ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
80 Anos years 1935 2015

REVSTAT
Statistical Journal

Special issue on
«Spatial-temporal Models in Epidemiology and Health»

Guest Editors:
Giovani Loiolda da Silva
Maria Antónia Amaral Turkman

Volume 13, No.1
March 2015

REVSTAT

Statistical Journal

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística = ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **PUBLISHER**
  - *Instituto Nacional de Estatística, I.P. (INE, I.P.)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: +351 218 426 100
    Fax: +351 218 454 084
    Web site: http://www.ine.pt
    Customer Support Service
    (National network): 808 201 808
    (Other networks): +351 218 440 695
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE, I.P., by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística, I.P.*
- **EDITION**
  - *150 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*
- **PRICE [VAT included]**
  - *€ 9,00*

FOREWORD

The One-day Workshop on Spatial-temporal Models in Epidemiology and Health was held in Lisbon, Portugal, on June 20th, 2014. It was organized by the Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), under the ambit of the FCT research projects: PTDC/MAT/118335/2010 and PEst-OE/MAT/UI0006/2014. The workshop aimed (i) to bring together researchers who work in epidemiology and health sciences; (ii) to promote a fruitful discussion on the role of statistical methods, such as time series analysis and spatial statistics.

Due to the proliferation of several studies involving data sets that are both spatially and temporally indexed, spatial-temporal modeling has received an increasing attention in the last few years. Spatial-temporal data are usually related to applied areas, such as epidemiology and health sciences. Six invited speakers presented their methodological advancements in those areas with a heavy emphasis on applications, whereas a poster session with contributed papers complemented the scientific program of the workshop.

This special issue of *REVSTAT — Statistical Journal* consists of peer-refereed papers generated from the research work presented at the workshop. This collection of papers reflects the stimulating research in the area of Epidemiology and Health Statistics, covering a wide range of topics, such as Spatio-temporal detection of influenza outbreaks, Assessing the evolution of territorial disparities in health, Space-time disease mapping, Bayesian projections, Longitudinal dynamic models and Estimating the long-term health effects of air pollution.

We would like to to thank all authors for their contributions and all the anonymous reviewers who helped to prepare this special issue. Furthermore, we are grateful to the Editor-in-Chief of *REVSTAT — Statistical Journal* for agreeing to publish this special issue, as well as to all members of the scientific and organizing committees who worked to make the workshop a very interesting event for discussing Spatial-temporal Models in Epidemiology and Health.

Giovani Loiola da Silva
CEAUL & Dep. Mathematics-IST
Universidade de Lisboa
giovani.silva@tecnico.ulisboa.pt

Maria Antónia Amaral Turkman
CEAUL & Faculdade de Ciências
Universidade de Lisboa
maturkman@fc.ul.pt

INDEX

Assessing the Evolution of Territorial Disparities in Health

Daniela Cocchi, Fedele Greco and *Francesco Scalone* 1

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

Jaione Etxeberria, María Dolores Ugarte, Tomás Goicoa and *Ana F. Militino* 21

Statistical Methods for Detecting the Onset of Influenza Outbreaks: A Review

Rubén Amorós, David Conesa, Miguel Angel Martinez-Beneito and *Antonio López-Quílez* 41

Longitudinal Analysis of Tumor Marker CEA of Breast Cancer Patients from Braga's Hospital

Ana Borges, Inês Sousa and *Luís Castro* 63

Alcohol Abuse Disorder Prevalence and its Distribution across Portugal. A Disease Mapping Approach

Helena Baptista, Jorge M. Mendes, José Caldas de Almeida and *Miguel Xavier* 79

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 13, Number 1, March 2015, 1–20

ASSESSING THE EVOLUTION OF TERRITORIAL DISPARITIES IN HEALTH

Authors: Daniela Cocchi
– Department of Statistical Sciences, University of Bologna, Italy
daniela.cocchi@unibo.it

Fedele Greco
– Department of Statistical Sciences, University of Bologna, Italy
fedele.greco@unibo.it

Francesco Scalone
– Department of Statistical Sciences, University of Bologna, Italy
francesco.scalone@unibo.it

Received: October 2014 Revised: November 2014 Accepted: December 2014

Abstract:

- The paper investigates spatio-temporal trends in health disparities through an empirical example. We deal with geographical health pattern in Italy from 1991 to 2010, starting from infant mortality data available at the provincial level and assessing the existent disparity among macro-regions (the conventional Northern, Central and Southern macro-regions). After a discussion concerning suitable inequality indices and their decompositions when dealing with small area data, we propose a model-based approach that allows to properly tackle sampling variability. Results give evidences of persisting spatial disparity in infant mortality along time.

Key-Words:

- *health inequality; Gini index; generalised entropy; hierarchical Bayesian modelling; small area estimation; spatio-temporal smoothing; sampling variability; INLA.*

AMS Subject Classification:

- 62F15, 62P25, 91B72, 91B82.

1. INTRODUCTION

In this note, we investigate spatio-temporal trends in health disparities, expressed by infant mortality, through an empirical example. We deal with geographical health pattern in Italy from 1991 to 2010, starting from data available at the provincial level and assessing the existent disparity among macro-regions (the conventional Northern, Central and Southern macro-regions). The evaluation of the temporal evolution of inequality requires the adoption of suitable indicators that, along with their decomposition, help in answering a couple of main questions: is inequality between small geographical units decreasing during the study period? Which is the trend of the inequality share explained by grouping the smaller geographical units in macro-regions? As in recent years spatial disparities are being investigated in depth, the above research questions are more and more crucial. Answers to such questions are critical to improve and implement better public policies.

As a matter of fact, persistent health inequalities in modern welfare states represent a great disappointment in public health, with widening disparities reported in many Western European countries (Mackenbach, 2012). Since health inequality could persist within the same country among different regions, appropriate statistical methods to describe the spatio-temporal evolution of this phenomenon are desirable. In this study, we propose suitable decompositions of inequality indices equipped with uncertainty measures as the mean to evaluate the temporal evolution of health inequality in Italy along a twenty-years period. In particular, we focus on infant mortality, that is one of the main indicators to measure the general health level. Health disparities can be described by means of a variety of statistical measures, such as dispersions measures or inequality indices (Wagstaff *et al.*, 1991). In order to assess the presence of geographic disparities in morbidity and mortality, various authors suggested measuring health inequality by means of the Generalized Entropy and Gini indices.

In the Italian case, despite a general declining trend, some studies found high dispersion in Infant Mortality Rates (IMRs) at provincial level, revealing evident and persisting geographical disparity in infant mortality. This persisting disparity was mainly related to differences in socio-economic and health care standards among Northern, Central and Southern Italian macro-regions (Fantini *et al.*, 2005; Lauria and De Stavola, 2003).

When studying provincial-level infant mortality, IMRs show large random fluctuations giving rise to relevant methodological issues concerning the evaluation and decomposition of health disparities: more precisely, because of the low birth counts observed in the provinces and because of the rarity of infant deaths, Italian provinces have to be considered as small areas and direct estimates (*i.e.* IMRs) are subject to high sampling variability that we will tackle

by means of a model-based approach. The consequences of sampling variability on the measurement and decomposition of appropriate inequality measures constitute the main methodological focus of the paper. Literature concerning sampling variability of Gini index and Generalised Entropy measures are centred on the classical case where available data constitute a sample from a larger population, but individual values of the study variable are considered as measured without error. Instead, due to the peculiarities of the motivating example, we address the situation where the whole population has been observed (*i.e.* data concerning birth and death counts are available for each province), but individual values of the study variable need to be estimated. In this situation, the sampling properties of inequality indices estimators depend on the sampling properties of individual-level estimators. According to our knowledge, this topic has been neglected in the literature concerning health inequalities. Proper smoothing techniques need to be used in order to limit potential biases due to sample variability (Congdon *et al.*, 2001; Congdon, 2010). Adopting a popular approach to spatio-temporal disease mapping (Kim and Lim, 2010; Knorr-Held, 2000; Blangiardo *et al.*, 2013), we estimate a Bayesian smoothing model exploiting spatial association of provincial IMRs and temporal correlation. The model allows reciprocal borrowing strength for area-level data, with the least reliable rates (based on the smallest birth counts) being mostly smoothed. Model fitting is performed via Integrated Nested Laplace Approximations (INLA, Rue and Martino, 2009).

The outline of this paper is as follows. Section 2 provides a brief description of the data concerning Italian infant mortality. Section 3 introduces inequality measures and their decomposition. A simulation study is discussed for highlighting some relevant features of inequality estimators. Section 4 describes the Bayesian spatio-temporal model adopted for smoothing mortality rates: computational details are provided. In Section 5, model-based inequality decomposition is presented. In the concluding section, evidences of persisting disparity in infant mortality are briefly discussed, illustrating the contribution of the differences among macro-regions.

2. MOTIVATING EXAMPLE

Yearly data about infant mortality, in our study referred to 95 provinces along 20 years (1991–2010), are published by the Italian Institute of Statistics (ISTAT). At each year $t = 1, ..., T$, province $s = 1, ..., S$, and macro-region $k = 1, ..., K$, infant death counts y_{skt} and birth counts P_{skt} are available. For each year, province and macro-region, the Infant Mortality Rate (IMR):

$$\hat{\theta}_{skt} = \frac{y_{skt}}{P_{skt}} , \qquad s = 1, ..., S, \quad k = 1, ..., K, \quad t = 1, ..., T , \tag{2.1}$$

is a quick measure of the infant mortality intensity. In particular, $\hat{\theta}_{skt}$ is the maximum likelihood estimator of the true mortality rate θ_{skt} according to the model

$$y_{skt}|\theta_{skt} \sim Poisson(\theta_{skt}P_{skt}) \ , \qquad s = 1, ..., S, \ \ k = 1, ..., K, \ \ t = 1, ..., T \, . \tag{2.2}$$

Estimator (2.1) is unbiased and its sampling variance is inversely proportional to the birth count P_{skt}, in fact $V(\hat{\theta}_{skt}|\theta_{skt}) = \theta_{skt}/P_{skt}$. It turns out that estimates referred to provinces with low birth counts are affected by huge sampling variability, while estimates based on high birth counts are more stable. This well-known feature of mortality rates gave rise to an extensive literature concerning spatial and spatio-temporal disease mapping, aiming at smoothing observed rates (Kim and Lim, 2010; Knorr-Held, 2000; Blangiardo *et al.*, 2013). The consequences of sampling variability on the measurement and decomposition of appropriate inequality measures are discussed in the following section and constitute the main methodological focus of the paper.

Figure 1 plots the IMRs series of all Italian provinces during the study period (reported in gray). Black lines refer to IMRs observed in the Northern, Central and Southern macro-regions. A general decline of the mortality level is observed along the study period in all the macro-regions, reflecting the general trend at the provincial level. At the national level, IMR declines from 0.008 to 0.003, but a general decline in the mortality intensity does not imply a decline in territorial inequalities.

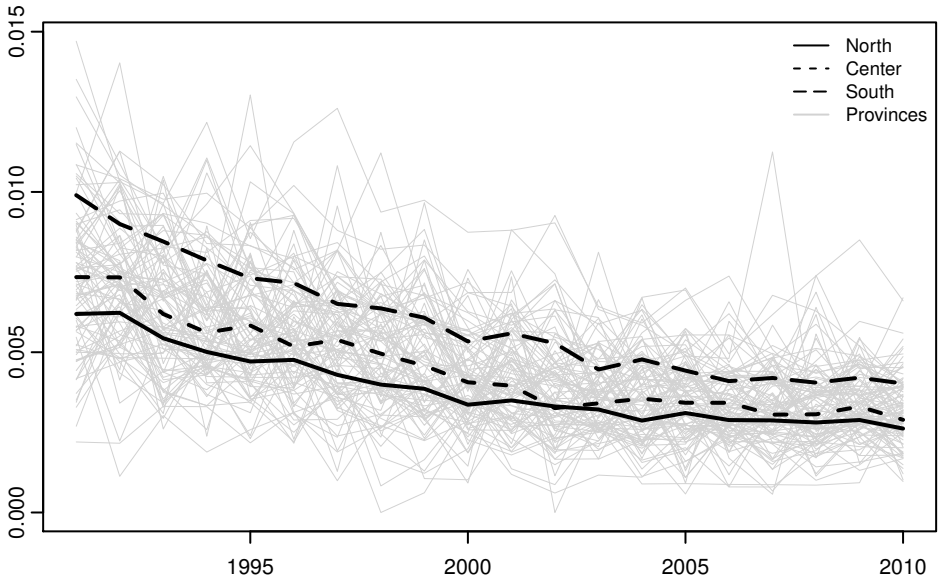

Figure 1: Temporal trend of IMRs at provincial and macro-region level.

Figure 2 reports the spatial distribution of IMRs classifying each province according to the octiles identified for the selected years 1991, 1994, 1998, 2002, 2006 and 2010. A persistent spatial trend occurs since southern provinces systematically register higher infant mortality with respect to northern provinces. In the following of this paper, we discuss how this territorial disparity can be evaluated focusing on both the overall inequality and the share of inequality explained by grouping provinces in macro-regions.

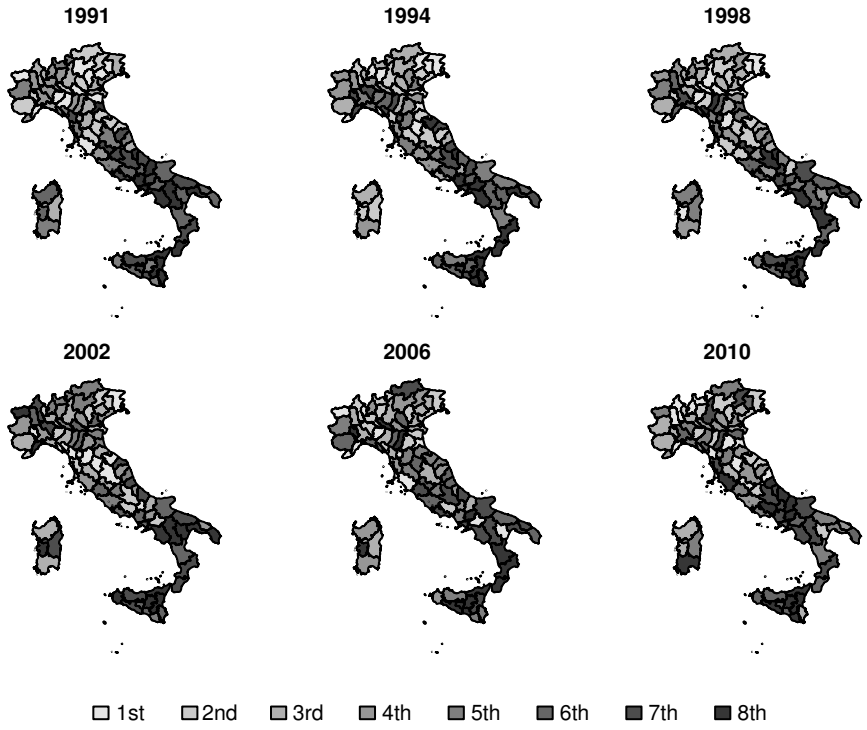

Figure 2: Octiles of the IMRs spatial distribution in selected years.

3. INEQUALITY MEASURES AND THEIR DECOMPOSITION

The theory concerning the measurement of inequalities has a long history and has been developed essentially in the framework of income distribution (Dalton, 1920; Atkinson, 1970; Dreher and Gaston, 2008). The same theory has been subsequently turned to the study of health inequalities at several spatial and temporal levels of aggregation. A popular approach defines health inequality as the uneven distribution of health across all units in a population and in population subgroups (see *e.g.*, Gakidou and King, 2002; Pradhan *et al.*, 2003). In this section, following this approach, we consider a framework where population

units are constituted by small geographical areas grouped in macro-regions and discuss some crucial statistical properties of inequality measures estimators when small area rates are involved. In particular, we focus on two popular inequality indicators: the Generalised Entropy class of indicators and the Gini coefficient. For simplifying notation, we drop the temporal subscript in what follows.

Given a population of S areas organised in K groups, the number of areas belonging to the k-th group is denoted as S_k, such that $\sum_{k=1}^{K} S_k = S$. Let $\boldsymbol{\theta}_k = (\theta_{1k}, ..., \theta_{sk}, ..., \theta_{S_k k})$ and $\mathbf{P}_k = (P_{1k}, ..., P_{sk}, ..., P_{S_k k})$ denote respectively the "true" mortality rates and the number of births referred to group k. Group-specific rates are weighted averages of area-specific rates denoted as $\bar{\theta}_k = \sum_{s=1}^{S_k} P_{sk}\theta_{sk}/P_k$, where $P_k = \sum_{s=1}^{S_k} P_{sk}$.

The Generalised Entropy is defined as:

$$GE(\boldsymbol{\theta}; \alpha) = \frac{1}{\alpha(\alpha-1)} \sum_{k=1}^{K} \sum_{s=1}^{S_k} \frac{P_{sk}}{P} \left(\left(\frac{\theta_{sk}}{\bar{\theta}} \right)^{\alpha} - 1 \right), \qquad \alpha \neq 0, 1 \ , \tag{3.1}$$

where α controls the weight assigned to the distance between mortality rates at different parts of the rates distribution: for negative/positive values of α, GE is more sensitive to changes in the lower/upper tail of the distribution. The GE class of inequality measures includes as special cases, among others, the Theil index ($\alpha = 0$) and the Coefficient of Variation ($\alpha = 2$, where the GE is equivalent to half times the squared coefficient of variation, or relative variance). The GE class of inequality measures is easily decomposable in the between and within group components. Namely, the between component is expressed as:

$$GE_B(\boldsymbol{\theta}; \alpha) = \frac{1}{\alpha(\alpha-1)} \sum_{k=1}^{K} \frac{P_k}{P} \left(\left(\frac{\bar{\theta}_k}{\bar{\theta}} \right)^{\alpha} - 1 \right),$$

a weighted average of the distances between the group means and the overall mean. The within component is expressed as a linear combination of the GEs in each sub-group

$$GE_W(\boldsymbol{\theta}; \alpha) = \sum_{k=1}^{K} \frac{P_k}{P} \left(\frac{\bar{\theta}_k}{\bar{\theta}} \right)^{\alpha} GE_{Wk} \ ,$$

where GE_{Wk} is the GE in the k-th group:

$$GE_{Wk} = GE(\boldsymbol{\theta}_k; \alpha) = \frac{1}{\alpha(\alpha-1)} \sum_{s=1}^{S_k} \frac{P_{sk}}{P_k} \left(\left(\frac{\theta_{sk}}{\bar{\theta}_k} \right)^{\alpha} - 1 \right).$$

Eventually, GE is decomposed in the between and within components as:

$$GE(\boldsymbol{\theta}; \alpha) = GE_B(\boldsymbol{\theta}; \alpha) + GE_W(\boldsymbol{\theta}; \alpha)$$

and the contribution of grouping to the global inequality can be evaluated as the ratio:

$$GE_B(\boldsymbol{\theta}; \alpha)/GE(\boldsymbol{\theta}; \alpha) \ . \tag{3.2}$$

As pointed out in Dagum (1997), the decomposition of GE-type inequality measures is essentially based on the hypotheses underlying one-way analysis of variance, neglecting differences in variances and asymmetry characterising subgroups, and delivering the between component from comparisons between group means. A nicer and more detailed picture of inequality decomposition can be obtained starting from the Gini index defined as:

$$G(\boldsymbol{\theta}) = \frac{1}{2\bar{\theta}} \sum_{h=1}^{K} \sum_{k=1}^{K} \sum_{s=1}^{S_h} \sum_{v=1}^{S_k} \frac{P_{sh}}{P} \frac{P_{vk}}{P} |\theta_{sh} - \theta_{vk}| \tag{3.3}$$

according to the proposal of Dagum (1997). This decomposition considers three components measuring respectively within-group inequality, net between-group inequality and transvariation (*i.e.* overlapping) between groups. The component due to transvariation represents one strong peculiarity of Gini's index with respect to the GE decomposition. For simplifying notation, it is assumed that the group means are ordered as $\bar{\theta}_1 \leq ... \leq \bar{\theta}_k \leq ... \leq \bar{\theta}_K$. The decomposition starts by defining the Gini index between the couple of groups h and k as:

$$G_{hk} = \frac{1}{\bar{\theta}_h + \bar{\theta}_k} \sum_{s=1}^{S_h} \sum_{v=1}^{S_k} \frac{P_{sh}}{P_h} \frac{P_{vk}}{P_k} |\theta_{sh} - \theta_{vk}| \ . \tag{3.4}$$

For $h = k$, expression (3.4) corresponds to the Gini index of the k-th group. It immediately turns out that (3.3) can be written as a function of (3.4) as:

$$G(\boldsymbol{\theta}) = \sum_{h=1}^{K} \sum_{k=1}^{K} \frac{P_h}{P} \left(\frac{P_k \bar{\theta}_k}{P\bar{\theta}} \right) G_{hk} = \sum_{h=1}^{K} \sum_{k=1}^{K} q_h r_k G_{hk} \ , \tag{3.5}$$

where $q_h = P_h/P$ is the population share of the h-th group and $r_k = (P_k\bar{\theta}_k)/(P\bar{\theta})$ can be interpreted as the share of expected death counts in the k-th group. Since $\sum_{h=1}^{K} \sum_{k=1}^{K} q_h r_k = 1$, the Gini index can be expressed as a weighted average of the between groups Gini indices G_{hk}; on the contrary, it is not possible to express GE-based decompositions as weighted averages, since the weights do not sum up to one. Coefficients G_{hk}, properly combined with weights q_h and r_k, allow to decompose the Gini index in three components. The first one is

$$G_W(\boldsymbol{\theta}) = \sum_{k=1}^{K} q_k r_k G_{kk} \ , \tag{3.6}$$

which measures the contribution of within group inequality. The following expression of the between component is due to Costa (2009):

$$G_B(\boldsymbol{\theta}) = \sum_{h=1}^{K-1} \sum_{k=h+1}^{K} \frac{r^*_{hk} - q^*_{hk}}{r^*_{hk} q^*_{kh} + r^*_{kh} q^*_{hk}} (q_h r_k + q_k r_h) \ , \tag{3.7}$$

where $r^*_{hk} = r_h/(r_h + r_k)$ and $r^*_{hk} = q_h/(q_h + q_k)$. The component due to transvariation, denoted in what follows as $G_T(\theta)$, can be obtained by difference. The

between component of the Gini index has the merit to take into account pairwise differences between individuals instead of being entirely based on comparisons among group means: for this reasons it should be preferred to GE-like indices. Eventually, the decomposition

$$G(\boldsymbol{\theta}) = G_W(\boldsymbol{\theta}) + G_B(\boldsymbol{\theta}) + G_T(\boldsymbol{\theta}) \tag{3.8}$$

is obtained.

For the purpose of this work, we consider $G(\boldsymbol{\theta})$, $G_B(\boldsymbol{\theta})$, $G_W(\boldsymbol{\theta})$, $G_T(\boldsymbol{\theta})$, $GE(\boldsymbol{\theta};\alpha)$, $GE_B(\boldsymbol{\theta};\alpha)$ and $GE_W(\boldsymbol{\theta};\alpha)$ as target population parameters to be estimated.

3.1. The use of direct estimates

Direct estimates $\hat{\theta}_{sk}$ plugged in the expression of population quantities are a popular way for estimating inequality indices. Figures 3 and 4 report estimates $\widehat{G}(\boldsymbol{\theta})$, $\widehat{G}_B(\boldsymbol{\theta})$, $\widehat{G}_W(\boldsymbol{\theta})$, $\widehat{G}_T(\boldsymbol{\theta})$, $\widehat{GE}(\boldsymbol{\theta};\alpha)$, $\widehat{GE}_B(\boldsymbol{\theta};\alpha)$ and $\widehat{GE}_W(\boldsymbol{\theta};\alpha)$ of these inequality measures for each year in the interval 1991–2010 concerning Italian infant mortality, obtained by simply plugging-in direct estimates of the mortality intensity. As an example, at each year, $\widehat{G}(\boldsymbol{\theta})$ is obtained as:

$$\widehat{G}(\boldsymbol{\theta}) = \frac{1}{2\hat{\bar{\theta}}} \sum_{h=1}^{K} \sum_{k=1}^{K} \sum_{s=1}^{S_h} \sum_{v=1}^{S_k} \frac{P_{sh}}{P} \frac{P_{vk}}{P} |\hat{\theta}_{sh} - \hat{\theta}_{vk}| \tag{3.9}$$

employing the direct estimates (2.1).

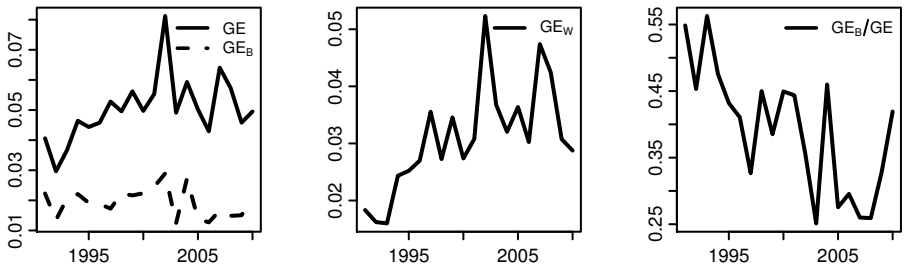

Figure 3: Plug-in estimates. Generalised Entropy and Between Generalised Entropy (left panel). Within Generalised Entropy (middle panel). Effect due to the between component (right panel).

Estimates of inequality indicators show a noisy temporal trend that suggests an increasing weight of the within components (see middle panels of the following Figures 3 and 4) and a decreasing weight of the between component

(see Figure 3 right panel and Figure 4 lower left panel). In particular, according to the GE index, the between component accounts for 55% of total inequality in 1991, decreases to 25% in year 2008 and then shows a further increase to 36% in 2010. The between component of the Gini index accounts for 70% of total inequality in 1991, decreases to 49% in year 2008 and then shows a further increase to 60% in 2010. Our purpose is to show that these estimates should not be considered as reliable pictures of territorial disparity in Italian infant mortality, since they are heavily affected by the sampling variability of direct estimates.

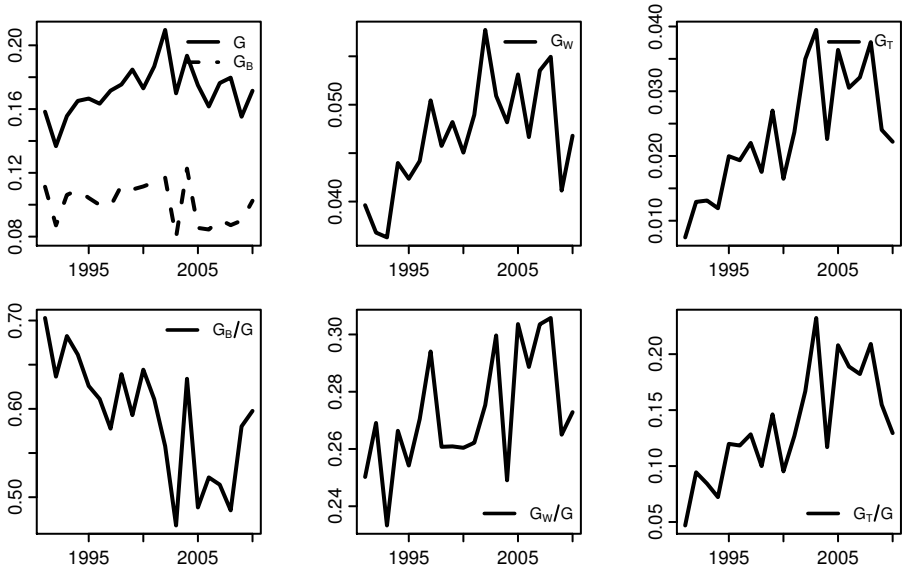

Figure 4: Plug-in estimates. Gini index and its between component (upper left panel). Within component (upper middle panel). Transvariation component (upper right panel). Lower panels report the contribution of each component to the total.

Literature concerning sampling variability of Gini index and Generalised Entropy measures focuses on the classical case where available data constitute a sample from a larger population, but individual values of the study variable are considered as measured without error (see for example Langel and Tillé, 2013). Instead, due to the peculiarities of the motivating example, we address the situation where the whole population has been observed (*i.e.* data concerning birth and death counts are available for each province), but the individual values of study variable (*i.e.* mortality intensity θ_{sk}) need to be estimated. In this situation, the sampling properties of inequality indices estimators depend on the sampling properties of individual-level estimators. According to our knowledge, this topic has been neglected in the literature concerning health inequalities. In Subsection 3.2 the effect of estimating inequality measures by simply plugging-in direct estimates $\hat{\theta}_{sk}$ is discussed.

