Statistical Methods in Medical Research
XX(X):2–25

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Dealing with risk discontinuities to estimate cancer mortality risks when the number of small areas is large

G Santafé[1,2], A Adin[1,2], D Lee[3] and MD Ugarte[1,2]

Abstract

Many statistical models have been developed during the last years to smooth risks in disease mapping. However, most of these modelling approaches do not take possible local discontinuities into consideration or if they do, they are computationally prohibitive or simply do not work when the number of small areas is large. In this paper we propose a two-step method to deal with discontinuities and to smooth noisy risks in small areas. In a first stage, a novel density-based clustering algorithm is used. In contrast to previous proposals, this algorithm is able to automatically detect the number of spatial clusters, thus providing a single cluster structure. In the second stage, a Bayesian hierarchical spatial model that takes the cluster configuration into account is fitted, which accounts for the discontinuities in disease risk. To evaluate the performance of this new procedure in comparison to previous proposals, a simulation study has been conducted. Results show competitive risk estimates at a much better computational cost. The new methodology is used to analyse stomach cancer mortality data in Spanish municipalities.

Keywords

Clustering, Disease Mapping, INLA, Small areas, Smoothing, Spanish municipalities, Stomach cancer

Introduction

Disease mapping is the field of spatial epidemiology[1] that studies the spatial and/or spatio-temporal distribution of disease incidence or mortality rates. It helps to design and evaluate the effect of public health policies and to detect high/low-risk areas or hot spots revealing health inequalities. Then, as stated by Waller and Carlin[2], the goal of disease mapping is somehow two-fold: on the one hand, being statistically precise when estimating local disease risk for each area and, on the other hand, detecting high/low-risk areas maintaining geographical resolution using small administrative units. To deal with these goals, many statistical models have been developed during the last years. Recent books covering statistical methods for disease mapping are Lawson[3] and Martinez-Beneito and Botella-Rocamora.[4] The models usually include spatial random effects with conditional autoregressive (CAR) priors[5,6,7] smoothing the risk locally by borrowing information from nearby regions to remove random noise. However, an excess of smoothing may prevent the identification of high-risk areas or hot spots as discontinuities in the smooth risk become blurred.

Since these two goals (a precise estimation of the risk and the identification of high/low-risk areas) are somehow contradictory, clustering methods to detect high/low-risk regions have been developed in parallel to smoothing methods.[8] Recent proposals[9,10] have considered a trade-off between both goals by first obtaining a clustering partition of the areal units, and then estimating spatial risks using Bayesian hierarchical models that take the cluster configuration structures into account. These methods adapt standard hierarchical clustering algorithms to the spatial (or spatio-temporal) areal data context, by obtaining a set of n spatially contiguous cluster structures with differing numbers of clusters. Each one of these cluster structures is evaluated by using it as a basis for a Bayesian hierarchical model and finally, the best model (in terms of certain statistical selection criteria) out of the *n* candidates is chosen to estimate the risk surface. Thus, the computational cost of this approach when dealing with a large number of small areas could be huge or even unfeasible. However, as said before, maintaining a high geographical resolution using small administrative units is important in disease

[1] Department of Statistics, Computer Science, and Mathematics, Public University of Navarre, Spain
[2] Institute for Advanced Materials (InaMat), Public University of Navarre, Spain
[3] School of Mathematics and Statistics, University of Glasgow, United Kingdom

Corresponding author:
María Dolores Ugarte, Department of Statistics, Computer Science, and Mathematics, Public University of Navarre, Campus de Arrosadia, 31006 Pamplona, Spain.
Email: lola@unavarra.es

Prepared using sagej.cls

mapping. Anderson et al.[9] and Adin et al.[10] apply the proposed methods to problems with a low/moderate number of areal units. For instance, Adin et al.[10] analyse cancer mortality in 47 Spanish provinces, and this low resolution lacks the fine spatial detail of disease risk and may mask the locations of possible hot spots. However, a higher geographical resolution such as municipalities would rocket the number of areal units and could make the method computationally unfeasible. Indeed, it is well-known that agglomerative hierarchical clustering methods are computationally very expensive for large clustering problems.[11]

In this paper, we propose a new two-step method to deal with discontinuities and to smooth noisy risks in small areas. In the first stage, a novel density-based clustering algorithm is proposed. In contrast to previous proposals, this algorithm is able to provide a single cluster configuration. In the second stage, a Bayesian hierarchical spatial model that takes the cluster configuration into account is fitted. This approach makes the algorithm suitable for analysing large spatial data sets, because only a single disease mapping model is fitted based on this single cluster structure. The performance of the new method is evaluated in a simulation study, where it is compared to recent proposals. Model fitting and inference is carried out using integrated nested Laplace approximations and numerical integration.[12] The new methodology will be used to analyse stomach cancer mortality in Spanish municipalities.

The rest of this paper is laid out as follows: Section 2 presents the motivating application. Section 3 introduces the density-based clustering algorithm proposed in this paper and describes the Bayesian hierarchical models that will be used after the cluster configuration is obtained. In Section 4 a simulation study based on a real data scenario is conducted to compare the new method with previous proposals. Section 5 uses the new algorithm to analyse stomach cancer data in Spanish municipalities. Finally, Section 6 ends with some conclusions.

Motivating application: stomach cancer mortality in Spanish municipalities

Cancer mortality data is an essential tool in both planning and assessing the impact of cancer control programmes at different levels. When this control needs to be done at a regional level, it is necessary to have precise information at that level. This is particularly important in a country like Spain where the health competence had been transferred to each of the autonomous communities. The administrative division of Spain into provinces and autonomous regions is shown in Figure 1.

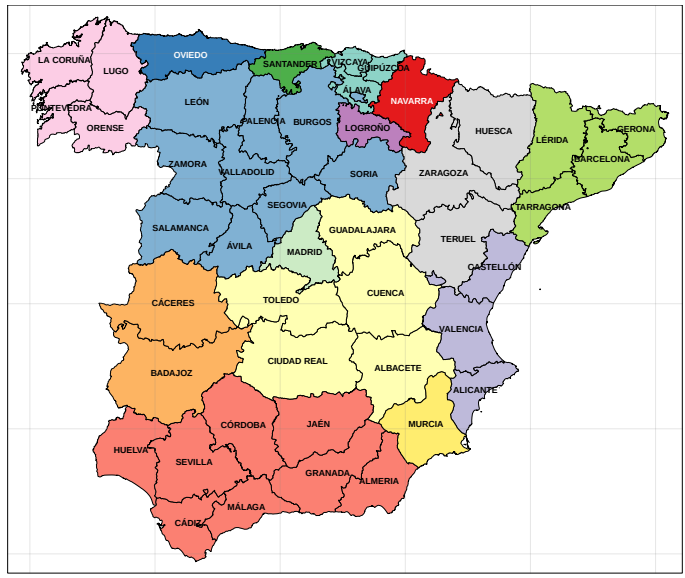

Figure 1. Administrative division of continental Spain into provinces. Provinces belonging to the same Autonomous Region share the same color.

In this work we plan to analyse the geographical pattern of stomach cancer mortality in Spain at municipality level in a recent period (2011-2015). However, the methodology developed here could be applied to any cancer sites or other diseases.

Stomach cancer mortality corresponds to code C16 of the 10th edition of the International Classification of Diseases.[13] According to the last cancer mortality estimates for 40 European countries in 2018,[14] stomach cancer is the fifth highest cause of death from cancer in males (61,880 deaths, 5.7% of the total cancer deaths in males), and the sixth cause in females (40,290 deaths, 4.7% of the total cancer deaths in females). In Spain, several studies have revealed a clear pattern in the geographical distribution of stomach cancer mortality, characterized by its persistence in time for both genders, which might be associated with dietary habits or territory-related environmental exposures.[15,16] Recently, Adin et al.[10] simultaneously smooth stomach cancer mortality risks and detect high/low risk clusters but at province level. Mortality data from cancer and other causes (60 in total) from 1975 by sex and province is available in the Interactive System of Epidemiological Information (ARIADNA, http://ariadna.cne.isciii.es/evindex.html) of the Spanish National Center for Epidemiology.

