35	0.05	0.06	0.82	0.30	0.46	0.12	0.14	0.65	0.30	0.52	0.17	0.16	0.80
ρ_{12}	0.76	0.75	0.04	0.04	0.93	0.58	0.59	0.09	0.10	0.89	0.71	0.65	0.11	0.12	0.93
ρ_{13}	0.52	0.52	0.07	0.08	0.93	0.30	0.36	0.14	0.18	0.84	0.34	0.35	0.18	0.18	0.93
ρ_{23}	0.47	0.48	0.08	0.08	0.91	0.37	0.42	0.15	0.15	0.92	0.51	0.45	0.17	0.18	0.93
	Castilla y León					Cataluña					Comunidad Valenciana				
α_1	−0.20	−0.20	0.02	0.03	0.88	−0.20	−0.20	0.02	0.02	0.93	−0.20	−0.19	0.02	0.02	0.86
α_2	−0.10	−0.11	0.03	0.03	0.91	−0.10	−0.10	0.02	0.02	0.94	−0.10	−0.10	0.02	0.03	0.81
α_3	0.10	0.10	0.03	0.03	0.96	0.10	0.10	0.03	0.03	0.94	0.10	0.10	0.03	0.03	0.96
σ_1^2	0.50	0.77	0.07	0.11	0.06	0.50	0.56	0.05	0.06	0.78	0.50	0.63	0.07	0.10	0.48
σ_2^2	0.40	0.63	0.07	0.10	0.12	0.40	0.45	0.05	0.06	0.81	0.40	0.52	0.06	0.08	0.53
σ_3^2	0.30	0.48	0.09	0.13	0.46	0.30	0.33	0.05	0.05	0.89	0.30	0.39	0.06	0.08	0.72
ρ_{12}	0.60	0.63	0.06	0.07	0.83	0.54	0.54	0.06	0.07	0.88	0.72	0.71	0.05	0.06	0.88
ρ_{13}	0.12	0.30	0.10	0.13	0.47	0.34	0.34	0.09	0.10	0.93	0.81	0.74	0.06	0.08	0.73
ρ_{23}	0.56	0.53	0.09	0.12	0.90	0.48	0.47	0.09	0.08	0.97	0.79	0.75	0.06	0.06	0.92
	La Rioja					Madrid					Murcia				
α_1	−0.20	−0.23	0.05	0.08	0.79	−0.20	−0.20	0.03	0.06	0.70	−0.20	−0.22	0.03	0.09	0.53
α_2	−0.10	−0.13	0.06	0.10	0.76	−0.10	−0.11	0.04	0.05	0.87	−0.10	−0.12	0.03	0.08	0.58

