Structured additive Regression Modeling of pulmonary tuberculosis infection

Bruno de Sousa[1], Carlos Pires[1], Dulce Gomes[2], Patrícia Filipe[2], Ana Costa-Veiga[3,4], Carla Nunes[3]

[1]Faculty of Psychology and Education Sciences, University of Coimbra, Portugal
[2]Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, Departamento de Matemática, Escola de Ciências e Tecnologia, Portugal
[3]CISP - Centro de Investigação em Saúde Pública, National School of Public Health, Universidade Nova de Lisboa, Portugal
[4]H&TRC - Health & Technology Research Center, ESTeSL, Lisbon School of Health Technology, Instituto Politécnico de Lisboa, Av. D. João II, Portugal

Abstract

Tuberculosis (TB) is one of the top 10 causes of death and the leading cause from a single infectious agent (above HIV/AIDS). In 2017, the World Health Organization (WHO) estimated 10.0 million people developed TB and 1.3 million deaths (range, 1.2–1.4 million) among HIV-negative people with an additional 300 000 deaths from TB (range, 266 000–335 000) among HIV-positive people. Studies that understand the socio-demographic characteristics, time and spatial distribution of the disease are vital to allocating resources in order to improve National TB Programs. The database includes information from all confirmed Pulmonary TB (PTB) cases notified in Continental Portugal between 2000 and 2010. Following a descriptive analysis of the main risk factors of the disease, a Structured Additive Regression (STAR) model is presented exploring possible spatial and temporal correlations in PTB incidence rates in order to identify the regions of increased incidence rates. Three main regions are identified as statistically significant areas of increased PTB incidence rates in Continental Portugal. STAR models proved to be a valuable and effective approach in identifying PTB incidence rates and will be used in future research to identify the associated risk factors in Continental Portugal, yielding high-level information for decision-making in TB control.

Keywords

Structured Additive Regression Models; Pulmonary Tuberculosis; Spatial-Temporal Epidemiology; Full Bayesian; Empirical Bayesian

1. Introduction

Pulmonary Tuberculosis (PTB) is an infectious disease which affects millions of people every year, being the second most deadly infectious disease worldwide after the human immunodeficiency virus (HIV) [1]. The disease is caused by the bacillus Mycobacterium tuberculosis that affects mainly the lungs, and can be transmitted through the air when the bacteria is expelled by coughing, sneezing or speaking.

From all notified cases in the WHO European Region in 2017, about 80% had pulmonary localization (PTB) [1,2], a fact also verified in Portugal, with 73.5% of the cases in our database being PTB. An earlier study conducted in Portugal in 2011 aimed at identifying critical areas for the joint occurrence of PTB and HIV/AIDS (Acquired Immune Deficiency Syndrome). The study, based on spatiotemporal clustering analyses, identified the Oporto and Lisbon Metropolitan Areas as critical areas for both diseases, either independently or jointly occurring [3].

Research on spatial and temporal correlations among PTB incidence rates together with disease factors are of the utmost importance from a Public Health perspective. This study will focus on analyzing through STAR (Structured Additive Regression) modeling temporal trends and geographic patterns of PTB incidence rates associated with notified PTB cases in Continental Portugal (278 municipalities) from 2000 to 2010.

2. Methodology

2.1 The data

This study was entirely based on data from registers with the permission of the National Program for Tuberculosis Control. The data was extracted from SVIG-TB (*Sistema de Vigilância da TB em Portugal*) database of the National Program for Tuberculosis Control and included information from all confirmed TB cases, whose notification is mandatory in Continental Portugal (henceforth referred to as Portugal) between 2000 and 2010. Ethics committee approval and informed consent were not required, as data was based on an Official National Surveillance System, provided by the General Directorate of Health, and was previously anonymized.

A total of 25,279 new cases with PTB were used, together with the information regarding municipality of residence, age, sex and disease risk factors, such as alcohol dependence, intravenous drug dependence (IV Drugs), other drug dependence, being an inmate, homeless, an immigrant and co-infected with HIV. This study considered a new case as one defined by WHO [1], that is, a patient with PTB disease involving lung parenchyma who has never received a treatment or who has been taking anti-TB drugs for less than one month. Yearly population data (global and per municipality) were taken from Statistics Portugal.

