Penalized composite link mixed models for two-dimensional count data

Diego Ayma*[1], María Durbán[1], Dae-Jin Lee[2], and Paul Eilers[1,3]

[1]Department of Statistics, Universidad Carlos III de Madrid, Spain

[2]BCAM - Basque Center for Applied Mathematics, Spain

[3]Erasmus University Medical Center, The Netherlands

8th May 2015

Abstract

Mortality data provide valuable information for the study of the spatial distribution of mortality risk, in disciplines such as spatial epidemiology, medical demography, and public health. However, they are often available in an aggregated form over irregular geographical units, hindering the visualization of the underlying mortality risk and the detection of meaningful patterns. Also, it could be of interest to obtain mortality risk estimates on a finer spatial resolution, such that they can be linked with potential risk factors — in a posterior correlation analysis — that are usually measured in a different spatial resolution than mortality data. In this paper, we propose the use of the penalized composite link model and its representation as a mixed model to deal with these issues. This model takes into account the nature of mortality rates by incorporating the population size at the finest resolution, and allows the creation of mortality maps at a desirable scale, reducing the visual bias resulting from the spatial aggregation within original units. We illustrate our proposal with the analysis of several datasets related with deaths by respiratory diseases, cardiovascular diseases, and lung cancer.

Keywords

Penalized composite link models; Mixed models; Mortality rates; Spatial disaggregation.

*Corresponding author:

Diego Ayma Anza, Departamento de Estadística, Universidad Carlos III de Madrid, Escuela Politécnica Superior,
Av. de la Universidad, 30, 28911 Leganés (Madrid), Spain.
E-mail: dayma@est-econ.uc3m.es.

1 Introduction

Disease maps deal with public health data that are usually available in an aggregated form over geographical units, like counties, districts, and municipalities. This is done to protect patients' privacy, making impossible the reconstruction of personal information. Epidemiologists, health care practitioners, and other related researchers use these data to study the spatial distribution of mortality caused by an specific disease, and thus identify areas of excess and their potential risk factors. Choropleth maps are then commonly used to display such distribution but they must be interpreted with caution, because the "small number problem" effect (Waller and Gotway, 2004) — that often affects health data — leads to a large uncertainty about rates calculated from small or sparsely populated areas, thus hindering the detection of meaningful spatial patterns. Another problem that could arise is the spatial misalignment between potential risk factors and health data: in general, the former are available on a finer spatial resolution than the latter. For example, most deprivation indices are built on the smallest possible geographical units of a certain region (see Rey et al., 2009; Salmond and Cramptom, 2012) or even on a fine grid (Caudeville et al., 2012). Environmental agents (such as air pollution, or electric and magnetic fields, to name a few) constitute examples of risk factors that vary continuously in space. Consequently, this situation prohibits their direct correlation analysis that is a critical step in a disease control intervention, which includes the implementation of appropriate control activities and the resource allocation of health funds. Therefore, it is important to develop spatial methodologies that circumvent those drawbacks, which filter the noise caused by the small number problem and allow the creation of mortality maps, from aggregated data, at a resolution compatible with the spatial support of risk factors.

It is noteworthy, in the disease mapping literature, the amount of statistical tools that deal with the reduction of the noise in mortality rates associated to geographical units (see Besag et al., 1991; MacNab and Dean, 2002; Fahrmeir et al., 2004; Goovaerts, 2005, 2006b; Lee and Durbán, 2009; among others). All of them give smoothed mortality estimates that are assumed constant over each unit, yielding a coarse spatial trend. In turn, several works about spatial disaggregation of health data appeared more recently. In a geostatistical framework, Kelsall and Wakefield (2002) obtained pointwise posterior medians of the underlying continuous risk surface, for colorectal cancer mortality in UK district of Birmingham, via a Gaussian random field (GRF) model. Goovaerts (2006a) generalized the Poisson kriging algorithm given by Monestiez et al. (2005, 2006), in which incorporates the size and shape of the units, as well as the population density, into the filtering of noisy mortality rates, and allows the mapping of the corresponding mortality risk at a fine resolution. The performance of his approach (called area-to-point Poisson kriging) was compared with two geostatistical methods that allow the creation of continuous mortality maps from aggregated data: the first one corresponds to the simple interpolation of raw rates to the nodes of a fine grid using ordinary kriging, and the second one is the approach proposed by Berke (2004), in which the raw rates are replaced by their global empirical Bayes estimates before the interpolation process. Local Bayes estimates were also considered in the analysis, to attenuate the smoothing effect produced by the global mean term involved in the calculation of those Bayes estimates. The performance comparison results showed that the area-to-point Poisson Kriging give more detailed spatial trends than the other geostatistical methods, when the geographical units vary widely in size and shape. Lately, and from a Bayesian inferential viewpoint, Diggle et al. (2013) used the class of log-Gaussian Cox processes (as models for spatial point process data) to construct a continuous map of lung cancer mortality risks in the Castile-La Mancha, region of Spain, from spatially discrete data.

In this paper, we propose the use of the penalized composite link model of Eilers (2007) for the case of spatial aggregation, and its representation as a mixed model. The resulting model, which we call penalized composite link mixed model, allows us to create mortality maps from aggregated health data at a fine spatial resolution, and to incorporate finer scale information into the filtering of noisy mortality rates. We illustrate two cases of spatial disaggregation: from coarse geographical units to smaller units (area-to-area (or ATA) case), and from coarse geographical units to a fine grid (area-to-point (or ATP) case). We will obtain a more refined spatial trend (represented as a choropleth map) in the first case, and a continuous surface (or isopleth map) without spatial boundaries for the second. The advantage of producing isopleth maps is to reduce the visual bias associated with the interpretation of choropleth maps (Cressie, 1993), which is produced by the variation in shape and size of the geographical units. Also, we illustrate the case where the aggregated count data have an array structure. In this context, we use the generalized linear array model (or GLAM) arithmetic given in Currie et al. (2006) and Eilers et al. (2006), for a fast and efficient implementation of our proposal in standard software as, for example, R or MATLAB©.

The rest of this paper is organized as follows. In Section 2, we present our proposal: the penalized composite link mixed model (or more briefly, PCLMM) for spatially aggregated data, and we indicate how the spatial disaggregation cases, which we discussed above, are accommodated by our model. Also, we provide in this section a parameter estimation approach and GLAM algorithms for the PCLMM. In Section 3, we illustrate our methodology for the case of aggregated count data with array structure, using American male death counts by respiratory diseases, and for ATA and ATP cases, using mortality data related with female deaths by cardiovascular diseases in the community of Madrid recorded over the period 2001-2007. Also, in this section, we compare the performance of our proposal (for the ATP case) with the ATP Poisson kriging of Goovaerts (2006a). For that purpose, we use age-adjusted lung cancer mortality rates for white females in the state of Indiana recorded over the period 1970-1994. Finally, we end up with a short discussion in Section 4.

2 The penalized composite link mixed model

2.1 The model

In the one-dimensional case, suppose that a vector of aggregated counts $\boldsymbol{y}$ follows a Poisson distribution with mean vector $\boldsymbol{\mu}$. These counts can be seen as indirect observations of a latent process that we want to model. The penalized composite link model (PCLM) approach of Eilers (2007) offers an elegant way to do this, by considering $\boldsymbol{\mu}$ composed by latent expectations. The Poisson PCLM is given by:

$$\boldsymbol{\mu} = \mathbf{C}\gamma = \mathbf{C}\exp(\mathbf{B}\boldsymbol{\theta}), \tag{2.1}$$

where γ represents the mean vector of the latent process at a desirable fine resolution, $\mathbf{C}$ is the composition matrix that describes how these latent expectations are combined to yield $\boldsymbol{\mu}$, $\mathbf{B} = \mathbf{B}(\boldsymbol{x})$ is a B-spline basis constructed from a covariate, $\boldsymbol{x}$, at fine resolution, and $\boldsymbol{\theta}$ is the associated vector of regression coefficients. Smoothness is imposed over adjacent regression coefficients, by subtracting a roughness penalty $\frac{1}{2}\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$ from the log-likelihood of $\boldsymbol{y}$, where $\mathbf{P} = \lambda\mathbf{D}'\mathbf{D}$ is based on a difference matrix $\mathbf{D}$ of order d, and a non-negative parameter λ that controls the amount of smoothness. The estimation of the model (2.1) is carried out by a penalized version of the iteratively reweighted least squares (IRWLS) algorithm, where an information criterion (such as AIC and BIC) is used to choose an optimal value for λ. Several applications of the PCLM can be found in Eilers (2012).