Table 7 continued

Parameter	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov
α_3	0.10	0.08	0.08	0.10	0.84	0.10	0.09	0.05	0.06	0.93	0.10	0.09	0.05	0.08	0.71
σ_1^2	0.50	1.12	0.24	0.29	0.19	0.50	0.87	0.13	0.19	0.16	0.50	1.06	0.23	0.25	0.14
σ_2^2	0.40	0.90	0.24	0.33	0.36	0.40	0.67	0.12	0.16	0.37	0.40	0.86	0.20	0.26	0.29
σ_3^2	0.30	0.81	0.29	0.33	0.48	0.30	0.51	0.12	0.16	0.53	0.30	0.65	0.19	0.23	0.45
ρ_{12}	0.65	0.61	0.13	0.16	0.92	0.66	0.66	0.08	0.11	0.88	0.80	0.73	0.08	0.11	0.86
ρ_{13}	0.26	0.39	0.19	0.20	0.81	0.52	0.49	0.12	0.15	0.90	0.42	0.45	0.15	0.18	0.85
ρ_{23}	0.11	0.27	0.21	0.26	0.83	0.12	0.26	0.14	0.18	0.82	0.49	0.54	0.14	0.16	0.90
	Cantabria					Castilla - La Mancha									
α_1	−0.20	−0.19	0.04	0.07	0.78	−0.20	−0.21	0.02	0.02	0.87					
α_2	−0.10	−0.08	0.05	0.07	0.83	−0.10	−0.11	0.02	0.02	0.93					
α_3	0.10	0.09	0.06	0.07	0.90	0.10	0.09	0.03	0.03	0.90					
σ_1^2	0.50	0.90	0.20	0.32	0.46	0.50	0.70	0.06	0.09	0.13					
σ_2^2	0.40	0.75	0.20	0.26	0.51	0.40	0.55	0.06	0.09	0.29					
σ_3^2	0.30	0.62	0.21	0.25	0.65	0.30	0.42	0.07	0.09	0.52					
ρ_{12}	0.37	0.43	0.15	0.18	0.87	0.60	0.60	0.06	0.08	0.86					
ρ_{13}	0.14	0.27	0.20	0.21	0.83	0.71	0.62	0.08	0.09	0.74					
ρ_{23}	0.69	0.62	0.16	0.15	0.97	0.38	0.39	0.10	0.13	0.84					
	Extremadura					Galicia									
α_1	−0.20	−0.19	0.02	0.04	0.77	−0.20	−0.20	0.02	0.03	0.75					
α_2	−0.10	−0.09	0.03	0.04	0.78	−0.10	−0.10	0.02	0.03	0.86					
α_3	0.10	0.09	0.04	0.05	0.92	0.10	0.10	0.03	0.03	0.91					
σ_1^2	0.50	0.71	0.10	0.11	0.41	0.50	0.58	0.07	0.08	0.81					
σ_2^2	0.40	0.57	0.10	0.11	0.51	0.40	0.45	0.06	0.07	0.84					
σ_3^2	0.30	0.41	0.11	0.11	0.83	0.30	0.33	0.07	0.08	0.87					
ρ_{12}	0.30	0.40	0.10	0.11	0.80	0.61	0.63	0.06	0.07	0.89					
ρ_{13}	0.27	0.32	0.14	0.15	0.93	0.36	0.36	0.10	0.11	0.94					
ρ_{23}	0.24	0.29	0.16	0.17	0.90	0.17	0.19	0.12	0.12	0.93					
	Navarra					País Vasco									
α_1	−0.20	−0.21	0.04	0.05	0.84	−0.20	−0.20	0.03	0.04	0.84					
α_2	−0.10	−0.11	0.04	0.06	0.83	−0.10	−0.10	0.03	0.04	0.88					
α_3	0.10	0.08	0.06	0.06	0.93	0.10	0.10	0.04	0.05	0.89					
σ_1^2	0.50	0.85	0.15	0.20	0.29	0.50	0.83	0.12	0.22	0.27					
σ_2^2	0.40	0.64	0.14	0.17	0.52	0.40	0.64	0.11	0.17	0.47					
σ_3^2	0.30	0.56	0.17	0.19	0.62	0.30	0.50	0.12	0.15	0.56					
ρ_{12}	0.73	0.71	0.09	0.10	0.90	0.73	0.71	0.07	0.10	0.83					
ρ_{13}	0.44	0.46	0.15	0.17	0.89	0.65	0.63	0.10	0.13	0.88					
ρ_{23}	0.65	0.50	0.15	0.19	0.81	0.47	0.47	0.13	0.15	0.90					

Average values of posterior mean, posterior standard deviation (SD), simulated standard errors (sim) and empirical coverage of the 95% credible intervals (Cov) for local estimates model parameters based on 100 simulated data sets for Scenario 2

