Prediction of cancer mortality risks in spatio-temporal disease mapping [1]

Goicoa, T.[1], Ugarte, M.D.[1], Militino, A.F.[1], Etxeberria, J.[1,2]
[1] Department of Statistics and O. R., Universidad Pública de Navarra
[2] CIBER in Epidemiology and Public Health

email:tomas.goicoa@unavarra.es

Abstract:
The main goal of spatio-temporal disease mapping is describing the evolution of geographical patterns of mortality or incidence risks (rates). This could give clues to epidemiologists and public health researchers to formulate etiologic hypothesis of the disease. However, the ability of disease mapping models to make predictions about future mortality or incidence risks has not been widely explored. In this work, a flexible spatio-temporal model is considered for risk estimation and forecasting. The prediction MSE of both fitted and forecast values, as well as estimators of those quantities, will be derived. Spanish cancer mortality data will be used for illustration.

Keywords: P-spline models, CAR models, smoothing risks, forecasting.

1 Introduction

Health agencies plan cancer prevention resources based on cancer mortality/incidence risk estimations available to date. However, these official numbers are available after three or four years. In this context, statistical procedures providing mortality/incidence risk predictions for different regions or health areas are very useful. Using jointpoint regression models, Malvezzi et al. (2011) present estimates of mortality for all cancers and for selected major cancer sites in the year 2011 in the whole European Union and in its six more populated countries. They use actual mortality data up to the most recent available year, which is between 2005 and 2007 for most EU countries.

In this work flexible spatio-temporal models are considered to predict risks. The prediction MSE of both fitted and forecast values, as well as estimators of those quantities, will be derived. P-splines have been proposed in small areas to forecast dwelling prices (see Ugarte et al., 2009), and here, we extend this work to disease mapping spatio-temporal models including interaction terms. The methodology will

[1]This research has been supported by the Spanish Ministry of Science and Innovation (MTM 2008-03085/MTM).

be used to analyze several mortality cancer data for all the Spanish provinces in the period 1975-2008. Risks predictions for future years will be also provided.

2 Materials and Methods

Two models are considered for forecasting. Firstly, a model with CAR distributions for both the spatial and temporal random effects is considered. This model includes a spatio-temporal interaction term similar to those described by Knorr-Held (2000). Secondly, a spatio-temporal P-spline model described in Ugarte et al. (2010) is used. Smoothing is carried out in three dimensions: longitude, latitude, and time, allowing for different smoothing parameter in each dimension. Predictions will be obtained by extending the marginal time B-spline basis.

Consider n contiguous regions labelled $i = 1, \ldots, n$, and T time periods denoted by $t = 1, \ldots, T$. Conditional on the random region effects r_{it}, the number of deaths in each area and time period, C_{it}, is assumed to be Poisson distributed with mean $\mu_{it} = e_{it} r_{it}$, where r_{it} represents the unknown relative risks of mortality from a rare disease, and e_{it} is the expected number of deaths. Namely

$$C_{it} | r_{it} \sim Poisson(\mu_{it} = e_{it} r_{it}), \quad \log \mu_{it} = \log e_{it} + \log r_{it}. \tag{1}$$

In the spatio-temporal CAR model, the log-risk is modeled as

$$u_{it} = \log r_{it} = \beta + \phi_i + \gamma_t + \delta_{it}, \tag{2}$$

where β is an overall risk level, ϕ_i represents spatial effects, γ_t denotes temporal effects, and δ_{it} are space-time interaction effects. The distributions for the random effects $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ are

$$\begin{aligned} \boldsymbol{\phi} &\sim N(\mathbf{0}, \sigma_s^2 \mathbf{D}_s) \quad ; \quad \mathbf{D}_s = (\lambda_s \mathbf{Q}_s + (1 - \lambda_s)\mathbf{I}_s)^-, \\ \boldsymbol{\gamma} &\sim N(\mathbf{0}, \sigma_t^2 \mathbf{D}_t) \quad ; \quad \mathbf{D}_t = \mathbf{Q}_t^-, \\ \boldsymbol{\delta} &\sim N(\mathbf{0}, \sigma_{st}^2 \mathbf{D}_{st}) \quad ; \quad \mathbf{D}_{st} = \mathbf{Q}_t^- \otimes \mathbf{Q}_s^-, \end{aligned}$$

where $\mathbf{Q}_s$ is determined by the spatial neighbourhood structure with the ith diagonal element equal to the number of neighbours of the ith region and for $i \neq j$, $\mathbf{Q}_{ij} = -1$ if i and j are neighbours and 0 otherwise; $\mathbf{I}_s$ is the $n \times n$ spatial identity matrix, and $\mathbf{Q}_t$ is determined by the temporal neighbourhood structure and it is analogously defined as $\mathbf{Q}_s$.