3.2. The effect of sampling variability on decomposition

In order to discuss the consequences of direct estimates sampling variability on the estimation of inequality measures decomposition, we design a simulation study that considers a population partitioned in $K = 3$ groups, where the whole inequality is explained by the between-group component, while equality within groups is postulated. We set $\bar{\theta}_1 = 0.6\ \bar{\theta}$, $\bar{\theta}_2 = \bar{\theta}$ and $\bar{\theta}_3 = 1.4\ \bar{\theta}$. Moreover, we set $\theta_{sk} = \bar{\theta}_k\ \forall s, k$, such that the within component of any inequality measure equals 0, *i.e.* $G_W(\boldsymbol{\theta}) = GE_W(\boldsymbol{\theta}; \alpha) = 0$. In order to investigate the effect of mortality intensity, we let $\bar{\theta}$ vary between .002 and .009, similarly to the national mortality levels observed between 1991 and 2010 in Italian infant mortality. For the sake of simplicity, from now on, we set $\alpha = 1.5$ when dealing with the GE index. With this setting, for any $\bar{\theta}$, we obtain $G(\boldsymbol{\theta}) = G_B(\boldsymbol{\theta}) = 0.196$ and, fixing $\alpha = 1.5$, $GE(\boldsymbol{\theta}; 1.5) = GE_B(\boldsymbol{\theta}; 1.5) = 0.069$. For each value of $\bar{\theta}$, and for each s and k, $M = 50,000$ death counts $\{y_{sk}^m\}_{m=1,...,M}$ are generated from the model:

$$y_{sk}|\bar{\theta}_k \sim Poisson(\bar{\theta}_k P_{sk})\ , \qquad k = 1, ..., K, \quad s = 1, ..., S_k\ ,$$

where P_{sk} is set at the number of births observed in the Italian provinces in 2010, in order to obtain simulation results relevant for highlighting the peculiarities of our case-study. For each simulated count y_{sk}^m, direct estimates $\hat{\theta}_{sk}^m = y_{sk}^m / P_{sk}$ are used to obtain plug-in estimates of the inequality measures and their components. Averaging over all simulated values, we obtain the expected value of the plug-in estimators, as an example:

$$E(\widehat{G}(\boldsymbol{\theta})|\boldsymbol{\theta}) = \frac{1}{M} \sum_{i=m}^{M} \widehat{G}^m(\boldsymbol{\theta})\ .$$

Simulation results are reported in Figures 5 and 6, with $\bar{\theta}$ values in abscissa. In all panels, true population values are reported as horizontal thin lines. The left panel of Figure 5 and the upper left panel of Figure 6 show that estimators $\widehat{G}(\boldsymbol{\theta})$ and $\widehat{GE}(\boldsymbol{\theta}; \alpha)$ of the global inequality are positively biased while both estimators of the between components, whose expected value is reported as a dashed line in the same panels, are approximately unbiased. Unbiasedness of the between-component estimators is not surprising and can be ascribed to the stability of the group-specific sample means $\hat{\bar{\theta}}_k$ as estimators of the population parameters $\bar{\theta}_k$, based on greater population sizes with respect to area-level estimates $\hat{\theta}_{sk}$. It turns out that the bias of the global estimators is essentially due to overestimation of the within component, as can be seen from the central panels of the figures. Moreover, the bias of the global measures decreases when $\bar{\theta}$ increases: the relative bias of $\widehat{GE}(\boldsymbol{\theta}; 1.5)$ ranges from 63% (when $\bar{\theta} = .002$) to 14% (when $\bar{\theta} = .009$), while the relative bias of $\widehat{G}(\boldsymbol{\theta})$ ranges from 34% (when $\bar{\theta} = .002$) to 13% (when $\bar{\theta} = .009$).

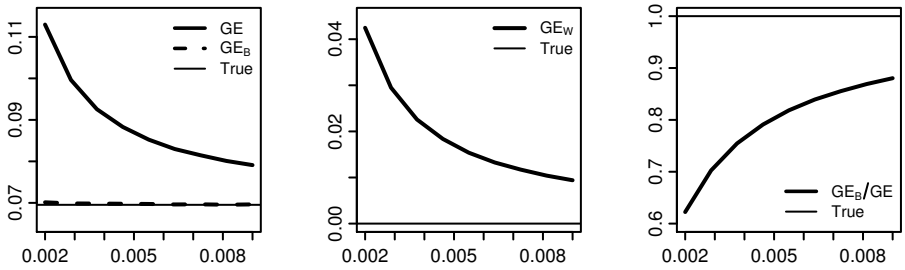

Figure 5: Expected values of the components of the Generalised Entropy index (left and central panels). Expected values of the contribution of the between component to the total (right panel). In abscissa $\bar{\theta}$ values.

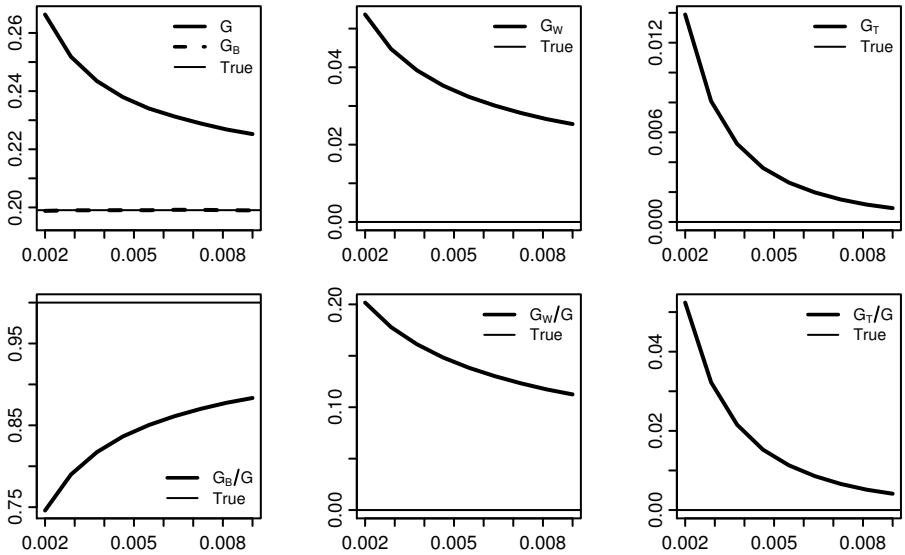

Figure 6: Expected values of the components of the Gini index (upper panels). Expected values of the contribution of each component to the total (lower panels). In abscissa $\bar{\theta}$ values.

This is a very relevant feature to bear in mind in our case study: since the average mortality intensity decreases along the study period (see Figure 1), it is very likely that overestimation of the within component is more severe at the end of the study period itself. In other words, inequality measures computed at the beginning and at the end of the period (reported in Figures 3 and 4) are not directly comparable since they are affected in a different way by sampling variability. An interesting feature of the Dagum's decomposition of the Gini index is its ability to capture (and to be affected by) transvariation: for low $\bar{\theta}$ values, simulated rates $\widehat{\theta}^m_{sk}$ are more likely overlapping between groups than for high $\bar{\theta}$ values: this intuitive behaviour induces the trend of $E(\widehat{G}_T(\boldsymbol{\theta})|\boldsymbol{\theta})$ plotted in the right panels of Figure 6. As a consequence of the overestimation of within

variability, the contribution of the between components to the global inequality turns out to be heavily underestimated (see Figure 5, right panel and Figure 6, lower left panel): $E(\widehat{GE}_B(\boldsymbol{\theta}; 1.5)/\widehat{GE}(\boldsymbol{\theta}; 1.5)|\boldsymbol{\theta})$ ranges from 0.62 to 0.88 as a function of $\bar{\theta}$, while $E(\widehat{G}_B(\boldsymbol{\theta})/\widehat{G}(\boldsymbol{\theta})|\boldsymbol{\theta})$ ranges from 0.75 to 0.89.

The dangers of a quick exploitation of direct estimates have been highlighted by the simulation study just described. A model-based approach where small area estimates are improved by a borrowing strength process is therefore needed. The Bayesian framework is particularly suitable to this aim and to easily obtain uncertainty measures concerning inequality decomposition.

4. SPATIO-TEMPORAL SMOOTHING

Spatio-temporal disease mapping models can be adopted as useful tools for attenuating the effects of sampling variability of individual-level estimates on inequality measures and their decomposition. Several spatio-temporal disease mapping models have been proposed including parametric or non-parametric time trend and different types of spatio-temporal interaction (see *e.g.* Blangiardo *et al.*, 2013; Schrödle and Held, 2011; Ugarte *et al.*, 2014). In this work, we adopt the well-known smoothing model proposed in Knorr-Held (2000), that is briefly sketched in what follows. Since our aim is limited to obtain smoothed mortality rates, we do not include group-specific parameters: between-group variation will be evaluated on the basis of the posterior distribution of the smoothed rates. According to the approach proposed in Knorr-Held (2000), the spatio-temporal trend is non-parametrically modelled: this delivers a very flexible model that can capture complex non-linear behaviours. Smoothing is achieved by borrowing strength along both space and time under the fairly reasonable hypothesis that rates variation is smooth along these dimensions. The model is hierarchically specified and is particularly suitable to be managed in a Bayesian framework. At the first level of the hierarchy, conditionally on model parameters involved in higher levels, mortality counts y_{skt} are assumed to follow independent Poisson distributions:

$$y_{skt}|\theta_{skt} \sim Poisson(\theta_{skt}P_{skt}), \qquad s = 1, ..., S, \;\; k = 1, ..., K, \;\; t = 1, ..., T \; . \tag{4.1}$$

In its most general formulation, the model includes both spatial and temporal structured and unstructured random effects and a spatio-temporal interaction term. All random effects are modelled as Gaussian Markov Random Fields (GMRF): the Markov property of GMRF models implies sparseness of the precision matrix, which allows fast computations. The linear predictor is specified as:

$$log(\theta_{skt}) = \mu + \phi_t + \nu_t + \psi_{sk} + u_{sk} + \delta_{skt}, \tag{4.2}$$

where μ captures the average log-rate; $\boldsymbol{\nu} = (\nu_1, ..., \nu_t, ..., \nu_T)$ and $\boldsymbol{u} = (u_{11}, ..., u_{S_1 1}, ..., u_{1K}, ..., u_{S_K K})$ are unstructured temporal and spatial random effects distributed as independent zero-mean Gaussian variables, *i.e.* $\boldsymbol{u} \sim N(\mathbf{0}, \tau_u \mathbf{I}_S)$ and $\boldsymbol{\nu} \sim N(\mathbf{0}, \tau_\nu \mathbf{I}_T)$. Intrinsic GMRF (IGMRF) are adopted for random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_t, ..., \phi_T)$ and $\boldsymbol{\psi} = (\psi_{11}, ..., \psi_{S_1 1}, ..., \psi_{1K}, ..., \psi_{S_K K})$, namely $\boldsymbol{\phi} \sim N(\mathbf{0}, \tau_\phi \mathbf{K}_T(\phi))$ and $\boldsymbol{\psi} \sim N(\mathbf{0}, \tau_\psi \mathbf{K}_S(\psi))$, where $\mathbf{K}_T(\phi))$ is structured in order to obtain a Random Walk 1 prior and $\mathbf{K}_S(\psi))$ depends on the neighbouring structure of the map, delivering the well-known Intrinsic Conditional AutoRegressive (ICAR) model. With regard to the spatio-temporal interaction random effects $\boldsymbol{\delta} = (\delta_{111}, ..., \delta_{skt}, ..., \delta_{SKT})$, four types of interaction can be postulated by specifying the structure matrix as the Kronecker product of the corresponding structure matrices of the main effects. Namely, $\boldsymbol{\delta} \sim N(\mathbf{0}, \tau_\delta \mathbf{K}_{ST}(\delta))$ where:

- Type I interaction: $\mathbf{K}_{ST}(\delta) = \mathbf{I}_T \otimes \mathbf{I}_S$;
- Type II interaction: $\mathbf{K}_{ST}(\delta) = \mathbf{I}_T \otimes \mathbf{K}_S(\psi)$;
- Type III interaction: $\mathbf{K}_{ST}(\delta) = \mathbf{K}_T(\phi) \otimes \mathbf{I}_S$;
- Type IV interaction: $\mathbf{K}_{ST}(\delta) = \mathbf{K}_T(\phi) \otimes \mathbf{K}_S(\psi)$.

To ensure model identifiability, appropriate linear constraints are needed for the random effects: with regard to IGMRFs, the number of required linear constraints equals the rank-deficiency of the precision matrix. As pointed out in Schrödle and Held (2011), identifiability can be ensured by computing the null space of the structure matrices and using the obtained eigenvectors as linear constraints: this is the strategy we adopt for model estimation. Unstructured random effects are constrained to zero sum in order to allow identification of the intercept term μ. Model hierarchy is completed by specifying a diffuse Gaussian distribution as a prior μ, while Gamma priors are specified for precision parameters $\tau_\phi, \tau_u, \tau_\nu, \tau_\psi$ and τ_δ.

4.1. Computations

Coherently with the Bayesian framework, we aim at evaluating and decomposing inequality measures (3.1) and (3.3) on the basis of their posterior distribution $p(G(\boldsymbol{\theta}_t)|\mathbf{y})$ and $p(GE(\boldsymbol{\theta}_t; \alpha)|\mathbf{y})$: this allows to easily obtain both point estimates and their associated uncertainty. When dealing with complex hierarchical Bayesian models, the joint posterior distribution is not available in closed form and needs to be approximated. Two alternative strategies are currently very popular for approximating the joint posterior distribution: Markov Chain Monte Carlo (MCMC) sampling and INLA (see Rue and Martino, 2009). The latter is particularly suitable for latent GMRF models and provides very accurate approximations of the posterior distribution. Moreover, INLA outper-

forms MCMC approaches in terms of computational time and accuracy. INLA has been made easily implementable by the R package INLA (Rue *et al.*, 2013), that we used for model estimation.

It is worth noting that inequality measures are non-linear combinations of model parameters: the R package INLA allows to approximate the posterior distribution of linear combinations of the model parameters, but does not allow to obtain approximations of non-linear combinations: as a consequence, posterior inference concerning inequality measures can only be performed by sampling from the *joint* posterior distribution, a task that is naturally addressed in an MCMC framework. Fortunately, the adoption of an MCMC algorithm can be avoided in our case study, since an experimental function implemented in the INLA package, `inla.posterior.sample`, allows to draw samples from the joint posterior distribution. We checked the coherence between the results obtained by the INLA experimental function and the posterior samples obtained by means of an MCMC algorithm, finding agreement between results for the estimated models, with an impressively lower computational time demanded by the INLA-based procedure. Once posterior samples from the joint posterior distribution are available, as is the case where MCMC sampling is performed, posterior distributions of any functions of the model parameters can be obtained on the basis of these samples. Given L samples $\{\boldsymbol{\theta}_t^l\}_{l=1,...,L}$ from the joint posterior distribution $\boldsymbol{\theta}_t|\mathbf{y}$, $l = 1, ..., l, ..., L$, $t = 1, ..., t, ..., T$, for each l, inequality measures and their decompositions can be computed, delivering an L-dimensional sample from their posterior distribution: as a byproduct, both posterior point estimates and credibility intervals can be easily obtained.

5. RESULTS

Model selection is performed by means of the Deviance Information Criterion (DIC, Spiegelhalter el at., 2002) according to the results of Table 1. Models without unstructured terms are preferred in terms of fitting, as the first column of results shows; the selected model includes a Type II interaction term.

Table 1: Model comparison: Deviance Information Criterion.

Interaction	**Without ν and u**	**With ν and u**
Type I	11185.09	11224.34
Type II	**11136.75**	11196.59
Type III	11245.67	11356.31
Type IV	11236.46	11286.92

On the basis of the selected model, we obtain posterior estimates of inequality measures and their decomposition of Figures 7 and 8, where posterior means are reported along with 90% credibility intervals.

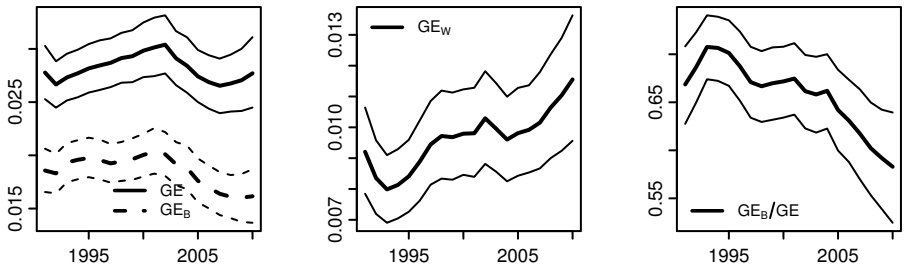

Figure 7: Posterior means and credibility intervals. Generalised Entropy and Between Generalised Entropy (left panel). Within Generalised Entropy (middle panel). Effect due to the between component (right panel).

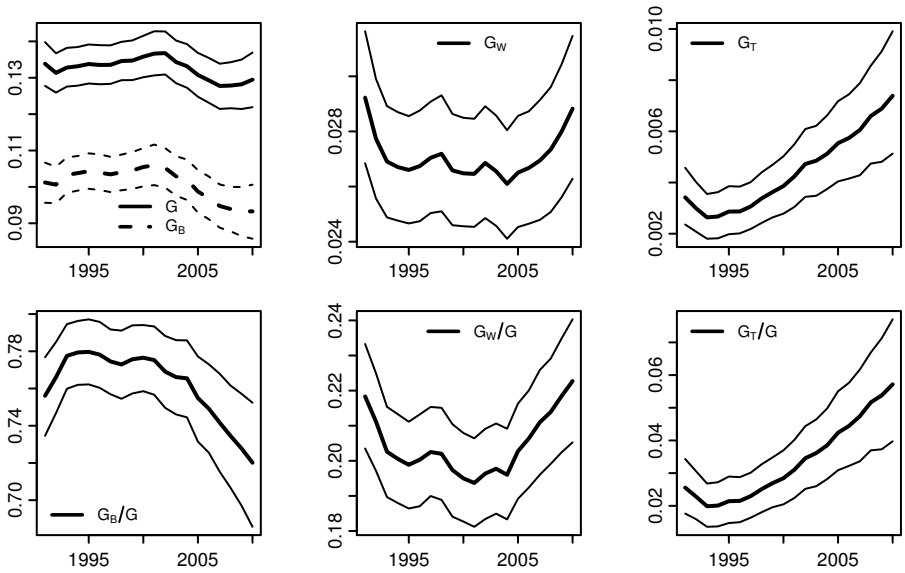

Figure 8: Posterior means and credibility intervals. Gini index and its between component (upper left panel). Within component (upper middle panel). Transvariation component (upper right panel). Lower panels report the contribution of each component to the total.

These figures should be compared with their counterparts based on direct estimates, already reported in Figures 3 and 4; all the comparisons discussed in what follows are coherent with the results of the simulation study reported in Figures 5 and 6, and should be interpreted in light of them. As a first difference,

smoothing of direct estimates turns out in smoothing of the temporal trend of inequality measures, delivering a clearer picture of the evolution of territorial disparities in Italian infant mortality. Secondly, posterior estimates of the overall level of inequality (see Figure 7 left panel and Figure 8 upper left panel) are sensibly lower than estimates obtained by means of plug-in estimators. The ratio $E(GE(\boldsymbol{\theta}_t)|\mathbf{y})/\widehat{GE}(\boldsymbol{\theta}_t)$, that can be interpreted as a quick measure of the effect due to the shrinkage of mortality rates estimates, ranges from 0.9 at the beginning of the study period, when mortality intensity is higher, to 0.45 at the end of the study period, characterised by lower mortality intensity. The same ratio referred to the Gini index ranges from about 0.9 at the beginning of the study period to about 0.7 in last years, witnessing a lower sensitivity of the Gini index to sampling variability of direct estimates. The difference between plug-in estimates and model-based posterior estimates is almost entirely due to the reduction of the within components (central panels of Figures 7 and 8) and, with regard to the Gini decomposition, to the reduction of the component measuring transvariation (Figure 8, right panel). Estimates of the between components remain basically unchanged for both indicators: as a result, the contribution of the between group variability is higher when considering model-based estimates. Despite some evidence of a decreasing trend, inequality between macro-regions explains a considerable share of global inequality: according to the Gini decomposition, which better captured the between-group component in the simulation study, such share ranges from 76% in 1991 to 72% in 2010. The same shares are reduced respectively to 66% and 57% when considering the Generalised Entropy decomposition.

6. CONCLUSIONS

In this paper, we studied the time trend of health disparity in Italy adopting a small area geographical scale. The analysis ranged over a number of methodological and empirical issues that emerge when combining traditional inequality indices, methods for their decomposition and Bayesian hierarchical models. We assumed provinces as units of analysis by grouping them in three main macro-regions: Northern, Central and Southern areas of Italy.

In order to evaluate the temporal evolution of health inequality in Italy, we focused on IMRs, since they are commonly considered as good proxies of health, environmental and socio-economic conditions. After defining, for the sake of brevity, health inequality as the uneven distribution of health across all units of a population, we took into account two popular classes of inequality indicators such as the Generalized Entropy and the Gini coefficient. We also measured the share of global inequality due to disparities among macro-regions, decomposing the total index in its basic components related to the within- and between-group inequality.

However, we preliminary showed that, when dealing with small area data, inequality measures based on direct IMRs tend to be severely affected by random fluctuations. In order to reduce the effect of sample variability and smooth direct IMRs, we estimated a Bayesian model that takes into account spatial, temporal and spatio-temporal interaction effects. Bayesian inference was carried out by means of INLA. Inequality measures based on posterior estimates come out to be less affected by random variations. The model-based Generalized Entropy and Gini coefficient appear stable over the study period, revealing a persistent inequality in infant mortality. In addition, it also comes out that the proportion of global inequality due to disparities among macro-regions tends to be higher when model based estimates are taken into account. We concluded that the persistent health disparity at provincial level is not due to small areas random variability, but is more evidently connected to relevant differences among macro-regions.

Since neonatal care given to mothers and newborns represents one of the main infant mortality causes (Scioscia *et al.*, 2007; Parazzini *et al.*, 1992), it is possible to ascribe the observed infant mortality disparity to different levels of health services (Bonati and Campi, 2005; Mazzucco *et al.*, 2011). In these terms, the persistent disparity in infant mortality between provinces may reflect the long-term socioeconomic inequalities between Northern and Southern Italy (Golini, 2014).

ACKNOWLEDGMENTS

We thank Andrea Riebler for her very useful comments and hints on the use of INLA. The research activity of Fedele Greco was funded by a FIRB 2012 grant (project no. RBFR12URQJ; title: Statistical modeling of environmental phenomena: pollution, meteorology, health and their interactions) for research projects by the Italian Ministry of Education, Universities and Research.

REFERENCES

[1] Atkinson, A.B. (1970). On the measurement of inequality, *Journal of Economic Theory*, **2**, 244–263.

[2] Bonati, M. and Campi, R. (2005). What can we do to improve child health in Southern Italy?, *PLOS Medicine*, **2**, E250.

[3] Blangiardo, M.; Cameletti M.; Baio, G. and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA, *Spatial and Spatio-Temporal Epidemiology*, **7**, 39–35.

[4] Congdon, P. (2010). Modelling trends and inequality in small area mortality, *Journal of Applied Statistics*, **31**, 6, 603–622.

[5] Congdon, P.; Campos, R.M.; Curtis, S.E.; Southall, H.R.; Gregory, I.N. and Jones, I.R. (2001). Quantifying and explaining changes in geographical inequality of infant mortality in England and Wales since the 1890s, *International Journal of Population Geography*, **7**, 35–31.

[6] Costa, M. (2009). Transvariation and inequality between subpopulations in the Dagum's Gini index decomposition, *Metron*, **3**, 229–241.

[7] Dagum, C. (1997). A new approach to the decomposition of the Gini income inequality ratio, *Empirical Economics*, **22**, 515–531.

[8] Dalton, H. (1920). Measurement of the inequality of incomes, *Economic Journal*, **30**, 9, 348–361.

[9] Dreher, A. and Gaston, N. (2008). Has globalization increased inequality?, *Review of International Economics*, **16**, 3, 516–536.

[10] Fantini, M.P.; Stivanello, E.; Dallolio, L.; Loghi, M. and Savoia, E. (2005). Persistent geographical disparities in infant mortality rates in Italy (1999-2001): comparison with France, England, Germany, and Portugal, *European Journal of Health*, **16**, 4, 429–432.

[11] Gakidou, E. and Gary, K. (2002). Measuring total health inequality: adding individual variation to group-level differences, *International Journal for Equity and Health*, **1**:3.

[12] Golini, A. (2014). *Demographic crisis and economic crisis for Italian Mezzogiorno: an iceberg detached from the continent?*. President's Invited Talk at the 47th Scientific Meeting of the Italian Statistical Society, Cagliari, June, 11–13.

[13] Kim, H. and Lim., H. (2010). Comparison of Bayesian spatio-temporal models for chronic diseases, *Journal of Data Science*, **8**, 2, 189–211.

[14] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in Medicine*, **19**, 2555–2567.

[15] Langel, M. and Tillé, Y. (2013). Variance estimation of the Gini index: revisiting a result several times published, *Journal of the Royal Statistical Society, Series A*, **176**, 2, 521–540.

[16] Lauria, L. and De Stavola, B.L. (2003). A district-based analyses of stillbirth and infant mortality rates in Italy: 1989-93, *Paediatric Perinatal Epidemiology*, **17**, 1, 22–23.

[17] Mackenbach, J.P. (2012). The persistence of health inequalities in modern welfare states: the explanation of a paradox, *Social Science & Medicine*, **75**, 4, 761–769.

[18] Mazzucco, W.; Cusimano, R.; Macaluso, M.; La Scola, C.; Fiumanó, G.; Scondotto, S.; Cernigliaro, A.; Corsello, G.; La Torre, G. and Vitale, F. (2011). A retrospective follow up study on maternal age and infant mortality in two Sicilian districts, *BMC Public Health*, **11**:817.

[19] Parazzini, F.; Pirotta, N.; La Vecchia, C.; Bocciolone, L. and Fedele, L. (1992). Determinants of perinatal and infant mortality in Italy, *Revue D'Epidemiologie et de Santé Publique*, **40**, 1, 15–24.

[20] Pradhan, M.; Sahn D.E. and Younger, S.D. (2003). Decomposing world health inequality, *Journal of Health Economics*, **40**, 1, 15–24.

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

[22] Rue, H.; Martino, S.; Lindgren, F.; Simpson, D. and Riebler, A. (2013). INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximation, *R Package Version* 001383402327.

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

[24] Scioscia, M.; Vimercati, A.; Maiorano, A.; Depalo, R. and Selvaggi, L. (2007). A critical analysis on Italian perinatal mortality in a 50-year span, *European Journal of Obstetrics & Gynecology and Reproductive Biology*, **130**, 1, 60–65.

[25] 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, 507–30.

[26] Wagstaff, A.; van Doorslaer, E. and Paci, P. (1991). Horizontal equity in the delivery of health care, *Journal of Health Economics*, **10**, 2, 251–256.

