Point pattern analysis with spatially varying covariate effects, applied to the study of cerebrovascular deaths

Jony A Pinto Junior,[1,2*] Dani Gamerman,[2], Marina S Paez[2], Regina H. F. Alves[2]

[1] Universidade Federal Fluminense,
[2] Universidade Federal do Rio de Janeiro

This article proposes a modelling approach for handling spatial heterogeneity present in the study of the geographical pattern of deaths due to cerebrovascular disease. The framework involves a point pattern analysis with components exhibiting spatial variation. Preliminary studies indicate that mortality of this disease and the effect of relevant covariates do not exhibit uniform geographic distribution. Our model extends [23] by allowing for spatial variation of the effect of non-spatial covariates. A number of relative risk indicators are derived by comparing different covariate levels, different geographic locations or both. The methodology is applied to the study of the geographical death pattern of cerebrovascular deaths in the city of Rio de Janeiro. The results compare well against existing alternatives, including fixed covariate effects. Our model is able to capture and highlight important data information that would not be noticed otherwise, providing information that is required for appropriate health decision making.

1 Introduction

Geo-referenced data is very common in the work of researchers from many areas such as Ecology, Geography, Seismology and Epidemiology. The desire to investigate connections between the point pattern and covariates that are possibly associated with the event of interest arises naturally.

Whatever the area, the most frequent question typically asked are: is there a spatial pattern governing the occurrence of the event of interest? Are there any variables that affect this pattern variation? For example, a cardiologist may be interested in the effect of individual (sex, age, etc) and non-individual (location-related) factors may have in the pattern of cerebrovascular deaths. This would allow him/her, for instance, to design plans of action to intervene where larger death risks are observed. Note that larger risks may be associated with a given combination of individual and non-individual factors.

Models for these types of data are usually built with point pattern processes. A good survey of its probabilistic properties is given by [7] where estimation

based on observed point patterns is described in [9]. The latter approach was initiated with exploratory methods based on distances, such as the Ripley's F, G and K functions. These methods did not specify likelihoods, which made hard to compare different alternatives. Likelihood-based methods include (non-)homogeneous Poisson processes, Cox processes and log-Gaussian Cox processes.

[4] suggested an important class of hierarchical models along these lines, with the introduction of covariate effects. Many papers follow this route presenting variations and/or extensions. [23] and [10] incorporate spatial effects and individual covariates. They showed the importance of the inclusion of non-spatial covariates, by adding their effect to the intensity rate in a standard log-linear fashion. They assumed the effects of the covariates to be fixed over space. This can be appropriate in many practical situations but may not be realistic assumption for datasets exhibiting large spatial heterogeneity. This leads naturally to the quest for point pattern models that allow for spatial variation not only of the baseline intensity rate but also of the effects of the covariates.

Thus, the purpose of this work is to propose a point pattern model that allow for the spatial variation of the regression coefficients. The regressors may be spatial, non-spatial or even an interaction between these two sets of variables. Gaussian processes (GP) are the main tool to accommodate spatial variations, by ensuring that this variation is smooth over space. Inference will be performed under the Bayesian point of view.

Disease epidemiology is one of the main areas of application of point pattern studies. Knowledge of the geographical distribution of the disease may have an important effect over its understanding and control. Chronic diseases are the main cause of deaths worldwide. The scene is even more worrying when underdeveloped countries are considered. According to [21], 80% of deaths in these countries are attributed to chronic diseases. Cardiovascular diseases are the largest cause of deaths and cerebrovascular diseases (CVD, hereafter) play a central role among them. Brazil has the largest CVD deaths rate in Latin America [25] and if the Caribbean is included, only Guiana, Jamaica and Haiti have larger death rates [21].

The epidemiologic literature indicates that the CVD death pattern does not exhibit uniform geographic distribution even in developed countries such as Canada [20] and USA [28]. These studies also show that this distribution does not seem to be related to classic risk factors, such as arterial hypertension and smoking [17].

There are a number of studies indicating the relevance of covariate information in the CVD deaths pattern. A large scale review of studies published between 1960 and 1993 conducted by [16] showed that individuals at lower strata in the social scale have higher death rates. [19] showed negative correlation between socioeconomic factors and cardiovascular diseases. [18] also showed negative correlation between mortality and socioeconomic levels, specially schooling. [22] also indicate that socioeconomic inequalities in Brazil play an important role in the mortality pattern of CVD in Brazil.

Our interest lies in understanding and explaining the CVD deaths pattern in the city of Rio de Janeiro. The database contains the location of the homes

of Rio de Janeiro inhabitants that had CVD indicated as primary death cause in their death certificates. Our analysis involves the sample of 18,237 individual deaths with complete individual records. Figure 1 presents the residential location of these deaths. The figure is visually dominated by the population demographics and any sensible analysis must incorporate that information. Also, it is worth recalling that the population distribution over the city is influenced by topography and, more importantly, by the social stratification, with wealthiest inhabitants occupying the southernmost part of the city, facing the Atlantic Ocean.

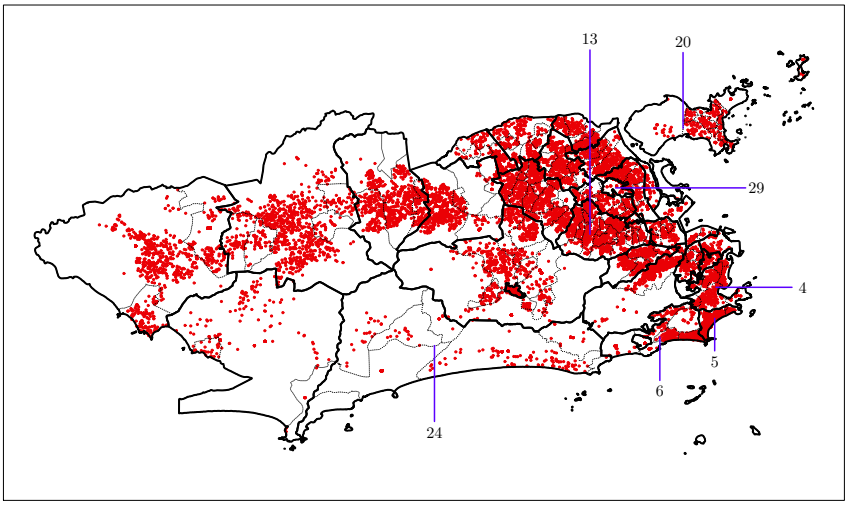

Figure 1: Residential locations of CVD deaths between 2002 and 2007 in the city of Rio de Janeiro. The city is divided in 33 administrative regions (RA) (shown in thick lines) and 160 boroughs (thin lines). The numbers of the RA´s that will be discussed in the sequel are also shown (in blue).

Modeling the spatial distribution of CVD deaths is important for implementation of efficient policy making strategies. Individual-specific covariates such as age, schooling and marital status are known to intervene and must be considered. We are also interested in the understanding and quantification of the impact the socio-cultural background has on this mortality pattern. An obvious first step in this process is the standardization alluded above to ensure fair comparisons. This operation ensures the effect of the population size is removed from the risk analysis.