Table 1. Number of deaths and ASRs (per 100,000 inhabitants) by sex and five-year period for stomach cancer in Spain.

Period	Both sexes		Males		Females	
	Cases	ASR	Cases	ASR	Cases	ASR
1991-1995	32,051	23.65	19,197	33.77	12,854	16.34
1996-2000	30,056	19.57	18,398	28.62	11,658	12.97
2001-2005	28,141	16.29	17,340	23.95	10,801	10.64
2006-2010	27,517	14.30	16,955	21.03	10,562	9.28
2011-2015	26,713	12.59	16,215	18.23	10,498	8.37

Our objective in this paper is to analyse the geographical variation in stomach cancer mortality more thoroughly using municipality-level mortality data. As noted before, the modelling approach described in Anderson et al.[9] and Adin et al.[10] requires to fit as many hierarchical models as small areas, which is computationally prohibitive when estimating risks in Spanish municipalities. Then, a new modelling approach is needed to estimate risks in the presence of local discontinuities when dealing with a large number of small areas.

Data on population and deaths for stomach cancer in $n = 7,907$ municipalities in continental Spain during the period 1991-2015 were obtained from records of the Spanish Statistical Institute. A total of 88,105 and 56,373 stomach cancer deaths were registered in this period in males and females, respectively. The number of deaths and age-standardized rates (ASRs) using the revised European Standard Population[17] by sex and five-year period are shown in Table 1. Clearly, a decline in mortality is observed during the period 1991-2015 in both sexes. However, previous studies show that the geographical pattern in Spanish municipalities changed very little over time.[16] In Section 5, mortality risks for the period 2010-2015 are analysed using data for the preceding four five-years periods as ancillary data for the clustering algorithm.

Cluster detection and mortality risk estimation

Suppose that the region under study is divided into n small areas labeled as $\{A_1, \ldots, A_n\}$. Let $\mathbf{O}^t = (O_1^t, \ldots, O_n^t)'$ and $\mathbf{E}^t = (E_1^t, \ldots, E_n^t)'$ denote the number of observed and expected disease cases in each area for a time period t, respectively. The aim of the spatial clustering algorithm is to obtain a partition of the n areal units into spatially contiguous groups, where areas within the same group present similar disease risk. In order to obtain this clustering partition, each area A_i is described by the vector of log-standardized mortality ratios (SMR) for q consecutive time periods,

$\mathbf{\Psi}_i = [\log(O_i^T/E_i^T), \log(O_i^{T-1}/E_i^{T-1}), \ldots, \log(O_i^{T-q}/E_i^{T-q})]$, where T is the time period under study and q is the number of previous time periods used as ancillary data.

In this paper, we propose a new spatial clustering algorithm based on the density clustering algorithm introduced by Rodriguez and Laio[18] and the posterior modification presented by Wang and Song.[19] The proposed algorithm is able to obtain a single clustering partition of the data by automatically detecting clustering centers and assigning each area to its nearest cluster centroid. This algorithm is compared to the spatially adjusted agglomerative hierarchical clustering (AHC) algorithm used by Anderson et al.[9] and Adin et al.[10] The spatial method proposed by these authors works in two stages. In a first stage, a set of n spatially contiguous cluster structure candidates are obtained using the AHC algorithm over ancillary data (the q time intervals preceding the study period). Then, in a second stage, spatial relative risks for the current study period are estimated by fitting independent Bayesian hierarchical models to each cluster partition candidate. Finally, the model with the cluster structure minimizing the Deviance Information Criterion (DIC; Spiegelhalter et al.[20]) is selected as the best model. By contrast, the clustering algorithm proposed in this paper uses all the available data $\mathbf{\Psi}_i$ (ancillary data for the preceding q time intervals and data for the current study period) to obtain a unique clustering partition. The use of ancillary data makes the algorithm more stable with respect to using only the current data to detect the single cluster structure. Consequently, a single spatial model is fitted to obtain relative risk estimates, which substantially reduces the computational time. The use of all the available data in the clustering process may lead to overfitting issues if the same score, the DIC criterion in this case, is used for both selecting and evaluating the final model. In the case of the density-based clustering algorithm described below, DIC, among other criteria, is used to evaluate the final model but it is not used to obtain the clustering partition, which mitigates overfitting issues.

Density-based spatial clustering algorithm

To automatically detect cluster centers, two metrics are defined in Wang and Song[19]: local density and minimum density-based distance. Thus, the algorithm has its basis in the assumption that cluster centers are points with high local density and relatively large distance to other points with higher local densities. The local density metric for an object A_i is defined as the inverse of the average distance between A_i and its K-nearest neighbors. Here, K is the size of the neighborhood, a user-defined parameter of the algorithm, and the K-nearest neighbors of A_i are defined as those K areas with the lowest distances to A_i.

The problem of estimating spatial disease risk in small areas presents some specific characteristics that can be exploited in the clustering algorithm. On the one hand, the data to be clustered correspond to the log-SMR values of the small areas from the region under study. These areas have a natural geographical arrangement in space, and all the areas within the same cluster have to be geographically connected. Therefore, this geographical arrangement of the areas may be used to define the neighborhood for each area A_i. Additionally, this neighborhood graph needs also to be considered when assigning each area to its nearest centroid to obtain the final clustering partition. On the other hand, small areas (especially low populated areas) usually present no observed cases associated with the disease of interest for long time periods. This might be also the case when analyzing rare diseases. Hence, a fixed clustering centroid ($\mathbf{\Psi}_{background}$) corresponding to an area with no observed deaths in all the time periods is introduced in the algorithm. This artificially introduced background cluster favours the identification of clusters, and avoids an over-smoothing of the estimated disease risk maps. However, as this background cluster is not always needed, the algorithm can also be fitted without considering it.

Bearing in mind the special characteristics of a disease-mapping study, we adapt both the local density and the minimum density-based distances. We define the *local density* as

$$\hat{\rho}_i = \frac{m_{i\ell}}{\sum\limits_{i \underset{\ell}{\sim} j} d_{ij}}$$

where $m_{i\ell}$ is the number of neighbors of area A_i in its ℓ-level neighborhood ($m_{i\ell}$ corresponds to the parameter K in the original algorithm[19]), $i \underset{\ell}{\sim} j$ indicates that area A_i and A_j are neighbors in an ℓ-level neighborhood, and d_{ij} is the Euclidean distance between vectors Ψ_i and Ψ_j. Here, ℓ is a user-given parameter to define the neighborhood. If $\ell = 1$ is considered, the neighborhood of an area A_i contains only its adjacent neighbors (i.e. areas that share a common border with A_i). However, greater values of ℓ can be used to consider ℓ-order neighborhoods. Additionally, the *minimum density-based distance* is defined as

$$\hat{\delta}_i = \min_j \{d_{ij} \,|\, \hat{\rho}_i < \hat{\rho}_j \text{ and } i \neq j\}.$$

According to the base idea of the clustering algorithm, clustering centers are areas with high local density ($\hat{\rho}_i$) and relatively large values of $\hat{\delta}_i$. Or similarly, areas with high values for $\hat{\gamma}_i = \hat{\rho}_i \cdot \hat{\delta}_i$ (with $i = 1, \ldots, n$). In Wang and Song[19], an outward statistical test based on long tailed distributions is used to automatically detect cluster centroids. However, this test tends to identify a small number of clusters. When only a few clusters are detected, low values for cluster-level spatial smoothing parameters

are usually estimated which leads the model to smooth over large risk discontinuities and provides worse risk estimates. Therefore the outward statistical test seems to be too restrictive to detect clusters when estimating the distribution of mortality rates with a large number of small areas. Thus our proposed algorithm uses a simple outlier detection criteria based on a boxplot, which obtains as cluster centroids those areas with $\hat{\gamma}_i$ values greater than twice the interquartile range. The final risk estimation is not very sensitive to the threshold to detect outliers since exploratory analyses showed very similar results were obtained using 1.5, 2, or 2.5 times the interquartile range. Once the clustering centroids are obtained, the final clustering partition is iteratively constructed by assigning each area to its closest cluster, taking into account that the areas within the same cluster must be geographically connected. Note that in this step, an area can be assigned to the background cluster instead of to one of the detected clusters. The algorithm is described as follows:

Density-based spatial clustering (DBSC) algorithm

1: Calculate $\hat{\rho}_i$, $\hat{\delta}_i$ and $\hat{\gamma}_i$ for $i = 1, \ldots, n$.
2: Obtain cluster centers $(\mathbf{c}^1, \ldots, \mathbf{c}^m)$ as the outliers in $\hat{\gamma}_i$, $i = 1, \ldots, n$. Note that cluster centers correspond to specific areas.
3: Construct $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ an initial clustering structure where each clustering centroid $(\mathbf{c}^1, \ldots, \mathbf{c}^m)$ is in its own a singleton cluster.
4: Construct an empty background cluster $\mathcal{C}_{m+1}$.
5: Compute $\mathbf{D}$, a $n \times m$ matrix with the distances between each areal unit A_i and each cluster $\mathcal{C}_j$ so that:

$$\mathbf{D}_{i\mathcal{C}_j} = \begin{cases} d_{i\mathbf{c}^j}, & \text{if} \quad i \sim \mathcal{C}_j \quad \text{and} \quad A_i \notin \{\mathcal{C} \cup \mathcal{C}_{m+1}\} \\ \infty, & \text{otherwise} \end{cases}$$

where $d_{i\mathbf{c}^j}$ is the Euclidean distance between $\mathbf{\Psi}_i$ and $\mathbf{\Psi}_{\mathbf{c}^j}$, and $i \sim \mathcal{C}_j$ means that cluster $\mathcal{C}_j$ contains at least one area that share a common border with A_i.
6: Repeat steps 7-14 until each areal unit $\{A_1, \ldots, A_n\}$ is assigned to a cluster:
7: Set $(s, r) = \arg\min(\mathbf{D}_{i\mathcal{C}_j})$, that is, the identifiers of the areal unit and cluster that have the minimum distance. In case of ties, (s, r) are taken at random from the set identifiers with minimum distance.
8: Compute $\mathbf{D}_{s\mathcal{C}_{m+1}}$ as the Euclidean distance between $\mathbf{\Psi}_s$ and $\mathbf{\Psi}_{background}$, where $\mathbf{\Psi}_{background}$ is the log-SMR vector describing the centroid of the background cluster.
9: **if** $\mathbf{D}_{s\mathcal{C}_r} \leq \mathbf{D}_{s\mathcal{C}_{m+1}}$ **then**
10: $\mathcal{C}_r = \{\mathcal{C}_r \cup A_s\}$
11: **else**
12: $\mathcal{C}_{m+1} = \{\mathcal{C}_{m+1} \cup A_s\}$
13: **end if**
14: Update $\mathbf{D}$
15: Obtain the clustering partition $\mathcal{C} = \{\mathcal{C} \cup \mathcal{C}_{m+1}\}$. Please note that background cluster, $\mathcal{C}_{m+1}$, might not be geographically connected.
16: Split $\mathcal{C}_{m+1}$ into geographically connected clusters to obtain the final configuration with k clusters ($k \geq m + 1$)

Disease risk estimation model

Once a cluster configuration has been obtained using the algorithm described in the previous section, a Bayesian hierarchical model is fitted to the data. Conditional on the

relative risk r_i, the number of observed cases is assumed to follow a Poisson distribution

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

Here, $\log E_i$ is an offset, and the specification of $\log r_i$ depends on the number of clusters in the partition $\mathbf{C} = \{C_1, \ldots, C_k\}$. If no cluster structure is detected the log-risk in Equation (1) is modelled as

$$\log r_i = \eta + \xi_i, \tag{2}$$

where η is an intercept representing an overall level of risk, and ξ_i is a spatially structured random effect with a Leroux et al.[6] CAR prior distribution (hereafter LCAR model). That is, the prior for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^{'}$ is given by

$$\boldsymbol{\xi} \sim N(\mathbf{0}, [\tau_\xi(\lambda_\xi \mathbf{R}_\xi + (1 - \lambda_\xi)\mathbf{I}_n)]^{-1}),$$

where τ_ξ is a precision parameter, λ_ξ is a spatial smoothing parameter taking values between 0 and 1, $\mathbf{R}_\xi$ is determined by the spatial neighbourhood with $(R_\xi)_{ij} = -1$ if areas i and j are neighbours and $(R_\xi)_{ij} = 0$ otherwise. The diagonal entries R_ξ equal the number of neighbours of the i-th area. $\mathbf{I}_n$ is an identity matrix of dimension $n \times n$. Note that the covariance matrix of the LCAR model is of full rank whenever $0 \leq \lambda_\xi < 1$, but a confounding problem still remains[21]. So, a sum-to-zero constraint $\sum_{i=1}^{n} \xi_i = 0$ is enforced in the model fitting.

When the obtained partition has more than one cluster, we extend the model of Equation (2) to include the cluster configuration structure as follows. If the number of clusters is low, cluster-level fixed effects are included in the model as

$$\log r_i = \eta + \xi_i + \sum_{j=1}^{k} I[A_i \in C_j]\beta_j, \tag{3}$$

where $(\beta_1, \ldots, \beta_k)'$ are the fixed effects associated to each cluster, and $I[\cdot]$ denotes an indicator function, so that $I[A_i \in C_j]$ equals 1 if area A_i lies in cluster C_j, and zero otherwise. If the number of clusters is high, cluster-level random effects are included as

$$\log r_i = \eta + \xi_i + \phi_{j(i)}, \tag{4}$$

where $j(i)$ denotes that area A_i is in cluster C_j. As for the area-level random effect, a LCAR prior distribution is assumed for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)^{'}$, so that

$$\boldsymbol{\phi} \sim N(\mathbf{0}, [\tau_\phi(\lambda_\phi \mathbf{R}_\phi + (1 - \lambda_\phi)\mathbf{I}_k)]^{-1}),$$

where τ_ϕ is a precision parameter, λ_ϕ is a spatial smoothing parameter, $\mathbf{I}_k$ is an identity matrix of dimension $k \times k$, and $\mathbf{R}_\phi$ is the spatial neighborhood matrix for the cluster configuration $\mathbf{C}$ based on whether an area in cluster C_i border an area in cluster C_j. An independent prior could have also been considered for the cluster effect, but the LCAR prior is more general and contemplates the independent prior as a particular case. To achieve identifiability, the following sum-to-zero constrains are placed over the area and cluster random effects respectively, $\sum_{i=1}^{n} \xi_i = \sum_{j=1}^{k} \phi_{j(i)} = 0$. Unlike the two-level spatial model considered in Adin et al.[10], the original neighborhood structure of the small areas $\mathbf{R}_\xi$ is taken into account in Model (4) to smooth risks borrowing information from spatial neighbors. Due to the characteristics of the problem analysed in the motivating application of Section 5, where the proportion of areas with zero counts could be high, a background cluster is also considered by the DBSC algorithm as described in the previous section. Usually, this background cluster (that can be split into several geographically connected clusters) is spread across all the region under study.

In all the models described above, improper uniform prior distributions on the positive real line are considered for the standard deviations, i.e., $\sigma = 1/\sqrt{\tau} \sim U(0, \infty)$. In addition, a standard uniform distribution is given to the spatial smoothing parameters λ_ξ and λ_ϕ. Finally, $N(0, 10)$ priors are given to the fixed effects β_j in Model (3). The code to fit these models with `R-INLA` (`www.r-inla.org`) is given as Supplementary Material.