Table 8 2nd-order neighbourhood model

Parameter	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov
	Andalucía					Aragón					Asturias				
α_1	−0.20	−0.20	0.01	0.03	0.63	−0.20	−0.19	0.02	0.06	0.52	−0.20	−0.19	0.03	0.10	0.52
α_2	−0.10	−0.09	0.01	0.03	0.71	−0.10	−0.09	0.03	0.05	0.68	−0.10	−0.09	0.04	0.09	0.58
α_3	0.10	0.10	0.02	0.03	0.79	0.10	0.10	0.04	0.04	0.88	0.10	0.11	0.05	0.09	0.70
σ_1^2	0.50	0.57	0.05	0.05	0.63	0.50	0.72	0.08	0.11	0.24	0.50	0.80	0.16	0.23	0.49
σ_2^2	0.40	0.46	0.05	0.06	0.78	0.40	0.58	0.09	0.11	0.43	0.40	0.64	0.15	0.18	0.59
σ_3^2	0.30	0.35	0.05	0.06	0.81	0.30	0.45	0.10	0.13	0.60	0.30	0.50	0.15	0.14	0.75
ρ_{12}	0.76	0.74	0.04	0.04	0.94	0.58	0.60	0.08	0.09	0.88	0.71	0.63	0.10	0.12	0.91
ρ_{13}	0.52	0.49	0.07	0.07	0.91	0.30	0.37	0.12	0.15	0.83	0.34	0.35	0.16	0.17	0.92
ρ_{23}	0.47	0.47	0.07	0.08	0.90	0.37	0.44	0.13	0.13	0.86	0.51	0.45	0.16	0.17	0.92
	Castilla y León					Cataluña					Comunidad Valenciana				
α_1	−0.20	−0.20	0.02	0.03	0.75	−0.20	−0.20	0.02	0.02	0.83	−0.20	−0.19	0.02	0.04	0.71
α_2	−0.10	−0.10	0.02	0.03	0.78	−0.10	−0.10	0.02	0.02	0.88	−0.10	−0.09	0.02	0.04	0.70
α_3	0.10	0.10	0.03	0.04	0.88	0.10	0.10	0.03	0.03	0.91	0.10	0.10	0.03	0.03	0.86
σ_1^2	0.50	0.74	0.06	0.09	0.03	0.50	0.56	0.05	0.06	0.76	0.50	0.63	0.06	0.10	0.47
σ_2^2	0.40	0.60	0.06	0.09	0.13	0.40	0.45	0.05	0.06	0.81	0.40	0.52	0.06	0.08	0.49
σ_3^2	0.30	0.46	0.07	0.11	0.40	0.30	0.34	0.05	0.05	0.88	0.30	0.39	0.06	0.08	0.67
ρ_{12}	0.60	0.63	0.05	0.07	0.79	0.54	0.54	0.06	0.07	0.90	0.72	0.71	0.05	0.06	0.87
ρ_{13}	0.12	0.35	0.08	0.12	0.29	0.34	0.35	0.09	0.10	0.91	0.81	0.69	0.06	0.08	0.48
ρ_{23}	0.56	0.50	0.08	0.11	0.85	0.48	0.47	0.08	0.08	0.97	0.79	0.72	0.06	0.07	0.81
	La Rioja					Madrid					Murcia				
α_1	−0.20	−0.23	0.04	0.10	0.60	−0.20	−0.21	0.03	0.09	0.53	−0.20	−0.24	0.02	0.17	0.21
α_2	−0.10	−0.12	0.05	0.11	0.55	−0.10	−0.11	0.04	0.07	0.72	−0.10	−0.12	0.03	0.15	0.34
α_3	0.10	0.07	0.07	0.11	0.75	0.10	0.09	0.05	0.07	0.79	0.10	0.08	0.04	0.13	0.39
σ_1^2	0.50	1.11	0.20	0.27	0.08	0.50	0.81	0.11	0.15	0.17	0.50	0.85	0.15	0.18	0.25
σ_2^2	0.40	0.89	0.20	0.27	0.26	0.40	0.64	0.10	0.14	0.36	0.40	0.70	0.14	0.18	0.36
σ_3^2	0.30	0.82	0.25	0.33	0.28	0.30	0.50	0.11	0.14	0.53	0.30	0.55	0.14	0.18	0.49
ρ_{12}	0.65	0.62	0.11	0.15	0.86	0.66	0.63	0.07	0.10	0.88	0.80	0.73	0.07	0.10	0.78
ρ_{13}	0.26	0.42	0.16	0.19	0.78	0.52	0.45	0.11	0.14	0.81	0.42	0.49	0.12	0.14	0.85
ρ_{23}	0.11	0.32	0.18	0.22	0.75	0.12	0.28	0.13	0.16	0.71	0.49	0.56	0.12	0.14	0.85
	Cantabria					Castilla - La Mancha									
α_1	−0.20	−0.19	0.04	0.09	0.60	−0.20	−0.20	0.02	0.02	0.80					
α_2	−0.10	−0.08	0.04	0.09	0.64	−0.10	−0.11	0.02	0.02	0.89					
α_3	0.10	0.10	0.06	0.09	0.79	0.10	0.10	0.02	0.03	0.93					
σ_1^2	0.50	0.86	0.17	0.27	0.48	0.50	0.71	0.05	0.08	0.06					
σ_2^2	0.40	0.72	0.17	0.23	0.40	0.40	0.57	0.06	0.09	0.21					
σ_3^2	0.30	0.57	0.18	0.19	0.65	0.30	0.44	0.07	0.08	0.36					
ρ_{12}	0.37	0.48	0.13	0.16	0.80	0.60	0.59	0.05	0.07	0.86					
ρ_{13}	0.14	0.34	0.17	0.19	0.75	0.71	0.57	0.07	0.09	0.48					
ρ_{23}	0.69	0.59	0.15	0.15	0.94	0.38	0.40	0.09	0.13	0.71					
	Extremadura					Galicia									
α_1	−0.20	−0.19	0.02	0.07	0.39	−0.20	−0.20	0.02	0.05	0.46					
α_2	−0.10	−0.09	0.03	0.07	0.54	−0.10	−0.10	0.02	0.05	0.55					
α_3	0.10	0.10	0.03	0.06	0.74	0.10	0.09	0.03	0.05	0.74					
σ_1^2	0.50	0.70	0.09	0.11	0.33	0.50	0.58	0.07	0.08	0.76					

