Penalized composite link mixed models for spatial count data

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

[1]dayma@est-econ.uc3m.es, Department of Statistics, Universidad Carlos III de Madrid
[2]mdurban@est-econ.uc3m.es, Department of Statistics, Universidad Carlos III de Madrid
[3]dlee@bcamath.org, BCAM - Basque Center for Applied Mathematics

Abstract

In this paper, we propose the use of penalized composite link model and its representation as a mixed model for filtering noisy mortality rates, and mapping the corresponding risk at a fine spatial resolution. To illustrate our proposal, we analyse mortality data recorded in the community of Madrid at municipality level, over the period 2001-2007; our model will disaggregate the data to obtain a continuous surface for mortality risk across municipalities.

Keywords: Penalized composite link models, mortality rates, spatial disaggregation.

1. Introduction

Disease maps deal with public health data that are usually available in an aggregated form over geographical units. This is done to protect patient 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 used to display such distribution but they must be interpreted with caution, since the "small number problem" effect [5] — that often affects health data — leads to a large uncertainty about rates calculated from small or sparsely populated areas. Another problem that could arise is the spatial misalignment between potential risk factors and health data: the former are, in general, available on a finer spatial resolution than the latter. This situation prohibits their direct use in correlation analysis, which is a critical step in a disease control intervention. Therefore, it is important to develop spatial tools 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.

In this paper, we propose to use the penalized composite link model [2] in the case of spatial aggregation, and its representation as a mixed model. The resulting model, which we call penalized composite link mixed model, allows to create mortality maps from aggregated health data at a desirable spatial resolution, and to incorporate fine scale information into the filtering of noisy mortality rates. Here, we illustrate the case when we seek to estimate the spatial trend of mortality in a fine grid, from aggregated data available at coarse geographical units (area-to-point (or ATP) case). We create a continuous surface, or isopleth map, that reduces the visual bias associated with the interpretation of choropleth maps, which is due to the variation in shape and size of the units.

2. The penalized composite link mixed 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, developed in [2], offers an elegant way to do this, by considering $\boldsymbol{\mu}$ composed by latent expectations. The Poisson PCLM is given by:

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

In (1), $\boldsymbol{\gamma}$ 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.

Now, suppose that the aggregated counts $\boldsymbol{y}$ are available over n non-overlapping geographical units. A first attempt to estimate the spatial trend of mortality could be the penalized generalized linear mixed model (or more briefly, PGLMM) approach given in [4], but it assumes that mortality rate estimates are constant over each unit. We propose an improved smooth spatial model, by extending (1) to the spatial setting, and representing it as a mixed model, as follows.

Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the geographical coordinates (longitude and latitude, respectively) of length m that define the fine spatial resolution. Then, in this new context, the regression basis $\mathbf{B}$ is defined as the "row-wise" Kronecker product [3] 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:

$$\mathbf{B} = \mathbf{B}_2 \square \mathbf{B}_1 = (\mathbf{B}_2 \otimes \mathbf{1}'_{c_1}) \odot (\mathbf{1}'_{c_2} \otimes \mathbf{B}_1), \tag{2}$$

and the two-dimensional penalty matrix is given by:

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

where $\mathbf{P}_i = \mathbf{D}'_i\mathbf{D}_i$ is based on the difference matrix $\mathbf{D}_i$ of order d_i, for $i = 1, 2$. Considering (2) and (3), it was shown in [4] that the expression $\mathbf{B}\boldsymbol{\theta}$ can be reformulated as $\mathbf{B}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$, where $\mathbf{X}$ and $\mathbf{Z}$ are the fixed and random effects matrices, respectively, and the new penalty becomes a block diagonal matrix denoted as $\mathbf{F}$, which depends on the smoothing parameters λ_1 and λ_2. Therefore, we can extend (1) by modifying $\boldsymbol{\gamma}$ as follows:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\boldsymbol{e}\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}), \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}(\lambda_1, \lambda_2)\right), \tag{4}$$

where $\mathbf{G}(\lambda_1, \lambda_2) = \sigma^2_\epsilon \boldsymbol{F}^{-1}$ and $\sigma^2_\epsilon = 1$ (Poisson case). We refer to (4) as the penalized composite link mixed model, or more briefly PCLMM. This model enables the inclusion of area specific random effects or further correlation structure if necessary, and includes a vector $\boldsymbol{e}$ of exposures

(e.g., expected number of deaths, population-at-risk), which allows to incorporate population information at the fine spatial resolution. The elements of the composition matrix $\mathbf{C}$ for the ATP case is chosen as:

$$c_{ij} = \begin{cases} 1 & \text{if } (x_{1j}, x_{2j}) \text{ belongs to municipality } i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $i = 1, ..., n$ and $j = 1, ..., m$.