The city of Rio de Janeiro has a well-known socioeconomic disparity between

its regions. Thus, a very heterogeneous spatial distribution of mortality is to be expected; regions with lower-strata individuals may present larger death rates and the death pattern of these regions may differ from the death pattern of wealthier regions. This may be explained by observed and also by unobserved factors. In summary, this problem may induce geographic heterogeneity of the explanatory processes of the response, requiring models that are flexible enough to capture this spatial variation.

The availability of geo-referenced data with the precise spatial location (home address) of the event of interest (CVD individual death) calls for a study of point patterns. Connections with explanatory variables that may be associated with the outcome of interest must also be established. This would allow the identification of effects that these conditions may have on CVD deaths.

Section 2 describes the hierarchical formulation of the model. A number of inference procedures, such as relative risk evaluation, are also discussed in this section. Section 3 describes the main results of the data analysis. Section 4 concludes the paper with some directions for further work.

2 Modeling point patterns with space-varying regression coefficients

Spatial point pattern processes are useful models for the statistical analysis of geo-referenced observed point patterns. They are stochastic processes denoted $X = \{X(\boldsymbol{s}) : \boldsymbol{s} \in S\}$, where $S \subseteq \Re^d, d > 0$ and

$$X(\boldsymbol{s}) = \begin{cases} 1, & \text{if the event of interest occurred in } \boldsymbol{s}. \\ 0, & \text{otherwise,} \end{cases}$$

The most common point pattern process is the (non-homogeneous) Poisson process with intensity function denoted by $\Lambda(\cdot) = \{\Lambda(\boldsymbol{s}) : \boldsymbol{s} \in S\}$. The notation used here will be

$$X \sim PP\left(\Lambda(\cdot)\right).$$

When $d = 2$, X is a spatial process on the plane. A realization of X can be unequivocally identified with a occurrence set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, $\boldsymbol{s}_i \in S$, $\forall i$ and $n \geq 0$, where all observed events take place.

[4] proposed a point process model where the only relevant covariates are related to the locations, denoted by $z(\boldsymbol{s})$. However, many areas of application of point process possess covariates associated with the individuals, denoted by $\boldsymbol{v}$. Models for this more general scenario were proposed by [23], where effects of all covariates are fixed over space.

Assume that a p_1-dimensional vector of spatial covariates $\boldsymbol{z}(\boldsymbol{s}) = (z_1(\boldsymbol{s}), \ldots, z_{p_1}(\boldsymbol{s}))'$, a p_2-dimensional vector of individual covariates $\boldsymbol{v} = (v_1, \ldots, v_{p_2})'$ and a p_3-dimensional vector of interaction between individual and spatial covariates are considered. Consider a collection $\{X_v(\boldsymbol{s}) : \boldsymbol{v} \in \mathcal{V}\}$ of Poisson point patterns and a corresponding collection $\Lambda(\cdot) = \{\Lambda_v(\boldsymbol{s}) : \boldsymbol{s} \in S \text{ and } \boldsymbol{v} \in \mathcal{V}\}$ of

intensity functions, where $S \subseteq \Re^d$ and $\mathcal{V}$ is the space of all individual covariate configurations. The likelihood is given by

$$L(\Lambda(\cdot)) = \prod_{v \in \mathcal{V}} L(\Lambda_v(\cdot)), \text{ where } L(\Lambda_v(\cdot)) = \prod_{i=1}^{n_v} \Lambda_v(\boldsymbol{s}_{v,i}) \exp\left\{ - \int_S \Lambda_v(\boldsymbol{s}) d\boldsymbol{s} \right\}, \tag{1}$$

where $\boldsymbol{s}_{v,i}$ is the location of the ith event, for $i = 1, \ldots, n_v$, and n_v is the number of events observed for the configuration $\boldsymbol{v}$ of the individual covariate.

2.1 Space-varying effects

It is well known that locations in Rio de Janeiro are associated with different socio-economic backgrounds. It is reasonable to expect that death patterns may be affected by this variation and thus it is recommended that effects are allowed to change over space. For example, some poorer locations may exhibit younger death patterns than other regions. This reasoning may be extended to other individual characteristics to address similar questions of interest to cardiologists; eg, is marital status more effective in protecting against CVD deaths in poorer regions? These questions of interest can only be appropriately answered by assuming interaction between the effects of covariates and space. Thus, point pattern models with spatially-varying effects of covariates must be considered.

The full model formulation is given by

$$\begin{aligned}
X_v &\sim PP(\Lambda_v(\cdot)), \forall\, \boldsymbol{v} \in \mathcal{V}, && (2)\\
\Lambda_v(\boldsymbol{s}) &= r(\boldsymbol{s}, \boldsymbol{v}) \lambda(\boldsymbol{s}, \boldsymbol{v}), \forall \boldsymbol{s} \in S \text{ and } \boldsymbol{v} \in \mathcal{V}, && (3)\\
\log \lambda(\boldsymbol{s}, \boldsymbol{v}) &= \boldsymbol{v}' \boldsymbol{\alpha}(\boldsymbol{s}) + \boldsymbol{z}(\boldsymbol{s})' \boldsymbol{\beta} + (\boldsymbol{v} \odot \boldsymbol{z}(\boldsymbol{s}))' \boldsymbol{\delta}(\boldsymbol{s}) + w(\boldsymbol{s}), && (4)\\
\boldsymbol{\alpha}_l(\cdot) &\sim GP(\mu_{\alpha_l}, \tau_{\alpha_l}, \rho_{\gamma_{\alpha_l}}), l = 1, \ldots, p_2, && (5)\\
\boldsymbol{\delta}_l(\cdot) &\sim GP(\mu_{\delta_l}, \tau_{\delta_l}, \rho_{\gamma_{\delta_l}}), l = 1, \ldots, p_3, && (6)\\
w(\cdot) &\sim GP(\mu_w, \tau_w, \rho_{\gamma_w}). && (7)
\end{aligned}$$

Equations (2)-(7) define a generalization of log-Gaussian Cox process, obtained after allowing effects of all components involving individual covariates to vary over space. [23] model is obtained as the special case where regression coefficients $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\delta}(\cdot)$ are fixed and do not vary over space.

Here, the coefficients $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\delta}(\cdot)$ and the intercept $w(\cdot)$ are allowed to vary over space according to isotropic Gaussian processes. This is in line with standard GP assumptions for the intercept of log-Gaussian Cox processes, when no covariates are present. A process $Y(\cdot)$, defined in S, is said to be isotropic Gaussian if $\forall n > 1$ and any set of locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \in S$, $(Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_m))' \sim N(\mu \mathbf{1}, \tau^{-1} \boldsymbol{R}_\gamma)$, denoted $Y(\cdot) \sim GP(\mu, \tau, \rho_\gamma)$, in which $\boldsymbol{R}_\gamma$ is a correlation matrix with elements $\boldsymbol{R}_{i,j} = \rho_\gamma(||\boldsymbol{s}_i - \boldsymbol{s}_j||)$ defined through a correlation function ρ_γ, depending on $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ only through their distance, for $i, j = 1, \ldots, m$. The correlation function usually decays monotonically with distance to reflect larger correlation between neighboring locations, thus ensuring smoothness of the spatial variation of the coefficients. The model above assumes independent GP for

each coefficient for parsimony but interactions between them can be easily built in [15].