Table 8 continued

Parameter	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov	Value	Mean	SD	Sim	Cov
σ_2^2	0.40	0.57	0.09	0.11	0.49	0.40	0.46	0.06	0.07	0.82					
σ_3^2	0.30	0.42	0.10	0.10	0.81	0.30	0.34	0.06	0.08	0.91					
ρ_{12}	0.30	0.43	0.09	0.11	0.66	0.61	0.63	0.06	0.07	0.88					
ρ_{13}	0.27	0.33	0.13	0.15	0.89	0.36	0.37	0.10	0.11	0.92					
ρ_{23}	0.24	0.33	0.14	0.16	0.85	0.17	0.20	0.11	0.12	0.90					
	Navarra					País Vasco									
α_1	−0.20	−0.21	0.03	0.07	0.64	−0.20	−0.20	0.03	0.06	0.64					
α_2	−0.10	−0.11	0.04	0.07	0.72	−0.10	−0.11	0.03	0.06	0.65					
α_3	0.10	0.09	0.05	0.07	0.80	0.10	0.10	0.04	0.06	0.78					
σ_1^2	0.50	0.80	0.12	0.16	0.21	0.50	0.82	0.11	0.20	0.25					
σ_2^2	0.40	0.60	0.11	0.14	0.51	0.40	0.64	0.11	0.17	0.40					
σ_3^2	0.30	0.52	0.13	0.15	0.56	0.30	0.50	0.11	0.15	0.54					
ρ_{12}	0.73	0.71	0.08	0.09	0.90	0.73	0.70	0.07	0.10	0.86					
ρ_{13}	0.44	0.48	0.13	0.15	0.88	0.65	0.61	0.10	0.13	0.87					
ρ_{23}	0.65	0.46	0.14	0.17	0.72	0.47	0.46	0.12	0.15	0.89					

Average values of posterior mean, posterior standard deviation (SD), simulated standard errors (sim) and empirical coverage of the 95% credible intervals (Cov) for local estimates model parameters based on 100 simulated data sets for Scenario 2