Simulation Study

A simulation study is conducted to compare the performance of the DBSC model (the two-stage method that uses the density based spatial clustering algorithm proposed in this paper) with the AHC model (the two-stage approach proposed by Adin et al.[10] based on an agglomerative hierarchical clustering algorithm described by Anderson et al.[9]). We use $n = 508$ municipalities of the Spanish autonomous regions of Navarre and the Basque Country as the simulation scenario, instead of $n = 7,907$ Spanish municipalities, due to unfeasibility of fitting the model proposed by Adin et al.[10] in this last case. Three different scenarios have been considered: a spatially smooth surface with no clusters (*Scenario 1*), a scenario with $k = 11$ non overlapping high/low risk clusters spread across the area of analysis (*Scenario 2*), and a last scenario with $k = 9$ non overlapping high-risk clusters whose centroids are located in four of the municipalities with the highest values of expected cases in addition to those ones corresponding approximately to the 5th, 25th, 50th, 75th, and 95th percentiles of the expected counts (*Scenario 3*). In this last scenario, in contrast to *Scenario 2*, the clusters have been generated so that the risks of the regions decrease gradually to imitate the behavior that is usually observed when

analyzing real data. The true risk surfaces for scenarios one, two, and three are displayed in Figure 2, Figure 3, and Figure 4 respectively.

Data generation

The simulated counts are generated from a Poisson distribution with mean $E_i r_i$, where E_i are the expected number of cases computed from global cancer mortality data for the period 2011-2015. In this case, the expected cases vary from 0.55 to 5540.09 with a mean of 77.27 and a median of 12.31. To imitate the case of analyzing rare diseases or very small domains, the expected number of cases have been multiplied by the scale factors 1, 1/10, and 1/30 giving rise to three different sub-scenarios (hereafter denoted as A, B, and C respectively). The true log-risk surfaces have been simulated as

$$\begin{aligned} \text{Scenario 1}: &\quad \log r_i = \xi_i, \\ \text{Scenario 2/3}: &\quad \log r_i = \xi_i + \sum_{j=1}^{k} I[A_i \in C_j]\beta_j, \end{aligned}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^{'}$ is a spatially structured random effect, and $(\beta_1, \ldots, \beta_j)'$ are the fixed effects associated to each cluster C_j. To obtain a spatially smooth surface, the vector of random effects $\boldsymbol{\xi}$ is generated from a LCAR prior distribution with spatial smoothing parameter $\lambda_\xi = 0.75$. The precision parameter τ_ξ is adequately chosen so that $\exp(\xi)$ takes values within the range $[0.91, 1.1]$. This means that the risk of a certain region is less than 10% lower/higher in comparison with the whole study area. In Scenario 2, values of $\beta_j = 0.5$ are established for the regions in a high-risk cluster, and values of $\beta_j = -0.5$ for regions in a low-risk cluster. In Scenario 3, values of $\beta_j = 0.5$ are established for the centroids of the high-risk clusters, and values of $\beta_j = 0.35$ for the surrounding areas. In both scenarios, values of $\beta_j = 0$ are considered for the rest of the municipalities.

A total of 100 simulations have been generated for each one of the nine sub-scenarios. As in Adin et al.[10], four data sets are generated for each simulation run, one set as study data and $q = 3$ sets of ancillary data corresponding to preceding time periods $T-1$, $T-2$, and $T-3$, respectively. Uniform random noise with parameters ± 0.05 has been sequentially added to generate the log-risk surfaces of each previous time periods. That is, for each simulation the following counts have been generated

$$\begin{aligned} O_i &\sim Poisson(E_i r_i), \\ O_i^{T-1} &\sim Poisson(E_i^{T-1} r_i^{T-1}), \quad \text{with } \log r_i^{T-1} = \log r_i + \text{Unif}(-0.05, 0.05), \\ O_i^{T-2} &\sim Poisson(E_i^{T-2} r_i^{T-2}), \quad \text{with } \log r_i^{T-2} = \log r_i^{T-1} + \text{Unif}(-0.05, 0.05), \\ O_i^{T-3} &\sim Poisson(E_i^{T-3} r_i^{T-3}), \quad \text{with } \log r_i^{T-3} = \log r_i^{T-2} + \text{Unif}(-0.05, 0.05), \end{aligned}$$

where E_i^{T-1}, E_i^{T-2}, and E_i^{T-3} are the expected number of cases computed from global cancer mortality data for the periods 2006-2010, 2001-2005, and 1996-2000 respectively. On average, the proportion of areas with values of O_i equal to zero is around 5%, 35%, and 57% in the sub-scenarios A, B, and C, respectively (corresponding to scale factors of 1, 1/10, and 1/30).

Results

Three different models have been fitted for each simulated data set: the LCAR model with no clustering effect (a standard model used in disease mapping studies), the AHC model described in Section 3.2 with a LCAR spatial prior distribution and a clustering effect for the second stage of the algorithm, and the DBCS model proposed in this paper (using different neighborhood levels, ℓ). We compare the models's performance in terms of mean absolute relative bias (MARB) and mean relative root mean square error (MRRMSE) defined as

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

where r_i and $\hat{r}_i^k$ are the true generated risk, and the posterior median estimates of relative risk in the k-th simulation, respectively. In addition, coverage probabilities and 95% credible intervals's lengths have been computed. The Logarithmic Score (LS, Gneiting and Raftery[22]), which is a cross validation measure based on the conditional predictive ordinate[23], is also given as a measure of the model predictive ability. Models with a lower log score value are preferred. Results are shown in Table 2.

In the scenario without cluster effects (Scenario 1), the LCAR model performs the best (as expected) in terms of relative risk estimates. Equal values of MARB and higher (although small) values of MRRMSE are obtained for the rest of the models in Scenario 1A, the one with the largest number of expected cases. In this scenario, the DBSC$_{l=3}$ is giving better values of MRRMSE than AHC. In a real scenario with a high number of expected cases where we really do not know if discontinuities in the risk surface are present, DBSC is very competitive with respect to LCAR. As the number of expected cases decreases, the models with a cluster effect behave worse, mainly in terms of MRRMSE. In addition, the AHC model shows the worst LS values. With respect to the DBSC model, it is observed that increasing the size of the neighborhood (the ℓ parameter) leads, in general, to worse MARB and MRRMSE results in this "middle size" problem. True risks, as well as average values of the relative risk estimates (posterior medians) for the LCAR, the AHC, and the best DBSC model, according to the Logarithmic Score,

are displayed in Figure 2. Maps for DBSC models with other ℓ values can be found in https://github.com/spatialstatisticsupna/DBSC_article.

Table 2. Average values of mean absolute relative bias (MARB), mean relative root mean square error (MRRMSE), coverage percentages and length of the 95% credible interval for the risks, and Logarithmic Score (LS).