The model is completed with prior distributions for the hyperparameters and the coefficients that are fixed over space. These coefficients are attributed Gaussian priors with suitably defined means and covariance matrices. Note that the intensity assumes a multiplicative decomposition of the intensity function with $r(\boldsymbol{s}, \boldsymbol{v})$ representing a known offset, usually required for standardization. Typical examples of offset are population size or populational density, used in our study. This is required to ensure that comparisons will be made at individual levels, as required for meaningful interpretation of results.

Inference means providing results about the values of the parameters and their transformations. One of the most important parameter transformations are given by relative risks. These are vital for comparing death rates between different locations, between different covariate configurations and different combinations of locations and covariate configurations.

2.2 Relative risks

Quantification of risk is one of the objectives of a point pattern study. In health studies, it is a fundamental tool. Variables such as age, sex and marital status of the deceased individuals are considered in our study. Questions of interest include: does location matter for CVD deaths rates of older married males? how much more likely to die from CVD are older married males in a given location than single older females from the same location?

Relative risks allow for the quantification of these comparisons. Thus, the relative risk between a given individual in location $\boldsymbol{s}_1$ and covariate configuration $\boldsymbol{v}_1$ and another individual in location $\boldsymbol{s}_2$ and covariate configuration $\boldsymbol{v}_2$ is

$$RR(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{\lambda(\boldsymbol{s}_1, \boldsymbol{v}_1)}{\lambda(\boldsymbol{s}_2, \boldsymbol{v}_2)} = \frac{\exp\left\{\boldsymbol{z}(\boldsymbol{s}_1)\boldsymbol{\beta} + \boldsymbol{v}_1'\boldsymbol{\alpha}(\boldsymbol{s}_1) + (\boldsymbol{v}_1 \odot \boldsymbol{z}(\boldsymbol{s}_1))'\boldsymbol{\delta}(\boldsymbol{s}_1) + w(\boldsymbol{s}_1)\right\}}{\exp\left\{\boldsymbol{z}(\boldsymbol{s}_2)\boldsymbol{\beta} + \boldsymbol{v}_2'\boldsymbol{\alpha}(\boldsymbol{s}_2) + (\boldsymbol{v}_2 \odot \boldsymbol{z}(\boldsymbol{s}_2))'\boldsymbol{\delta}(\boldsymbol{s}_2) + w(\boldsymbol{s}_2)\right\}}. \tag{8}$$

Note that populational densities are removed from risk evaluations to ensure comparisons at the level of the individuals.

Special cases of (8) are obtained if comparisons are made within locations, ie, for individuals with different covariate configurations ($\boldsymbol{v}_1 \neq \boldsymbol{v}_2$) but same locations ($\boldsymbol{s}_1 = \boldsymbol{s}_2$). Further simplification in the comparison is obtained if it is additionally assumed that covariate configurations differ only in the lth covariate, $l = 1, \ldots, p_2$. If v_l is a binary variable with $v_l = 1$ indicating presence and $v_l = 0$ indicating absence of the factor, the relative risk becomes

$$RR(\boldsymbol{s}) = e^{\alpha_l(\boldsymbol{s})}. \tag{9}$$

Analogously, if v_l is a continuous regressor, an increase of m units in this covariate will result in a multiplicative effect of $e^{m\alpha_l(\boldsymbol{s})}$ in the expected number of CVD deaths in this location.

2.3 Discretizing log-Gaussian Cox processes

The likelihood function for model (2)-(7) is given in compact form in (1). It depends on the uncountable random functions $\boldsymbol{\alpha}(\cdot)$, $\boldsymbol{\delta}(\cdot)$ and $w(\cdot)$. This poses a difficult problem to handle. Exact solutions are only available in limited cases and even then, they depend on a number of issues. Some of these issues are associated with the dimension of the number of occurrences, which in our cases is fairly large ($O(10^4)$).

Thus, alternatives must be sought. Reasonable options may be obtained through approximations at the modeling level. [27] assumed that the region of interest S can be partitioned into sub-regions $\{S_1, \ldots, S_N\}$ satisfying $\bigcup_{k=1}^{N} S_k = S$ and $S_k \cap S_{k'} = \emptyset$, for $k \neq k'$, and S_k has centroid located at $\boldsymbol{s}_k^*$, $k = 1, \ldots, N$ where

$$\boldsymbol{\alpha}(\boldsymbol{s}) = \boldsymbol{\alpha}_{[k]}, \boldsymbol{\delta}(\boldsymbol{s}) = \boldsymbol{\delta}_{[k]} \quad \text{and} \quad w(\boldsymbol{s}) = w_{[k]}, \forall \boldsymbol{s} \in S_k, k = 1, \ldots, N.$$

Assuming further that $r(\boldsymbol{s}, \boldsymbol{v}_j) = r_{k,j}, \forall \boldsymbol{s} \in S_k$, enforces homogeneity of the intensity rate within the regions and

$$\lambda(\boldsymbol{s}, \boldsymbol{v}_j) = \lambda_{k,j} = \exp\{\boldsymbol{v}_j'\boldsymbol{\alpha}_{[k]} + \boldsymbol{z}_{[k]}'\boldsymbol{\beta} + (\boldsymbol{v}_j \odot \boldsymbol{z}_{[k]})'\boldsymbol{\delta}_{[k]} + w_{[k]}\}, \forall \boldsymbol{s} \in S_k, k = 1, \ldots, N, \tag{10}$$

where $j = 1, \ldots, J$ and $J = \#\mathcal{V}$ is the cardinality of the observed covariate space $\mathcal{V}$.

The integral in (1) can be rewritten as

$$\int_S r(\boldsymbol{s}, \boldsymbol{v}_j)\lambda(\boldsymbol{s}, \boldsymbol{v}_j)d\boldsymbol{s} = \sum_{k=1}^{N} r_{k,j}\lambda_{k,j}|S_k|, \tag{11}$$

where $|S_k|$ is the volume of the kth sub-region, for $k = 1, ..., N$.

The spatial discretization of the intensity above can also be found elsewhere. [13] used it to make inference in point processes over time. Similar approaches for the spatial domain may also be found in [6], for example. If interest lies in the effect of a covariate at the region level rather than at a specific location, the discretization does not cause any limitation in the results. [32] showed that the posterior distributions of the intensities are well approximated and converge to the posterior distribution of the exact, continuously-varying intensity when the volumes of the sub-regions tend to 0.