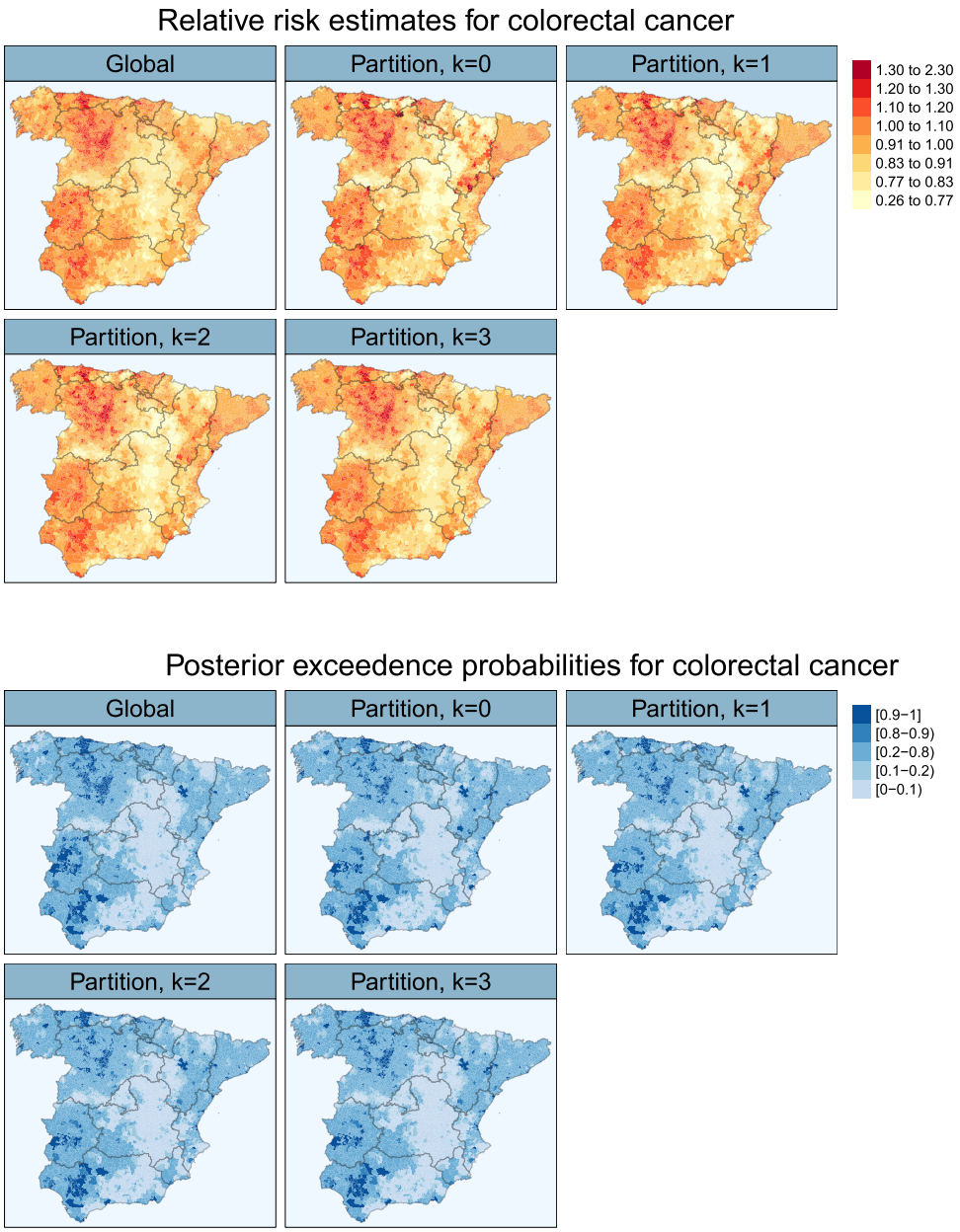

Fig. 6 Maps of posterior median estimates of mortality relative risk for colorectal cancer (top) and posterior exceedance probabilities $P(R_{ij} > 1|\mathbf{O})$ (bottom) in continental Spain

