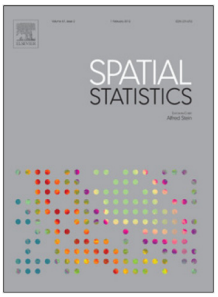

Exploring risk factors in breast cancer screening program data using structured geoadditive models with high order interaction

Elisa Duarte, Bruno de Sousa, Carmen Cadarso-Suárez, Thomas Kneib, Vítor Rodrigues

PII: S2211-6753(16)30162-2
DOI: http://dx.doi.org/10.1016/j.spasta.2017.07.004
Reference: SPASTA 246

To appear in: *Spatial Statistics*

Received date : 30 November 2016
Accepted date : 18 July 2017

Please cite this article as: Duarte, E., Sousa, B.d., Cadarso-Suárez, C., Kneib, T., Rodrigues, V., Exploring risk factors in breast cancer screening program data using structured geoadditive models with high order interaction. *Spatial Statistics* (2017), http://dx.doi.org/10.1016/j.spasta.2017.07.004

Exploring risk factors in breast cancer screening program data using structured geoadditive models with high order interaction

Elisa Duarte[a,*], Bruno de Sousa[b], Carmen Cadarso-Suárez[a], Thomas Kneib[c], Vítor Rodrigues[d]

[a] *Unit of Biostatistics, Department of Statistics and Operations Research, School of Medicine, University of Santiago de Compostela, C/ San Francisco s/n, 15782-Santiago de Compostela, Spain*
[b] *Faculty of Psychology and Education Sciences, University of Coimbra, CINEICC, Rua do Colégio Novo, Apartado 6153, 3001-802 Coimbra, Portugal*
[c] *Institute of Statistics and Econometrics, Dept. of Economics, Georg-August-Universität Göttingen, Humboldtallee 3, 37073 Göttingen, Germany*
[d] *Faculty of Medicine, University of Coimbra, Rua Larga, 3004-504 Coimbra, Portugal*

Abstract

When analyzing data from cancer screening programs, flexible regression specifications are required to account for the highly complex structure in such data. We analyzed data from a breast cancer screening program conducted in central Portugal and considered an extension of structured additive regression models where, in addition to the possibility to include nonlinear and spatial effects, we can include a trivariate interaction between attendance rate, detection rate and mortality rate in the screening program. While spatial effects capture unobserved heterogeneity at the municipality level, the trivariate interaction proves important for the understanding of the complex interaction effects resulting from the diversity in municipality coverage and attendance rates. The trivariate interaction is implemented based on a Markov random field representation which enables efficient Bayesian inference and, when modeling breast cancer incidence rates, showed a significant improvement in terms of model fit when compared to a simpler geoadditive regression model.

*Please address correspondence to Elisa Duarte
Email addresses: duarte.elisa@gmail.com (Elisa Duarte), bruno.desousa@fpce.uc.pt (Bruno de Sousa), carmen.cadarso@usc.es (Carmen Cadarso-Suárez), tkneib@uni-goettingen.de (Thomas Kneib), vrodrigues@fmed.uc.pt (Vítor Rodrigues)

Keywords: Structured Geoadditive Regression (STAR) models, Triple interaction, Breast Cancer Screening

1. Introduction

It is widely accepted that the detection of breast cancer at an early stage, together with adequate and prompt treatment, increases a woman's chances of survival. This makes screening tests one of the most popular prevention strategies conducted to identify breast cancer prior to the development of any symptoms. Screening programs, besides routinely providing these tests to people who appear to be healthy and are not suspected of having breast cancer, also gather valuable inputs that can be useful when studying the risk of the disease. Beyond the pure identification of risk and protective factors and the determination of the marginal magnitude of the impact of each of these factors individually, a thorough understanding of disease etiology can only be acquired when studying the simultaneous potential of multiple risk factors in predicting the probability of developing the disease with the ultimate goal of identifying inequalities between population groups.

By using a geoadditive regression model with a binary response, Duarte et al. (2014a) explored the relationship between variables identified as risk factors (e.g. birth year, age at menopause and menarche, reproductive factors, family history, purchasing power index, region of residency) and the probability of having the disease. Although the study identified an increase in breast cancer risk when going from east to west in the region considered, a pattern also reported in the early years of the screening program by Rodrigues (1993), this effect was not found to be statistically significant. The lack of significance in these results may be due to self-selection bias that naturally occurs in screening programs, along with the complex relationship of variables such as mortality, attendance and detection rates that were not considered in Duarte et al. (2014a). While attendance and detection rates are considered monitoring measurements of a screening program, mortality rates can be seen as an assessment measure of the

success of such programs.

High mortality rates can be a result of a combination of three main factors: high incidence rates (directly associated with the high detection rates of the disease), late diagnosis, and poor accessibility of diagnosis. The last two effects could be diminished, at least to some extent, by the screening program. In the particular case of Central Portugal, the coastal regions are, in general terms, wealthier than the interior ones. This could imply that women in the coastal areas are more likely to use private health facilities as opposed to taking part in the screening program. The natural implication of this phenomenon is twofold: not only will these areas show lower screening attendance rates, but also the women from the coastal areas taking part in the screening program will tend to be of a lower socio-economic status. This behavior will be reflected in deviations from the ideal situation where higher attendance rates will imply that detection rates can be seen as good estimates of incidence rates of a disease in a population. Therefore, the coastal areas will show detection rates as high as their interior counterparts, only due to the fact that women will choose the parallel private health care system instead of the screening program. Due to the complex relationship between detection, attendance and mortality rates, the present work will consider the interaction of these three epidemiologic variables in order to improve the understanding of breast cancer incidence rates while minimizing the impact of the natural bias present in a screening program. Thus, assuming the same context of previous works from the author's (Duarte et al., 2014a), this study will focus on how the inclusion of the triple interaction will be able to capture the expected spatial effects in this region.

Substantially increasing the complexity of the model is the integration of variables so distinct in nature in a single model and the cohort's dimension of 172 334 women. Such complexity demands the use of models that are able to efficiently explore the relationship between different kinds of variables and the probability of having the disease. Therefore, the study uses a structured additive regression (STAR) model (Brezger and Lang, 2006; Kneib, 2006; Fahrmeir et al., 2013) that includes nonlinear effects of continuous covariates,

spatial information, and nonlinear interaction effects in a unified framework. In addition, in order to reliably estimate the spatial effects in our data, this work proposes the inclusion of an interaction term involving more than two variables in the structured additive predictor. Fahrmeir et al. (2013) suggest the construction of Markov random fields (MRF) to be applied in problems of higher dimensions as an alternative approach to the methodology based on an extension of the univariate P-spline to higher dimensions using tensor product smooths (Wood, 2006). In this work, we propose an automatic procedure that creates a neighbor structure to estimate high-order interaction using the MRF methodology as a way to include the interaction between mortality, attendance and detection rates in the model. This structure was created using R (R Core Team, 2013) and inference was carried out using the BayesX software version 3.0.2 (Belitz et al., 2015)

The rest of this paper is organized as follows: Section 2 provides the data provided by the Breast Cancer Screening Program and describes the structured geoadditive regression. Section 3 presents the method to include high-order interaction terms in a structured additive predictor. The results of how this new feature was applied to the breast cancer screening data is presented in Section 4. Finally, Section 5 and 6 present the discussion of the results from this study and the conclusion, respectively.