For illustration purposes, consider the death counts of Danish females in 2006, from ages 1 to 100 (these data was taken from the Human Mortality Database, 2015). In Figure 1, the histogram reports these death counts as totals of five-year age classes (that is, the aggregated counts $\boldsymbol{y}$), while the points depict these death counts for each one-year age class. In this case, $\boldsymbol{x} = (1, ..., 100)'$, γ

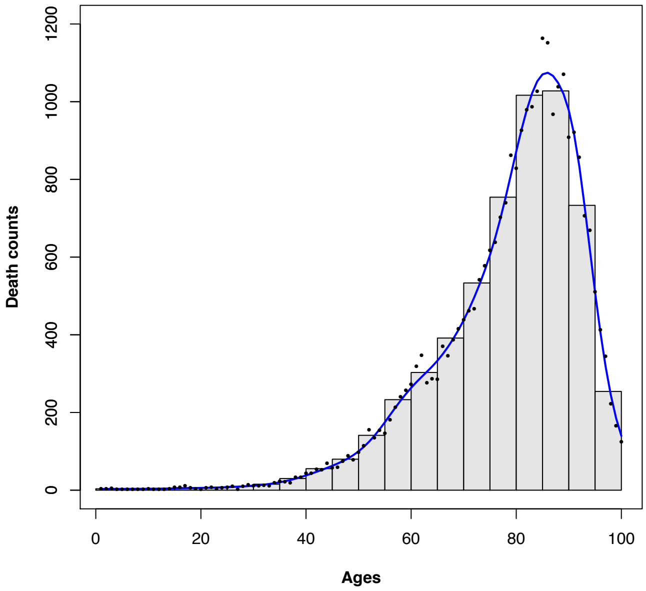

Figure 1: *Death counts of Danish females in* 2006*, from ages* 1 *to* 100*, represented as totals of five-year age classes (histogram) and as one-year age classes (points). The solid blue line represents the estimated distribution using the Poisson PCLM approach (BIC was used here as the optimal selection criterion for λ).*

represents the vector of expected numbers in one-year age classes and has 100 elements. The expected mean vector $\boldsymbol{\mu}$ has 20 elements, while the composition matrix $\mathbf{C}$ has 20 rows and 100 columns. Most of the elements of $\mathbf{C}$ are zeroes, but in the first row we find a 1 in columns 1 to 5, in the second row we find a 1 in columns 6 to 10, and so on. If we apply the Poisson PCLM approach to this aggregated count data $\boldsymbol{y}$, we obtain the blue smooth curve in Figure 1. We observe that this curve follows the trend indicated by the points, reflecting graphically the potential of the model (2.1) when we want to estimate the underlying distribution behind aggregated count data.

Now, in a two-dimensional setting, the aggregated counts $\boldsymbol{y}$ can be classified into two categories: (i) as areal or regional data (that is, they are available over n non-overlapping geographical units) or (ii) as array data (that is, they are recorded in a coarse grid of values as, for example, mortality tables classified by age and year classes with different levels of aggregation). Since our goal is to estimate the underlying mortality trend behind these aggregated data, we extend the PCLM (2.1) to the two-dimensional setting, and representing it as a mixed model, as follows.

First, suppose that $\boldsymbol{y}$ are classified as areal data. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the geographical coordinates (longitude and latitude, respectively) of length m that define the desirable fine spatial resolution. Then, in this new context, the regression basis $\mathbf{B}$ is defined as the Box-product or "row-wise" Kronecker product (Eilers et al., 2006) of the marginal B-spline bases $\mathbf{B}_1 = \mathbf{B}(\boldsymbol{x}_1)$ and $\mathbf{B}_2 = \mathbf{B}(\boldsymbol{x}_2)$

(of dimension $m \times c_1$ and $m \times c_2$, respectively), denoted by $\Box$:

$$\mathbf{B} = \mathbf{B}_2 \Box \mathbf{B}_1 = (\mathbf{B}_2 \otimes \mathbf{1}'_{c_1}) \odot (\mathbf{1}'_{c_2} \otimes \mathbf{B}_1), \tag{2.2}$$

where the matrix operators $\otimes$ and $\odot$ represent the Kronecker and the Hadamard (or "element-wise") products, respectively. When $\boldsymbol{y}$ are classified as array data, the adequate regression basis $\mathbf{B}$ correspond to,

$$\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1, \tag{2.3}$$

where now the B-spline bases $\mathbf{B}_1$ and $\mathbf{B}_2$ are constructed from two covariates, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, of lengths m_1 and m_2, respectively, each of one at a desirable fine resolution. In either case, the construction of $\mathbf{B}_1$ and $\mathbf{B}_2$ depend on the number of selected (equally spaced) knots for each coordinate, ndx_1 and ndx_2, and the degree of the B-splines used, $bdeg_1$ and $bdeg_2$. The two-dimensional penalty matrix is given by:

$$\mathbf{P} = \lambda_1 \mathbf{I}_{c_2} \otimes \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 \otimes \mathbf{I}_{c_1}, \tag{2.4}$$

where $\mathbf{P}_i = \mathbf{D}'_i \mathbf{D}_i$ is the marginal penalty matrix based on the difference matrix $\mathbf{D}_i$ of order d_i, for $i = 1, 2$. The penalty matrix (2.4) allows for anisotropy (i.e., different amount of smoothing for each dimension) and is valid whether we are dealing with areal or array data, since its definition is independent of data structure. Note here that we have to make choices about ndx_i, $bdeg_i$, and d_i, with $i = 1, 2$. For ndx_1 and ndx_2, it is enough to choose a moderate number of knots that cover the study area (up to a maximum of about 40 knots as suggested by Ruppert, 2002), and for the other quantities is often sufficient to use cubic B-splines (that is, $bdeg_1 = bdeg_2 = 3$) and quadratic penalties ($d_1 = d_2 = 2$) (see Eilers and Marx, 1996, and Currie and Durbán, 2002, for a further discussion).

Considering the regression basis (2.2) for areal data (or (2.3) for array data) and its associated regression coefficients $\boldsymbol{\theta}$, it was shown in Lee and Durbán (2009) that the expression $\mathbf{B}\boldsymbol{\theta}$ can be reformulated as $\mathbf{B}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$, using a suitable orthogonal transformation matrix $\mathbf{T}$ such that $\mathbf{BT} = [\mathbf{X} : \mathbf{Z}]$ and $\mathbf{T}'\boldsymbol{\theta} = \left[\begin{smallmatrix}\boldsymbol{\beta}\\ \boldsymbol{\alpha}\end{smallmatrix}\right]$, where $\mathbf{X}$ and $\mathbf{Z}$ are the fixed and random effects matrices, and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are their associated coefficients, respectively. The construction of these mixed model matrices, $\mathbf{X}$ and $\mathbf{Z}$, is described below (for more details, see Lee and Durbán, 2009 and Lee, 2010, pp. 63-65).

Consider the singular value decomposition (SVD) of the marginal penalty matrix $\mathbf{P}_i$ in (2.4),

$$\mathbf{P}_i = \mathbf{U}_i \boldsymbol{\Sigma}_i \mathbf{U}'_i,$$

where $\mathbf{U}_i$ is the matrix of eigenvectors, and $\boldsymbol{\Sigma}_i$ is a diagonal matrix that contains the eigenvalues of the SVD of $\mathbf{P}_i$, for $i = 1, 2$. Each matrix $\mathbf{U}_i$ can be splitted in two parts:

$$\mathbf{U}_i = [\mathbf{U}_{in} : \mathbf{U}_{is}],$$

where $\mathbf{U}_{in}$ contains the null part (of dimension $c_i \times d_i$) and $\mathbf{U}_{is}$ contains the non-null part of the decomposition (of dimension $c_i \times (c_i - d_i)$), for $i = 1, 2$. With these partitions, we can decompose each marginal penalty matrix as follows:

$$\mathbf{P}_i = [\mathbf{U}_{in} : \mathbf{U}_{is}] \begin{bmatrix} \mathbf{O}_{d_i} & \\ & \tilde{\mathbf{\Sigma}}_i \end{bmatrix} [\mathbf{U}_{in} : \mathbf{U}_{is}]',$$

where $\mathbf{O}_{d_i}$ denotes a $d_i \times d_i$ square matrix of zeroes, and $\tilde{\mathbf{\Sigma}}_i$ is a diagonal matrix that contains the $c_i - d_i$ positive eigenvalues, for $i = 1, 2$. Then, defining the matrices $\mathbf{X}_i = \mathbf{B}_i\mathbf{U}_{in}$ and $\mathbf{Z}_i = \mathbf{B}_i\mathbf{U}_{is}$, for $i = 1, 2$, the mixed model matrices for areal data are obtained as:

$$\mathbf{X} = \mathbf{X}_2 \Box \mathbf{X}_1, \tag{2.5}$$

$$\mathbf{Z} = [\mathbf{Z}_2 \Box \mathbf{X}_1 : \mathbf{X}_2 \Box \mathbf{Z}_1 : \mathbf{Z}_2 \Box \mathbf{Z}_1], \tag{2.6}$$

while the mixed model matrices for array data are obtained as in (2.5) and (2.6), replacing the "row-wise" Kronecker products □ by Kronecker products ⊗. Moreover, due to the transformation matrix $\mathbf{T}$ and the penalty matrix given in (2.4), it can be shown that the mixed model penalty is the block diagonal matrix:

$$\boldsymbol{F} = \begin{bmatrix} \lambda_2 \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{d_1} & & \\ & \lambda_1 \mathbf{I}_{d_2} \otimes \tilde{\mathbf{\Sigma}}_1 & \\ & & \lambda_1 \mathbf{I}_{c_2 - d_2} \otimes \tilde{\mathbf{\Sigma}}_1 + \lambda_2 \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{c_1 - d_1} \end{bmatrix}, \tag{2.7}$$

with matrices $\tilde{\mathbf{\Sigma}}_i$, $i = 1, 2$, previously defined.

