Abstract:

- Extrapolating cancer mortality trends can be very valuable as a tool to predict cancer burden. National Health Agencies use different models to figure out future evolution of cancer, but they mainly work at national level. However, developed countries are divided into different regions with their own governments and health care systems, and this should be taken into account. In this paper, an ANOVA-type P-spline model is considered to predict the number of mortality cases in forthcoming years in regions within a country. The model is very interesting as it allows to split the predictions into components representing region-specific features and characteristics common to the whole country. Prediction variability is also calculated to provide prediction intervals. Real data on cancer mortality are used for illustration.

Key-Words:

- region-specific predictions; smoothing and predicting counts; space-time interactions; prostate cancer.

AMS Subject Classification:

- 62M20, 62G08, 62H11.
1. INTRODUCTION

Prediction of future events has always been a challenge in modern societies, and statistical methods are valuable tools to forecast outcomes in many fields of daily life. For example, in economy, we are constantly receiving predictions about employment, rate of growth, income, expenses and many other quantities. In medicine, it is common to make predictions about the evolution of a disease, the spread of an epidemic, the outbreak of influenza, or the number of new HIV cases. The main reason why governments, institutions or private companies demand predictions is that advance knowledge about the future allows to make plans, to think about business strategies and management, or to allocate resources efficiently.

Future information about cancer incidence or mortality is essential for Public Health Agencies since this illness brings huge expenses in developed countries involving diagnosis, treatment, research, loss of productivity because of sick leaves, or pensions due to premature deaths in a family. These figures are also important to efficiently organize cancer screening programs and to prioritize prevention activities. A cancer situation assessment requires an appraisal of the problem in terms of the number of incidence or mortality cases. This should be done based on an updated collection of cancer figures provided by population-based cancer registries or censuses. Regarding the official figures, these are available with a delay of approximately three or four years due to the complexity of updating cancer registries. Hence, Health Agencies substitute this lack of information with projections of cancer cases based on statistical models. Most of these agencies use models at country level, and hence, they are essentially temporal models. To show some examples, Lee et al. (2011) provide a comparison of the different methods using Canadian cancer mortality data for twelve cancer sites. The authors compare a temporal Poisson log-linear model used by the Public Health Agency of Canada; age-period-cohort models considered by the Association of the Nordic Cancer Registries; autoregressive with time trends models used by the American Cancer Society, or state-space models used by the National Cancer Institute. Joint point regression models implemented in the Jointpoint Regression Programme by the National Cancer Institute are also studied. According to these authors, no model can be used for all cancer sites, and the performance also depends on the number of observed cases. Moreover, the same models can show different behavior in different countries. For example, for testis, thyroid and ovary cancers, different performance is observed with Canadian and American data.

There has been additional academic research on predicting cancer mortality cases mainly based on time models. For example, Chen et al. (2012) and Zhu et al. (2012) evaluate different models to provide 4-year-ahead cancer counts projections in USA. Tiwari et al. (2004) consider state-space methods to improve
current projection methods used by the American Cancer Society. Ghosh and Tiwari (2007) proposed a local linear model for short term projections and a quadratic local model for longer prediction periods. Ghosh et al. (2008) develop a projection method based on state-space models combining the best features of a local quadratic model and an autoregressive model with fixed trend. Some more work on temporal models includes Dyba and Hakulinen (2000) or Malvezzi et al. (2012, 2013) to cite some of them. These models consider the calendar year as the relevant time axis. However, for some cancer sites, the relevant time axis is not the calendar year but the cohort of birth, and consequently age-period-cohort models could be used. Research about cancer projections using age-period-cohort models without spatial correlation can be found in Knorr-Held and Rainer (2001), Clements et al. (2005), Riebler and Held (2010) or Riebler et al. (2012). On the other hand, Schmid and Held (2004) provide stomach cancer mortality projections using age-period-cohort models including spatial correlation. Very recently Ugarte et al. (2012a) consider a three-dimensional P-spline model to project prostate cancer mortality counts in fifty Spanish provinces. The authors conclude that the P-spline model, that takes into account spatial dependencies, is preferable to individual P-spline temporal models fitted separately in each province. Etxeberria et al. (2014) compare different conditional autoregressive models (CAR), P-spline models, and a combination of both in terms of their predictive performance using cancer mortality data. Results reveal that models combining CAR random effects for space and P-splines for time perform slightly worse than models based only on P-splines or CAR models. The key point of these papers is that the authors provide a unified framework of smoothing and predicting under the mixed model theory using the mixed model representation of P-spline models. In a different context, Currie et al. (2004) use P-splines to smooth and forecast mortality rates for the pension industry, but they do not use the mixed model reformulation. In an economic setting, Ugarte et al. (2009) forecast dwelling prices in different neighbourhoods of Vitoria, a Spanish city.

The goal of this paper is to provide guidelines on how to extend an ANOVA-type P-spline model to predict cancer mortality counts. Recently, Ugarte et al. (2012b) used this model to smooth prostate cancer mortality risks in Spain. One interesting feature of this model is that it allows to split the relative risk into a smooth trend common to all regions, a smooth spatial surface constant along the time period, and a smooth interaction term representing the region-specific temporal evolution of the risk. Projections can be then decomposed into the same components. This is of great interest from an epidemiological point of view, since the decomposition of the predicted risks into these components allows to assess if the increase/decrease of those risks is mainly attributable to a common temporal behavior of all the regions or is due to an area-specific behavior during the oncoming years. This information could lead to a better organization of cancer prevention programs, open up new research lines to investigate the differences among the areas, or just help to speculate about new risk factors.
The ANOVA-type P-spline model can also be reformulated as a generalized linear mixed model where the strategy to avoid identifiability problems is very simple. In this paper, predictions of future mortality counts derived from this model are provided under the mixed model framework such that smoothing, predicting and assessing variability are jointly accomplished. The methodology will be illustrated using Spanish prostate cancer mortality data during the period 1975–2008. This will allow us to make comparisons with alternative models previously used in the literature.

