UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
16-07
ISSN 2387-0303
July 2016

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

Modelling latent trends from spatio-temporally grouped data using composite link mixed models

Diego Ayma[a], María Durbán[a], Dae-Jin Lee[b], and Jan van de Kassteele[c]

Abstract

Epidemiological data are frequently recorded at coarse spatio-temporal resolutions. The aggregation process is done for several reasons: to protect confidential patients' information, to compare with other datasets at a coarser resolution than the original, or to summarize data in a compact manner. However, we lose detailed patterns that follow the original data, which can be of interest for researchers and public health officials. In this paper we propose the use of the penalized composite link model (Eilers, 2007), together with its mixed model representation, to estimate the underlying trend behind grouped data at a finer spatio-temporal resolution. Also, this model allows the incorporation of fine-scale population into the estimation procedure. We assume the underlying trend is smooth across space and time. The mixed model representation enables the use of sophisticated algorithms such as the SAP algorithm of Rodríguez-Álvarez et al. (2015) for fast estimation of the amount of smoothness. We illustrate our proposal with the analysis of data obtained during the largest outbreak of Q fever in the Netherlands.

Keywords: Penalized composite link models; Q fever incidence; SAP algorithm; Spatio-temporal disaggregation.

[a] Department of Statistics, Universidad Carlos III de Madrid, Spain
[b] BCAM - Basque Center for Applied Mathematics, Spain.
[c] National Institute for Public Health and the Environment, The Netherlands.
[*] Corresponding author. E-mail: dayma@est-econ.uc3m.es

Acknowledgements: The first and the second authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-02 and MTM2014-52184-P. The third author acknowledges financial support from 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.

Modelling latent trends from spatio-temporally grouped data using composite link mixed models

Diego Ayma*[1], María Durbán[1], Dae-Jin Lee[2], and Jan van de Kassteele[3]

[1]Department of Statistics, Universidad Carlos III de Madrid, Spain
[2]BCAM - Basque Center for Applied Mathematics, Spain
[3]National Institute for Public Health and the Environment, the Netherlands

25th July 2016

Abstract

Epidemiological data are frequently recorded at coarse spatio-temporal resolutions. The aggregation process is done for several reasons: to protect confidential patients' information, to compare with other datasets at a coarser resolution than the original, or to summarize data in a compact manner. However, we lose detailed patterns that follow the original data, which can be of interest for researchers and public health officials. In this paper we propose the use of the penalized composite link model (Eilers, 2007), together with its mixed model representation, to estimate the underlying trend behind grouped data at a finer spatio-temporal resolution. Also, this model allows the incorporation of fine-scale population into the estimation procedure. We assume the underlying trend is smooth across space and time. The mixed model representation enables the use of sophisticated algorithms such as the SAP algorithm of Rodríguez-Álvarez et al. (2015) for fast estimation of the amount of smoothness. We illustrate our proposal with the analysis of data obtained during the largest outbreak of Q fever in the Netherlands.

Keywords

Penalized composite link models; Q fever incidence; SAP algorithm; Spatio-temporal 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

In the last decades, the development of spatio-temporal statistical methods has experienced a remarkable growth in the field of epidemiology: the cornerstone of public health. This increase has been favored due mostly to the advances in geographic information systems (GISs), the access to reliable health data registers, and the disposition of powerful software capable to processing and analysing large amount of data. Methodological contributions for the analysis of spatio-temporal health data come from several interdisciplinary researchers, whose backgrounds are mostly related with statistics, geography, and environmental sciences.

Among the types of epidemiological inquiries, disease mapping has received a great interest in public health since it allows the visualization of disease incidence (or mortality risk) patterns over a specific study area. For that, appropriate statistical methods are applied to health data — which are usually recorded over geographical units — to provide smoothed disease incidences by unit. The spatial smoothing is performed in order to obtain more stable and less noisy estimates of the incidence rates associated with each unit (Waller and Gotway, 2004), thus allowing the identification of meaningful patterns. The incorporation of the temporal dimension in this context enables the study of the disease incidence evolution in each unit, during a certain period of time (generally divided in years). However, its inclusion implies a challenge when it comes to smoothing data, in terms of computational time and storage. Several techniques has been proposed for the spatio-temporal smoothing of health data; most of them developed under an empirical Bayes approach where B-splines are used (MacNab and Dean, 2001; Ugarte et al., 2010) or a hierarchical Bayesian framework where conditional autoregressive (CAR) structures are included (Waller et al., 1997; Martínez-Beneito et al., 2008). Within the latter approach, methods using integrated nested Laplace approximations (INLA, Rue et al., 2009) has recently been proposed (see Schrödle and Held, 2011; Ugarte et al., 2014; Bauer et al., 2016; among others).

All the works cited above provide smoothed estimates that are assumed constant over each unit and year. Moreover, most of them can be extended in order to include explanatory variables, which must be at the same spatio-temporal resolution as health data. Thus, they restrict the incorporation of fine-scale population information and other relevant risk factors recorded at a finer resolution (an issue known as the *modifiable areal unit problem*; see, for example, Gelfand et al., 2010, Ch. 29). To overcome this and other limitations, we propose the use of the *penalized composite link model* (PCLM, Eilers, 2007), together with its *mixed model representation*, for the disaggregation of spatio-temporally grouped count data. The *composite link mixed model* (CLMM, Ayma et al., 2016) approach allows to obtain smoothed estimates at a finer resolution, from data aggregated over space and/or time. Moreover, it allows to include fine-scale population information and specific random effects or further correlation structure if necessary. We assume here the underlying spatio-temporal trend behind aggregated data is smooth. Thus no further hypotheses (as either stationarity and isotropy) are required to define the latent spatio-temporal process. The

flexibility of the model is given by the use of B-spline bases and a discrete penalty on the regression coefficients, following the P-spline methodology (Eilers and Marx, 1996). Smoothness is controlled by three smoothing parameters (two for the spatial dimension and one for the temporal dimension) that have to be estimated together with the global trend at the fine scale. To carry out the spatio-temporal disaggregation, the CLMM uses a composition matrix that links both coarse and fine resolutions, which is expressed as a Kronecker product of two marginal composition matrices (one that acts at spatial level and the other at temporal level).

In the literature we can find several techniques that allow to disaggregate health data spatially. For example, the use of an interpolation process from empirical Bayes estimates (Berke, 2004), (generalized) Poisson kriging methods (Goovaerts, 2006), log-Gaussian Cox processes (Diggle et al., 2013), and spatial CLMMs (Ayma et al., 2016) have been suggested to produce a continuous smooth surface (across the study area) from regional health data. However, as far as we know, no appropriate model exists addressing the problem of disaggregation of health data both in space and time (although there exist works about the spatio-temporal disaggregation of Gaussian data; see, for example, Prairie et al., 2007; Segond et al., 2007; Schleiss and Berne, 2012; Bindhu and Narasimhan, 2015). The approach presented here allows to obtain detailed dynamic maps for disease incidence data, which are publicly available in an aggregated form over space and time.

The rest of the paper is organized as follows. In Section 2, we introduce the CLMM approach for spatio-temporally grouped count data. Here we include how we can estimate the parameters of the model and approximate standard errors associated to the estimated latent trend. Also, we adapt the *separation of anisotropic penalties* (SAP) algorithm of Rodríguez-Álvarez et al. (2015) in the context of CLMMs for estimating smoothing parameters, and we use *generalized linear array model* (GLAM) methods (Currie et al., 2006; Eilers et al., 2006) to speed up computations. In Section 3, we apply the CLMM to a Q fever dataset recorded over municipalities (of a specific Dutch study area) and months in 2009, in order to visualize the Q fever incidence on a fine spatial grid along all weeks. In Section 4, we conduct a simulation study to examine the performance of the CLMM approach. The paper ends with a short discussion in Section 5.

2 The spatio-temporal composite link mixed model

The PCLM approach of Eilers (2007), which is based on the work by Thompson and Baker (1981), allows to estimate the underlying or latent distribution of grouped count data at a finer resolution. Since his work is developed in an one-dimensional context, we first extend the PCLM to the spatio-temporal case as follows.

Let y_{it}, $i = 1, ..., n$, $t = 1, ..., T_1$, denote count data that are recorded over n non-overlapping geographical units, which form the study area of interest, at T_1 time periods. Here, the unit

boundaries are assumed to be fixed during these periods. Now suppose that we want to estimate the latent distribution of the vector of counts $\boldsymbol{y} = (y_{11}, ..., y_{n1}, ..., y_{1T_1}, ..., y_{nT_1})'$ at a spatio-temporal support that is a (nested) refinement of the original one. The fine support is determined by three covariates: $\boldsymbol{x}_1 = (x_{11}, ..., x_{1m})'$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2m})'$, with $m > n$, which represent the longitude and latitude coordinates of the spatial refinement, respectively; and $\boldsymbol{x}_3 = (x_{31}, ..., x_{3T_2})'$, with $T_2 > T_1$, which represents the temporal refinement. Assuming that $\boldsymbol{y}$ is distributed Poisson with mean vector $\boldsymbol{\mu} = (\mu_{11}, ..., \mu_{n1}, ..., \mu_{1T_1}, ..., \mu_{nT_1})'$, the spatio-temporal PCLM is given by:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\exp(\mathbf{B}\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\gamma}$ denotes the latent mean at the fine resolution, $\mathbf{C}$ is the composition matrix that describes how these latent expectations are combined to yield $\boldsymbol{\mu}$, and $\mathbf{B}$ is a full regression basis constructed from covariates $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. To achieve smoothness, the vector of regression coefficients $\boldsymbol{\theta}$ is penalized by a discrete penalty matrix $\mathbf{P}$ in the form $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$. Matrices $\mathbf{B}$, $\mathbf{P}$, and $\mathbf{C}$ in Eq. (1) are described below.