The aim of our analysis here is the understanding of the spatial distribution of the mortality over the city and the way that auxiliary information can be used to describe this pattern. This knowledge is required for efficient policy making by the health officials in the central administration of the city. Decision in this case is rarely, if ever, taken at a point level. Thus, the assumption of homogeneity of the mortality within sub-regions does not seem to affect the main goals of the analysis. Of course, the number and the sizes of the homogeneous sub-regions must be appropriately chosen to ensure as much fidelity to the data variation as possible. Natural candidates for this partition for the specific case of this analysis are the partition into 33 RA's and the finer partition into 160

boroughs. The RA structure is used for policy making by the health authorities in the city of Rio de Janeiro, but comparisons with results at the borough level will also be made.

2.4 Inference

The vector of coefficients of non-spatial covariates will be denoted $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_{[1]}, \ldots, \boldsymbol{\alpha}'_{[N]})'$, where $\boldsymbol{\alpha}_{[k]} = (\alpha_{1[k]}, \ldots, \alpha_{p_2[k]})'$, $k = 1, \ldots, N$. Similarly, $\boldsymbol{\delta} = (\boldsymbol{\delta}'_{[1]}, \ldots, \boldsymbol{\delta}'_{[N]})'$, where $\boldsymbol{\delta}_{[k]} = (\delta_{1[k]}, \ldots, \delta_{p_3[k]})'$, $k = 1, \ldots, N$ and $\boldsymbol{w} = (w_{[1]}, \ldots, w_{[N]})'$. Different prior distributions could now be used for these model components. An example is the conditional autoregressive (CAR) prior of [5]. However, the GP prior is kept to retain characteristics of the continuum model and enable easy change of sub-region structure. We have retained the point pattern approach over the areal data for the same reason, despite the similarities between them.

Thus, application of the likelihood (1) in model (2)-(7) with discretizations (10) and (11) leads to the full likelihood

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{w}) = & \prod_{j=1}^{J} \prod_{i=1}^{n_j} r_{i,j} \exp\{\boldsymbol{v}'_j \boldsymbol{\alpha}_{[i]} + \boldsymbol{z}'_{[i]} \boldsymbol{\beta} + (\boldsymbol{v}_j \odot \boldsymbol{z}_{[i]})' \boldsymbol{\delta}_{[i]} + w_{[i]}\} \\ & \times \exp\left\{ - \sum_{k=1}^{N} r_{k,j} \exp\{\boldsymbol{v}'_j \boldsymbol{\alpha}_{[k]} + \boldsymbol{z}'_{[k]} \boldsymbol{\beta} + (\boldsymbol{v}_j \odot \boldsymbol{z}_{[k]})' \boldsymbol{\delta}_{[k]} + w_{[k]}\} |S_k| \right\} \\ & \propto \exp\left\{ \sum_{j=1}^{J} \sum_{i=1}^{n_j} \boldsymbol{v}'_j \boldsymbol{\alpha}_{[i]} + \boldsymbol{z}'_{[i]} \boldsymbol{\beta} + (\boldsymbol{v}_j \odot \boldsymbol{z}_{[i]})' \boldsymbol{\delta}_{[i]} + w_{[i]} \right. \\ & \left. - \sum_{j=1}^{J} \sum_{k=1}^{N} r_{k,j} |S_k| \exp\{\boldsymbol{v}'_j \boldsymbol{\alpha}_{[k]} + \boldsymbol{z}'_{[k]} \boldsymbol{\beta} + (\boldsymbol{v}_j \odot \boldsymbol{z}_{[k]})' \boldsymbol{\delta}_{[k]} + w_{[k]}\} \right\}. \end{aligned} \tag{12}$$

The main objective here is to perform inference about the likelihood parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{w})$. Additionally, one may also be interested in the hyperparameters $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma})$ associated with the GP priors for the likelihood parameters, where $\boldsymbol{\mu} = (\mu_{\alpha_1}, \ldots, \mu_{\alpha_{p_2}}, \mu_{\delta_1}, \ldots, \mu_{\delta_{p_3}}, \mu_w)'$, $\boldsymbol{\tau} = (\tau_{\alpha_1}, \ldots, \tau_{\alpha_{p_2}}, \tau_{\delta_1}, \ldots, \tau_{\delta_{p_3}}, \tau_w)'$ and $\boldsymbol{\gamma} = (\gamma_{\alpha_1}, \ldots, \gamma_{\alpha_{p_2}}, \gamma_{\delta_1}, \ldots, \gamma_{\delta_{p_3}}, \gamma_w)'$.

The prior independence assumption for regression coefficients leads to the prior distribution

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{w} | \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma})\ \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\mu}) \pi(\boldsymbol{\tau}) \pi(\boldsymbol{\gamma}). \tag{13}$$

The corresponding posterior distribution is

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma} | \{s\}_{i=1}^{n}) \propto & L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{w}; \{s\}_{i=1}^{n}) \pi(\boldsymbol{\beta}) \\ & \times \pi(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{w} | \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma}) \pi(\boldsymbol{\mu}) \pi(\boldsymbol{\tau}) \pi(\boldsymbol{\gamma}), \end{aligned} \tag{14}$$

where $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{w}; \{s\}_{i=1}^{n})$ is the full likelihood in (12) and $n = \sum_j n_j$ is the total number of events observed for all covariate configurations.

The posterior distribution (14) is the basis for inference about all model unknowns, and for their transformations, specially relative risks presented in Sub-section 2.2. The prior distribution for the likelihood parameters is given in (5)-(6)-(7). Whatever the prior used for the hyperparameters, the posterior distribution will not be the tractable analytically and approximations will be required. There are a few methods available for this approximation. We have opted for MCMC methods [14].

3 Results

This section presents the results of the analysis of the CVD data in the city of Rio de Janeiro with model (2)-(7) discretized according to the RA structure. Previous studies suggested the relevance of individual and socioeconomic (location-related) variables. Indication of socioeconomic status is location-driven and is provided by the human development index (HDI) compiled by the United Nations [31]. This index is a summary of human development of a given location by weighing its health, education and wealth indicators, is only available at the RA level, and is denoted here by z. Individual-specific covariates are: v_1, age at death, v_2, schooling indicator (1, for individuals with 8 or more years of education and 0, otherwise), v_3, gender (0 = female, 1 = male), and an interaction term between schooling and marital status with indicators

$$v_4 = \begin{cases} 1, & \text{if individual lives alone and had adequate education } (v_2 = 1), \\ 0, & \text{otherwise,} \end{cases}$$

$$v_5 = \begin{cases} 1, & \text{if individual does not live alone and did not have adequate education } (v_2 = 0), \\ 0, & \text{otherwise,} \end{cases}$$

and

$$v_6 = \begin{cases} 1, & \text{if individual does not live alone and had adequate education } (v_2 = 1), \\ 0, & \text{otherwise,} \end{cases}$$

and reference category given by individuals that live alone and did not have adequate education. Thus, $\boldsymbol{v} = (v_1, v_2, v_3, v_4, v_5, v_6)'$ is the vector of individual covariates and the number of different configurations of these variables is $J = \#\mathcal{V} = 136$. More details about this study may be found in [12].