Therefore, we can extend the model given in (2.1) by modifying γ as follows:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\boldsymbol{e}\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}), \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}\right), \tag{2.8}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the mixed model matrices defined above (depending on data structure), $\boldsymbol{e}$ is a vector of exposures at the fine resolution, and $\mathbf{G}$ is the covariance matrix of random effects given by $\mathbf{G} = \sigma^2_\epsilon \boldsymbol{F}^{-1}$, where $\sigma^2_\epsilon = 1$ (in Poisson case) and $\mathbf{F}$ is the penalty matrix defined in (2.7). We refer to (2.8) as the (Poisson) penalized composite link mixed model (or more briefly, PCLMM), which allows to incorporate population information at the fine spatial resolution and to include specific random effects or further correlation structure if necessary.

The vector $\boldsymbol{e}$ in (2.8) has to be known in advance; otherwise, it has to be estimated. If $\boldsymbol{e}$ is available at the aggregated level, we can obtain exposure estimates at the required disaggregated level assuming that these aggregated exposures are evenly distributed throughout the finer spatial resolution. The resulting disaggregated exposures have to sum up the same quantity on each coarse area that conform the coarse spatial resolution. We call these resulting estimates as naive estimates and we will use them in Section 3 for illustrative purpose.

The composition matrix $\mathbf{C}$ in (2.8) is fixed and its structure depends on the process that generates the aggregated data. Note that if we take $\mathbf{C}$ as the identity matrix, then $\boldsymbol{\mu} = \boldsymbol{\gamma}$ in (2.8). In such case, the PCLMM is reduced to the penalized generalized linear mixed model (or more briefly, PGLMM) approach of Lee and Durbán (2009) for Poisson data. On the other hand, when we are dealing with array data, the composition matrix $\mathbf{C}$ can be obtained as $\mathbf{C} = \mathbf{C}_2 \otimes \mathbf{C}_1$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are the composition matrices associated with the (disaggregated) covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Finally, in Section 1 we pointed out that the PCLMM approach can handle two types of spatial disaggregation of interest: area-to-area and area-to-point cases. For the former case, we can choose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as the coordinates of the centroids of the smaller units, and thus the elements of the associated composition matrix $\mathbf{C}$ become:

$$c_{ij} = \begin{cases} 1 & \text{if } (x_{1j}, x_{2j}) \text{ belongs to unit } i \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

where $i = 1, ..., n$, and $j = 1, ..., m$. For the later case, the coordinates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ correspond to the dense grid points that fall inside coarse units, and may be known by the user in advance or not. The composition matrix in this case is constructed in a similar fashion as in (2.9).

2.2 Parameter estimation for PCLMM

Consider the joint density function of $\boldsymbol{y}$ in a PCLMM context, that is:

$$f(\boldsymbol{y}|\boldsymbol{\alpha}) = \exp\left\{\boldsymbol{y}' \log(\boldsymbol{\mu}) - \mathbf{1}'\boldsymbol{\mu} - \mathbf{1}' \log(\Gamma\left(\boldsymbol{y} + \mathbf{1}\right))\right\}, \tag{2.10}$$

where $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$, $\boldsymbol{\gamma} = \boldsymbol{e} \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})$, and $\boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}\right)$. From (2.7), we see that $\mathbf{G} = \sigma_\epsilon^2 \boldsymbol{F}^{-1}$ depends on two smoothing parameters, λ_1 and λ_2, which are interpreted as ratio of variances that have to be estimated. In a mixed model framework, numerical integration techniques are usually demanded to evaluate (2.10) for a full likelihood analysis. To deal with this inconvenient, we use the penalized quasi-likelihood approach (PQL) of Breslow and Clayton (1993) that is described below.

Taking into account the joint density function in (2.10) and for given values of λ_1 and λ_2, we obtain estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by maximizing the penalized log-likelihood:

$$\ell_{pen} = \log\left\{f(\boldsymbol{y}|\boldsymbol{\alpha})\right\} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\boldsymbol{\alpha}. \tag{2.11}$$

Differentiating (2.11) with respect to β_k and α_l, we obtain:

$$\begin{aligned} \frac{\partial \ell_{pen}}{\partial \beta_k} &= \sum_{i=1}^{n}\left((y_i-\mu_i)\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}x_{jk}\gamma_j\right),\ k=1,...,p; && (2.12)\\ \frac{\partial \ell_{pen}}{\partial \alpha_l} &= \sum_{i=1}^{n}\left((y_i-\mu_i)\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}z_{jl}\gamma_j\right) - \mathbf{G}_i^{-1}\boldsymbol{\alpha},\ l=1,...,r, && (2.13)\end{aligned}$$

where $\mathbf{G}_i^{-1}$ denotes the ith row of matrix $\mathbf{G}^{-1}$. Writing the terms $\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}x_{jk}\gamma_j$ in (2.12) and $\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}z_{jl}\gamma_j$ in (2.13) as $\breve{x}_{ik}$ and $\breve{z}_{il}$, respectively, and equating the expressions above to zero, we obtain:

$$\begin{aligned} \sum_{i=1}^{n}(y_i-\mu_i)\breve{x}_{ik} &= 0,\ \text{for } k=1,...,p; && (2.14)\\ \sum_{i=1}^{n}(y_i-\mu_i)\breve{z}_{il} &= \mathbf{G}_i^{-1}\boldsymbol{\alpha},\ \text{for } l=1,...,r. && (2.15)\end{aligned}$$

Moreover, (2.14) and (2.15) can be rewritten, in matrix form, as:

$$\begin{aligned} \breve{\mathbf{X}}'(\boldsymbol{y}-\boldsymbol{\mu}) &= \mathbf{0}; && (2.16)\\ \breve{\mathbf{Z}}'(\boldsymbol{y}-\boldsymbol{\mu}) &= \mathbf{G}^{-1}\boldsymbol{\alpha}, && (2.17)\end{aligned}$$

where $\breve{\mathbf{X}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\breve{\mathbf{Z}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$, with $\mathbf{W} = diag(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = diag(\boldsymbol{\gamma})$. Defining the working vector:

$$\boldsymbol{z} = \breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{Z}}\boldsymbol{\alpha} + \mathbf{W}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{2.18}$$

the solution of (2.16) and (2.17) via Fisher scoring algorithm (see Green, 1987) can be expressed as the iterative solution of the system:

$$\begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z} \\ \breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z} \end{bmatrix}. \tag{2.19}$$

This yields to a modified version of the standard mixed model estimators:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}, && (2.20)\\ \widehat{\boldsymbol{\alpha}} &= \mathbf{G}\breve{\mathbf{Z}}'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}), && (2.21)\end{aligned}$$

where:

$$\mathbf{V} = \mathbf{W}^{-1} + \breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'. \tag{2.22}$$

Conditioning on the estimates obtained in (2.20) and (2.21), the smoothing parameters λ_1 and λ_2 can be estimated by maximizing the residual maximum log-likelihood (REML):

$$-\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}| - \frac{1}{2}\boldsymbol{z}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1})\boldsymbol{z}. \tag{2.23}$$

Therefore, the PQL solution is achieved by iteration between (2.20), (2.21), and (2.23), until convergence. The following useful equivalences can be used in the iterative procedure (Searle et al., 1992, p. 453):

$$\begin{aligned} |\mathbf{V}| &= |\mathbf{W}|^{-1}|\mathbf{G}|\,|\mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}|, & (2.24)\\ \mathbf{V}^{-1} &= \mathbf{W} - \mathbf{W}\breve{\mathbf{Z}}(\mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}. & (2.25)\end{aligned}$$

Finally, it is possible to approximate the covariance matrix of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ by its Bayesian counterpart. Following the work of Lin and Zhang (1999), the approximate covariance matrix of $\left[\begin{smallmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\alpha}}\end{smallmatrix}\right]$ is given by

$$\mathbf{M} = \begin{bmatrix}\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}\\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} + \mathbf{G}^{-1}\end{bmatrix}^{-1}, \tag{2.26}$$

which corresponds to the inverse of the matrix on the left-hand side of equation (2.19). Therefore, we can obtain approximate standard errors for $\widehat{\boldsymbol{\eta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}$, using the square root of diagonal elements of $\mathbb{V}ar(\widehat{\boldsymbol{\eta}})$, which are obtained as:

$$\mathbb{V}ar(\widehat{\eta}_i) = diag([\mathbf{X}:\mathbf{Z}]\mathbf{M}[\mathbf{X}:\mathbf{Z}]')_{ii}, \tag{2.27}$$

with $\mathbf{M}$ defined in (2.26).