2. Material and methods

2.1. Breast cancer screening program database from central Portugal

The data set for our analyses was provided by the Portuguese Cancer League (LPCC/NRC – *Liga Portuguesa Contra o Cancro/Núcleo Regional do Centro*) and consists of information on all women who attended the Breast Cancer Screening Program in central Portugal between 1990 and 2009. The central region of Portugal (Figure 1) represents approximately 25% of the population of Portugal and consists of a total of 78 municipalities of different sizes.

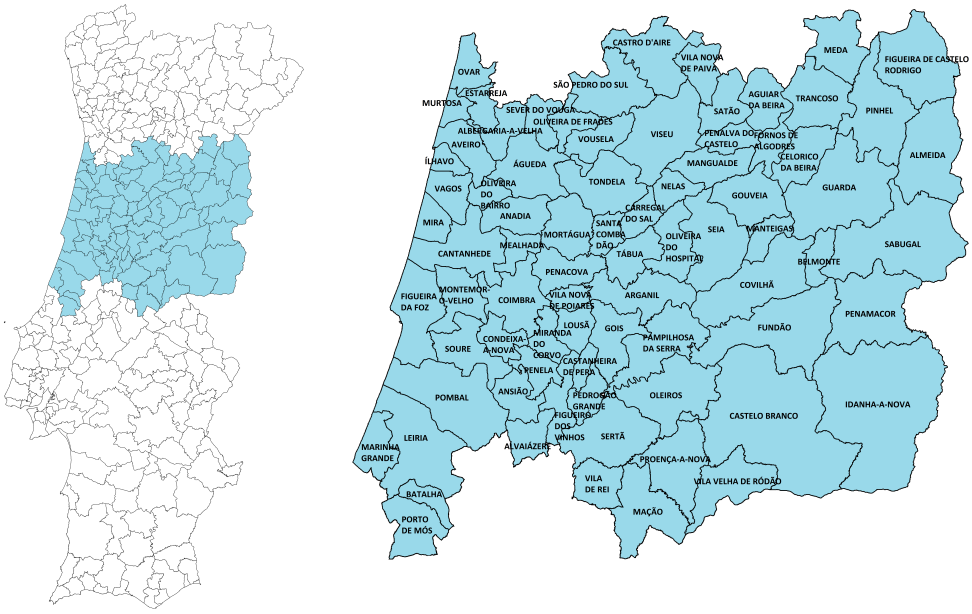

Figure 1: The map of Portugal, with the regions representing the municipalities under study.

The screening program recruitment is done by invitation with different coverage rates achieved in each round. Moreover, since the screening program did not start simultaneously in all municipalities, municipalities can be operating in different rounds in any given year. Once a woman is invited to be part of the screening program in a municipality, follow-up rounds will occur every two years for all women who fall within the inclusion criteria. To facilitate the follow-up, information on all women is continuously updated, including information on cancer diagnosis. Once a woman is diagnosed with breast cancer, she will automatically enter the National Health Service system and will no longer be part of the subsequent rounds of the screening program. The information considered in this study concerns all data collected in the 78 municipalities that comprise the central region of Portugal, representing the most recent data collected from the last visit that a woman made to the screening

program. Due to missing values regarding detection rates, 6 municipalities will be excluded from the analysis, corresponding to a total of 2 996 registries.

The covariates considered in our study are: year of birth, age at menarche and menopause, pregnancy status, nursing status, the use of oral contraceptives, the municipality purchasing power index (PPI), and the place of residence (in terms of the municipality where a woman lives). For age at menopause and menarche, we will consider a bivariate interaction effect while a trivariate interaction will be constructed for detection, mortality, and attendance rates. For purposes of comparison, we will also present the results of the reduced model without considering the high order triple interaction.

The database consists of women born after 1920 and with screening ages between 44 and 70 years old. There are a total of 134 530 women (79%) who reached menopause. A total of 1 716 women (roughly 1%) have been diagnosed with breast cancer while 132 814 women were disease-free throughout the complete screening period. Oral contraceptives, pregnancy and nursing status are coded as binary covariates (1 = Yes, 0 = No). Only 8% of the women in our database were never pregnant. Sixty percent of women breastfed and 34% of women used oral contraceptives. An important covariate considered in this study is family history. This variable is coded with levels from 0 to 3, where level 0 (89% of the women) means a woman with no directly related family member with a breast cancer diagnosis, and the remaining levels representing women with a family history of the disease. Level 1 (5%) indicates that the relatives with the disease were aunts and grandmothers, level 2 (4%) identifies women with a sister that had the disease, and level 3 (2.5%) relates to women where the mother or daughter were affected by breast cancer. Thus, breast cancer family history is represented by three dummy variables, with level 0 as the reference category. The age at menarche and menopause are considered continuous covariates and some descriptive statistics are reported in Table 1. According to INE - Statistics Portugal, the PPI expresses the daily manifested purchasing power, per capita, in the different municipalities, and it is also considered as a continuous covariate in this study. The municipality's baseline is

100 such that municipalities with values under 100 represent regions with lesser economic power. The values of the PPI observed in the data vary between 24 and 145, with the central region of Portugal considered in general a region with a lower purchasing power (89% of the municipalities have a PPI under 100). For the spatial effects analysis, we considered information on the municipality code where a woman lives.

Mortality, detection, and attendance rates were calculated according to the European guidelines for quality assurance in breast cancer screening and diagnosis (Health & Consumer Protection Directorate-General, 2006). These rates were defined by screening round per municipality, with the range and median values presented in Table 1.