Following the proposal of Lee and Durbán (2011), the regression basis $\mathbf{B}$ in Eq. (1) is given by:

$$\mathbf{B} = \mathbf{B}_3 \otimes (\mathbf{B}_2 \square \mathbf{B}_1), \tag{2}$$

where $\mathbf{B}_1 = \mathbf{B}(\boldsymbol{x}_1)$, $\mathbf{B}_2 = \mathbf{B}(\boldsymbol{x}_2)$, and $\mathbf{B}_3 = \mathbf{B}(\boldsymbol{x}_3)$ are univariate B-spline bases of dimensions $m \times c_1$, $m \times c_2$ and $T_2 \times c_3$, respectively. The matrix operators $\otimes$ and $\square$ represent the Kronecker and the 'row-wise' Kronecker (Eilers et al., 2006) products, respectively. The construction of the B-spline bases $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{B}_3$ in Eq. (2) depends on the number of selected equally-spaced knots for each covariate, and the degree of the B-spline used (see Eilers and Marx (1996) and Eilers et al. (2015), for a further discussion). For the spatio-temporal case, Lee and Durbán (2011) propose to use the following penalty matrix:

$$\mathbf{P} = \lambda_1 \mathbf{I}_{c_3} \otimes \mathbf{I}_{c_2} \otimes \mathbf{P}_1 + \lambda_2 \mathbf{I}_{c_3} \otimes \mathbf{P}_2 \otimes \mathbf{I}_{c_1} + \lambda_3 \mathbf{P}_3 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_1}, \tag{3}$$

where $\mathbf{I}_k$ denotes an identity matrix of dimension $k \times k$, λ_d is the smoothing parameter that controls the amount of smoothing along the covariate $\boldsymbol{x}_d$, and $\mathbf{P}_d = \mathbf{D}_d'\mathbf{D}_d$ is the marginal penalty matrix based on the matrix $\mathbf{D}_d$ that computes q_d-th differences, i.e, $\Delta^{q_d}\boldsymbol{\theta} = \mathbf{D}_d\boldsymbol{\theta}$ $(d = 1, 2, 3)$. The penalty matrix in Eq. (3) allows for anisotropy, i.e., a different amount of smoothing for $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, which is a desirable characteristic in the spatio-temporal context.

The composition matrix $\mathbf{C}$ in Eq. (1) can be expressed as:

$$\mathbf{C} = \mathbf{C}_t \otimes \mathbf{C}_s, \tag{4}$$

where $\mathbf{C}_s$ and $\mathbf{C}_t$ are the spatial and temporal composition matrices of dimensions $n \times m$ and $T_1 \times T_2$, respectively. The definition of these matrices depends on the type and level of aggregation. For example, if we want to estimate the latent distribution at a fine spatial grid (over the study area),

the entries of the composition matrix $\mathbf{C}_s$ will be:

$$[\mathbf{C}_s]_{ij} = \begin{cases} 1 & \text{if } (x_{1j}, x_{2j}) \text{ belongs to unit } v_i \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where (x_{1j}, x_{2j}) are the cell centroid coordinates of the fine grid, for $i = 1, ..., n$ and $j = 1, ..., m$. Another option to construct $\mathbf{C}_s$ would be to consider its entries as the proportions of area that each grid cell shares with a specific unit. The temporal composition matrix $\mathbf{C}_t$ can be used to disaggregate coarse time intervals into detailed time periods (for example, from years or trimesters to months, weeks, or days). In Section 3, we show the structure that $\mathbf{C}_t$ will have, specifically for our purposes. Notice that if $\mathbf{C}_s = \mathbf{I}_n$, $\mathbf{C}_t = \mathbf{I}_{T_1}$, and the unit centroids are used as spatial covariates, the presented methodology is reduced to the Poisson version of the proposal given by Lee and Durbán (2011), for the smoothing of spatio-temporal count data. In this paper we refer to this particular model as (Poisson) penalized generalized linear mixed model or, more briefly, as PGLMM.

When spatial data are recorded over a coarse regular grid, a more appropriate definition for the regression basis $\mathbf{B}$ in Eq. (1) is $\mathbf{B} = \mathbf{B}_3 \otimes (\mathbf{B}_2 \otimes \mathbf{B}_1)$, where the spatial refinement correspond to the cell centroid coordinates of a fine grid. The penalty matrix $\mathbf{P}$ in Eq. (3) is still valid in this case, because its definition is independent of data structure. Moreover, the spatial composition matrix can be written as $\mathbf{C}_s = \mathbf{C}_2 \otimes \mathbf{C}_1$, where each $\mathbf{C}_d$, $d = 1, 2$, is constructed according to the disaggregation of the coarser grid cells into small ones.

2.1 Mixed model representation

To estimate the model in Eq. (1), subject to the penalization given in Eq. (3), we reformulate it as a mixed model — in fact, as a generalized linear mixed model (GLMM). The reformulated model is called (Poisson) CLMM (Ayma et al., 2016), which not only allows to analyse count data but also rates. The CLMM is given by:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\left(\boldsymbol{e}_{\mathrm{f}} * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})\right), \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \tag{6}$$

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 random effects have covariance matrix $\mathbf{G}$ that depends on the smoothing parameters. The CLMM incorporates population information at the fine scale by means of the vector $\boldsymbol{e}_{\mathrm{f}}$ in Eq. (6). If the vector of exposures is only available at the aggregated scale, i.e., $\boldsymbol{e} = (e_{11}, ..., e_{n1}, ..., e_{1T_1}, ..., e_{nT_1})'$, a naive approach to estimate $\boldsymbol{e}_{\mathrm{f}}$ assumes that these aggregated exposures are evenly distributed throughout the fine resolution. Therefore, we can compute these naive estimates as follows:

$$\hat{\boldsymbol{e}}_{\text{naive}} = \mathbf{C}^{-}\boldsymbol{e},$$

where $\mathbf{C}^{-}$ denotes the Moore-Penrose inverse of $\mathbf{C}$.

To obtain the mixed model representation given in Eq. (6) under a spatio-temporal context, we follow the proposal given by Lee and Durbán (2011) where the singular value decomposition (SVD) of each discrete penalty matrix $\mathbf{P}_d$ in Eq. (3), for $d = 1, 2, 3$, is used. Let $\mathbf{P}_d = \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{U}_d'$ be the SVD of $\mathbf{P}_d$, where $\mathbf{U}_d$ is the matrix of singular vectors and $\mathbf{\Sigma}_d$ is a diagonal matrix that contains the singular values of the decomposition, for $d = 1, 2, 3$. Each matrix $\mathbf{U}_d$ can be partitioned in two parts so that:

$$\mathbf{P}_d = \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{U}_d' = [\mathbf{U}_{dn} : \mathbf{U}_{ds}] \begin{bmatrix} \mathbf{0}_{q_d} & \\ & \tilde{\mathbf{\Sigma}}_d \end{bmatrix} [\mathbf{U}_{dn} : \mathbf{U}_{ds}]',$$

where $\mathbf{0}_q$ denotes a square matrix of zeroes of dimension $q \times q$, $\mathbf{U}_{dn}$ and $\mathbf{U}_{ds}$ are matrices of singular vectors corresponding to zero and non-zero singular values, respectively, and $\tilde{\mathbf{\Sigma}}_d$ is a diagonal matrix with the non-zero singular values in the main diagonal. Using this decomposition, the mixed model matrices of the model in Eq. (6) are obtained as:

$$\begin{aligned} \mathbf{X} &= \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1)\,, \\ \mathbf{Z} &= [\,\mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1) : \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1) : \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1) : \\ &\quad \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1) : \mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1) : \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1) : \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1)], \end{aligned} \tag{7}$$

where $\mathbf{X}_d = \mathbf{B}_d \mathbf{U}_{dn}$ and $\mathbf{Z}_d = \mathbf{B}_d \mathbf{U}_{ds}$, for $d = 1, 2, 3$. Moreover, the inverse of the covariance matrix $\mathbf{G}$ of the random effects $\boldsymbol{\alpha}$ in Eq. (6) becomes the block-diagonal matrix:

$$\begin{aligned} \mathbf{G}^{-1} = \text{blockdiag}\, (&\lambda_3 \mathbf{F}_{3u}, \lambda_2 \mathbf{F}_{2u}, \lambda_1 \mathbf{F}_{1u}, \\ &\lambda_2 \mathbf{F}_{22} + \lambda_3 \mathbf{F}_{32}, \lambda_1 \mathbf{F}_{12} + \lambda_3 \mathbf{F}_{31}, \lambda_1 \mathbf{F}_{11} + \lambda_2 \mathbf{F}_{21}, \\ &\lambda_1 \mathbf{F}_{1t} + \lambda_2 \mathbf{F}_{2t} + \lambda_3 \mathbf{F}_{3t})\,, \end{aligned} \tag{8}$$

where:

$$\begin{aligned} &\mathbf{F}_{1u} = \mathbf{I}_{q_3} \otimes \mathbf{I}_{q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{2u} = \mathbf{I}_{q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{q_1}, && \mathbf{F}_{3u} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{q_2} \otimes \mathbf{I}_{q_1}, \\ &\mathbf{F}_{11} = \mathbf{I}_{q_3} \otimes \mathbf{I}_{c_2 - q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{12} = \mathbf{I}_{c_3 - q_3} \otimes \mathbf{I}_{q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{21} = \mathbf{I}_{q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{c_1 - q_1}, \\ &\mathbf{F}_{22} = \mathbf{I}_{c_3 - q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{q_1}, && \mathbf{F}_{31} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{q_2} \otimes \mathbf{I}_{c_1 - q_1}, && \mathbf{F}_{32} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{c_2 - q_2} \otimes \mathbf{I}_{q_1}, \\ &\mathbf{F}_{1t} = \mathbf{I}_{c_3 - q_3} \otimes \mathbf{I}_{c_2 - q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{2t} = \mathbf{I}_{c_3 - q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{c_1 - q_1}, && \mathbf{F}_{3t} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{c_2 - q_2} \otimes \mathbf{I}_{c_1 - q_1}\,. \end{aligned}$$

If the data are spatially recorded over a regular coarse grid, the corresponding mixed model matrices are obtained as in Eq. (7), replacing the 'row-wise' Kronecker products $\square$ by Kronecker products $\otimes$. Under this context, the formulation of $\mathbf{G}^{-1}$ is the same as in Eq. (8).

2.2 Parameter estimation

As is usual in the theory of GLMMs, the estimation of the CLMM is carried out iteratively. The estimation procedure involves two interrelated stages: (i) fixed and random effects coefficients estimation ($\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$); and (ii) smoothing parameters estimation (λ_1, λ_2, and λ_3). For fixed values

of the smoothing parameters, the estimation of the model's fixed and random effects are obtained using the penalized quasi-likelihood (PQL) approach of Breslow and Clayton (1993). We briefly describe this point in Section 2.2.1. Since several matrix cross-products appear in the estimation procedure, in Section 2.2.2 we include the use of the so-called GLAM methods (Currie et al., 2006; Eilers et al., 2006) in the context of CLMMs, to speed up computations, and to avoid storage problems. For the estimation of the smoothing parameters, an adaptation of the SAP algorithm (Rodríguez-Álvarez et al., 2015) is used in the context of the CLMMs. This adaptation is presented in Section 2.2.3.

2.2.1 Fixed and random effects coefficients estimation

For given values of λ_1, λ_2, and λ_3, the estimation of the fixed and random effects coefficients of the CLMM in Eq. (6) can be obtained by maximizing the approximate penalized log-likelihood:

$$\boldsymbol{y}' \log(\boldsymbol{\mu}) - \mathbf{1}'_{nT_1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\boldsymbol{\alpha}, \tag{9}$$

where $\mathbf{1}_n$ denotes a vector of ones of length n and $\boldsymbol{\mu} = \mathbf{C}\left(\boldsymbol{e}_{\mathrm{f}} * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})\right)$. As is was shown in Ayma et al. (2016), differentiation of Eq. (9) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ leads to the score equations:

$$\begin{aligned} \breve{\mathbf{X}}'\left(\boldsymbol{y} - \boldsymbol{\mu}\right) &= \mathbf{0}, \\ \breve{\mathbf{Z}}'\left(\boldsymbol{y} - \boldsymbol{\mu}\right) &= \mathbf{G}^{-1}\boldsymbol{\alpha}, \end{aligned} \tag{10}$$

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 $\mathbf{W} = \mathrm{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. Matrices $\breve{\mathbf{X}}$ and $\breve{\mathbf{Z}}$ are called 'working' mixed model matrices, since $\mathbf{W}$ and $\boldsymbol{\Gamma}$ change during the estimation procedure. Thus defining the working vector:

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

the solution of the score equations given in Eq. (10) via Fisher scoring algorithm can be expressed as the iterative solution of the linear system:

$$\begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}\mathbf{G} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{I} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}\mathbf{G} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z} \\ \breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z} \end{bmatrix}, \tag{11}$$

where $\boldsymbol{b} = \mathbf{G}^{-1}\boldsymbol{\alpha}$. This yields to a modified version of the standard mixed model estimators:

$$\widehat{\boldsymbol{\beta}} = (\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}, \tag{12}$$

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} &= \mathbf{G}\breve{\mathbf{Z}}'\mathbf{V}^{-1}(\boldsymbol{z} - \breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{G}\breve{\mathbf{Z}}'\mathbf{N}\boldsymbol{z}, \end{aligned} \tag{13}$$

where:

$$\mathbf{V} = \mathbf{W}^{-1} + \breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}', \tag{14}$$

$$\mathbf{N} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}. \tag{15}$$

2.2.2 GLAM methods for CLMMs

When we are dealing with the estimation of latent trends in multiple dimensions, we are susceptible to encounter problems with storage and computational burden. In the case of data arranged in multidimensional grids, it is possible to circumvent these problems using GLAM methods developed in Currie et al. (2006) and Eilers et al. (2006). These methods are designed to avoid the direct computation of matrix cross-products where Kronecker operations are involved, by using sequences of nested matrix operations. In this section we show the use of these methods in the spatio-temporal CLMM context.

In the previous section we have seen that several matrix cross-products have to be computed as, for example, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$, and $\breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z}$. The GLAM methods offer a fast and an efficient way to compute them as follows.

The matrix-by-vector products that we need to compute are: $\mathbf{X}\boldsymbol{\beta}$, $\mathbf{Z}\boldsymbol{\alpha}$, and $\mathbf{C}\boldsymbol{\gamma}$. These expressions are computed as follows:

$$
\begin{aligned}
\mathbf{X}\boldsymbol{\beta} &\equiv \rho(\mathbf{X}_3, \rho(\mathbf{R}_1, \tilde{\mathbf{Q}})),\\
\mathbf{Z}\boldsymbol{\alpha} &\equiv \rho(\mathbf{Z}_3, \rho(\mathbf{R}_1, \tilde{\mathbf{A}}_1)) + \rho(\mathbf{X}_3, \rho(\mathbf{R}_2, \tilde{\mathbf{A}}_2)) + \rho(\mathbf{X}_3, \rho(\mathbf{R}_3, \tilde{\mathbf{A}}_3)) + \rho(\mathbf{Z}_3, \rho(\mathbf{R}_2, \tilde{\mathbf{A}}_4)) +\\
&\quad \rho(\mathbf{Z}_3, \rho(\mathbf{R}_3, \tilde{\mathbf{A}}_5)) + \rho(\mathbf{X}_3, \rho(\mathbf{R}_4, \tilde{\mathbf{A}}_6)) + \rho(\mathbf{Z}_3, \rho(\mathbf{R}_4, \tilde{\mathbf{A}}_7)),\\
\mathbf{C}\boldsymbol{\gamma} &\equiv \rho(\mathbf{C}_t, \rho(\mathbf{C}_s, \tilde{\boldsymbol{\Gamma}})),
\end{aligned}
$$

with $\mathbf{R}_1 = \mathcal{G}(\mathbf{X}_2, \mathbf{X}_1)$, $\mathbf{R}_2 = \mathcal{G}(\mathbf{Z}_2, \mathbf{X}_1)$, $\mathbf{R}_3 = \mathcal{G}(\mathbf{X}_2, \mathbf{Z}_1)$, $\mathbf{R}_4 = \mathcal{G}(\mathbf{Z}_2, \mathbf{Z}_1)$, where ρ and $\mathcal{G}$ denote the rotated $\mathcal{H}-$transform and the row-tensor product, respectively, which are defined in Appendix A, and the symbol $\equiv$ means that both sides have the same elements but in a different order. The matrices $\tilde{\mathbf{Q}}$, $\tilde{\boldsymbol{\Gamma}}$, and the $\tilde{\mathbf{A}}_k$'s (for $k = 1, ..., 7$) are arrangements of the vectors $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}_k$'s, respectively, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', ..., \boldsymbol{\alpha}_7')'$, whose dimensions correspond to the number of columns of the first matrix where ρ acts, times the number of columns of the second matrix where ρ acts (i.e, $\tilde{\mathbf{Q}}$ has dimension $\text{ncol}(\mathbf{R}_1) \times \text{ncol}(\mathbf{X}_3) = q_1 q_2 \times q_3$, $\tilde{\boldsymbol{\Gamma}}$ has dimension $\text{ncol}(\mathbf{C}_s) \times \text{ncol}(\mathbf{C}_t) = m \times T_2$, and so on). Therefore, it holds that $\text{vec}(\tilde{\mathbf{Q}}) = \boldsymbol{\beta}$, $\text{vec}(\tilde{\boldsymbol{\Gamma}}) = \boldsymbol{\gamma}$, and $\text{vec}(\tilde{\mathbf{A}}_k) = \boldsymbol{\alpha}_k$, for $k = 1, ..., 7$.

