Spatial Disaggregation in Epidemiology

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

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

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

January 20, 2015

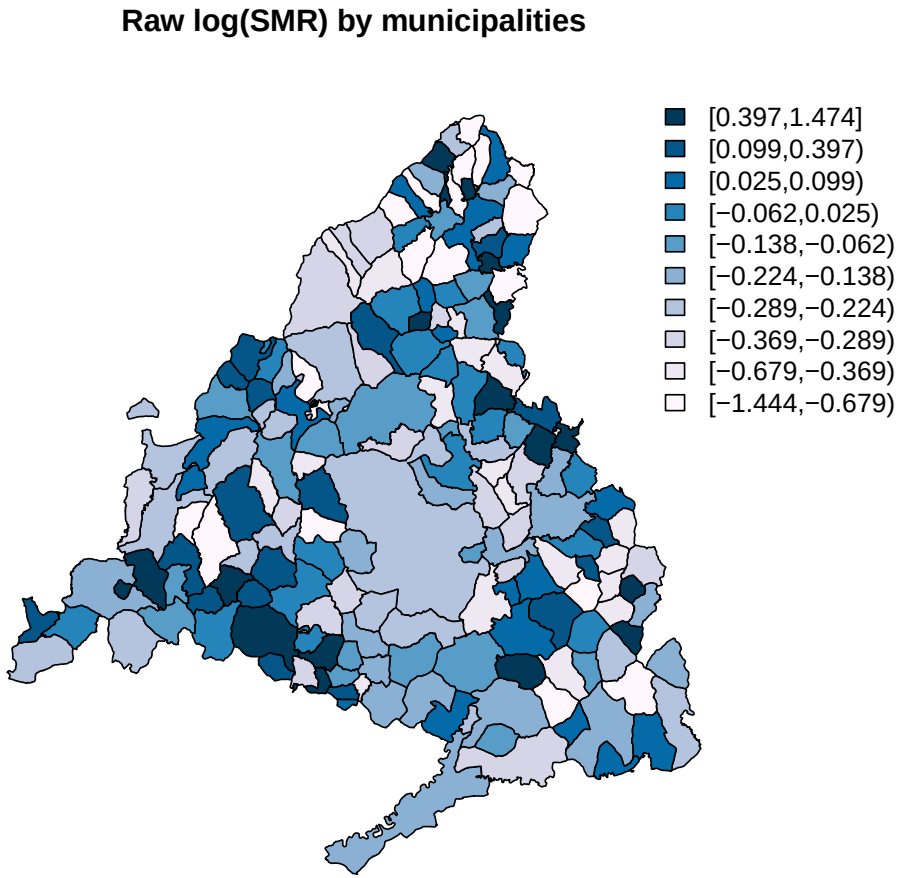
Motivation

- Disease mapping studies deal with data that are, in general, aggregated over geographical units.
- A choropleth map provides an essential tool for the description of these data:
  - ✓ Quick visualization of spatial distribution of rates or risks.
  - ⚠ Choice of the colour scale in the legend map is arbitrary! Epidemiologists prefer to use quantiles as class boundaries.

Motivation

Example

- Standardized mortality ratio (SMR) for female deaths by cardiovascular diseases in the Community of Madrid (2001).
- $\log(SMR_k) = \log(o_k / e_k)$, where o_k and e_k are the observed and expected numbers of female deaths in municipality k, for $k = 1, \ldots, 179$.
- The class boundaries correspond to the deciles of raw $\log(SMR)$s.



Motivation

- Raw $\log(SMR)$s vary abruptly between geographical units, making hard to detect meaningful underlying patterns.

✓ **Solution**: use **smoothing methods** to enhance the visibility of underlying trends in noisy data.

- One approach: use a spatial two dimensional P-spline (represented as a mixed model), at the centroids of the geographical units, to model that trend (Lee and Durbán, 2009).

  ⋆ We will apply the Poisson version of this approach (Poisson P-GLMM) to the previous raw data.