REVSTAT – Statistical Journal
Volume 13, Number 1, March 2015, 21–40

ON PREDICTING CANCER MORTALITY USING ANOVA-TYPE P-SPLINE MODELS

Authors: Jaione Etxeberria
– Department of Statistics and O. R., Public University of Navarre, Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Spain
jaione.etxeberria@unavarra.es

María Dolores Ugarte
– Department of Statistics and O. R., Public University of Navarre, Inst. for Advanced Materials (InaMAT), Public Univ. of Navarre, Spain
lola@unavarra.es

Tomás Goicoa
– Department of Statistics and O. R., Public University of Navarre, Research Network on Health Services in Chronic Diseases (REDISSEC), Inst. for Advanced Materials (InaMAT), Public Univ. of Navarre, Spain
tomas.goicoa@unavarra.es

Ana F. Militino
– Department of Statistics and O. R., Public University of Navarre, Inst. for Advanced Materials (InaMAT), Public Univ. of Navarre, Spain
militino@unavarra.es

Received: October 2014 Revised: November 2014 Accepted: December 2014

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

$$C_{st}|r_{st} \sim Poisson(\mu_{st} = e_{st} r_{st}), \qquad \log \mu_{st} = \log e_{st} + \log r_{st}. \tag{2.1}$$

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

$$u^*_{st} = \log r^*_{st} = \delta + f_s(x_1, x_2) + f_t(t^*) + f_{st}(x_1, x_2, t^*) = \mathbf{B}^*\boldsymbol{\theta}^* \,. \tag{2.2}$$

The term δ 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 ith 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 $\boldsymbol{\theta}^*$ is a vector of coefficients. The matrix $\mathbf{B}^*$ is explicitly defined as

$$\mathbf{B}^* = [\mathbf{1}^*_{st} : \mathbf{1}^*_t \otimes \mathbf{B}_s : \mathbf{B}^*_t \otimes \mathbf{1}_s : \mathbf{B}^*_t \otimes \mathbf{B}_s] \,, \tag{2.3}$$

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. $\mathbf{B}_s = \mathbf{B}_{s_2} \square \mathbf{B}_{s_1}$ is the spatial B-spline basis defined by the row-wise ($\square$) 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

$$\mathbf{B}^*_t = \begin{pmatrix} \mathbf{B}_t & \mathbf{0} \\ \mathbf{B}_{t_1} & \mathbf{B}_{t_2} \end{pmatrix}. \tag{2.4}$$

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 $\boldsymbol{\theta}^*$. 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, $\mathbf{P}^* = \text{diag}\,(\mathbf{0}, \mathbf{P}_s, \mathbf{P}^*_t, \mathbf{P}^*_{st})$, where

$$\begin{aligned} \mathbf{P}_s &= \lambda_{s_1}\mathbf{I}_{m_2} \otimes \mathbf{P}_{s_1} + \lambda_{s_2}\mathbf{P}_{s_2} \otimes \mathbf{I}_{m_1} \,, \\ \mathbf{P}^*_t &= \lambda_t \mathbf{P}_{t^*} \,, \\ \mathbf{P}^*_{st} &= \tau_{s_1}\mathbf{I}^*_{m_3} \otimes \mathbf{I}_{m_2} \otimes \mathbf{P}_{s_1} + \tau_{s_2}\mathbf{I}^*_{m_3} \otimes \mathbf{P}_{s_2} \otimes \mathbf{I}_{m_1} + \tau_t \mathbf{P}_{t^*} \otimes \mathbf{I}_{m_2} \otimes \mathbf{I}_{m_1} \,. \end{aligned} \tag{2.5}$$

In these expressions, $\mathbf{I}_{m_j}$, $j = 1, 2, 3$ are identity matrices of dimension $m_j \times m_j$, where m_j is the number of columns of $\mathbf{B}_j, j = s_1, s_2, t$, and $\mathbf{I}^*_{m_3} = \begin{pmatrix} \mathbf{I}_{m_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. $\mathbf{P}_{s_1}$ and $\mathbf{P}_{s_2}$ are penalty matrices for longitude and latitude respectively defined by $\mathbf{P}_{s_j} = \mathbf{D}'_{s_j}\mathbf{D}_{s_j}$, $j = 1, 2$ where $\mathbf{D}_{s_j}$ are second order difference matrices to achieve smoothness over adjacent marginal coefficients (see Eilers and Marx, 1996). Matrix $\mathbf{P}_{t^*}$ is defined using the extended difference matrix $\mathbf{D}^*_t$ for the time component given by the next expression

$$\mathbf{D}^*_t = \begin{pmatrix} \mathbf{D}_t & \mathbf{0} \\ \mathbf{E}_t & \mathbf{L}_t \end{pmatrix}, \qquad \mathbf{P}_{t^*} = \mathbf{D}^{*\prime}_t\mathbf{D}^*_t = \begin{pmatrix} \mathbf{P}_t + \mathbf{E}'_t\mathbf{E}_t & \mathbf{E}'_t\mathbf{L}_t \\ \mathbf{L}'_t\mathbf{E}_t & \mathbf{L}'_t\mathbf{L}_t \end{pmatrix}. \tag{2.6}$$

$\mathbf{D}_t$ and $\mathbf{P}_t$ are the difference matrix and the penalty matrix for the observed time period, $\mathbf{E}_t$ and $\mathbf{L}_t$ are the rows used to obtain the penalty for the oncoming years, and λ_{s_1}, λ_{s_2}, λ_t, τ_{s_1}, τ_{s_2} and τ_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 $\mathbf{T}^*$ is used to transform $\mathbf{B}^*$ into $[\mathbf{X}^* : \mathbf{Z}^*]$ and $\boldsymbol{\theta}^*$ into $(\boldsymbol{\beta}', \boldsymbol{\alpha}^{*'})'$. In this paper we provide the definition of this transformation matrix $\mathbf{T}^*$ which is based on matrices of eigenvectors corresponding to non-zero and zero eigenvalues respectively obtained from the eigen decomposition of the matrices $\mathbf{P}_i$, $i = s_1, s_2, t$. The key point in this process is to choose an extended transformation matrix preserving the original transformation matrix $\mathbf{T}$ used to fit the data. Based on the transformation matrix $\mathbf{T}$, the following extended transformation matrix is considered

$$\mathbf{T}^* = \begin{pmatrix} 1 & & & \\ & \mathbf{T}_s & & \\ & & \mathbf{T}^*_t & \\ & & & \mathbf{T}^*_{st} \end{pmatrix},$$

where $\mathbf{T}^*_t$ and $\mathbf{T}^*_{st}$ are defined by

$$\mathbf{T}^*_t = \begin{pmatrix} \mathbf{T}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_t^{-1} \end{pmatrix}, \qquad \mathbf{T}^*_{st} = \begin{pmatrix} \mathbf{T}_{st} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_t^{-1} \otimes \mathbf{I}_{m_2} \otimes \mathbf{I}_{m_1} \end{pmatrix},$$

and

$$\begin{aligned} \mathbf{T}_s &= [1 \otimes [\mathbf{u}_{2n} \otimes \mathbf{1}_1 : \mathbf{1}_2 \otimes \mathbf{u}_{1n} : \mathbf{u}_{2n} \otimes \mathbf{u}_{1n}] : \mathbf{R}_s], \\ \mathbf{T}_t &= [\mathbf{u}_{3n} \otimes 1 : \mathbf{R}_t], \\ \mathbf{T}_{st} &= [\mathbf{u}_{3n} \otimes [\mathbf{u}_{2n} \otimes \mathbf{1}_1 : \mathbf{1}_2 \otimes \mathbf{u}_{1n} : \mathbf{u}_{2n} \otimes \mathbf{u}_{1n}] : \mathbf{R}_{st}]. \end{aligned}$$

The matrices $\mathbf{R}_s$, $\mathbf{R}_t$, and $\mathbf{R}_{st}$ are given by

$$\begin{aligned}
\mathbf{R}_s &= [1 \otimes [\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}]] , \\
\mathbf{R}_t &= [\mathbf{U}_{3s} \otimes 1] , \\
\mathbf{R}_{st} &= [\mathbf{u}_{3n} \otimes [\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}] : \\
&\quad \mathbf{U}_{3s} \otimes [\mathbf{1}_2 \otimes \mathbf{u}_{1n} : \mathbf{u}_{2n} \otimes \mathbf{1}_1 : \mathbf{u}_{2n} \otimes \mathbf{u}_{1n} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}]] .
\end{aligned}$$

Note that $\mathbf{T}_s$, $\mathbf{T}_t$ and $\mathbf{T}_{st}$ represent the components of the original transformation matrix corresponding to the observed data. $\mathbf{U}_{in} = [\mathbf{1}_i : \mathbf{u}_{in}]$ and $\mathbf{U}_{is}$, $i = 1, 2, 3$, are matrices of eigenvectors corresponding to zero and non-zero eigenvalues obtained from the eigen-decomposition of the penalty matrix $\mathbf{P}_j$, $j = s_1, s_2, t$. 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

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

and the extended model is expressed as

$$\begin{aligned}
\begin{pmatrix} \mathbf{u}^o \\ \mathbf{u}^p \end{pmatrix} &= \delta + \begin{bmatrix} \mathbf{X}_s & \mathbf{Z}_s \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{\alpha}_s \end{pmatrix} + \begin{bmatrix} \mathbf{X}^o_t & \mathbf{Z}^o_t & \mathbf{0} \\ \mathbf{X}^p_t & \mathbf{Z}^p_{t_1} & \mathbf{Z}^p_{t_2} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{\alpha}_t \\ \boldsymbol{\alpha}^p_t \end{pmatrix} \\
&\quad + \begin{bmatrix} \mathbf{X}^o_{st} & \mathbf{Z}^o_{st} & \mathbf{0} \\ \mathbf{X}^p_{st} & \mathbf{Z}^p_{st_1} & \mathbf{Z}^p_{st_2} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_{st} \\ \boldsymbol{\alpha}_{st} \\ \boldsymbol{\alpha}^p_{st} \end{pmatrix} ,
\end{aligned} \tag{2.7}$$

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; $\boldsymbol{\beta}_s$, $\boldsymbol{\beta}_t$, $\boldsymbol{\beta}_{st}$ are the fixed effects; $\boldsymbol{\alpha}_s$, $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{st}$ are the random effects for space, time and space-time interaction respectively, corresponding to the observed data, and $\boldsymbol{\alpha}^p_t$ and $\boldsymbol{\alpha}^p_{st}$ 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 $\mathbf{C} = \text{diag}\,(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$ where

$$\mathbf{C}_1 = \text{Cov}(\boldsymbol{\alpha}_s) = \mathbf{R}'_s\mathbf{P}_s\mathbf{R}_s = \begin{pmatrix} \mathbf{F}_1^{-1} & & \\ & \mathbf{F}_2^{-1} & \\ & & \mathbf{F}_3^{-1} \end{pmatrix} ,$$

$$\mathbf{C}_2 = \mathrm{Cov}(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_t^p) = \mathbf{R}_t'^{*}\mathbf{P}_t^{*}\mathbf{R}_t^{*} = \begin{pmatrix} \mathbf{F}_4^{-1} & -\mathbf{F}_4^{-1}\mathbf{R}_t'\mathbf{E}_t' \\ -\mathbf{E}_t\mathbf{R}_t\mathbf{F}_4^{-1} & \mathbf{I}_r + \mathbf{I}_r\mathbf{E}_t\mathbf{R}_t\mathbf{F}_4^{-1}\mathbf{R}_t'\mathbf{E}_t'\mathbf{I}_r \end{pmatrix},$$

$$\begin{aligned}\mathbf{C}_3 &= \mathrm{Cov}(\boldsymbol{\alpha}_{st}, \boldsymbol{\alpha}_{st}^p) = \mathbf{R}_{st}'^{*}\mathbf{P}_{st}^{*}\mathbf{R}_{st}^{*} \\ &= \begin{pmatrix} \mathbf{F}^{-1} & -\mathbf{F}^{-1}\mathbf{R}_{st}'(\mathbf{E}_t' \otimes \mathbf{I}_s) \\ -(\mathbf{E}_t \otimes \mathbf{I}_s)\mathbf{R}_{st}\mathbf{F}^{-1} & \mathbf{I}^{*} + \mathbf{I}^{*}(\mathbf{E}_t \otimes \mathbf{I}_s)\mathbf{R}_{st}\mathbf{F}^{-1}\mathbf{R}_{st}'(\mathbf{E}_t' \otimes \mathbf{I}_s)\mathbf{I}^{*} \end{pmatrix},\end{aligned}$$

Here, $\mathbf{R}_t^{*}$ and $\mathbf{R}_{st}^{*}$ are part of the transformation matrix $\mathbf{T}^{*}$ and they are given by

$$\mathbf{R}_t^{*} = \begin{pmatrix} \mathbf{R}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_t^{-1} \end{pmatrix}, \qquad \mathbf{R}_{st}^{*} = \begin{pmatrix} \mathbf{R}_{st} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_t^{-1} \otimes \mathbf{I}_{m_2} \otimes \mathbf{I}_{m_1} \end{pmatrix}.$$

Expressions for $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$ and $\mathbf{F}_4$, and $\mathbf{F}$ 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 $\boldsymbol{\alpha}_t^p$ and $\boldsymbol{\alpha}_{st}^p$ are given by

$$\begin{aligned}\widehat{\boldsymbol{\alpha}_t^p} &= -\mathbf{E}_t\mathbf{R}_t\mathbf{F}_4^{-1}\mathbf{F}_4\hat{\boldsymbol{\alpha}}_t = -\mathbf{E}_t\mathbf{R}_t\widehat{\boldsymbol{\alpha}_t}\,, \\ \widehat{\boldsymbol{\alpha}_{st}^p} &= -(\mathbf{E}_t \otimes \mathbf{I}_s)\mathbf{R}_{st}\mathbf{F}^{-1}\mathbf{F}\hat{\boldsymbol{\alpha}}_{st} = -(\mathbf{E}_t \otimes \mathbf{I}_s)\mathbf{R}_{st}\widehat{\boldsymbol{\alpha}_{st}}\,.\end{aligned} \tag{2.8}$$

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 $\boldsymbol{\lambda} = (\lambda_{s_1}, \lambda_{s_2}, \lambda_t, \tau_{s_1}, \tau_{s_2}, \tau_t)'$. Finally, using Equation (2.8), the estimated (corresponding to observed values) and the predicted (corresponding to future values) log-relative risks are given by

$$\begin{aligned}\begin{pmatrix} \hat{\mathbf{u}}^o \\ \hat{\mathbf{u}}^p \end{pmatrix} &= \hat{\delta} + \begin{bmatrix} \mathbf{X}_s & \mathbf{Z}_s \end{bmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}}_s \\ \hat{\boldsymbol{\alpha}}_s \end{pmatrix} + \begin{bmatrix} \mathbf{X}_t^o & \mathbf{Z}_t^o \\ \mathbf{X}_t^p & \mathbf{Z}_{t_1}^p - \mathbf{Z}_{t_2}^p\mathbf{E}_t\mathbf{R}_t \end{bmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}}_t \\ \hat{\boldsymbol{\alpha}}_t \end{pmatrix} \\ &\quad + \begin{bmatrix} \mathbf{X}_{st}^o & \mathbf{Z}_{st}^o \\ \mathbf{X}_{st}^p & \mathbf{Z}_{st_1}^p - \mathbf{Z}_{st_2}^p(\mathbf{E}_t \otimes \mathbf{I}_s)\mathbf{R}_{st} \end{bmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{st} \\ \hat{\boldsymbol{\alpha}}_{st} \end{pmatrix}.\end{aligned} \tag{2.9}$$

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.

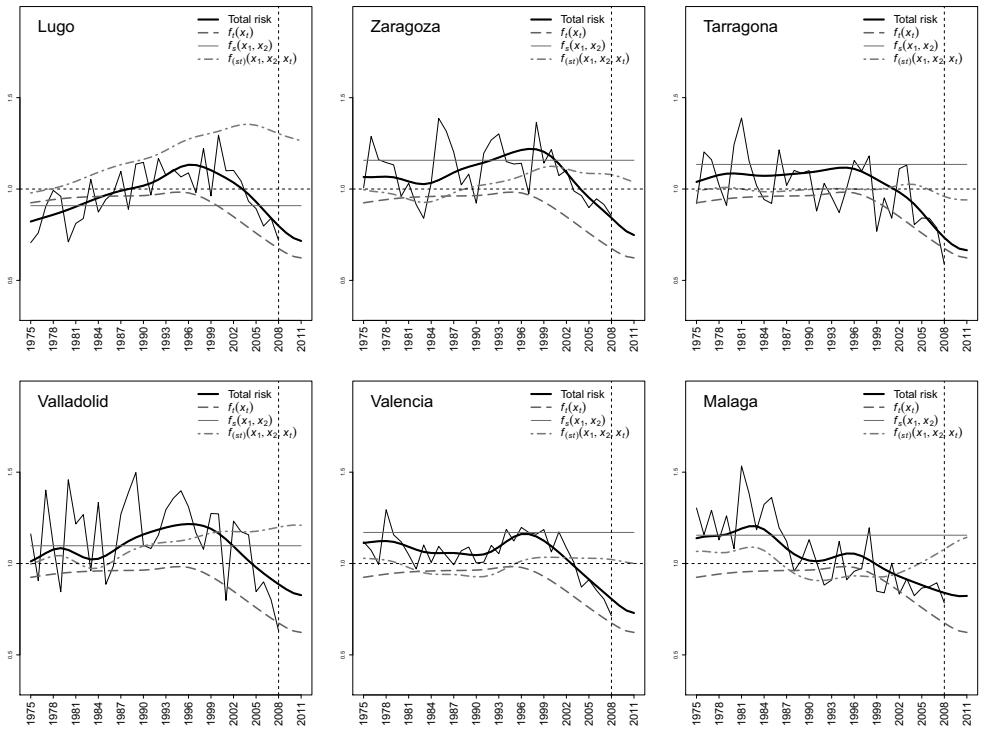

Figure 1: Risks temporal evolution of the different terms of the ANOVA-type P-spline model and predictions for the years 2009, 2010, and 2011. The smooth thick solid line corresponds to the total risk estimates and predictions, the dashed line represents the temporal trend common to all areas, and the 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; counts predictions for 2011 and their corresponding 95% prediction intervals.

Province	**Counts 2008**	**Risks 2011**	**95% C.I. Risks 2011**	**Expected 2011**	**Counts 2011**	**95% C.I. Counts 2011**
Coruña	161.00	0.76	[0.62, 0.93]	224.63	170.39	[123.85, 216.94]
Lleida	71.00	0.72	[0.59, 0.87]	86.24	61.73	[41.04, 82.41]
Ourense	91.00	0.76	[0.63, 0.92]	96.26	72.98	[49.72, 96.24]
Pontevedra	130.00	0.83	[0.68, 1.01]	163.42	135.11	[98.10, 172.12]
Oviedo	196.00	0.82	[0.68, 0.99]	226.20	185.59	[138.76, 232.43]
Santander	69.00	0.83	[0.69, 1.00]	107.40	89.58	[63.54, 115.62]
Lugo	53.00	0.71	[0.60, 0.86]	101.68	72.69	[49.80, 95.58]
Álava	39.00	0.73	[0.61, 0.87]	52.96	38.62	[23.97, 53.27]
Guipúzcoa	112.00	0.70	[0.58, 0.84]	125.16	87.48	[60.92, 114.05]
Vizcaya	162.00	0.78	[0.65, 0.94]	215.64	168.24	[124.91, 211.57]
Navarra	83.00	0.71	[0.60, 0.86]	112.94	80.64	[55.98, 105.29]
Huesca	54.00	0.73	[0.61, 0.88]	55.59	40.82	[25.50, 56.14]
Zaragoza	147.00	0.75	[0.62, 0.90]	183.76	137.45	[100.39, 174.51]
Teruel	31.00	0.65	[0.54, 0.78]	40.76	26.52	[14.72, 38.32]
Burgos	69.00	0.74	[0.62, 0.89]	84.56	62.84	[42.49, 83.18]
Palencia	32.00	0.79	[0.66, 0.95]	41.38	32.81	[19.68, 45.94]
León	95.00	0.76	[0.64, 0.91]	127.74	97.28	[69.36, 125.21]
Zamora	45.00	0.80	[0.66, 0.97]	60.62	48.72	[31.60, 65.84]
Valladolid	57.00	0.83	[0.69, 1.00]	98.57	81.50	[57.12, 105.88]
Soria	21.00	0.67	[0.55, 0.81]	26.60	17.72	[8.50, 26.95]
Salamanca	56.00	0.73	[0.61, 0.89]	88.01	64.57	[43.22, 85.93]
Ávila	29.00	0.70	[0.58, 0.84]	47.07	33.03	[19.59, 46.46]
Segovia	28.00	0.70	[0.59, 0.84]	38.86	27.34	[15.61, 39.06]
Logroño	62.00	0.66	[0.55, 0.79]	61.95	40.85	[25.23, 56.46]
Girona	88.00	0.72	[0.59, 0.87]	117.73	84.59	[58.33, 110.84]
Barcelona	524.00	0.64	[0.54, 0.76]	863.63	553.89	[426.37, 681.41]
Tarragona	70.00	0.66	[0.55, 0.80]	129.64	86.18	[59.70, 112.66]
Castellón	74.00	0.74	[0.62, 0.89]	94.70	69.97	[47.80, 92.14]
Valencia	253.00	0.73	[0.61, 0.87]	382.96	279.33	[213.30, 345.36]
Alicante	181.00	0.74	[0.62, 0.89]	288.08	213.82	[159.92, 267.73]
Murcia	172.00	0.73	[0.61, 0.88]	193.45	141.63	[103.02, 180.25]
Baleares	115.00	0.68	[0.55, 0.84]	145.72	99.38	[67.62 , 131.13]
Madrid	544.00	0.61	[0.52, 0.73]	871.18	533.29	[408.12, 658.47]
Cáceres	63.00	0.73	[0.60, 0.88]	85.99	62.48	[41.43, 83.53]
Badajoz	92.00	0.82	[0.68, 1.00]	119.86	98.79	[69.84, 127.74]
Guadalajara	36.00	0.57	[0.48, 0.69]	43.32	24.84	[13.41, 36.28]
Toledo	106.00	0.65	[0.55, 0.78]	121.00	78.83	[54.31, 103.34]
Cuenca	45.00	0.61	[0.51, 0.73]	54.43	33.15	[19.49, 46.81]
Ciudad Real	71.00	0.71	[0.59, 0.86]	101.42	72.14	[48.95, 95.34]
Albacete	65.00	0.77	[0.64, 0.93]	73.63	56.88	[37.86, 75.91]
Huelva	63.00	0.91	[0.74, 1.13]	71.24	65.15	[43.75, 86.54]
Sevilla	201.00	0.84	[0.69, 1.01]	240.17	201.03	[151.42, 250.63]
Cádiz	114.00	0.85	[0.70, 1.04]	149.34	126.89	[91.66, 162.12]
Córdoba	89.00	0.66	[0.55, 0.80]	130.23	86.48	[59.36, 113.60]
Málaga	147.00	0.82	[0.68, 1.00]	210.21	172.95	[128.03, 217.88]
Jaén	95.00	0.66	[0.55, 0.79]	115.29	75.98	[51.57, 100.39]
Granada	75.00	0.66	[0.54, 0.80]	138.34	91.21	[62.45, 119.97]
Almería	52.00	0.69	[0.56, 0.84]	81.31	55.83	[35.71, 75.95]
Las Palmas	113.00	0.80	[0.64, 1.01]	119.57	96.25	[65.44, 127.05]
Tenerife	110.00	0.75	[0.60, 0.94]	134.76	101.48	[68.43, 134.53]

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 2: A comparison of empirical coverage probabilities in the period 1995–2008 (excluding 2006–2007).

	Three year-ahead predictions		**Two year-ahead predictions**		**One year-ahead predictions**	
	95%	99%	95%	99%	95%	99%
Additive CAR RW2	87.67	95.00	85.00	95.00	86.50	95.50
Interaction CAR RW2 (struc.)	94.33	98.83	93.00	97.50	91.17	96.83
Interaction CAR RW2(unstruc.)	93.83	99.17	92.83	97.50	90.67	97.00
CAR(s)+Pspline(t)	85.83	94.67	87.50	94.00	88.33	95.00
CAR(s)+Pspline(t) + Int	85.50	93.67	89.33	95.50	91.33	97.50
Pspline(t) + Pspline Int	94.83	98.17	93.17	97.67	93.33	98.17
Pure interaction Pspline	93.33	99.17	93.00	98.17	91.00	97.00
ANOVA-type Pspline	95.33	99.00	92.33	97.66	91.83	97.33

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 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 $\mathbf{T}^*$, the fixed and random effect matrices of the extended generalized linear mixed model are given by

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

where

$$\begin{aligned}
(\mathbf{1}^*_t \otimes \mathbf{B}_s)\mathbf{T}_s &= \mathbf{1}^*_t \otimes [\mathbf{x}_s : [\mathbf{Z}_2 \square \mathbf{X}_1 : \mathbf{X}_2 \square \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1]] = \mathbf{1}^*_t \otimes [\mathbf{x}_s : \mathbf{Z}_a] \\
&= [(\mathbf{1}^*_t \otimes \mathbf{x}_s) : (\mathbf{1}^*_t \otimes \mathbf{Z}_a)] = [\mathbf{X}_s : \mathbf{Z}_s], \\
(\mathbf{B}^*_t \otimes \mathbf{1}_s)\mathbf{T}^*_t &= \begin{bmatrix} \mathbf{B}_t\mathbf{u}_{3n} & \mathbf{B}_t\mathbf{U}_{3s} & \mathbf{0} \\ \mathbf{B}_{t_1}\mathbf{u}_{3n} & \mathbf{B}_{t_1}\mathbf{U}_{3s} & \mathbf{B}_{t_2}\mathbf{L}_t^{-1} \end{bmatrix} \otimes \mathbf{1}_s = \begin{bmatrix} \mathbf{x}^o_t \otimes \mathbf{1}_s & \mathbf{Z}^o_t & \mathbf{0} \\ \mathbf{x}^p_t \otimes \mathbf{1}_s & \mathbf{Z}^p_{t_1} & \mathbf{Z}^p_{t_2} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}^o_t & \mathbf{Z}^o_t & \mathbf{0} \\ \mathbf{X}^p_t & \mathbf{Z}^{p_1}_t & \mathbf{Z}^{p_2}_t \end{bmatrix}, \\
(\mathbf{B}^*_t \otimes \mathbf{B}_s)\mathbf{T}^*_{st} &= \begin{bmatrix} \mathbf{x}^o_t \otimes \mathbf{x}_s : & [\mathbf{x}^o_t \otimes \mathbf{Z}_a : & \mathbf{B}_t\mathbf{U}_{3s} \otimes \mathbf{x}_s & : \mathbf{B}_t\mathbf{U}_{3s} \otimes \mathbf{Z}_a &] : & \mathbf{0} \\ \mathbf{x}^p_t \otimes \mathbf{x}_s : & [\mathbf{x}^p_t \otimes \mathbf{Z}_a : & \mathbf{B}_{t_1}\mathbf{U}_{3s} \otimes \mathbf{x}_s & : \mathbf{B}_{t_1}\mathbf{U}_{3s} \otimes \mathbf{Z}_a &] : & (\mathbf{B}_{t_2}\mathbf{L}_t^{-1}) \otimes \mathbf{B}_s \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}^o_{st} & \mathbf{Z}^o_{st} & \mathbf{0} \\ \mathbf{X}^p_{st} & \mathbf{Z}^p_{st_1} & \mathbf{Z}^p_{st_2} \end{bmatrix}.
\end{aligned}$$

Here, $\mathbf{Z}_s = (\mathbf{1}^*_t \otimes \mathbf{Z}_a)$, $\mathbf{Z}_a = [\mathbf{Z}_2 \square \mathbf{X}_1 : \mathbf{X}_2 \square \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1]$, $\mathbf{X}_s = (\mathbf{1}^*_t \otimes \mathbf{x}_s)$, $\mathbf{x}_s = [\mathbf{1}_n \square \mathbf{x}_1 : \mathbf{x}_2 \square \mathbf{1}_n : \mathbf{x}_2 \square \mathbf{x}_1]$, $\mathbf{X}_1 = [1 : \mathbf{x}_1]$, $\mathbf{X}_2 = [1 : \mathbf{x}_2]$, $\mathbf{Z}_1 = \mathbf{B}_{s_1}\mathbf{U}_{1s}$, $\mathbf{Z}_2 = \mathbf{B}_{s_2}\mathbf{U}_{2s}$ and $\mathbf{Z}_3 = \mathbf{B}_t\mathbf{U}_{3s}$. Finally, $\mathbf{x}_1$ and $\mathbf{x}_2$ are column vectors of longitude and latitude respectively, and $\mathbf{x}^o_t$ and $\mathbf{x}^p_t$ are column vector of time corresponding to observed and prediction period respectively. Using these results, the extended model is (2.7) is attained.

B. APPENDIX

In this section, and to make the reading easier, expressions for matrices $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, $\mathbf{F}_4$, and $\mathbf{F} = \text{blockdiag}(\mathbf{F}_5, \mathbf{F}_6, \mathbf{F}_7, \mathbf{F}_8, \mathbf{F}_9, \mathbf{F}_{10}, \mathbf{F}_{11})$ 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 $\mathbf{F}_i$, $i = 1, ..., 8$ are exactly the same as those in Ugarte *et al.* (2012b). $\mathbf{F}_9$, $\mathbf{F}_{10}$, $\mathbf{F}_{11}$ are not the same because in this paper we have considered $\mathbf{B}_t^* \otimes \mathbf{B}_s$, the last term in the extended basis (2.3), instead of the other way around $\mathbf{B}_s \otimes \mathbf{B}_t^*$. 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

$$\mathbf{F}_1 = \lambda_2\widetilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_2, \quad \mathbf{F}_2 = \lambda_1\mathbf{I}_2 \otimes \widetilde{\mathbf{\Sigma}}_1, \quad \mathbf{F}_3 = \lambda_1\mathbf{I}_{m_2-2} \otimes \widetilde{\mathbf{\Sigma}}_1 + \lambda_2\widetilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{m_1-2},$$

$$\mathbf{F}_4 = \lambda_t\widetilde{\mathbf{\Sigma}}_3, \quad \mathbf{F}_5 = \tau_2\widetilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_2, \quad \mathbf{F}_6 = \tau_1\mathbf{I}_2 \otimes \widetilde{\mathbf{\Sigma}}_1,$$

$$\mathbf{F}_7 = \tau_1\mathbf{I}_{m_2-2} \otimes \widetilde{\mathbf{\Sigma}}_1 + \tau_2\widetilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{m_1-2}, \quad \mathbf{F}_8 = \tau_t\mathbf{I}_3 \otimes \widetilde{\mathbf{\Sigma}}_3,$$

$$\mathbf{F}_9 = \tau_2\mathbf{I}_{m_3-2} \otimes \widetilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_2 + \tau_t\widetilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{m_2-2} \otimes \mathbf{I}_2,$$

$$\mathbf{F}_{10} = \tau_1\mathbf{I}_{m_3-2} \otimes \mathbf{I}_2 \otimes \widetilde{\mathbf{\Sigma}}_1 + \tau_t\widetilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_2 \otimes \mathbf{I}_{m_1-2},$$

$$\mathbf{F}_{11} = \tau_1\mathbf{I}_{m_3-2} \otimes \mathbf{I}_{m_2-2} \otimes \widetilde{\mathbf{\Sigma}}_1 + \tau_2\mathbf{I}_{m_3-2} \otimes \widetilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{m_1-2} + \tau_t\widetilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{m_2-2} \otimes \mathbf{I}_{m_1-2}.$$

where $\widetilde{\mathbf{\Sigma}}_i$, $i = 1, 2, 3$ are diagonal matrices of non zero eigenvalues coming from the eigen-decomposition of the marginal penalties $\mathbf{P}_{s_1}$, $\mathbf{P}_{s_2}$ and $\mathbf{P}_t$ respectively. $\mathbf{F}^{-1} = blockdiag(\mathbf{F}_5^{-1}, \mathbf{F}_6^{-1}, \mathbf{F}_7^{-1}, \mathbf{F}_8^{-1}, \mathbf{F}_9^{-1}, \mathbf{F}_{10}^{-1}, \mathbf{F}_{11}^{-1})$, $\mathbf{I}_s = \mathbf{I}_{m_2} \otimes \mathbf{I}_{m_1}$, $\mathbf{I}^* = \mathbf{I}_r \otimes \mathbf{I}_s$, and $\mathbf{I}_r$ is the identity matrix of dimension $r \times r$ where r is the number of columns of $\mathbf{L}_t$.