Also, the following matrix cross-products are needed: $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}$, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}}$ (which is equal to $(\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}})'$), $\breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z}$, and $\breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z}$. First, note that they can be reduced as:

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

Thus, we only need to compute $\mathbf{C\Gamma X}$ and $\mathbf{C\Gamma Z}$. These expressions are computed as follows:

$$\begin{aligned}
\mathbf{C\Gamma X} &\equiv \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_1, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})), \\
\mathbf{C\Gamma Z} &\equiv [\rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_1, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})) : \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_2, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})) : \\
&\quad \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_3, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})) : \rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_2, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})) : \\
&\quad \rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_3, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})) : \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_4, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}})) : \\
&\quad \rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_4, \mathbf{C}_s')', \tilde{\mathbf{\Gamma}}))]
\end{aligned}$$

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

As we seen before, the GLAM methods have overheads of rearrangement and redimensioning, but these are efficient operations in comparison to the Kronecker product. Table 1 gives some comparatives on timing for the computation of the matrix products $\mathbf{C\Gamma X}$ and $\mathbf{C\Gamma Z}$, with and without using GLAM methods. These computations were performed using hypothetical matrices $\mathbf{C}$, $\mathbf{\Gamma}$, $\mathbf{X}$, and $\mathbf{Z}$, of dimension 40×165160, 165160×165160, 165160×8, and 165160×839, respectively. In both cases the use of GLAM methods improve the computational burden.

Matrix product	Times (s) without GLAM	Times (s) with GLAM	Ratio
$\mathbf{C\Gamma X}$	1.95	0.08	24 : 1
$\mathbf{C\Gamma Z}$	183.09	8.74	21 : 1

Table 1: *User CPU times to calculate* $\mathbf{C\Gamma X}$ *and* $\mathbf{C\Gamma Z}$.

2.2.3 Smoothing parameters estimation

Conditioning on the estimators given in Eq. (12) and Eq. (13), we can obtain estimates for λ_1, λ_2, and λ_3 by numerically maximizing the approximate restricted maximum log-likelihood (REML, Patterson and Thompson, 1971) given in Eq. (19). To avoid the use of any numerical optimization method, we adapt the SAP algorithm of Rodríguez-Álvarez et al. (2015) to the CLMM setting. This algorithm is a generalization of the work by Schall (1991) that deals with non-standard structures of the covariance matrix of the random effects, where the smoothing parameters are seen as ratios of variance components, i.e., $\lambda_d = \frac{\phi}{\tau_d^2}$, for $d = 1, 2, 3$. Since we are working under a Poisson framework, $\phi = 1$. Thus, the problem is reduced to obtain estimates for the variance components τ_1^2, τ_2^2, and τ_3^2.

Following the proposal of Rodríguez-Álvarez et al. (2015), we can derive closed-form expressions for the REML estimates of the variance components τ_d^2 (for $d = 1, 2, 3$). These estimates are given by:

$$\widehat{\tau}_d^2 = \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{\Lambda}_d \widehat{\boldsymbol{\alpha}}}{\mathrm{ed}_d}, \tag{16}$$

where

$$\text{ed}_d = \text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\mathbf{\Lambda}_d}{\tau_d^2}\mathbf{G}\right), \tag{17}$$

with $\mathbf{N}$ defined as in Eq. (15), and

$$\begin{aligned}
\mathbf{\Lambda}_1 &= \text{blockdiag}\left(\mathbf{0}_{q_1q_2(c_3-q_3)}, \mathbf{0}_{q_1q_3(c_2-q_2)}, \mathbf{F}_{1u}, \mathbf{0}_{q_1(c_2-q_2)(c_3-q_3)}, \mathbf{F}_{12}, \mathbf{F}_{11}, \mathbf{F}_{1t}\right), \\
\mathbf{\Lambda}_2 &= \text{blockdiag}\left(\mathbf{0}_{q_1q_2(c_3-q_3)}, \mathbf{F}_{2u}, \mathbf{0}_{q_2q_3(c_1-q_1)}, \mathbf{F}_{22}, \mathbf{0}_{q_2(c_1-q_1)(c_3-q_3)}, \mathbf{F}_{21}, \mathbf{F}_{2t}\right), \\
\mathbf{\Lambda}_3 &= \text{blockdiag}\left(\mathbf{F}_{3u}, \mathbf{0}_{q_1q_3(c_2-q_2)}, \mathbf{0}_{q_2q_3(c_1-q_1)}, \mathbf{F}_{32}, \mathbf{F}_{31}, \mathbf{0}_{q_3(c_1-q_1)(c_2-q_2)}, \mathbf{F}_{3t}\right).
\end{aligned}$$

The proof of this result is provided in Appendix B. Notice that the matrix $\mathbf{G}^{-1}$ defined in Eq. (8) can be decomposed as $\mathbf{G}^{-1} = \frac{1}{\tau_1^2}\mathbf{\Lambda}_1 + \frac{1}{\tau_2^2}\mathbf{\Lambda}_2 + \frac{1}{\tau_3^2}\mathbf{\Lambda}_3$, where the capital lambdas are defined above. An algorithm for the CLMM estimation (which is an adaptation of the algorithm provided in Rodríguez-Álvarez et al., 2015, p. 945) is provided in Appendix C.

The computation of the trace given in Eq. (17) can be efficiently obtained by taking into account that $\mathbf{G}\mathbf{\Lambda}_d\mathbf{G}$ is a diagonal matrix. Thus, we only need to compute the diagonal of $\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}$ in order to obtain this trace. From Harville (1977, Eq. (5.3)) we have that:

$$\breve{\mathbf{Z}}'\mathbf{N} = [\mathbf{0}_{(c_1c_2c_3-q_1q_2q_3)\times q_1q_2q_3}|\mathbf{I}_{(c_1c_2c_3-q_1q_2q_3)}]\mathbf{M}^{-1}[\breve{\mathbf{X}}|\breve{\mathbf{Z}}]\mathbf{W},$$

where:

$$\mathbf{M} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}\mathbf{G} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{I} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}\mathbf{G} \end{bmatrix}.$$

Therefore, the diagonal elements of the matrix $\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}$ are obtained by the column-wise addition of

$$\left(\mathbf{0}_{(c_1c_2c_3-q_1q_2q_3)\times q_1q_2q_3}|\mathbf{I}_{(c_1c_2c_3-q_1q_2q_3)}]\mathbf{M}^{-1}\right)' \odot \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} \end{bmatrix},$$

where $\odot$ represents the Hadamard product. Notice that $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, and the submatrices of $\mathbf{M}$ have been already computed for the estimation of the fixed and random coefficients. To speed up the computational time even more, we can derive an explicit formulation of the inverse of $\mathbf{M}$, following the ideas given in Harville (1997, p. 99).

2.3 Effective dimension and approximate standard errors

The effective dimension (ED) of the model given in Eq. (6) can be obtained as:

$$\text{ED} = \sum_{d=1}^{3} \text{ed}_d + q_1q_2q_3, \tag{18}$$

where the ed_d expressions are computed from Eq. (17). The first term in the right-hand side of Eq. (18) corresponds to the dimension of the penalized part (see Appendix B), whereas the second to the unpenalized part.

Once the model parameter estimates at convergence are obtained, we can derive standard errors for $\widehat{\boldsymbol{\eta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}$ by using the Bayesian approximation of the covariance matrix for $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})'$ (Lin and Zhang, 1999):

$$\mathbb{C}\text{ov}((\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})') = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} + \mathbf{G}^{-1} \end{bmatrix}^{-1}.$$

Thus, we can obtain approximate standard errors for $\widehat{\boldsymbol{\eta}}$ by taking the square root of $\mathbb{V}\text{ar}(\widehat{\boldsymbol{\eta}})$, which is obtained as:

$$\mathbb{V}\text{ar}(\widehat{\boldsymbol{\eta}}) = \text{diag}([\mathbf{X} : \mathbf{Z}]\text{Cov}((\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})')[\mathbf{X} : \mathbf{Z}]').$$

3 Real data application: Q fever outbreak in the Netherlands

In this section we illustrate our proposal using data related with Q fever outbreaks in the Netherlands. First we briefly describe these data, and then we analyse them using the CLMM approach described in Section 2.