				No background cluster			With background cluster		
		LCAR	AHC	DBSC$_{\ell=1}$	DBSC$_{\ell=2}$	DBSC$_{\ell=3}$	DBSC$_{\ell=1}$	DBSC$_{\ell=2}$	DBSC$_{\ell=3}$
Scenario 1A	MARB	0.012	0.012	0.012	0.013	0.012	0.013	0.014	0.012
	MRRMSE	0.014	0.041	0.028	0.033	0.027	0.034	0.039	0.028
	Cov.(%)	84.24	85.07	85.19	85.37	85.42	85.51	85.71	85.40
	Length	0.082	0.202	0.112	0.118	0.127	0.129	0.137	0.133
	LS	1395.1	1400.4	1401.9	1414.3	1398.6	1423.7	1442.0	1401.0
Scenario 1B	MARB	0.012	0.015	0.014	0.014	0.025	0.014	0.028	0.075
	MRRMSE	0.021	0.099	0.041	0.049	0.082	0.040	0.081	0.146
	Cov.(%)	99.97	99.72	99.79	99.52	99.20	99.89	99.40	98.21
	Length	0.219	0.493	0.338	0.350	0.402	0.354	0.438	0.561
	LS	776.7	794.4	776.3	776.6	776.7	776.2	775.8	778.7
Scenario 1C	MARB	0.012	0.017	0.031	0.035	0.043	0.043	0.061	0.079
	MRRMSE	0.031	0.109	0.090	0.095	0.116	0.101	0.130	0.155
	Cov.(%)	93.99	93.73	93.53	93.40	92.35	93.49	92.08	90.95
	Length	0.399	0.648	0.601	0.599	0.626	0.650	0.667	0.698
	LS	519.6	538.9	518.9	519.3	519.6	518.8	518.8	521.0
Scenario 2A	MARB	0.060	0.044	0.055	0.056	0.061	0.055	0.056	0.060
	MRRMSE	0.125	0.122	0.126	0.130	0.130	0.130	0.132	0.131
	Cov.(%)	97.53	96.12	96.11	96.21	96.35	96.08	96.20	96.34
	Length	0.648	0.482	0.592	0.604	0.618	0.605	0.614	0.622
	LS	1533.7	1434.2	1517.6	1523.3	1520.7	1541.1	1534.3	1521.5
Scenario 2B	MARB	0.102	0.090	0.098	0.102	0.106	0.097	0.107	0.121
	MRRMSE	0.174	0.201	0.206	0.212	0.220	0.209	0.237	0.263
	Cov.(%)	98.24	98.48	96.41	96.43	96.58	96.83	96.10	95.64
	Length	1.130	1.258	1.022	1.066	1.111	1.080	1.124	1.167
	LS	828.2	845.6	819.3	821.1	823.9	818.4	821.2	831.1
	Rand Index	–	0.59	0.58	0.61	0.58	0.56	0.58	0.50
Scenario 2C	MARB	0.134	0.106	0.119	0.120	0.115	0.132	0.132	0.134
	MRRMSE	0.198	0.231	0.263	0.262	0.259	0.265	0.264	0.277
	Cov.(%)	95.35	95.43	93.57	93.81	94.02	93.16	93.82	93.89
	Length	1.417	1.508	1.254	1.311	1.384	1.285	1.319	1.379
	LS	551.9	562.5	544.4	546.7	549.3	545.0	548.4	553.7
Scenario 3A	MARB	0.050	0.039	0.044	0.046	0.048	0.043	0.045	0.047
	MRRMSE	0.097	0.095	0.099	0.102	0.101	0.102	0.106	0.102
	Cov.(%)	96.84	95.86	95.92	96.36	96.16	95.93	96.36	96.14
	Length	0.580	0.531	0.548	0.564	0.560	0.558	0.574	0.563
	LS	1524.3	1472.1	1516.2	1525.3	1517.6	1533.7	1543.6	1520.2
Scenario 3B	MARB	0.113	0.083	0.090	0.088	0.085	0.086	0.074	0.101
	MRRMSE	0.143	0.165	0.161	0.167	0.182	0.157	0.185	0.237
	Cov.(%)	96.22	96.10	93.84	95.10	95.26	94.07	94.80	94.34
	Length	0.956	1.152	0.869	0.946	0.991	0.923	1.016	1.063
	LS	834.5	849.3	826.5	828.9	831.7	826.2	827.4	834.8
Scenario 3C	MARB	0.146	0.093	0.079	0.081	0.086	0.080	0.081	0.092
	MRRMSE	0.170	0.182	0.205	0.214	0.204	0.204	0.221	0.234
	Cov.(%)	98.68	98.35	95.64	96.27	97.17	95.69	95.19	95.73
	Length	1.19	1.358	1.096	1.161	1.219	1.134	1.201	1.238
	LS	558.7	569.5	551.9	553.7	556.4	551.5	554.2	559.0

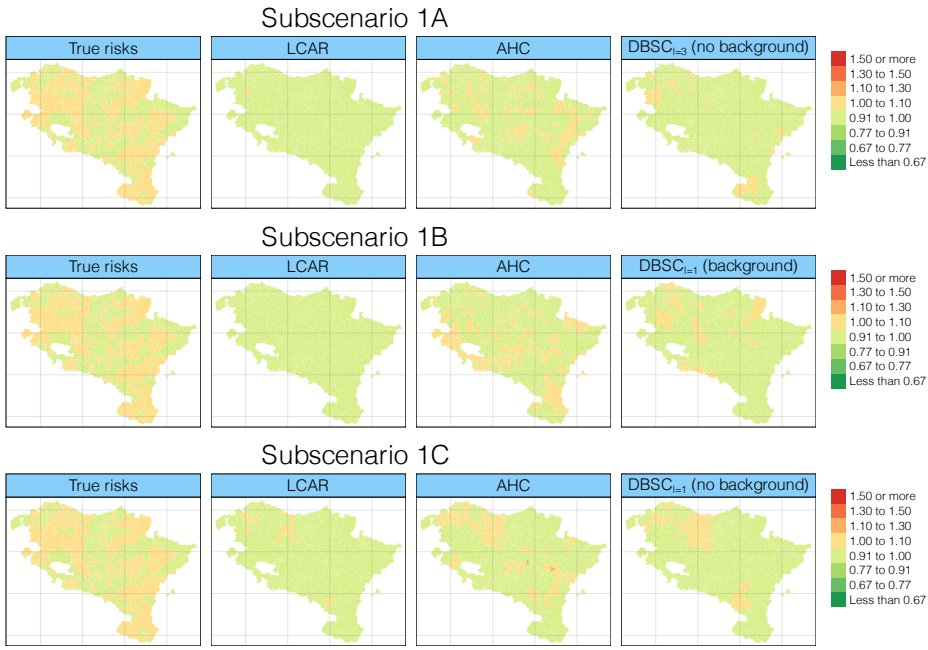

Figure 2. Average values of the relative risks r_i posterior median estimates for the simulation study of Scenario 1A (top), Scenario 1B (center), and Scenario 1C (bottom).

In Scenario 2A, the AHC model seems to be the best recovering the simulated high/low risk cluster structures, while the rest of the models perform similarly. However, in Scenario 2B, where the generated data contains a higher percentage of zeros, better coverage probabilities, narrower 95% credible intervals, and lower values of LS are obtained with the DBSC models. The maps with the true risks and the average values of the relative risks posterior median estimates are shown in Figure 3. In Scenario 2C, the one that contains almost 60% of areas with zero cases, an over-smoothing effect is observed for the LCAR model, providing a high number of background areas with relative risks over 1.1. In contrast, the DBSC$_{\ell=1}$ and the AHC models seem to recover better the high/low risk cluster structures of the simulated risks. Additionally, the new model proposal is always better in terms of LS.

Similar results are observed in Scenario 3. When the number of expected cases is high (Scenario A) the AHC model performs slightly better. However, as the number of cases is reduced, the DBSC model provides better values of MARB, more accurate 95% credibility intervals, and lower values of LS. In general, clusters with centroids located in the municipalities with the highest values of expected cases are quite well recovered by

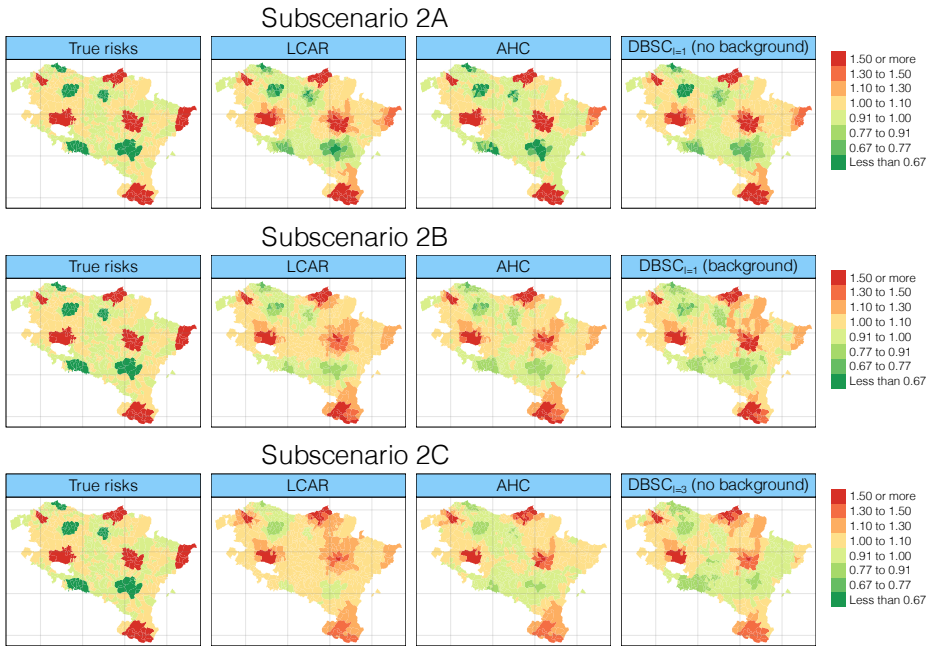

Figure 3. Average values of the relative risks r_i posterior median estimates for the simulation study of Scenario 2A (top), Scenario 2B (center), and Scenario 2C (bottom).

all the models. However, the LCAR and the AHC models seem to over-smooth the risk surfaces, especially in Scenario 3C, where better results are obtained with the DBSC$_{\ell=1}$ model (see Figure 4).