Table 1: Summary statistics for the continuous covariates

Variable	*Range*	*Median*
Year of Birth	1920 - 1964	1943
Menarche Age	8.0 - 18.0	13.0
Menopause Age	20.0 - 59.0	49.0
Purchasing Power Index (PPI)	24.0 - 144.9	72.3
Attendance Rate (per 100)	8.7 - 92.6	59.7
Detection Rate (per 1 000)	0 - 9.4	2.8
Mortality Rate (per 100 000)	0 - 579.4	74.3

2.2. Structured geoadditive regression models

Structured Additive Regression (STAR) models provide a generic framework for modelling complex semiparametric regression data. They are particularly useful in our setting since they are flexible enough to deal with different and complex structures of data sets, but also are able to consider a multitude of covariates while exploring possible spatial and temporal correlations. They use a semi-parametric structured additive predictor, representing in one single equation nonlinear effects of continuous covariates, time trends and seasonal

effects, two-dimensional or spatially correlated effects, and covariates with parametric effects. The geoadditive regression model is a special case of the structured additive predictor when considering the spatial correlations.

$$\eta = f_1(\nu_1) + \ldots + f_q(\nu_q) + f_{spat}(s) + \boldsymbol{x'\beta}, \tag{1}$$

where $\nu_1, \ldots, \nu_q$ are one or multidimensional covariates of different types, and $f_{spat}(s)$ is a spatially correlated effect of the location s. The predictor η in equation (1) is linked to the expectation of the response $\mu = \mathrm{E}(y)$ via the link function $\eta = g(\mu)$ as in generalized linear models proposed by (McCullagh and Nelder, 1989).

In this study, we analyzed data on 134 530 women that reached menopause. The probability of each woman having breast cancer, $y_i \in \{0, 1\}$ with $i = 1, \ldots, 134\,530$, is a binary response. A logit model is considered, with the expectation of y_i being given by the probability $\pi \in [0, 1]$. Thus, the response has a logistic distribution with success probability $\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1-\exp(\eta_i)}$ and the link function is given by $\eta_i = g(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$.

The unified treatment that the structured additive regression models approach provides to the different components in a model allows for the direct inclusion of a high order interaction term in the geoadditive predictor equation. Thus, to model the probability of having breast cancer, we are able to consider all the covariates of screening data involved in this study in predictor (1) as follows:

$$\begin{aligned}
\eta_i \;=\; & \beta_0 + \beta_1 famhist1_i + \beta_2 famhist2_i + \beta_3 famhist3_i \\
& + \beta_4 pregnancystatus_i + \beta_5 oralcontraceptives_i \\
& + \beta_6 nursingstatus_i + f_1(birthyear_i) + f_2(PPI_i) \\
& + f_3(menopause_i) + f_4(menarche_i) + f_5(menopause_i, menarche_i) \\
& + f_{str}(municipality_i) + f_{unstr}(municipality_i) \\
& + f_{3Di}(attendance_i, detection_i, mortality_i)
\end{aligned} \tag{2}$$

Nursing and pregnancy status, use of oral contraceptives, and family history are considered via parametric effects. Age at menopause and menarche, year of birth and the municipality purchasing power index are considered continuous covariates, and are modeled using nonlinear functions. This structure is also used for the interaction between age at menopause and menarche. The spatial correlated effect of the municipality where a woman resides can be separated into a spatially structured, $f_{str}(municipality_i)$, and a spatially unstructured, $f_{unstr}(municipality_i)$ part. The aim of this split is to distinguish between factors that obey a strong spatial structure and the ones that are only present locally. The term $f_{3Di}(attendance, detection, mortality)$ refers to the trivariate interaction effect.

For all effects in a geoadditive regression model, the vectors of function evaluations at observed values of the covariates can be expressed as the product of a suitable design matrix $\mathbf{V}_j$ and the coefficient vectors $\boldsymbol{\gamma}_j$ yielding

$$\boldsymbol{\eta} = \boldsymbol{f}_1 + \ldots + \boldsymbol{f}_q + \boldsymbol{X\beta} = \mathbf{V}_1\gamma_1 + \ldots + \mathbf{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X\beta}. \tag{3}$$

This representation is the key to treating covariates of different types in a unifying framework. In a Bayesian formulation, priors are assigned to the regression coefficients $\boldsymbol{\gamma}_j$ to enforce desirable properties of the estimates. In generic form, this prior is of the multivariate Gaussian form with density

$$p(\boldsymbol{\gamma}_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\mathbf{K}_j\boldsymbol{\gamma}_j\right), \tag{4}$$

where the variance parameter τ_j^2 controls the flexibility and the smoothness, and the precision matrix $\mathbf{K}_j$ acts as a penalty matrix. The form of the design matrix $\mathbf{V}_j$ and of the penalty matrix $\mathbf{K}_j$ will depend on the type of model term used.

Bayesian P-splines (Lang and Brezger, 2004) were used to estimate the smooth effects of continuous covariates, age at menopause and menarche, year of birth and the municipality purchasing power index. This approach uses the penalized splines (P-splines) introduced by Eilers and Marx (1996), and assumes that unknown smooth function f_j of a covariate ν_j in equation (1)

can be approximated by a polynomial spline using a rich set of B-spline basis functions. The basis functions evaluated at the observations define the design matrix $\mathbf{V}$ in equation (3), thus the function f_j can be represented by:

$$f_j(\nu_j) = \sum_{m=1}^{d_j} \gamma_{jm} B_{jm}(\nu_j), \tag{5}$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd_j})'$ represents the vector of unknown regression coefficients and B_{jm} is the m-th basis function. In the frequentist approach of Eilers and Marx (1996), penalized likelihood estimation with penalties based on squared k-th order differences of adjacent basis coefficients is used to guarantee sufficient smoothness of the fitted curves. To control function wiggliness in Bayesian P-spline estimates, we assign random walk priors of order k to the regression coefficients (which is the stochastic analogue to a difference penalty in the frequentist setting). The first order random walk is defined as

$$\gamma_{jm} = \gamma_{j,m-1} - u_{jm}, \quad m = 2, \ldots, d_j, \tag{6}$$

while a second order random walk is defined as

$$\gamma_{jm} = 2\gamma_{j,m-1} - \gamma_{j,m-2} + u_{jm}, \quad m = 3, \ldots, d_j, \tag{7}$$

with $u_{jm} \sim N(0, \tau_j^2)$ Gaussian errors and, for initial values, the diffuse priors $p(\gamma_{j1}) \propto const$ and $p(\gamma_{j1}, \gamma_{j2}) \propto const$ are considered.

The term associated with the interaction between age at menopause and menarche is estimated using tensor product P-splines with bivariate first order random walk penalty, an extension of the univariate P-spline to two dimensions (Lang and Brezger, 2004; Brezger and Lang, 2006). The basic idea is to approximate the unknown function $f(\nu) = f(\nu_1, \nu_2)$, where ν_1 and ν_2 are continuous covariates, by the tensor product of one dimensional B-splines, i.e.

$$f(\nu_1, \nu_2) = \sum_{j=1}^{d_1} \sum_{r=1}^{d_2} \gamma_{jr} B_j^{(1)}(\nu_1) B_r^{(2)}(\nu_2),$$

where $B_j^{(1)}(\nu_1), j = 1, \ldots, d_1$ and $B_r^{(2)}(\nu_2), r = 1, \ldots, d_2$ are the univariate basis functions for ν_1 and ν_2. Then the design matrix $\mathbf{V}$ is obtained by evaluating

the products of the basis functions at the observations. The penalty matrix $\mathbf{K}$ can be constructed from Kronecker products of the univariate penalty matrices as

$$\mathbf{K} = \mathrm{I}_{d1} \otimes \mathbf{K}_1 + \mathrm{I}_{d2} \otimes \mathbf{K}_2$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are penalty matrices for first order random walks in ν_1 and ν_2 directions, respectively, and I_{d1} and I_{d2} are d_1 and d_2 dimensional identity matrices, respectively.