3.1 Q fever data

Q fever is a widespread zoonotic disease caused by the bacterium *Coxiella burnetii*. The transmission to humans of *C. burnetii* is primary associated with ruminants like cattle, sheep, and goats. During parturition or abortion of infected animals, high numbers of *C. burnetii* are shed within the amniotic fluids and the placenta. These organisms end up in the environment where they may survive for long periods of time due by their resistant to heat, drying, and many common disinfectants. Humans are often very susceptible to the disease, and very few organisms may be required to cause infection. More information about this infectious disease is provided in Maurin and Raoult (1999).

The south of the Netherlands faced large outbreaks of human Q fever from 2007 to 2010. In this country, local municipal health services (MHSs) are responsible for recording every confirmed diagnosis of acute Q fever. The information collected is then entered into the electronic national infectious diseases surveillance database. Due to confidentiality, these data are not publicly available and, in some cases, may be provided in an aggregated form.

Figure 1 shows the temporal distribution of Q fever cases (in months) from January 2007 to July 2010. A total of 3807 acute Q fever cases were registered in this period: 192 in 2007, 980 in 2008, 2309 in 2009, and 325 in 2010. The epidemic peaks of each year are observed every spring, specifically during May. This coincides with small ruminants (sheep and goats) birth period — a fact that was pointed out in several studies about this exceptionally large Q fever outbreaks in

the Netherlands (see, for example, van der Hoek et al., 2010; Roest et al., 2011). Since the largest outbreak was observed during 2009, we will study the distribution of Q fever incidence in this year.

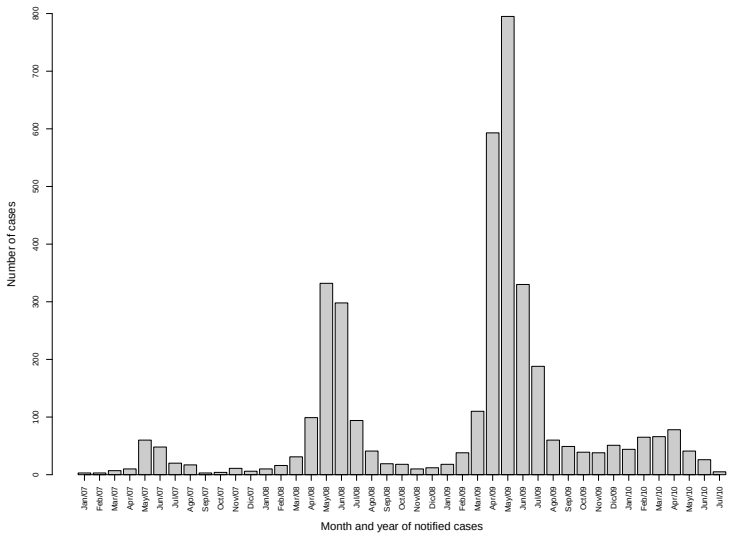

Figure 1: *Human Q fever cases in the Netherlands grouped per months, from January 2007 to July 2010.*

Figure 2a shows the geographical distribution of the residential addresses of human Q fever cases (red points) in 2009. If we select a 60 $\times$ 60 km area in the south of the Netherlands (see black square in Figure 2a), 72 municipalities overlap with this study area. The total number of Q fever cases reported in these municipalities is 1798. Aggregating these cases per municipality, and taking into account the number of inhabitants of each municipality, we can calculate the Q fever incidence (per 100000 inhabitants). Figure 2b shows the spatial distribution of the resulting Q fever incidence, where higher incidence values are observed around the municipalities of Landerd (1439.676), Lith (562.546), and Heusden (295.006).

3.2 Detailed smooth incidence maps

Figure 2b shows the choropleth map of raw Q fever incidences at municipality level. However, it would be desirable to estimate latent incidence maps at detailed periods of time, in order to obtain a better insight of the evolution of the incidence. The CLMM approach, developed in Section 2, allows to visualize the Q fever incidence at a finer spatio-temporal resolution, and also to incorporate population information at fine scale into the estimation of the latent process. Here, we illustrate the application of our methodology using Q fever data collected over municipalities and months in 2009. Our goal is to obtain Q fever incidence estimates at a fine spatial grid over the study area, and at each week of 2009, from data collected over municipalities and months (i.e., we want to simultaneously disaggregate in space and time).

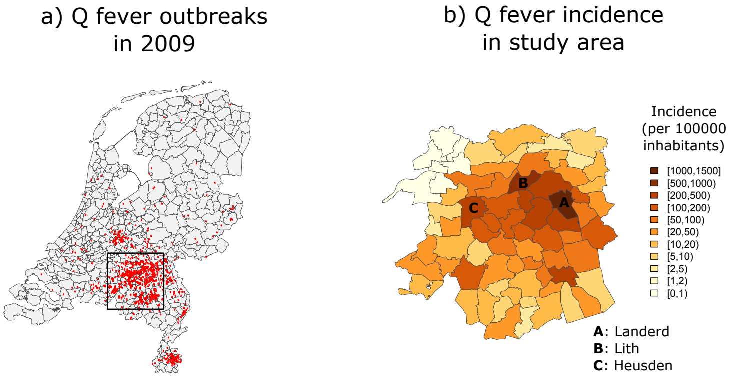

Figure 2: *Map of human Q fever cases in the Netherlands, 2009. Left: the red points indicate the residential addresses of human cases (2309 in total). Right: study area in the south of the Netherlands showing (crude) incidence (per 100000 inhabitants) of Q fever by municipality in 2009.*

Figure 3a shows a fine grid composed of 4871 regular cells of size 1000×1000 m. The blue dots represent the spatial coordinates of the centroids of these cells, and Figure 3b shows the spatial distribution of the population on this fine grid (obtained from CBS - Statistics Netherlands, https://www.cbs.nl/nl-nl.), which is heterogeneous across municipalities. To provide a more detailed impression of the Q fever incidence in 2009, we apply the CLMM approach on data recorded over municipalities and months. We choose the fine grid displayed in Figure 3a as the spatial refinement, and the weeks of 2009 as the temporal fine scale. To set up the CLMM formulation, we use the spatial coordinates of the grid points as spatial covariates at fine resolution, i.e., $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\boldsymbol{x}_3 = (1, ..., 53)'$ (since 53 weeks are observed in 2009), second order penalties, 12 equally-spaced knots for the marginal cubic B-spline bases $\mathbf{B}_1$ and $\mathbf{B}_2$, and 8 equally-spaced knots for the marginal cubic B-spline basis $\mathbf{B}_3$. We assume here that the population at the fine grid is constant over the time period; thus $\boldsymbol{e}_\mathrm{f}$ in Eq. (6) is considered as a vector obtained by repeating the fine-scale population fifty three times. The spatial composition matrix is obtained as is described in Eq. (5), whereas the temporal composition matrix for this case has the following form:

$$\mathbf{C}_t = \left[\begin{array}{*{53}{c}}
1&1&1&1&1 \\
&&&&&1&1&1&1 \\
&&&&&&&&&1&1&1&1&\frac{3}{7} \\
&&&&&&&&&&&&&\frac{4}{7}&1&1&1&\frac{5}{7} \\
&&&&&&&&&&&&&&&&&\frac{2}{7}&1&1&1&1&\frac{1}{7} \\
&&&&&&&&&&&&&&&&&&&&&&\frac{6}{7}&1&1&1&\frac{3}{7} \\
&&&&&&&&&&&&&&&&&&&&&&&&&&\frac{4}{7}&1&1&1&\frac{6}{7} \\
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&\frac{1}{7}&1&1&1&1&\frac{2}{7} \\
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&\frac{5}{7}&1&1&1&\frac{4}{7} \\
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&\frac{3}{7}&1&1&1&1 \\
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&1&1&1&1&\frac{2}{7} \\
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&\frac{5}{7}&1&1&1&1
\end{array}\right],$$

where Sunday is considered as the first day of the week. As opposed to the spatial composition matrix, matrix $\mathbf{C}_t$ has some entries that are fractions. This is because some months share parts of a specific week (for example, some days of week 14 belong to March and other to April).

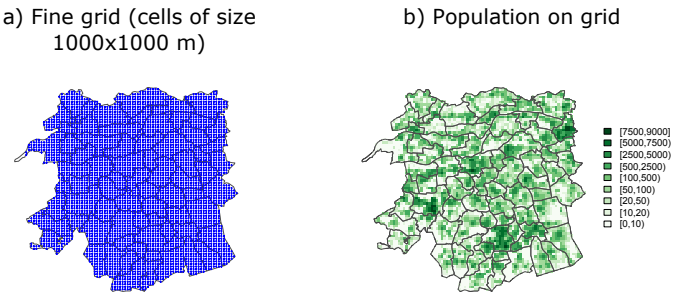

Figure 3: *The left map shows the fine grid of cell sizes 1000× 1000 m in the study area showed in Figure 2b. The right map shows the spatial distribution of the population on this fine grid.*

The dynamic map in Figure 4a shows the resulting CLMM incidence (per 100000 inhabitants) for Q fever disease at the desire spatio-temporal resolution. These incidences are obtained as $\widehat{inc} = 100000 \times \exp(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}})$. The evolution of the incidence vary across municipalities and weeks, where higher incidences are mostly observed between weeks 13 and 28. Most of these weeks belong to the months of April, May, and June, which have the largest observed number of Q fever outbreaks in 2009 (see Figure 1). Notice also that most of the higher incidences in these weeks are spatially concentrated around the area that involves points A and C in Figure 4a, which are located in the municipalities of Landerd and Heusden, respectively (see Figure 2b). The dynamic map in Figure 4b shows the approximate standard error maps associated with Figure 4a. Higher errors are located at the boundary of the study area and at the beginning and the end of the time period. This is because we have less information in these parts, and thus the CLMM estimates are more variable there.