C. APPENDIX

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 $\mathbf{M}$ matrix in the ANOVA-type P-spline model. An estimator for the MSE of the predicted log-risk is given by

$$\widehat{MSE}[\hat{u}^p_{st}] = g^*_{1st}(\hat{\boldsymbol{\lambda}}) + g^*_{2st}(\hat{\boldsymbol{\lambda}}) + 2g^*_{3st}(\hat{\boldsymbol{\lambda}}) \ .$$

where

$$g^*_{1st}(\boldsymbol{\lambda}) \ = \ \mathbf{z}^p_{st}(\mathbf{C} - \mathbf{M}\mathbf{Z}^{o'}\mathbf{V}^{-1}\mathbf{Z}^o\mathbf{M}')\mathbf{z}^{p\,\prime}_{st} \ ,$$

$$g^*_{2st}(\boldsymbol{\lambda}) \ = \ (\mathbf{x}^p_{st} - \mathbf{z}^p_{st}\mathbf{M}\mathbf{Z}^{o'}\mathbf{V}^{-1}\mathbf{X}^o)(\mathbf{X}^{o'}\mathbf{V}^{-1}\mathbf{X}^o)^{-1}(\mathbf{x}^p_{st} - \mathbf{z}^p_{st}\mathbf{M}\mathbf{Z}^{o'}\mathbf{V}^{-1}\mathbf{X}^o)' \ ,$$

$$g^*_{3st}(\boldsymbol{\lambda}) \ = \ \mathrm{tr}[\mathbf{S}^*\mathbf{V}\mathbf{S}^{*\prime}\mathcal{I}^{-1}] \ .$$

Here $\mathbf{V}$ 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 $\mathbf{z}^p_{st}$ and $\mathbf{x}^p_{st}$ are the st row of the matrices $\mathbf{Z}^p = [\mathbf{Z}_s : \mathbf{Z}^p_{t_1} : \mathbf{Z}^p_{t_2} : \mathbf{Z}^p_{st_1} : \mathbf{Z}^p_{st_2}]$ and $\mathbf{X}^p = [\mathbf{X}_s : \mathbf{X}^p_t : \mathbf{X}^p_{st}]$ respectively, and finally, $\mathbf{Z}^o = [\mathbf{Z}_s : \mathbf{Z}_t : \mathbf{Z}_{st}]$ and $\mathbf{X}^o = [\mathbf{X}_s : \mathbf{X}^o_t : \mathbf{X}^o_{st}]$. An explicit expression for $\mathbf{M}$ is given by

$$\mathbf{M} = \begin{pmatrix} \mathbf{C}_1 & & \\ & \begin{pmatrix} \mathbf{F}_4^{-1} \\ -\mathbf{E}_t\mathbf{R}_t\mathbf{F}_4^{-1} \end{pmatrix} & \\ & & \begin{pmatrix} \mathbf{F}^{-1} \\ -(\mathbf{E}_t \otimes \mathbf{I}_s)\mathbf{R}_{st}\mathbf{F}^{-1} \end{pmatrix} \end{pmatrix} .$$

If λ_j denotes the jth entry of the vector of variance components $\boldsymbol{\lambda} = (\lambda_{s_1}, \lambda_{s_2}, \lambda_t, \tau_{s_1}, \tau_{s_2}, \tau_t)'$, the matrix $\mathbf{S}^*$ is given by

$$\mathbf{S}^*_j = \mathbf{z}^*_{st}\left(\frac{\partial \mathbf{M}}{\partial \lambda_j}\mathbf{Z}'\mathbf{V}^{-1} + \mathbf{M}\mathbf{Z}'\frac{\partial \mathbf{V}^{-1}}{\partial \lambda_j}\right), \qquad j = 1, 2, 3, 4, 5, 6 \, .$$

Finally, the variance for predicted counts is calculated as

$$\mathrm{Var}[C^p_{st}] = \mathrm{E}[\mathrm{Var}[C^p_{st}|r^p_{st}]] + \mathrm{Var}[\mathrm{E}[C^p_{st}|r^p_{st}]] = e^p_{st}\mathrm{E}[r^p_{st}] + e^{p2}_{st}\mathrm{Var}[r^p_{st}] \ ,$$

where e^p_{st} are projections of the number of expected cases for future years. $\mathrm{Var}[r^p_{st}]$ is easily estimated from $\widehat{MSE}[\hat{u}^p_{st}]$ 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

[1] Belitz, C. and Lang, S. (2005). Simultaneous selection of variables and smoothing parameters in structured additive regression models, *Computational Statistics and Data Analysis*, **53**, 61–81.

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

[3] Chen, H.S.; Portier, K.; Ghosh, K.; Naishadham, D.; Kim H.J.; Zhu, L.; Pickle, L.W.; Krapcho, M.; Scoppa, S.; Jemal, A. and Feuer, E.J. (2012). Predicting US- and state-level cancer counts for the current calendar year. Part I: Evaluation of temporal projection methods for mortality, *Cancer*, **118**, 1091–1099.

[4] Clements, M.S.; Armstrong, B.K. and Moolgavkar, S.H. (2005). Lung cancer rate predictions using generalized additive models, *Biostatistics*, **6**, 576–589.

[5] Currie, I.D.; Durbán, M. and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates, *Statistical Modelling*, **4**, 279–298.

[6] Dyba, T. and Hakulinen, T. (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques, *Statistics in Medicine*, **19**, 1741–1752.

[7] 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.

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

[9] Escaramís, G.; Carrasco, J.L. and Ascaso, C. (2008). Detection of significant disease risks using a spatial conditional autoregressive model, *Biometrics*, **64**, 1043–1053.

[10] Etxeberria, J.; Goicoa, T.; Ugarte, M.D. and Militino, A.F. (2014). Evaluating space-time models for short-term cancer mortality risk predictions in small areas, *Biometrical Journal*, **56**, 383–402.

[11] Ghosh, K. and Tiwari, R.C. (2007). Prediction of U.S. cancer mortality counts using semiparametric Bayesian techniques, *Journal of the American Statistical Association*, **102**, 7–15.

[12] Ghosh, K.; Tiwari, R.C.; Feuer, E.J.; Cronin, K. and Jemal, A. (2008). *Predicting US cancer mortality using state-space models.* In "Computational Methods in Biomedical Research" (R. Khattree and D.N. Naik, Eds.), Chapman and Hall/CRC, Boca Raton, 131–151.

[13] Gilmour, A.; Cullis, B.; Welham, S.; Gogel, B. and Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models, *Computational Statistics and Data Analysis*, **44**, 571–586.

[14] Goicoa, T.; Ugarte, M.D.; Etxeberria, J. and Militino, A.F. (2012). Comparing CAR and P-spline models in spatial disease mapping, *Environmental and Ecological Statistics*, **19**, 573–599.

[15] Gu, C. (2002). *Smoothing Spline ANOVA Models.* Springer, New York.

[16] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease riskm *Statistics in Medicine*, **19**, 2555–2567.

[17] Knorr-Held, L. and Rainer, E. (2001). Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction, *Biostatistics*, **2**, 109–129.

[18] Lee, T.C.K.; Dean, C.B. and Semenciw, R. (2011). Short-term cancer mortality projections: A comparative study of prediction methods, *Statistics in Medicine*, **30**, 3387–3402.

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

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

[21] MacNab, Y.C. and Gustafson, P. (2007). Regression B-spline smoothing in Bayesian disease mapping: with an application to patient safety surveillance, *Statistics in Medicine*, **26**, 4455–4474.

[22] Malvezzi, M.; Bertuccio, P.; Levi, F.; La Vecchia, C. and Negri, E. (2012). European cancer mortality predictions for the year 2012, *Annals of Oncology*, **23**, 1044–1052.

[23] Malvezzi, M.; Bertuccio, P.; Levi, F.; La Vecchia, C. and Negri, E. (2013). European cancer mortality predictions for the year 2013, *Annals of Oncology*, **24**, 792–800.

[24] Riebler, A. and Held, L. (2010). The analysis of heterogeneous time trends in multivariate age-period-cohort models, *Biostatistics*, **11**, 57–69.

[25] Riebler, A.; Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data-Borrowing strength from related populations, *The Annals of Applied Statistics*, **6**, 304–333.

[26] Schmid, V. and Held, L. (2004). Bayesian extrapolation of space-time trends in cancer registry data, *Biometrics*, **60**, 1034–1042.

[27] Silva, G.L.; Dean, C.B.; Niyonsenga, T. and Vanasse, A. (2008). Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines, *Statistics in Medicine*, **27**, 2381–2401

[28] Tiwari, R.C.; Ghosh, K.; Jemalm A.; Hachey, M.; Ward, E.; Thun, M.J. and Feuer, E.J. (2004). A new method of predicting U.S. and state-level cancer mortality counts for the current calendar year, *CA: A Cancer Journal for Clinicians*, **54**, 30–40.

[29] Ugarte, M.D.; Goicoa, T.; Etxeberria, J. and Militino, A.F. (2012a). Projections of cancer mortality risks using spatio-temporal Pspline models, *Statistical Methods in Medical Research*, **21**, 545–560.

[30] Ugarte, M.D.; Goicoa, T.; Etxeberria, J. and Militino, A.F. (2012b). A P-spline ANOVA type model in space-time disease mapping, *Stochastic Environmental Research and Risk Assesment*, **26**, 835–845.

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

[32] Ugarte, M.D.; Goicoa, T.; Militino, A.F. and Durbán, M. (2009). Spline smoothing in small area trend estimation and forecasting, *Computational Statistics and Data Analysis*, **53**, 3616–3629.

[33] Ugarte, M.D.; Militino, A.F. and Goicoa, T. (2008). Prediction error estimators in empirical Bayes disease mapping, *Environmetrics*, **19**, 287–300.

[34] Wood, S.N., Scheipl, F. and Faraway, J.J. (2013). Straifhtforward intermediate rank tensor product smoothing in mixed models, *Statistics and Computing*, **23**, 341–360.

[35] Zhu, L.; Pickle, L.W.; Ghosh, K.; Naishadham, D.; Portier, K.; Chen, H.S.; Kim, H.J; Zou, Z.; Cucinelli, J.; Kohler, B.; Edwards, B.K.; King, J.; Feuer, E.J. and Jemal, A. (2012). Predicting US- and state-level cancer counts for the current calendar year. Part II: Evaluation of spatiotemporal projection methods for incidence, *Cancer*, **118**, 1100–1109.

REVSTAT – Statistical Journal
Volume 13, Number 1, March 2015, 41–62

STATISTICAL METHODS FOR DETECTING THE ONSET OF INFLUENZA OUTBREAKS: A REVIEW

Authors: Rubén Amorós
– Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana, Spain
ruben.amoros@uv.es

David Conesa
– Depart. d'Estadística i Investigació Operativa, Univ. de València, Spain
david.v.conesa@uv.es

Miguel Angel Martinez-Beneito
– Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana,
CIBER de Epidemiología y Salud Pública (CIBERESP), Spain
martinez_mig@gva.es

Antonio López-Quílez
– Depart. d'Estadística i Investigació Operativa, Univ. de València, Spain
antonio.lopez@uv.es

Received: October 2014 Revised: November 2014 Accepted: December 2014

Abstract:

- This paper reviews different approaches for determining the epidemic period from influenza surveillance data. In the first approach, the process of differenced incidence rates is modeled either with a first-order autoregressive process or with a Gaussian white noise process depending on whether the system is in an epidemic or a non-epidemic phase. The second approach allows us to directly model the process of the observed cases via a Bayesian hierarchical Poisson model with Gaussian incidence rates whose parameters are modeled differently, depending on the epidemic phase of the system. In both cases transitions between both phases are modeled with a hidden Markov switching model over the epidemic state. Bayesian inference is carried out and both models provide the probability of being in epidemic state at any given moment. A comparison of both methodologies with previous approaches in terms of sensitivity, specificity and timeliness is also performed. Finally, we also review a web-based client application which implements the first methodology for obtaining the posterior probability of being in an epidemic phase.

Key-Words:

- *autoregressive modeling; Bayesian inference; influenza; hidden Markov models; public health; temporal surveillance.*

AMS Subject Classification:

- 62F15, 62P10.

1. INTRODUCTION

Influenza is an infectious disease that affects the upper and/or lower parts of the respiratory tract and is caused by the influenza virus. Influenza spreads around the world in seasonal epidemics, resulting in severe infections and the deaths of hundreds of thousands worldwide annually, and millions in pandemic years (good descriptions about its impact can be found in Simonsen *et al.* [52], Fleming *et al.* [16], and Monto [36]). This propensity for causing large scale seasonal epidemics and pandemics has clearly turned influenza surveillance into a challenging issue in public health practice.

Prospective local and national influenza surveillance systems can provide important and timely information to health service providers on the circulation of the influenza virus in a population. The emergence of the A-H1N1 influenza virus in 2009 and its subsequent rapid global spread was a clear and real example of the need of having good surveillance infrastructures available. But at the same time it has also focused attention on surveillance capabilities worldwide (Lipsitch *et al.*, [31]). In particular, it has shown the need for sentinel surveillance systems that could strengthen a country's capacity for seasonal, novel, and pandemic influenza detection and prevention (Ortiz *et al.* [44]).

There are several surveillance information sources that make it possible to track influenza virus activity such as real-time internet surveys [58], queries [14], microblogging [30], over the counter sales [33], prescription pharmaceutical sales [45], absenteeism registers [13], syndromic/sentinel surveillance [21], laboratory test isolations [43], emergency room visit rates [10], hospital admissions [8], pneumonia and influenza mortality rates [48], etc. Nevertheless, there is no such thing as the "best" surveillance information source. As Cheng *et al.* [5] state, "each method only captures a portion of infections within the community with different timeliness and specificity". On one hand, laboratory test results are highly specific, but take days or even weeks and only capture a small fraction of the infected population. In contrary, data may be collected instantly from the internet, typically from searching engines or social networks, as Broniatowski *et al.* [2], Gesualdo *et al.* [18], Li and Cardie [30] or Grover and Aujla [22] do with Twitter data, or Google Flutrends does with Google queries (Ginsbert *et al.* [19]). A discussion about the role of internet data sources in disease surveillance can be found in Milinovich *et al.* [35]. However, these sources of information provide just indirect measures of the influenza incidence levels in the population so their accuracy may be sometimes poor. Nevertheless, the high amount and the immediate availability of the data from this non-conventional sources make its analysis particularly interesting.

No matter what kind of data the surveillance system uses, there is always the need for an algorithm that, applied to the data, could quickly detect meaningful increases in reported influenza incidence. This would make it possible health

services to get prepared for the incipient outbreak, which could have a great impact on the number of lives saved, outbreak management and the effective setting of prevention measures. This has made the statistical literature to pay considerable attention to early detection methods for influenza, or infectious diseases in general.

Methods based on historical limit are the most widely used for detecting the onset of influenza epidemics and with longer tradition in the epidemiologic literature. These methods are based on the model of Shewhart [50], where a warning is triggered when the difference between the current observation and a theoretical mean of the process surpasses a determined threshold, usually set using the estimated standard error. One way to determine this theoretical mean and threshold is to consider a window of observations of times $t-m, ..., t-1$ from the present year and/or $t-m, ..., t+m$ times from previous years and compute some central estimator and standard error for the observations in these windows, as Stroup *et al.* [55] or Farrington *et al.* [15] do. Another option would be using all non epidemic data as training, fitting a regression model which includes time trend and Fourier periodical terms as proposed by Serfling's method [49], the approach used for influenza surveillance by the Center for Disease Control and Prevention (CDC) of the United States (Muscatello *et al.* [38]). This approach or modifications are also used in other works like Costagliola *et al.* [11], Simonsen *et al.* [52] and Boyle *et al.* [1].

These approaches have some drawbacks in practice (Rath *et al.* [47]). Firstly, a predefinition of epidemic and non-epidemic periods is needed for most of them in order to characterize the observations of the non-epidemic phase, when that division is precisely the final outcome that we want to draw. Secondly, time observations are treated as completely independent values, when we would expect that their temporal arrangement could induce some kind of dependence. Thirdly, the baseline (non-epidemic) period is often estimated with national data that maybe does not properly fit if we are mostly interested in a local influenza surveillance system. Finally, Goddard *et al.* [20] also point out as a fourth drawback that the use of temporally fixed threshold values to describe the levels of influenza activity can be misleading due to time trends in consultations for influenza. Specifically, they pointed out a decline in the number of influenza-related consultations in recent years that could reduce the sensitivity of these methods.

Le Strat and Carrat [28] pioneered the use of hidden Markov models to segment time series of influenza indicators into epidemic and non-epidemic phases. Hidden Markov models are a particular case of Markov switching models, which are stochastic models that consider a set of non observed variables Z_t (hidden states, usually $Z_t = 0$ for the non epidemic state and $Z_t = 1$ for the epidemic state) and a set of observed values y_t (observations), one for each time unit $t \in \{1, ..., T\}$, so that $\{Z_t\}$ is a Markov chain

$$P(Z_t|Z_1, ..., Z_{t-1}, p_{ij}) = P(Z_t|Z_{t-1}, p_{ij}) \tag{1.1}$$

where p_{ij} are the transition probabilities, and the observations y_t are dependent on previous observations, usually through an autoregressive process. The present state Z_t affects both the present observation y_t through the transition probabilities $p(y_t|y_{t-1})$. The conditional relationships in a Markov switching model are represented in Figure 1.

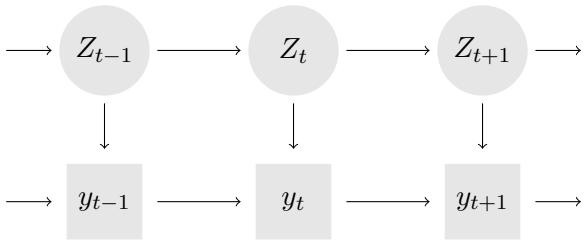

Figure 1: Diagrams of the conditional dependencies in a Markov switching model.

Hidden Markov models are usually formulated with an added restriction, so that the value of the observed variable at each time y_t is only dependent on the hidden state for that time, given the past observations and the present and past states

$$P(y_t|Z_1, ..., Z_t, y_1, ..., y_{t-1}, \boldsymbol{\theta}) = P(y_t|Z_t, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ are the remaining parameters of the model.

Le Strat and Carrat's approach has two advantages, the first being that the method can be applied to historical data without the need to previously distinguish between epidemic and non-epidemic periods in the data. The second one is considering observations as dependent on the past observations and states or, at least, on the last epidemic state, whereas Serfling's method assumes marginal independence of the data [47]. In subsequent papers, Rath *et al.* [47], Madigan [32] and Sun and Cai [56] further developed that modeling. Nevertheless, we find also convenient to mention some other contributions beyond Markov Switching models such Cowling *et al.* [12], Griffin *et al.* [21] and Boyle *et al.* [1] who use models based on statistical quality control and time series methods like CUSUM, EWMA or auto-regressive processes, Frisen *et al.* [17] who search for change points on the monotony of the process, Shmueli [51] who uses wavelet-based methods or Nuño and Pagano who adjust one or several Gaussian peaks in different locations for each year [42]. Three comprehensive reviews of statistical algorithms for the detection of infectious disease outbreaks can be found in Le Strat [27], Burkom [3] and Unkel *et al.* [57].

Bayesian methodology provides a unified theory for handling uncertainty in very different areas such as statistical inference, forecasting, decision-making

under uncertainty, analysis of expert systems, etc. This ability to deal with uncertainties is what makes Bayesian analysis a very advisable tool for many issues that arise in the decision-making process of a surveillance system. Specifically, Bayesian analyses enable to quantify whatever feature of interest of any variable in the model by means of its posterior distribution. In our setting, this makes the Bayesian methodology to be perfectly suited for quantifying the probability of being in an epidemic phase at any given moment. Bayesian studies are not new in surveillance literature, but in recent years there has been increasing interest in them (see for example Mugglin *et al.* [37], Cooper *et al.* [9], Sebastiani *et al.* [48], Niemi *et al.* [40], Zhou and Lawson [61], Charland *et al.* [4] and Neill and Cooper [39]).

Our main goal in this paper is to review two alternative approaches to influenza surveillance that avoid some of the above-mentioned disadvantages and take advantage of the ability to quantify epidemic probabilities of Bayesian methodology. The first proposal is to use a Markov switching model in order to determine the epidemic and non-epidemic periods from influenza surveillance data (Martinez-Beneito *et al.* [34]). This approach differs from those hidden Markov models previously mentioned in the sense that it models the series of differenced incidence rates rather than the series of incidence rates. The new differenced series is detrended, allowing us to take advantage of autoregressive (stationary) modeling to analyze the data. In particular, depending on whether the system is in an epidemic or a non-epidemic phase, the differenced series are, respectively, modeled either with a first-order autoregressive process or with a Gaussian white noise process. The transition between the phases of the disease is considered to follow a Markovian process. The Bayesian paradigm is used to estimate the probability of being in an epidemic phase at any given moment, which is the key to detecting influenza epidemics at their onset.

Two features of this model have proved to be very convenient in the influenza surveillance context. The first one is the use of Markov switching models to segment the time series of influenza into epidemic and non-epidemic phases. The second is the use of the variability of data as main tool to distinguish between both epidemic and non-epidemic phases. Thus, the underlying hypothesis of this model is that non-epidemic dynamics are characterized by small, time-independent random changes (since, supposedly, there is no underlying active process) meanwhile, in epidemic dynamics, changes are greater and possibly correlated.

Nevertheless, although the variability of data may enable to distinguish both dynamics, incorporating the magnitude of the incidence rates and not just their differences could also be very advantageous because this magnitude would also inform on the state of the illness (low incidence clearly meaning a non-epidemic phase). This could increase the capability of the method to distinguish between both epidemic and non-epidemic phases and so it could be easier to

determine the onset of epidemics. As a result, we also review here an enhanced version of that modeling (Conesa *et al.* [7]) that also incorporates the magnitude of the incidence rates. Moreover, this proposal directly models the weekly observed counts as a Poisson distribution depending on the incidence rates, thus these incidence rates are not considered deterministic quantities but, on the contrary, (random) variables in a Bayesian model. As a consequence, this proposal also takes into account the uncertainty in the available incidence rates.

The remainder of this paper is organized as follows. In Sections 2 and 3 we present, respectively, the model based on the differenced rates and on the observed cases. In Section 4 we describe the results obtained when applying the proposed methodologies in a particular setting related to sentinel surveillance data. That section also includes a validation of the performance of both models compared with other existing methodologies. Section 5 describes `fludetweb`, a web-based implementation of the first methodology for obtaining the posterior probability of being in an epidemic phase. Finally, in Section 6 we present some concluding remarks and some lines of development of the current methodology.

2. MODELING OF THE DIFFERENCED INCIDENCE RATES

We first review the modeling introduced by Martinez-Beneito *et al.* [34]. This model performs a segmentation of the differenced incidence rates series into epidemic and non-epidemic phases by using a Markov switching model. Specifically, let $Y = \{Y_{i,j}, i = 1, ..., I; j = 1, ..., J\}$ denote the set of differences between the rates of weeks $i + 1$ and i in year j. We consider a set of retrospective years so that the system has previous information about epidemic periods before observing it in the current year. The underlying idea of Markov switching models is to associate each $Y_{i,j}$ with a random variable $Z_{i,j}$ that determines the conditional distribution of $Y_{i,j}$ given $Z_{i,j}$. In this case, each $Z_{i,j}$ is an unobserved random variable that indicates which phase the system is in (1, epidemic; 0, non-epidemic). The unobserved sequence of $Z_{i,j}$ follows a first order two-state Markov chain with transition probabilities:

$$\mathrm{P}(Z_{i+1,j} = l | Z_{i,j} = k) = P_{k,l}$$

where $k, l \in \{0, 1\}$, $i \in \{1, ..., I - 1\}$ and $j \in \{1, ..., J\}$. This Markovian feature enables epidemic, respectively non-epidemic, weeks to be followed by epidemic, respectively non-epidemic, weeks with a high probability if the data required it. This performance could not be achieved with an independent modeling of the $Z_{i,j}$'s and it makes the epidemic/non-epidemic state to be more robust to sudden, although slight, changes in the differenced series.

The next step is to model the behavior of the differenced series for both epidemic and non-epidemic periods. It seems reasonable to assume no underlying

process beyond Gaussian noise for the non-epidemic period since, supposedly, no underlying mechanism should be inducing dependence among the observations. On the other hand, the epidemic phase should show greater variability, and possibly dependent observations. Therefore, the conditional distribution of $Y_{i,j}$ is modeled either as a Gaussian white noise process or as an autoregressive process of order 1, depending on whether the system is in, respectively, non-epidemic or epidemic phase, *i.e.*

$$
\begin{aligned}
&Y_{1,j}|(Z_{1,j}=0) \sim \mathrm{N}(0,\sigma^2_{0,j}) \\
&Y_{1,j}|(Z_{1,j}=1) \sim \mathrm{N}(0,\sigma^2_{1,j}) \\
&Y_{i,j}|(Z_{i,j}=0) \sim \mathrm{N}(0,\sigma^2_{0,j}) \qquad i\in\{2,...,I\}, \qquad j\in\{1,...,J\}, \\
&Y_{i,j}|(Z_{i,j}=1) \sim \mathrm{N}(\rho Y_{i-1,j},\sigma^2_{1,j}) \qquad i\in\{2,...,I\}, \qquad j\in\{1,...,J\},
\end{aligned}
\tag{2.1}
$$

where the first subindex of the variance $\sigma^2_{k,j}$ represents whether the system is in the epidemic phase ($k=1$) or not ($k=0$). This model assumes a different variance for each season in order to reflect that the variability in any of the phases is not necessarily the same in different years, as a consequence of differences in the shape of the corresponding epidemic waves. Note also that the conditional distribution of the first difference of rates cannot be modeled as an autoregressive process as there is no previous value to condition on.

Once the model is determined, the following step is to estimate its parameters. Martinez-Beneito *et al.* [34] propose using the following prior distributions for the parameters involved in the model:

$$
\begin{aligned}
\rho &\sim \mathrm{Unif}(-1,1) & \theta_{\mathrm{low}} &= \lambda_{[1]} \\
P_{1,1} &\sim \mathrm{Beta}(0.5,0.5) & \theta_{\mathrm{mid1}} &= \lambda_{[2]} \\
P_{0,0} &\sim \mathrm{Beta}(0.5,0.5) & \theta_{\mathrm{mid2}} &= \lambda_{[3]} \\
\sigma_{0,j} &\sim \mathrm{Unif}(\theta_{\mathrm{low}},\theta_{\mathrm{mid1}}) & \theta_{\mathrm{sup}} &= \lambda_{[4]} \\
\sigma_{1,j} &\sim \mathrm{Unif}(\theta_{\mathrm{mid2}},\theta_{\mathrm{sup}})
\end{aligned}
\tag{2.2}
$$

where $\{\lambda_{[1]},\lambda_{[2]},\lambda_{[3]},\lambda_{[4]}\}$ corresponds to the ordered sequence of the variables $\{\lambda_1,\lambda_2,\lambda_3,\lambda_4\}$ which follow as prior distribution:

$$
\lambda_j \sim \mathrm{Unif}(a,b) \;\; j=1,...,4\ , \tag{2.3}
$$

where a and b are hyperparameters to be fixed by the modeler, typically expressing vague prior knowledge.

Expressions (2.1) and (2.2) contain all the knowledge about the system but they do not yield analytical estimates. Therefore, Markov Chain Monte Carlo (MCMC) methods are necessary, WinBUGS [54] being an option for carrying out the inference. See Martinez-Beneito *et al.* [34] for more details on the specific implementation of this model.