In summary, when there are no clusters and the percentage of zeros is small (Scenario 1A) all the models behave pretty well, but the new DBSC algorithm is always better than the AHC model. In the presence of clusters and a relatively high number of expected cases (Scenarios 2A y 3A), the AHC model seems to perform somehow better than the LCAR and the DBSC models (in particular when the number of expected cases is big-Scenario 2A-). Differences in terms of relative bias and variability between AHC and DBSC modes are pretty small in Scenarios 2B and 2C. When the percentage of zeros increases (Scenarios B and C), a common situation when analyzing rare disease and very small domains (as it is the case of the motivating application considered in this paper), the DBSC model is very competitive and performs better than AHC model in terms of LS. Although the LCAR model seems to behave always pretty well in terms of MRRMSE, in Scenarios 2B, 2C, 3B, and 3C, it gives big relative biases (sometimes unacceptably big,

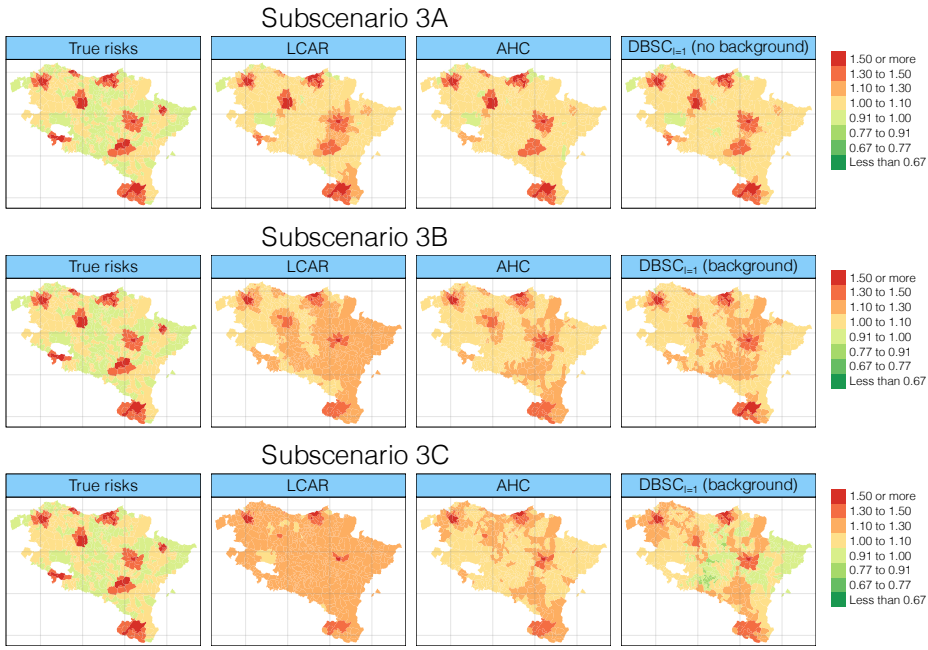

Figure 4. Average values of the relative risks r_i posterior median estimates for the simulation study of Scenario 3A (top), Scenario 3B (center), and Scenario 3C (bottom).

see also rows 2 and 3 of Figure 4), worse coverages (in general), and worse lengths of the 95% credible intervals than the DBSC model.

With respect to computing times in the simulation study (a medium size problem), the new algorithm takes on average about 0.07 minutes whereas the AHC algorithm takes on average about 60 minutes. Consequently, it is important to stress again that when the number of small areas is large (as it is the case of the motivating application) it is unfeasible to fit the AHC model.

Data analysis: stomach cancer mortality in Spanish municipalities

Stomach cancer rates have been continuously falling in the most part of the world since the middle of the last century.[15] However, stomach cancer still remains the fifth most malignant cancer in the world. In Spain, about 8,500 new cases are diagnosed every year. The number of mortality cases in the studied five-year period (2011-2015) is about 54% higher in males than in females (16,215 male cases against 10,498 females cases). One of the main features of stomach cancer epidemiology is its geographical variability. Indeed,

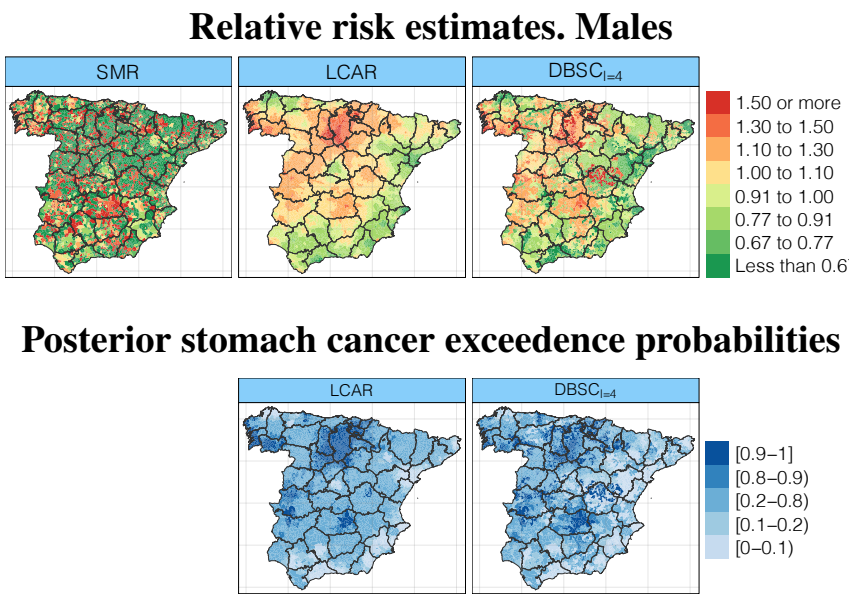

Figure 5. Maps of the SMRs and posterior median estimates for r_i (top) and posterior exceedence probabilities $P(r_i > 1|\mathbf{O})$ (botton) of male stomach cancer mortality risks in the municipalities of Spain during the period 2011-2015.

although it is known that the risk has decreased in Spain, the strong geographical pattern remains stable over time.[16] Here, we analyze males and females mortality by stomach cancer in the municipalities of Spain using a model that takes spatial discontinuities into account. As it was said before, the AHC modelling approach is computationally prohibitive to estimate mortality risks in Spain at municipality level (see comments below). Therefore, only the LCAR and the new model proposal, named the DBSC model, are considered to analyze the data.

The indirect age-standardization method has been used to compute the number of expected cases. So, the excess of risk in each municipality will be interpreted with respect to Spain in the studied period. The expected number of cases in males ranges from 0 to 1,574 (mean value of 2.23 and median value of 0.42), while the number of expected cases in females varies from 0 to 1,195 (mean value of 1.43 and median value of 0.23). The maps with the SMRs for each small area are shown in Figure 5 and Figure 6 for males and females respectively. The expected number of cases for the preceding $q = 4$ time periods used as ancillary data to determine the cluster structure has been computed similarly.