From the previous CLMM estimates, we can also visualize the temporal evolution of the Q fever disease at a specific spatial coordinate of the fine grid. Figure 5a shows the smoothed temporal incidence (per week) at three random locations A, B, and C, of the study area. We observe the temporal evolution of the incidence in point B is constant and almost zero, whereas, in points A and C, the temporal smoothed incidence present a unimodal behaviour, where the peak is reached around week 19 (of the month of May). Figure 5b shows the spatial trend of the Q fever incidence in this week.

4 Simulation study

In this section we perform a simulation study in order to examine the prediction performance of the CLMM approach described in Section 2. For that we use a detailed dataset, where the full postal code of the home address and the date of onset of illness of the patients suffering from acute Q fever in 2009 are available. This dataset is publicly restricted due to confidentiality reasons, and

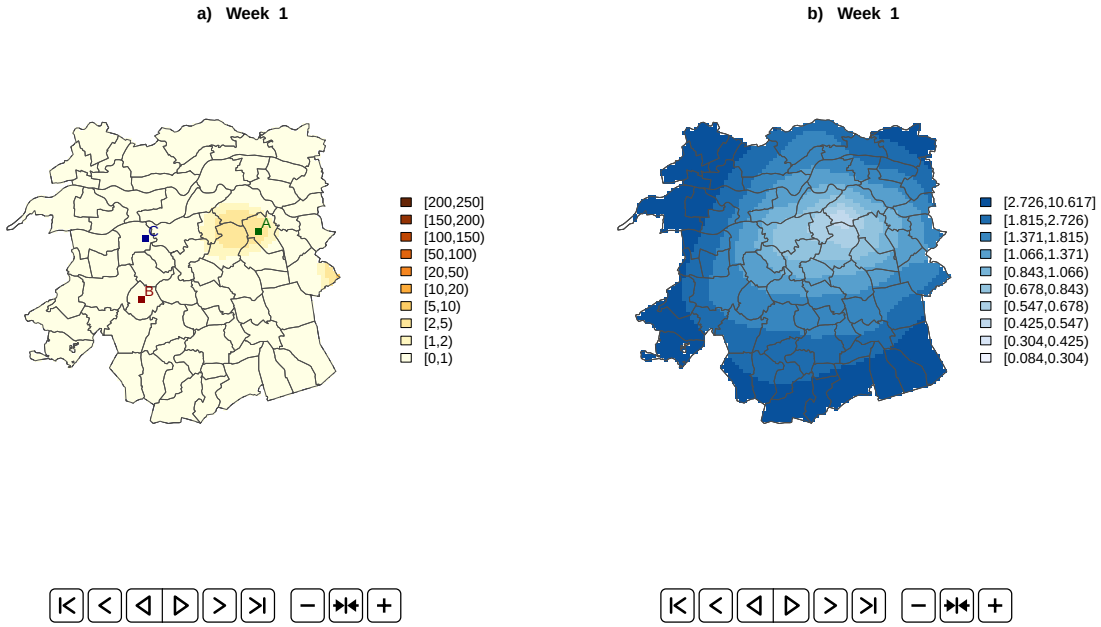

Figure 4: *The left dynamic map shows the smoothed Q fever incidence at a detailed spatio-temporal scale, resulting from the CLMM approach. The right dynamic map shows the approximate standard errors associated with the left dynamic map (at log scale).*

it is used here to asses how well the CLMM approach recovers the latent Q fever incidence, when data are available at a specific coarse scale.

Using the full postal code of home address of patients, we can determine the number of cases occurred per weeks and in each cell of the fine grid depicted in Figure 3a. Thus, using the population on the fine grid depicted in Figure 3b, we can obtain smoothed incidence estimates at this fine scale using the PGLMM approach (not showed here), where the spatial coordinates of the fine grid in Figure 3a are used as spatial covariates, and the number of weeks is used as temporal covariate. These smoothed incidences are considered here as estimates for the true latent incidence trend at the fine resolution, which are denoted as $inc(\boldsymbol{u}_j)$, where $\boldsymbol{u}_j$, $j = 1, ..., J$, with $J = 4371 \times 53 = 258163$, represents the spatio-temporal coordinates at fine resolution. To study the prediction performance of the CLMM approach, we artificially aggregate these Q fever incidence trends in several coarse periods of time.

The simulation study was conducted as follows:

1. The fine-scale smoothed incidences and the population on the fine grid were used to compute the Q fever incidence for each municipality v_i, $i = 1, ..., 72$ (which are depicted in Figure 2b), and in different coarse temporal resolutions: 1) fortnights (two weeks); 2) months; and 3) bimesters (two months). Thus we have three types of spatio-temporal aggregation: $g =$

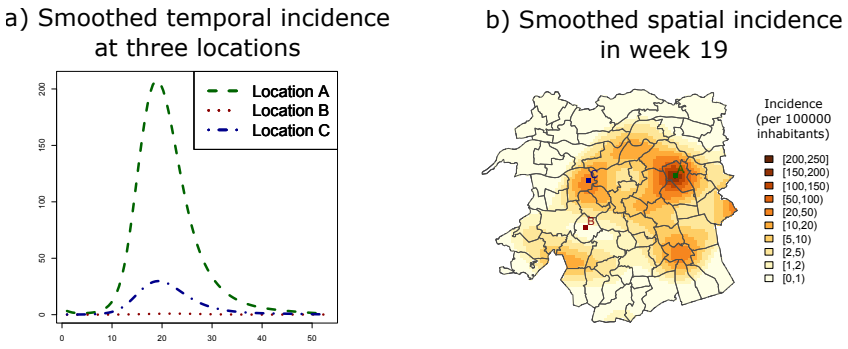

Figure 5: *The right figure shows the temporal evolution of Q fever incidence in three specific points (A, B, and C), spatially depicted on the map at the left.*

1 where data are summarized over municipalities and fortnights, $g = 2$ where data are summarized over municipalities and months, and $g = 3$ where data are summarized over municipalities and bimesters.

2. 100 realizations of the number of cases recorded over each municipality and each coarse temporal resolution were generated by random drawing of a Poisson distribution whose mean parameter is calculated as the corresponding aggregated incidence (obtained in the previous step) times the population recorded at the appropriate coarse resolution.

3. For each realization, we apply the CLMM approach using the population on the fine grid (repeated 53 times) as the vector $\boldsymbol{e}_{\mathrm{f}}$ of exposures at the fine resolution.

For all $l = 1, ..., 100$ realizations, the predicted incidence $inc_{\mathrm{P}_g}^{(l)}(\boldsymbol{u}_j)$ obtained from the CLMM approach of each type of aggregation g, with $g = 1, 2, 3$, were compared to the smoothed incidences $inc(\boldsymbol{u}_j)$, $j = 1, ..., J$, using the following criteria:

- Mean absolute error (MAE):

$$\mathrm{MAE}_g^{(l)} = \frac{1}{J}\sum_{j=1}^{J}\left|inc_{\mathrm{P}_g}^{(l)}(\boldsymbol{u}_j) - inc(\boldsymbol{u}_j)\right|$$

- Root mean squared error (RMSE):

$$\mathrm{RMSE}_g^{(l)} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(inc_{\mathrm{P}_g}^{(l)}(\boldsymbol{u}_j) - inc(\boldsymbol{u}_j)\right)^2}$$

Figure 6 shows these resulting errors via box-plots for the different types of aggregations, and Table 2 gives the averages and the standard deviations of the resulting errors (for each criterion)

derived from the simulation study. As we could expect we have found the CLMMs estimates that were obtained from the most coarser spatio-temporal aggregation (type of aggregation 3) are less similar to the true incidences. Notice that these are overall results and the spatial support remains the same in all type of aggregations (municipalities). Similar performances as in Figure 6 are obtained if the errors are analysed by weeks. Since no methodologies exists for the simultaneous disaggregation of health data in space and time, we cannot compare the prediction performance of our approach with other techniques.