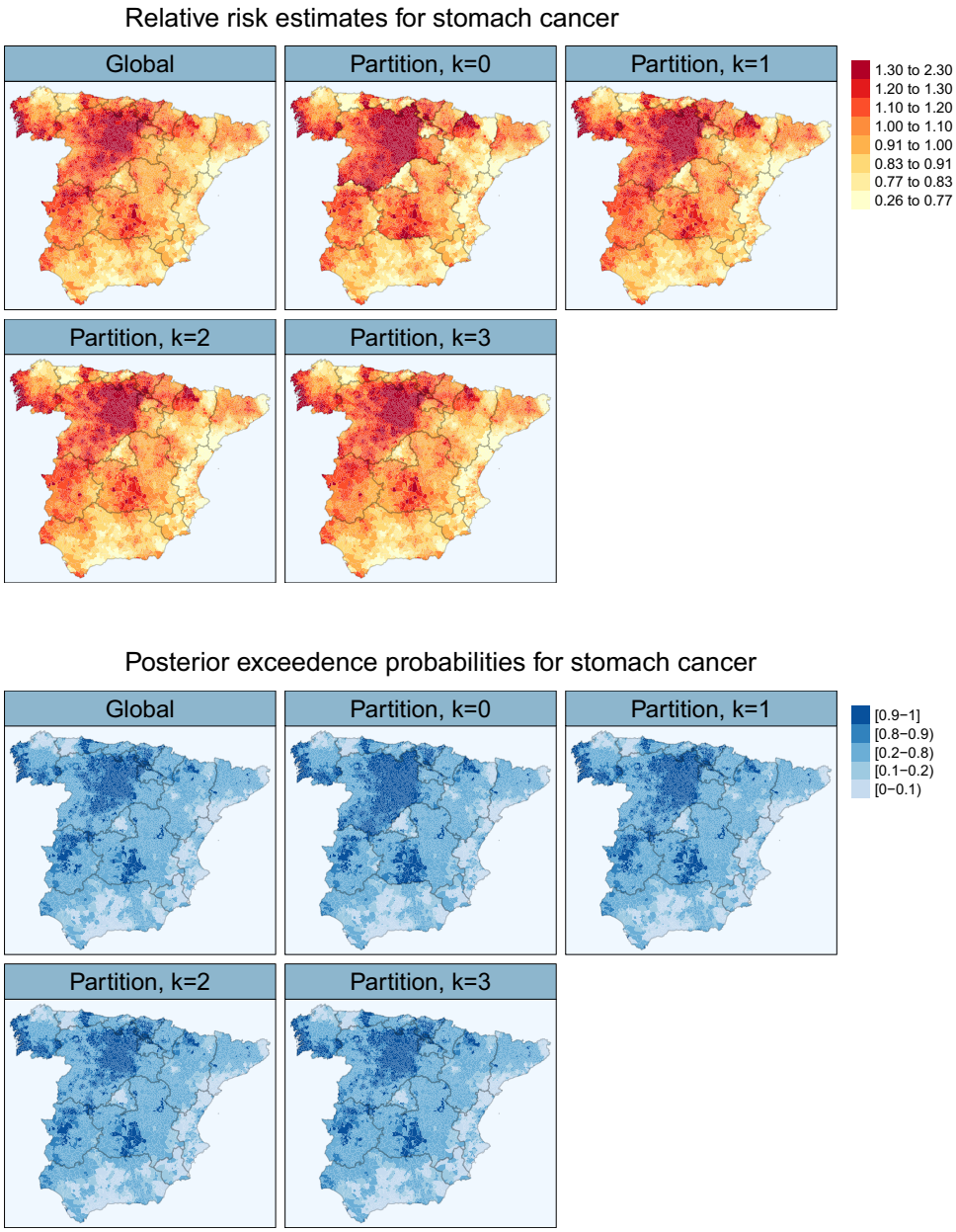

Fig. 7 Maps of posterior median estimates of mortality relative risk for stomach cancer (top) and posterior exceedance probabilities $P(R_{ij} > 1|\mathbf{O})$ (bottom) in continental Spain