2.2 The model

Structured Additive Regression Models (STAR) enable the placement within the same framework of nonlinear effects of continuous covariates, spatial effects, time trends and the usual linear or fixed effects in regression models with non-Gaussian responses [4]. A suitable STAR model for spatiotemporal data is given by

$$\eta_{it} = f_1(x_{it1}) + \ldots + fk\,(x_{itk}) + f_{trend}(t) + f_{spat}\,(s_{it}) + u'_{it}\,\gamma, \qquad (1)$$

where η_{it} is the additive predictor for observation i at time t, $f_1(x_{it1})$, ..., $f_k(x_{itk})$ are smooth functions of k continuous covariates x_{it1},... x_{itk}, $f_{trend}(t)$ is a temporal trend, $u'_{it}\,\gamma$ represents the parametric component with γ being the parameter vector of the fixed effects, and $f_{spat}\,(s_{it})$ is a spatially correlated effect of the location (s) where the observation belongs. The spatial effect can furthermore be split into a spatially correlated part and a spatially uncorrelated part: f_{spat} (.) = f_{srt} (.)+ f_{unstr} (.), allowing for a distinction to be made between the unobserved influential factors which obey a global spatial structure and those which may be present only locally [5]

For smooth non-linear effects of continuous covariates and time trends Bayesian penalized splines are used [6, 7]. Correlated and uncorrelated spatial effects follow a Gaussian Markov random field and an independent identically distributed (iid) Gaussian random effects priors, respectively [8].

Inference in the above STAR model can be made through a full (FB) or empirical Bayesian (EB) approach. In a FB approach the unknown variance or smoothing parameters are considered as random variables with suitable hyperpriors and are estimated together with the unknown functions and covariate effects, using MCMC (Markov chain Monte Carlo) simulation techniques [9]. EB approach is based on penalized likelihood inference for the regression coefficients and restricted maximum likelihood estimation (REML) for the variance components [4, 5, 9].

The model here presented analyzes the temporal trend and the spatial distribution of PTB incidence rates in Portugal between 2000 and 2010. The main goal was to identify areas with different risk levels in terms of PTB incidence rates, if they exist.

For this model, municipality was considered as the statistical unit and Y_{it}, the number of new PTB cases in the i^{th} municipality at year t, as the response variable. To be able to model PTB incidence rates, an offset term with regression coefficient fixed to 1 is included in the model and is defined as $log(P_{it}/100{,}000)$, where P_{it} represents the number of habitants in the municipality i at the year t. The final model can then be specified as:

$$\eta_{it} = log(P_{it}/100\,000)\,(offset) + f_{year}\,(t) + f_{srt}\,(s_{it}) + f_{unstr}\,(s_{it}), \qquad (2)$$

where $\eta_{it} = log(E(Y_{it}))$ represents the additive predictor for the $i = 1,\ldots,278^{th}$ municipally at year $t = 2000,\ldots,2010$. The function f_{year} is a smooth function estimated using a Bayesian cubic P-spline [6, 7] with second order random walk penalty with 20 inner knots. For the spatial components, a Gaussian Markov random field is used for the structured effects, $f_{srt}(.)$, and an iid Gaussian random effects for the unstructured effects, $f_{unstr}(.)$ [10]. To take into account the excess of zeros and possible overdispersion of the data, a zero-inflated negative binomial distribution for the response variable was assumed [11]. Inference results are obtained considering a FB approach.

3. Results and Discussion

3.1 Descriptive analysis

Portugal shows a decrease of 42.3% in PTB incidence rates from 28.6 cases per 100 000 population in 2000 to 16.5 cases in 2010 (Figure 1a). When looking at sex differences (Figure 1b), the ratio man to woman was 2.4 in the period 2000-2010, being stable over this time period. Regarding the ratio by age group, Figure 1b, there is almost the same number of new cases for men and women before the age of 25, with over 3 times more new cases of men between the ages of 35 and 64. It is also worth noting that, although there is a decrease in the sex ratio for the age group greater than 64 years of age, this ratio is still equal to 2 for this class.

With respect to changes in age over time, the consistent decrease in incidence is followed by a consistent increase of the median age, Figure 1a, suggesting a decrease in PTB endemic in Portugal.