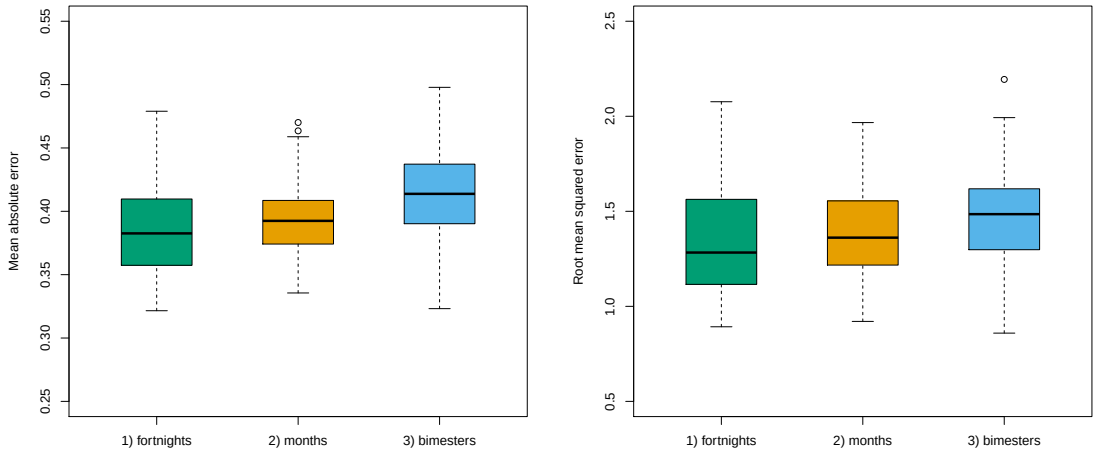

Figure 6: *Performance comparison of the CLMM approach under three different types of aggregation (1: municipalities and fortnights; 2: municipalities and months; 3: municipalities and bimesters), using different criteria: mean absolute errors (left) and root mean squared errors (right).*

Type of	MAE		RMSE	
aggregation	avg	std	avg	std
$g = 1$: fortnights	0.3856	0.0321	1.3413	0.2774
$g = 2$: months	0.3935	0.0276	1.3933	0.2327
$g = 3$: bimesters	0.4142	0.0326	1.4625	0.2469

Table 2: *Performance comparison of the CLMM approach under three different types of aggregation (1: municipalities and fortnights; 2: municipalities and months; 3: municipalities and bimesters), using different criteria: mean absolute errors (MAE) and root mean squared errors (RMSE). These errors are summarized in terms of the average (avg) and standard deviation (std).*

5 Discussion

We presented and applied the composite link mixed model approach to the disaggregation of grouped data both in space and time. The model allows to obtain detailed trends of disease incidence, mortality risks, or any other vital rates at a desirable fine resolution. Thus the resulting CLMM outcomes can be displayed as a dynamic map. Also it allows to include population information at a fine resolution into the estimation. Here we illustrate the case when this fine-scale population is available at a fine regular grid. Our proposal can also with situations when population information is recorded over small geographical units that are nested in coarser units (for example, from municipalities to census tracts). In that case the centroids of the small units can be used as covariates at the fine scale. It is possible to incorporate other covariates (such as, socio-economic, demographic, and environmental factors) into the CLMM formulation in order to improve the estimation of the latent trend. We performed a simulation study to see the prediction accuracy of the CLMM using Q fever data recorded at different temporal resolutions (keeping fixed the spatial support). As it was expected, our approach is able to properly capture the underlying trend when the original spatio-temporal resolution is not so coarse.

It is important to acknowledge the use of GLAM methods (whose usage is described in Section 2.2.2), together with SAP algorithm, to avoid storage problems and to speed up computations. However, we are aware that the disaggregation of grouped data into a very detailed resolution could lead to the incrementation of the computational burden and storage problems. The sparsity of the marginal composition matrices can be exploited to deal with these issues (see, for example, Bates and Maechler, 2016).

In the CLMM formulation we assume the geographical unit boundaries remain fixed through time. But it can occur that some of these boundaries change during the years. This issue is known in the statistical literature as the *spatio-temporal misalignment problem*, where some works were proposed to solve it (see, for example, Zhu et al., 2000; Zhu and Carlin, 2000; Hund et al., 2012). As future work we plan to extend the CLMM approach in order to handle this problem, where the marginal composition matrices will play an important role.

The implementation of our proposal and all data analyses were carried out using the statistical software R (R Core Team, 2016).

Acknowledgements

This research was supported by the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-02 and MTM2014-52184-P. 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

Ayma, D., Durbán, M., Lee, D.-J., and Eilers, P. H. C. (2016). Penalized composite link models for aggregated spatial count data: a mixed model approach. *Spatial Statistics*.

Bates, D. and Maechler, M. (2016). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-4.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in Medicine*, 35(11):1848–1865.

Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(18).

Bindhu, V. M. and Narasimhan, B. (2015). Development of a spatio-temporal disaggregation method (DisNDVI) for generating a time series of fine resolution NDVI images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 101:57–68.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 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. *Statistical Science*, 28(4):542–563.

Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7(3):239–254.

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

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

Eilers, P. H. C., Marx, B. D., and Durbán, M. (2015). Twenty years of P-splines. *Statistics and Operations Research Transactions (SORT)*, 39(2):149–186.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, Florida.

Goovaerts, P. (2006). 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. *International Journal of Health Geographics*, 5(52).

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hund, L., Chen, J. T., Krieger, N., and Coull, B. A. (2012). A geostatistical approach to large-scale disease mapping with temporal misalignment. *Biometrics*, 68(3):849–858.

Lee, D.-J. and Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, 11(1):49–69.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):381–400.

MacNab, Y. C. and Dean, C. B. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57:949–956.

Martínez-Beneito, M. A., López-Quilez, A., and Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27(15):2874–2889.

Maurin, M. and Raoult, D. (1999). Q fever. *Clinical Microbiology Reviews*, 12(4):518–553.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Prairie, J., Rajagopalan, B., Lall, U., and Fulp, T. (2007). A stochastic nonparametric technique for space-time disaggregation of streamflows. *Water Resources Research*, 43(3).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M., and Eilers, P. H. C. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, 25(5):941–957.

Roest, H. I. J., Tilburg, J. J. H. C., van der Hoek, W., Vellema, P., van Zijderveld, F. G., Klaassen, C. H. W., and Raoult, D. (2011). The Q fever epidemic in The Netherlands: history, onset, response and reflection. *Epidemiology & Infection*, 139(1):1–12.

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

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

Schleiss, M. and Berne, A. (2012). Stochastic space-time disaggregation of rainfall into DSD fields. *Journal of Hydrometeorology*, 13(6):1954–1969.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6):725–734.

Segond, M.-L., Neokleous, N., Makropoulos, C., Onof, C., and Maksimovic, C. (2007). Simulation and spatio-temporal disaggregation of multi-site rainfall data for urban drainage applications. *Hydrological Sciences Journal*, 52(5):917–935.

Thompson, R. and Baker, R. J. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 30(2):125–131.

Ugarte, M. D., Adin, A., Goicoa, T., and Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate bayesian inference. *Statistical Methods in Medical Research*, 23(6):507–530.

Ugarte, M. D., Goicoa, T., and Militino, A. F. (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics*, 21(3–4):270–289.

van der Hoek, W., Dijkstra, F., Schimmer, B., Schneeberger, P. M., Vellema, P., Wijkmans, C., ter Schegget, R., Hackert, V., and van Duynhoven, Y. (2010). Q fever in the Netherlands: an update on the epidemiology and control measures. *Eurosurveillance*, 15(12):pii:19520.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438):607–617.

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

Zhu, L. and Carlin, B. P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19(17-18):2265–2278.

Zhu, L., Carlin, B. P., English, P., and Scalf, R. (2000). Hierarchical modeling of spatio-temporally misaligned data: relating traffic density to pediatric asthma hospitalizations. *Environmetrics*, 11(1):43–61.

Appendices

Appendix A Definitions used within GLAM methods

In this section we include some definitions and notations related with GLAM methods, which are used in Section 2.2.2.

Definition 1 (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 $\mathbf{1}_k$ *denotes a vector of ones of length* k *and* $\odot$ *represents the Hadamard product.*

Definition 2 ($\mathcal{H}$-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* $r \times c_2 \times \cdots \times c_d$ *that is obtained from* $\mathbf{X}\mathbf{A}^*$ *by reinstating dimensions* 2*-*d *of* $\mathbf{A}$*.*

If $\mathbf{A}$ is a vector, i.e., $\mathbf{A} = \boldsymbol{a}$, then we have that $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\boldsymbol{a}$, whereas if $\mathbf{A}$ is a matrix, $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\mathbf{A}$. Thus, the $\mathcal{H}$-transform generalizes premultiplication of vector and matrices by a matrix. The following definition generalizes the transpose of a matrix.