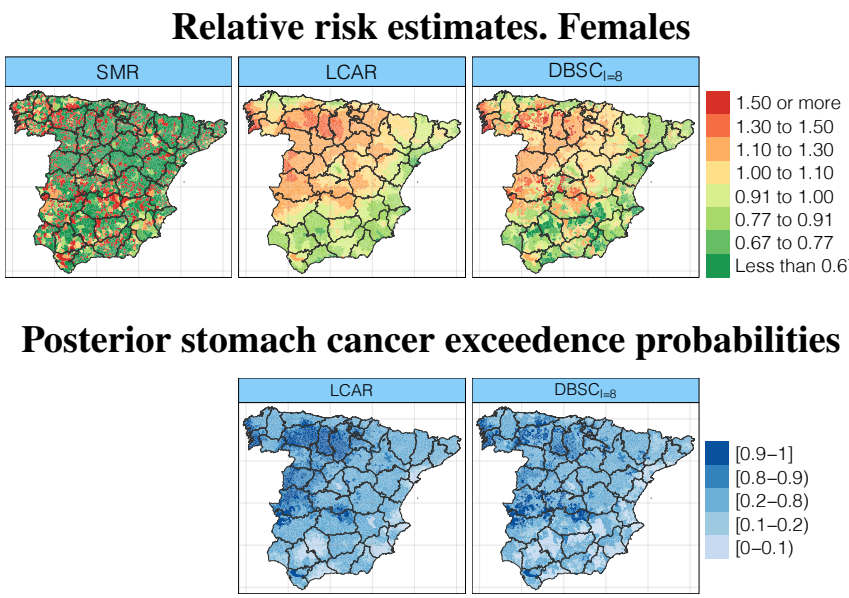

Figure 6. Maps of the SMRs and posterior median estimates for r_i (top) and posterior exceedence probabilities $P(r_i > 1|\mathbf{O})$ (botton) of female stomach cancer mortality risks in the municipalities of Spain during the period 2011-2015.

Using the observed and expected cases, we run the DBSC algorithm to obtain the clustering partition, and then, we fitted the Bayesian hierarchical models described in Section 3.2 using the simplified Laplace approximation strategy of INLA. Calculations were mader on a twin superserver with four processors, Intel Xeon 6C and 96GB RAM, using the R-INLA (stable) version 18.07.12 of R-3.5.3. In addition, the usual LCAR model without cluster effects was also fitted. The high computational cost of fitting the AHC model to 7,907 small areas made unfeasible to include it for comparison purposes. For example, the computing time to obtain the 7,907 cluster structures with the AHC algorithm proposed by Anderson et al.[9] took about 38 hours. Note also that to implement the whole AHC proposal one still needs to fit 7,907 models, one for each cluster structure. This is computationally prohibitive even in the above cited machine. The new DBSC algorithm took about 3-4 minutes to obtain the cluster structure and less than 13 minutes to fit any of the Bayesian hierarchical models considered here. Although the ℓ-order neighborhood matrices must be computed, and the computational cost increases for high ℓ values (due to the lost of sparsity), these matrices can be previously computed and storage since they only depend on the spatial configuration of the small areas.

Table 3. Stomach cancer mortality data: average value of effective number of parameters (p_D), deviance information criterion (DIC), Watanabe-Akaike information criterion (WAIC), Logarithmic Score (LS) and computational time (Time) in minutes. Models were fitted using the simplified Laplace approximation strategy of INLA.

	Male population					Female population				
	p_D	DIC	WAIC	LS	Time	p_D	DIC	WAIC	LS	Time
LCAR	307.8	15162.8	15146.2	7594.4	13.0	216.9	12200.9	12198.4	6112.5	12.8
DBSC$_{\ell=1}$	368.9	15098.3	15069.9	7563.7	16.3	266.2	12160.8	12152.2	6095.4	16.5
DBSC$_{\ell=2}$	370.0	15075.1	15044.1	7550.9	16.2	271.2	12162.9	12154.2	6097.3	16.6
DBSC$_{\ell=3}$	369.3	15010.4	14979.7	7517.1	16.3	268.3	12149.1	12141.7	6090.1	16.5
DBSC$_{\ell=4}$	396.6	**14973.6**	**14942.8**	**7501.9**	16.3	265.8	12133.2	12126.4	6082.5	16.6
DBSC$_{\ell=5}$	363.5	15028.6	15006.3	7529.9	16.1	268.8	12128.5	12119.8	6078.9	16.6
DBSC$_{\ell=6}$	372.6	15008.5	14988.6	7522.0	16.1	276.7	12118.5	12115.1	6077.5	16.4
DBSC$_{\ell=7}$	362.4	14984.7	14974.3	7514.4	16.2	267.4	12100.8	12098.5	6067.8	16.5
DBSC$_{\ell=8}$	361.5	14980.4	14971.2	7512.9	16.4	274.0	**12091.6**	**12087.4**	**6062.2**	16.4
DBSC$_{\ell=9}$	361.5	14980.4	14971.2	7512.9	16.3	274.0	12091.6	12087.4	6062.2	16.7
DBSC$_{\ell=10}$	370.0	15075.1	15044.1	7550.9	16.3	271.2	12162.1	12154.2	6097.3	16.6

The LCAR and the new model proposal (DBSC) with several neighborhood orders are compared in Table 3 in terms of several model selection criteria such as the DIC, the Logarithmic Score and the Watanabe-Akaike information criterion[24]. All criteria point to the DBSC model as the best (DBSC$_{\ell=4}$ for males and DBSC$_{\ell=8}$ for females). The maps with posterior median estimates for r_i, and posterior exceedence probabilities $P(r_i > 1|\mathbf{O})$ of stomach cancer mortality risks for males and females are shown in Figure 5 and Figure 6, respectively. The disease risk surfaces estimated by the LCAR and DBSC models are somehow similar. However, the geographical pattern obtained with the LCAR model is much smoother than that obtained with the DBSC model. The new model seems to be more effective capturing local discontinuities between municipalities and avoiding over-smoothing.

Regarding results, slightly different geographical risk surfaces are estimated for males and females, however, in both cases, the municipalities with highest and lowest relative risk estimates are located in similar regions. The areas with higher relative risks for both males and females are located in some coastal municipalities of Pontevedra and La Coruña provinces, with estimated relative risks even above 2 (see the map of administrative division of continental Spain in Figure 1). In addition, high-risk spots are detected between Burgos and Palencia provinces for male population (relative risks greater than 1.75), while municipalities with high-risks located in the north of Palencia are also observed for females (relative risks around 1.5). In contrast, most of the municipalities with the lowest estimated risks are located across the Mediterranean coast,

highlighting some areas in the eastern part of the province of Alicante with relative risks below 0.5 in both sexes.

Table 4 shows stomach cancer mortality SMRs and the highest/lowest estimated risks for Spanish provinces' capitals. Again, Burgos and Palencia are between the capitals with the highest risks in comparison with the whole Spain for both sexes. Other capitals like Vitoria, Cuenca, Teruel, and Soria also show a significant risk excess for males whereas Pontevedra, Salamanca, and León have high significant risks in females. On the other hand, Barcelona, Tarragona, Murcia, Málaga and Sevilla are low-risk capitals for both sexes, all of them located on the east coast of Spain. Córdoba, which is not located on the coast, is also one of the capitals with the lowest relative risks in both sexes. Further research is needed to analyze possible risk factors associated with high relative risks regions.

Discussion

Smoothing risks and detecting high/low-risk clusters are somehow two contradictory goals which have usually been treated separately in disease mapping. However, recent proposals try to reconcile both objectives avoiding over-smoothing. More specifically, the AHC method is a two-stage approach to obtain a clustering partition of the areal units (first stage) and to estimate spatial risks by fitting Bayesian hierarchical models that take the cluster configuration structures into account. Nevertheless, this proposal obtains a set of candidate clusters to be evaluated, and the cost of this process is computational prohibitive in problems with a large number of areal units. In this paper, a new two-stage method is proposed to overcome these limitations. The first step of the new proposal is based on a new algorithm that is able to automatically obtain a single clustering partition by detecting the centroids of the clusters accounting for the special characteristics of the spatial data. Then, in the second stage, only one model needs to be fitted.