2.2.1 Array methods for PCLMM

When we are dealing with the estimation of the underlying distribution in several dimensions, we are susceptible to present runaway problems with storage and computational time. In the case of data arranged in multidimensional grids, it is possible to circumvent these problems using the generalized linear array model (or GLAM) algorithms developed by Currie et al. (2006) and Eilers et al. (2006). In this section, we show the use of GLAM algorithms in the PCLMM context, when the aggregated data have array structure. In Section 2.2, we proposed the use of the restricted (or residual) maximum log-likelihood (REML) for the estimation of the variance parameters. Given (2.23) and the definitions of $\mathbf{V}$, $|\mathbf{V}|$ and $\mathbf{V}^{-1}$ in (2.22), (2.24), and (2.25), we may use the GLAM algorithms for a fast and efficient computation of the matrix cross-products: $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z}$, $\breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z}$, etc., and estimate the variance components by REML.

To illustrate the implementation of the GLAM algorithms, we divide the REML in four parts as:

$$-\frac{1}{2}\underbrace{\log|\mathbf{V}|}_{\text{part I}}-\frac{1}{2}\underbrace{\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}_{\text{part II}}-\frac{1}{2}(\underbrace{\boldsymbol{z}'\mathbf{V}^{-1}\boldsymbol{z}}_{\text{part III}}-\underbrace{\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}}_{\text{part IV}}).$$

Here, we use some GLAM notation and definitions proposed by Currie et al. (2006) and Eilers et al. (2006), as for example, the row tensor of two matrices, $\mathcal{G}$, and the rotated $\mathcal{H}$-transform of an array by a matrix, ρ (for their definitions, see Appendix).

Part I: Array computation of $\log|\mathbf{V}|$

Given the covariance matrix $\mathbf{G}=\sigma_\epsilon^2\boldsymbol{F}^{-1}$, with $\sigma_\epsilon^2=1$ (Poisson case) and $\mathbf{F}$ defined in (2.7), and considering (2.24), the term $\log|\mathbf{V}|$ can be written as:

$$\log|\mathbf{V}|=-\log|\mathbf{W}|+\log|\mathbf{F}^{-1}|+\log|\mathbf{F}+\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}|. \tag{2.28}$$

Since $\mathbf{W}$ is a diagonal matrix and $\mathbf{F}$ is a block-diagonal matrix, the first two terms in (2.28) are calculated as $-\log|\mathbf{W}|=-\sum\log(\mu_i)$ and $\log|\mathbf{F}^{-1}|=-\sum\log(\mathbf{F}_{ii})$, where $\mathbf{F}_{ii}$ denote the diagonal elements of $\mathbf{F}$.

For the computation of $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$ in (2.28), note that we can reduce this expression as:

$$\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}=(\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}). \tag{2.29}$$

Since the composition matrix $\mathbf{C}$ is given by $\mathbf{C}=\mathbf{C}_2\otimes\mathbf{C}_1$ and the model matrix $\mathbf{Z}$ can be rewritten as $\mathbf{Z}=[\mathbf{Z}_2\otimes\mathbf{X}_1:\tilde{\mathbf{Z}}_2\otimes\mathbf{Z}_1]$, where $\tilde{\mathbf{Z}}_2=\mathbf{X}_2\otimes\mathbf{Z}_2$, the product of matrices $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$ in (2.29) can be computed as:

$$\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}\equiv[\rho(\mathcal{G}(\mathbf{Z}_2,\mathbf{C}_2')',\rho(\mathcal{G}(\mathbf{X}_1,\mathbf{C}_1')',\tilde{\boldsymbol{\Gamma}})):\rho(\mathcal{G}(\tilde{\mathbf{Z}}_2,\mathbf{C}_2')',\rho(\mathcal{G}(\mathbf{Z}_1,\mathbf{C}_1')',\tilde{\boldsymbol{\Gamma}}))], \tag{2.30}$$

where $\tilde{\boldsymbol{\Gamma}}$ is a matrix of dimension $m_1\times m_2$, whose entries are the elements of the diagonal of $\boldsymbol{\Gamma}$, that is, $\tilde{\boldsymbol{\Gamma}}$ is an arrangement of the vector $\boldsymbol{\gamma}$.

Part II: Array computation of $\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|$

Using the equivalence (2.25), we can rewrite $\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}$ as:

$$\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}=\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}-\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}(\mathbf{F}+\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}}. \tag{2.31}$$

Since $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$ was calculated previously and $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} = (\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}})'$, we only need to compute the expressions $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}$ and $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$ in (2.31). Note that we can reduce them as:

$$\begin{aligned} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} &= (\mathbf{C\Gamma X})'\mathbf{W}^{-1}(\mathbf{C\Gamma X}), && (2.32)\\ \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} &= (\mathbf{C\Gamma X})'\mathbf{W}^{-1}(\mathbf{C\Gamma Z}). && (2.33) \end{aligned}$$

The expression $\mathbf{C\Gamma Z}$ was calculated in (2.30). Considering the model matrix $\mathbf{X} = \mathbf{X}_2 \otimes \mathbf{X}_1$, the expression $\mathbf{C\Gamma X}$, which appears in (2.32) and (2.33), can be computed as:

$$\mathbf{C\Gamma X} \equiv \rho(\mathcal{G}(\mathbf{X}_2, \mathbf{C}_2')', \rho(\mathcal{G}(\mathbf{X}_1, \mathbf{C}_1')', \tilde{\mathbf{\Gamma}})), \qquad (2.34)$$

with $\tilde{\mathbf{\Gamma}}$ defined above.

Part III: Array computation of $\boldsymbol{z}'\mathbf{V}^{-1}\boldsymbol{z}$

Given (2.25), we can write $\boldsymbol{z}'\mathbf{V}^{-1}\boldsymbol{z}$ as:

$$\boldsymbol{z}'\mathbf{V}^{-1}\boldsymbol{z} = \boldsymbol{z}'\mathbf{W}\boldsymbol{z} - \boldsymbol{z}'\mathbf{W}\breve{\mathbf{Z}}(\mathbf{F} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z}, \qquad (2.35)$$

where $\boldsymbol{z}'\mathbf{W}\boldsymbol{z}$ is calculated as $\sum \mu_i z_i^2$. We can rewrite the expression $\boldsymbol{z}'\mathbf{W}\breve{\mathbf{Z}}$ in (2.35) as:

$$\boldsymbol{z}'\mathbf{W}\breve{\mathbf{Z}} = \boldsymbol{z}'\mathbf{C\Gamma Z},$$

where $\mathbf{C\Gamma Z}$ was calculated in (2.30).

Part IV: Array computation of $\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}$

We already shown how to compute $\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}$ in (2.31). Thus, we only need to compute $\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}}$ (since $\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z} = (\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}})'$). Given (2.25), we can write $\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}}$ as:

$$\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}} = \boldsymbol{z}'\mathbf{W}\breve{\mathbf{X}} - \boldsymbol{z}'\mathbf{W}\breve{\mathbf{Z}}(\mathbf{F} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}\mathbf{X}, \qquad (2.36)$$

where all the quantities were computed previously, except to $\boldsymbol{z}'\mathbf{W}\breve{\mathbf{X}}$, which is computed as:

$$\boldsymbol{z}'\mathbf{W}\breve{\mathbf{X}} = \boldsymbol{z}'\mathbf{C\Gamma X},$$

where $\mathbf{C\Gamma X}$ was calculated in (2.34).

3 Applications

In this section, we apply our methodology to three mortality datasets. The first one is related with American male deaths (indexed by age and year at death) and we will use them to illustrate how the PCLMM approach is applied when the data have array structure. The second dataset comes from a large European epidemiological project called MEDEA (see http://www.proyectomedea.org), whose aim was to study the impact of socio-economic and environmental inequalities on the mortality rates by different causes. Deaths are not only related to individual factors, but also to contextual factors, most of them related to the area of residence. Therefore, it is of great interest to estimate spatial trends present in the data that can help to identify areas that may need intervention. We will use this dataset to illustrate how our proposal is applied in the area-to-area and area-to-point cases. Finally, the third dataset is part of the Atlas of Cancer Mortality in the United States (Pickle et al., 1999) and was downloaded from http://ratecalc.cancer.gov. This dataset was analysed by Goovaerts (2006a) and we will use them to compare the performance of our proposal with his in the area-to-point case.

3.1 Deaths by respiratory diseases

Consider the death counts by respiratory diseases of American males from ages 1 to 100, and from 1959 to 1998 (for more details about this data, see Currie et al., 2006). These raw data are displayed in the left panel of Figure 2. Suppose that we observe aggregated death counts, recorded in five-year age and four-year classes, instead of the previous raw data. The middle panel of Figure 2 shows the bivariate histogram for these aggregated counts, which is conformed by 200 classes (that is, the resulting product of the 20 age groups and 10 year groups).