As mentioned above, the spatial covariate s is the municipality code of a woman's residence, where it is assumed that neighboring municipalities have more similar characteristics than non-adjoining locations. The definition of a neighborhood is based on two municipalities sharing a common boundary. As mentioned in the description of equation (2), the spatial effect is the sum of two components: the structured (spatially correlated) part, f_{str}, and the unstructured (spatially uncorrelated) part, f_{unstr}, where the index $s \in \{1, \ldots, S\}$ represents a location in the connected geographical regions. For the smooth spatial effect, $f_{str}(s)$, the spatial smoothness prior for the function evaluations $f_{str}(s) = \gamma_s$ is a Markov random field (MRF) Rue and Held (2005). Let therefore ∂_s be the set of all neighbors of region s and let $N_s = |\partial_s|$ denote the number of adjacent regions. Then the spatial smoothness prior for the functions evaluations of neighboring sites is given by:

$$\gamma_s | \gamma_r, r \neq s, \tau^2_{str} \sim N\left(\frac{1}{N_s}\sum_{r \in \partial_s} \gamma_r, \frac{\tau^2_{str}}{N_s}\right),$$

where $r \in \partial_s$ denotes that site r is a neighbor of site s and τ^2_{str} is the variance parameter. Large values of the variance parameter, allows for ample deviations from the prior expectation, while for small values, all effects will tend to be the same which corresponds to a flat surface estimate.

The design matrix $\mathbf{V}$ is now a 0/1 incidence matrix that links observations and regions:

$$\mathbf{V}[i, s] = \begin{cases} 1, & \text{if } y_i \text{ was observed in region} s \\ 0, & \text{otherwise.} \end{cases}$$

The penalty matrix $\mathbf{K}$ has a form of an adjacency matrix:

$$\mathbf{K}[s,r] = \begin{cases} -1, & s \neq r, s \sim r \\ 0, & s \neq r, s \nsim r \\ N_s, & s = r \end{cases},$$

where $s \sim r$ denotes that site s is neighbor of site r.

The prior used for modeling the unstructured spatial effects, $f_{unstr}(s) = \gamma_s$, is a standard Gaussian random effects prior given by $\gamma_s \sim N(0, \tau^2_{unstr})$, that defines i.i.d random effects in a classical point of view. The design matrix $\mathbf{V}$ for this term is a 0/1 incidence matrix that links observations and municipalities, and the penalty matrix $\mathbf{K}$ is given by the identity matrix $\mathbf{I}$.

2.3. Inference

Inference is performed with empirical Bayes posterior analysis based on a generalized linear mixed model representation (Kneib, 2006), using the software BayesX v.3.0.2 (Belitz et al., 2015) The model is reparameterized as a proper mixed model based on the decomposition of the vector $\boldsymbol{\gamma}_j$ of each model component, $\boldsymbol{f}_j = \mathbf{V}_j\boldsymbol{\gamma}_j, j = 1, \ldots, q$ into a unpenalized and a penalized part, i.e.

$$\boldsymbol{\gamma}_j = \mathbf{V}_j^{unpen}\boldsymbol{\beta}_j + \mathbf{V}_j^{pen}\tilde{\boldsymbol{\gamma}}_j,$$

with $\mathbf{V}_j^{unpen}$ a $d_j \times (d_j - k_j)$ matrix, and $\mathbf{V}_j^{pen}$ a $d_j \times k_j$ matrix, where k_j denotes the rank of the associated penalty matrix.

The vectors $\boldsymbol{\beta}_j$ of fixed effects and $\tilde{\boldsymbol{\gamma}}_j \sim \mathrm{N}(\mathbf{0}, \tau_j^2 I)$ of i.i.d random effects result from an appropriate choice of the design matrices $\mathbf{V}_j^{unpen}$ and $\mathbf{V}_j^{pen}$. By defining the matrices $\mathbf{X}_j = \mathbf{V}_j\mathbf{V}_j^{unpen}$ and $\mathbf{U}_j = \mathbf{V}_j\mathbf{V}_j^{pen}$, it is possible to rewrite the predictor (3) as

$$\eta = \sum_{j=1}^{q} \mathbf{V}_j\boldsymbol{\gamma}_j + \boldsymbol{X}\boldsymbol{\beta} = \sum_{j=1}^{q} \mathbf{X}_j\beta_j + \mathbf{U}_j\tilde{\gamma}_j.$$

With this mixed model representation, the regression coefficients are estimated using penalized maximum likelihood, and the variance parameters are

estimated using restricted maximum likelihood (REML). Thus, the variances τ_j^2 are treated as constants to be estimated from the marginal posterior, while the posterior only depends on the regression terms of the model. This posterior is given by

$$p(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}, \mid y) \propto L(y, \boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}) \prod_{j=1}^{q} p(\tilde{\boldsymbol{\gamma}} \mid \tau_j^2), \tag{8}$$

where $L(.)$ denotes the likelihood, which is the product of individual likelihood contributions defined by the Bernoulli density, and $p(\tilde{\boldsymbol{\gamma}} \mid \tau_j^2)$ is the prior for the regression coefficients as given in equation (4).

3. High order interaction term

In this section the procedure to include the interaction between the covariates attendance, detection and mortality rates in the geoadditive predictor 1 will be described. Although the application of the procedure is done using a geoadditive predictor, the procedure can be used to cast a high order term in the general form of the structured additive predictor. The inclusion of a high order interaction term, $f(\nu_1, \ldots, \nu_q)$, where $\nu_1, \ldots, \nu_q$ are continuous covariates, in the structured additive predictor can be based on appropriate extensions of the approaches used in the bivariate case: tensor product P-splines, radial basis functions or kriging terms based on isotropic correlation functions.