The rest of the paper is laid out as follows. Section 2 describes the extension of the ANOVA-type P-spline model and how predictions are obtained. The technique is illustrated in Section 3. A validation study is presented in Section 4. Finally, the paper ends with a discussion.

2. TIME EXTENDED ANOVA-TYPE P-SPLINE MODEL

ANOVA decompositions of smooth functions have been already considered in the literature. See for example Gu (2002) and Belitz and Lang (2005). Recently, Wood et al. (2013) propose new penalties that allow ANOVA models to be fitted using existing mixed model software. In this section, a spatio-temporal ANOVA-type P-spline model with B-spline bases is considered to estimate and predict cancer mortality figures. This model was initially used by Lee and Durbán (2011) to estimate ozone levels in Europe and by Ugarte et al. (2012b) to smooth risk in space-time disease mapping. Different approaches using B-splines have also been considered in the disease mapping literature (see for example MacNab and Dean, 2001; MacNab and Gustafson, 2007; Silva et al., 2008). In this paper we focus on extending the ANOVA-type model to estimate and predict risks jointly using a mixed model reformulation. Suppose we have a big area (e.g. a country) divided into smaller regions (e.g. provinces), for which mortality (or incidence) counts in different time points are available. Denoting the province by the subindex $s = 1, ..., S$, the time period for observed data by $t = 1, ..., T$, and conditional on the unknown relative risk $r_{st}$, the number of deaths $C_{st}$ is assumed to be Poisson distributed with mean $\mu_{st} = e_{st}r_{st}$, where $e_{st}$ is the expected number of deaths calculated on the basis that the $s$-th province in time $t$ behaves as the whole country in the studied period. Then

$$C_{st}|r_{st} \sim \text{Poisson}(\mu_{st} = e_{st}r_{st}), \quad \log \mu_{st} = \log e_{st} + \log r_{st}.$$ 

In this work, our interest lies in estimating and predicting risks and counts for each province. An extension of an ANOVA-type P-spline model will be considered. The model includes additive terms for space (longitude and latitude), time, and space-time interactions, and hence the log-risk (log $r_{st}$) is modeled as
the sum of an intercept, a smooth term for the spatial surface, a temporal smooth trend, and a smooth term for the space-time interaction.

Let us define the extended time period encompassing observed and future values. This is denoted by \( t^* = 1, ..., T, T+1, T+2, ..., T+p \), where \( p \) is the number of years to predict. Log-risks for observed and predicted values are modeled as

\[
(2.2) \quad u_{st}^* = \log r_{st}^* = \delta + f_s(x_1, x_2) + f_t(t^*) + f_{st}(x_1, x_2, t^*) = \mathbf{B}^* \theta^*.
\]

The term \( \delta \) is an intercept, \( f_s(x_1, x_2) \) represents the smooth spatial effect constant along the period, \( f_t(t^*) \) is an extended temporal trend common to all areas, and \( f_{st}(x_1, x_2, t^*) \) is the extended interaction term that can be interpreted as the specific temporal trend for each area. In these expressions \( x_1 \) and \( x_2 \) are the coordinates of the centroid of the \( i \)th small area (longitude and latitude respectively), \( t^* \) is the time (for observed and predicted values), and \( f_i, i = s, t, st \) are smooth functions to be estimated using P-splines with B-spline bases. \( \mathbf{B}^* \) is the extended B-spline basis and \( \theta^* \) is a vector of coefficients. The matrix \( \mathbf{B}^* \) is explicitly defined as

\[
(2.3) \quad \mathbf{B}^* = [\mathbf{1}_{st}^* : \mathbf{1}_t^* \otimes \mathbf{B}_s : \mathbf{B}_t^* \otimes \mathbf{1}_s : \mathbf{B}_t^* \otimes \mathbf{B}_s],
\]

where \( \mathbf{1}_{st}^*, \mathbf{1}_t^*, \) and \( \mathbf{1}_s \) are column vectors of ones of length \( S \times (T+p) \), \( T+p \), and \( S \) respectively. \( \mathbf{B}_s = \mathbf{B}_{s_2} \mathbf{B}_{s_1} \) is the spatial B-spline basis defined by the row-wise (\( \otimes \)) Kronecker product (Eilers et al., 2006) of the marginal basis for longitude (\( \mathbf{B}_{s_1} \)) and latitude (\( \mathbf{B}_{s_2} \)). \( \mathbf{B}_t^* \) represents the extended marginal basis for time and it is a lower block-triangular partitioned matrix given by

\[
(2.4) \quad \mathbf{B}_t^* = \begin{pmatrix} \mathbf{B}_t & \mathbf{0} \\ \mathbf{B}_{t_1} & \mathbf{B}_{t_2} \end{pmatrix}.
\]

In this expression, \( \mathbf{B}_t \) is the time marginal basis corresponding to the observed period \( (t = 1, ..., T) \), and \( \mathbf{B}_{t_1} \) and \( \mathbf{B}_{t_2} \) are the rows corresponding to the extended data.

To ensure that \( f_i, i = s, t, st \) are smooth functions, the P-spline approach places penalties on the coefficients \( \theta^* \). The extended penalty matrix \( \mathbf{P}^* \) is given by a block-diagonal matrix whose components are penalties for the two-dimensional spatial component, the one dimensional time component and the three-dimensional component (space-time interactions). More precisely, \( \mathbf{P}^* = \text{diag}(\mathbf{P}_s, \mathbf{P}_t^*, \mathbf{P}_{st}^*) \), where

\[
(2.5) \quad \mathbf{P}_s = \lambda_s \mathbf{I}_{m_2} \otimes \mathbf{P}_{s_1} + \lambda_{s_2} \mathbf{P}_{s_2} \otimes \mathbf{I}_{m_1},
\]

\[
\mathbf{P}_t^* = \lambda_t \mathbf{P}_t^*,
\]

\[
\mathbf{P}_{st}^* = \tau_s \mathbf{I}_{m_3} \otimes \mathbf{I}_{m_2} \otimes \mathbf{P}_{s_1} + \tau_{s_2} \mathbf{I}_{m_3} \otimes \mathbf{P}_{s_2} \otimes \mathbf{I}_{m_1} + \tau_t \mathbf{P}_t^* \otimes \mathbf{I}_{m_2} \otimes \mathbf{I}_{m_1}.
\]
In these expressions, $I_{m_j}$, $j = 1, 2, 3$ are identity matrices of dimension $m_j \times m_j$, where $m_j$ is the number of columns of $B_j, j = s_1, s_2, t$, and $I_{m_3} = \begin{pmatrix} I_{m_3} & 0 \\ 0 & 0 \end{pmatrix}$.

$P_{s_1}$ and $P_{s_2}$ are penalty matrices for longitude and latitude respectively defined by $P_{s_j} = D_{s_j}' D_{s_j}, j = 1, 2$ where $D_{s_j}$ are second order difference matrices to achieve smoothness over adjacent marginal coefficients (see Eilers and Marx, 1996). Matrix $P^*$ is defined using the extended difference matrix $D_t^*$ for the time component given by the next expression

\begin{equation}
D_t^* = \begin{pmatrix} D_t & 0 \\ E_t & L_t \end{pmatrix}, \quad P_t^* = D_t'^* D_t^* = \begin{pmatrix} P_t + E_t' E_t & E_t' L_t \\ L_t' E_t & L_t' L_t \end{pmatrix}.
\end{equation}

$D_t$ and $P_t$ are the difference matrix and the penalty matrix for the observed time period, $E_t$ and $L_t$ are the rows used to obtain the penalty for the oncoming years, and $\lambda_{s_1}, \lambda_{s_2}, \lambda_t, \tau_{s_1}, \tau_{s_2}$, and $\tau_t$ are different smoothing parameters corresponding to space, time, and interaction components respectively. The extended B-spline basis for time in Equation (2.4) and the extended difference and penalty matrices in Equation (2.6) are equal to those obtained in a three-dimensional P-spline model by Ugarte et al. (2012a). However, the extended transformation matrix is different. The next step is to reformulate the P-spline model (2.2) as a generalized linear mixed model. To do this, a matrix $T^*$ is used to transform $B^*$ into $[X^* : Z^*]$ and $\theta^*$ into $(\beta', \alpha'^r)'$. In this paper we provide the definition of this transformation matrix $T^*$ which is based on matrices of eigenvectors corresponding to non-zero and zero eigenvalues respectively obtained from the eigen decomposition of the matrices $P_i, i = s_1, s_2, t$. The key point in this process is to choose an extended transformation matrix preserving the original transformation matrix $T$ used to fit the data. Based on the transformation matrix $T$, the following extended transformation matrix is considered

\[
T^* = \begin{pmatrix} 1 \\ T_s \\ T_t^* \\ T_{st}^* \end{pmatrix},
\]

where $T_t^*$ and $T_{st}^*$ are defined by

\[
T_t^* = \begin{pmatrix} T_t & 0 \\ 0 & L_t^{-1} \end{pmatrix}, \quad T_{st}^* = \begin{pmatrix} T_{st} & 0 \\ 0 & L_t^{-1} \otimes I_{m_2} \otimes I_{m_1} \end{pmatrix},
\]

and

\[
T_s = [1 \otimes [u_{2n} \otimes 1 : 1_2 \otimes u_{1n} : u_{2n} \otimes u_{1n}] : R_s],
\]

\[
T_t = [u_{3n} \otimes 1 : R_t],
\]

\[
T_{st} = [u_{3n} \otimes [u_{2n} \otimes 1 : 1_2 \otimes u_{1n} : u_{2n} \otimes u_{1n}] : R_{st}].
\]
The matrices $R_s$, $R_t$, and $R_{st}$ are given by
\[
R_s = [1 \otimes [U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}]],
\]
\[
R_t = [U_{1s} \otimes 1],
\]
\[
R_{st} = [u_{1n} \otimes [U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}]:
U_{3n} \otimes [1_2 \otimes u_{1n} : u_{2n} \otimes 1_1 : u_{2n} \otimes u_{1n} : U_{2n} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}]].
\]

Note that $T_s$, $T_t$ and $T_{st}$ represent the components of the original transformation matrix corresponding to the observed data. $U_{in} = [1_i : u_{in}]$ and $U_{is}$, $i = 1, 2, 3$, are matrices of eigenvectors corresponding to zero and non-zero eigenvalues obtained from the eigen-decomposition of the penalty matrix $P_j$, $j = s_1, s_2, t$. Using this transformation, the generalized mixed model reformulation of the extended ANOVA-type P-spline model (2.2) can be obtained. More precisely, the fixed and random effect matrices of the extended generalized linear mixed model are given by
\[
B^*T^* = [1^*_{st} : (1^*_{s} \otimes B_s)T_s : (B^*_{t} \otimes 1_s)T^*_{t} : (B^*_{t} \otimes B_s)T^*_{st}],
\]
and the extended model is expressed as
\[
\begin{pmatrix}
\mathbf{u}^o \\
\mathbf{u}^p
\end{pmatrix} = \delta + \begin{pmatrix} X^o \ Z^o_s \\ X^p \ Z^p_{st} \end{pmatrix} \begin{pmatrix} \beta^o_s \\ \alpha^o_s \end{pmatrix} + \begin{pmatrix} X^o_t \ Z^o_t \\ X^p_t \ Z^p_{t1} \ Z^p_{t2} \end{pmatrix} \begin{pmatrix} \beta^o_t \\ \alpha^o_t \end{pmatrix} + \begin{pmatrix} X^o_{st} \ Z^o_{st} \\ X^p_{st} \ Z^p_{st1} \ Z^p_{st2} \end{pmatrix} \begin{pmatrix} \beta^o_{st} \\ \alpha^o_{st} \end{pmatrix},
\]
(2.7)

where detailed expressions for each of the components are given in Appendix A. Super-indexes $o$ and $p$ refer to matrices for observed and predicted values respectively. Note that repeated columns have been removed to avoid identifiability problems. Here $\mathbf{u}^p$ are the log-risks to be predicted; $\beta^o_s$, $\beta^o_t$, $\beta^o_{st}$ are the fixed effects; $\alpha^o_s$, $\alpha^o_t$ and $\alpha^o_{st}$ are the random effects for space, time and space-time interaction respectively, corresponding to the observed data, and $\alpha^p_t$ and $\alpha^p_{st}$ denote random effects corresponding to predicted values.

To predict these random effects, some results on forecasting using mixed models are required, but first the covariance matrix of the random effects corresponding to the observed and predicted random effects are needed. The covariance matrix is given by $C = \text{diag} (C_1, C_2, C_3)$ where
\[
C_1 = \text{Cov}(\alpha_s) = R_s^T P_s R_s = \begin{pmatrix} F_1^{-1} \\ F_2^{-1} \end{pmatrix},
\]
and here, the vector of variance components is (Clayton, 1993) is used. The smoothing parameters become variance components, (2.8)

Finally, using Equation (2.8), the estimated (corresponding to observed values) α about prediction in mixed models, estimators for 

Expressions for using these covariance matrices and the results provided by Gilmour et al. (2004) about prediction in mixed models, estimators for α and α are given by

Here, R* and R* are part of the transformation matrix T* and they are given by

Expressions for F, F, F, and F are left out in Appendix B. Then, using these covariance matrices and the results provided by Gilmour et al. (2004) about prediction in mixed models, estimators for α and α are given by

(2.8)

To estimate model parameters, penalized quasi-likelihood (Breslow and Clayton, 1993) is used. The smoothing parameters become variance components, and here, the vector of variance components is λ = (λ, λ, λ, λ, λ, λ)'.

Finally, using Equation (2.8), the estimated (corresponding to observed values) and the predicted (corresponding to future values) log-relative risks are given by

(2.9)

3. ILLUSTRATION

To illustrate results, Spanish prostate cancer mortality data from 1975 to 2008 are considered. This data set has been described elsewhere (see Ugarte et al., 2012a, 2012b; Etxeberria et al., 2014) to study different disease mapping models in terms of smoothing and prediction. We use this data set here to make comparisons with the ANOVA-type P-spline model presented in this paper. In brief, a total of 150,616 prostate cancer deaths were registered in Spain during the study period. The number of observed cases ranges from 6 to 651 depending on the province, while the number of expected cases varies from 13.76 to 794.14.
Figure 1 shows the different components of the ANOVA-type P-spline model for some Spanish provinces. Risk projections for 2009–2011 are also provided after fitting the model for the observed data using penalized quasi-likelihood. The smooth thick solid line is used for the total risk estimates and predictions, the dashed line corresponds to the temporal trend common to all areas, and the dashed-dotted line represents the area specific temporal trend. Finally, the non smooth line corresponds to the SMR’s and the thin solid horizontal line is the spatial effect constant along the period. The common temporal trend is below one, and hence it contributes to decrease the mortality risk. The specific temporal trend (dashed-dotted line) can be above or below one increasing or decreasing the risk. For example, in Lugo, it is above one producing an increase in risk, even though it starts to decrease at the end of the period. It is interesting to look at Malaga or Valladolid, where the specific trend contributes to increase the risk in future years, but this is compensated for the global trend which makes the risk decrease.

Figure 1: Risks temporal evolution of the different terms of the ANOVA-type P-spline model and predictions for the years 2009, 2010, and 2011. The smooth thick solid line corresponds to the total risk estimates and predictions, the dashed line represents the temporal trend common to all areas, and the dashed-dotted line is used for the area specific temporal trend. The non smooth line represents the SMR’s, and finally the thin solid horizontal line is the spatial effect constant along the period.
Table 1: Observed counts in 2008; risks predictions for 2011 and their corresponding 95% prediction intervals; expected counts in 2011 and their corresponding 95% prediction intervals.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coruña</td>
<td>161.00</td>
<td>0.76</td>
<td>[0.62, 0.93]</td>
<td>224.63</td>
<td>170.39</td>
<td>[123.85, 216.94]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lleida</td>
<td>71.00</td>
<td>0.72</td>
<td>[0.59, 0.87]</td>
<td>86.24</td>
<td>61.73</td>
<td>[41.04, 82.41]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ourense</td>
<td>91.00</td>
<td>0.76</td>
<td>[0.63, 0.92]</td>
<td>96.26</td>
<td>72.98</td>
<td>[49.72, 96.24]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pontevedra</td>
<td>130.00</td>
<td>0.81</td>
<td>[0.68, 1.11]</td>
<td>136.42</td>
<td>125.13</td>
<td>[89.10, 172.53]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oviedo</td>
<td>196.00</td>
<td>0.82</td>
<td>[0.68, 0.99]</td>
<td>226.20</td>
<td>185.59</td>
<td>[138.76, 232.43]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santander</td>
<td>69.00</td>
<td>0.83</td>
<td>[0.69, 1.00]</td>
<td>107.40</td>
<td>89.58</td>
<td>[63.54, 115.62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lugo</td>
<td>53.00</td>
<td>0.71</td>
<td>[0.60, 0.86]</td>
<td>101.68</td>
<td>72.69</td>
<td>[49.80, 95.58]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Álava</td>
<td>39.00</td>
<td>0.73</td>
<td>[0.61, 0.87]</td>
<td>52.96</td>
<td>38.62</td>
<td>[25.50, 53.27]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guipúzcoa</td>
<td>112.00</td>
<td>0.70</td>
<td>[0.58, 0.84]</td>
<td>125.16</td>
<td>97.28</td>
<td>[63.54, 115.62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vizcaya</td>
<td>83.00</td>
<td>0.71</td>
<td>[0.60, 0.86]</td>
<td>112.94</td>
<td>80.64</td>
<td>[55.08, 105.29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ourense</td>
<td>91.00</td>
<td>0.76</td>
<td>[0.63, 0.92]</td>
<td>96.26</td>
<td>72.98</td>
<td>[49.72, 96.24]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pontevedra</td>
<td>130.00</td>
<td>0.81</td>
<td>[0.68, 1.11]</td>
<td>136.42</td>
<td>125.13</td>
<td>[89.10, 172.53]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oviedo</td>
<td>196.00</td>
<td>0.82</td>
<td>[0.68, 0.99]</td>
<td>226.20</td>
<td>185.59</td>
<td>[138.76, 232.43]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santander</td>
<td>69.00</td>
<td>0.83</td>
<td>[0.69, 1.00]</td>
<td>107.40</td>
<td>89.58</td>
<td>[63.54, 115.62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lugo</td>
<td>53.00</td>
<td>0.71</td>
<td>[0.60, 0.86]</td>
<td>101.68</td>
<td>72.69</td>
<td>[49.80, 95.58]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Álava</td>
<td>39.00</td>
<td>0.73</td>
<td>[0.61, 0.87]</td>
<td>52.96</td>
<td>38.62</td>
<td>[25.50, 53.27]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guipúzcoa</td>
<td>112.00</td>
<td>0.70</td>
<td>[0.58, 0.84]</td>
<td>125.16</td>
<td>97.28</td>
<td>[63.54, 115.62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vizcaya</td>
<td>83.00</td>
<td>0.71</td>
<td>[0.60, 0.86]</td>
<td>112.94</td>
<td>80.64</td>
<td>[55.08, 105.29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ourense</td>
<td>91.00</td>
<td>0.76</td>
<td>[0.63, 0.92]</td>
<td>96.26</td>
<td>72.98</td>
<td>[49.72, 96.24]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pontevedra</td>
<td>130.00</td>
<td>0.81</td>
<td>[0.68, 1.11]</td>
<td>136.42</td>
<td>125.13</td>
<td>[89.10, 172.53]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oviedo</td>
<td>196.00</td>
<td>0.82</td>
<td>[0.68, 0.99]</td>
<td>226.20</td>
<td>185.59</td>
<td>[138.76, 232.43]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santander</td>
<td>69.00</td>
<td>0.83</td>
<td>[0.69, 1.00]</td>
<td>107.40</td>
<td>89.58</td>
<td>[63.54, 115.62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lugo</td>
<td>53.00</td>
<td>0.71</td>
<td>[0.60, 0.86]</td>
<td>101.68</td>
<td>72.69</td>
<td>[49.80, 95.58]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Álava</td>
<td>39.00</td>
<td>0.73</td>
<td>[0.61, 0.87]</td>
<td>52.96</td>
<td>38.62</td>
<td>[25.50, 53.27]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guipúzcoa</td>
<td>112.00</td>
<td>0.70</td>
<td>[0.58, 0.84]</td>
<td>125.16</td>
<td>97.28</td>
<td>[63.54, 115.62]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vizcaya</td>
<td>83.00</td>
<td>0.71</td>
<td>[0.60, 0.86]</td>
<td>112.94</td>
<td>80.64</td>
<td>[55.08, 105.29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For illustration purposes, Table 1 displays the observed counts in 2008 (the last year of the study period), risk predictions for 2011 (three year ahead predictions) together with their 95% prediction intervals; the number of expected cases for 2011 (obtained from projections of population provided by the Spanish Statistical Office), and count predictions for 2011 with their corresponding 95% prediction intervals. Confidence intervals for risks and counts are based on an appropriate estimator of the mean squared error (MSE). Traditionally, the variability associated to the estimation of the variance components have been ignored in empirical Bayes disease mapping and hence, the MSE was underestimated. This can be particularly relevant for the ANOVA-type P-spline model considered here as six smoothing parameters (variance components) are involved. The MSE for the log-risks corresponding to observed data has been derived in a spatial context by Ugarte et al. (2008), Escaramís et al. (2008) and Goicoa et al. (2012), and in a spatio-temporal context by Ugarte et al. (2010) and Ugarte et al. (2012b) when considering CAR, P-splines and ANOVA-type P-spline models. The MSE for predicted log-risks has also been obtained for an interaction P-spline model (Ugarte et al., 2012a), and for CAR and mixtures of CAR and P-spline models (Etxeberria et al., 2014). Using similar tools, the MSE estimator for projections of log-risks derived from the ANOVA-type P-spline model is computed here. The empirical coverage of confidence intervals based on this estimator reveals a good performance. To facilitate the reading of the paper technical details are given in Appendix C.

4. VALIDATION

To assess the predicted ability of the model, a validation study is conducted. We consider the period 1995–2008 to compare the observed with the predicted counts. In brief, data from 1975–1992 are used to fit the model and to predict counts for 1995. Using data till 1993, we forecast counts for 1996 and so on. Three year ahead predictions are considered as this is normally the delay in the registers. Hence, observed counts and three-year ahead predictions from 1995 till 2008 are compared. In this validation period, predictions for 2006 and 2007 were excluded due to computational instabilities in the variance component estimates. Additionally, the models described in Etxeberria et al. (2014) are taken into account for comparison purposes. Namely, an additive model with a CAR structure for space and a random walk of order 2 (RW2) for time; two models with the same structure for space and time and structured and unstructured interactions (Knorr-Held, 2000); an additive model with a CAR structure for space and a P-spline for time; the same additive models with space-time interactions; a model with a common P-spline for time and specific P-splines to describe the temporal evolution of each region, and finally a pure interaction P-spline model. To make the comparison with the ANOVA-type P-spline model fair, predictions for 2006 and 2007 have been also excluded in the previous models.
Table 2 displays empirical coverage rates for prediction intervals corresponding to three, two and one year ahead predictions at nominal values 95%, and 99%. The ANOVA-type P-spline model achieves the nominal values for three year-ahead predictions, the most interesting case from a practical point of view as it is the usual delay in mortality registers. For two and one year ahead predictions, the last three models in the table (all based on P-splines) seem to attain empirical coverage rates closer to the nominal ones. The ANOVA-type P-spline model offers great flexibility and it allows to explicitly split the predictions into components representing region-specific features and characteristics common to the whole country.

<table>
<thead>
<tr>
<th>Model</th>
<th>Three year-ahead predictions</th>
<th>Two year-ahead predictions</th>
<th>One year-ahead predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive CAR RW2</td>
<td>87.67</td>
<td>95.00</td>
<td>85.00</td>
</tr>
<tr>
<td>Interaction CAR RW2 (struc.)</td>
<td>94.33</td>
<td>98.83</td>
<td>93.00</td>
</tr>
<tr>
<td>Interaction CAR RW2 (unstruc.)</td>
<td>93.83</td>
<td>99.17</td>
<td>92.83</td>
</tr>
<tr>
<td>CAR(s)+Pspline(t)</td>
<td>85.83</td>
<td>94.67</td>
<td>87.50</td>
</tr>
<tr>
<td>CAR(s)+Pspline(t)+Int</td>
<td>85.50</td>
<td>93.67</td>
<td>89.33</td>
</tr>
<tr>
<td>Pspline(t) + Pspline Int</td>
<td>94.83</td>
<td>98.17</td>
<td>93.17</td>
</tr>
<tr>
<td>Pure interaction Pspline</td>
<td>93.33</td>
<td>99.17</td>
<td>93.00</td>
</tr>
<tr>
<td>ANOVA-type Pspline</td>
<td>95.33</td>
<td>99.00</td>
<td>92.33</td>
</tr>
</tbody>
</table>

5. DISCUSSION

Statistical methods represent a valid scientific tool to make predictions about future events taking into account past information. These statistical methods gain importance in an epidemiological context since official cancer death figures are available after approximately three years from current date due to the delay in administrative procedures of data collection and registration.

Some models including CAR, P-splines and combinations of both have been studied in the literature (see for example Etxeberria et al., 2014) to provide predictions of mortality or incidence counts. In this paper, an ANOVA-type P-spline model is studied to complete the P-spline alternatives within a generalized linear mixed model framework. An extended transformation matrix, including the spatial and temporal additive terms, and the spatio-temporal interaction is derived in order to express risks related to observed and future time periods in a single mixed model. The MSE of the predicted log-risks is also provided accounting for prediction intervals.
for all sources of variability, including the one coming from the estimation of the smoothing parameters, and is used in turn to calculate the count prediction error. The model has good empirical coverage rates for three year ahead predictions and in addition, it is very attractive as it explicitly considers one smooth term for space, another one for time, and a final interaction term, each one with its own smoothing parameters. This allows to split the predicted risk into a spatial component constant along the time period, a smooth temporal term common to all regions and an area specific term representing the specificity of a region. This is of practical interest as the area specific term indicates whether the region contributes to increase or decrease its own risk, and hence it helps to plan prevention or intervention measures and epidemiological policies in general.
A. APPENDIX

To understand how the extended mixed model (2.7) is obtained, detailed expressions for the different matrices are provided in this section. Using the transformation matrix $T^*$, the fixed and random effect matrices of the extended generalized linear mixed model are given by

$$B^*T^* = [1^*_s: (1^*_t \otimes B_s)T_s: (B^*_t \otimes 1_s)T^*_t: (B^*_t \otimes B_s)T^*_st]$$

where

$$ (1^*_t \otimes B_s)T_s = 1^*_t \otimes [x_s: [Z_2 \otimes X_1: X_2 \otimes Z_1: Z_2 \otimes Z_1]] = 1^*_t \otimes [x_s: Z_a]$$

$$ (B^*_t \otimes 1_s)T^*_t = \begin{bmatrix} B_t u_{3s} & B_t U_{3s} & 0 \\ B_{t1} u_{3n} & B_{t1} U_{3s} & B_{t1} L_{t1}^{-1} \end{bmatrix} \otimes 1_s = \begin{bmatrix} x^o_t \otimes 1_s & Z^o_t & 0 \\ x^p_t \otimes 1_s & Z^p_t & Z^o_t_s \\ Z^p_t_s & Z^o_t_s \\ 0 \\ Z^p_t_s & 0 \end{bmatrix}$$

Here, $Z_s = (1^*_t \otimes Z_a)$, $Z_a = [Z_2 \otimes X_1: X_2 \otimes Z_1: Z_2 \otimes Z_1]$, $X_a = (1^*_t \otimes x_s)$, $x_s = [1_n \otimes x_1: x_2 \otimes 1_n : x_2 \otimes x_1]$, $X_1 = [1: x_1]$, $X_2 = [1: x_2]$, $Z_1 = B_{t1} U_{1s}$, $Z_2 = B_{t2} U_{2s}$ and $Z_3 = B_t U_{3s}$. Finally, $x_1$ and $x_2$ are the column vectors of longitude and latitude respectively, and $x^o_t$ and $x^p_t$ are the column vector of time corresponding to observed and prediction period respectively. Using these results, the extended model is (2.7) is attained.
In this section, and to make the reading easier, expressions for matrices $F_1, F_2, F_3, F_4,$ and $F = \text{blockdiag}(F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11})$ are given. These matrices are different blocks of the covariance matrix of the random effects coming from the mixed model representation of the ANOVA-type P-spline model. Note that $F_i, i = 1, ..., 8$ are exactly the same as those in Ugarte et al. (2012b). $F_9, F_{10}, F_{11}$ are not the same because in this paper we have considered $B_t^* \otimes B_s$, the last term in the extended basis (2.3), instead of the other way around $B_s \otimes B_t^*$. This has been done because it is more natural and convenient when extending the time basis to make predictions. Expressions for these matrices are given by

\[
F_1 = \lambda_2 \tilde{\Sigma}_2 \otimes I_2, \quad F_2 = \lambda_1 I_2 \otimes \tilde{\Sigma}_1, \quad F_3 = \lambda_1 I_{m_2-2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes I_{m_1-2},
\]

\[
F_4 = \lambda_1 \tilde{\Sigma}_2, \quad F_5 = \tau_2 \tilde{\Sigma}_2 \otimes I_2, \quad F_6 = \tau_1 I_2 \otimes \tilde{\Sigma}_1,
\]

\[
F_7 = \tau_1 I_{m_2-2} \otimes \tilde{\Sigma}_1 + \tau_2 \tilde{\Sigma}_2 \otimes I_{m_1-2}, \quad F_8 = \tau_1 I_2 \otimes \tilde{\Sigma}_3,
\]

\[
F_9 = \tau_2 I_{m_2-2} \otimes \tilde{\Sigma}_2 \otimes I_2 + \tau_1 \tilde{\Sigma}_3 \otimes I_{m_2-2} \otimes I_2,
\]

\[
F_{10} = \tau_1 I_{m_3-2} \otimes I_2 \otimes \tilde{\Sigma}_1 + \tau_2 I_{m_2-2} \otimes \tilde{\Sigma}_2 \otimes I_{m_1-2},
\]

\[
F_{11} = \tau_1 I_{m_3-2} \otimes I_{m_2-2} \otimes \tilde{\Sigma}_1 + \tau_2 I_{m_3-2} \otimes \tilde{\Sigma}_2 \otimes I_{m_1-2} + \tau_1 \tilde{\Sigma}_3 \otimes I_{m_2-2} \otimes I_{m_1-2}.
\]

where $\tilde{\Sigma}_i, i = 1, 2, 3$ are diagonal matrices of non zero eigenvalues coming from the eigen-decomposition of the marginal penalties $P_{s1}, P_{s2}$ and $P_t$ respectively. $F^{-1} = \text{blockdiag}(F_5^{-1}, F_6^{-1}, F_7^{-1}, F_8^{-1}, F_9^{-1}, F_{10}^{-1}, F_{11}^{-1}), I_s = I_{m_2} \otimes I_{m_1}, I^* = I_r \otimes I_s,$ and $I_r$ is the identity matrix of dimension $r \times r$ where $r$ is the number of columns of $L_t$. 