In order to estimate the underlying distribution behind these aggregated data, we apply the PCLMM approach (for array data) described in the previous section. In this case, $\boldsymbol{x}_1 = (1, ..., 100)'$ and $\boldsymbol{x}_2 = (1959, ..., 1998)'$ are the vectors of ages and years at fine resolution, and $\mathbf{C} = \mathbf{C}_2 \otimes \mathbf{C}_1$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are the (marginal) composition matrices for ages and years, of dimensions 20×100 and 10×40, respectively. The right panel of Figure 2 shows the smoothed bivariate distribution obtained by the PCLMM approach, choosing $ndx1 = 25$ and $ndx2 = 10$ as the number of equally-spaced knots for each dimension. We observe that the smoothed distribution closely follows the bivariate trend displayed by the original raw data. This is due in part by the levels of aggregation of each dimension. In general, the smoothed PCLMM distribution will lose precision,

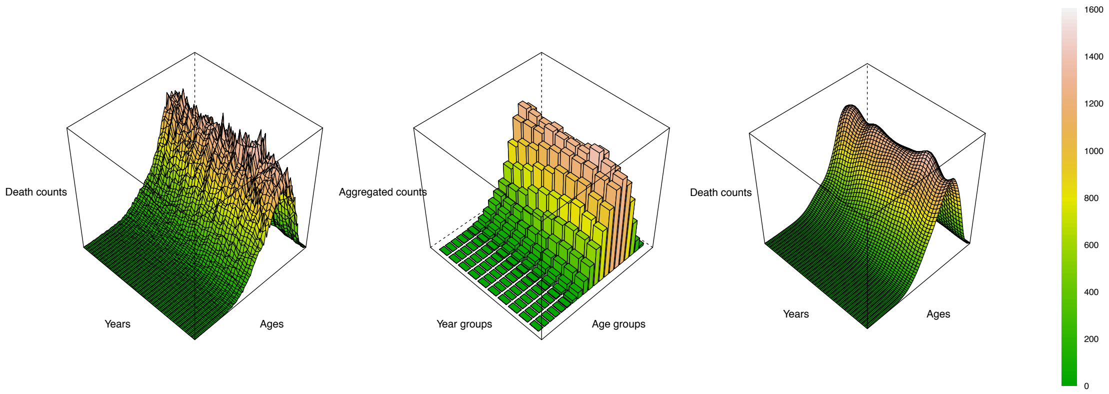

Figure 2: *American male deaths by respiratory diseases during the period* $1959-1998$*, from ages* 1 *to* 100*. The left and middle panels represent these deaths as totals of one-year age/one-year classes and five-year age/four-year classes respectively. The right panel shows the estimated distribution using Poisson PCLMM approach for array data.*

if we observe wide classes at the edge of the histogram. Since we are only considering death counts, a way to improve the description of the mortality is considering a vector of exposures. In the following examples, we will include these values in the analysis and we will see how to deal with them when they area only available at the aggregated level.

Considering the array methods described in Section 2.2.1 into the iterative procedure (to obtain the estimated distribution in the right panel of Figure 2), the resulting computing time took about 61.840 seconds (Intel® Core$^{\text{TM}}$ $i7$, 1.80 GHz, Windows 8.1). On the other hand, if we disregard the use of these array methods, the computing time took about 221.360 seconds; that is, for this case, the computing time was reduced in about 3.6 times. This shows the usefulness of the adapted GLAM algorithms developed here, for PCLMM estimation, in terms of computational speed.

3.2 Deaths by cardiovascular diseases

Our data correspond to the number of observed and expected female deaths by cardiovascular diseases in the community of Madrid, Spain, over the period 2001-2007, which are available at different (aggregated) spatial levels. The left panel in Figure 3 shows the spatial distribution of the (raw) natural logarithm of standardized mortality rates (denoted by log(SMR)) for 179 municipalities of this community. We use a sequential map color scheme with 10 equally-weighted classes (that is, the class boundaries correspond to deciles of the raw log(SMR)) to readily identify which values are higher or lower than others on the mortality map (Brewer, 1999). For 2001-2007

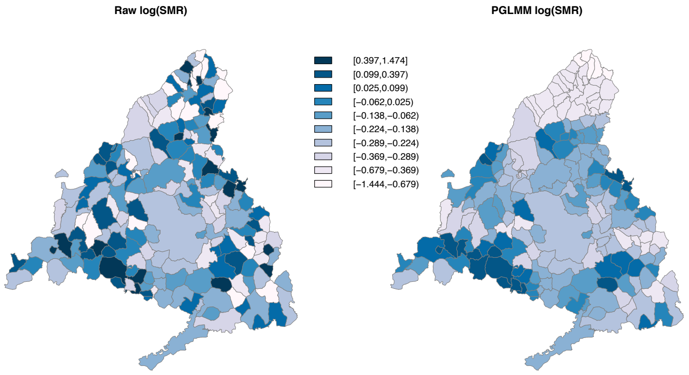

Figure 3: *Spatial distribution of raw log(SMR) for* 179 *municipalities in the community of Madrid over the period* 2001*-*2007 *(left panel) and the resulting smoothed log(SMR) by applying the PGLMM approach to raw data (right panel). The color legend applies to both maps, where the class boundaries correspond to the deciles of raw log(SMR).*

period, the number of expected deaths ranges from 0.916 to 44715.610 over these municipalities, while the number of observed deaths varies from 0 to 34884.

Since raw log(SMR) vary abruptly between municipalities, we apply the PGLMM approach of Lee and Durbán (2009) to enhance the visibility of underlying trends. The right panel in Figure 3 shows the estimated spatial trend obtained from PGLMM approach (using $ndx_1 = ndx_2 = 20$ equally spaced knots), where the coordinates of the centroids of the municipalities are used as covariates. We observe that most of the higher rates are in the boundaries of the community of Madrid, specially in the south-western area. They correspond to areas with difficult access to health facilities, or industrialized areas where environmental conditions are poor.

3.2.1 Area-to-area case

Now, suppose that we seek to visualize the spatial distribution of log(SMR) at census tract level, assuming that we only have mortality data at municipality level. The total number of census tracts for the community of Madrid is 3906. To estimate the desired spatial distribution, we use the model given in (2.8) for the area-to-area case (ATA-PCLMM), were we must consider the exposures (the number of expected deaths, in this case) at census tract level. For this dataset, we in fact have these quantities, which we denote as $\boldsymbol{e}_{\text{true}}$; otherwise the user has to estimate them in advance. A naive way to do this is to assume that the exposures are evenly distributed throughout the census tracts at each municipality. We denote these resulting estimates as $\boldsymbol{e}_{\text{naive}}$, and we will use them for comparison purpose. The top-left and top-right maps in Figure 4 show

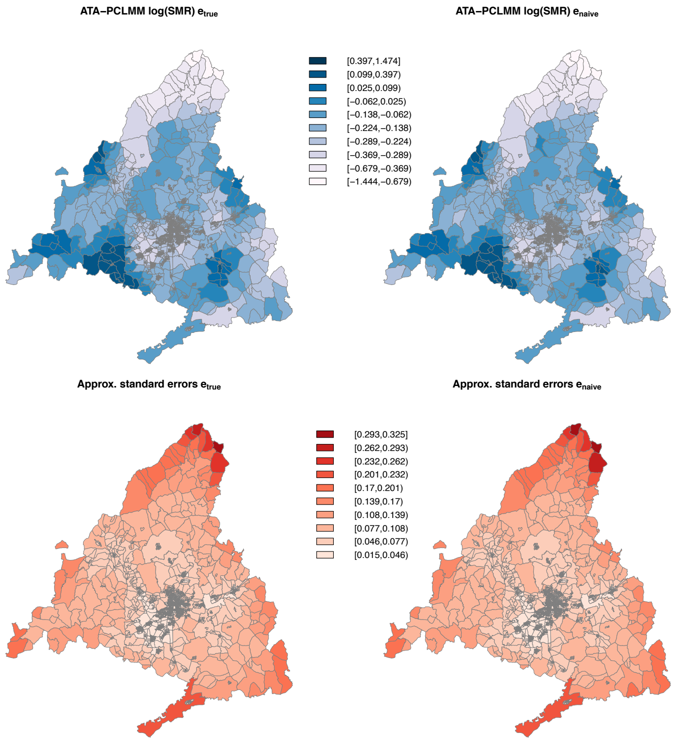

Figure 4: *Smoothed log(SMR) and their approximate standard errors at census tract level, using ATA-PCLMM with the true number of expected deaths at census tract level (top-left) and its naive estimator (top-right). The color legend applies to all the maps that show the same quantity; the class boundaries for smoothed log(SMR) correspond to the deciles of raw log(SMR) at municipality level, and the class boundaries for standard errors correspond to the cuts of the range of all errors in ten equal parts.*

the resulting smoothed log(SMR) at census tract level, using ATA-PCLMM approach ($ndx_1 = ndx_2 = 20$ equally spaced knots) with $\boldsymbol{e}_{\text{true}}$ and $\boldsymbol{e}_{\text{naive}}$, respectively. These maps have a similar spatial distribution and are consistent with the smoothed trend obtained at municipality level (right panel of Figure 3). The approximate standard errors for these smoothed ATA-PCLMM log(SMR) were obtained using (2.27), and are displayed at the bottom of each mortality map in Figure 4. For comparison purpose, we select the class boundaries for these maps as the cuts of the range of all errors in ten equal parts. We observe that these error maps are very similar and both present high values in the northern area of the community of Madrid. The later is due the fact that both smoothed maps are unable to capture more precise mortality trends over the census tracts in this part of the map, where we have less information.