In this model, $\boldsymbol{\alpha}_1$ measures the effect of age, $\boldsymbol{\alpha}_2$ measures the effect of schooling, $\boldsymbol{\alpha}_3$ measures the effect of gender and ($\boldsymbol{\alpha}_4$, $\boldsymbol{\alpha}_5$, $\boldsymbol{\alpha}_6$) measure the effect of the iteraction between marital status and schooling with the same GP prior parameters, where $\boldsymbol{\alpha}_l = (\alpha_{l[1]}, \ldots, \alpha_{l[33]})', l = 1, \ldots, 6$. The other likelihood parameters are β, measuring the effect of HDI, and w, measuring the remaining spatial effect. If the socio-economic factor HDI were able to capture the spatial heterogeneity, then the spatially-varying intercept $\boldsymbol{w}$ would be irrelevant and there would not be any spatial variation of the other regression coefficients.

The hyperparameters were given reasonably vague prior distributions: $\mu_x \sim N(0, 100), x = \alpha_1, \ldots, \alpha_4, w, \tau_x \sim G(1, 0.01), x = \alpha_1, \ldots, \alpha_4, w$ and $\beta \sim N(0, 100)$. The exponential correlation function $\rho(||\boldsymbol{s}_i - \boldsymbol{s}_j||; \gamma) = \exp\{-||\boldsymbol{s}_i - \boldsymbol{s}_j||/\gamma\}$ was

used with $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ representing distance between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and γ is the range parameter. Following [11], the prior for the range parameters were $\gamma_x \sim G(1, 0.3/\text{med}(d_{\boldsymbol{S}}))$, where $\text{med}(d_{\boldsymbol{S}})$ is the median of the distances between the 33 regions, for $x = \alpha_1, \ldots, \alpha_4, w$.

The offsets $r_{k,j}$ were taken populational density, given by

$$r_{k,j} = \frac{n_{k,j}}{|S_k|}, k = 1, \ldots, 33 \quad \text{and} \quad j = 1, \ldots, 136,$$

where $n_{k,j}$ is the population size for configuration j in region k and $|S_k|$ is the area of region k. Obtaining the counts $n_{k,j}$ is not an easy task. It requires knowledge of the population sizes of the regions for all 136 configurations. The time span of the data is evenly placed between 2000 and 2010, years where official national censuses were carried out in Brazil. Thus, simple averages between the 2 censuses counts were used to represent the population sizes. Note that offsets $r_{k,j}, k = 1, \ldots, 160$ and $j = 1, \ldots, 136$ must be calculated for each regional partition.

The fit of the above model showed weak identification of the range parameters γ. This well-known difficulty of spatial models was reported in many papers with a single Gaussian process, including [23]. They suggested instead to fix the ranges at the median of the distances. This problem is even more pronounced in models where there is a collection of latent Gaussian processes. Their suggestion did not imply significant changes in the resulting posterior distribution for the likelihood parameters in our analyses but stabilized results for the hyperparameters and was thus followed here.

Results were obtained via MCMC methods implemented in Winbugs [26]. Convergence was ascertained by using 2 chains with different starting values. Correlation between successive chain draws was alleviated by thinning at every 100 iterations, after a burn-in period of 5,000 draws. The resulting sample consisted on 2,000 draws.

Figure 2 shows the results for the main effect coefficients under both partitions for the space-varying (SVC) model. The age coefficients are consistently positive, indicating that intensity of CVD deaths increases with age, as expected. It is well known that age is a risk factor for cardiovascular diseases. The strength of this association significantly changes across the city. The largest age coefficients are encountered in the southerly, wealthiest regions (4, 5, 6 and 24) of the city. Thus, larger differences of CVD deaths intensities between older and younger people in these regions are found in these regions than in other regions. The socioeconomic variation across the city is highlighted by noticing that the smaller difference between ages is observed in RA 29, the slum-town Complexo do Alemão. The largest range of the credibility interval is observed at region 21, due to the scarcity of information (smaller population size) in this region.

The remaining individual covariates also exhibit noticeable spatial variation. In the RA partition, some regional effects are significant while others are not. The variations exhibited by these coefficients across regions is a compelling evidence in favor of allowing them to vary over space. The results show a protective effect of adequate education and being female, as expected. The

results obtained with the borough partition are basically in the same direction. The only notable discrepancies between posterior means under the different partitions are the age effects of regions 20 (Ilha do Governador), an island, and 29 (Complexo do Alemão).

The main difference between RA and borough results are the larger uncertainty bounds of the latter. This is to be expected: parameters that would share the same information at the RA level must now split it between boroughs. As a result, the schooling and gender effects still vary over space but here virtually all 95% credibility intervals include 0. Thus, these important effects become basically irrelevant, under the borough partition.

A summary that can be drawn here is that the borough granulation produces too many details that data is not able to inform about whereas a standard model, with fixed effects, unnecessarily removes important variation, needed for appropriately addressing the spatial variability present in the data. The DIC values [30] reinforce this point with smaller values for the model with RA partition (16.300) as opposed to larger values for the fixed effect (FE) model (17.260) and the borough partition (37.820). The results from the interaction covariates are not shown for conciseness.

The effect of age over CVD deaths intensity can also be assessed through relative risks $RR(k) = e^{10\alpha_{1[k]}}, k = 1, \ldots, 33$, associated with an increase in death intensity after an increase of 10 years of age. Inference for these risks is summarized in Figure 3. They range between an increase around 50% for Complexo do Alemão to around 270% for wealthier regions. Among the many possible risks comparisons, we have decided to single out a few within region comparisons. If one compares having adequate schooling against not having it, the median relative risk for RA 18 is 0.41 with $[0.36, 0.47]$ as 95% credibility interval while for RA 6 the median risk is 0.77 and the 95% limits are $[0.63, 0.94]$. These results indicate that there is a significant difference between the relative risks of these regions, with adequate schooling being more far less protective in the poorer RA 18. In the end of this scale, slum-towns regions do not exhibit relative risks that are significantly different from 1.

Further risk comparisons can be made with relative risks between regions for the more protective (against CVD deaths) covariate configuration. Figure 4 summarizes the results for comparisons of all regions against the wealthiest region 6. Relative risks range from 1 to 14 and increase as one moves northwardly, toward poorer regions. The highest relative risks were observed for the isolated, small regions consisting of slum-towns, where poor quality of life seems to be reflected into substantially higher risks. These regions however show the largest credibility intervals. So, the risks must be quantified with caution but they are definitely higher in these regions. On the other hand, the credibility intervals for the risks against the other wealthy regions (4, 5 and 24) include 0, indicating similarity of risks.

The effect of the socioeconomic background, represented in the model by HDI, revealed a strong detrimental effect in death intensity with posterior median of -10 and 95% credibility intervals given by $[-10.2, -9.7]$ in SVC model. Regions with better quality of life (higher HDI) have a much reduced CVD

deaths rate, when compared against less priviledged regions (low HDI). The relative risk between two RA's with a difference of 0.05 in the HDI is 1.65 and the 95% credibility intervals given by $[1.62, 1.67]$, showing an average 65% increase in death risk after a 0.05 decrease in HDI.