Motivation

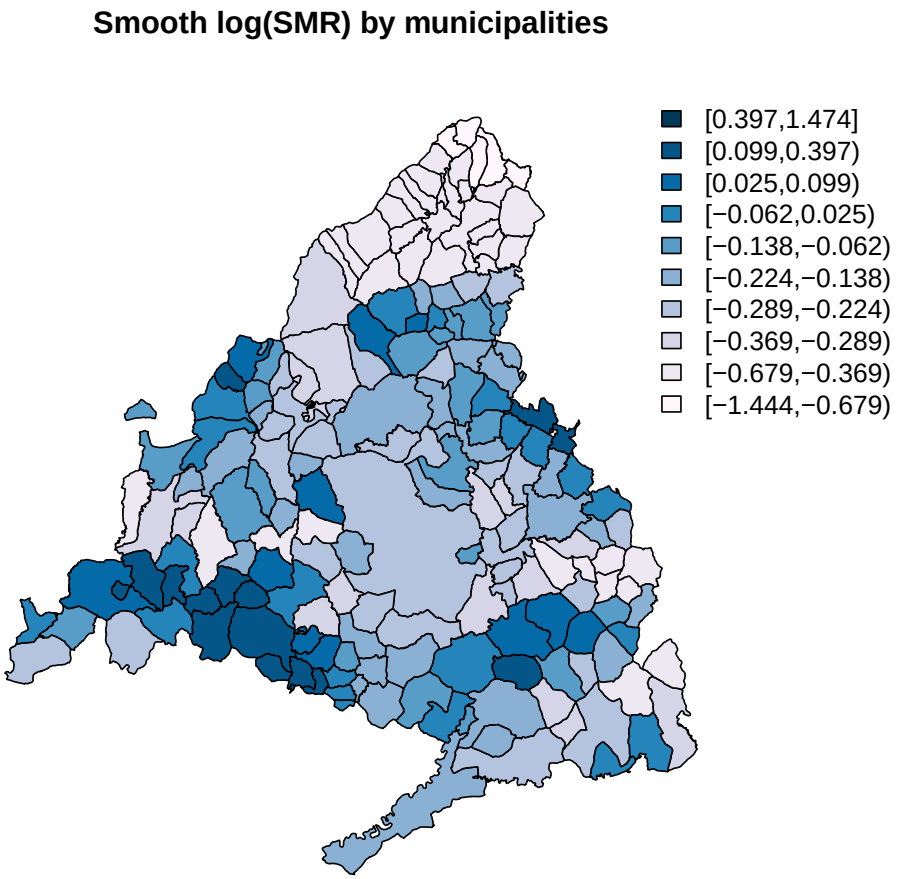
Example (continued)

- Poisson P-GLMM. Given $\boldsymbol{o} \rightsquigarrow Poiss(\boldsymbol{\mu})$,

$$\boldsymbol{\mu} = \boldsymbol{e} \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha}),$$

with $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}(\lambda_1, \lambda_2))$, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are constructed from the centroid coordinates of municipalities.

- Parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, λ_1, and λ_2 are estimated using the Penalized quasi-likelihood (PQL) approach of Breslow and Clayton (1993).



Motivation

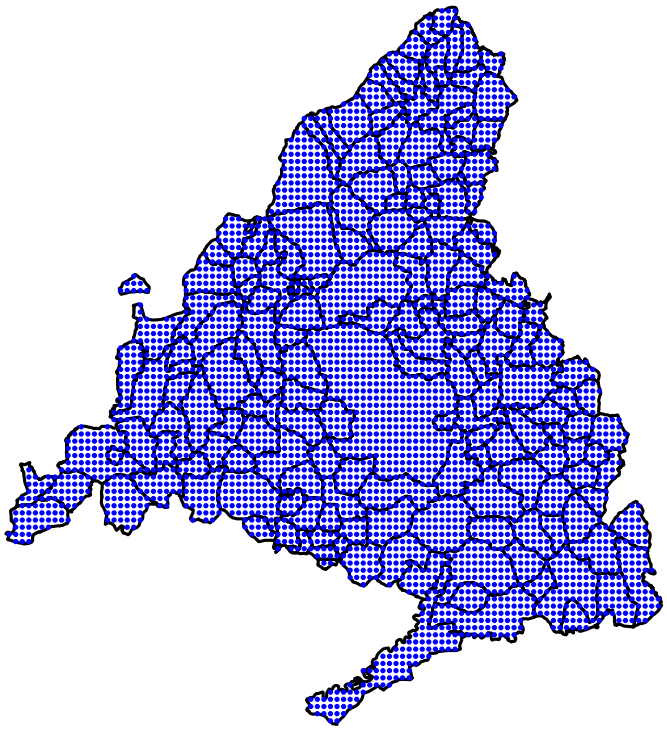
Some limitations:

- Choropleth maps present the common biases visual perception that larger rural and sparsely populated areas are of greater importance.
- Usually, there exists a mismatch of spatial supports between health data and relevant explanatory variables, making impossible the direct analysis of their possible relationship.

Motivation

Goal: Estimate a *latent* spatial trend for $\log(SMR)$s at a desirable fine scale, using available data. For example:

From municipalities (179) to a fine grid (4359)



Motivation

- **Proposal**: Extend the penalized composite link model (P-CLM, Eilers, 2007) to the spatial case, into a mixed model framework.
- The resulting model will be called (spatial) penalized composite link mixed model, or Poisson P-CLMM.