It is clear that the municipalities of the community of Madrid vary greatly in shape and size, especially when we compare the municipality of Madrid (which is located at the center of the community) with the rest of them. Figure 5 displays the district and census tract boundaries for this municipality, which was conformed by 21 districts and 2358 census tracts, and the spatial distribution of raw log(SMR) at district level. The zoom in on the center of this municipality provides a more detailed geographical distribution of the census tracts. Suppose that we only have the number of observed female deaths by cardiovascular diseases for each district in the municipality of Madrid, and we want to estimate mortality rates for the selected districts at census tract level, using the additional information of the number of expected deaths at this level. The ATA-PCLMM approach fits to this situation, in which we want to move from district to census tracts. Figure 6 shows the resulting smoothed log(SMR) using ATA-PCLMM approach ($ndx_1 = ndx_2 = 20$ equally spaced knots) with the true vector of exposures. For the area of interest, we observe a more detailed spatial distribution of mortality, where the highest log(SMR) are mostly concentrated around Madrid Centro.

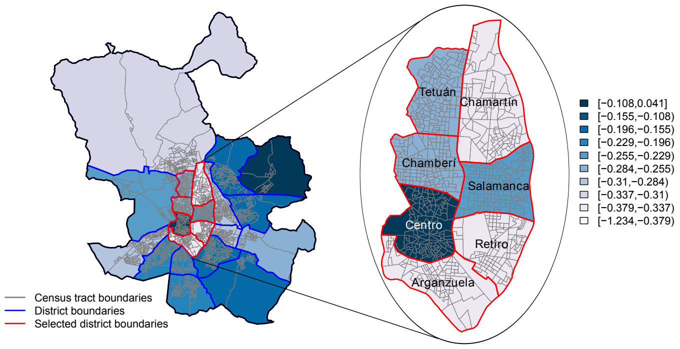

Figure 5: *Spatial distribution of raw log(SMR) for the 21 districts in the municipality of Madrid. The zoom shows 7 centric districts of interest and their 780 census tracts. The class boundaries correspond to the deciles of raw log(SMR).*

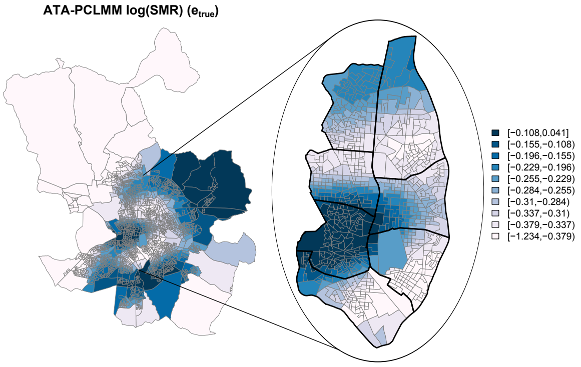

Figure 6: *Smoothed log(SMR) using ATA-PCLMM approach with the true number of expected deaths at census tract level. The class boundaries for the smoothed log(SMR) correspond to the deciles of the raw log(SMR) at district level.*

3.2.2 Area-to-point case

To illustrate the area-to-point case, suppose that we want to create a continuous mortality trend across municipalities in the community of Madrid. We discretize this region by imposing a 100×100 fine grid over it, and we select the points that fall inside of each municipality (which leads to 4359 points). Figure 7 shows the map of the municipalities in the community of Madrid and the 100×100 grid chosen. Then we can use the model given in (2.8) for the area-to-point case (ATP-PCLMM), to produce such continuous trend. Due to the lack of expected deaths at this point-level, we estimate them using the naive approach described previously. The right map of Figure 8 shows the resulting smoothed log(SMR) using ATP-PCLMM with these naive estimates (where we use $ndx_1 = ndx_2 = 20$ equally spaced knots for the creation of B-spline bases), and the left map of Figure 8 displays the corresponding approximate standard errors. We observe that the ATP-PCLMM log(SMR) map gives more details than the previous ATA-PCLMM log(SMR) maps, but some of their associated standard errors are larger than the maximum of the ATA-PCLMM standard errors. These higher values are located at the boundary of the community of Madrid, as would be expected.

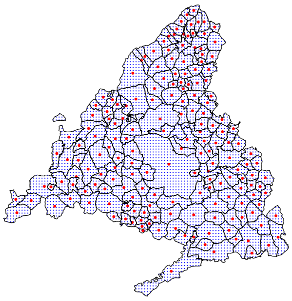

Figure 7: *Map of the community of Madrid. The red and blue points represent the centroids of the* 179 *municipalities and the* 4359 *grid points selected, respectively*

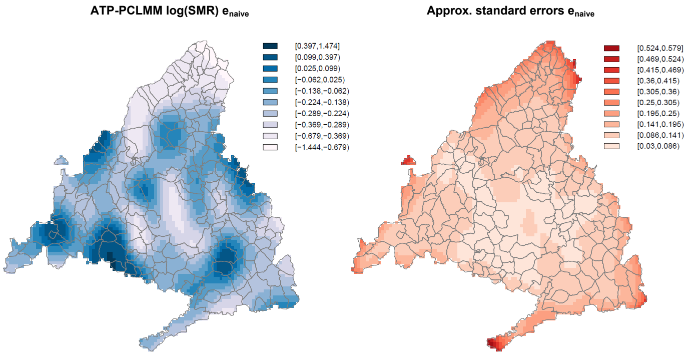

Figure 8: *Smoothed log(SMR) using ATA-PCLMM approach with the true number of expected deaths at census tract level. The class boundaries correspond to the deciles of these smoothed log(SMR).*

3.3 Deaths by lung cancer

This dataset contains the number of white female deaths by lung cancer and the corresponding age-adjusted mortality rates (per 100000 person-years), recorded over the period 1970-1994 in the state of Indiana, United States, at county level (92 counties). The population at risk in each county can be estimated as: $100000\times$ the total number of deaths over the period 1970-1994 divided by the corresponding age-adjusted mortality rate. Goovaerts (2006a) allocated these county-level population estimates (according to the 2000 census block level data) to a fine grid of 25 km^2 cells, leading to 3751 grid points. These high-resolution population estimates were kindly provided by Dr. Pierre Goovaerts (BioMedware Inc., MI, USA) and we will use them in subsequent analysis.

The top maps of Figure 9 shows the spatial distribution of (raw) age-adjusted lung cancer mortality rates of the Indiana county data previously described (left) and the smoothed mortality rates using the PGLMM approach with $ndx_1 = ndx_2 = 23$ equally-spaced knots. The class boundaries correspond to the deciles of the raw mortality rates and the color legend applies to all these maps. We observe that the highest lung cancer mortality rates are still observed in the central counties of Indiana after to apply the PGLMM approach. The first two rows of Table 1 summarize the corresponding statistics for the raw and smoothed PGLMM lung cancer mortality rates, respectively, which reflect numerically the increase of the minimum raw rates observed in a few north-western and north-eastern counties after to apply the PGLMM approach. This situation was pointed out by Goovaerts (2006a), when he analysed the Indiana county data (at county level) with different kriging methods.

Considering now the population at risk over the fine grid of 25 km^2 cells, we can apply the ATP-PCLMM approach on this dataset to obtain a continuous mortality risk map, thus eliminating the visual bias associated with the interpretation of the top-right choropleth map in Figure 9. The bottom-left map of Figure 9 shows the resulting ATP-PCLMM mortality rates, using $ndx_1 = ndx_2 = 23$ equally-spaced knots. These estimates are calculated as $\widehat{r}_{\text{PCLMM}}(\boldsymbol{x}_s) = 100000\exp(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}})$, where $\boldsymbol{x}_s = (\boldsymbol{x}_{1s}, \boldsymbol{x}_{2s})$, $s = 1, \ldots, 3751$, represent the points of the fine grid. This isopleth map shows more in detail the presence of delineated areas of lower and higher mortality rates.

In order to compare our proposal with other existing methods, we apply the area-to-point (or ATP) Poisson kriging approach of Goovaerts (2006a) on this dataset. The ATP kriging estimator is given by:

$$\widehat{r}_{\mathrm{PK}}(\boldsymbol{x}_s) = \sum_{i=1}^{K} \lambda_i(\boldsymbol{x}_s) r(\boldsymbol{v}_i),$$

where $\lambda_i(\boldsymbol{x}_s)$ is the weight assigned to each raw mortality rate $r(\boldsymbol{v}_i)$ at county $\boldsymbol{v}_i$. The K weights are computed by solving a system of linear equations, where a point-support covariance, or equivalently, a point-support semivariogram of the risk is needed. This function cannot be estimated directly from the observed rates, since only aggregated data are available. Goovaerts (2008) developed a procedure to conduct the derivation of the point-support semivariogram from the "regularized" experimental semivariogram computed from areal data (deconvolution process), in presence of irregular geographical units and heterogeneous population distribution.