The approach used in this study is based on the concept of neighborhoods, applying the construction of Markov random fields as a smoother of the interaction between attendance, detection and mortality rates in order to take into account this interaction term in the structured additive predictor. The idea behind this relies on the discretization of the attendance, mortality and detection rates into categories, in which the possible combinations will work as a code of position in a three-dimensional structure. Through this structure it is possible to define for each position a set of neighbors where Markov random fields can be applied. In the present case, for the discretization process is calculated the percentiles 10%, 25%, 50%, 75% and 90% for each rate. Then, the observed value of each rate is associated to the matching class: 0-10%, if the value is less

or equal to the percentile 10; 10%-25% if the value is greater than the percentile 10 and lesser or equal to the percentile 25; following the same logic for the classes consisting of the percentiles 25-50, 50-75, 75-90 and greater than 90. This association is quantified with the values in the set $\{1, 2, 3, 4, 5, 6\}$, corresponding in increasing order to the percentiles classes. As each observation in the database will have one of these values for the attendance, the detection and the mortality rate, the classification of each rate is combined into categories. In this application there are three covariates, with six possibilities of classifications, which corresponds to $6^3 = 216$ possible categories. This new information is recorded into a new covariate called *cat*.

Applying a coordinate system that associates one dimension with each rate classification domain, the new covariate *cat* with values $1, ..., 216$ works as a spatial index representing a location in a three-dimensional structure. As such, it is possible to define a set ∂_{cat} of all neighbors of a category *cat* and the number $N_{cat} = |\partial_{cat}|$ of adjacent categories in the three-dimensional structure. With this setting, it is possible to define a MRF as a smoothness prior for the interaction term. In the same way as was described in the previous section for the spatial term,

$$\gamma_{cat}|\gamma_{nei}, nei \neq cat, \tau^2_{3Di} \sim N\left(\frac{1}{N_{cat}} \sum_{nei \in \partial_{cat}} \gamma_{nei}, \frac{\tau^2_{3Di}}{N_{cat}}\right),$$

define the smoothness prior for the functions evaluations of neighboring categories. Similarly, as described for the spatial term, the design matrix $\mathbf{V}$ is a $0/1$ incidence matrix linking the observations and the categories is given by

$$\mathbf{V}[i, cat] = \begin{cases} 1, & \text{if } y_i \text{ is an observation with category } cat \\ 0, & \text{otherwise,} \end{cases}$$

and the penalty matrix $\mathbf{K}$ has a form of an adjacency matrix:

$$\mathbf{K}[cat, nei] = \begin{cases} -1, & s \neq nei, cat \sim nei \\ 0, & cat \neq nei, cat \nsim nei \\ N_{cat}, & cat = nei. \end{cases}$$

,

where $cat \sim nei$, denotes that cat and nei are neighbors in the three-dimensional structure.

One critical aspect for our approach to modelling 3D interactions is of course the number of percentiles to consider. Using a very large number will lead to only a minor loss of information but will also induce an excessively larger number of potential covariate combinations. On the other hand, choosing the number of percentiles too small will considerably reduce the flexibility and therefore the possibility to adapt to the data. Importantly, our approach involves a neighborhood-based prior such that it can also accommodate situations where no only a small number of observations is available in certain categories. In such cases, the neighborhood-based prior allows us to borrow strength from neighboring regions such that the effect is still well identified even with data sets of only moderate size.

The 3D structure based on the percentiles was created using a function developed in R (R Core Team, 2013) environment and its code is disclosed in the Appendix A of this paper.

4. Results

As previously described, the trivariate interaction term was introduced in the intent to capture the screening data complexity which results from, among other things, the heterogeneity present in municipalities coverage and attendance rates. In order to evaluate the need for this high order interaction term, the results of the model of equation (2) will be compared to those obtained with a simpler model performed without this trivariate interaction term. Henceforth, to distinguish between these two models, the first is labeled as TIM and the second as M0.

Table 2 summarizes the estimation results for the fixed effects. The use of oral contraceptives and breastfeeding appear as a risk factor, while ever being

pregnant shows to be a protective factor in the model TIM. Despite agreement between TIM and M0 models on the effects of the fixed effects, none of these effects was statistically significant in contrast to the latter two for the M0 model. Regarding family history, the results show that categories 2 (sisters with cancer) and 3 (mother or daughters with cancer) are statistically significant in both models, classifying these two categories as a risk factor of the disease.

Table 2: Fixed effects results: the estimated coefficients are presented in terms of the odds-ratio (OR) and the corresponding 95% credible intervals (CI).

Variable	*Model*	*OR*	*95% CI*	*p-value*
Oral contraceptives	TIM	1.068	(0.956, 1.193)	0.245
	M0	1.115	(0.999, 1.243)	0.051
Nursing status	TIM	1.010	(0.898, 1.138)	0.860
	M0	1.145	(1.019, 1.286)	0.022
Pregnancy status	TIM	0.896	(0.744, 1.281)	0.250
	M0	0.745	(0.621, 0.895)	0.002
Family History 1	TIM	1.184	(0.947, 1.481)	0.139
	M0	1.174	(0.942, 1.464)	0.153
Family History 2	TIM	1.784	(1.453, 2.190)	<0.001
	M0	1.706	(1.394, 2.087)	<0.001
Family History 3	TIM	1.655	(1.291, 2.122)	<0.001
	M0	1.730	(1.354, 2.209)	<0.001

TIM - trivariate interaction model; M0 - simpler model

Figure 2 shows the effect of the continuous covariates in the probability of having breast cancer.

An increasing effect is observed for women born until 1945, followed afterwards by a decrease in both models. It is worth noting that the TIM model clearly detects a constant negative effect on having the disease for women born before 1936. The plot of age at menarche shows no effect of this variable, while an upward trend in age at menopause effect is observed. Both

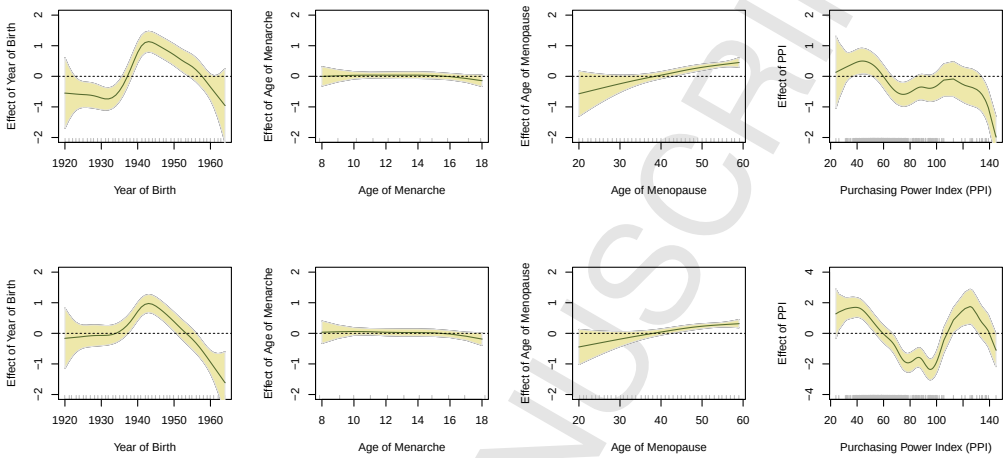

Figure 2: Non-linear effect of the continuous covariates together with 95% pointwise credible intervals for the TIM model (upper) and the M0 model (lower).

models suggest that the risk of breast cancer increases for women who reached menopause after 40 years of age. The purchasing power index (PPI) is the continuous covariate with more differences between the trivariate interaction model (TIM) the simpler model (M0). The marked fluctuation observed in the simpler model is lessened with the introduction of the triple interaction term. An increased risk of the disease is present in both models for municipalities with lower PPI values, but for the ones with a PPI between 100 and 130 there is no effect of this index for the model TIM.