The model

Poisson P-CLMM. Given $\boldsymbol{o} \backsim Poiss(\boldsymbol{\mu})$,

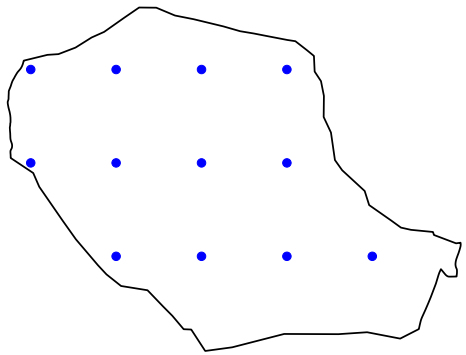
$$\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} = \boldsymbol{e}\exp(\boldsymbol{\eta}) \leftarrow \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha},$$

with $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}(\lambda_1, \lambda_2))$. Here:

- $\boldsymbol{\gamma}$: Latent expectation at the fine scale.
- $\boldsymbol{C}$: Composition matrix that describes how latent expectations was combined before generating the data (must be known).
- $\boldsymbol{\eta}$: Latent spatial trend at the fine scale, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the fixed and random effects matrices, constructed from fine grid coordinates.

The model

- $\boldsymbol{e}$: expected values at the fine scale. If it is not known, we may split e_k into equal parts, over the corresponding fine scale ($\widehat{\boldsymbol{e}}_{naive}$).
- For example, for municipality 4, we have $e_4 = 49.9158$. Then, we split it in 12 equal parts.



Parameters estimation and approximate errors

Parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, λ_1, and λ_2 are estimated using a modified PQL approach. The estimated fixed and random parameters are:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\check{\boldsymbol{X}}^{\mathsf{T}} \boldsymbol{V}^{-1} \check{\boldsymbol{X}}\right)^{-1} \check{\boldsymbol{X}}^{\mathsf{T}} \boldsymbol{V}^{-1} \boldsymbol{z}\,, \\ \widehat{\boldsymbol{\alpha}} &= \boldsymbol{G} \check{\boldsymbol{Z}}^{\mathsf{T}} \boldsymbol{V}^{-1} \left(\boldsymbol{z} - \check{\boldsymbol{X}} \widehat{\boldsymbol{\beta}}\right), \end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{W}^{-1} + \check{\boldsymbol{Z}} \boldsymbol{G} \check{\boldsymbol{Z}}^{\mathsf{T}}$, $\check{\boldsymbol{X}} = \boldsymbol{W}^{-1} \boldsymbol{C} \boldsymbol{\Gamma} \boldsymbol{X}$, $\check{\boldsymbol{Z}} = \boldsymbol{W}^{-1} \boldsymbol{C} \boldsymbol{\Gamma} \boldsymbol{Z}$, with matrices $\boldsymbol{W} = diag\,(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = diag\,(\boldsymbol{\gamma})$ and working vector $\boldsymbol{z} = \check{\boldsymbol{X}} \boldsymbol{\beta} + \check{\boldsymbol{Z}} \boldsymbol{\alpha} + \boldsymbol{W}^{-1} (\boldsymbol{o} - \boldsymbol{\mu})$.

Parameters estimation and approximate errors

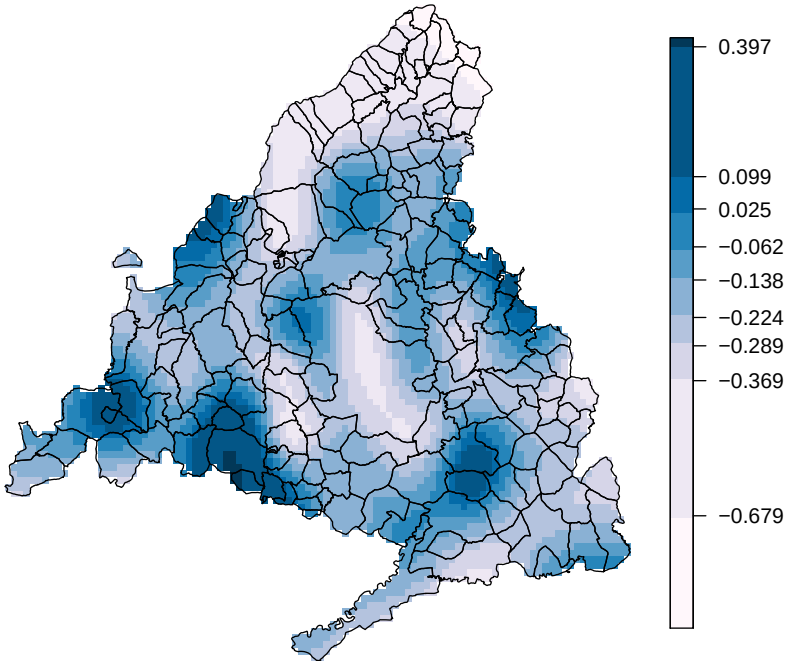
The approximate variance-covariance matrix for $\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}$ is:

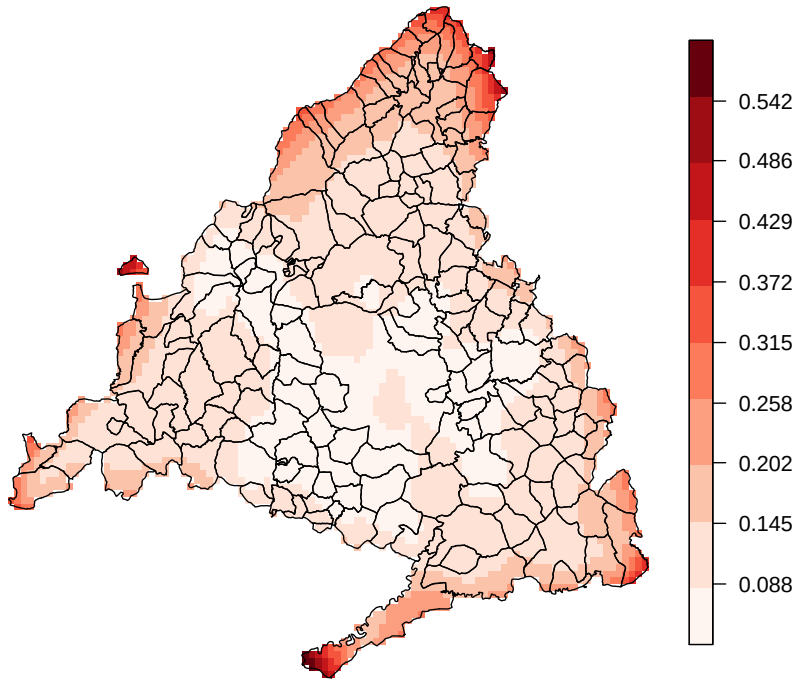
$$Var(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}) \approx [\boldsymbol{X} \mid \boldsymbol{Z}] \begin{bmatrix} \check{\boldsymbol{X}}^{\mathsf{T}} \boldsymbol{W} \check{\boldsymbol{X}} & \check{\boldsymbol{X}}^{\mathsf{T}} \boldsymbol{W} \check{\boldsymbol{Z}} \\ \check{\boldsymbol{Z}}^{\mathsf{T}} \boldsymbol{W} \check{\boldsymbol{X}} & \check{\boldsymbol{Z}}^{\mathsf{T}} \boldsymbol{W} \check{\boldsymbol{Z}} + \boldsymbol{G}^{-1} \end{bmatrix}^{-1} [\boldsymbol{X} \mid \boldsymbol{Z}]^{\mathsf{T}}.$$

The approximate errors for $\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}$ correspond to the squared root of the diagonal values of this matrix.

Application

From municipalities to a fine grid:

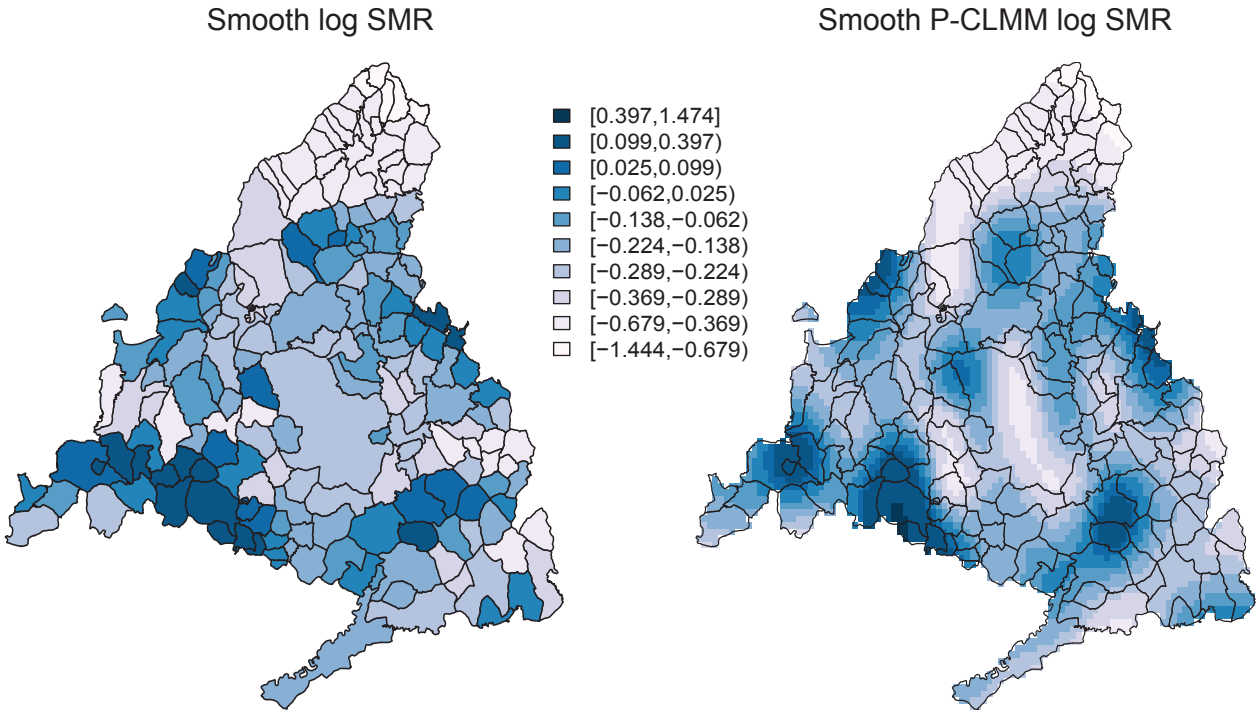




Here, we use equispaced cut-points for the error map legend.

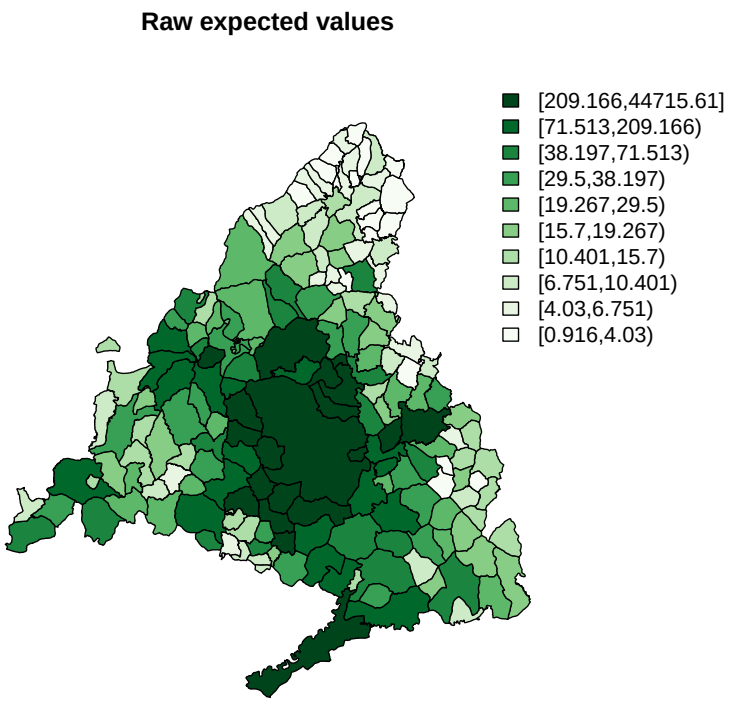
Application

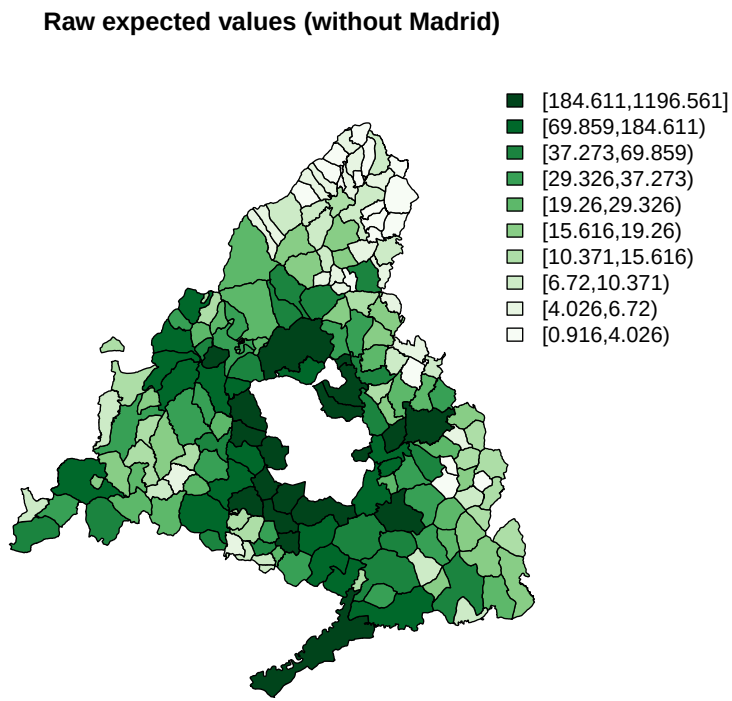
Compare (coarse) smooth P-GLMM log(SMR) with (continuous) smooth P-CLMM log(SMR) ...