The isopleth map at the bottom-right of Figure 9 shows the resulting ATP Poisson kriging mortality rates. This estimation was conducted by following the indications given in Goovaerts (2006a) and performed by using SpaceStat 4.0 software (`http://www.biomedware.com/`). Summary statistics for both ATP-PCLMM and ATP Poisson kriging estimates are displayed in Table 1, where we observe that the variance of ATP-PCLMM estimates is higher than the variance obtained with the ATP Poisson kriging. Note that both methods have similar minimum values, but some of their estimates exceed the maximum raw lung mortality rate (31.795 deaths/10000 habitants).

Table 1: *Summary statistics for county-level and point-level estimates of lung cancer mortality in Indiana over the period* 1970-1994.

Spatial level	Quantity	Mean	Variance	Min	Max
County	Raw rates	21.188	18.443	9.084	31.795
County	PGLMM rates	21.589	11.516	13.202	31.624
Point	ATP-PCLMM rates ($\boldsymbol{e}_{\mathrm{true}}$)	21.204	11.488	12.232	34.067
Point	ATP Poisson kriging rates	21.198	9.481	12.112	33.896

Mean, variance, minimum and maximum values for quantities related with age-adjusted mortality rates (per 100000 person-years) of white female deaths by lung cancer in Indiana at different spatial resolutions.

3.3.1 Simulation study

The isopleth maps at the bottom of Figure 9 illustrate different estimates for the latent or underlying spatial distribution of lung cancer mortality, which is unknown in practice. To assess the prediction performance of ATP-PCLMM and ATP Poisson kriging approaches, we conducted a

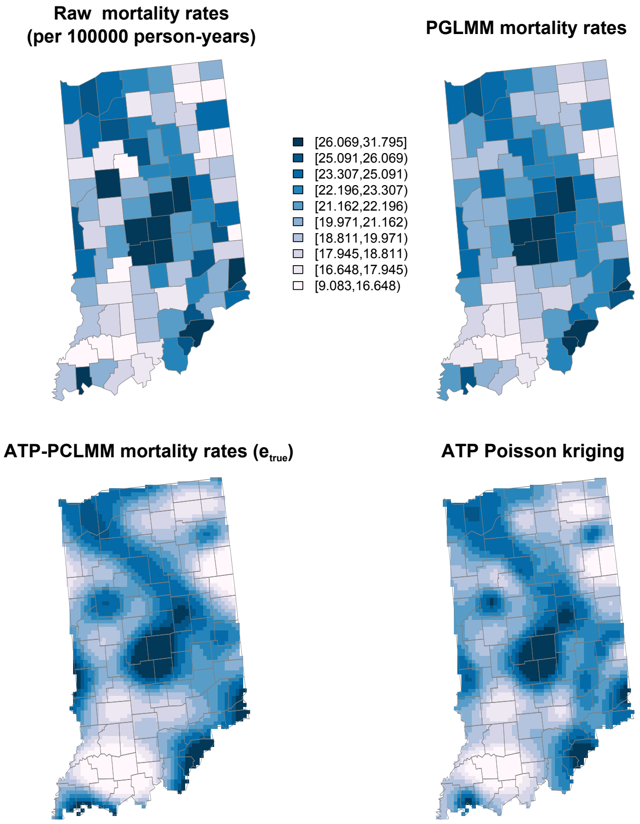

Figure 9: *Maps of age-adjusted lung cancer mortality rates in Indiana. The top-left map shows the raw lung cancer mortality rates per* 100000 *person-years recorded over the period* 1970-1994*, and the top-right map shows the resulting smoothed mortality rates by applying the PGLMM approach for raw data. The bottom maps shows the resulting smoothed mortality estimates using ATP-PCLMM and ATP Poisson kriging approaches (with the high-resolution population estimates), respectively. The color legend applies to all maps; the class boundaries correspond to the deciles of the raw mortality rates*

simulation study in the following way:

1. The continuous mortality surface obtained with the ATP Poisson kriging approach was considered here as the true underlying mortality trend over the fine grid of 25 km^2 cells in Indiana. We denoted these mortality rates as $r(\boldsymbol{u}_s)$, where $\boldsymbol{u}_s$, $s = 1, ..., 3751$, represent the points of the fine grid.

2. These quantities and the population at risk over the fine grid of 25 km^2 cells (denoted as

$e(\boldsymbol{u}_s)$) were used to calculate the mortality rate for each county $\boldsymbol{v}_\delta$, $\delta = 1, ..., 92$:

$$r(\boldsymbol{v}_\delta) = \frac{1}{e(\boldsymbol{v}_\delta)} \sum_{s=1}^{P_\delta} e(\boldsymbol{u}_s) r(\boldsymbol{u}_s),$$

where P_δ denotes the number of points $\boldsymbol{u}_s$ used to discretize the county $\boldsymbol{v}_\delta$, and $e(\boldsymbol{v}_\alpha) = \sum_{s=1}^{P_\delta} e(\boldsymbol{u}_s)$.

3. 100 realizations of the number of deaths recorded over each county were generated by random drawing of a Poisson distribution whose mean parameter is $r(\boldsymbol{v}_\delta) \times e(\boldsymbol{v}_\delta)$.

4. For each realization, we apply ATP-PCLMM and ATP Poisson kriging approaches, using the population at risk over the fine grid of 25 km^2 cells as the vector $\boldsymbol{e}$ of exposures at the fine resolution.

For all $l = 1, ..., 100$ realizations, the predicted risks $r_{\text{P}}^{(l)}(\boldsymbol{u}_s)$ obtained from both approaches were compared to the underlying risk $r(\boldsymbol{u}_s)$, $s = 1, ..., 3751$, using the following criteria:

- Mean error:

$$ME^{(l)} = \frac{1}{W} \sum_{s=1}^{3751} e(\boldsymbol{u}_s) \left[r_{\text{P}}^{(l)}(\boldsymbol{u}_s) - r(\boldsymbol{u}_s) \right] \text{ with } W = \sum_{s=1}^{3751} e(\boldsymbol{u}_s)$$

- Mean absolute error:

$$MAE^{(l)} = \frac{1}{W} \sum_{s=1}^{3751} e(\boldsymbol{u}_s) \left| r_{\text{P}}^{(l)}(\boldsymbol{u}_s) - r(\boldsymbol{u}_s) \right| \text{ with } W = \sum_{s=1}^{3751} e(\boldsymbol{u}_s)$$

- Mean squared error:

$$MSE^{(l)} = \frac{1}{3751} \sum_{s=1}^{3751} \left(r_{\text{P}}^{(l)}(\boldsymbol{u}_s) - r(\boldsymbol{u}_s) \right)^2$$

Note that both the mean and the absolute mean errors penalize more the errors that affect a larger population, and were used as criteria for the study of the performance of alternative smoothing techniques in Goovaerts (2005) and Goovaerts (2006a). Figure 10 shows these resulting errors via box-plots, in which we observe that our approach gives slightly more prediction accuracy than the ATP Poisson kriging, for each criterion. Table 2 shows the simulation results obtained on average over 100 realizations in this study.

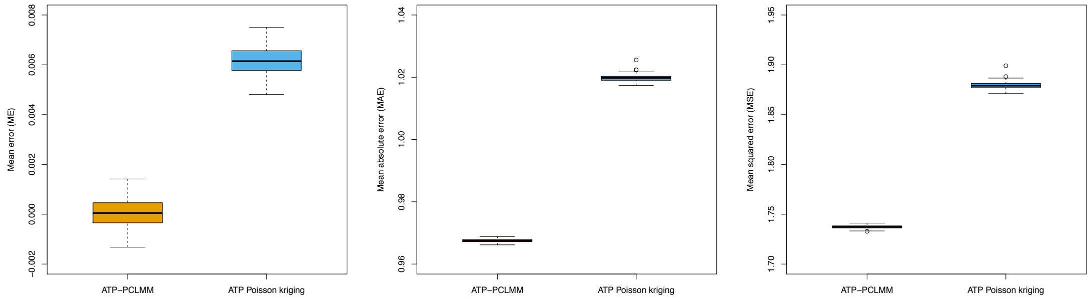

Figure 10: *Performance comparison between ATP-PCLMM and ATP Poisson kriging approaches: mean errors (left), mean absolute errors (middle), and mean squared errors (right) of predictions.*

Table 2: *Perfomance comparison of ATP-PCLMM and ATP Poisson kriging approaches.*

Model	Mean error	Mean absolute error	Mean squared error
ATP Poisson kriging	0.006	1.020	1.879
ATP-PCLMM ($\boldsymbol{e}_{\text{true}}$)	$5.661e^{-05}$	0.968	1.737

Results obtained on average over 100 *realizations generated for Indiana county data.*

4 Discussion

In this paper, the penalized composite link mixed model (PCLMM) for spatially aggregated data was developed and applied to the disaggregation of mortality rates. It provides a flexible descriptive tool for epidemiological studies, when the aim is to visualize the spatial distribution of certain rates at a desirable spatial resolution. The PCLMM approach filters the existing noise in raw rates, which is caused by the small number problem, and allows the creation of more refined mortality maps by including the distribution of the exposure variable at fine resolution. Moreover, the resulting PCLMM estimates may be linked with potential risks factors that are available over the fine resolution, allowing a posterior correlation analysis between them.