Regarding the interaction between age at menopause and menarche (Figure 3), the results are very similar for both models. The interaction effect is positive for women with late menopause being stronger for women with earlier menarche (black regions). Nevertheless these effects are not statistically significant.

The structured spatial effect are shown in Figure 4. A high breast cancer risk is observed in the TIM model, for municipalities of the center north and center west of Portugal, with the municipalities of Tondela, Mortágua, Carregal do Sal

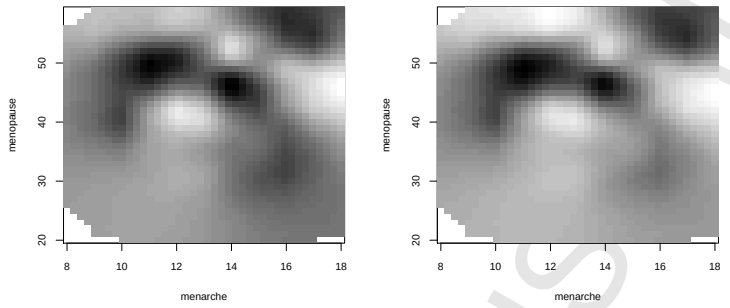

Figure 3: Surface effect of menopause× menarche for the TIM (left) and M0 (right) models. Black regions denote an increased risk of the disease, while white regions represent a decreased one.

and Águeda (red regions) standing out. The results provide a distinct behavior of the structured spatial effect when the trivariate interaction is included in the model along with the observed significance of these effects shown in the upper right side of Figure 4. There were no significant unstructured spatial effects in the model with the trivariate interaction (TIM) as opposed to the simpler model (M0) that showed some local statistically significant effects (graphs not shown here). This fact indicates that in the latter case, the model shows some possible local effects caused by unobserved covariates.

In Table 3, model fit indices are presented for the two different models. The inclusion of the trivariate term improves the model fit, both in terms of the Akaikes information criterion and the Bayesian information criterion. The effect of the trivariate term is significant, with strictly positive and negative credible intervals for the 95% posterior probabilities of the estimated effects. As an example, Figure 5 shows the changes of the effect on breast cancer between attendance and detection rates over the considered six categories of the mortality rate, with the black zones representing an increased effect on the breast cancer risk, while an opposite effect is depicted by the white zones.

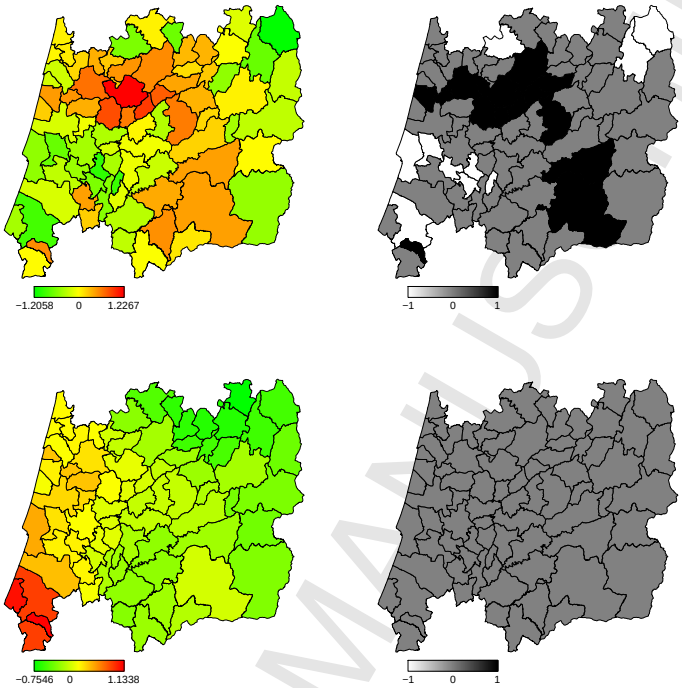

Figure 4: Structured spatial effects (left side) for the model TIM (upper) and the model M0 (lower), together with the posterior probabilities for a nominal level of 95% credible intervals (right side). Black areas denote regions with strictly positive credible intervals, while white areas denote regions with strictly negative credible intervals.

Table 3: Information criteria for the models with and without the trivariate interaction term

Model	*TIM*	*M0*
-2*log-likelihood	15898.7	17337.6
Degrees of freedom	181.923	97.857
AIC (conditional)	16262.6	17533.3
BIC (conditional)	18047.2	18493.2

In addition, the good performance of the TIM model is reinforced by the ROC curve plotted in Figure 6. There is an increase of approximately 10% of the area under the curve for the model TIM comparing with simpler model, M0.

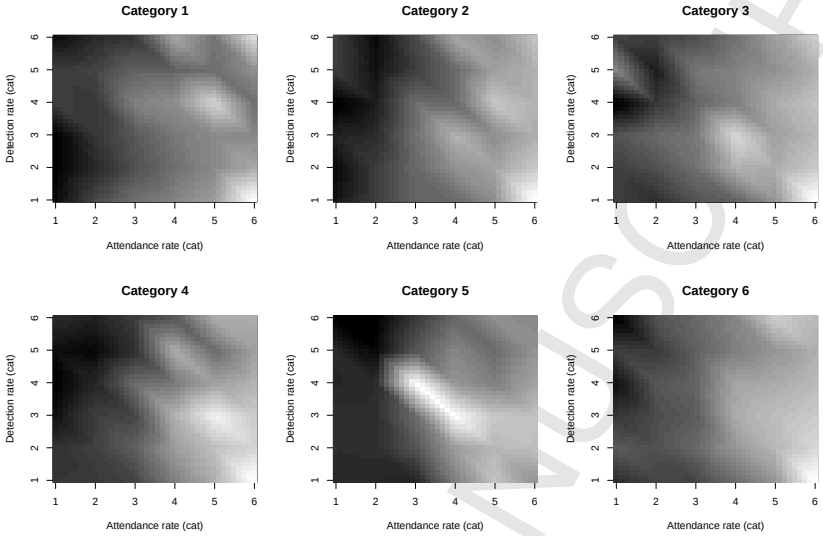

Figure 5: Trivariate interaction effect for the six categories of the Mortality rate. Category 1: Mortality rates $\leqslant$ percentil 10 ($P10$); Category 2: Mortality rates in $]P10, P25]$; Category 3: Mortality rates in $]P25, P50]$; Category 4: Mortality rates in $]P50, P75]$; Category 5: Mortality rates in $]P75, P90]$; Category 6: Mortality rates $> P90$.