3. DIRECT MODELING OF THE OBSERVED COUNTS

The key for developing a method that could be adapted to most kinds of surveillance data is to use a common feature to all of them. This feature is that most surveillance system are usually feed of counting incidence data. In fact, most of the available surveillance data consist of time series of daily/weekly rates directly obtained by transforming time series of observed cases (such as the number of daily/weekly deaths due to influenza, the number of daily/weekly hospital admissions, etc.). Attending to this comment, Conesa *et al.* [7], model the weekly observed cases, which are subjected to sampling variation that should be taken into account in our proposal, instead of directly modeling the rates. In particular, if $O_{i,j}$ denotes the number of observed cases of influenza during week i in season j, they model $O_{i,j}$ by means of a Poisson distribution whose parameter is a function of the incidence rate $r_{i,j}$ of the week i in season j via the following hierarchical structure:

$$\begin{aligned} O_{i,j} &\sim \text{Poisson}(\nu_{i,j}) \\ \nu_{i,j} &= f(r_{i,j}) \\ r_{i,j} &\sim \mathcal{N}(R_{i,j}(Z_{i,j}), \sigma_j^2(Z_{i,j})) \ . \end{aligned} \tag{3.1}$$

The function in the second line in (3.1) depends on the type of data we are working with. For instance, when working with sentinel data in which data are formed by the weekly percentage of patients with influenza, the function in the second line in (3.1) will be:

$$f(r_{i,j}) = \frac{\mathrm{N}_{i,j} \cdot r_{i,j}}{100} \ , \tag{3.2}$$

where $\mathrm{N}_{i,j}$ represents the total number of patients seen in the corresponding week.

The rates now are modeled again as a Normal distribution with both mean and variance depending on $Z_{i,j}$, an unobserved random variable that indicates which phase the system is in (1, epidemic; 0, non-epidemic). As in the previous model, this is the idea of a Markov switching model in which the unobserved sequence of $Z_{i,j}$ follows a two-state Markov chain of first order with transition probabilities:

$$P(Z_{i+1,j} = l | Z_{i,j} = k) = P_{k,l} \quad k, l \in \{0, 1\} \ , \tag{3.3}$$

with $P_{0,1} = 1 - P_{0,0}$ and $P_{1,0} = 1 - P_{1,1}$ for suitable probabilities $P_{0,0}$ and $P_{1,1}$.

The model assumes constant but different variances for each phase of each season: $\{\sigma_j^2(0), \sigma_j^2(1) : j = 1, ..., J\}$. Moreover, the variance of the epidemic phase will be assumed higher than that in the non-epidemic phase. As already mentioned, non-epidemic dynamics are characterized by small random changes while

in epidemic dynamics the changes are greater. This will help once again to separate the series in 2 different periods. Different variances are assumed for each season in order to reflect the different features of the different epidemic seasons since, for example, some years have higher and steeper incidence peaks in contrast to other years with flatter epidemic waves.

The next step is to model the mean of the rates in both states. Note that $R_{i,j}(0)$ and $R_{i,j}(1)$ represent the level of magnitude of the incidence in case of being, respectively, at non-epidemic or epidemic phase. The model at every week decides which of them (jointly with the corresponding variance) fits better to the new observed data. The first and easiest way to model both means is to consider them as independent but constant. This modeling can be denoted as AR0-AR0, *i.e.* two order 0 (independent) autoregressive processes, these terms representing respectively the non-epidemic and epidemic phases. A second option would be to consider the mean of the rates as temporally dependent processes. In this setting, it could be convenient to model the mean of the rates as a first order autoregressive process. As a result, three more modelings could be considered combining the AR0 and AR1 proposals, that is: AR0-AR1, AR1-AR0 and AR1-AR1. In a similar way, we could think that the rates are related to rates from two or more previous weeks. Then a suitable option would clearly be to consider the mean of the rates as an autoregressive process of higher order. When dealing with second order autoregressive processes, five more modelings could be considered, that is AR2-AR0, AR2-AR1, AR2-AR2, AR1-AR2 and AR0-AR2. For simplicity, here we just present the AR1-AR1 model, with both means (those corresponding to non-epidemic and epidemic settings) being first order autoregressive processes, *i.e.*:

$$\begin{aligned} R_{i,j}(0)|r_{1,j},...,r_{(i-1),j} &= \mu_0 + \rho_0 \cdot (r_{i-1,j} - \mu_0) \\ R_{i,j}(1)|r_{1,j},...,r_{(i-1),j} &= \mu_1 + \rho_1 \cdot (r_{i-1,j} - \mu_1) \end{aligned} \tag{3.4}$$

with $\mu_0 < \mu_1$ in order to set the epidemic period as that having a higher expected rate.

The specification proposed by Conesa *et al.* [7] for the prior distributions of each of the parameters involved in this model is the following:

$$\begin{aligned} P_{0,0} &\sim \text{Beta}(0.5, 0.5) & \theta_{\text{low}} &= \lambda_{[1]} \\ P_{1,1} &\sim \text{Beta}(0.5, 0.5) & \theta_{\text{mid1}} &= \lambda_{[2]} \\ \sigma_j(0) &\sim \text{Unif}(\theta_{\text{low}}, \theta_{\text{mid1}}) & \theta_{\text{mid2}} &= \lambda_{[3]} \\ \sigma_j(1) &\sim \text{Unif}(\theta_{\text{mid2}}, \theta_{\text{sup}}) & \theta_{\text{sup}} &= \lambda_{[4]} \end{aligned}$$

where $\{\lambda_{[1]}, \lambda_{[2]}, \lambda_{[3]}, \lambda_{[4]}\}$ corresponds to the ordered sequence of the variables $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ which follow the non-informative prior distribution:

$$\lambda_j \sim \text{Unif}(0, c) \;\; j = 1, ..., 4 \;, \tag{3.5}$$

where c is usually high enough so that it does not condition the posterior distribution of $\lambda_1, ..., \lambda_4$.

With respect to the parameters involved in the modeling of the mean of the rates, that is, μ_0 and μ_1, they propose a slightly different structure than the one used for the variance:

$$\mu_0 = \theta_{[1]} \ , \qquad \mu_1 = \theta_{[2]} \ , \tag{3.6}$$

where $\{\theta_{[1]}, \theta_{[2]}\}$ corresponds to the ordered sequence of the variables $\{\theta_1, \theta_2\}$ which follow the non-informative prior distribution:

$$\theta_j \sim \text{Unif}(0, d) \ , \qquad j = 1, 2 \ , \tag{3.7}$$

where d is again a hyperparameter chosen in a way that makes vague the former distribution.

Finally, they also propose considering flat prior distributions for any of the parameters of the autoregressive processes. Specifically, they choose uniform distributions in the region where the processes are stationary. As an example, the selection in the AR1-AR1 modeling would be:

$$\rho_0, \rho_1 \sim \text{Unif}(-1, 1) \ . \tag{3.8}$$

Again, expressions from (3.1) to (3.8) contain all the knowledge of the system but these expressions do not yield analytical expressions for the posterior distribution of the parameters. MCMC methods and WinBUGS are again a good option for carrying out the inference. See Conesa *et al.* [7] for more details on the specific implementation of this model.

4. COMPARING METHODOLOGIES ON SENTINEL SURVEILLANCE DATA

One of the most popular kinds of influenza surveillance data comes from sentinel systems. In spite of their limitations, these data have proved to be very useful to follow up influenza during the last decade, rapid information transmission being one of their main advantages. Basically, sentinel systems are formed by volunteer practitioners that, depending on the system, report in a weekly basis the percentage of patients with Influenza-like illness (ILI), usually defined as fever plus acute respiratory symptoms such as cough and/or sore throat, from the total number of patients seen, or just the number of consultations with patients reporting ILI symptoms. Data are collected at least during the periods of non-negligible influenza activity. Thus in Western countries in the temperate climate

zone, data are typically collected in seasons (lasting around 30–35 weeks) that extend over two consecutive years (as the epidemic activity usually extends across both of them), while in other places data are collected throughout the year.

In what follows we review a case study presented in Conesa *et al.* [7] in order to check how these two methodologies can be applied in real settings. In particular data analyzed in this example are retrieved from the North Carolina Influenza Sentinel Surveillance Program. The dataset is formed by the weekly percentage of ILI patients from the total number of seen patients. Information about this system can be found in its website [41], which allows to export raw data in standard format. In this system, sentinels also collect representative samples for virus strain identification. As we mention later, this laboratory data on influenza isolates can be very useful as a gold standard for validating the performance of a method.

It is worth noting that the methodology introduced in Section 3 requires the selection of the model that fits each phase better from the different options introduced in the previous section. As a result, the first step must be to use a model selection criterion. One option is to use DIC, a Bayesian model comparison criterion introduced by Spiegelhalter *et al.* [53], based on the trade-off between the fit of the data and the corresponding complexity of the model. In our experience, models with different structures in both phases tend to give more weight to the phase with more structure. As a result we usually consider just models with both phases having the same structure, in order to avoid decompensation between phases. In this case, we have only analyzed those three options with the same structure in both phases (AR0-AR0, AR1-AR1 and AR2-AR2). For this data, results indicate that models AR1-AR1 and AR2-AR2 outperform (as they have lower DIC, 2097.6 and 2095.9 respectively) the AR0-AR0 model (with DIC 2125.8). As a result the AR0-AR0 model can be discarded, the next step being to choose between the remaining two models.

The assessment of models can be done using the measures proposed by Cowling *et al.* [12] and Kleinman and Abrams [25]. These measures summarize the information on sensitivity, specificity and timeliness for a detection method and a particular data set in a unique value between 0 and 1, achieving 1 for a perfect performance. In particular, the measures used here are `AUWROC1` and `VUTROS1`, which weigh the information given by the ROC curve by the timeliness achieved for each method, building a 3D ROC curve which height is a function of the timeliness of detection and measuring the area left behind this curve, and `VUTROS3` and `VUTROCS`, which construct several reference ROC curves restricted by different maximum delay of detection and integrate the information from them all.

In order to perform this comparison we need to know approximately the real epidemic periods in all the seasons. The approach used here to obtain the gold-standard is that presented in Cowling *et al.* [12]. In particular, using laboratory

data on influenza isolates from the North Carolina Sentinel Network, the period between the first and last week in which the proportion of positive isolations for influenza surpassed 30% of the maximum seasonal level has been taken as the gold standard epidemic phase.

This criteria may also be used beyond the comparison of the AR1-AR1 and AR2-AR2 models to compare these proposals with alternative models in the literature. In particular, here we present the results of applying those metrics to the on-line results obtained with both models and those obtained using four alternative methods for the automatic detection of influenza, again using information only from past weeks. The first approach is the widely-used method proposed by Serfling [49], while the second one uses the `depmix` package [59] of R [46] to reproduce Le Strat and Carrat's [28] hidden Markov model. The third approach is a simple regression method which is a slight modification of Stroup *et al.* [55] and the last one is the model based on the differences of the rates (MBDR from now on) presented in Section 2.

It is worth noting that this validation has been done using an on-line version of all the models instead of the results obtained by applying them to the whole data set. That is, by applying all models sequentially, starting from the fourth season (in order to use at least three seasons as training data) and then predicting, for each week, the probability of being in an epidemic phase by only taking into account the information from past weeks (within the current season and all the weeks from previous seasons). The reason for this is that the on-line version reproduces the sequential arrival of data that is habitual in real situations.

Table 1 shows the values of the metrics `AUWROC1`, `VUTROS1`, `VUTROS3` and `VUTROCS` obtained when applying the selected models (AR1-AR1 and AR2-AR2) and the four alternative methods to the North Carolina data set. As can be appreciated, results indicate that both models perform better than the other alternative methods, since the values of the four metrics are greater and closer to 1. Moreover, the AR2-AR2 behaves substantially better and so it will be our selection from now on for analyzing the North Carolina data set.

Table 1: Comparison of metrics AUWROC1, VUTROS1, VUTROS3 and VUTROCS when applying different early warning systems (higher being better) to the North Carolina data set.

	AUWROC1	VUTROS1	VUTROS3	VUTROCS
Serfling	0.612	0.553	0.349	0.698
depmix	0.608	0.556	0.341	0.682
Stroup	0.540	0.517	0.404	0.807
MBDR	0.601	0.544	0.356	0.713
AR1-AR1	0.676	0.595	0.412	0.824
AR2-AR2	0.726	0.649	0.420	0.840

The results of applying the on-line version of the AR2-AR2 model can be seen in Figure 2, which shows the weekly ILI incidence rates, the white, grey and black dots indicating those weeks where the posterior probability of being in an epidemic phase exceeds the values 0.25, 0.5 and 0.75, respectively. As mentioned above, these probabilities have been sequentially obtained, starting from the fourth season and only taking into account the information from past weeks (within the current season and the all weeks from previous seasons). This is the kind of graph that Health Authorities could use to raise the alarm at those precise moments in which there is a high probability of being in an epidemic phase. In particular, values exceeding 0.5 indicate that we are observing for that week a higher probability of being in an epidemic phase than of being in a non-epidemic one, and so an alarm could be triggered if considered convenient.

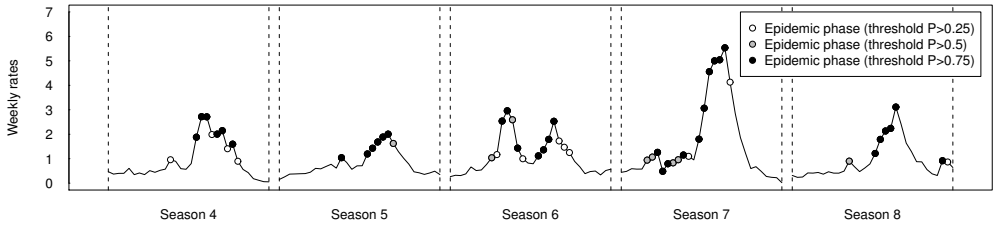

Figure 2: On-line weekly results for the North Carolina data set from seasons 4 to 8. Graphs show the influenza incidence rates in which the white, grey and black dots indicate those weeks where the posterior probability of being in an epidemic phase exceeds the values 0.25, 0.5 and 0.75 respectively.

5. FLUDETWEB

The complexity of disease surveillance methods has been progressively increasing. In fact, most of the methods mentioned in the Introduction are not easy to implement. On the contrary, most of them and, in general, the most advanced surveillance systems require skilled personnel to implement, fine-tune and maintain them. These requirements have kept these new developments from common usage. In order to resolve this issue, there has been a recent interest in enhancing existing disease surveillance methodologies by using tools for presenting data and information to users. Hauenstein *et al.* [23] describe in detail the processes and tools (such as system architecture, web-based applications, etc.) needed to do so.

In this section, we review `fludetweb` (Conesa *et al.* [6]), an enhanced web implementation of the MBDR, the surveillance methodology described in Section 2. This implementation is available on-line at: `http://www.geeitema.org/meviepi/fludetweb/`.

`Fludetweb`'s implementation has been done using a thin client application design for ease of user interaction with the program, that is through a web application that could be accessed by any network-enabled device (PC's, tablets, cell phones, etc.) with a web browser. But moreover, the computational requirements of the detection algorithm, which could need several minutes to return the results, also motivated the use of a master-slave intranet architecture to take advantage of other secondary available computers.

Figure 3 shows the internal architecture of the server and its connections with the slaves and clients. The system has been implemented as a three-tier architecture by separating its functions into three separate layers. The top tier corresponds to the presentation layer and is responsible for interaction between the user and the system through data and personal information querying, visualization of results, etc. The second tier is the business logic tier, which is the core of the system as it controls the running of the influenza surveillance algorithm. The final layer is the data tier and it is responsible for data storage, not only of the influenza rates but also of the user's personal information, the availability and state of slaves, IP addresses, assigned tasks, etc.

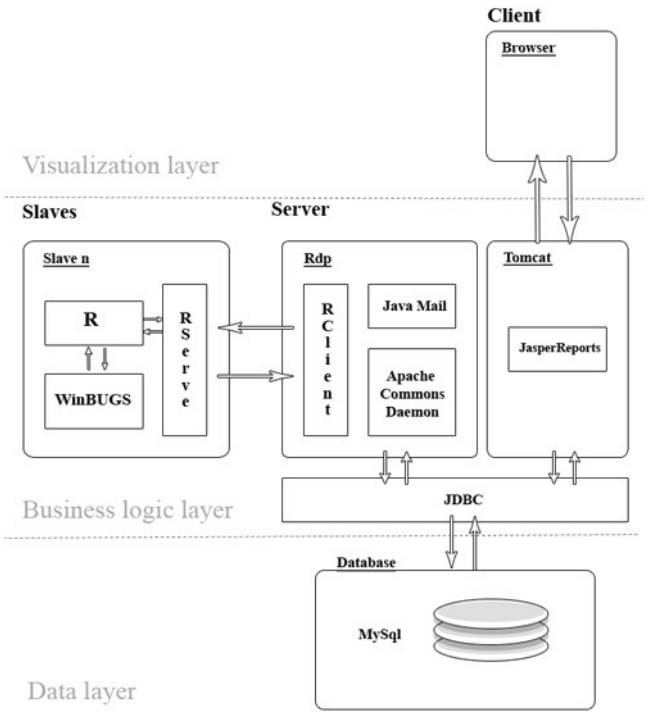

Figure 3: Internal architecture of the `fludetweb` implementation, including its internal connections with slaves and clients.

In practice, users can introduce and edit their own data consisting of a series of weekly influenza incidence rates. Users may also obtain estimates of

the probability of being in an epidemic phase for the weeks of interest. The estimation process is not immediate, so the system has been designed to respond to requests from a multi-user environment on a first-come, first-served basis. After completion of the process, the system returns the probability of being in an epidemic phase jointly with a forecast of the probability of an increase in the incidence rate during the following week. `Fludetweb` also provides two further graphs. The first one shows the weekly rates of the last two seasons, indicating whether the posterior probability of being in an epidemic phase in the analyzed week is greater than 0.5 or not. The second graph shows all the weekly rates with flags only for requested weeks. In particular, flags indicate whether the posterior probability of being in an epidemic phase is greater than 0.5 or not. Its ease of use and on-line availability should make `fludetweb` a valuable tool for public health practitioners.

6. FURTHER LINES OF DEVELOPMENT

Our interest in this paper has been to review two possible methodologies for detecting influenza epidemics at the very moment of their onset. Both modelings can be used for any kind of surveillance data that show either epidemic periods, describing an increase and a decrease of the epidemic curve, or non-epidemic periods, where the data follows a random noise process. Both methodologies provide the posterior probability of being in an epidemic phase, and so they can be very useful for health authorities who could use them to raise the alarm at those precise moments in which there is a high probability of being in, or even better starting, an epidemic phase. We have also reviewed a web-based implementation of the first methodology for obtaining the posterior probability of being in an epidemic phase.

With respect to possible extensions of these proposals, a first improvement could be to include a spatial component that would help us to raise geographically-referenced alarms. Now, we would have as many time series as areal units in the region of study and the goal would be to induce spatial dependence on these series in order to take advantage of the information of neighboring series. As an example Zou *et al.* [62] and Heaton *et al.* [24] propose to use, for the epidemic phase, spatio-temporally correlated random effects, with every random effect being conditional dependent to its temporal and spatial neighbors. On the other hand, they consider either a white noise or a spatially correlated process for the non-epidemic phase. Finally, the transition probability between states for every region is considered to depend on the number of neighbors in epidemic state. Knorr-Held and Richardson [26], Watkins *et al.* [60], Li *et al.* [29] or Li and Cardie [30] are just some other examples of studies addressing the incorporation of a spatial component in outbreak detection studies.

Spatio-temporal models improve traditional temporal outbreak detection models by using a second source of information, that provided by neighboring regions. A second alternative would be to incorporate complementary information sources, not necessarily linked to other geographical locations. Thus, Nunes *et al.* [43] consider a bivariate information source (for every week) instead of the traditional univariate rates for distinguishing between epidemic and non-epidemic periods. Thus, on one hand, they use for each week the number of laboratory-confirmed cases of the previous week and, on the other hand, the number of reported ILI cases of the current week. That is, we have two information sources in this system one of higher quality (confirmed cases in the previous week) and a second one corresponding to more recent data. This kind of analysis would be more sensible than traditional univariate studies as they are based on a larger amount of information in order to decide the current epidemic/non-epidemic state.

A particular case of multivariate data use is that of Twitter. Twitter is a microblogging social network in which users publicly post short messages of less than 140 characters that may also be geolocated. Several words or sets of words related to influenza can be used to predict the onset, spread and decay of the epidemic. Works like those of Li and Cardie [30] or Grover and Aujla [22] deal with the necessary preprocessing of the raw data and the use of Markov chains to model several phases like rising, stationarity and declining of the epidemic.

Flexibility is one of the main advantages of Bayesian hierarchical models. These models can be easily adapted to any specific feature of any dataset, what has made them particularly common in outbreak detection studies. Nevertheless, the main drawback of Bayesian hierarchical models is that they usually resort to (frequently slow) MCMC simulation to carry inference out. Although WinBUGS usually makes relatively easy the inference process, once a new weekly observation arrives, we are forced to repeat the whole simulation with the new (and all the previous) observation(s). This makes simulations to become progressively slower as new observations are incorporated into the system. Sequential MCMC methods, such as particle filters, would be a solution to this, avoiding to run again the model each week with all historical and, obviously, the new data. The incorporation of this kind of inference tools to influenza outbreak detection problems could be very advantageous for a problem where the computing time is a limitation if results want to be used for on-line epidemiological surveillance.

Finally, we would like to mention a general drawback of this kind of models. Model selection is a delicate issue in outbreak detection studies. Most model selection criteria are based on fit or predictive properties of models. Nevertheless, in our particular case we would be mostly interested in some other particular criteria such as sensitivity, specificity or timeliness in the detection of outbreaks. As mentioned in this paper, several measures have been proposed in the literature (Cowling *et al.* [12] and Kleinman and Abrams [25]) paying specific attention to these aspects and, therefore, particularly suited for outbreak detection.

Regretfully these methods depend on the availability of a gold standard, that make it possible to assess the goodness of any particular method. Acceptable gold standards are not generally available excepting maybe for very few publicly accessible datasets. This makes model selection in this area a particularly cumbersome issue where more research (or more publicly accessible datasets for this purpose) would be very welcome.

ACKNOWLEDGMENTS

This paper constitutes part of the work developed by the authors during their visit as Research Fellows to the Statistical and Applied Mathematical Sciences Institute in North Carolina in the working group "Fundamentals of Spatial Modeling" of the program "Space-time analysis for environmental mapping, epidemiology and climate change". Financial support from the Conselleria de Sanitat of the Generalitat Valenciana (the Valencian Regional Health Authority) is gratefully acknowledged. The authors would also like to acknowledge financial support from the Ministerio de Educación y Ciencia (the Spanish Ministry of Education and Science) via research grants MTM2010-19528, MTM2013-42323-P (jointly financed with the European Regional Development Fund) and FUT-C2-0047 (as part of the INGENIO-MATHEMATICA research project) and from the Generalitat Valenciana via the research grant EVES-015/2008.

REFERENCES

[1] Boyle, J.R.; Sparks, R.S.; Keijzers, G.B.; Crilly, J.L.; Lind, J.F. and Ryan, L.M. (2011). Prediction and surveillance of influenza epidemics, *The Medical Journal of Australia*, **194**, 4, S28–33.

[2] Broniatowski, D.A.; Paul, M.J. and Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic, *PloS ONE*, **8**, 12, e83672.

[3] Burkom, H.S. (2007). *Alerting algorithms for biosurveillance*. In "Disease Surveillance, a Public Health Informatics Approach" (J.S. Lombardo and D.L. Buckeridge, Eds.), John Wiley & Sons, Ltd, 143–192.

[4] Charland, K.M.L.; Buckeridge, D.L.; Sturtevant, J.L.; Melton, F.; Reis, B.Y.; Mandl, K.D. and Brownstein, J.S. (2009). Effect of environmental factors on the spatio-temporal patterns of influenza spread, *Epidemiology and Infection*, **137**, 10, 1377–1387.

[5] Cheng, C.K.Y.; Lau, E.H.Y.; Ip, D.K.M.; Yeung, A.S.Y.; Ho, L.M. and Cowling, B.J. (2009). A profile of the online dissemination of national influenza surveillance data, *BMC Public Health*, **9**, 339.

[6] Conesa, D.; López-Quílez, A.; Martinez-Beneito, M.A.; Miralles, M.T. and Verdejo, F. (2009). FluDetWeb: an interactive web-based system for the early detection of the onset of influenza epidemics, *BMC Medical Informatics and Decision Making*, **9**, 36.

[7] Conesa, D.; Martinez-Beneito, M. A.; Amorós, R. and López-Quílez, A. (2015). Bayesian hierarchical Poisson models with a hidden Markov structure for the detection of influenza epidemic outbreaks, *Statistical Methods in Medical Research*, online first as doi: 10.1177/0962280211414853.

[8] Congdon, P. (2005). *Bayesian Models for Categorical Data*, John Wiley & Sons, Inc.

[9] Cooper, G.F.; Dash, D.H.; Levander, J. D.; Wong, W.K.; Hogan, W.R. and Wagner, M.M. (2004). *Bayesian biosurveillance of disease outbreaks*. In "UAI'04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence" (J.S. Lombardo and D.L. Buckeridge, Eds.), Arlington, Virginia, AUAI Press, 94–103.

[10] Corberán-Vallet, A. and Lawson, A.B. (2014). Prospective analysis of infectious disease surveillance data using syndromic information, *Statistical Methods in Medical Research*, **23**, 6, 572–590.

[11] Costagliola, D.; Flahault, A.; Galinec, D.; Garnerin, P.; Menares, J.; Valleron, A.J. and Gamenn, P. (1991). A routine tool for detection and assessment of epidemics of influenza-like syndromes in France, *American Journal of Public Health*, **81**, 11, 97–99.

[12] Cowling, B.J.; Wong, I.O.L.; Ho, L.M.; Riley, S. and Leung, G.M. (2006). Methods for monitoring influenza surveillance data, *International Journal of Epidemiology*, **35**, 5, 1314–1321.

[13] Crawford, G.B.; McKelvey, S.; Crooks, J.; Siska, K.; Russo, K. and Chan, J. (2011). Influenza and school-based influenza-like illness surveillance: a pilot initiative in Maryland, *Public Health Reports*, **126**, 4, 591–596.

[14] Dukic, V.; Lopes, H.F. and Polson, N.G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model, *Journal of the American Statistical Association*, **107**, 500, 1410–1426.

[15] Farrington, C.P.; Andrews, N.J.; Beale, A.D. and Catchpole, M.A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **159**, 3, 547–563.

[16] Fleming, D.M.; Zambon, M.; Bartelds, A.I. and de Jong, J.C. (1999). The duration and magnitude of influenza epidemics: a study of surveillance data from sentinel general practices in England, Wales and the Netherlands, *European Journal of Epidemiology*, **15**, 5, 467–473.

[17] Frisén, M.; Andersson, E. and Schiöler, L. (2009). Robust outbreak surveillance of epidemics in Sweden, *Statistics in Medicine*, **28**, 3, 476–493.

[18] Gesualdo, F.; Stilo, G.; Agricola, E.; Gonfiantini, M.V.; Pandolfi, E.; Velardi, P. and Tozzi, A.E. (2013). Influenza-like illness surveillance on Twitter through automated learning of naïve language, *PloS ONE*, **8**, 12, e82489.

[19] Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data, *Nature*, **457**, 1012–1014.

[20] Goddard, N.L.; Kyncl, J. and Watson, J.M. (2003). Appropriateness of thresholds currently used to describe influenza activity in England, *Communicable Disease and Public Health / PHLS*, **6**, 3, 238–245.

[21] Griffin, B.A.; Jain, A.K.; Davies-Cole, J.; Glymph, C.; Lum, G.; Washington, S.C. and Stoto, M.A. (2009). Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system, *BMC Public Health*, **9**, 483.

[22] Grover, S. and Aujla, G.S. (2014). Prediction model for influenza epidemic based on Twitter data, *International Journal of Advanced Research in Computer and Communication Engineering*, **3**, 7, 7541–7545.

[23] Hauenstein, L.; Wojcik, R.; Loschen, W.; Ashar, R.; Sniegoski, C. and Tabernero, N. (2007). *Putting it together: the biosurveillance information system*. In "Disease Surveillance: A Public Health Informatics Approach" (J.S. Lombardo and D.L. Buckeridge, Eds.), John Wiley & Sons, Ltd, 193–261.

[24] Heaton, M.J.; Banks, D.L.; Zou, J.; Karr, A.F.; Datta, G.; Lynch, J. and Vera, F. (2012). A spatio-temporal absorbing state model for disease and syndromic surveillance, *Statistics in Medicine*, **31**, 19, 2123–2136.

[25] Kleinman, K.P. and Abrams, A.M. (2006). Assessing surveillance using sensitivity, specificity and timeliness, *Statistical Methods in Medical Research*, **15**, 5, 445–464.

[26] Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 2, 169–183.

[27] Le Strat, Y. (2005). *Overview of temporal surveillance*. In "Spatial and Syndromic Surveillance for Public Health" (A.B. Lawson and K. Kleinman, Eds.), John Wiley & Sons, Ltd, 13–29.

[28] Le Strat, Y. and Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models, *Statistics in Medicine*, **18**, 24, 3463–3478.