We used the statistical software R (R Core Team, 2014) for data analysis with the PCLMM approach. Our plan is to implement the presented methodology in a future R package, in such a way that it can be accessible by any user. Although we have omitted any kind of CPU time analysis, associated with computation of mortality trends at fine resolution for the applications presented in Sections 3.2 and 3.3, the computing time of our procedure in such cases are relatively short. Of course, this time will be reduced, if we disaggregate at spatial resolutions that are not so fine. A possibility to improve the computational speed is to consider the generalization of the Schall algorithm (Schall, 1991) presented by Rodríguez-Álvarez et al. (2015), into a PCLMM context. In the other hand, it might be possible to improve the estimation of the mortality estimates in the ATP case, considering new structures for the composition matrix **C**. An attempt could be the new composition matrix $\mathbf{C}^*$, whose entries are determined as the amount of area (measured between 0 and 1) that each grid cell shares with a specific geographical units.

We performed a simulation study to compare the area-to-point Poisson kriging of Goovaerts (2006a) with our proposal, using aggregated data measured over the 92 counties of the Indiana and the high-resolution population estimates over a fine grid. The simulation results showed that our proposal is competitive with respect to this geostatistical technique. However, further simulation studies should be done, especially for the case when the geographical units vary greatly in shape and size (for the state of Indiana, we have fairly similar counties).

Finally, the proposed methodology can be generalized to the spatio-temporal setting, in which the temporal dimension could be available only in aggregated form. In this context, the implementation of efficient and fast algorithms for the estimation procedure of PCLMMs will be critical. The resulting estimates will be displayed as dynamic maps, and will allow the comparison of mortality in the finest spatio-temporal resolution.

Acknowledgements

We thank again to Dr. Pierre Goovaerts who provided the high resolution population estimates that were used in Section 3.3. The research of Diego Ayma and María Durbán was supported by the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-02 and MTM2014-52184. The research of Dae-Jin Lee was also supported by the Basque Government through the BERC 2014-2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

References

Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. *Int J Health Geogr*, 3(18).

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Ann Inst Statist Math*, 43:1–59.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J Am Statist Assoc*, 88(421):9–25.

Brewer, C. A. (1999). Color use guidelines for data representation. In *Proceedings of The American Statistical Association's Section on Statistical Graphics*, pages 55–60.

Caudeville, J., Bonnard, R., Boudet, C., Denys, S., Govaert, G., and Cicolella, A. (2012). Development of a spatial stochastic multimedia model to assess population exposure at regional scale. *Sci Total Environ*, 432:297–308.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (revised edition)*. John Wiley & Sons, New York.

Currie, I. D. and Durbán, M. (2002). Flexible smoothing with P-splines: a unified approach. *Stat Modelling*, 4:333–349.

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J R Statist Soc Ser B*, 68(2):259–280.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Stat Sci*, 28(4):542–563.

Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Stat Model*, 7:239–254.

Eilers, P. H. C. (2012). Composite link, the neglected model. In Komárek, A. and Nagy, S., editors, *Proceedings of 27th International Workshop on Statistical Modelling*, pages 11–23, Prague, Czech Republic.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Comput Stat Data An*, 50(1):61–76.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat Sci*, 11(2):89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression: a Bayesian perspective. *Stat Sinica*, 14:731–761.

Goovaerts, P. (2005). Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *Int J Health Geogr*, 4(31).

Goovaerts, P. (2006a). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *Int J Health Geogr*, 5(52).

Goovaerts, P. (2006b). Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. *Int J Health Geogr*, 5(7).

Goovaerts, P. (2008). Kriging and semivariogram deconvolution in presence of irregular geographical units. *Math Geol*, 40(1):101–128.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int Stat Rev*, 55:245–259.

Human Mortality Database (2015). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on April 2015).

Kelsall, W. and Wakefield, J. (2002). Modelling spatial variation in disease risk: a geostatistical approach. *J Am Statist Assoc*, 97(459):692–701.

Lee, D.-J. (2010). *Smoothing mixed models for spatial and spatio-temporal data*. PhD thesis, Department of Statistics, Universidad Carlos III de Madrid, Spain.

Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Comput Stat Data An*, 53:2968–2979.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J Roy Statist Soc B*, 61(2):381–400.

MacNab, Y. C. and Dean, C. B. (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Stat Med*, 21:347–358.

Monestiez, P., Dubroca, L., Bonin, E., Durbec, J. P., and Guinet, C. (2005). Comparison of model based geostatistical methods in ecology: application to fin whale spatial distribution in northwestern Mediterranean Sea. In Leuangthong, O. and Deutsch, C., editors, *Geostatistics Banff 2004 Volume 2*, pages 777–786, Dordrecht, The Netherlands. Kluwer Academic Publishers.

Monestiez, P., Dubroca, L., Bonin, E., Durbec, J. P., and Guinet, C. (2006). Geostatistical modelling of spatial distribution of Balenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecol Model*, 193:615–628.

Pickle, L. W., Mungiole, M., Jones, G. K., and White, A. A. (1999). Exploring spatial patterns of mortality: the new Atlas of United States mortality. *Stat Med*, 18:3211–3220.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rey, G., Jougla, E., Fouillet, A., and Hemón, D. (2009). Ecological association between a deprivation and mortality in France over the period 1997-2001: variation with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC Public Health*, 9(33).

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M., and Eilers, P. H. C. (2015). Fast algorithm for smoothing parameter selection in multidimensional generalized *P*-splines. *Stat Comput*. (in press).

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J Comput Graph Stat*, 11:735–757.

Salmond, C. E. and Cramptom, P. (2012). Development of New Zealand's deprivation index (NZDep) and its uptake as a national policy tool. *Can J Public Health*, 103(8 Suppl 2):S7–11.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–721.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics, New Jersey.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, New York.

Appendix

In this section, we introduce some notation and definitions of array methods proposed in Currie et al. (2006) and Eilers et al. (2006) that we have used in Section 2.2.1.

Definition 1 (Row tensor) *The row tensor of a matrix* $\mathbf{X}$ *with* c *columns is defined as:*

$$\mathcal{G}(\mathbf{X}) = (\mathbf{X} \otimes \mathbf{1}'_c) \odot (\mathbf{1}'_c \otimes \mathbf{X}),$$

where $\mathbf{1}_c$ *is a vector of 1's of length* c*, and* $\odot$ *is the element-by-element product.*

The previous definition can be extended in the following way.

Definition 2 (Row tensor of two matrices) *The row tensor of the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, of dimensions $n \times c_1$ and $n \times c_2$, respectively, is defined as:*

$$\mathcal{G}(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1 \otimes \mathbf{1}'_{c_2}) \odot (\mathbf{1}'_{c_1} \otimes \mathbf{X}_2),$$

where $\mathbf{1}_{c_1}$ and $\mathbf{1}_{c_2}$ are vectors of 1's of lengths c_1 and c_2, respectively.

Note that the previous definition denotes the "row-wise" Kronecker product of two matrices, which we have introduced in Section 2.1.

Definition 3 ($\mathcal{H}$-transform) *The $\mathcal{H}$-transform of the d-dimensional array $\mathbf{A}$ of size $c_1 \times c_2 \times \cdots \times c_d$ by the matrix $\mathbf{X}$ of dimension $r \times c_1$, denoted as $\mathcal{H}(\mathbf{X}, \mathbf{A})$, is defined as follows. Let $\mathbf{A}^*$ be the matrix of dimension $c_1 \times c_2 c_3 \cdots c_d$ that is obtained by flattening dimensions 2-d of $\mathbf{A}$; form the matrix product $\mathbf{X}\mathbf{A}^*$ of dimension $r \times c_2 c_3 \cdots c_d$; then $\mathcal{H}(\mathbf{X}, \mathbf{A})$ is the d-dimensional array of size $r \times c_1 \times c_2 c_3 \cdots c_d$ that is obtained from $\mathbf{X}\mathbf{A}^*$ by reinstating dimensions 2-d of $\mathbf{A}$.*

In one-dimension, we have that $\mathbf{A} = \boldsymbol{a}$ and $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\boldsymbol{a}$, whereas in two-dimensions $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\mathbf{A}$. The following definition generalizes the transpose of a matrix.

Definition 4 (Array rotation) *The rotation of the d-dimensional array $\mathbf{A}$ of size $c_1 \times c_2 c_3 \cdots c_d$ is the d-dimensional array $\mathcal{R}(\mathbf{A})$ of size $c_2 \times c_3 \times \cdots \times c_d \times c_1$ that is obtained by permuting the indices of $\mathbf{A}$.*

From the two last definitions, we obtain:

Definition 5 (Rotated $\mathcal{H}$-transform) *The rotated $\mathcal{H}$-transform of the array $\mathbf{A}$ by the matrix $\mathbf{X}$ is given by:*

$$\rho(\mathbf{X}, \mathbf{A}) = \mathcal{R}(\mathcal{H}(\mathbf{X}, \mathbf{A})).$$