Our simulation study indicates that the proposed DBSC model provides competitive results at a much lower computational time. Indeed, when the number of expected cases is relatively low, which is common when analyzing rare diseases or very small domains, our proposal performs better than both the AHC model and the LCAR model, the latter providing unacceptably high relative biases. When the the number of small areas is big and discontinuities are expected, an alternative to the LCAR and AHC model is really needed. On one hand, because the LCAR model is over-smoothing and on the other hand, because the AHC model is computationally prohibitive.

Regarding the real data analysis, it is important to stress that concerning cancer, mortality is the only comprehensive source of information in Spain as there are not cancer incidence registers covering the whole country. In addition, the quality of cancer

Table 4. Observed and expected stomach cancer death cases, SMRs, relative risks r_i, posterior exceedence probabilities $P(r_i > 1|\mathbf{O})$ and 95% credible intervals for the capitals of Spanish provinces with highest/lowest estimated risks.

	Municipality	Observed	Expected	SMR	Risk (r_i)	$P(r_i > 1\|\mathbf{O})$	95% CI
Males	Vitoria	155	92.4	1.68	1.60	1.00	[1.39, 1.84]
	Cuenca	28	19.9	1.41	1.56	1.00	[1.22, 1.96]
	Burgos	119	69.1	1.72	1.56	1.00	[1.32, 1.83]
	Palencia	53	33.4	1.59	1.51	1.00	[1.21, 1.86]
	Teruel	17	14.1	1.21	1.46	0.99	[1.09, 1.93]
	Soria	23	15.9	1.45	1.34	0.99	[1.07, 1.67]
	...	...	...	...	...	...	...
	Barcelona	545	647.6	0.84	0.86	0.00	[0.79, 0.93]
	Málaga	151	173.7	0.87	0.85	0.01	[0.73, 0.98]
	Córdoba	86	108.8	0.79	0.85	0.02	[0.72, 0.99]
	Tarragona	36	46.3	0.78	0.84	0.06	[0.66, 1.05]
	Murcia	113	131.4	0.86	0.84	0.01	[0.72, 0.96]
	Sevilla	190	228.8	0.83	0.83	0.00	[0.74, 0.94]
Females	Burgos	77	49.7	1.55	1.54	1.00	[1.26, 1.87]
	Pontevedra	34	21.2	1.61	1.48	1.00	[1.18, 1.85]
	Palencia	43	24.3	1.77	1.48	1.00	[1.16, 1.87]
	Salamanca	85	52.1	1.63	1.42	1.00	[1.17, 1.71]
	Orense	50	34.1	1.47	1.32	0.99	[1.04, 1.65]
	León	54	45.4	1.19	1.32	0.99	[1.07, 1.60]
	...	...	...	...	...	...	...
	Sevilla	142	167.7	0.85	0.87	0.02	[0.76, 1.00]
	Madrid	752	902.9	0.83	0.84	0.00	[0.79, 0.90]
	Barcelona	397	486.0	0.82	0.84	0.00	[0.76, 0.91]
	Málaga	110	121.8	0.90	0.82	0.01	[0.69, 0.97]
	Tarragona	24	30.5	0.79	0.76	0.03	[0.55, 1.01]
	Córdoba	47	76.3	0.62	0.75	0.00	[0.62, 0.90]

mortality data has been stated to be accurate with quality indicators comparable to other industrialised countries.[25] Stomach cancer rates for females are smaller that those for men, but the reasons why this happens remain still unknown. Several risks factors could be responsible for the risk differences observed among Spanish municipalities. Some authors suggest associations between stomach cancer and diet and smoking, while it seems to be proved the relationship between the existence of the H. pylori and certain types of gastric cancer. Alternative explanations have to be with the presence of environmental exposures such as nitrates, arsenic or other metals, but these associations have not been proved yet.[26]. Further research is needed at this respect.

Acknowledgements

We would like to thank Spanish Statistical Office for providing the data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Adin, A., Santafé, G., and Ugarte, M.D. research has been supported by project MTM2017-82553-R (AEI/FEDER, UE). Lee, D. research has been supported by the UK Medical Research Council (Grant number MR/L022184/1).

References

1. Lawson A, Banerjee S, Haining RP and Ugarte MD. *Handbook of Spatial Epidemiology*. Chapman and Hall/CRC, 2016.
2. Waller L and Carlin B. Disease mapping. In Gelfand AE, Diggle PJ, Fuentes M and Guttorp P (eds) *Handbook of Spatial Statistics*. Chapman and Hall/CRC, 2010.
3. Lawson A. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. 3th ed. Chapman and Hall/CRC, 2018.
4. Martinez-Beneito MA and Botella-Rocamora P. *Disease Mapping: From Foundations to Multidimensional Modeling*. CRC Press, 2019.
5. Besag J, York J, and Mollié A. Bayesian image restoration, with two applications in spatial statistics. *An I Stat Math* 1991; **43(1)**: 1-20.
6. Leroux BG, Lei X, and Breslow N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Halloran M and Berry D (eds) *Statistical Models in Epidemiology, the Environment and Clinical Trials*. New York: Springer-Verlag, 1999; pp. 179-191.
7. Dean CB, Ugarte MD, Militino, AF. Detecting interaction between random regions and fixed age effects in disease mapping. *Biometrics* 2001; **57**: 197-202.
8. Mclafferty S. Disease cluster detection methods: recent developments and public health implications. *Ann. GIS* 2015; **21(2)**:127-133.
9. Anderson C, Lee D and Dean N. Identifying clusters in Bayesian disease mapping. *Biostatistics* 2014; **15(3)**: 457-469.

10. Adin A, Lee D, Goicoa T and Ugarte MD. A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. *Stat Meth Med Res* 2019; **28(9)**:2595-2613.
11. Embrechts M, Gatti C, Linton J and Roysam B. Hierarchical Clustering for Large Data Sets. In Georgieva P, Mihaylova L and Jain LC (eds) *Advances in Intelligent Signal Processing and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013; pp. 197-233.
12. Rue H, Martino S, and Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Roy Stat Soc B* 2009; **71(2)**: 319-392.
13. World Health Organization. *International Statistical Classification of Diseases and related Health problems (10th revision)*. ICD-10 online version available https://icd.who.int/browse10/2016/en (accessed 21 February 2019).
14. Ferlay J, Colombet M, Soerjomataram I, Dyba T, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. *Eur J Cancer* 2018; **103**:356-387.
15. Aragonés N, Goicoa T, Pollán M, et al. Spatio-temporal trends in gastric cancer mortality in Spain: 19752008. *Canc Epidemiol* 2013; **37(4)**:360369.
16. López-Abente G, Aragonés N, Pérez-Gómez B, et al. Time trends in municipal distribution patterns of cancer mortality in Spain. *BMC Canc* 2014; **14(1):** 535.
17. Pace M, Lanzieri G, Glickman M and Zupanič T. *Revision of the European Standard Population: report of Eurostat's task force*. Publications Office of the European Union, 2013.
18. Rodriguez A and Laio A. Clustering by fast search and find of density peaks. *Science* 2014; **344(6191)**:1492-1496.
19. Wang G and Song Q. Automatic clustering via outward statistical testing on density metrics. *IEEE T Knowl Data En* 2016; **28(8)**:1971-1985.
20. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *J R Stat Soc Series B (Stat Methodol)* 2002; **64(4)**:583-639.
21. Goicoa T, Adin A, Ugarte MD et al. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stoch Env Res Risk A* 2018; **32(3)**:749-770.
22. Gneiting T and Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 2007; **102(477)**:359378.
23. Pettit L. The conditional predictive ordinate for the normal distribution. *J R Stat Soc Series B (Stat Methodol)* 1990; **52(1)**:175-184.

24. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 2010; **11**:3571-3594.
25. Pérez-Gómez B, Aragonés N, Pollán M et al. Spatio-temporal trends in gastric cancer mortality in Spain: 1975–2008. *Gaceta Sanitaria*, 2006; **20**:42-51.
26. Aragonés N, Pérez-Gómez B, Pollán M et al. The striking geographical pattern of gastric cancer mortality in Spain: environmental hypothesis revisited. *BMC Cancer* 2009; **9**:316.