References

Adin, A., Orozco-Acosta, E., Ugarte, M.D.: bigDM: Scalable Bayesian Disease Mapping Models for High-Dimensional Data. R package version 0.5.1 (2023)

Besag, J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. Ser. B (Stat. Methodol.) **36**(2), 192–225 (1974)

Besag, J., York, J., Mollié, A.: A Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. **43**(1), 1–21 (1991)

Botella-Rocamora, P., Martinez-Beneito, M.A., Banerjee, S.: A unifying modeling framework for highly multivariate disease mapping. Stat. Med. **34**(9), 1548–1559 (2015)

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., Dorie, V.: Weakly informative prior for point estimation of covariance matrices in hierarchical models. J. Educ. Behav. Stat. **40**(2), 136–157 (2015)

Corpas-Burgos, F., Botella-Rocamora, P., Martinez-Beneito, M.A.: On the convenience of heteroscedasticity in highly multivariate disease mapping. TEST **28**(4), 1229–1250 (2019)

Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **70**(1), 209–226 (2008)

Dean, C.B., Ugarte, M.D., Militino, A.F.: Detecting interaction between random region and fixed age effects in disease mapping. Biometrics **57**(1), 197–202 (2001)

Eberly, L.E., Carlin, B.P.: Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Stat. Med. **19**(17–18), 2279–2294 (2000)

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, Berlin (2006)

Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Stat. Comput. **24**(6), 997–1016 (2014)