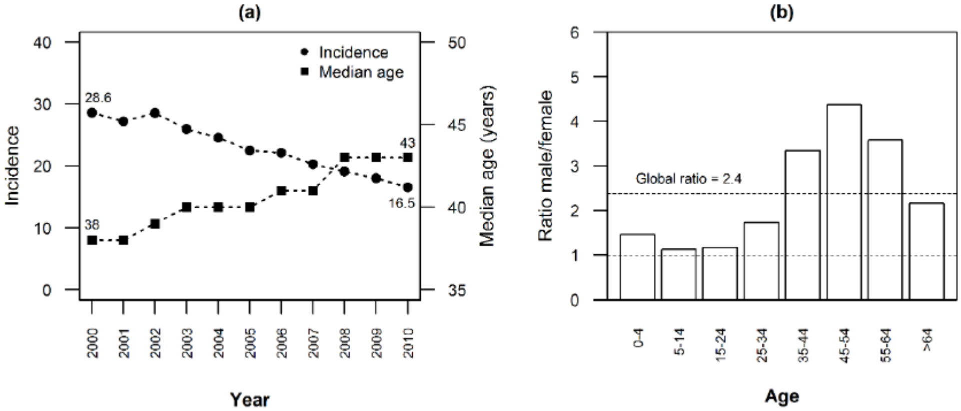

Figure 1: (a) Incidence (new cases per 100 000 population) versus median age, by year; (b) Sex ratio (men to women) of new PTB cases by age group for the period 2000-2010

Factors such as alcohol or drug dependency, HIV co-infection, being an inmate, homeless or an immigrant could contribute to the increased risk of infection with TB, as well as of disseminating it if already ill. Figure 2 shows the yearly evolution of these factors in our database.

Worth note is the steady decrease in the proportion of HIV diagnosed individuals, from 22.3% in 2000 to 10.7% in 2010. A similar trend was observed in IV drugs (Intravenous drugs) dependents that decreased from 12.8% to 7.1% in the same period. Although more moderate, the proportion of new PTB cases being alcohol dependent is also decreasing over time. Notice the increase of the proportion of immigrants after the year 2005.

When looking at the risk factors by sex (Figure 3), it is very clear that the percentage of men with a certain risk factor is always higher when compared to women, except when an immigrant. This difference is quite remarkable when looking at alcohol, where almost 25% of the men in the database are alcohol dependent.

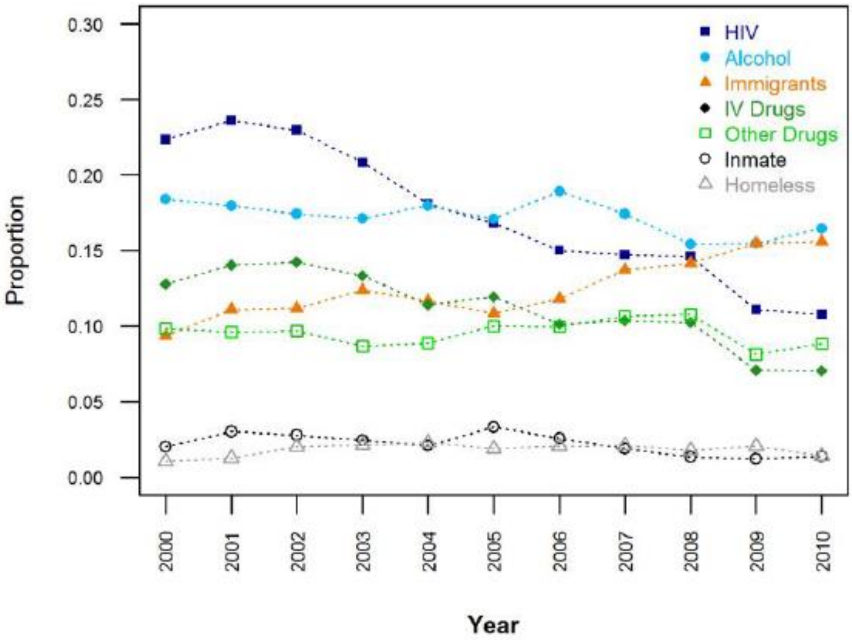

Figure 2: Proportion of risk factors per year in PTB new cases, 2000-2010

Although the total number of men and women are quite different (numbers in brackets in Figure 3), these differences are indeed statistically significant with a p < 0.001 for all the comparisons, with the exception of the proportion of being an immigrant, which showed statistically significant differences between sexes with a p = 0.040.