[29] Li, G.; Best, N.; Hansell, A.L.; Ahmed, I. and Richardson, S. (2012). BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice, *Biostatistics*, **13**, 4, 695–710.

[30] Li, J. and Cardie, C. (2013). Early Stage Influenza Detection from Twitter. Working paper.

[31] Lipsitch, M.; Riley, S.; Cauchemez, S.; Ghani, A.C. and Ferguson, N.M. (2009). Managing and reducing uncertainty in an emerging influenza pandemic, *The New England Journal of Medicine*, **361**, 2, 112–115.

[32] Madigan, D. (2011). *Bayesian data mining for health surveillance*. In "Spatial and Syndromic Surveillance for Public Health" (A.B. Lawson and K. Kleinman, Eds.), John Wiley & Sons, Ltd, 203–221.

[33] Magruder, S.F. (2003). Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease, *Johns Hopkins Applied Physics Laboratory Technical Digest*, **24**, 4, 349–353.

[34] Martinez-Beneito, M.A.; Conesa, D.; López-Quílez, A. and López-Maside, A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics, *Statistics in Medicine*, **27**, 22, 4455–4468.

[35] Milinovich, G.J.; Williams, G.M.; Clements, A.C.A. and Hu; W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases, *The Lancet Infectious diseases*, **14**, 2, 160–168.

[36] Monto, A.S. (1999). Individual and community impact of influenza, *Pharmacoeconomics*, **16**, Suppl 1, 1–6.

[37] Mugglin, A.S.; Cressie, N. and Gemmell, I. (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time, *Statistics in Medicine*, **21**, 18, 2703–2721.

[38] Muscatello, D.J.; Morton, P.M.; Evans, I. and Gilmour, R. (2008). Prospective surveillance of excess mortality due to influenza in New South Wales: feasibility and statistical approach, *Communicable Diseases Intelligence*, **32**, 4, 435–442.

[39] Neill, D.B. and Cooper, G.F. (2010). A multivariate Bayesian scan statistic for early event detection and characterization, *Machine Learning*, **79**, 3, 261–282.

[40] Niemi, J.B.; Smith, M. and Banks, D.L. (2008). *Test power for drug abuse surveillance*. In "Bio-surveillance and Biosecurity", Springer, Ltd, 131–142.

[41] North Carolina Influenza Sentinel Surveillance Program. (2014). Communicable disease: North Carolina Influenza update. Available in *http://www.flu.nc.gov/data/*. Data of access: 2014 October 22.

[42] Nuño, M. and Pagano, M. (2007). *A model for characterizing annual flu cases*. In "Intelligence and Security Informatics: Biosurveillance", Springer Berlin Heidelberg, 37–46.

[43] Nunes, B.; Natário, I. and Carvalho, M.L. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models, *Statistics in Medicine*, **32**, 15, 2643–2660.

[44] Ortiz, J.R.; Sotomayor, V.; Uez, O.C.; Oliva, O.; Bettels, D.; McCarron, M.; Bresee, J.S. and Mounts, A.W. (2009). Strategy to enhance influenza surveillance worldwide, *Emerging Infectious Diseases*, **15**, 8, 1271–1278.

[45] Patwardhan, A. and Bilkovski, R. (2012). Comparison: Flu prescription sales data from a retail pharmacy in the US with Google Flu trends and US ILINet (CDC) data as flu activity indicator, *PloS ONE*, **7**, 8, e43611.

[46] R Core Team. (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[47] Rath, T.M.; Carreras, M. and Sebastiani, P. (2003). *Automated detection of influenza epidemics with hidden Markov hodels*. In "Advances in Intelligent Data Analysis V" (M.R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, Eds.), Springer, Berlin, 521–532.

[48] Sebastiani, P.; Mandl, K.D.; Szolovits, P.; Kohane, I.S. and Ramoni, M.F. (2006). A Bayesian dynamic model for influenza surveillance, *Statistics in Medicine*, **25**, 11, 1803–1825.

[49] Serflin, R.E. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths, *Public Health Reports*, **78**, 6, 494–506.

[50] Shewhart, W.A. (1931). *Economic Control of Quality of Manufactured Product*, Princeton: Van Nostrand Reinhold.

[51] Shmueli, G. (2005). Wavelet-based monitoring for modern biosurveillance. Technical report, Robert H. Smith School of Business, University of Maryland, College Park.

[52] Simonsen, L.; Clarke, M.J.; Williamson, G.D.; Stroup, D.F.; Arden, N.H. and Schonberger, L.B. (1997). The impact of influenza epidemics on mortality: introducing a severity index, *American Journal of Public Health*, **87**, 12, 1944–1950.

[53] Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 4, 583–639.

[54] Spiegelhalter, D.J.; Thomas, A.; Best, N. and Lunn, D.J. (2003). *WinBUGS User Manual, Version 1.4.*, MRC Biostatistics Unit.

[55] Stroup, D.F.; Williamson, G.D.; Herndon, J.L. and Karon, J.M. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data, *Statistics in Medicine*, **8**, 3, 323–332.

[56] Sun, W. and Cai, T.T. (2009). Large-scale multiple testing under dependence, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 2, 393–424.

[57] Unkel, S.; Farrington, C.P.; Garthwaite, P.H.; Robertson, C. and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **175**, 1, 49–82.

[58] Vandendijck, Y.; Faes, C. and Hens, N. (2013). Eight years of the Great Influenza Survey to monitor influenza-like illness in Flanders, *PloS ONE*, **8**, 5, e64156.

[59] Visser, I. (2007). depmix: An R-package for fitting mixture models on mixed multivariate data with Markov dependencies. R-package manual.

[60] Watkins, R.E.; Eagleson, S.; Veenendaal, B.; Wright, G. and Plant, A.J. (2009). Disease surveillance using a hidden Markov model, *BMC Medical Informatics and Decision Making*, **9**, 39.

[61] Zhou, H. and Lawson, A.B. (2008). EWMA smoothing and Bayesian spatial modeling for health surveillance, *Statistics in Medicine*, **27**, 28, 5907–5928.

[62] Zou, J.; Karr, A.F.; Banks, D.; Heaton, M.J.; Datta, G.; Lynch, J. and Vera, F. (2012). Bayesian methodology for the analysis of spatial-temporal surveillance data, *Statistical Analysis and Data Mining*, **5**, 3, 194–204.

REVSTAT – Statistical Journal
Volume 13, Number 1, March 2015, 63–78

LONGITUDINAL ANALYSIS OF TUMOR MARKER CEA OF BREAST CANCER PATIENTS FROM BRAGA'S HOSPITAL

Authors: Ana Borges
– CIICESI, ESTGF — Polytechnic Institute of Porto, DMA-ECUM, Minho's University, Portugal
aib@estgf.ipp.pt

Inês Sousa
– Department of Mathematics and Applications, Minho's University, Centre of Mathematics (CMAT), Minho's University, Portugal
isousa@math.uminho.pt

Luís Castro
– Senology Unit of Braga's Hospital, Portugal
luis.castro@hospitaldebraga.pt

Received: October 2014 Revised: November 2014 Accepted: December 2014

Abstract:

- Allied to an epidemiological study of population of the Senology Unit of Braga's Hospital that have been diagnosed with malignant breast cancer, we describe the progression in time of repeated measurements of tumor marker Carcinoembryonic antigen (CEA). Our main purpose is to describe the progression of this tumor marker as a function of possible risk factors and, hence, to understand how these risk factors influences that progression. The response variable, values of CEA, was analyzed making use of longitudinal models, testing for different correlation structures. The same covariates used in a previous survival analysis were considered in the longitudinal model. The reference time used was time from diagnose until death from breast cancer. For diagnostic of the models fitted we have used empirical and theoretical variograms. To evaluate the fixed term of the longitudinal model we have tested for a changing point on the effect of time on the tumor marker progression. A longitudinal model was also fitted only to the subset of patients that died from breast cancer, using the reference time as time from date of death until blood test.

Key-Words:

- *breast cancer; carcinoembryonic antigen; longitudinal models.*

AMS Subject Classification:

- 62P10, 62J10.

1. INTRODUCTION

Oncological diseases are the second highest cause of death in Portugal, and they have a big social impact in patients and their families [12]. In Europe breast cancer is the tumor with highest incidence in women [1]. In Portugal there are not many published studies on breast cancer. However, Pinheiro *et al.* (2003) ([12]) refer that, since 1995, mortality due to breast cancer has been decreasing in Portugal. They argue that this improvement is a consequence of earlier diagnostic and better quality of treatment.

According to results presented by the European Cancer Observatory [5], the estimated incidence for Breast Cancer in Portuguese women in 2012 is 85.6% and the estimated mortality rate due to this type of cancer is 18.4%, both values are quite lower than the European average (94.2% and 23.1% respectively). At the moment, the existing recommendations and guidelines from the National Health Service are mainly based on European studies. However, it is not clear that the behavior of the disease is similar among European countries. Therefore, it is of great importance the continuous investment on statistical and epidemiological studies in oncological diseases for understanding the progression of the disease in Portugal.

This study aims to answer at least some of the questions on a specific Portuguese population, particularly the population of the Senology Unit of Braga's Hospital, located in the north of Portugal, that were diagnosed with malignant breast cancer.

The tumor marker Carcinoembryonic antigen (CEA) is usually used for therapy monitoring in advanced disease ([6]), although recent reports, *e.g.* Fiorella *et al.* (2001) ([6]), discourage its routine use because of low sensitivity. The authors conclude that its use should be considered as an inefficient method of follow-up evaluation for breast cancer patients, and it provides no additional value when used in combination with another tumor marker Carcinoma Antigen 15-3 (CA 15-3). Nevertheless, as Sturgeon *et al.* ([16]) point out, on occasion, it can be informative when levels of CA 15-3 remain below the cutoff point.

Since it is a usual medical procedure to be alert for possible tumor recurrence in the case of detecting a rise in levels of this marker above a certain reference value, our main purpose is to describe the progression of this tumor marker, on patients who were followed and treated in this Unit, as a function of possible risk factors. We intend to estimate on average the time to the increase of this tumor marker, and to characterize the degree of heterogeneity between patients.

2. METHODOLOGY

2.1. Motivation and data set

Data were collected directly from the medical records of each patient, listed in the computer system of Braga's Hospital — Glintt HS. We therefore have access to baseline and clinical history of each patient (a roll of information such as diagnosis; pre-surgery, post-surgery, group meetings; follow-up and medical exams). The authorization to collect and use of senology data was approved by the Ethical Committee of Hospital de Braga.

From the information gathered in the medical reports we were able to collect more than 50 variables that can be grouped into two categories: (i) explanatory variables at individual level, which are a group of demographic characteristics that include a set of prognostic factors reported by Rodrigues (2011) ([14]), for example: age, menopause, age at first full term pregnancy; (ii) explanatory variables at tumor level, that include characteristics of the tumor, some of them important prognostic factors which were already reported in the literature and resumed by Fitzgibbons *et al.* (2000) ([7]) and Cianfrocca and Goldstein (2004) ([3]), such as TNM stage, histological type of tumor, hormonal receptors or vascular or lymphatic invasion, among others.

We collected data from 577 female patients diagnosed with a malignant tumor in the period of 2008 until 2012 (or before, but alive at 2008 and all patients at follow up on group meetings at 2008). Patients at follow up on group meetings were diagnosed as late as 1998. Patients' age at the time of diagnosis varies between 20 and 89 years. However patients with no information regarding tumor markers CEA were excluded for the present analysis, as well patients with no follow up information. We handled all missing values as missing completely at random ([10]).

For the longitudinal analysis of the tumor marker CEA, we considered data of 532 patients. Since 19 patients had bilateral breast cancer, and bilateral breast cancer is treated as independent case in this study, it translates into a total number of 551 cases analyzed. The total number of deaths from breast cancer is 54. There were 4166 measurements of tumor marker CEA, with a number of observations per patient varying between 1 and 23 measurements, as shown in Figure 1. The median number of measurements per person is 7.

It is an unbalanced study for the tumor marker, since patients measurements were not made at the same moment.

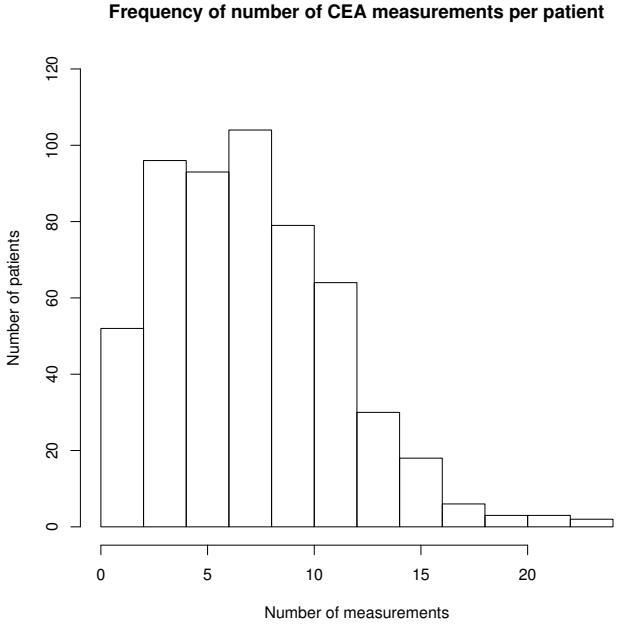

Figure 1: Histogram for the number of measurements per patients for tumor marker CEA.

2.2. Statistical methodology

The response variable, value of CEA, was analyzed making use of longitudinal models as defined in Diggle et al (2002) ([4]), where different correlation structures were tested.

The same covariates used in the survival model, previous adjusted in an earlier study ([2]), were tested in the longitudinal model fitted. The reference time used was time, in years, since diagnose of breast cancer. We have used the reference value of 5,0 ng/mL ([14]) for the response variable. According to the usual medical procedures, physicians stay alert to a possible recurrence of breast cancer for patients that present values of CEA above this reference value.

In general, we denote each patient in this analysis by the subscript $i = 1, ..., n$. Repeated tumor marker measurements for each patient i, at corresponding time t_{ij}, are denoted by Y_{ij}, where $j = 1, ..., m_i$. Note that for this particular study, measurement times are not common to all subjects (unbalanced study). Let $N = \sum_i^n m_i$, be the total number of measurements in the data set.

For the analysis, we began with an exploratory analysis and point estima-

tion by modeling a saturated ordinary least square (OLS) ([4]) model with the variables that had shown significant effect on patients' survival, given by:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} \ , \tag{2.1}$$

where $E[Y_{ij}] = \mu_{ij}$ and ε_{ij} are N independent and identically distributed (i.i.d.) realizations of $N(0, \xi^2)$.

Since the OLS model assumes independence between any two measurements, from the same or different subject, it is important to consider different models in the context of longitudinal analysis, that take into account the correlation that usually exists in the measurements of the same subject.

A longitudinal model was also fitted only to the subset of patients that died, using the reference time, in years, from blood tests until date of death.

To model the correlation structure for each model we analyzed the empirical variogram of OLS residuals from the saturated model for the mean response ([4]). These patterns suggested the existence of variability between subjects (as random effects), and a possible variability within subjects (serial correlation). Hence, maintaining the same mean structure we compared two nested models with different covariance structures with three components, such as: (i) random effects, exponential serial correlation and measurement error; (ii) random effects, Gaussian serial correlation and measurement error.

In many medical studies it is important to consider not only random effects but also a possible variability within subjects as it may have important medical implications. In fact, Liang and Zeger (1986) ([9]) alert that treating the correlation as a nuisance may be less appropriate when the time course of the outcome for each subject is of primary interest or when the correlation itself has scientific relevance.

Both longitudinal models are given by:

$$Y_{ij} = \mu_{ij} + U_i + W_i(t_{ij}) + Z_{ij} \ , \tag{2.2}$$

where U_i are n i.i.d. realizations of $N(0, \nu^2)$, representing the random effects at individual level, $W_i(t_{ij})$ is a continuous time Gaussian Process with $E[W_i(t_{ij})] = 0$ and $\mathrm{Var}(W_i(t_{ij})) = \sigma^2$ and, Z_{ij} are N i.i.d. realizations of $N(0, \tau^2)$, representing the measurement error (variability non specified).

To model the fixed term of the longitudinal model, μ_{ij}, we have tested for a changing point δ on the effect of time on the tumor markers. In practice, the changing point is the moment where there is an alteration on the slope of the linear marker's progression, on average. Considering δ the changing point, we

have $E[Y_{ij}] = \mu_{ij}$ with:

$$\mu_{ij} = \begin{cases} X_{ij}\beta + \alpha_1 t_{ij}\,, & \text{if } t_{ij} < \delta\ , \\ X_{ij}\beta + \alpha_2(t_{ij} - \delta)\,, & \text{if } t_{ij} \geq \delta\ , \end{cases} \tag{2.3}$$

where X_{ij} represents the vector of covariates, β the vector of unknown regression coefficients, α_1 and α_2 the coefficients representing the slope before and after the changing point, respectively.

For parameter estimation we use the maximum likelihood method, whose associated likelihood function is given by:

$$L(\theta; Y) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} \frac{1}{2\pi|V_{ij}|} \exp\left\{-\left(\frac{1}{2}\right)(y_{ij} - \mu_{ij})V_{ij}^{-1}(y_{ij} - \mu_{ij})^T\right\}\ , \tag{2.4}$$

where V_{ij} is the variance/covariance positions on the variance/covariance matrix of all data.

We then conducted a backward elimination to delete variables not significant, until the mean structure was well defined with only significant covariates.

What distinguishes these two longitudinal models is how two different realizations of W_i are correlated in time. That is, if we consider the correlation among $W_i(t_{ij})$, let say between $W(t)$ and $W(t-u)$, determined by the autocorrelation function $\rho(u)$, we will have for the REE model $\rho(u) = \exp(-\frac{1}{\phi}.|u|)$, and for the REG model $\rho(u) = \exp(-\frac{1}{\phi}.u^2)$, where ρ is the range parameter that specifies the rate at which the correlation stables.

The validation of the correlation structure was made by graphical comparison between the empirical variogram and the theoretical fitted ones, and comparing their maximized log likelihood values.

The variogram ([4]) of a stochastic process $Y(t)$ is given by:

$$V(u) = \frac{1}{2}\,\mathrm{Var}\big\{Y(t) - Y(t-u)\big\}\ , \qquad u \geq 0\ . \tag{2.5}$$

For a stationary process, the autocorrelation function, $\rho(u)$, and the variance of $Y(t)$, σ^2, are related by:

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}\ . \tag{2.6}$$

The estimation of the empirical variogram is based on the calculation of the observed half-squared-differences between pair of residuals, $\nu_{ij} = \frac{1}{2}(r_{ij} - r_{ik})^2$, and the corresponding time-differences, $u_{ijk} = t_{ij} - t_{ik}$, where $r_{ij} = Y_{ij} - \mu_{ij}$, and $j < k = 1, ..., m_i$.

The autocorrelation function at any lag u is estimated from the sample variogram by:

$$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2} \ , \tag{2.7}$$

where $\hat{\gamma}(u)$ is the average of all the ν_{ij} corresponding to that particular value of u, and $\hat{\sigma}^2$ is the estimated process variance.

The entire analysis was performed using R software ([15]), in particular making use of both *nlme* ([11]) and *JoineR* ([13]) packages.

3. RESULTS

Since the normality assumption of the response variable failed, we used a log-transformation of the tumor marker CEA values. It is, in fact, a usual transformation in biological markers. The spaghetti plot (Figure 2) presents the progression of the CEA values for each patient, against the reference, and the non parametric smooth spline line, indicating the average trend of progression. The smooth spline suggests that, on average, the marker progression stays below the reference value with a non accentuated slope in its increase. However, it is possible to see that there are individuals with values above the reference value of log (5.0) ng/mL. Nevertheless a linear modeling approach appears to be reasonable. Also, it does not point out to the existence of a changing point in its progression in time.

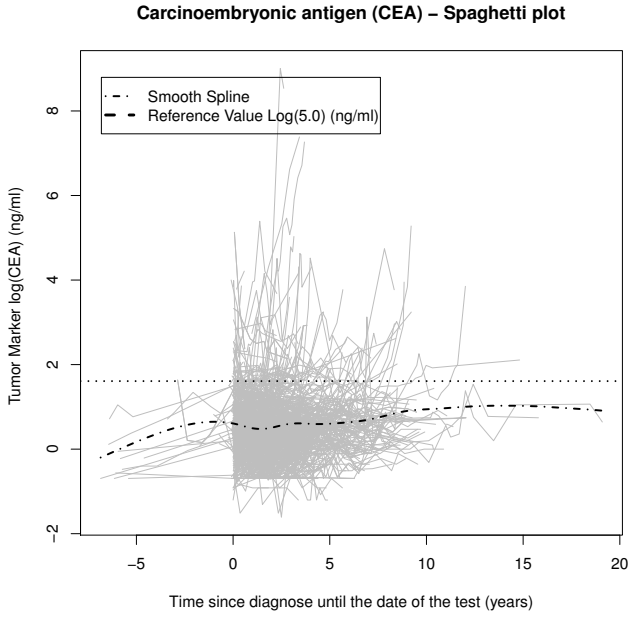

Figure 2: Spaghetti plot for tumor marker CEA values.

In fact, after fitting several saturated parametric models considering various changing points' values, its existence was not significant in the mean time trend of the tumor progression.

Table 1 presents the estimated parameters of the fitted longitudinal model that best represent the tumor marker progression in time, and compares the estimates to those obtained by fitting the simple OLS model and the respective log Likelihood values.

Table 1: Estimated parameters values for General Linear Model and Longitudinal Model.

	OLS Model		**REE Model**	
	Estimate	**p-value**	**Estimate**	**p-value**
Intercept	0.0689	0.7170	0.7355	0.0405
Time	−0.1304	<0.0001	−0.1049	0.0038
Tumor stage (III or IV)	0.2132	<0.0001	0.2655	0.0038
Primary tumor size (Tx or T1 or T2 or T3 or Tis)	−0.2063	0.2660	−1.0383	0.0023
Age at diagnosis	0.0095	<0.0001	0.0117	<0.0001
Venous vascular invasion (Yes) * Time	0.1355	<0.0001	0.0967	0.0175
Tumor degree (G3) * Time	0.1281	<0.0001	0.1179	<0.0001
Estrogen receptor expression (positive) * Time	0.1548	<0.0001	0.1455	<0.0001
$\hat{\nu}^2$			0.2849	
$\hat{\sigma}^2$			0.3295	
$\hat{\phi}$			2.1912	
$\hat{\tau}^2$			0.0239	
$\hat{\xi}^2$	0.6770			
Log Likelihood	−3792.429		−1853.366	

The fixed part of the longitudinal model, which describes the mean progression of the marker, is composed by the following significant covariates on the intercept component of the model: tumor stage (0/I/II versus III/IV), primary tumor size (Tx/T1/T2/T3/Tis versus T4), and age at diagnosis. The intercept component of the model, in this particular case, means that a patient with a tumor stage of 0, I or II, a T4 primary tumor size at an earlier age of diagnosis will start the progression of the tumor marker with a value of 0.7355, on a logarithmic scale.

A patient with a tumor on stage III or IV implicates an increasing of the log value of the tumor marker by an increment of 0.2655, comparing to those

with a tumor on stage 0, I or II. Also, a tumor that presented a primary tumor size different from the classification T4 has a decrease in the starting point of the marker value by an increment of -1.0383. The age at diagnosis affects the log value of the marker at a rate of 0.0117 per year of age at diagnosis.

The covariates that affect the slope (-0.1049) of the linear progression of the tumor are: images of vascular invasion (Yes versus No), Bloom-Richardson degree of differentiation (Gx/G1/G2 versus G3) and estrogen receptors expression (Positive versus Negative).

According to the estimated values, cases that present a venous vascular invasion of the tumor, a tumor degree G3 and a positive estrogen receptor expression increase the progression slope at a rate of, respectively, 0.0967, 0.1179 and 0.1455.

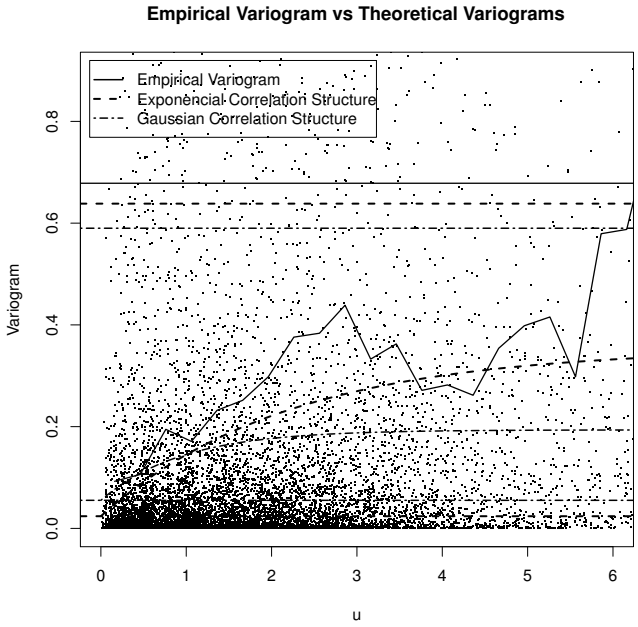

Figure 3: Superposition of empirial variogram and theoretical variogram.

The correlation structure that best represents the variability of the data is, in fact, the one that incorporates random effects at individual level with $\hat{\nu}^2 \approx 0.2849$, an exponential correlation structure to describe the variability within patients with $\rho(u) = \exp(\frac{-1}{2.1912}.|u|)$ and $\hat{\sigma}^2 \approx 0.3295$, and a measurement error with variance $\hat{\tau}^2 \approx 0.0239$. That fact can be easily accessed by the superposition of the theoretical fitted variogram of both exponential and Gaussian correlation structures with the empirical variogram (Figure 3).

When fitting the saturated general linear model for the subset of patients who died from breast cancer, we detected a changing point at 2 years before death. The smooth spline of the spaghetti plot (Figure 4) is consistent with that result and informs a transposition of the reference value nearly after that changing point.

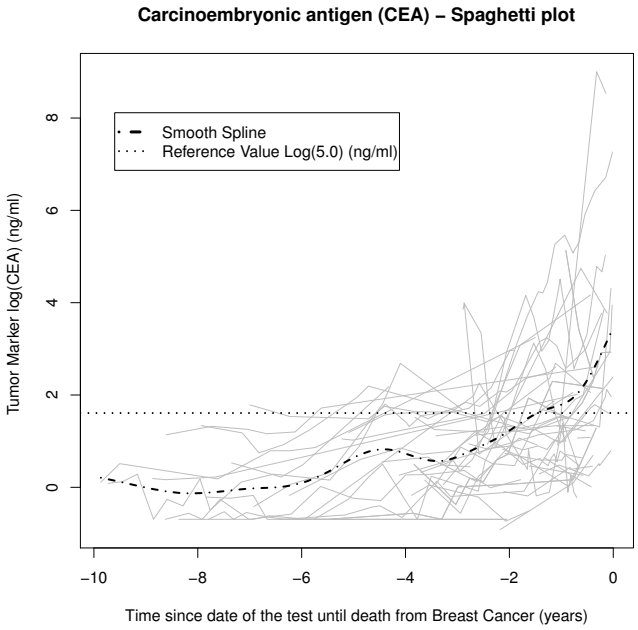

Figure 4: Spaghetti plot for tumor marker CEA values of patients who died from breast cancer.

Note that, as we are analyzing the marker values from date of blood tests until death, we are dealing with duration at a negative scale.

Table 2 summarizes and compares the estimated parameters for the longitudinal model which best fitted the data with those of the general linear model (OLS Model). As expected, the presence of venous vascular invasion has an increasing effect on the average CEA linear progression in time, as it is related to a worst prognostic case in the previous survival analysis ([2]).

Contradictory results are the decreasing effect of a bilateral type of tumor and the presence of lymphatic invasion and the increasing effect of a positive estrogen and HER-2neu expression. The mentioned covariates have a statistical significant effect on the intercept component of the model (1.4622). Bilateral cancer cases have a decrease of 0.5981 on the intercept component, and a case with lymphatic invasion a decrease of 0.7322 compared to those with no lymphatic invasion. A case that presents images of vascular invasion increases of the start value of the tumor marker by an increment of 0.7322, comparing to those that do

not present any image. A positive estrogen receptor expression has an increasing effect on the intercept component by 1.2177, compared to those with a negative expression. A positive expression of HER-2neu has an increment of 0.4882.

Table 2: Estimated parameters values for General Linear Model and Longitudinal Model, for the patients who died from breast cancer.