Issues

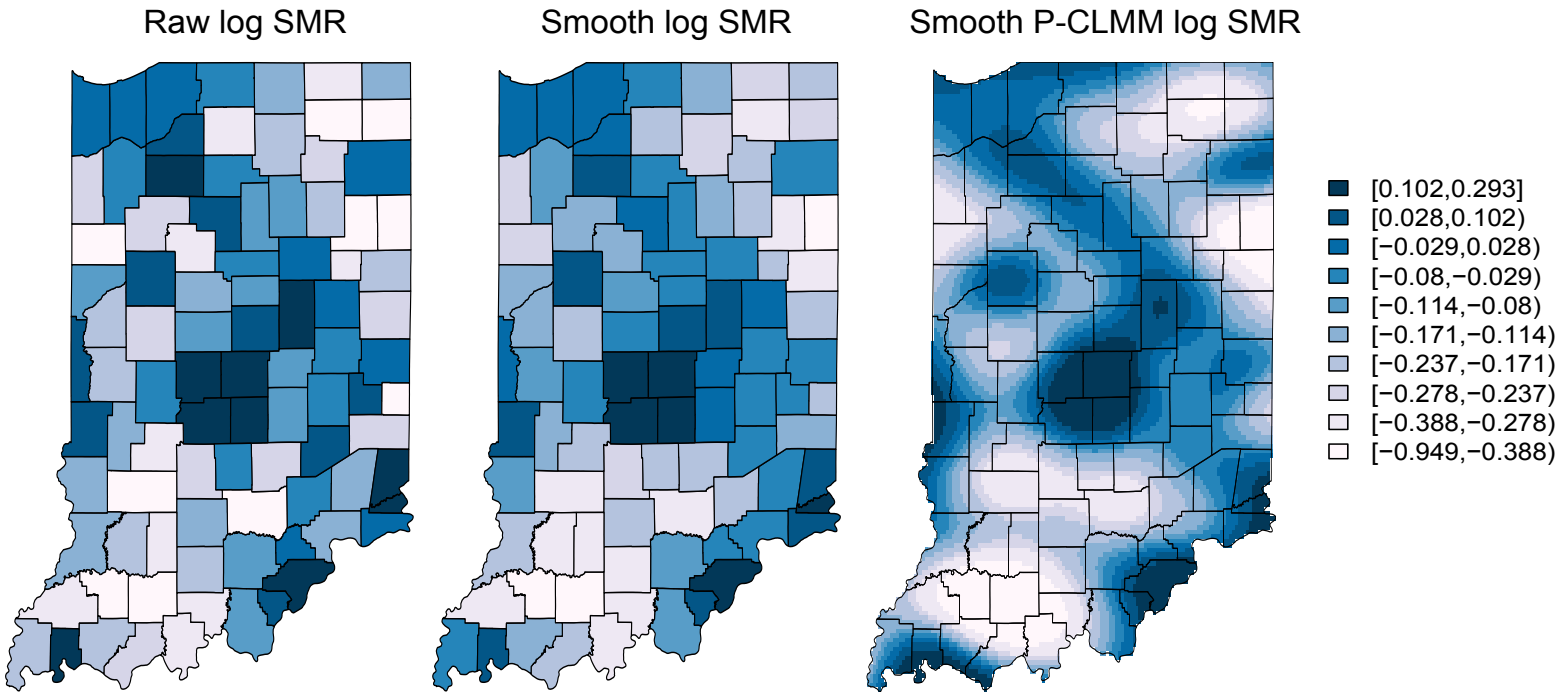
- Other alternatives to estimate e?
- We tried to estimate it using a Poisson and Gaussian spatial P-CLMM approach, but we had numerical problems. There are extreme raw expected values (for example, Madrid) ...