Figure 5 presents a summary of the space-varying intercept w. Recall that w is responsible for capturing the remaining spatial variation, that was not captured by the covariates. The spatial variation of w is still significant, despite the competition for handling spatial variation by the regression coefficients of this model. It seems that spatial heterogeneity is captured partly by the spatial variation of the regression coefficients and partly by this intercept.

Figures 6 and 7 summarize posterior inference for the hyperparameters. The mean for the age effect is concentrated around 0.082 and for gender effect around 0, as most variation is present in the estimated coefficients. These means are a grand mean of these coefficients and are similar to the same estimated coefficients under the FE model. The estimated precisions provide an indication of the magnitude of the spatial variation of the corresponding coefficient. They seem to indicate that the intercepts w for the FE model vary on average twice as much as the intercepts w for the SVC model. The larger magnitude of the precision of the age coefficient is merely a consequence of the larger nominal variation of this covariate when compared to the other individual variables, that are binary indicators.

The results above indicated that the effects of some covariates do not exhibit substantial spatial variation. Models with fixed effect for these covariates were also considered. They did not produce any noticeable change in the results for the other covariates, for the remaining hyperparameters nor for the fit criteria used.

4 Conclusions

This work presents a hierarchical formulation to handle point patterns that are subject to the effect of covariates with large spatial heterogeneity. This heterogeneity is modeled with isotropic Gaussian processes to ensure smooth variation of covariate coefficients over space.

The results showed the relevance of allowing also for spatial variation of the effect of covariates. They confirmed hypotheses about the important impact that socio-economic conditions may have over death rates but, more importantly here, over the effect of the covariates. Being old in a wealthy neighborhood is more protective against CVD deaths than being old in a poor neighborhood and much more protective than being old in a slum-town. Age was found to be the covariate with strongest spatial variation. The other covariates did not show strong changes over the city regions. Quantification of these effects was found to be more effectively reported to cardiologists through relative risks and examples were provided. Many other relevant comparisons can be made and used by health policy makers.

The data provided information about the model hyperparameters. Mean-

ingful, concentrated posterior distributions were obtained after the specification of vague prior distributions for them.

This work can be extended to additionally accommodate temporal heterogeneity in the intensity function [29]. It is well known that mortality patterns are changing over time. Larger time spans (the database used covered a period of only 6 years) may be required to identify significant demographic changes.

Finally, the analysis of datasets of these types could be improved from a model formulation without discretization. This is currently an active area of research; an example is provided in [1]. Note that once this is appropriately solved one will be faced with handling highly dimensional covariance matrices associated with Gaussian processes. This is another important topic of immediate practical relevance, with recent contributions from [3], [2], [8] and [24], to name a few. The specific problem analysed here did not require these approaches but they may well be useful for other problems with similar data structure.

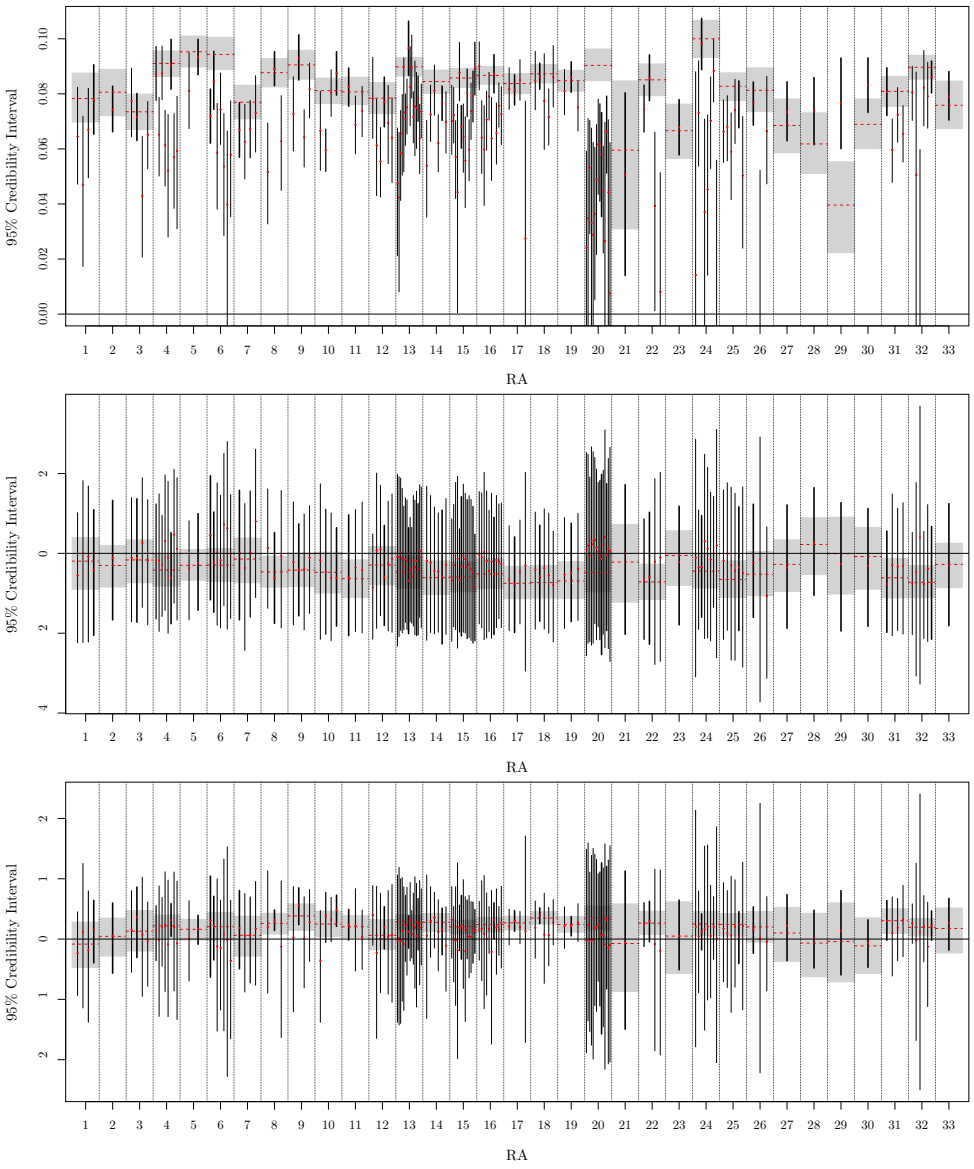

Figure 2: Posterior medians (red dots) and 95% credibility intervals (vertical bars) for the age coefficient for the 160 boroughs in the SVC model. Also represented in the figure are the median (red dashed line) and 95% credibility interval (shaded area) of the SVC model by RA. The widths of the vertical bars and dots increase with the total death counts of the boroughs to provide an account of the information provided by each borough. top - age; middle - schooling and botttom - gender