---

**B. APPENDIX**

In this section, and to make the reading easier, expressions for matrices $F_1, F_2, F_3, F_4,$ and $F = \text{blockdiag}(F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11})$ are given. These matrices are different blocks of the covariance matrix of the random effects coming from the mixed model representation of the ANOVA-type P-spline model. Note that $F_i, i = 1, ..., 8$ are exactly the same as those in Ugarte et al. (2012b). $F_9, F_{10}, F_{11}$ are not the same because in this paper we have considered $B_t^* \otimes B_s$, the last term in the extended basis (2.3), instead of the other way around $B_s \otimes B_t^*$. This has been done because it is more natural and convenient when extending the time basis to make predictions. Expressions for these matrices are given by

\[
F_1 = \lambda_2 \tilde{\Sigma}_2 \otimes I_2, \quad F_2 = \lambda_1 I_2 \otimes \tilde{\Sigma}_1, \quad F_3 = \lambda_1 I_{m_2-2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes I_{m_1-2},
\]

\[
F_4 = \lambda_1 \tilde{\Sigma}_2, \quad F_5 = \tau_2 \tilde{\Sigma}_2 \otimes I_2, \quad F_6 = \tau_1 I_2 \otimes \tilde{\Sigma}_1,
\]

\[
F_7 = \tau_1 I_{m_2-2} \otimes \tilde{\Sigma}_1 + \tau_2 \tilde{\Sigma}_2 \otimes I_{m_1-2}, \quad F_8 = \tau_1 I_2 \otimes \tilde{\Sigma}_3,
\]

\[
F_9 = \tau_2 I_{m_2-2} \otimes \tilde{\Sigma}_2 \otimes I_2 + \tau_1 \tilde{\Sigma}_3 \otimes I_{m_2-2} \otimes I_2,
\]

\[
F_{10} = \tau_1 I_{m_3-2} \otimes I_2 \otimes \tilde{\Sigma}_1 + \tau_2 I_{m_2-2} \otimes \tilde{\Sigma}_2 \otimes I_{m_1-2},
\]

\[
F_{11} = \tau_1 I_{m_3-2} \otimes I_{m_2-2} \otimes \tilde{\Sigma}_1 + \tau_2 I_{m_3-2} \otimes \tilde{\Sigma}_2 \otimes I_{m_1-2} + \tau_1 \tilde{\Sigma}_3 \otimes I_{m_2-2} \otimes I_{m_1-2}.
\]
The MSE for predicted log-risk has already been proposed for a three-dimensional P-spline model (Ugarte et al., 2012a). Here we reproduce the expressions and make explicit the specific formula for the $M$ matrix in the ANOVA-type P-spline model. An estimator for the MSE of the predicted log-risk is given by

$$
\hat{MSE}[\hat{u}_{st}^p] = g_{1st}^*(\lambda) + g_{2st}^*(\lambda) + 2g_{3st}^*(\lambda).
$$

where

$$
g_{1st}^*(\lambda) = z_{st}^p(C - MZ\lambda V^{-1}Z'M)z_{st}^p',
$$

$$
g_{2st}^*(\lambda) = (x_{st}^p - z_{st}^pMZ\lambda Xo)(X'o'V^{-1}X'o)^{-1}(x_{st}^p - z_{st}^pMZ\lambda Xo)'',
$$

$$
g_{3st}^*(\lambda) = \text{tr}[S^*VS'^{'}\mathcal{I}^{-1}].
$$

Here $V$ and $\mathcal{I}^{-1}$ are the covariance matrix of the working vector and the asymptotic covariance matrix of the variance components estimators arising from the PQL algorithm. Vectors $z_{st}^p$ and $x_{st}^p$ are the $st$ row of the matrices $Z^p = [Z_s : Z_t^p : Z_{st}^p : Z_{stp}]$ and $X^p = [X_s : X_t^p : X_{st}^p]$ respectively, and finally, $Z^o = [Z_s : Z_t : Z_{st}]$ and $X^o = [X_s : X_t^o : X_{st}^o]$. An explicit expression for $M$ is given by

$$
M = \begin{pmatrix}
C_1 \\
F_4^{-1} \\
-\mathcal{E}_t\mathcal{R}_tF_4^{-1} \\
F^{-1} \\
-(\mathcal{E}_t \otimes I_s)\mathcal{R}_{st}F^{-1}
\end{pmatrix}
$$

If $\lambda_j$ denotes the $j$th entry of the vector of variance components $\lambda = (\lambda_{s1}, \lambda_{s2}, \lambda_t, \tau_s, \tau_{s2}, \tau_t)'$, the matrix $S^*$ is given by

$$
S_j^* = z_{st}^p(\frac{\partial M}{\partial \lambda_j}Z'V^{-1} + MZ'\frac{\partial V^{-1}}{\partial \lambda_j}), \quad j = 1, 2, 3, 4, 5, 6.
$$

Finally, the variance for predicted counts is calculated as

$$
\text{Var}[C_{st}^p] = \text{E}[\text{Var}[C_{st}^p|r_{st}^p]] + \text{Var}[\text{E}[C_{st}^p|r_{st}^p]] = e_{st}^p\text{E}[r_{st}^p] + e_{st}^{p2}\text{Var}[r_{st}^p],
$$

where $e_{st}^p$ are projections of the number of expected cases for future years. $\text{Var}[r_{st}^p]$ is easily estimated from $\hat{MSE}[\hat{u}_{st}^p]$ using the delta method.
ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Science and Innovation (project MTM 2011-22664 co-funded with FEDER grants) and by the Health Department of the Navarre Government (project 113, Res.2186/2014). We would like to thank the National Epidemiology Center (area of Environmental Epidemiology and Cancer) for providing the data, originally created by the Spanish Statistical Office.

REFERENCES


On Predicting Cancer Mortality using ANOVA-type P-spline Models