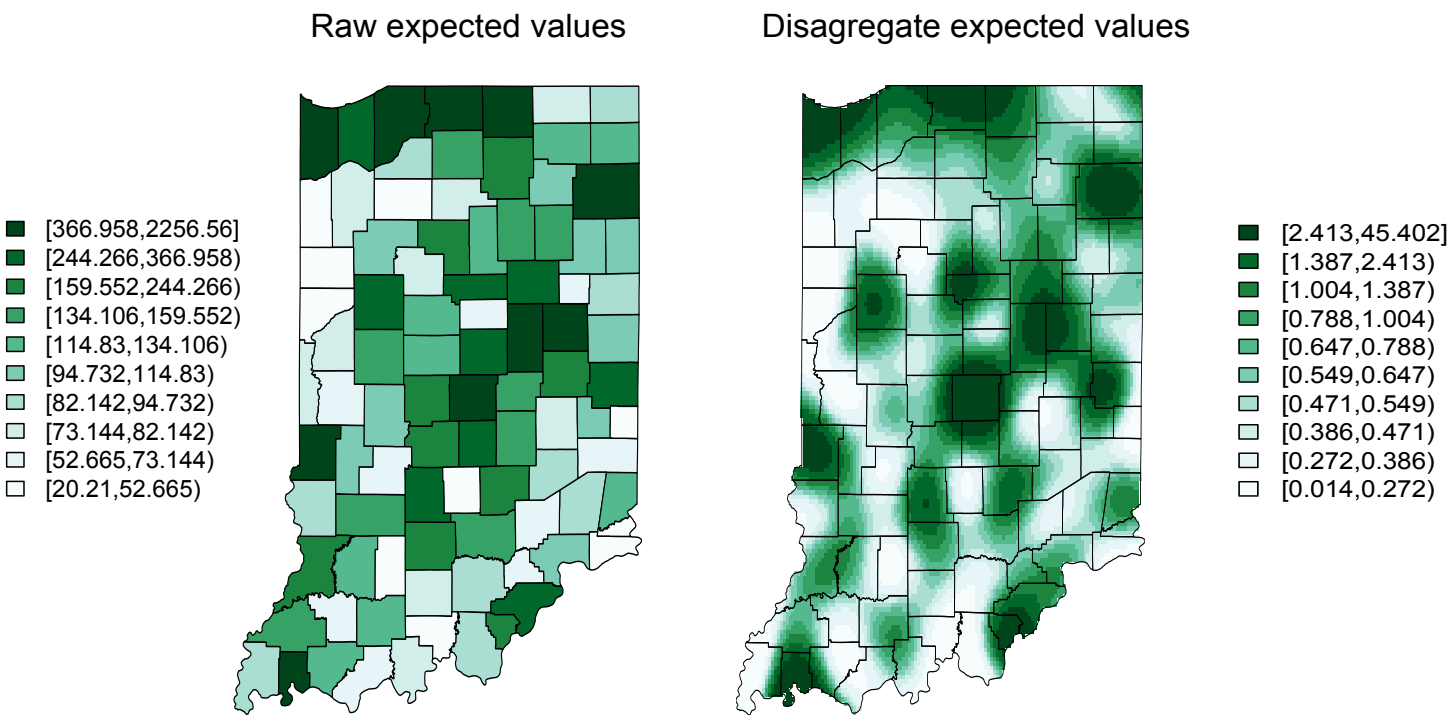
Indiana data set

But, look at this new data set. The Raw $log(SMRs)$ correspond to white female deaths by lung cancer in Indiana, USA, by (92) states. The smooth P-CLMM $log(SMRs)$ were obtained using $\widehat{\boldsymbol{e}}_{naive}$



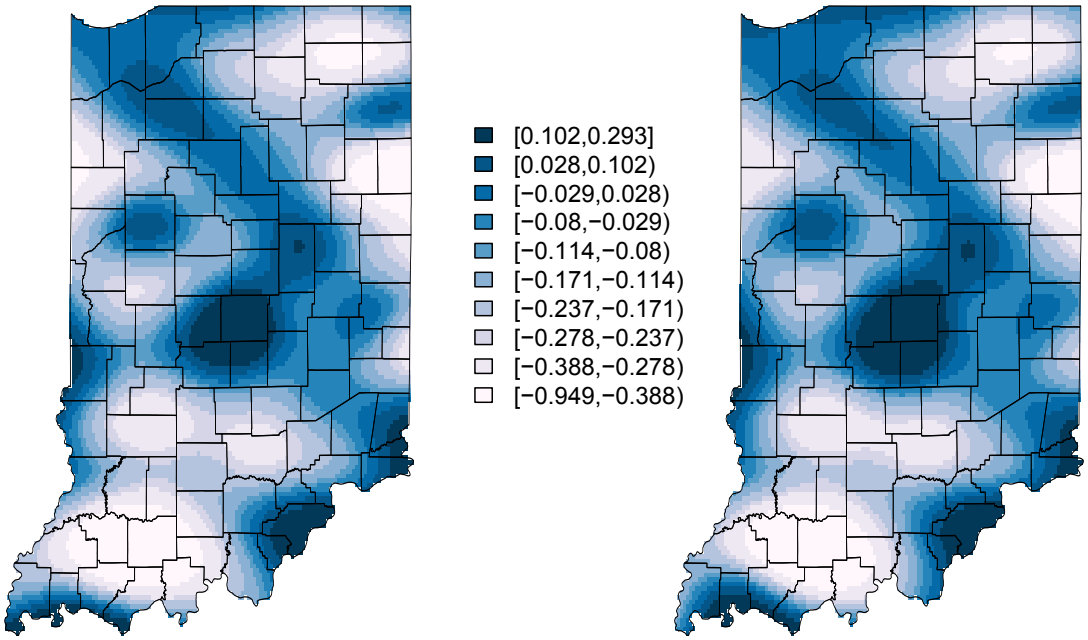
Indiana data set

In this case, we were able to estimate an underlying surface for expected values using a P-CLMM approach. Look at the raw expected values and its disaggregation:



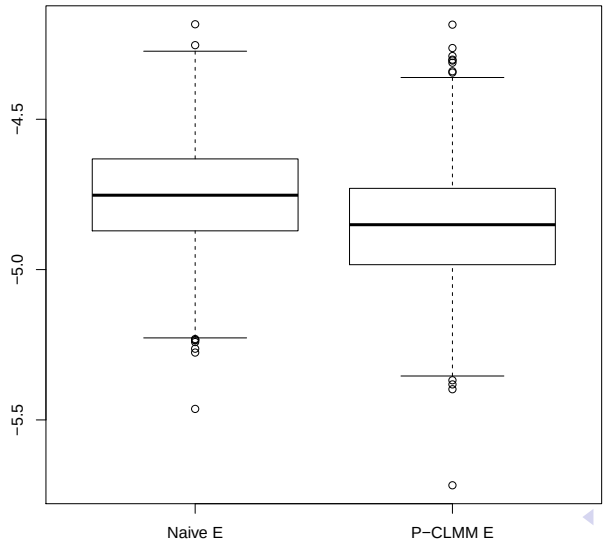
Indiana data set

If we use this estimate ($\widehat{e}_{P-CLMM}$), we obtain the smooth surface in left panel. Compare with the smooth surface obtained using $\widehat{e}_{naive}$ (right panel):



Simulation

If we consider the smooth surface obtained using $\widehat{e}_{P-CLMM}$, we can simulate $n = 1000$ observed values (at the fine grid) as Poisson realizations, with intensity $\widehat{e}_{P-CLMM} * \widehat{\eta}_{\widehat{e}_{P-CLMM}}$, aggregated them, and then apply our method using $\widehat{e}_{naive}$ and $\widehat{e}_{P-CLMM}$ as expected values at the fine scale. The boxplot of the MSEs (in log scale) for each case is the following:



Thanks!

References

- **Lee**, D.-J. and **Durbán**, M.(2009) Smooth-CAR mixed models for count data. Computational Statistics & Data Analysis, 53:2958-2979.
- **Eilers**, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. Statistical Modelling, 7:239-254.
- **Breslow**, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88:9-25.

Appendix

Mixed model formulation:

Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the lon-lat coordinates, at the fine scale. Consider B-spline bases $\boldsymbol{B}_i = \boldsymbol{B}(\boldsymbol{x}_i)$, and difference matrices $\boldsymbol{D}_i = \boldsymbol{D}_i(q_i)$, $i = 1, 2$. Using the SVD of $\boldsymbol{D}_i^{\mathsf{T}} \boldsymbol{D}_i$,

$$\boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{U}_i^{\mathsf{T}},$$

with partitions $\boldsymbol{U}_i = [\boldsymbol{U}_{in} : \boldsymbol{U}_{is}]$ and $\boldsymbol{\Sigma}_i = BlockDiag(\boldsymbol{0}_{q_i}, \tilde{\boldsymbol{\Sigma}}_i)$, we can define:

$$\boldsymbol{X} = \boldsymbol{X}_2 \square \boldsymbol{X}_1 \text{ and } \boldsymbol{Z} = [\boldsymbol{Z}_2 \square \boldsymbol{X}_1 : \boldsymbol{X}_2 \square \boldsymbol{Z}_1 : \boldsymbol{Z}_2 \square \boldsymbol{Z}_1],$$

where $\boldsymbol{X}_i = \boldsymbol{B}_i \boldsymbol{U}_{in}$ and $\boldsymbol{Z}_i = \boldsymbol{B}_i \boldsymbol{U}_{is}$.

Appendix

It can be shown that $\boldsymbol{G}(\lambda_1, \lambda_2)$ is reformulated as $\boldsymbol{G}(\lambda_1, \lambda_2) = \boldsymbol{F}^{-1}$, where:

$$\boldsymbol{F} = \begin{bmatrix} \lambda_2 \tilde{\boldsymbol{\Sigma}}_2 \otimes \boldsymbol{I}_{q_1} & & \\ & \lambda_1 \boldsymbol{I}_{q_2} \otimes \tilde{\boldsymbol{\Sigma}}_1 & \\ & & \lambda_1 \boldsymbol{I}_{c_2 - q_2} \otimes \tilde{\boldsymbol{\Sigma}}_1 + \lambda_2 \tilde{\boldsymbol{\Sigma}}_2 \otimes \boldsymbol{I}_{c1\_q_1} \end{bmatrix}$$

with $c1 = ncol(\boldsymbol{B}_1)$ and $c2 = ncol(\boldsymbol{B}_2)$.