Model 2 can be expressed in matrix form as

$$\mathbf{u} = \mathbf{X}\beta + \mathbf{Z}_1 \boldsymbol{\phi} + \mathbf{Z}_2 \boldsymbol{\gamma} + \mathbf{Z}_3 \boldsymbol{\delta} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}),$$

Using a P-spline spatio-temporal model the log-risk is modeled as

$$u_{it} = \log r_{it} = f(x_{1i}, x_{2i}, x_t), \tag{3}$$

where x_{1i} and x_{2i} are the coordinates of the centroid of the ith small area (longitude and latitude respectively), x_t is the time, and f is a smooth function to be estimated using P-splines with B-spline bases. One of the most interesting aspects of the P-spline models is that they can be expressed as linear mixed models using a one-to-one (orthogonal) transformation. Hence, the P-spline model can be represented as

$$\mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{F}^{-1}).$$

The $\mathbf{X}$ and $\mathbf{Z}$ matrices are given by

$$\begin{aligned} \mathbf{X} &= \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1) \\ \mathbf{Z} &= [\mathbf{Z}_1^* : \mathbf{Z}_2^* : \mathbf{Z}_3^* : \mathbf{Z}_4^* : \mathbf{Z}_5^* : \mathbf{Z}_6^* : \mathbf{Z}_7^*], \end{aligned}$$

and

$$\begin{aligned} \mathbf{Z}_1^* &= \mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1), \quad \mathbf{Z}_2^* = \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1), \quad \mathbf{Z}_3^* = \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1), \\ \mathbf{Z}_4^* &= \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1), \quad \mathbf{Z}_5^* = \mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1), \quad \mathbf{Z}_6^* = \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1), \\ \mathbf{Z}_7^* &= \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1), \end{aligned}$$

where the symbol $\square$ denotes the "row-wise" Kronecker product of two matrices (see for example, Eilers et al., 2006). Here, $\mathbf{X}_1 = [1 : \mathbf{x}_1]$, $\mathbf{X}_2 = [1 : \mathbf{x}_2]$, $\mathbf{X}_3 = [1 : \mathbf{x}_3]$, $\mathbf{Z}_1 = \mathbf{B}_1\mathbf{U}_{1s}$, $\mathbf{Z}_2 = \mathbf{B}_2\mathbf{U}_{2s}$, and $\mathbf{Z}_3 = \mathbf{B}_3\mathbf{U}_{3s}$. The matrices $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{B}_3$ are the marginal B-spline bases for longitude, latitude and time; $\mathbf{U}_{1s}$, $\mathbf{U}_{2s}$ and $\mathbf{U}_{3s}$ come from the singular value decomposition of the penalty matrices for longitude, latitude and time, and the covariance matrix $\mathbf{F}^{-1}$ is a diagonal matrix arising from the representation of the P-spline model as a mixed model (see Ugarte et al., 2010 for more details).

The models are estimated using the well known penalized quasi-likelihood technique (PQL)(Breslow and Clayton, 1993). Risk predictions and their standard errors are obtained by extending the $\mathbf{X}$ and $\mathbf{Z}$ matrices.

3 Results

The methodology is illustrated analyzing prostate cancer in Spain from 1975 to 2008. Figure 1 displays relative risks estimates (1975-2008) and predictions (2009-2011) for four selected Spanish provinces, together with 95% confidence bands obtained with the P-spline model (3). A decreasing trend in mortality can be observed.

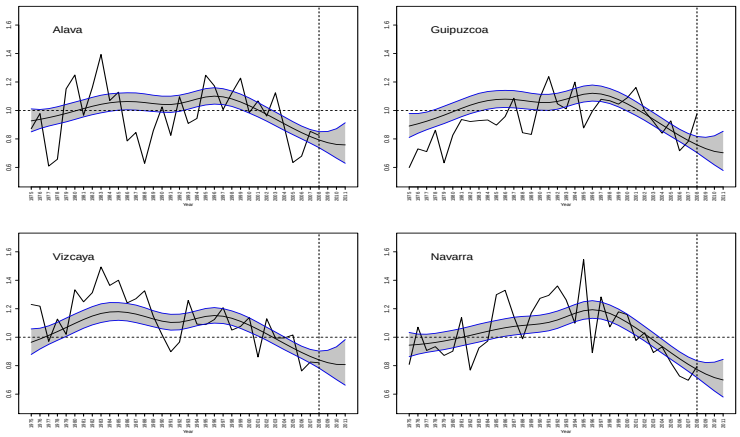

Figure 1: Smoothed prostate cancer mortality risks estimations and predictions with 95% confidence bands.

Acknowledgments

This research has been supported by the Spanish Ministry of Science and Innovation (MTM 2008-03085/MTM). The authors would like to thank to Marina Pollán from the National Epidemiology Center (area of Environmental Epidemiology and Cancer) for providing the data.

References

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

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

Knorr-Held L. 2000. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* **19**, 2555-2567.

Ugarte M.D. , Goicoa T., Militino A.F., Durban M. (2009). Spline smoothing in small area trend estimation and forecasting, *Computational Statistics and Data Analysis*, **53**, 3616-3629.

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