Definition 3 (Array rotation) *The rotation of the* d*-dimensional array* $\mathbf{A}$ *of size* $c_1 \times c_2 \times \cdots \times c_d$ *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* $\mathbf{A}$*.*

Combining the last two definitions, we obtain:

Definition 4 (Rotated $\mathcal{H}$-transform) *The rotated* $\mathcal{H}$*-transform of the array* $\mathbf{A}$ *by the matrix* $\mathbf{X}$ *is given by:*

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

Appendix B Proof for closed-form expressions of the variance components

In this section we provide the proof for the closed-form expressions of the variance components given in Eq. (16). We should note that the following proof is similar to that given in Rodríguez-Álvarez et al. (2015), but now the 'working' mixed model matrices $\breve{\mathbf{X}}$ and $\breve{\mathbf{Z}}$ (as well as related matrices like $\mathbf{V}$ and $\mathbf{N}$) appear. Here the formulation of the matrix $\mathbf{G}^{-1}$ in Eq. (8) with $\lambda_d = \frac{1}{\tau_d^2}$ $(d = 1, 2, 3)$ is used.

Consider the approximate restricted maximum log-likelihood (REML) version of Patterson and Thompson (1971):

$$l^* = -\frac{1}{2}\underbrace{\log|\mathbf{V}|}_{\textbf{Part I}} - \frac{1}{2}\underbrace{\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}_{\textbf{Part II}} - \frac{1}{2}\underbrace{(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})}_{\textbf{Part III}}, \tag{19}$$

in which the dependence of the matrix of weights $\mathbf{W}$ on τ_d^2 $(d = 1, 2, 3)$ is ignored. To obtain the REML estimates of the variance components τ_d^2, we first differentiate each part of Eq. (19) with respect to them. Using the fact that $\frac{\partial \mathbf{V}}{\partial \tau_d^2} = \breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'$, with $\mathbf{V}$ defined in Eq. (14), we obtain:

Part I: Here we use property (8.6) given in Harville, 1997, p.305:

$$\frac{\partial \log|\mathbf{V}|}{\partial \tau_d^2} = \text{trace}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \tau_d^2}\right) = \text{trace}\left(\mathbf{V}^{-1}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'\right)$$

Part II: Here we use properties (8.6) and (8.18) given in Harville, 1997, pp. 305, 308:

$$\begin{aligned}
\frac{\partial \log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}{\partial \tau_d^2} &= \text{trace}\left((\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\frac{\partial(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})}{\partial \tau_d^2}\right) \\
&= -\text{trace}\left((\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \tau_d^2}\mathbf{V}^{-1}\breve{\mathbf{X}}\right) \\
&= -\text{trace}\left(\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \tau_d^2}\right) \\
&= -\text{trace}\left(\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'\right)
\end{aligned}$$

Part III: Here we use property (8.6) given in Harville, 1997, p. 308:

$$\begin{aligned}
\frac{\partial(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})}{\partial \tau_d^2} &= -(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \tau_d^2}\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}) \\
&= -(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}) \\
&= -\boldsymbol{b}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\boldsymbol{b} \\
&= -\boldsymbol{\alpha}'\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\mathbf{G}^{-1}\boldsymbol{\alpha}
\end{aligned}$$

Adding the derivatives obtained from Part I and Part II, we obtain:

$$\begin{aligned}\frac{\partial \log|\mathbf{V}|}{\partial \tau_d^2}+\frac{\partial \log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}{\partial \tau_d^2} &= \text{trace}\left((\mathbf{V}^{-1}-\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1})\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'\right)\\ &= \text{trace}\left(\mathbf{N}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'\right)\\ &= \text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\right),\end{aligned}$$

It follows that:

$$\frac{\partial l^*}{\partial \tau_d^2}=-\frac{1}{2}\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\right)+\frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\mathbf{G}^{-1}\boldsymbol{\alpha}, \tag{20}$$

where:

$$\frac{\partial \mathbf{G}}{\partial \tau_d^2}=-\mathbf{G}\frac{\partial \mathbf{G}^{-1}}{\partial \tau_d^2}\mathbf{G}=-\mathbf{G}\left(-\frac{1}{\tau_d^4}\boldsymbol{\Lambda}_d\right)\mathbf{G}=\frac{1}{\tau_d^4}\mathbf{G}\boldsymbol{\Lambda}_d\mathbf{G}, \tag{21}$$

for $d=1,2,3$.

By replacing Eq. (21) in Eq. (20), we obtain:

$$\frac{\partial l^*}{\partial \tau_d^2}=-\frac{1}{2\tau_d^2}\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\boldsymbol{\Lambda}_d}{\tau_d^2}\mathbf{G}\right)+\frac{1}{2\tau_d^4}\boldsymbol{\alpha}'\boldsymbol{\Lambda}_d\boldsymbol{\alpha}.$$

Therefore, the REML estimates of the variance components τ_d^2 are found by equating the expression above by zero. These estimates are given by:

$$\hat{\tau}_d^2=\frac{\boldsymbol{\alpha}'\boldsymbol{\Lambda}_d\boldsymbol{\alpha}}{\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\boldsymbol{\Lambda}_d}{\tau_d^2}\mathbf{G}\right)},\ \text{for } d=1,2,3.$$

Notice that if we add the denominators of the previous expressions, we obtain:

$$\begin{aligned}\sum_{d=1}^{3}\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\boldsymbol{\Lambda}_d}{\tau_d^2}\mathbf{G}\right) &= \text{trace}\left(\sum_{d=1}^{3}\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\boldsymbol{\Lambda}_d}{\tau_d^2}\mathbf{G}\right)\\ &= \text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\right)\\ &= \text{trace}\left(\breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'\mathbf{N}\right),\end{aligned}$$

where $\breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'\mathbf{N}$ is the 'hat matrix' (Hastie and Tibshirani, 1990) of the unpenalized (or random) part of the fitted CLMM (see Eq. (13)).

Appendix C SAP Algorithm for CLMM's parameters estimation

Algorithm 1 SAP Algorithm for CLMM's parameters estimation

Require: Convergence tolerances ν_1 and ν_2 (e.g., 1×10^{-6}), and maximum number of iterations maxit_1 and maxit_2 (e.g., 100).

1: Set initial values for the mixed model coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, and the variance components τ_1^2, τ_2^2, and τ_3^2 (for example, $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ with length $q_1 q_2 q_3$, $\widehat{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}$ with length $(c_1 c_2 c_3 - q_1 q_2 q_3)$, and $\widehat{\tau}_1^{2(0)} = \widehat{\tau}_2^{2(0)} = \widehat{\tau}_3^{2(0)} = 1$). Set $k = 0$

2: **for** 1 **to** maxit_1 **do**

3: Given the current mixed model coefficients' estimates, construct the matrix of weights $\mathbf{W}$ and the working vector $\boldsymbol{z}$ as follows:

$$\begin{aligned}
\boldsymbol{\Gamma} &= \text{diag}(\widehat{\boldsymbol{\gamma}}^{(k)}) \text{ , with } \widehat{\boldsymbol{\gamma}}^{(k)} = \boldsymbol{e}_{\text{f}} * \exp(\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}^{(k)}); \\
\mathbf{W} &= \text{diag}(\widehat{\boldsymbol{\mu}}^{(k)}) \text{ , with } \widehat{\boldsymbol{\mu}}^{(k)} = \mathbf{C}\widehat{\boldsymbol{\gamma}}^{(k)}; \\
\boldsymbol{z} &= \breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}^{(k)} + \breve{\mathbf{Z}}\widehat{\boldsymbol{\alpha}}^{(k)} + \mathbf{W}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}^{(k)}) = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}^{(k)} + \mathbf{W}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}^{(k)}),
\end{aligned}$$

with $\widehat{\boldsymbol{\eta}}^{(k)} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}^{(k)}$.

4: **for** 1 **to** maxit_2 **do**

5: Given the current estimates for the variance components, obtain new estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by solving the linear system given in Eq. (11). The resulting estimates are denoted as $\widehat{\boldsymbol{\beta}}^{(k+1)}$ and $\widehat{\boldsymbol{\alpha}}^{(k+1)}$, respectively.

6: Obtain new estimates for the variance components from Eq. (16). The resulting estimates are denoted as $\widehat{\tau}_d^{2(k+1)}$, for $d = 1, 2, 3$.

7: Compare new variance components' estimates with the previous ones, using the following convergence criterion:

$$\frac{\sum_{d=1}^{3} |\widehat{\tau}_d^{2(k+1)} - \widehat{\tau}_d^{2(k)}|}{3} \leq \nu_1.$$

8: If the convergence tolerance is achieved, **break**, otherwise set $\widehat{\tau}_d^{2(k)} = \widehat{\tau}_d^{2(k+1)}$ and repeat steps 5, 6, and 7 until convergence.

9: **end for**

10: Compute new estimates for the vector of smooth trends at fine scale using the model's fixed and random effects estimates obtained in the last iteration of step 5. The resulting vector is denoted as $\widehat{\boldsymbol{\eta}}^{(k+1)}$. Compare new estimates with the previous ones, using the following convergence criterion:

$$\frac{\|\widehat{\boldsymbol{\eta}}^{(k+1)} - \widehat{\boldsymbol{\eta}}^{(k)}\|^2}{\|\widehat{\boldsymbol{\eta}}^{(k+1)}\|^2} \leq \nu_2,$$

11: If the convergence tolerance is achieved, **break**, otherwise set $k = k + 1$ and repeat the previous steps until convergence.

12: **end for**