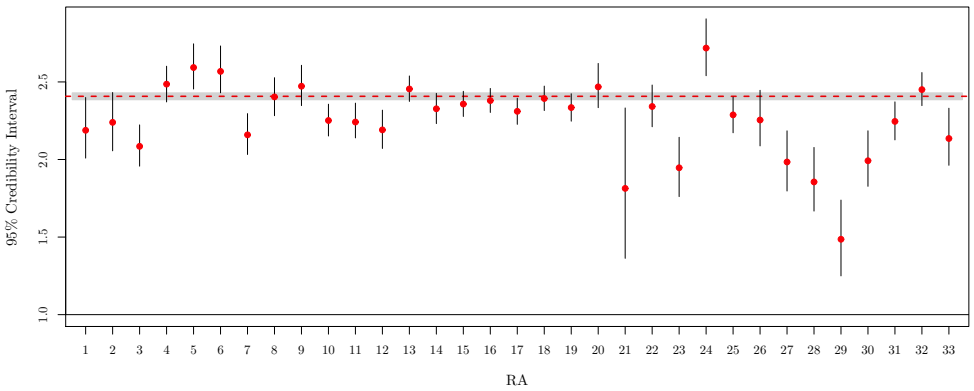

Figure 3: Posterior medians (red dots) and 95% credibility intervals (vertical bars) for the relative risk associated with a 10 year increase in age for the 33 RA's in the SVC model. Also represented in the figure are the median (red dashed line) and 95% credibility interval (shaded area) of the model with fixed effects.

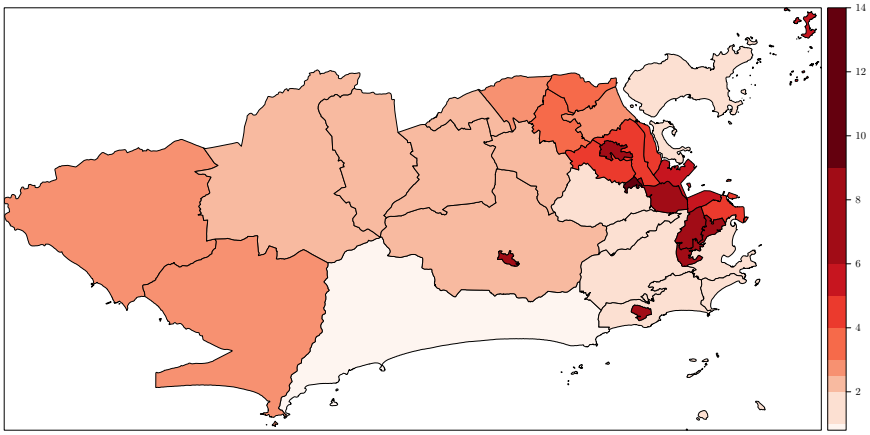

Figure 4: Posterior medians for the relative risks between regions for the more protective configuration: female, living with somebody, educated and young. Comparison are made against the region with the highest HDI value in the city, region 6, located in the southesterly corner of the map.

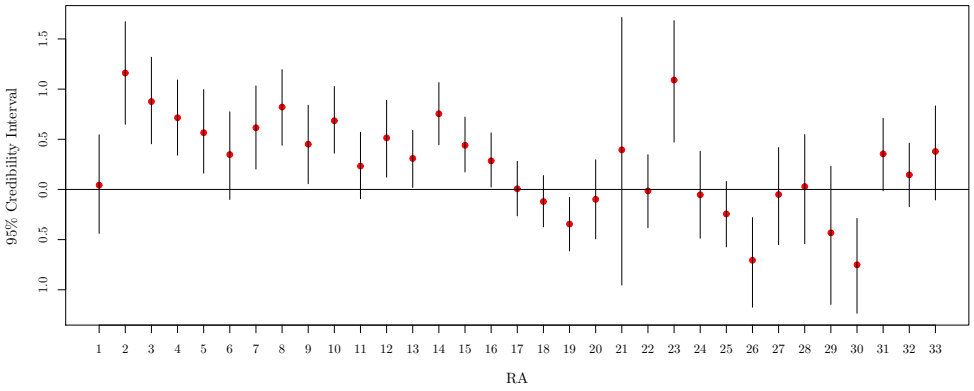

Figure 5: Posterior medians (red dots) and 95% credibility intervals (vertical bars) for the space-varying intercept w for the 33 RA's.

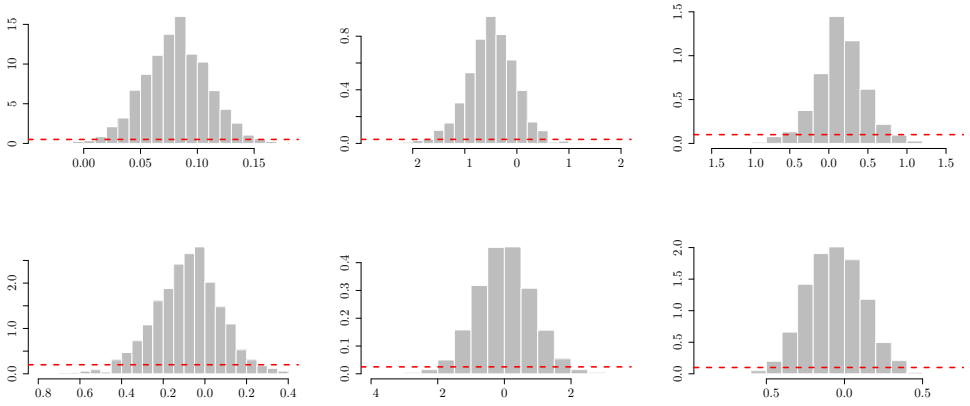

Figure 6: Posterior histograms for the means μ of the GP processes: top row (from left to right) - age, schooling and gender; bottom row (from left to right) - interaction of marital status and schooling, intercept w for the space-varying and intercept w for FE model. The dashed lines indicate the vague prior densities used in the analysis.

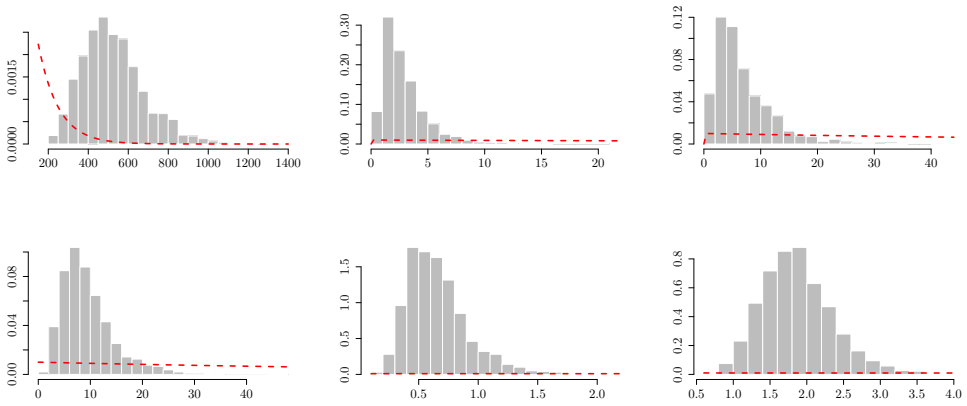

Figure 7: Posterior histograms for the precisions τ of the GP processes: top row (from left to right) - age, schooling and gender; bottom row (from left to right) - interaction of marital status and schooling, intercept w for the space-varying and intercept w for FE model. The dashed lines indicate the vague prior densities used in the analysis.