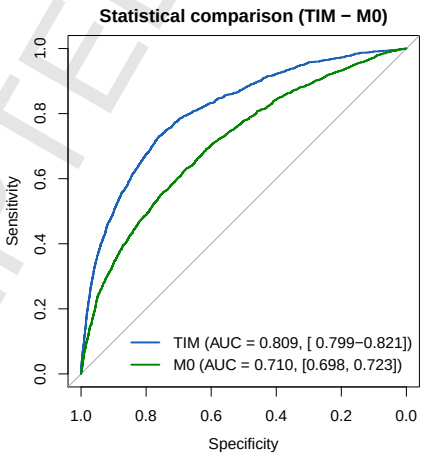

Figure 6: ROC curves and corresponding AUCs for TIM and M0 models. In brackets, 95% Credible Intervals for the AUCs.

4.1. Computational aspects

Regardless of the approach chosen, MRF or higher dimensional tensor product splines, the estimation of the higher-order term is a cumbersome process, as pointed out in Fahrmeir et al. (2013), and can be seen as a computational challenge. Besides the considerable number of observations combined with the complexity of the predictor components, the BayesX software has a high performance when dealing with such complexity, as proved in Duarte et al. (2014b) and Duarte et al. (2014a). This kind of complexity was augmented by the introduction of the trivariate interaction term, demanding huge processing memory, more than fast processors, which was not a hurdle for the BayesX performance. The models compared in this study, with and without the high interaction term were performed using an 8x Intel 3.33GHz machine, with 22Gb of RAM. The model with the trivariate interaction term converged in 15 hours, after 17 iterations.

The script to perform the trivariate interaction model in BayesX can be seen in the Appendix B.

5. Discussion

The breast cancer screening program data analyzed in this study was collected over more than 20 years, where technical changes had occurred over this period of time. Factors such as whether a film-screen or digital mammography was used, or a whether a skilled technician and radiologist performed and reported a woman's exam, can influence the results registered in the database. This study not only looks into the most recent data collected from the last visit made by a woman to the screening program, but tries to address some of these complexities inherent to a screening program by introducing a trivariate interaction term of the attendance, detection and mortality rates in a structured geoadditive predictor.

With the inclusion of the trivariate interaction term, the proposed TIM model identifies family history, in particular women with a sister, mother or daughter with the disease, as the only statistically significant fixed effect that contributes to an increased factor risk for breast cancer. This model (TIM) disregarded being pregnant and having breastfed as protective and increasing risk factors of the disease, respectively, as did the simple model (M0). Although both models agreed on how these factors influence the diagnose of the disease, the lack of information such as the duration of breast feeding for each child, the mothers age at first childs birth, or the number of children she had, can rightly validate the non-significant results obtained in the TIM model.

For the non-linear effects of the continuous covariates it is worth noting that the main differences between the two models, TIM and M0, are the effects of the purchasing power index, for values greater than 60. Again, this could be a result inherent to the dynamics of the screening program, where a woman can migrate internally or externally among the municipalities while being part of the screening program, or belong to a municipality in the screening program different from the one that she lived all her life, thus being influenced by the risk factors of the municipality of residence and not by the ones present in the municipality where she is registered in the screening program.

The most important result achieved with the introduction of the trivariate interaction in the model concerns the ability of this model (TIM) to detect the significance of the spatial effects across the central part of Portugal. The inclusion of the triple interaction together with the risk factors considered in this study allowed the model to clearly detect the spatial patterns, where those characteristics intrinsic to a woman and of her place of residence were taken into consideration.

Nevertheless, challenges remain to be addressed in future research such as considering extensions of distributional models to censored responses which will allow for women that have not yet reached menopause and those who left the screening program before reaching menopause to be included.

6. Conclusion

As a final note, the inclusion of the trivariate interaction term proved to be important in the control of possible bias that may arise from data of a screening program and that are not directly controlled by the program design, such as the coverage of the screening program. The proposed cube representation of the three covariates - attendance, coverage, and mortality rates - using a MRF prior proved to be an easy and effective way to represent the trivariate interaction term in the geoadditive predictor. The model performance increased, and the effects of the factors that can influence the diagnosis of breast cancer together with the spatial effects were captured.

7. Appendices

Appendix A.