Goicoa, T., Ugarte, M., Etxeberria, J., Militino, A.: Comparing CAR and P-spline models in spatial disease mapping. Environ. Ecol. Stat. **19**(4), 573–599 (2012)
Goicoa, T., Adin, A., Ugarte, M.D., Hodges, J.S.: In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. Stoch. Env. Res. Risk Assess. **32**(3), 749–770 (2018)
Held, L., Natário, I., Fenton, S.E., Rue, H., Becker, N.: Towards joint disease mapping. Stat. Methods Med. Res. **14**(1), 61–82 (2005)
Jin, X., Banerjee, S., Carlin, B.: Order-free co-regionalized areal data models with application to multiple-disease mapping. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **69**(5), 817–838 (2007)
Katzfuss, M.: A multi-resolution approximation for massive spatial datasets. J. Am. Stat. Assoc. **112**(517), 201–214 (2017)
Katzfuss, M., Guinness, J.: A general framework for Vecchia approximations of Gaussian processes. Stat. Sci. **36**(1), 124–141 (2021)
Knorr-Held, L., Best, N.G.: A shared component model for detecting joint and selective clustering of two diseases. J. R. Stat. Soc. A. Stat. Soc. **164**(1), 73–85 (2001)
Leroux, B.G., Lei, X., Breslow, N.: 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, pp. 179–192 (1999)
Li, G., Haining, R., Richardson, S., Best, N.: Space-time variability in burglary risk: a Bayesian spatio-temporal modelling approach. Spat. Stat. **9**, 180–191 (2014)
Lindgren, F., Rue, H.: Bayesian spatial modelling with R-INLA. J. Stat. Softw. **63**, 1–25 (2015)
Lindgren, F., Rue, H., Lindström, J.: 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 (2011)
Lindsay, B.G.: Mixture models: theory, geometry, and applications. In: NSF-CBMS Regional Conference Series in Probability and Statistics, JSTOR (1995)
MacNab, Y.C.: On Bayesian shared component disease mapping and ecological regression with errors in covariates. Stat. Med. **29**(11), 1239–1249 (2010)
MacNab, Y.C.: Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate CAR models. Stat. Med. **35**(21), 3827–3850 (2016)
MacNab, Y.C.: Some recent work on multivariate Gaussian Markov random fields. TEST **27**(3), 497–541 (2018)
MacNab, Y.C.: Bayesian disease mapping: past, present, and future. Spat. Stat. **50**, 100593 (2022)
Mardia, K.: Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. J. Multivar. Anal. **24**(2), 265–284 (1988)
Martinez-Beneito, M.A.: A general modelling framework for multivariate disease mapping. Biometrika **100**(3), 539–553 (2013)
Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S.: A multiresolution Gaussian process model for the analysis of large spatial datasets. J. Comput. Graph. Stat. **24**(2), 579–599 (2015)
Orozco-Acosta, E., Adin, A., Ugarte, M.D.: Scalable Bayesian modelling for smoothing disease risks in large spatial data sets using INLA. Spat. Stat. **41**, 100496 (2021)
Orozco-Acosta, E., Adin, A., Ugarte, M.D.: Big problems in spatio-temporal disease mapping: methods and software. Comput. Methods Programs Biomed. **231**, 107403 (2023)
Peña, V., Irie, K.: On the relationship between Uhlig extended and beta-Bartlett processes. J. Time Ser. Anal. **43**(1), 147–153 (2022)
Pettit, L.: The conditional predictive ordinate for the normal distribution. J. R. Stat. Soc. Ser. B (Methodol.) **52**(1), 175–184 (1990)
Riebler, A., Sørbye, S.H., Simpson, D., Rue, H.: An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Stat. Methods Med. Res. **25**(4), 1145–1165 (2016)
Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B (Methodol.) **71**(2), 319–392 (2009)
Sain, S.R., Furrer, R., Cressie, N.: A spatial analysis of multivariate output from regional climate models. Ann. Appl. Stat. **5**(1), 150–175 (2011)
Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., McCulloch, R.E.: Bayes and big data: the consensus Monte Carlo algorithm. Int. J. Manag. Sci. Eng. Manag. **11**(2), 78–88 (2016)
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A.: Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B (Methodol.) **64**(4), 583–639 (2002)
Ugarte, M.D., Goicoa, T., Militino, A.F.: Spatio-temporal modeling of mortality risks using penalized splines. Environmetrics **21**(3–4), 270–289 (2010)
Ugarte, M.D., Adin, A., Goicoa, T.: One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: Model fitting and model identifiability. Spat. Stat. **22**, 451–468 (2017)
Van Niekerk, J., Rue, H.: Correcting the Laplace Method with Variational Bayes. (2021) arXiv preprint arXiv:2111.12945
Van Niekerk, J., Krainski, E., Rustand, D., Rue, H. (2023). A new avenue for Bayesian inference with INLA. Comput. Stat. Data Anal. p. 107692
Vicente, G., Goicoa, T., Puranik, A., Ugarte, M.D.: Small area estimation of gender-based violence: rape incidence risks in Uttar Pradesh, India. Stat. Appl. **16**(1), 71–90 (2018)
Vicente, G., Goicoa, T., Fernández-Rasines, P., Ugarte, M.D.: Crime against women in India: unveiling spatial patterns and temporal trends of dowry deaths in the districts of Uttar Pradesh. J. R. Stat. Soc. A. Stat. Soc. **183**(2), 655–679 (2020a)
Vicente, G., Goicoa, T., Ugarte, M.D.: Bayesian inference in multivariate spatio-temporal areal models using INLA: analysis of gender-based violence in small areas. Stoch. Environ. Res. Risk Assess. **34**(10), 1421–1440 (2020b)
Vicente, G., Goicoa, T., Ugarte, M.D.: Multivariate Bayesian spatio-temporal P-spline models to analyze crimes against women. Biostatistics (in press) (2021). https://doi.org/10.1093/biostatistics/kxab042
Wang, X., Dunson, D.B.: Parallelizing MCMC via Weierstrass sampler (2013). arXiv preprint arXiv:1312.4605
Watanabe, S.: Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. **11**, 3571–3594 (2010)