References

[1] [author] Adams, R. P.R. P., Murray, I.I. MacKay, D. J. C.D. J. C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In Proceedings of the 26th Annual International Conference on Machine Learning. ICML'09 9-16. ACM, New York, NY, USA.

[2] [author] Banerjee, A.A., Dunson, D. B.D. B. Tokdar, S.S. (2013). Efficient Gaussian process regression for large datasets. Biometrika 100 75-89.

[3] [author] Banerjee, A.A., Gelfand, A.A., Finley, A.A. Sang, H.H. (2008). Gaussian predictive process models for large spatial process datasets. Journal of the Royal Statistical Society Series B 70 825-848.

[4] [author] Beněs, V.V., Boldak, K.K., Møller, J.J. Waagepetersen, R.R. (2005). A case study on point process modelling in disease mapping. Image Analysis and Stereology 24 159-168.

[5] [author] Besag, J.J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society Series B 36 192-236.

[6] [author] Brix, A.A. Møller, J.J. (2001). Space-time Multi Type Log Gaussian Cox Processes with a View to Modelling Weeds. Scandinavian Journal of Statistics 28 471–488.

[7] [author] Cox, D.D. Isham, V.V. (1980). Point Processes. Chapman & Hall.

[8] [author] Cressie, N.N. Johannesson, G. .G. . (2008). Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society Series B 70 209-226.

[9] [author] Diggle, P.P. (2003). Statistical Analysis of Spatial Point Patterns, 2nd ed. Arnold, London.

[10] [author] Diggle, P. J.P. J., Guan, Y.Y., Hart, A. C.A. C., Paize, F.F. Stanton, M.M. (2010). Estimating Individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. Journal of the American Statistical Association 105 1394-1402.

[11] [author] Fonseca, T.T. Steel, M.M. (2011). A general class of nonseparable space-time covariance models. Environmetrics 22 224-242.

[12] [author] Fonseca, R. H. A.R. H. A., Pedroso, J. M. A.J. M. A., Pinto Junior, J. A.J. A., Souza e Silva, N. A.N. A., Gamerman, D.D. Duarte, M. M. T.M. M. T. (2013). Spatial analysis of mortality from cerebrovascular disease in Rio de Janeiro, 2002 to 2007. Demographic and socioeconomic correlations Technical Report, Cardiology Service - HUCFF/UFRJ.

[13] [author] Gamerman, D.D. (1992). A dynamic approach to the statistical analysis of point process. Biometrika 79 39-50.

[14] [author] Gamerman, D.D. Lopes, H.H. (2006). Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall/CRC, New York.

[15] [author] Gamerman, D.D., Salazar, E.E. Reis, E. A.E. A. (2007). Dynamic Gaussian Process Priors, with Applications to The Analysis of Space-time Data (with discussion) In: Bernardo, J. M.; Bayarri, M. J.; Berger, J. O.; Dawid, A. P.; Heckerman, D.; Smith, A. F. M.; West, M. (Org.). Bayesian Statistics 8. Oxford: Oxford Univeristy Press 149-174.

[16] [author] Gonzales, M.M., Rodrigues, A.A. Calero Jr, F.F. (1998). Relationship between socioeconomic status and ischaemic heart disease in cohort and case-control studies: 1960-1993. International Journal of Epidemiology 27 350-358.

[17] [author] Howard, GG., Cushman, M.M., Prineas, R.R., Howard, V.V., Moy, C.C., Sullivan, L.L., D'Agostino, R.R., McClure, L.L., Pulley, L.L. Safford, M.M. (2009). Advancing the hypothesis that geographic variations in risk factors contribute relatively little to observed geographic variations in heart diseases and stroke mortality. Preventive Medicine 49 129-132.

[18] [author] Ishitani, L. H.L. H., Franco, G. C.G. C., Perpétuo, I. H. O.I. H. O. França, E.E. (2006). Social inequality and precocious mortality by cardiovascular disease in Brazil. Revista de Saúde Pública 40 684-691 (In Portuguese).

[19] [author] Kaplan, GG. Keil, J.J. (1993). Socioeconomic factors and cardiovascular disease: a review of the literature. Circulation 88 1973-1998.

[20] [author] Kapral, M.M., Wang, H.H., Muhammad, M.M. Tu, J.J. (2002). Effect of socioeconomic status on treatment and mortality after stroke. Stroke 33 268-275.

[21] [author] Lavados, P.P., Hennis, A.A., Fernandes, J.J., Medina, M.M., Legetic, B.B., Hoppe, A.A., Sacks, C.C., Jadue, L.L. Salinas, R.R. (2007). Stroke epidemiology, prevention, and management strategies at a regional level: Latin America and the Caribbean. Lancet Neurology 6 362-372.

[22] [author] Lessa, I.I. (1990). Social aspects of early mortality (15 to 59 anos) by cerebrovascular diseases. Arquivos de Neuro-Psiquiatria 48(3) 296-300 (In Portuguese).

[23] [author] Liang, S.S., Carlin, B.B. Gelfand, A.A. (2008). Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. The Annals of Applied Statistics 3 943-962.

[24] [author] Lindgren, F.F., Rue, H.H. Lindstrom, J.J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastics partial differential equation approach (with discussion). Journal of the Royal Statistical Society Series B 73 423-498.

[25] [author] Lotufo, P.P. (2005). Stroke in Brazil: a neglected disease. São Paulo Medical Journal 123 3-4.

[26] [author] Lunn, D. J.D. J., Thomas, A.A., Best, N.N. Spiegelhalter, D.D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10 325-337.

[27] [author] Møller, J.J., Syversveen, A.A. Waagepetersen, R.R. (1998). Log Gaussian Cox process. Scandinavian Journal of Statistics 25 451-482.

[28] [author] Pedigo, A.A., Aldrich, T.T. Odoi, A.A. (2011). Neighborhood disparities in stroke and myocardial infarction mortality: a GIS and spatial scan statistics approach. BioMed Central Public Health 11 644.

[29] [author] Reis, E. A.E. A., Gamerman, D.D., Paez, M. S.M. S. Martins, T. G.T. G. (2013). Bayesian dynamic models for space-time point processes. Computational Statistics & Data Analysis 60 146-156.

[30] [author] Spiegelhalter, D. J.D. J., Best, N. G.N. G., Carlin, B. P.B. P. Linde, A.A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64 583-639.

[31] [author] UNPD (1990). Human Development Report 1990. New York: Oxford University Press.

[32] [author] Waagepetersen, R.R. (2004). Convergence of posterioris for discretized log Gaussian Cox process. Statistics and Probability Letters 66(3) 229-235.