The Cube function is presented here to create the 3D structure for the implementation of the trivariate interaction as a Markov random field term. The function inputs are: the datafile d; x1,x2 and x3 are the datafile covariates involved in the interaction. The outputs are datacat, a new datafile with the interaction covariates categories codes and cube, the adjacency matrix created as a graph file.

```
cube = function(d,x1,x2,x3,datacat,cube){
  #d is the datafile
  #x1 = the first
  #Create categories
  #x1
  x1c=rep(0,length(x1))
  q=quantile(x1,probs=c(0.1,0.25,0.5,0.75,0.9),na.rm=TRUE)
  x1c[x1<=q[1]]=1
  x1c[x1>q[1]&x1<=q[2]]=2
```

```
x1c[x1>q[2]&x1<=q[3]]=3
x1c[x1>q[3]&x1<=q[4]]=4
x1c[x1>q[4]&x1<=q[5]]=5
x1c[x1>q[5]]=6

#x2
x2c=rep(0,length(x2))
q=quantile(x2,probs=c(0.1,0.25,0.5,0.75,0.9),na.rm=TRUE)
x2c[x2<=q[1]]=1
x2c[x2>q[1]&x2<=q[2]]=2
x2c[x2>q[2]&x2<=q[3]]=3
x2c[x2>q[3]&x2<=q[4]]=4
x2c[x2>q[4]&x2<=q[5]]=5
x2c[x2>q[5]]=6

#x3
x3c=rep(0,length(x3))
q=quantile(x3,probs=c(0.1,0.25,0.5,0.75,0.9),na.rm=TRUE)
x3c[x3<=q[1]]=1
x3c[x3>q[1]&x3<=q[2]]=2
x3c[x3>q[2]&x3<=q[3]]=3
x3c[x3>q[3]&x3<=q[4]]=4
x3c[x3>q[4]&x3<=q[5]]=5
x3c[x3>q[5]]=6

#Create the combinations
z=expand.grid(x1 = c(1:max(x1c)), x2 = c(1:max(x2c)),x3 = c(1:max(x3c)))

#write file with the created categories
dcat=data.frame(d,x1c,x2c,x3c,cat=rep(0,nrow(d)))
for(i in 1:nrow(dcat))
```

```
{l=(z$x1==dcat$x1c[i]&z$x2==dcat$x2c[i]&z$x3==dcat$x3c[i])
 if(sum(l[l==TRUE])==1)
 {dcat$cat[i]=as.numeric(rownames(z[l,]))}
 else{dcat$cat[i]=NA} }
 write.table(dcat, datacat,row.names = F,quote=F)

#create cube
adjmat=matrix(data = 0, nrow = nrow(z), ncol = nrow(z),
 dimnames = list(rownames(z), rownames(z)))

region=0
for (i in 1:nrow(z)) {
  region=region+1
  l11=(z$x1==z$x1[i]-1&z$x2==z$x2[i]&z$x3==z$x3[i])
  l12=(z$x1==z$x1[i]+1&z$x2==z$x2[i]&z$x3==z$x3[i])
  l21=(z$x1==z$x1[i]&z$x2==z$x2[i]-1&z$x3==z$x3[i])
  l22=(z$x1==z$x1[i]&z$x2==z$x2[i]+1&z$x3==z$x3[i])
  l31=(z$x1==z$x1[i]&z$x2==z$x2[i]&z$x3==z$x3[i]-1)
  l32=(z$x1==z$x1[i]&z$x2==z$x2[i]&z$x3==z$x3[i]+1)
  nei=c(as.numeric(rownames(z[l11,])),as.numeric(rownames(z[l12,])),
  as.numeric(rownames(z[l21,])),
  as.numeric(rownames(z[l22,])),as.numeric(rownames(z[l31,])),
  as.numeric(rownames(z[l32,])))
  j=0
  while (j <=length(nei)){
    adjmat[i,nei[j]]=adjmat[nei[j], i]=-1;j=j+1
  }
  adjmat[i,i]=length(nei)
}
class(adjmat) = "gra"
```

```
  lcube <- as.integer(rownames(adjmat))
  S <- length(lcube)
  write(S, file)
  for (i in 1:S) {
    write(lcube[i], cube, append = TRUE)
    write(adjmat[i, i], cube, append = TRUE)
    ind <- which(adjmat[i, ] == -1)-1
    write(ind, cube, ncolumns = length(ind), append = TRUE)
  }
}
```

Appendix B.

The BayesX script used to perform the trivariate interaction model is as follows.

```
%create a map object
map m

%read the Portugal Central region map
m.infile,graph using C:\temp\PTCentralN.gra

%create a map object to assign the graph file (cube) created using the Cube function
map c
c.infile,graph using C:\temp\cube.gra

%create a dataset
dataset d

%read the data
%datacat.txt is the datafile augmented with the trivariate categories codes,
%created with the Cube function.
```

```
d.infile using C:\temp\datacat.txt

%remove women without menopause register
d.drop if MENOPAUSE=0

%create the categories for the "Family history" covariate
%reference category GCANFAM=0

d.generate GCANFAM1= 1*(GCANFAM=1)
d.generate GCANFAM2= 1*(GCANFAM=2)
d.generate GCANFAM3= 1*(GCANFAM=3)

%create a regression object using the mixed model methodology
remlreg modelname

%indicate the path to the directory where will be save the results
modelname.outfile = C:\temp\Results\resfit
%fit the model
modelname.regress DIAGNOSTIC = PREGNANCYSTATUS + ORALCONTRACEPTIVES
                             + NURSINGSTATUS + GCANFAM1 + GCANFAM2
                             + GCANFAM3 + BIRTHYEAR(psplinerw2)
                             + PPI(psplinerw2) + MENOPAUSE(psplinerw2)
                             + MENARCHE(psplinerw2)
                             + MENARCHE*MENOPAUSE(pspline2dimrw1)
                             + MUNICIPALITY(spatial, map=m)+ MUNICIPALITY(random)
                             + 3DCATEGORY(spatial,map=c),
                             family=binomial lowerlim=0.01 eps=0.0005 using d

%delete the dataset d
drop d
```

Acknowledgments

This work was financed by POPH-QREN, the European Social Fund and national funds MCTES - Portuguese Ministry of Science, Technology and Higher Education [grant number [SFRH/BD/64761/2009]; Spanish Ministry of Science and Innovation grant number [MTM2015-69068-REDT]; 2014 Competitive Call for "Acciones Conjuntas Hispano-Alemans/PPP Spain" in cooperation with the Foundation for the International Promotion of Spanish Universities Unversidad.es; Ministry of Economy and Competitiveness (SPAIN) and the European Regional Development Fund (FEDER) grant number [MTM2014-52975-C2-1-R].

References

Belitz, C., Brezger, A., Kneib, T., Lang, S., Umlauf, N., 2015. BayesX: Software for Bayesian Inference in Structured Additive Regression Models. URL: `http://www.BayesX.org/`. version 3.0.2.

Brezger, A., Lang, S., 2006. Generalized structured additive regression based on Bayesian P-splines. Computational Statistics and Data Analysis 50, 967–991.

Duarte, E., de Sousa, B., Cadarso-Suárez, C., Rodrigues, V., Kneib, T., 2014a. Structured additive regression (STAR) models applied in analysis of breast cancer risk in central portugal, in: Kneib, T., Sobotka, F., Fahrenholz, J., Imer, H. (Eds.), Proccedings of the 29th International Workshop on Statistical Modelling, Statistical Modelling Society. pp. 111–116.

Duarte, E., de Sousa, B., Cadarso-Suárez, C., Rodrigues, V., Kneib, T., 2014b. Structured additive regression (STAR) modeling of age of menarche and the age of menopause in breast cancer screening program. Biometrical Journal 56, 416–427.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing using B-splines and penalties (with comments and rejoinder). Statistical Science. 11, 89–121.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: Models, Methods and Applications. Springer-Verlag Berlin Heidelberg.

Health & Consumer Protection Directorate-General, 2006. European guidelines for quality assurance in breast cancer screening and diagnosis - Fourth edition. Technical Report. European Commission - Office for Official Publications of the European Communities: Luxembourg.

Kneib, T., 2006. Mixed model based inference in structured additive regression. Ph.D. thesis. Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München.

Lang, S., Brezger, A., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics. 13, 183–212.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. Chapman & Hall, London.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Rodrigues, V., 1993. Geographical epidemiology of cancer. Application of empirical Bayesian estimation to the analysis of the geographical distribution of mortality from malignant tumors in Portugal. Ph.D. thesis. University of Coimbra, Faculty of Medicine, Portugal.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields. Chapman & Hall/CRC, Boca Raton, FL.

Wood, S., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62, 1025–1036.