Finally, considering the joint density function of $\boldsymbol{y}$ in a PCLMM context,

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

with $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} = \boldsymbol{e} \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})$, we can obtain estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by maximizing the penalized log-likelihood $\ell_{pen} = \log\{f(\boldsymbol{y}|\boldsymbol{\alpha})\} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\boldsymbol{\alpha}$, which yields to a modified version of the standard mixed model estimators:

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

where $\breve{\mathbf{X}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$, $\breve{\mathbf{Z}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$, and $\mathbf{V} = \mathbf{W}^{-1} + \breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'$, with $\mathbf{W} = diag(\boldsymbol{\mu})$, $\boldsymbol{\Gamma} = diag(\boldsymbol{\gamma})$, and working vector $\boldsymbol{z} = \breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{Z}}\boldsymbol{\alpha} + \mathbf{W}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. Smoothing parameters λ_1 and λ_2 are estimated using the penalized quasi-likelihood approach (PQL) given in [1].

3. Illustration

Our data come from a large European epidemiological study called MEDEA (for more information, see http://www.proyectomedea.org) and correspond to the number of observed and expected female deaths by cardiovascular diseases in the community of Madrid, over the period 2001-2007, which are available at municipality level. Suppose that we want to create a continuous mortality trend across these municipalities. For that, we can impose a fine grid over the community and select the points (with coordinates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) that fall inside of each municipality. Figure 1 shows the map of the municipalities in the community of Madrid and the 100×100 grid chosen. Now, we use the model given in (4) for the area-to-point case (ATP-PCLMM) to produce such

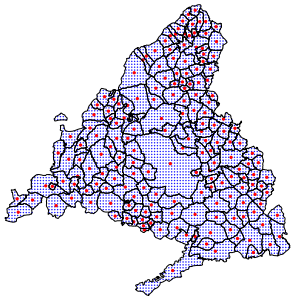

Figure 1: Map of the community of Madrid. The red and blue points represent the centroids of the 179 municipalities and the 4359 grid points selected, respectively.

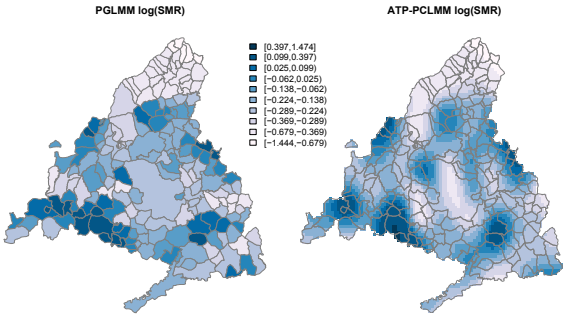

Figure 2: Spatial distribution of the resulting smoothed log(SMR) using the PGLMM approach to raw data (left) and ATP-PCLMM approach (with $\hat{\boldsymbol{e}}_{\text{naive}}$) (right). The color legend applies to both maps, where the class boundaries correspond to the deciles of raw log(SMR).

continuous trend. Due to the lack of expected deaths at this point-level, we have to estimate them. A naive way is to assume that the exposures are evenly distributed throughout the selected grid points at each municipality. We denote these estimates as $\hat{\boldsymbol{e}}_{\text{naive}}$. In Figure 2, the left map shows the spatial distribution of the natural logarithm of standardized mortality rates (log(SMR)) obtained by applying the PGLMM approach to raw data, and the right map shows the resulting smoothed log(SMR) using the ATP-PCLMM approach. The ATP-PCLMM log(SMR) map gives more details than the PGLMM log(SMR) map, where most of the higher rates are in the boundaries of the community of Madrid, specially in the south-western area.

Finally, the presented PCLMM approach can also be used to accommodate the area-to-area (or ATA) case, that is, when we seek to visualize mortality trend at smaller geographical units, using aggregated data available at coarse units. In this case, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be chosen as the centroids of the smaller units, and the matrix $\mathbf{C}$ will have a similar structure as in (5).

4. Acknowledgments

This research has been funded by the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-02 and MTM2014-52184.

5. Bibliography

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

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

[3] Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Comput Stat Data An*, 11(2):89-121.

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

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