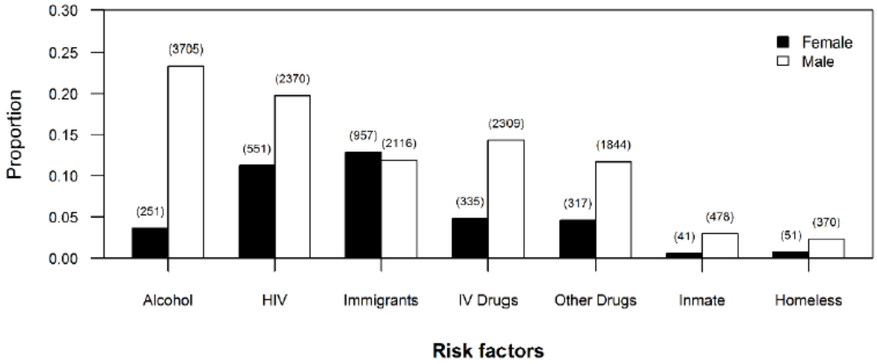

Figure 3: Proportions of risk factors per sex in PTB cases, 2000-2010. In brackets are presented the absolute number of cases.

3.2 Spatial and temporal analysis

Figure 4 shows a clear spatial pattern, with the Metropolitan Area of Porto/Upper North (Region I - MAP), Metropolitan Area of Lisbon (Region II - MAL) and Algarve/Lower Alentejo (Region III) areas (red/darker and black areas in Figure 4) being the higher risk regions that significantly contribute to an increase of the PTB incidence rates. On the contrary, it shows some regions with lower risk in the interior north, center, and Alentejo (lighter areas in Figure 4), that are significantly decreasing PTB incidence rates. The model did not show any significant unstructured (local) spatial effects (not shown here).

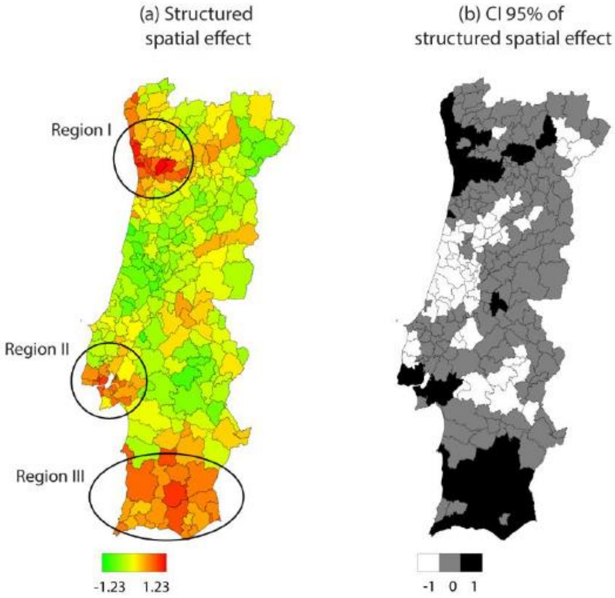
(a) Structured spatial effect
(b) CI 95% of structured spatial effect
Region I
Region II
Region III
-1.23
1.23
-1
0
1

Figure 4: For the period 2000 -2010, (a) Spatial distribution of the posterior means of the global spatial effect; (b) 95% posterior probabilities. Black areas on (b) denote municipalities with strictly positive credible intervals; white areas representing municipalities with strictly negative credible intervals; and

grey areas represent municipalities of non-significant effects for PTB incidence rates (credible intervals containing zero)

Regarding time, Figure 5 shows a slightly non-linear decreasing effect between 2000 and 2010, confirming the capacity of the model to pick up the decreasing effect of PTB new cases shown in the previous descriptive analysis (Figure 1a).

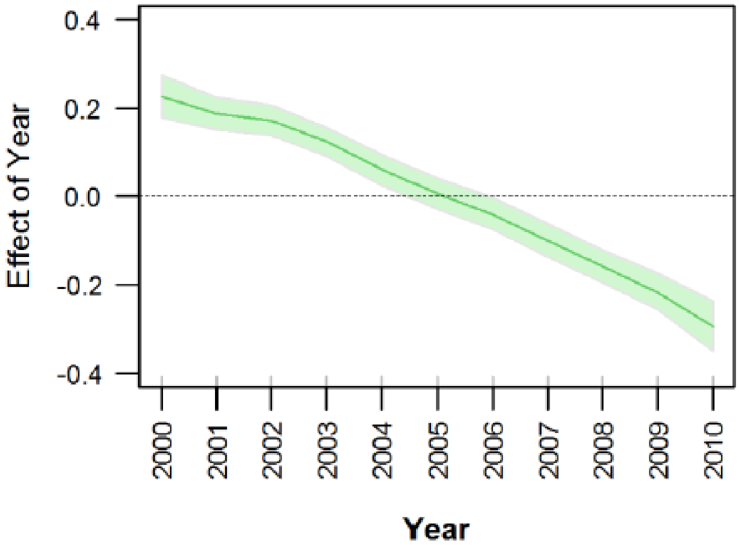

Figure 5: Estimated nonlinear effect of year in PTB incidence rates, together with 95% credible intervals.

4. Conclusion

Nunes et al. [3] identified two main regions, MAL and MAP, as being high risk areas for contracting PTB in Portugal in 2001. The results of our study also suggest a clear urban problem, with MAL (Region II) and the MAP (Region I) being two of the main areas identified as statistically significant areas of increased PTB incidence rates (Figure 4). Although with smaller numbers of new cases of PTB, Algarve and Lower Alentejo (Region III) also emerge as a region within this category. The metropolitan areas of Lisbon (Region II) and Oporto (Region I) correspond to two regions with high population density, resulting immediately in an agglomeration of the main risk groups associated with high incidence of tuberculosis (e.g. homeless, unemployed, IV drug addicts and other drugs). On the other hand, Region III which includes Algarve, not corresponding to an area of high population density throughout the year, it is associated with seasonal tourism and workers particularly through the months of April to September, when it also becomes a high density populated region. It is worth noticing that, after Lisbon with 52% of the total of foreigners living in Portugal, Algarve, North and Center of Portugal, are the three regions with the highest percentage of foreigners (13%, each). In addition, 12% of

Algarve's population is foreigner, making it the region with the greatest representativeness of foreigners' residents (Census 2011, Statistics Portugal).

Future research will focus on the risk factors associated with the identified four regions, namely Region I – Metropolitan Area of Porto and Upper North (34 municipalities), Region II –Metropolitan Area of Lisbon (20 municipalities), Region III – Algarve and Lower Alentejo (17 municipalities), and the Low Risk region with the remaining municipalities (207 municipalities).

As a final note, it is essential to emphasize how Structured Additive Regression (STAR) models offer a rich framework that allows the presence of a wide range of covariates while simultaneously exploring possible spatial and temporal correlations within a very diverse type of response variables.

Acknowledgments

This work was supported by the Portuguese National Funding Agency for Science, Research and Technology, *Fundação para a Ciência e Tecnologia – Ministério da Educação e Ciência*, through the research project [PTDC/SAU-SAP/116950/2010]. The third and fourth authors are also supported through the project [UID/MAT/04674/2019].

References

1. World Health Organization. Global Tuberculosis Report 2018. Geneva: World Health Organization, 2018.
2. European Centre for Disease Prevention and Control/WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in Europe 2014. Stockholm: European Centre for Disease Prevention and Control, 2014.
3. Nunes C, Briz T, Gomes D, & Filipe PA (2011). Pulmonary Tuberculosis and HIV/AIDS: joint space-time clustering under an epidemiological perspective. In: Cafarelli B, editors. Proceedings of the Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition; Foggia e Gargano, p. 1-4.
4. Kneib T (2006). *Mixed model based inference in structured additive regression*. PhD Thesis. Munchen: Universität Munchen, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians.
5. Fahrmeir L, Kneib T, & Lang, S (2004). Penalized structured additive regression for space-time data - a bayesian perspective. *Stat Sinica*, 14:731-761.
6. Lang S, & Brezger A (2004). Bayesian P-splines. *J Comput Graph Stat.*, 13:183-212.
7. Brezger A, & Lang S (2006). Generalized additive regression based on Bayesian P-splines*. Comput Stat Data An.*, 50:967-991.
8. Rue H, & Held L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: Chapman & Hall/CRC.
9. Fahrmeir L, & Lang S (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Fields Priors. *J R Stat Soc C-Appl.*, 50(2):201-220.
10. Osei FB, Duker AA, & Stein A (2012). Bayesian structured additive regression modeling of epidemic data: application to cholera. *BMC Med Res Methodol.*, 12(118).
11. Klein N, Kneib T, & Lang S (2015). Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110:509, 405-419, DOI: 10.1080/01621459.2014.912955.