	OLS Model		**REG Model**	
	Estimate	**p-value**	**Estimate**	**p-value**
Intercept	2.0376	<0.0001	1.8507	<0.0001
Time before changing point (2 years before death)	0.2540	<0.0001	0.2128	<0.0001
Time after changing point (2 years before death)	0.9453	<0.0001	0.8815	<0.0001
Bilateral (Yes)	−0.9290	<0.0001	−0.5981	0.0471
Lymphatic invasion (Yes)	−0.8821	<0.0001	0.7769	<0.0001
Venous vascular invasion (Yes)	1.0350	<0.0001	0.7769	0.0266
Estrogen receptor expression (positive)	1.5675	<0.0001	1.2177	<0.0001
$\hat{\nu}^2$			0.2404	
$\hat{\sigma}^2$			0.8239	
$\hat{\phi}$			0.3762	
$\hat{\tau}^2$			0.0415	
$\hat{\xi}^2$	1.2499			
Log Likelihood	−1089.503		−621.695	

For this subset, the correlation structure that best represent the variability of the data is the structure that incorporates random effects at individual level with $\hat{\nu}^2 \approx 0.2404$, a Gaussian correlation structure to describe the variability within patients with $\rho(u) = \exp(\frac{-1}{0.3762.}u^2)$ and $\hat{\sigma}^2 \approx 0.8239$, and a measurement error with variance $\hat{\tau}^2 \approx 0.0415$. The superposition of the theoretical variogram of both exponential and Gaussian correlation structures with the empirical variogram (Figure 5) validates the choice of an exponential correlation structure.

Both REE and REG models were compared with a longitudinal model only with an intercept random effect component U_i, and the serial correlation component $W_i(t_{ij})$ shown to be significant in the models. This result reinforces the need to take into account correlation within subject measurements.

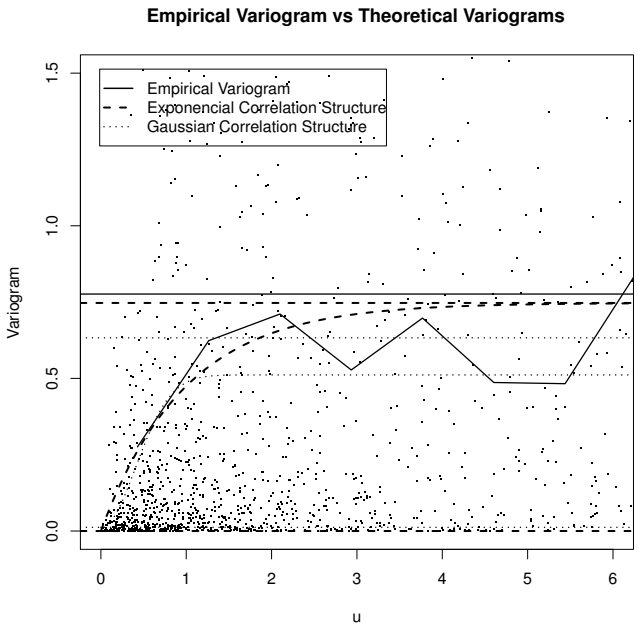

Figure 5: Superposition of empirial variogram and theoretical variogram, for patients who died from breast cancer.

4. DISCUSSION

An abrupt rise in values of CEA tumor marker, over a reference value, is an alert sign to a possible recurrence of breast cancer.

When analyzing all patients that were diagnosed with breast cancer, in our study, the only variables that have a statistically significant effect on the linear progression of the tumor marker are: tumor stage (III/IV versus 0/I/II), primary tumor size (Tx/T1/T2/T3/Tis versus T4), age at diagnosis, venous vascular invasion (Yes versus No), tumor degree (Gx/G1/G2 versus G3) and estrogen receptor expression (positive versus negative). As expected, a III or IV tumor stage, a T4 type of tumor, a G3 type of tumor, the presence of venous vascular invasion and age at diagnosis have an increasing effect on the average tumor marker progression in time, as they are related to a worst prognostic case ([2]). One unexpected result was the fact that a positive expression of the estrogen receptor has an increasing effect on that progression, contradicting the results from a previous survival analysis ([2]), where the same patients' cases of positive estrogen receptor shown a lower probability of dying from breast cancer than those who presented a negative expression.

It was detected a changing point on the linear progression of the tumor marker for the subset of patients that died from breast cancer two years before the death. This means that, at that point, there is an abrupt rise on the rate of its progression.

The risk factors for the progression of the marker, for that subset of patients are: bilateral (Yes versus No), lymphatic invasion (Yes versus No), venous vascular invasion (Yes versus No), estrogen receptor expression (positive versus negative) and HER-2neu expression (positive versus negative). As expected, the presence of venous vascular invasion has an increasing effect on the average CEA linear progression in time, as it is related to a worst prognostic case in the previous survival analysis ([2]). A bilateral type of tumor and the presence of lymphatic invasion have a decreasing effect. A positive estrogen and HER-2neu expression has an increasing effect. These two last results contradict the results from the previous survival analyses ([2]) since bilateral cases and lymphatic invasion are related to lower survival probability and, a positive estrogen and Her-2neu expression are both related to a higher probability of survival.

For both models fitted, the fact that the estimated variance of the measurement error is quite lower that the estimated variance of the OLS model, means that the fitted REE longitudinal model explains the variability of the data mainly by means of variability between patients and within patients assigning a very low value for measurement error (or *white noise* as usually mentioned in literature).

The fact that, when comparing the REE and the REG models to a longitudinal model with only an intercept random effect, the component the serial correlation was significant, stresses the importance incorporating a variability component that translates within subject measurements correlation, in this type of biological data.

The presented longitudinal analysis of this tumor marker, in combination with the previous survival analysis is going to be proceeded, in future work, with a joint modeling of the longitudinal and survival process of the present data.

ACKNOWLEDGMENTS

This work is funded by FEDER funds through the Operational Program for Competitiveness Factors – COMPETE and National Funds through FCT – Foundation for Science and Technology under the project EXPL/MAT-STA/0313/2013. The author Ana Borges has a PhD scholarship from FCT SFRH/BD/74166/2010.

REFERENCES

[1] Boyle, P. and Ferlay, J. (2004). Cancer incidence and mortality in Europe, *Annals of Oncology*, **16**, 3, 481–488.

[2] Borges, A.; Sousa, I. and Castro, L. (2013). *Breast cancer survival at Braga's hospital – Portugal*. In "Theoretical and Applied Issues in Statistics and Demography" (C.H. Skiadas, Ed.), Book of Abstracts - ASMDA2013, 34–35.

[3] Cianfrocca, M. and Goldstein, L.J. (2004). Prognostic and predictive factors in early-stage breast cancer, *The Oncologist*, **9**, 6, 606–616.

[4] Diggle, P.J.; Heagerty, P.; Liang K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, University Oxford Press.

[5] *European Cancer Observatory*, (http://eco.iarc.fr/).

[6] Fiorella, G.; Ferroni, P.; Carlini, S.; Mariotti, S.; Spila, A.; Aloe, S.; D'Alessandro, R.; Carone, M.D.; Cicchetti, A.; Ricciotti, A.; Venturo, I.; Perri, P.; Filippo, F.; Cognetti, F.; Botti, c. and Roselli, M. (2001). A re-evaluation of carcinoembryonic antigen (CEA) as a serum marker for breast cancer: a prospective longitudinal study, *Clinical Cancer Research*, **7**, 8, 2357–2362.

[7] Fitzgibbons, P.L.; Page, D.L.; Weaver, D.; Thor, A.D.; Allred, D.C. and Clark, G.M. (2000). Prognostic factors in breast cancer, College of American Pathologists Consensus Statement 1999, *Archives of Pathology & Laboratory Medicine*, **124**, 7, 966–978.

[8] Harris, L.; Fritsche, H.; Mennel, R.; Norton, L.; Ravdin, P.; Taube, S.; Somerfield, MR.; Hayes, D.F. and Bast, R.C. Jr (2007). American Society of Clinical Oncology: American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer, *Journal of Clinical Oncology*, **25**, 5287–5312.

[9] Liang, K.Y. and Zeger, S.L. (2007). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

[10] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd Edition, Wiley-Interscience.

[11] Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, D. and R Core Team (2014). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-118, http://CRAN.R-project.org/package=nlme.

[12] Pinheiro, P.S.; Tyczynski, J.E.; Bray, F.; Amado, J.; Matos, E. and Parkin, D.M. (2003). Cancer incidence and mortality in Portugal, *European Journal of Cancer*, **39**, 17, 2507–2520.

[13] Philipson, P.; Sousa, I.; Diggle, P., Williamson, P.; Kolamunnage-Dona, R.; Henderson, R. and R Core Team (2014). JoineR: Joint modelling of repeated measurements and time-to-event data. R package version: 1.0-3, http://CRAN.R-project.org/package=joineR.

[14] Rodrigues, V. (2011). *Chapter 34*. In "Manual de Ginecologia", Permanyer, Portugal, 175–191.

[15] R Development Core Team (2008). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org.

[16] Sturgeon, C.M.; Duffy, M.J.; Stenman, U.H.; Lilja, H.; Brünner, N.; Chan, D.W.; Babaian, R.; Bast, R.C. Jr.; Dowell, B. and Esteva, F.J. (2008). National Academy of Clinical Biochemistry *et al.*: National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast and ovarian cancers, *Clinical Chemistry Journal*, **54**, e11–e79.

REVSTAT – Statistical Journal
Volume 13, Number 1, March 2015, 79–101

ALCOHOL ABUSE DISORDER PREVALENCE AND ITS DISTRIBUTION ACROSS PORTUGAL. A DISEASE MAPPING APPROACH

Authors: Helena Baptista
– NOVA IMS, Universidade Nova de Lisboa, Lisboa, Portugal
d2011073@novaims.unl.pt

Jorge M. Mendes
– NOVA IMS, Universidade Nova de Lisboa, Lisboa, Portugal
jmm@novaims.unl.pt

José Caldas de Almeida
– Department of Mental Health, NOVA Medical School, Portugal
jcaldasalmeida@gmail.com

Miguel Xavier
– Department of Mental Health, NOVA Medical School, Portugal
migxavier@gmail.com

Received: October 2014 Revised: November 2014 Accepted: December 2014

Abstract:

- Disease mapping is linked to two other scientific areas: small area estimation and ecological-spatial regression. This paper reviews similarities and differences among them. Bayesian hierarchical models are typically used in this context, using a combination of covariate data and a set of spatial random effects to represent the risk surface. The random effects are typically modeled by a conditional autoregressive prior distribution, and a number of alternative specifications have been proposed in the literature. The four models assessed here are applied to a study on alcohol abuse in Portugal, using data collected by the World Mental Health Survey Initiative.

Key-Words:

- *alcohol abuse; Bayesian hierarchical models; disease mapping; generalized linear models; small area estimation.*

AMS Subject Classification:

- 62F15, 62J12, 62M30, 62P10.

1. INTRODUCTION

The availability of disease data in sets of non-overlapping and contiguous spatial areal units has increased over the last few decades. Concepts such as small area estimation (SAE), disease mapping (DM) and ecological-spatial regression (ESR) are linked and are used in the context of the analysis of this type of data.

The purpose of this work is two-fold; first we clarify those concepts, and second, after focusing on DM, we apply several models to Portuguese alcohol abuse disorder (AAD) data, collected by the World Mental Health Survey Initiative (WMHSI), as specified in [39]. Harmful use of alcohol was considered by the World Health Organization (WHO) as one of the world's leading risk factors for disease and disability ([38]).

The goal of DM is to estimate the spatial pattern in disease risk over a geographical region, so that small areas with elevated risk can be identified. This term was first used in [5]. It uses the spatial setting and assumes positive spatial correlation between observations, essentially 'borrowing' more information from neighboring areas than from areas far away, smoothing local rates toward local neighboring values ([37]).

The remainder of this paper is organized as follows. Section 2 introduces the Portuguese data, as well as some background information on AAD. Section 3 provides the DM definition highlighting the differences and common aspects among DM, SAE and ESR. Section 4 deals with the most common and widely used models for DM, providing some basic information on those, as well as some challenges and recent methodological advances. Section 5 contains the results of the models, reviewed in Section 4, applied to the data defined in Section 2. Finally, Section 6 contains a concluding discussion and areas of future work.

2. DATA

The WMHSI was administered at the households of a nationally representative sample of respondents, between October 2008 and December 2009. The target population for the survey was defined as the resident, non-institutionalized, Portuguese-speaking population of the Portuguese mainland, aged 18 or above, residing in permanent private dwellings. Details regarding the design, target population, sampling, tools, measures, fieldwork organization, procedures, and weighting are reported in detail elsewhere ([39]). This is a cross-sectional study, meaning that both disease cases and possible risk factors are collected at the same time. As reported in [36] that restricts the conclusions that can be drawn from

the models. It is not possible to establish causal relationships between disease cases and possible covariates.

Data collected by cross-sectional studies may have several types of biases. In the present case the possibility of selection bias is particularly evident, as only the non-institutionalized population and the population above 18 years of age was selected, and accordingly to the WHO ([38]) the alcohol consumption is rising between adolescents (13–18 years of age) and young adults. Therefore inferences can only be made on the study population and not on the global Portuguese population. Another possible common bias is the misclassification bias, *i.e.*, the incorrect assignment of a disease to the study participants. This type of bias may occur in studies like this one, because there is no intervention of a medical doctor during the questionnaire's self-administration. This problem seems to have been solved in France, Italy, Spain and United States of America, as [14] and [10] provide evidence that the diagnoses of substance abuse disorders identified by the questionnaire used in this initiative, the CIDI 3.0 (Composite International Diagnoses Interview) have generally good concordance with diagnoses based on blinded clinical reappraisal interviews. Unfortunately those tests have not been conducted in Portugal. Although the alcohol consumption and related disorders are very much connected with cultural aspects ([28]), we think that the performance while identifying the actual presence of disease has not been seriously affected.

According to the WHO ([38]) approximately 5.1% of the global burden of disease, and 5.9% of all deaths worldwide are attributable to alcohol consumption. Furthermore, harmful use of alcohol inflicts significant social and economic losses on individuals and society at large.

In accordance with the DSM-IV ([8]) criteria there are two possible diagnoses of alcohol disorders, the alcohol abuse disorder and the alcohol dependence disorder. In the six European countries ([1]) covered by the ESEMeD project[1], 5.2% of the respondents report a lifetime history of alcohol abuse and/or dependence disorders. In the WMHSI, lifetime and 12-month alcohol disorder diagnoses are provided. From the data collected in Portugal the prevalence rate of a lifetime history of alcohol abuse and/or dependence disorders is 10.0%, while the last 12-month prevalence rate is 1.6%; the lifetime prevalence rate of alcohol dependence disorder is 1.3%, while the last 12-month prevalence rate is 0.26%; the lifetime prevalence rate of alcohol abuse disorder is 8.7%, whereas the last 12-month prevalence rate is 1.3%. The high prevalence of alcohol abuse disorder found in Portugal reiterates the need to maintain alcohol abuse as a public health priority in the country, and therefore more detailed studies are needed.

[1]The ESEMeD Project was created to fully study the results of the WMHSI on the following countries: Belgium, France, Germany, Italy, the Netherlands and Spain. As Portugal joined the WMHSI later than others, most of the publications, including the [1], do not include Portuguese results.

The study region is mainland Portugal partitioned into 28 units called NUTS 3[2], corresponding to the 3rd level territorial units aggregation. There are 30 NUTS 3 in Portugal, from which 28 are in mainland Portugal and 2 are in the Islands. The *response variable* is the number of lifetime AAD cases per NUTS 3. Differences in the size and demographic structure of the population living in each NUTS 3 are accounted for by computing the expected number of AAD cases using indirect internal standardization, based on age specific AAD rates for the whole study region.

The age standardization process, as defined in [36], can be direct or indirect. The choice between direct and indirect standardization is usually defined by the type of data available. Age-specific rates for the disease at each NUTS 3 are not available and therefore the indirect method is used, by applying the age-specific disease rate for the global population to the NUTS 3 age-specific population, provided by the Portuguese Statistics organization for the year of 2008. As this standardization is done using the age-specific disease rate for the global population, as it was collected by the survey itself, the standardization is internal (external standardization only occurs when standard tables of age-specific rates for the disease are available). As mentioned in [2] internal standardization is 'cheating' in some sense, since 'a degree of freedom is lost' by estimating the age-specific disease rate from the current data.

Accordingly, the following notations and/or definitions are introduced:

a) Y_k the random variable representing the number of observed cases (y_k) in each k age group;

b) n_k representing the number of people at risk in each k age group;

c) $r_k = \frac{y_k}{n_k}$ representing the observed prevalence proportion for each k age group;

d) n_{ik} representing the number of people at risk in each k age group in the i^{th} NUTS 3;

e) E_{ik} and y_{ik} representing the expected and observed number of cases for the k age group in the i^{th} NUTS 3, respectively, where $E_{ik} = r_k n_{ik}$;

f) $E_i = \sum_k r_k n_{ik}$ and $y_i^\star = \sum_k y_{ik}$ representing the total number of expected and observed cases in the i^{th} NUTS 3, respectively;

g) $SMR_i = \frac{Y_i^\star}{E_i}$, the *standardized morbidity ratio*, representing the risk of each i^{th} NUTS 3. A value of SMR greater (less) than one indicates that the area i has a higher (lower) than average disease risk. If the $SMR_i = 1.15$, it can be said that area i has a 15% increased risk of the disease.

[2] *Nomenclatura Comum das Unidades Territoriais Estatísticas*, in Portuguese language as defined by Eurostat, the European statistical organization.

Figure 1 shows the raw SMR values for the 28 NUTS 3.

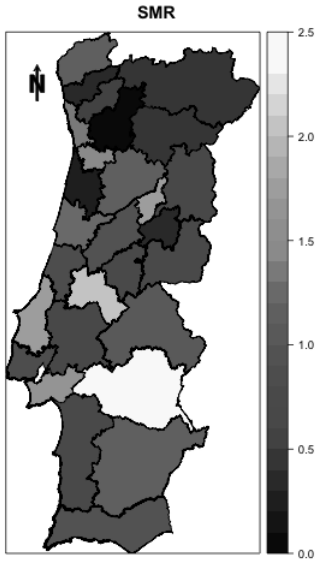

Figure 1: AAD Raw SMRs per NUTS 3. The four regions, which had originally missing values, are shown already with the imputed mean values resulting from the GLM (see Section 5).

Our illustrative example also considers two ecological covariates that are widely known as being associated with the AAD ([13, 18, 27, 33]), which are (a) proportion of population aged 18 to 34, (b) proportion of males. AAD is more prevalent in younger men. These data are only available per NUTS 3, for the year of 2011, as provided by the latest census conducted in Portugal, which we find to be temporally misaligned with WMHSI data used in this work. However as population age and gender structures do not significantly change in 3 years, no corrective measures have been implemented.

3. SMALL AREA ESTIMATION, DISEASE MAPPING AND ECOLOGICAL-SPATIAL REGRESSION

DM joins together three different disciplines: statistics/biostatistics, epidemiology and geography. DM focuses on the challenge of obtaining reliable statistical estimates (statistics/biostatistics) of local disease risk based on counts of observed cases (epidemiology) within small administrative districts or regions (geography) coupled with potentially relevant background information. DM goals are twofold: obtain statistically precise local estimates of disease risk for each region and maintain the regions 'small' in order to keep the geographic resolution.

The areas are not only small in size (relative to the area of the full spatial domain of interest), but are also small in terms of local sample size, resulting in deteriorated local statistical precision. To solve this problem the classical design-based solutions are often infeasible since the local sample sizes within each region, required for the desired level of statistical precision, are often unavailable or unattainable. The model-based approaches can help overcome this problem by the mechanism of 'borrowing strength' across small areas to improve local estimates.

3.1. DM as a special case of SAE

Nowadays sample survey data are extensively used to provide reliable direct estimates of parameters of interest for the whole population. When it comes to getting the same estimates for domains of that population, and due to the small sample sizes in those domains, direct survey estimates are likely to yield unacceptably large standard errors. This makes it necessary to combine survey data collected from the small areas with auxiliary information from sources external to the survey. In this context, named as SAE, several indirect estimators have been extensively used. Some of the most common are the traditional indirect estimators based on implicit models, which include synthetic and composite estimators, and the Empirical Best Linear Unbiased Prediction approach. Most of these approaches also consider a contiguity matrix that describes the neighborhood structure between small areas 'borrowing strength' from related areas to find more accurate estimates for a given area. The works [29] and [6] provide respectively an overview of the foundations of SAE and a comparison between several traditional estimators and some proposed estimators using a Monte Carlo simulation.

DM is a special case of SAE, since the goal is to find reliable statistical estimates of local disease risk. As mentioned by [37] DM refers to a collection of methods extending SAE to directly utilize the spatial setting and assumed positive spatial correlation between observations. The data used are aggregated or averaged values at the small area level, representing disease incidence, prevalence or mortality rates, frequently not coming from surveys but coming from counts of disease cases from hospital admissions ([21, 24]), counts of cancer cases or cancer deaths ([3, 16, 34]), and mortality data ([7, 24, 25]). In the present work we use counts of disease cases from a survey.

3.2. DM and ESR apply the same methodologies to reach different goals

By combining data from administrative registries and/or surveys with auxiliary data, DM goal is to predict area-level outcome summaries, to identify areas

of elevated risk. ESR uses the same type of data and the same methodologies but its objective is the estimation of associations between covariates and the disease cases.

Therefore, two common problems found in ESR are not of a concern in DM: (a) ecological bias and (b) the inclusion of spatially correlated errors changing the association between disease cases and fixed effects.

Ecological bias is the difference between estimated associations on ecological- and individual-level data ([35]). Data used in DM and ESR, both for the number of cases and for the covariates are found rarely at individual-level, mainly due to confidentiality reasons, and therefore the association found at the aggregated level might not be the same if we would have used individual-level data. Aggregated data is usually designated as areal data ([2]). The objective of DM is not to estimate the associations between the cases and the covariates or to improve predictions, and therefore ecological bias is not a concern (for more details on the subject see [35]).

The inclusion of spatially correlated errors, changing the association between disease cases and fixed effects, has been studied by [34] and [12]. Often the study of ESR has provided estimates of the fixed-effect coefficients substantially different from those of ecological regressions. ESR is an ecological regression augmented with the inclusion of random effects modeled by a globally smooth conditional autoregressive model. If the covariates are also globally smooth, collinearity problems might change dramatically the coefficients of the fixed-effects. As before, the coefficients of association are not of direct interest in DM, and therefore this aspect is not a concern.

4. DISEASE MAPPING MODELS

DM methodologies are explained in [37] and [2]. DM methodologies for areal data are usually divided in frequentist methods and hierarchical Bayesian models [2]. To provide a wide comparison of methods, [15] presents some preliminary results concerning the goodness-of-fit of a variety of DM models applied to simulated disease incidence data. These simulated models cover simple risk gradients and more complex true risk structures, including spatial correlation. Authors conclude that full Bayesian hierarchical models are the most robust across a range of diverse models. A number of hierarchical Bayesian models have been proposed in the literature, including the following two, which have been widely used: a) the model developed by Besag, York and Mollié ([3]), from now on designated as BYM model and b) the model developed by Leroux, Lei and Breslow ([22]), from now on designated as LLB model. These two models will be used in Section 5.

Authors of [4] review the main classes of Bayesian models, among which the BYM model is included (but not the LLB model) and conclude that the BYM model has good properties for modeling a single disease and 'appears to be the only fully Bayesian spatial model to have been used in published applications of disease mapping outside of the statistical literature' (page 57). Recently, [24] and [16] published comparisons between hierarchical Bayesian models and both conclude that the LLB model is the best overall, because it produces consistently good results across a range of spatial correlation scenarios, is more parsimonious on parameters, and has less undesirable features (this subject will be further developed in Subsection 4.1).

One of the challenges posed at the DM level arises from its basic goal, the smoothing of local rates toward local neighboring values. When real discontinuities exist between neighboring areas, the models will lead to oversmoothing blurring the edges, which may not be appropriate. If the goal is to identify boundaries or regions of rapid change, the methods of *boundary analysis* or *wombling* need to be applied. For more detail see the recent works of [19] and [20].

A general formulation for the first level of the hierarchical Bayesian models used in DM is given by

$$Y_i|E_i, R_i \sim Poisson(E_i R_i) \qquad \text{for} \quad i = 1, ..., n \ ,$$

$$\ln(R_i) = \mu + \mathbf{x_i^T}\boldsymbol{\beta} + \phi_i \ . \tag{4.1}$$

If E_i is not too large (as it is the case of rare diseases) or the regions i are sufficiently small, the usual model for the Y_i is a Poisson model ([2]). In the model, R_i denotes the risk of disease in area i, which is modeled by an intercept term μ, a set of p covariates $\mathbf{x_i^T} = (x_{i1}, ..., x_{ip})$ and a random effect ϕ_i. The random effects are included to model any overdispersion and/or spatial correlation that might remain in the data after have being accounted for by the included covariate information. Most studies of this type show overdispersion, meaning that $Var[Y_i] > \mathbb{E}[Y_i]$, which has several possible causes: subject heterogeneity; correlation between individual responses; omitted unobserved variables; and/or excess zero counts. Inference for this type of model is based on Markov chain Monte-Carlo (MCMC) simulation, using a combination of Gibbs sampling and Metropolis-Hastings steps and more recently using Integrated nested Laplace approximations ([31]).

The random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are usually modeled by the class ([2]) of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model ([11]). Instead of a specification of a single multivariate distribution $f(\boldsymbol{\phi})$, the above models are specified by a set of univariate full conditional distributions $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. To determine the spatial correlation between the random effects, we use the neighborhood matrix

W, which is a binary $n \times n$ matrix, with elements w_{ji}:

$$w_{ji} = \begin{cases} 1, & \text{if } j \sim i\,, \\ 0, & \text{otherwise}\,, \end{cases}$$

where $j \sim i$ represents contiguous areas, and therefore j and i are considered neighbors. Other *adjacency-based* weights are available but are much less widely applied ([37]). If two areas are neighbors we believe their random effects are correlated, while non-neighboring areas are modeled as being conditionally independent given the remaining elements of $\boldsymbol{\phi}$.

4.1. BYM model

The BYM model combines the intrinsic CAR (ICAR) with an additional set of independent random effects.

The full conditional distributions of ICAR, as proposed by [3] are given by

$$u_i|\boldsymbol{u}_{-i}, \sigma^2 \sim N\left(\frac{1}{n_i}\sum_{j\sim i} u_j, \frac{\sigma^2}{n_i}\right). \tag{4.2}$$

The conditional expectation of u_i is equal to the mean of the random effects in neighborhood areas, while the conditional variance is inversely proportional to the number of neighbors n_i. The variance parameter σ^2 controls the amount of variation between the random effects. The ICAR model has three main drawbacks:

1) Its simplicity turns it into a very restrictive prior. Its single parameter does not determine the strength of the spatial correlation (for example multiplying each u_i by 10, will only increase σ^2 leaving the spatial correlation unchanged). If data are weakly correlated, the ICAR is not the most appropriate model ([16]).

2) The joint distribution for $f(\mathbf{u})$ corresponding to (4.2) is improper (it does not determine a legitimate probability distribution, one that integrates to 1). Nevertheless, this is easily solved by enforcing a constrain such as, $\sum_{j=1}^{n} u_j = 0$, which can be *numerically* imposed by recentering each sampled **u** vector around its own mean following each Gibbs iteration ([2]).

3) According to [24] the ICAR has an undesirable *global* (*i.e.* large-scale) property of tending to a negative pair-wise risk dependence as the 'spatial proximity' of the two regions is further apart.

The BYM model defines ϕ in (4.1) by

$$\phi_i = \theta_i + \psi_i \ , \tag{4.3}$$

$$\theta_i|\sigma_\theta^2 \sim N(0, \sigma_\theta^2) \ ,$$

$$\boldsymbol{\psi} = (\psi_1, ..., \psi_n)|\mathbf{W} \ , \qquad \sigma_\psi^2 \sim ICAR(\mathbf{W}, \sigma_\psi^2) \ ,$$

where $\mathbf{W}$ is defined in Section 4). More details on the BYM model are provided by [21] and [3].

The set of random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ is independent between areas. Different strengths of spatial correlation can be represented by varying the relative sizes of the two components $(\boldsymbol{\theta}, \boldsymbol{\psi})$. In practice, it will often be the case that either $\boldsymbol{\theta}$ or $\boldsymbol{\psi}$ dominates the other depending upon the strength of the spatial structure and the relative sizes of σ_θ^2 and the σ_ψ^2. This flexibility is also a disadvantage, as each data point is represented by two random effects while only their sum $(\theta_i + \psi_i)$ is identifiable. In order to attain model identification and achieve convergence when MCMC is used, at least one considerably informative hyper prior has to be assumed either for σ_θ^2 or σ_ψ^2. Several authors have studied this aspect ([37, 2]), and ([24]) implemented a model that can 'attain model identifiability, allow the data to inform risk decomposition, and facilitate principled attribution of the relative risk variability to spatially varying *clustering* effects and randomly varying *heterogeneity* effects based on the *given data*' (page 66), hereafter called Modified BYM (MBYM). This model replaces (4.3) by

$$\boldsymbol{\phi} = \sqrt{\lambda}\boldsymbol{\psi} + \sqrt{1-\lambda}\boldsymbol{\theta} \ , \qquad \boldsymbol{\psi} \perp \boldsymbol{\theta}, \quad \lambda \in (0, 1) \ . \tag{4.4}$$

One interpretation of the above is that it represents a re-parameterized BYM prior with $\sigma_\psi^2 = \lambda\sigma^2$ and $\sigma_\theta^2 = (1-\lambda)\sigma^2$. The new prior interpolates between the ICAR prior and the Gaussian prior for θ. λ serves as a spatial smoothing parameter and determines the proportion of the spatially structured risk variability over the total risk variability.

4.2. LLB model

The LLB model is based on a single set of random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, represented by a multivariate Gaussian distribution

$$\boldsymbol{\phi}|\mathbf{W}, \sigma^2, \rho, \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2[\rho\mathbf{W}^* + (1-\rho)I_n]^{-1}) \ . \tag{4.5}$$

The prior above has a constant non-zero mean $\boldsymbol{\mu} = (\mu, ..., \mu)$, avoiding the use of the intercept term in (4.1). In the matrix, $\sigma^2[\rho\mathbf{W}^* + (1-\rho)I_n]^{-1}$, I_n is an

$n \times n$ identity matrix and the elements of $\mathbf{W}^*$ are equal to

$$w_{ji}^* = \begin{cases} n_i, & \text{if } j = i\,, \\ -1, & \text{if } j \sim i\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

The precision matrix is a weighted average of the spatially dependent correlation structures, represented by the matrix $\mathbf{W}^*$, the independent correlation structures, represented by the identity matrix, and the weight represented by the parameter ρ. When $\rho = 0$ the model becomes a simple independent random effects model and when $\rho = 1$ the model becomes the ICAR as in (4.1). When $0 \leqslant \rho < 1$ the joint distribution (4.5) is proper. The full conditional distributions corresponding to (4.5) are given by

$$\phi_i|\boldsymbol{\phi}_{-i}, \mathbf{W}, \sigma^2, \rho, \mu \sim N\left(\frac{\rho \sum_{j\sim i} \phi_j + (1-\rho)\mu}{n_i\rho + 1 - \rho}, \frac{\sigma^2}{n_i\rho + 1 - \rho}\right). \tag{4.6}$$

The conditional expectation is the weighted average of the random effects in the neighboring areas and the overall mean μ. The conditional variance, in the presence of strong spatial correlation is approximately σ^2/n_i, the same as the ICAR, but if the random effects are independent then it is a constant (σ^2).

4.3. Localized conditional autoregressive model

All three models defined above use CAR priors that are globally smooth. The random effects are forced to exhibit a single global level of spatial smoothness determined only by geographical adjacency. With real data such a uniform level of smoothness for the entire region is unrealistic. It is more realistic to think that sub-areas of spatial autocorrelation co-exist with areas of discontinuity. As an example, areas of wealth and poverty, sharing boundaries, are very common in the biggest cities of the world, showing different patterns in the disease risk. A possible solution to this problem is presented by [21], and is called Bayesian localized conditional autoregressive model, LCAR from now on. This model was initially applied to a ESR, but as explained in Subsection 3.2 the same methodology can be applied in the DM field.

The LCAR treats the elements in the neighborhood matrix, representing contiguous areas, as a set of binary random quantities and not as fixed values. The elements of this new neighborhood matrix, $\tilde{\boldsymbol{W}}$, continue to be set to zero for non adjacent areas but adjacency is no longer the only reason for those elements to be set to one. When all adjacencies are kept, the model simplifies to the ICAR, while if all adjacencies are removed the random effects are independent.

The model defines $\boldsymbol{\phi}$ in (4.2) as $\tilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}, \phi_{\circledast})$ where $\phi_{\circledast}$ is a global random effect that is potentially common to all areas and prevents any unit from having no information to 'borrow strength' from. Based on the extended matrix, the proposal is to model $\tilde{\boldsymbol{\phi}}$ as $\tilde{\boldsymbol{\phi}} \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}(\tilde{\mathbf{W}}, \epsilon)^{-1})$, with the precision matrix given by

$$\mathbf{Q}(\tilde{\boldsymbol{W}}, \epsilon) = \text{diag}(\tilde{\boldsymbol{W}}\boldsymbol{I}) - \tilde{\boldsymbol{W}} + \epsilon \boldsymbol{I} \ , \tag{4.7}$$

The component $\text{diag}(\tilde{\boldsymbol{W}}I) - \tilde{\boldsymbol{W}}$ corresponds to the ICAR model applied to the extended random effects vector $\tilde{\boldsymbol{\phi}}$ and the component ϵ ensures that the matrix is diagonally invertible. This restriction is now needed because $\mathbf{Q}(\tilde{\boldsymbol{W}})$ is no longer fixed. The parameter ϵ is recommended to be set as $\epsilon = 0.001$. The full conditional distributions corresponding to the LCAR model are given by

$$\phi_j | \boldsymbol{\phi}_{-j} \sim N\left(\frac{\sum_{i=1}^{n} w_{ij}\phi_i + w_{i\circledast}\phi_{\circledast}}{\sum_{i=1}^{n} w_{ij} + w_{i\circledast} + \epsilon}, \frac{\sigma^2}{\sum_{i=1}^{n} w_{ij} + w_{i\circledast} + \epsilon} \right), \quad j = 1, ..., n \ , \tag{4.8}$$

$$\phi_{\circledast} | \boldsymbol{\phi}_{-\circledast} \sim N\left(\frac{\sum_{i=1}^{n} w_{i\circledast}\phi_{\circledast}}{\sum_{i=1}^{n} w_{i\circledast} + \epsilon}, \frac{\sigma^2}{\sum_{i=1}^{n} w_{i\circledast} + \epsilon} \right).$$

In (4.8) the conditional expectation is a weighted average of the global random effect $\phi_{\circledast}$ and the random effects in the neighboring areas, with the binary weights depending on the current value of $\tilde{\mathbf{W}}$. The conditional variance is approximately (due to ϵ) inversely proportional to the number of neighbors remaining in the model, including the global random effect $\phi_{\circledast}$.

The matrix $\tilde{\boldsymbol{W}}$ is treated by the LCAR model as a single random quantity, which avoids several problems identified by other authors (for more details see [21], Subsection 3.2). The authors propose eliciting the set of candidate values of $\tilde{\boldsymbol{W}}$ from data having a similar spatial structure as the response variable.

The increased flexibility provided by the LCAR model inevitably means that it is more computationally demanding than the common BYM model.

5. ALCOHOL ABUSE DISORDER DISTRIBUTION ACROSS PORTUGAL

The number of lifetime AAD cases vary between 2679 (16A – Cova da Beira) and 136789 (171 – Grande Lisboa). There are four NUTS 3 (164 – Pinhal Interior Norte, 166 – Pinhal Interior Sul, 169 – Beira Interior Sul and 181 – Alentejo Litoral) where no cases were identified. The national nature of the survey sampling design creates situations where very small or even zero samples at the NUTS 3 level occur. In this situation it might happen that no cases are estimated, which does not mean that no disease diagnoses exist. Therefore, these

areas are treated as having missing values and not as having a null number of cases. The first level of the Bayesian hierarchical model, as seen in (4.1), involves complex calculations, very difficult to run on such numbers, therefore numbers of cases per 100 inhabitants, as well as expected number of cases per 100 inhabitants are used (this change does not eliminate the need of using the expected number of cases because only the size of the population is accounted for, not the structure).

The R software (version 3.1.1), with the package **CARBayes** ([17]) is used to fit the hierarchical models. The main advantages of this package are: (1) the spatial adjacency information is easy to specify as a binary neighborhood matrix; (2) given the neighborhood matrix the models can be implemented by a single function call in R; (3) maps with the disease risk estimates can easily be produced. The package has predefined the following models that will be used: BYM, LLB and LCAR. By running the same model on R and on the BUGS software ([23]) the package's author shows that there is good agreement between the two sets of point estimates, as we confirm in the present work. One disadvantage of the package is that it cannot handle missing values at the response variable level. To overcome this, a Generalized Linear model (GLM), Poisson (quasi-likelihood) model ([26]), is fitted using as response variable the number of lifetime observed cases per NUTS 3 and as covariates the ecological variables defined before, namely the proportion of men and the proportion of population aged 18 to 34. The mean estimated number of lifetime observed cases achieved for the four areas with missing data are incorporated in the response variable vector $\mathbf{Y}$. This methodology is debatable and more work needs to be done, in order to evaluate all possible consequences of this approach.

The MBYM model is fitted using the OpenBUGS software ([23]). Even though the Bayesian methodology could handle the missing values, for comparison purposes the missing values are also replaced by the mean estimated values.

As mentioned in Subsection 4.3, the authors of LCAR propose that, for the elicitation of $\tilde{\boldsymbol{W}}$, data having a similar spatial structure as the response variable should be used. In their case, the prior elicitation was based on response variable data from previous years. Our decision was to use the number of cases of four other related mental disorders, chosen as follows. Disorders considered in the Portuguese version of the WMHSI include a comprehensive range of mental disorders, and a GLM is fitted (Binomial model) to understand which mental disorders are most commonly present with AAD. The response variable is two-level categorical, taking value one if the individual suffers from AAD and taking value zero otherwise, and the covariates are of the same type, for all other disorders. At a lower than 5% significance level, the following disorders have an Odds Ratio larger than one: Alcohol Dependence, Oppositional Defiant Disorder, Hypomania, and Intermittent Explosive Disorder. In the cases where values are missing the procedure followed is the one defined before, using as covariates the remaining disorders. For example, for alcohol dependence disorder as response variable, the

covariates are: alcohol abuse disorder, oppositional defiant disorder, hypomania and intermittent explosive disorder. The mean estimated number of cases are imputed in the response variable vectors. There are two reasons to use a different approach in the present case. First, in Portugal, data on ADD from previous surveys is not available. Second, this work is on DM and not on ESR, therefore the decision is to use data from related mental disorders.

5.1. Hyperpriors

Table 1 shows the prior distributions implemented in the four models. In the LCAR model, on top of the already mentioned information for the $\tilde{\boldsymbol{W}}$ matrix, the parameter ϵ is set to 0.001.

Table 1: Prior distributions for the models.

Model	Parameter	Prior Distribution	Mean/Shape	Variance/Scale
BYM	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	μ	Gaussian	0	1000
	σ_θ^2 and σ_ψ^2	Inverse-Gamma	0.001	0.001
MBYM	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	100000
	μ	Flat	-	-
	σ^2	Inverse-Gamma	0.001	0.001
	λ	Uniform [0,1)	0.5	0.5
LLB	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	σ^2	Inverse-Gamma	0.001	0.001
	ρ	Uniform [0,1)	0.5	0.5
LCAR	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	σ^2	Uniform [0,1000)	500	500

5.2. Inference

Posterior inference for all models is based on Markov Chain Monte-Carlo simulation, using a combination of Gibbs sampling and Metropolis-Hastings algorithms. Posterior inference is based on 8000 MCMC samples, which are obtained by running one chain for 100000 samples, by which convergence is assumed to have occurred. We ignore the first 20000 samples as burn-in, and use the remaining 80000 subsequent samples to obtain the posterior distributions of the parameters of interest (a thin of 10 is used to reduce the autocorrelation).

Pilot runs are carried out to establish appropriate burn-in using the Geweke's diagnostic ([9]). Convergence is assessed by visually monitoring the trace and the posterior density plot for each of the parameters.

5.3. Results

Each model is assessed by the resulting Deviance Information Criterion (DIC) ([32]), where a smaller value represents a better fitting model. Table 2 shows the results of the four models.

Table 2: DIC results, which include the effective number of parameters in the model (p.D.).

	BYM	**MBYM**	**LLB**	**LCAR**
DIC	155.3	145.0	159.2	158.0
p.D.	14.3	5.8	18.5	19.5

Table 2 shows that, according to DIC, the MBYM model exhibits the best fit. BYM model is the second best. Following [24], $\lambda = 1$ represents spatial/local smoothing and $\lambda = 0$ represents non-spatial/local smoothing, based on the disease mapping data at hand. In the MBYM the posterior mean value of $\lambda = 0.58$, shows that the data has an higher spatially structured variance than unstructured variance. As already proved by [16], the BYM model shows more robust results in the presence of strong spatial correlation structures, as it seems to be the case here.

Figure 2 shows the posterior median SMR values for the 28 NUTS 3, produced by the MBYM model. Table 3 shows summary measures of the marginal posterior of the parameters of interest obtained by the MBYM model.

Figure 3 displays histograms of the (a) raw SMR and the (b to d) smooth posterior median SMR values for the 28 NUTS 3, produced by the models. The concentration around the interval $[0.5, 1.5]$ on the latter can clearly be seen. Mapping the raw SMRs gives a misleading picture of the risk pattern, whereas any of the four models (plus LLB, which is not presented, but shows the same overall results) give posterior median relative risks less dispersed. This ability of the Bayesian models to "clean" adequately the SMRs from the false patterns created by the Poisson noise had been already referred by [30].

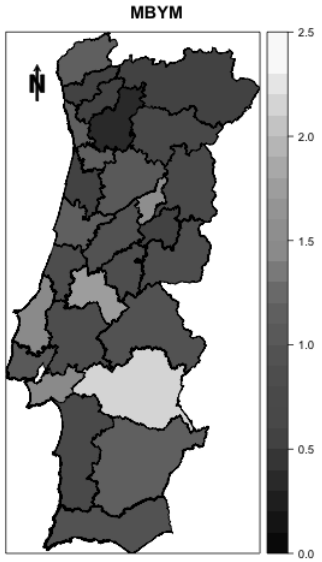

Figure 2: MBYM AAD posterior median SMRs per NUTS 3.

Table 3: MBYM model parameters summary.

Parameter	**Prior distribution**	**Prior mean**	**Prior std**	**MCMC Posterior mean (std)**	**2.5%**	**Median**	**97.5%**
β_0	Flat	0		−0.11 (0.10)	−0.32	−0.11	0.08
β_1	N(0, 100000)	0	100000	−0.23 (0.14)	−0.52	−0.22	0.06
β_2	N(0, 100000)	0	100000	−0.8 (0.13)	−0.34	−0.07	0.18
λ	U[0,1)	0.5	0.5	0.58 (0.25)	0.07	0.61	0.97
σ^2	IG(0.001, 0.001)	1	10	0.61 (0.17)	0.35	0.59	1

The LCAR model is the only one that does not have a single global level of smoothness and therefore any existing discontinuities in the risk pattern can only be concluded from this model. There are 122 neighborhoods (or connections) between the 28 NUTS 3. When applying the LCAR model, the 95% credibility interval of the number of removed connections is [2, 54]. This fact provides evidence that there is information in the data to estimate the number of connections to be removed. Results confirm the known deep cultural roots in the country on the differences between the coast- and the country-side NUTS 3. This is the case of Península de Setúbal and Algarve, two coast-side NUTS 3 sharing physi-

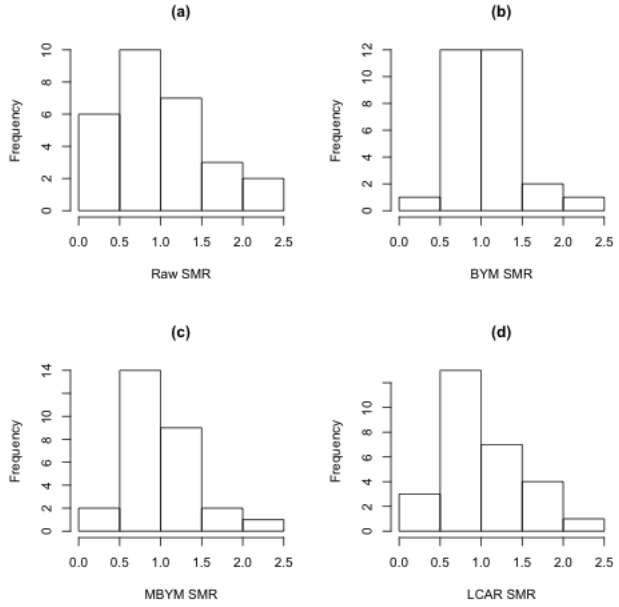

Figure 3: Histograms of the (a) raw SMRs and posterior medians of the (b,c,d) SMRs, for all areas derived by each of the three models, (b) BYM, (c) MBYM and (d) LCAR.

cal borders with the country-side NUTS 3 Alentejo, which are no longer present when data is used to estimate connections.

As mentioned in Subsection 3.2 the goal of DM is not the estimation of associations between covariates and the disease cases, but is to estimate the pattern of disease risk over a geographical region. Nevertheless, due to the fact that the two coefficients (β_1 and β_2) did not show to be significantly different from zero (contrary to expectations mentioned in Section 2), one must remember that this is an ecological study design, and the results must not be interpreted in terms of individual level cause and effect. One possible explanation is ecological bias as the prevalence rate of AAD is higher in younger men. Another possible explanation is that both the random and the covariate effects are confounded, because both are globally smooth in the MBYM model.

6. DISCUSSION

In the past years hierarchical Bayesian models have been developed and refined to achieve statistically precise local estimates of disease risk for each small region. In this study four of those models are assessed and used to estimate the disease risk of AAD at the NUTS 3 level, in Portugal.

In terms of DIC, the MBYM model achieves the best results. The MBYM model derives from the BYM model in an attempt to overcome the known deficiency of the latter, the lack of identifiability. The MBYM is identifiable and facilitates hierarchical modeling of the additive effects of unobserved covariates that might be spatially and randomly varying ([24]). In the present case its superior performance is likely to result from the BYM (and MBYM) model ability of achieving the best results in cases when the spatial correlation structure is strong, as seems to be this case.

The LLB model has consistently shown good results across a variety of cases but in this study, in terms of DIC, it proves to be the most poorly performing. While other authors show that the LLB model is the one achieving the best results ([16, 24]), our study shows otherwise. The performance of each model will depend on the type of data at hand, and none can be defined as the 'gold standard' over others.

The LCAR model is the only model that does not take the neighborhoods as fixed but those emerge from real data, as a random quantity. By doing that, in this example, the known cultural differences (between country- and coast-side) in the country are confirmed.

This study has some particularities when compared with the majority of the published applications:

a) The data use emerged from a survey, which was not plan to have local (at NUTS 3 level) samples with the proper size to allow designed-based estimation, and therefore presents some missing values. To overcome this a frequentist model is used.

b) The complex computations of the first level of the hierarchical Bayesian models do not allow the direct use of the survey estimates. To overcome this the number of lifetime cases of AAD per 100 inhabitants is used.

c) The LCAR model is used as a DM and not as a ESR, and therefore the type of data used for the elicitation of the $\tilde{\mathbf{W}}$ matrix is not previous periods data for the same disease but data from correlated disorders.

The epidemiological study presented in this paper shows substantial evidence of some 'hot spots' in the Center and South of the country allowing the authorities to focus interventions on these 'excess risk' areas.

There are still many opportunities for future work in this area. First the global ICAR's property of tending to negative pair-wise risk dependence as the 'spatial proximity' between two regions is further apart and its potential impact on posterior inference has not been yet sufficiently explored and understood.

Second [7] showed that region effects can be greater (smaller) for specific age groups. We know that AAD is more prevalent in young adult men ([13, 27]). Further research on the region effects on this age-gender group is needed. Third the four models used in this work were GLMM (Generalized linear mixed models), but the linear assumption on the covariate effects might be too restrictive, the usage of a GAMM (generalized additive mixed model) should be explored as it can eventually reveal non-linear relationships.

ACKNOWLEDGMENTS

The Portuguese Mental Health Survey Initiative is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative which is supported by the National Institute of Mental Health (NIMH; R01 MH070884), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13-MH066849, R01-MH069864, and R01-DA016558), the Fogarty International Center (FIRCA R03-TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, GlaxoSmithKline, and Bristol-Myers Squibb. We thank the staff of the WMH Data Collection and Data Analysis Coordination Centers for assistance with instrumentation, fieldwork, and consultation on data analysis. None of the funder had any role in the design, analysis, interpretation of results, or preparation of this paper. A complete list of all within-country and cross-national WMH publications can be found at http://www.hcp.med.harvard.edu/wmh/.

The Portuguese Mental Health Study was carried out by the Department of Mental Health, Faculty of Medical Sciences, NOVA University of Lisbon, with collaboration of the Portuguese Catholic University, and was funded by Champalimaud Foundation, Gulbenkian Foundation, Foundation for Science and Technology (FCT) and Ministry of Health.

The authors would like to thank Vitor Lobo and Ales Popovic for their useful comments and suggestions.

The authors gratefully acknowledge the valuable comments and suggestions made by the two referees and the editor, all of which have greatly improved the focus and presentation of this work.

REFERENCES

[1] Alonso, J.; Angermeyer, M.C.; Bernert, S.; Bruffaerts, R.; Brugha, T.S.; Bryson, H.; de Girolamo, G.; Graaf, R.; Demyttenaere, K.; Gasquet, I.; Haro, J.M.; Katz, S.J.; Kessler, R.C.; Kovess, V.; Lépine, J.P.; Ormel, J,; Polidori, G.; Russo, L.J.; Vilagut, G.; Almansa, J.; Arbabzadeh-Bouchez, S.; Autonell, J.; Bernal, M.; Buist-Bouwman, M.A.; Codony, M.; Domingo-Salvany, A.; Ferrer, M.; Joo, S.S.; Martínez-Alonso, M.; Matschinger, H.; Mazzi, F.; Morgan, Z.; Morosini, P.; Palacín, C.; Romera, B.; Taub, N. and Vollebergh, W.A. (2004). Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project, *Acta Psychiatrica Scandinavica. Supplementum*, **420**, 21–7.

[2] Banerjee, S.; Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman&Hall/CRC, New York.

[3] Besag, J.; York, J. and Mollié, A. (1991). Bayesian image restoration,with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics*, **43**, 1, 1–20.

[4] Best, N.; Richardson, S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping, *Statistical Methods in Medical Research*, **14**, 1, 35–59.

[5] Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 3, 671–81.

[6] Coelho, P.S. and Pereira, L.N. (2011). A spatial unit level model for small area estimation, *REVSTAT*, **9**, 2, 155–180.

[7] Dean, C.B.; Ugarte, M.D. and Militino, A.F. (2001). Detecting interaction between random region and fixed age effects in disease mapping, *Biometrics*, **57**, 1, 197–202.

[8] Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition: DSM-IV-TR (1994). American Psychiatric Association.

[9] Geweke, J. (1992). *Evaluating the accuracy of sampling–based approaches to the calculation of posterior moments.* In "Bayesian Statistics 4" (J.M. Bernardo; J.O. Berger; A.P. Dawid and A.F.M. Smith, Eds.), Oxford University Press, Oxford, 169–193.

[10] Haro, J.M.; Arbabzadeh-Bouchez, S.; Brugha, T.S.; de Girolamo, G.; Guyer, M.E.; Jin, R.; Lepine, J.P.; Mazzi, F.; Reneses, B.; Vilagut, G.; Sampson, N.A. and Kessler, R.C. (2006). Concordance of the Composite International with standardized clinical assessments in the WHO World Mental Health Surveys, *Archives of General Psychiatry*, **62**, 6, 593–602.

[11] Held, L. and Rue, H. (2010). *Conditional and intrinsic autoregressions.* In "Handbook of Spatial Statistics" (A.E. Gelfand; P.J. Diggle; M. Fuentes and P. Guttorp, Eds.), CRC Press, Boca Raton, 201–215.

[12] Hodges, J.C. and Reich, B.J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love, *The American Statistician*, **64**, 4, 325–334.

[13] Kalaydjian, A.; Swendsen, J.; Chiu, W.T.; Dierker, L.; Degenhardt, L.; Glantz, M.; Merikangas, K.R.; Sampson, N. and Kessler, R.C. (2009). Sociodemographic predictors of transitions across stages of alcohol use, disorders, and remission in the National Comorbidity Survey Replication, *Comprehensive Psychiatry*, **50**, 4, 299–306.

[14] Kessel, R.C.; Berglund, P.; Demler, O.; Jin, R.; Merikangas, K.R. and Walters, E.E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication, *International Journal of Methods in Psychiatry Research*, **15**, 4, 167–180.

[15] Lawson, A.B.; Biggeri, A.B.; Boehning, D.; Lesaffre, E.; Viel, J.-F.; Clark, A.; Schlattmann, P. and Divino, F. (2000). Disease mapping models: an empirical evaluation, *Statistics in Medicine*, **19**, 9, 2217–1142.

[16] Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping, *Spatial and Spatio-Temporal Epidemiology*, **2**, 79–89.

[17] Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors, *Journal os Statistical Software*, **55**, 13, 1–24.

[18] Lee, S.; Guo, W.J.; Tsang, A.; He, Y.L.; Huang, Y.Q.; Zhang, M.Y.; Liu, Z.R.; Shen, Y.C. and Kessler, R.C. (2009). Associations of cohort and socio-demographic correlates with transitions from alcohol use to disorders and remission in metropolitan China, *Addiction (Abingdon, England)*, **104**, 8, 1313–23.

[19] Lee, D. and Mitchell, R. (2012). Boundary detection in disease mapping studies, *Biostatistics (Oxford, England)*, **13**, 3, 415–426.

[20] Lee, D. and Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 4, 593–608.

[21] Lee, D.; Rushworth, A. and Sahu, S.K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution, *Biometrics*, **70**, 2, 419–29.

[22] Leroux, B.G.; Lei, X. and Breslow, N. (2000). *Estimation of disease rates in small areas: a new mixed model for spatial dependence*. In "Statistical Models in Epidemiology, the Environment, and Clinical Trials" (M.E. Halloran and D. Berry, Eds.), Springer, New York, 179–191.

[23] Lunn, D.; Spiegelhalter, D.; Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique, and future directions, *Statistics in Medicine*, **28**, 3049–3067.

[24] MacNab, Y.C. (2011). On Gaussian Markov random fields and Bayesian disease mapping, *Statistical Methods in Medical Research*, **20**, 1, 49–68.

[25] MacNab, Y.C. (2014). On identification in Bayesian disease mapping and ecological-spatial regression models, *Statistical Methods in Medical Research*, **23**, 2, 134–55.

[26] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman & Hall/CRC, New York.

[27] Neumark, Y.D.; Lopez-Quintero, C.; Grinshpoon, A. and Levinson, D. (2007). Alcohol drinking patterns and prevalence of alcohol-abuse and dependence in the Israel National Health Survey, *The Israel Journal of Psychiatry and Related Sciences*, **44**, 2, 126–35.

[28] Neuroscience of Psychoactive Substance Use and Dependence (2004). World Health Organization, Geneva, **99**, 10, 1361–1362

[29] Rao, J.N.K. (2003). *Small Area Estimation*, Wiley, New York.

[30] Richardson, S.; Thomson, A.; Best, N. and Elliot, P. (2004). Interpreting posterior relative risk estimates in disease mapping studies, *Environmental Health Perspectives*, **112**, 9, 1016–1025.

[31] 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.

[32] Spiegelhalter, D.J.; Best, N.; Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Society Statistical Society. Series B (Statistical Methodology)*, **64**, 4, 583–639.

[33] Swendsen, J.; Conway, K.P.; Degenhardt, L.; Dierker, L.; Glantz, M.; Jin, R.; Merikangas, K.R., Sampson, N. and Kessler, R.C. (2009). Socio-demographic risk factors for alcohol and drug dependence: the 10-year follow-up of the national comorbidity survey, *Addiction*, **104**, 8, 1346–1355.

[34] Wakefield, J. (2007). Disease mapping and spatial regression with count data, *Biostatistics (Oxford, England)*, **82**, 2, 158–83.

[35] Wakefield, J. and Lyons, H. (2010). *Spatial aggregation and the ecological fallacy*. In "Handbook of Spatial Statistics" (A.E. Gelfand; P.J. Diggle; M. Fuentes and P. Guttorp, Eds.), CRC Press, Boca Raton, 541–558.

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

[37] Waller, L. and Carlin, B. (2010). *Disease mapping*. In "Handbook of Spatial Statistics" (A.E. Gelfand; P.J. Diggle; M. Fuentes and P. Guttorp, Eds.), CRC Press, Boca Raton, 217–244.

[38] WHO, S. A. T. in the D. of M. H. and S. A. of the W. H. O. (2014). Global Status Report on alcohol and health, *Retrieved from http://www.who.int*.

[39] Xavier, M.; Baptista, H.; Mendes, J.M.; Magalhães, J.M. and Caldas-de-Almeida, J. (2013). Implementing the World Mental Health Survey Initiative in Portugal - rationale, design and fieldwork procedures, *International Journal of Mental Health Systems*, **7**, 1, 19.

REVSTAT – Statistical Journal

Background

Statistical Institute of Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT – STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another, external.

— The only working language allowed will be English.

— Three volumes are scheduled for publication, one in April, one in June and the other in November.

— On average, four articles will be published per issue.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract / indexed in

REVSTAT is expected to be abstracted / indexed at least in *Current Index to Statistics*, *Statistical Theory and Method Abstracts* and *Zentralblatt für Mathematik*.

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

— By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

— By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Submitted manuscripts (text, tables and figures) should be typed only in black, on one side, in double spacing, with a left margin of at least 3 cm and not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style.

This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to *PC Windows System* (Zip format), *Mackintosh*, *Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's Website: http://www.ine.pt/revstat/inicio.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: liliana.martins@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Also send always the final paper version to:

Maria José Carrilho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística, I.P.
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright