

Douglas Roberto Mesquita Azevedo

**Spatial Confounding Beyond Generalized Linear  
Mixed Models: Extension to Shared  
Components and Spatial Frailty Models**

Belo Horizonte, Brasil

2020

Douglas Roberto Mesquita Azevedo

**Spatial Confounding Beyond Generalized Linear Mixed  
Models: Extension to Shared Components and Spatial  
Frailty Models**

Tese apresentada ao Programa de Pós-graduação em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutor em Estatística.

Universidade Federal de Minas Gerais – UFMG

Departamento de Estatística

Orientador: Marcos Oliveira Prates

Coorientador: Dipankar Bandyopadhyay

Belo Horizonte, Brasil

2020

© 2020, Douglas Roberto Mesquita Azevedo  
. Todos os direitos reservados

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende  
Costa CRB 6ª Região nº 1510

Azevedo, Douglas Roberto Mesquita.

A994s      Spatial confounding beyond generalized linear mixed models: extension to shared components and spatial frailty models / Douglas Roberto Mesquita Azevedo — Belo Horizonte, 2020.  
90 f. il.; 29 cm.

Tese(doutorado) - Universidade Federal de Minas Gerais – Departamento de Estatística.

Orientador: Marcos Oliveira Prates.  
Coorientador: Dipankar Bandyopadhyay

1. Estatística - Teses. 2. Análise de sobrevivência (Biometria) - Teses. 3. Mapas de doenças - Teses. 4. Multicolinearidade - Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.2 (043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

## ATA DA DEFESA DE TESE DO ALUNO DOUGLAS ROBERTO MESQUITA AZEVEDO

Realizou-se, no dia 28 de fevereiro de 2020, às 14:00 horas, 2040 ICEx, da Universidade Federal de Minas Gerais, a 61ª defesa de tese, intitulada *Spatial Confounding Beyond Generalized Linear Models: Extension to Shared Components and Spatial Frailty Model*, apresentada por DOUGLAS ROBERTO MESQUITA AZEVEDO, número de registro 2016675246, graduado no curso de ESTATÍSTICA, como requisito parcial para a obtenção do grau de Doutor em ESTATÍSTICA, à seguinte Comissão Examinadora: Prof(a). Marcos Oliveira Prates - Orientador (DEST/UFMG), Prof(a). Dipankar Bandyopadhyay - Coorientador (Virgina Commonwealth University), Prof(a). Wagner Hugo Bonat (UFPR), Prof(a). Leonardo Soares Bastos (FIOCRUZ), Prof(a). Renato Martins Assunção (DCC/UFMG), Prof(a). Vinícius Diniz Mayrink (DEST/UFMG).


A Comissão considerou a tese:

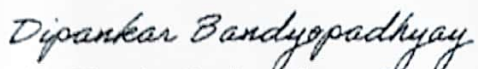
Aprovada

Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Belo Horizonte, 28 de fevereiro de 2020.

  
Prof(a). Marcos Oliveira Prates (Doutor)

  
Prof(a). Dipankar Bandyopadhyay - Coorientador (Doutor)

  
Prof(a). Wagner Hugo Bonat (Doutor)

  
Prof(a). Leonardo Soares Bastos (Doutor)

  
Prof(a). Renato Martins Assunção (Doutor)

  
Prof(a). Vinícius Diniz Mayrink (Doutor)



*A todos que direta ou indiretamente me apoiaram nessa jornada.*

# Agradecimentos

Em primeiro lugar agradeço aos meus pais e aos meus irmãos que apesar da distância sempre me incentivaram a seguir em frente. Vocês foram essenciais nessa caminhada.

Agradeço à minha namorada, Larissa Sayuri, pela paciência, carinho, companheirismo e prontidão de sempre. Agradeço em especial pelos momentos em que pudemos nos desligar e fazer alguma viagem divertida e renovadora.

Agradeço ao Marcos Prates por ter sido um excelente orientador, por dedicar seu tempo para me transmitir um pouco do seu conhecimento. Também agradeço pela oportunidade de trabalhar em problemas e em projetos tão distintos. Tenho certeza que essas experiências contribuíram de forma grandiosa para o meu crescimento profissional.

Agradeço ao Dipankar Bandyopadhyay, por ter me acolhido nos Estados Unidos durante o meu período de Doutorado sanduíche e por ter me apresentado diversos pratos da culinária indiana.

Agradeço ao Augusto Marcolin, Lucas Godoy, Luís Gustavo Silva e Rodrigo Reis pelas risadas, conversas e nerdices desses últimos seis anos. Essa etapa seria muito mais difícil sem vocês. Vocês são os grandes amigos que a UFMG me deu!

Agradeço ao Douglas Vargas, Gulliti Sena, Lucas Pereira, Maicon Fontes e Mateus Esswein, meus amigos de longa data, pela amizade que transcende o tempo e a distância.

Agradeço aos colegas de Mestrado/Doutorado Ali Abolhassani, Ana Gabriela, Bruno Barbarioli, Estevão Prado, Fernanda Gabriely, Guilherme Oliveira, Juliana Freitas e Magno Tairone pela troca de experiências e conversas entre um estudo e outro.

Agradeço também ao apoio financeiro da CAPES que não só viabilizou a execução deste trabalho como também me proporcionou a experiência de morar nos Estado Unidos por seis meses.

Agradeço também a todos que não foram citados mas que torcem por mim e que de alguma maneira contribuem ou contribuíram para o meu crescimento.

*“We know accurately only when we know little,  
with knowledge doubt increases”  
Johann Wolfgang von Goethe*

# Resumo

Confundimento espacial é o nome dado para o confundimento entre efeitos fixos e aleatórios espaciais em modelos lineares generalizados mistos (MLGMs). O confundimento espacial vem sendo amplamente estudado e vem ganhando atenção na literatura nos últimos anos visto que esta limitação pode gerar resultados inesperados na modelagem. As abordagens baseadas em projeção, conhecidas por modelos restritos, aparecem como uma boa alternativa para contornar as limitações do confundimento espacial em MLGMs. Entretanto, quando o suporte dos efeitos fixos difere do suporte do efeito espacial ou então quando diversos efeitos espaciais estão presentes na análise, os modelos baseados em projeção não são diretamente aplicáveis. Neste trabalho são introduzidas soluções para amenizar o confundimento espacial em duas famílias de modelos estatísticos. Em modelos de componente compartilhado, diversas variáveis resposta de contagem são observadas em cada região em estudo e muitas vezes apresentam padrões espaciais similares. Desta forma, os efeitos espaciais podem ser compartilhados entre as respostas além da possível presença de efeitos espaciais específicos. Neste contexto, nossa proposta se baseia no uso de estruturas espaciais modificadas para cada um dos componentes compartilhados e também dos efeitos espaciais específicos. Já modelos de fragilidade espacial permitem incorporar efeitos espacialmente estruturados através de um termo de fragilidade. Além disso, é comum observar-se mais de um indivíduo por região o que implica que o número de observações é maior que o número de regiões em estudo. Neste contexto propomos um modelo de projeção reduzindo a dimensionalidade dos dados. Como um produto deste trabalho, foi criado um pacote em R chamado `RASCO: An R package to Alleviate Spatial Confounding` que fornece à comunidade uma ferramenta para aliviar o confundimento espacial em MLGMs, modelos de componente compartilhado e modelos de fragilidade espacial. Para uma inferência Bayesiana à um custo computacional baixo, a metodologia INLA foi utilizada. Casos de câncer de pulmão e brônquios na Califórnia foram estudados em ambos os modelos mostrando a eficiência dos métodos propostos.

**Key-words:** análise de sobrevivência. mapa de doenças. efeitos latentes. multicolinearidade. SPOCK. projeção.

# Abstract

Spatial confounding is the name given to the confounding between fixed and spatial random effects in generalized linear mixed models (GLMMs). It has been widely studied and it gained attention in the past years in the spatial statistics literature, as it may generate unexpected results in modeling. The projection-based approach, also known as restricted models, appears like a good way to overcome the spatial confounding in this kind of models. However, when the support of fixed effects is different from the spatial effect one or when multiple spatial effects are present in the modeling, this approach can no longer be applied directly. In this work, we introduce solutions to alleviate the spatial confounding for two families of statistical models. In shared component models, multiple count responses are recorded at each spatial location, which may exhibit similar spatial patterns. Therefore, the spatial effect terms may be shared between the outcomes in addition to specific spatial patterns. In this case, our proposal relies on the use of modified spatial structures for each shared component and specific effects. Spatial frailty models can incorporate spatially structured effects and it is common to observe more than one sample unit per area which means that the support of fixed and spatial effects differ. In this case, we introduce a projection-based approach reducing the dimensionality of the data. As a product of this work an R package named **RASCO: An R package to Alleviate Spatial Confounding** is provided and it allows the community to alleviate the spatial confounding in GLMMs, shared component models and spatial frailty models. To provide a fast inference for the parameters, we used the INLA methodology. Lung and bronchus cancer in the California state is investigated under both methodologies and the results prove the efficiency of the proposed models.

**Key-words:** survival analysis. disease mapping. latent effects. multicollinearity. SPOCK. projection.

# List of figures

Figure 1 – Undirected graph and its adjacency matrix. . . . .	21
Figure 2 – Five regions and its respective <i>CAR</i> precision matrix. . . . .	21
Figure 3 – Eigenvectors 3, 8, 16, 20 and 26 from the RHZ and HH basis disposed in a $30 \times 30$ lattice for a simulated example. . . . .	25
Figure 4 – Centroids and neighborhood structure. Left: Original centroids and respective neighborhood; Center: SPOCK correction without spatial confounding; Right: SPOCK correction under spatial confounding . . .	28
Figure 5 – Covariates from CHRR in California (US) in 2016. . . . .	41
Figure 6 – Relative risk of lung and bronchus cancer in California (US). Source: SEER. . . . .	43
Figure 7 – Boxplot of $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ for $\boldsymbol{\theta} = \{\beta_{10}, \beta_{20}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$ in the shared component model. Dashed line represents the value 0. . . . .	51
Figure 8 – Boxplot of $\sigma_{\theta}$ for $\boldsymbol{\theta} = \{\beta_{10}, \beta_{20}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$ in the shared compo- nent model where $\sigma_{\theta}$ represents the standard deviation of $\theta$ . . . . .	52
Figure 9 – Estimated spatial patterns of a simulated dataset without spatial con- founding for the shared component model. . . . .	52
Figure 10 – Shared and specific spatial patterns estimates for the incidence of lung and bronchus cancer in California (US). . . . .	55
Figure 11 – Aggregated spatial pattern estimates for the incidence of lung and bronchus cancer in California (US). . . . .	56
Figure 12 – Time spent to fit the Weibull proportional hazard model with and without the reduction operator. Right: Original scale in seconds; Left: Logarithmic scale. . . . .	64
Figure 13 – Boxplot of $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ for $\boldsymbol{\theta} = \{\alpha, \beta_1, \beta_2\}$ in the spatial frailty model. Dashed line represents the value 0. . . . .	66
Figure 14 – Boxplot of $\sigma_{\theta}$ for $\boldsymbol{\theta} = \{\beta_1, \beta_2\}$ for the SFM and RSFM where $\sigma_{\theta}$ represents the standard deviation of $\theta$ . . . . .	67
Figure 15 – Boxplot of the SVIF (log scale) between spatial models (SFM and RSFM) and the baseline model (Weibull proportional hazard model). Dashed line marks the value 0, which in the log scale represent the equality of variances. . . . .	68
Figure 16 – Spatial risk effects for death by lung and bronchus cancer in California (US). . . . .	70

# List of tables

Table 1 – Distributions and its respective set of parameters ( $\theta$ ), probability density function ( $f_\theta$ ), survival function ( $S_\theta$ ) and hazard function ( $h_\theta$ ). The term $\Phi(t)$ represents the cumulative distribution function of a standard Normal distribution. . . . .	35
Table 2 – Summary statistics of CHRR covariates (SD: Standard deviation). . . .	42
Table 3 – Summary statistics of SEER covariates. For categorical variables the sample size and the percentage. For continuous variables median and quantiles 25% and 75%. . . . .	44
Table 4 – Simulation results for scenarios 1 and 4 for the shared component model experiment. The results are shown by mean, standard deviation (SD), coverage rate for a nominal rate of 95 % (Cov) and mean square error (MSE). . . . .	50
Table 5 – Analysis of the incidence of lung and bronchus cancer for men and women in California (US). Results are presented as mean, standard deviation (SD) and the 95% credibility interval (ICr). . . . .	54
Table 6 – Simulation results the spatial frailty model experiment. The results are shown by mean, standard deviation (SD), coverage rate for a nominal rate of 95 % (Cov) and mean square error (MSE). . . . .	65
Table 7 – Time until death by lung and bronchus cancer in California (US). Results are presented as mean, standard deviation (SD) and the 95 %credibility interval (ICr). . . . .	69
Table 8 – Simulation results for the shared component model experiment (Scenarios S2 and S3). The results are shown by mean, standard deviation (SD), coverage rate for a nominal rate of 95 % (Cov) and mean square error (MSE). . . . .	82

# Contents

	<b>Introduction</b> . . . . .	<b>14</b>
<b>1</b>	<b>METHODS REVIEW</b>	<b>17</b>
<b>1.1</b>	<b>GENERALIZED LINEAR MIXED MODELS</b> . . . . .	<b>18</b>
<b>1.2</b>	<b>SPATIAL MODELS</b> . . . . .	<b>20</b>
<b>1.2.1</b>	<b>Intrinsic Conditional autoregressive - ICAR</b> . . . . .	<b>20</b>
<b>1.2.2</b>	<b>Spatial confounding</b> . . . . .	<b>22</b>
1.2.2.1	Reich, Hodges and Zadnik (2006) . . . . .	23
1.2.2.2	Hughes and Haran (2013) . . . . .	24
1.2.2.3	Hanks et al. (2015) . . . . .	26
1.2.2.4	Prates, Assunção and Rodrigues (2019) . . . . .	27
1.2.2.5	Measures of spatial confounding . . . . .	28
<b>1.3</b>	<b>SHARED COMPONENT MODEL</b> . . . . .	<b>30</b>
<b>1.3.1</b>	<b>Disease mapping</b> . . . . .	<b>30</b>
<b>1.3.2</b>	<b>Shared component model</b> . . . . .	<b>31</b>
<b>1.4</b>	<b>SPATIAL FRAILTY MODEL</b> . . . . .	<b>33</b>
<b>1.4.1</b>	<b>Survival models</b> . . . . .	<b>33</b>
<b>1.4.2</b>	<b>Frailty models</b> . . . . .	<b>36</b>
<b>1.4.3</b>	<b>Spatial frailty models</b> . . . . .	<b>36</b>
<b>1.5</b>	<b>BAYESIAN INFERENCE</b> . . . . .	<b>37</b>
<b>1.5.1</b>	<b>Introduction</b> . . . . .	<b>37</b>
<b>1.5.2</b>	<b>Integrated nested Laplace approximation - INLA</b> . . . . .	<b>37</b>
1.5.2.1	Marginal distribution for $\theta_k$ . . . . .	38
1.5.2.2	Marginal distribution for $u_j$ . . . . .	39
<b>2</b>	<b>DATA SOURCES</b>	<b>40</b>
<b>2.1</b>	<b>DATA SOURCES</b> . . . . .	<b>41</b>
<b>2.1.1</b>	<b>CHRR dataset</b> . . . . .	<b>41</b>
<b>2.1.2</b>	<b>SEER dataset</b> . . . . .	<b>42</b>
2.1.2.1	Incidence of bronchus and lung cancer . . . . .	42



2.1.2.2	Time until death by bronchus and lung cancer . . . . .	43
<b>3</b>	<b>SPATIAL CONFOUNDING IN SHARED COMPONENT MODELS</b>	<b>45</b>
3.1	METHOD . . . . .	46
3.2	SIMULATION . . . . .	49
3.3	MALE VS FEMALE LUNG AND BRONCHUS CANCER INCIDENCE IN CALIFORNIA . . . . .	53
<b>4</b>	<b>SPATIAL CONFOUNDING IN SPATIAL FRAILTY MODELS</b>	<b>57</b>
4.1	METHOD . . . . .	58
4.1.1	Reduction operator . . . . .	60
4.1.1.1	HH model . . . . .	62
4.1.1.2	SPOCK model . . . . .	62
4.2	SIMULATION . . . . .	63
4.2.1	Computational improvement . . . . .	63
4.2.2	Confounding alleviation . . . . .	64
4.3	TIME UNTIL DEATH BY LUNG AND BRONCHUS CANCER IN CALIFORNIA . . . . .	69
<b>5</b>	<b>RASCO: AN R PACKAGE TO ALLEVIATE SPATIAL CONFOUNDING</b>	<b>71</b>
5.1	RASCO: AN R PACKAGE TO ALLEVIATE SPATIAL CONFOUNDING	72
5.1.1	Installation . . . . .	72
5.1.2	Generalized Linear Mixed models . . . . .	73
5.1.3	Shared Component models . . . . .	75
5.1.4	Survival models . . . . .	77
	Final remarks . . . . .	79

<b>APPENDICES</b>	<b>81</b>
<b>Appendix A – RSCM - SIMULATION</b> . . . . .	<b>82</b>
<b>Appendix B – REDUCTION OPERATOR PROOFS</b> . . . . .	<b>83</b>
<b>REFERENCES</b> . . . . .	<b>86</b>

# Introduction

Spatial models are widely studied in the literature and are important in practice to model spatially correlated data. In addition to the development of robust models, many works focus on identifying and solving the limitations of those models. One of the spatial model's limitation is called spatial confounding (CLAYTON; BERNARDINELLI; MONTOMOLI, 1993; REICH; HODGES; ZADNIK, 2006; HUGHES; HARAN, 2013; HANKS et al., 2015; HEFLEY et al., 2017; THADEN; KNEIB, 2018; PRATES; ASSUNÇÃO; RODRIGUES, 2019). This problem resembles the multicollinearity in linear models which can distort the results and even lead to wrong conclusions. Spatial confounding occurs when the spatial effect contains similar information to the one coming from the fixed effects. Thus, point estimates of the regression coefficients become biased and their variance gets inflated (REICH; HODGES; ZADNIK, 2006).

For traditional spatial models, the most common approach to alleviate spatial confounding is the restricted spatial regression model (REICH; HODGES; ZADNIK, 2006). It is based on the projection of spatial effects onto the orthogonal space of the covariates. This approach is well accepted and some alternatives to the original idea have been reported (HUGHES; HARAN, 2013; HANKS et al., 2015; GUAN; HARAN, 2018).

Alternatively to these projection-based approaches, Prates, Assunção and Rodrigues (2019) provided a tool, named spatial orthogonal centroid “k”orrection (SPOCK). Its main idea is to alleviate the spatial confounding misplacing the regions centroids, creating a new neighborhood structure that leads to a model where the spatial confounding is less intense. The main advantage of this approach is that the correction is made before fitting the model, which allows the user to choose its preferred software.

Another set of alternatives to overcome the spatial confounding is available in the literature using various statistical tools such as the lasso regression (HEFLEY et al., 2017), structural equation models (THADEN; KNEIB, 2018) and, causal inference (PAPADO-GEORGOU; CHOIRAT; ZIGLER, 2018; DAVIS et al., 2019; OSAMA; ZACHARIAH; SCHÖN, 2019).

Advances in technology, data storage and the good quality of data allow fitting increasingly complex models that more realistically represent the phenomenon of interest, providing better fit and interpretation. These models are in use by non-statisticians and might suffer from spatial confounding. Thus, it is important to investigate the existence of this limitation in more general settings.

In epidemiology, it is often observed that disease incidences and counts exhibit spatial clustering, i.e., counts observed in geographically proximal areas (such as counties

in a state) have similar patterns and may differ from those observed in distant areas. Disease mapping is a widely used tool to model and evaluate risk factors of diseases from spatially structured counts of deaths, or new cases of a disease. Inference derived from disease mapping studies may offer the researcher a new look into the reasons for such spatial clustering, which may pave way for new policy decisions to contain the disease spread.

For areal (count) responses recorded on a single disease at spatial locations, the common approach is to use a (univariate) Poisson model, and covariates regressed via appropriate link function on the Poisson intensity. However, in practice, multiple count responses pertaining of multiple diseases, may be recorded at the same spatial locations, thereby experiencing similar spatial patterns. Modeling multiple disease counts can be readily accomplished via a shared component model (KNORR-HELD; BEST, 2001), where spatial effects may be shared between multiple diseases for modeling the association, with additional spatial terms accounting for disease-specific effects.

Frailty models are a useful and flexible class of survival models. It allows the inclusion of latent effects related to a subject or a group to accommodate possible non-observed covariates. The simplest way to employ a frailty term is through an unstructured effect, ensuring that the hazard is positive for each sample unit. However, this approach does not handle structured effects as the case of spatial effects. Papers as Henderson, Shimakura and Gorst (2002), Li and Ryan (2002), Banerjee, Wall and Carlin (2003), Bastos and Gamerman (2006) propose the use of frailty models to incorporate spatial structure into the latent effects. Banerjee, Wall and Carlin (2003) show how to include consolidated spatial models such as the conditional autoregressive (CAR) model (BESAG, 1974) for areal data and the Gaussian model for georeferenced data (CRESSIE, 1992).

The usual projection-based approaches for areal data cannot be directly applied to these families of models. For the shared component model, multiple spatial effects are present. Also, each disease may have specific covariates which makes the projection a non-trivial task. On the other hand, for frailty models, there is a difference in the support of the spatial structure (areal level) and the fixed effects (sample unit level) imposing a limitation on applying a projection-based approach.

Several solutions are available to alleviate the spatial confounding for generalized linear mixed models (GLMM). However, there is no software available which unifies the solutions for GLMM and, to the best of our knowledge, there are no investigation studies of spatial confounding in models beyond the GLMM family.

In this work, we provide a tool to alleviate the impact of spatial confounding in spatial frailty models as well as in shared component models. For the shared component model, we propose multiple applications of SPOCK for specific and shared spatial effects considering the specific and all covariates, respectively. For spatial frailty models, we

propose an alternative way to apply the projection-based approach reducing the dimension of the covariate matrix using a reduction operator. It also reduces the computational cost, being an interesting tool for spatial frailty models in big data.

As a product of this work, we have an R package called “**RASCO: An R package to Alleviate Spatial Confounding**”. It includes functions to alleviate the spatial confounding in shared component models, spatial frailty models and also generalized linear mixed models. Some projection-based approaches are provided in the package as the cases of Prates, Assunção and Rodrigues (2019) and Reich, Hodges and Zadnik (2006) proposals. The package can be found at <https://github.com/douglasmesquita/RASCO>.

We applied the methodology to respiratory cases (lung and bronchus) in the California state. The dataset was provided by Surveillance, Epidemiology, and End Results Program (SEER, 2019). For the shared component analysis, counts of new cases of lung and bronchus cancers are modeled for men and women. For the spatial frailty model, the time until death by lung and bronchus cancer is investigated.

This Ph.D. thesis is organized as follows. In Part 1 a review is presented with a discussion about spatial confounding, shared component models and spatial survival models. Part 2 shows the data sources for the applications and some summary statistics. Part 3 describes our proposal to alleviate the spatial confounding in shared component models. Part 4 presents our proposal to alleviate the spatial confounding in spatial frailty models. Part 5 introduces the RASCO package and its basic usage. Finally, we present the main conclusions and final remarks.

# Part 1

## Methods review

## 1.1 Generalized linear mixed models

Linear regression is a simple and widely applied method in Statistics (NETER et al., 1996). It connects the unknown mean  $\boldsymbol{\mu}$  of a random vector  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  with a set of  $q$  covariates aiming to explain the behavior of  $\mathbf{Y}$  through the behavior of  $\mathbf{X}$ , in which  $\mathbf{X}$  is a  $n \times q$  matrix of covariates. To simplify calculations, it is usual to assume that  $Y_i | \mu_i, \sigma^2 \sim \text{Normal}(\mu_i, \sigma^2)$ . In this case, we have that

$$\begin{aligned} \mathbf{Y} | \boldsymbol{\mu}, \sigma^2 &\sim \text{Normal}_n(\mathbf{Y}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}_n) \\ \boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

where  $\boldsymbol{\beta}$  is a column vector of  $q$  coefficients and  $\mathbf{I}_n$  is a  $n \times n$  identity matrix. This tool is useful because it is simple to interpret each  $\beta_j$  in terms of variations in  $\boldsymbol{\mu}$ . However, this technique has some assumptions like the independence and the distribution of  $\mathbf{Y}$ . Frequently, the outcomes are not in the Gaussian family and in other cases the assumption of independence is not achieved. This way, more flexible methods are needed.

The generalized linear model (GLM) (NELDER; WEDDERBURN, 1972) is a natural extension of linear regression to distributions belonging to the exponential family. Nelder and Wedderburn (1972) cite three main characteristics of this model family in their work: 1) A dependent variable  $\mathbf{Y}$  whose distribution belongs to the exponential family; 2) A matrix of independent covariates represented by  $\mathbf{X}$  that linearly predict a function of  $\boldsymbol{\mu}$  by  $\mathbf{X}\boldsymbol{\beta}$ ; 3) A link function  $g(\cdot)$  that connects  $\boldsymbol{\mu}$  with the linear predictor  $\mathbf{X}\boldsymbol{\beta}$ . Then the definition becomes

$$\begin{aligned} \mathbf{Y} | \boldsymbol{\mu}, \boldsymbol{\theta} &\sim f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\theta}) \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

where  $f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\theta})$  is the density or probability function of the  $\mathbf{Y}$  in the exponential family and  $\boldsymbol{\theta}$  is a vector of any other parameters of  $f(\cdot)$ .

Under this framework, it is possible to fit, for example, binomial, Poisson and gamma regressions (NELDER; WEDDERBURN, 1972). Currently, due to computational improvement, the exponential family requirement is no longer necessary.

This set of distributions allows us to employ this methodology in a huge number of real situations. However, the assumption of independence is strong and not realistic in some cases. It is common to observe clustered data as the case of data collected in time or space. It is realistic to assume that data collected in closed periods of time or regions may have some similarities.

The generalized linear mixed model (GLMM) (BRESLOW; CLAYTON, 1993) extends the GLM family in which beside fixed effects one can also include latent effects of

unobserved covariates. The general formulation is given by:

$$\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\theta} \sim f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\theta})$$

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\epsilon},$$

where the vector  $\boldsymbol{\epsilon}$  contains possible random effects and  $\mathbf{Z}$  is a design matrix linked to the latent effects.

Some useful models are encompassed on the GLMM formulation. For example, for the previous linear regression,  $\mathbf{Y}$  is a Gaussian outcome,  $g(\boldsymbol{\mu}) = \boldsymbol{\mu}$  is the identity link function,  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are the covariate matrix and coefficients, respectively, and there are no latent effects.

Disease mapping is widely used in the epidemiology literature. In the simpler case,  $Y_i$  is a Poisson outcome,  $g(\mu_i)$  is generally the logarithm function,  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are the covariate matrix and coefficients respectively,  $\mathbf{Z}$  is a  $n \times n$  diagonal matrix with entries 1 and  $\boldsymbol{\epsilon}$  is a vector of spatial effects.

Another example is the longitudinal analysis where the same subject is observed multiple times in a period of time. It is realistic to imagine that a specific subject may evolve on time according to its historical observations. In this case, the observations of this subject are not independent. Thus, under the GLMM it is possible to consider a temporal structure to its observations (DIGGLE et al., 2002).

Similarly, in data collected on the space, observations may be influenced by its location. For example, in an epidemic, it is realistic to think that it is going to spread following some spatial patterns. Therefore, it is important to include spatial dependence into the model. Under the GLMM it is also possible to consider a spatial structure of the observations through a latent effect (BANERJEE; CARLIN; GELFAND, 2014).

Because of the importance of this class of models, several methodological improvements are made to develop this family making it even more general and robust to possible limitations. One can cite the choice of the link function and the implications of misspecification (CZADO; SANTNER, 1992) or recently, the spatial confounding (REICH; HODGES; ZADNIK, 2006; HODGES; REICH, 2010; HUGHES; HARAN, 2013; HEFLEY et al., 2017; GUAN; HARAN, 2018; THADEN; KNEIB, 2018; PAPADOGEORGOU; CHOIRAT; ZIGLER, 2018; PRATES; ASSUNÇÃO; RODRIGUES, 2019; DAVIS et al., 2019; OSAMA; ZACHARIAH; SCHÖN, 2019) that is the limitation under investigation in this work.



## 1.2 Spatial models

Modeling the sources of variation is important for countless fields and can help researchers identifying spatial patterns and make decisions. In many cases, the data is spatially structured which makes the models capable of identifying the influence of these structures indispensable.

In Statistics the most common types of spatial data are those in which observations are collected in a continuous space (geostatistical data) or when they represent a region in space (areal data). In the areal data context, it is desired to add the neighborhood structure information into the modeling. This is done to accommodate the spatial behavior of a possible unobserved or latent covariate.

There are several approaches for modeling spatially structured data in the areal context as the cases of conditional autoregressive (CAR) (BESAG, 1974; BANERJEE; CARLIN; GELFAND, 2014), simultaneous autoregressive (SAR) (WHITTLE, 1954; ORD, 1975), Leroux (LEROUX; LEI; BRESLOW, 1999), mixture neighborhood structure (RODRIGUES; ASSUNÇÃO, 2012) and, directed acyclic graph autoregressive (DAGAR) (DATTA et al., 2019).

In applied sciences, the most common approach for areal data is the CAR model due to the model simplicity. Although any of the mentioned methodologies are valid, in this work we will focus on the ICAR model.

### 1.2.1 Intrinsic Conditional autoregressive - ICAR

Let  $\mathbf{Y} = Y_1, \dots, Y_n$  be a random vector observed into  $n$  regions and let  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)^T$  denote random effects with zero mean related to each one of the  $n$  regions. The ICAR model is specified by the following conditional distributions

$$(\psi_i | \boldsymbol{\psi}_{-i}) \sim \text{Normal} \left( \sum_{j \sim i} b_{ij} \psi_j, \sigma_i^2 \right),$$

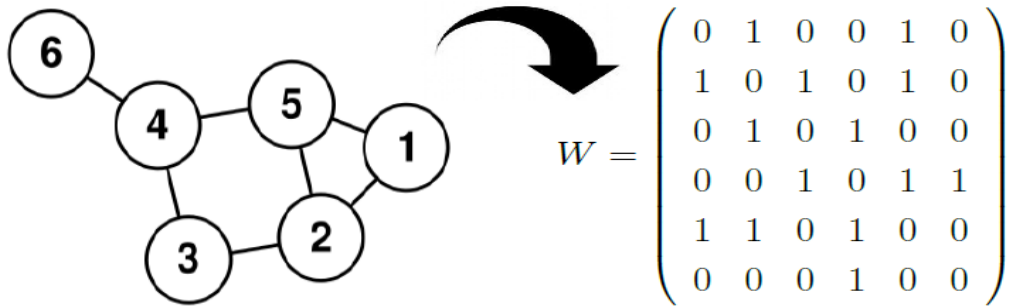
where  $\boldsymbol{\psi}_{-i}$  represents the vector  $\boldsymbol{\psi}$  without the  $i$ -th element,  $b_{ij}$  is a weight relating the regions  $i$  and  $j$  and,  $j \sim i$  indicates that regions  $j$  and  $i$  are neighbors. One can show that the joint distribution of  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)^T$  is given by:

$$\pi(\boldsymbol{\psi}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\psi}^T \mathbf{D}^{-1} (\mathbf{I}_n - \mathbf{B}) \boldsymbol{\psi} \right\}, \quad (1.2.1)$$

where  $\mathbf{D}$  is a diagonal matrix with entries  $\sigma_i^2$ ,  $i = 1, \dots, n$  and  $\mathbf{B}$  is a  $n \times n$  matrix of weights with entries  $b_{ij}$ .

Equation (1.2.1) resembles a multivariate Gaussian distribution with mean vector  $\mathbf{0}$  and covariance matrix given by  $\Sigma = (\mathbf{I}_n - \mathbf{B})^{-1} \mathbf{D}$ . However, one needs to show that  $\Sigma$  is symmetric and positive definite.

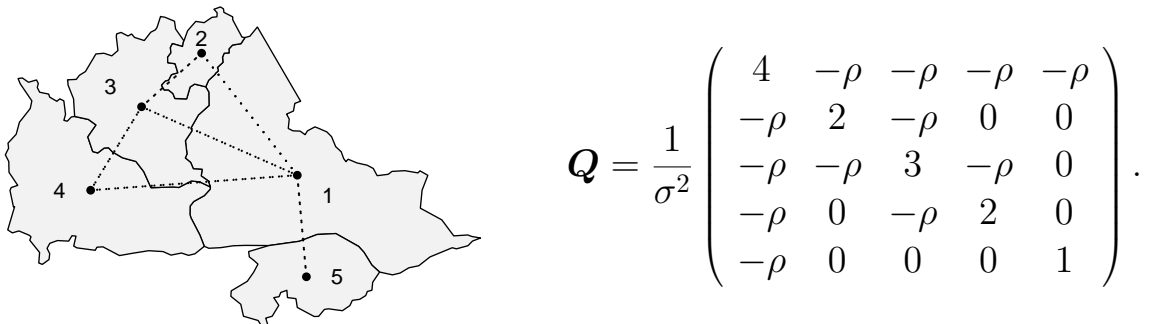
The traditional ICAR model defines an adjacency matrix  $\mathbf{W}$  composed of zeros and ones where  $w_{ij} \neq 0$  if and only if  $j \sim i$ . Thus,  $\mathbf{W}$  comprises zeros and ones, where 1 indicates that areas  $i$  and  $j$  are neighbors, and 0 otherwise. Figure 1 shows for a given graph its respective adjacency matrix.



**Figure 1** – Undirected graph and its adjacency matrix.

Then, a common approach is to take  $b_{ij} = \frac{w_{ij}}{w_{i+}}$ , where  $w_{i+}$  is the sum of the elements of the  $i$ -th row of matrix  $\mathbf{W}$ , in other words,  $w_{i+}$  is the number of neighbors of region  $i$ . Another definition is to take the marginal variance in each region as  $\sigma_i^2 = \frac{\sigma^2}{w_{i+}}$ .

In this situation,  $\mathbf{Q} = \Sigma^{-1} = \frac{1}{\sigma^2}(\mathbf{D}_w - \mathbf{W})$ , where  $\mathbf{D}_w$  is a diagonal matrix with values  $w_{i+}$ . This set of assumptions only guarantees that  $\mathbf{Q}$  is symmetric, but, it still does not guarantee positive definiteness and therefore it does not guarantee a proper joint distribution. Despite that, this formulation is still useful and is known as intrinsic conditional autoregressive model (ICAR) (BESAG; YORK; MOLLIE, 1991).



**Figure 2** – Five regions and its respective CAR precision matrix.

A proper version of the latter model is obtained by inserting a dependence parameter  $\rho$  in the previous dependence matrix (BESAG, 1974). Thus, the new dependence structure  $\mathbf{Q} = \frac{1}{\sigma^2}(\mathbf{D}_w - \rho\mathbf{W})$  is proper and therefore the joint distribution is proper where  $\rho \in (\lambda_{min}^{-1},$

$\lambda_{max}^{-1}$ ) and  $\lambda_{min}$  and  $\lambda_{max}$  are the smallest and largest eigenvalues of  $\mathbf{D}_w^{\frac{1}{2}} \mathbf{W} \mathbf{D}_w^{\frac{1}{2}}$ , respectively (BANERJEE; CARLIN; GELFAND, 2014). Figure 2 shows for a given map and a specific neighborhood structure the respective proper CAR precision matrix.

Both ICAR and CAR models can be accomplished in the GLMM through the latent effect. In this case, the latent effect is going to accommodate any spatially structured information not observed. In our work we are focusing on the ICAR model.

## 1.2.2 Spatial confounding

A current limitation in spatial statistics is the so-called spatial confounding. This problem resembles what occurs in linear models when two or more covariates bring the same information about the response variable. In linear models, we call it multicollinearity. One of the main problems related with multicollinearity is the variance inflation of the regression coefficient estimators. This inflation, in some cases, changes the model interpretation leading the researcher, occasionally, to wrong conclusions about a covariate in the model.

A simple way to investigate the multicollinearity is using the variance inflation factor (VIF). The variance of a given coefficient can be written as

$$\widehat{\text{var}}(\hat{\beta}_j) = s^2 (\mathbf{X}^T \mathbf{X})_{j,j}^{-1} = \frac{s^2}{(n-1) \widehat{\text{var}}(\mathbf{X}_j)} \cdot \frac{1}{1 - R_j^2}, \quad (1.2.2)$$

where  $s^2$  is the variance of the residuals,  $R_j^2$  is the coefficient of determination for the regression of  $\mathbf{X}_j$  over other covariates but not  $\mathbf{Y}$ . Thus,

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}.$$

If  $R_j^2$  is near 1, it suggests that a linear combination of the covariates is bringing similar information to that coming from  $X_j$ . In this case,  $VIF(\beta_j) \rightarrow \infty$ . On the other hand, if  $R_j^2$  is near 0, then,  $VIF(\beta_j) \rightarrow 1$  suggesting that there is no multicollinearity in this model.

The extension of the multicollinearity problem to the spatial context is called spatial confounding and it occurs when the spatial effect brings similar information to one or a linear combination of the covariates in the model. Differently from the multicollinearity, spatial confounding determines an inflation in the variance of the regression estimators and also a bias in the point estimate, possibly changing conclusions drastically.

Recently several works as Reich, Hodges and Zadnik (2006), Hughes and Haran (2013), Hanks et al. (2015), Hefley et al. (2017), Guan and Haran (2018), Thaden and Kneib (2018), Prates, Assunção and Rodrigues (2019) approached the spatial confounding problem either for generalized linear mixed models areal or geostatistical data. We are going to describe some of these approaches in the next subsections.

### 1.2.2.1 Reich, Hodges and Zadnik (2006)

Reich, Hodges and Zadnik (2006) mathematically formulate the problem of spatial confounding for spatial linear regression models. In their work, the marginal distributions of regression coefficients have closed form. The authors employed the structure of the ICAR model as a spatial component in the following model:

$$\begin{aligned} \mathbf{Y} | \boldsymbol{\beta}, \boldsymbol{\psi}, \tau_\epsilon &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\psi}, \tau_\epsilon \mathbf{I}_n), \\ \boldsymbol{\psi} | \tau_\psi &\sim \text{Normal}(0, \tau_\psi \mathbf{Q}), \end{aligned} \quad (1.2.3)$$

where  $\boldsymbol{\psi}$  is the ICAR spatial effect,  $\mathbf{Q}$  is the precision matrix presented in Section 1.2.1,  $\tau_\epsilon$  and  $\tau_\psi$  are precision parameters related to the Gaussian observations and the ICAR, respectively. Therefore, the Gaussian distribution is being represented by its precision matrix rather than its covariance matrix.

In the case of spatial linear regression, it is possible to analytically calculate the expected mean and variance integrating out the latent effect as:

$$\begin{aligned} E(\boldsymbol{\beta} | \tau_\epsilon, \tau_\psi, \mathbf{y}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \hat{\boldsymbol{\psi}}) = \boldsymbol{\beta}_{ols} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\psi}}, \\ \text{Var}^{-1}(\boldsymbol{\beta} | \tau_\epsilon, \tau_\psi, \mathbf{y}) &= \tau_\epsilon (\mathbf{X}^T \mathbf{X}) - \mathbf{X}^T \text{Var}(\boldsymbol{\psi} | \boldsymbol{\beta}, \tau_\epsilon, \tau_\psi, \mathbf{y}) \mathbf{X}, \end{aligned} \quad (1.2.4)$$

where  $\boldsymbol{\beta}_{ols} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  and  $\hat{\boldsymbol{\psi}} = E(\boldsymbol{\psi} | \tau_\epsilon, \tau_\psi, \mathbf{y})$ .

One can notice that the expectation of  $\boldsymbol{\beta}$ , integrating out the spatial effect, is the same as those obtained by the ordinal least squares minus a component that relates the covariates and the latent effect  $\boldsymbol{\psi}$ . Similarly, the precision is the same as in linear models, see Equation (1.2.2), minus a quantity involving the spatial effect. In other words, under spatial models, the expected mean and variance are different from those from linear models by a quantity involving the latent effect.

Carefully investigating Equation (1.2.4) it is possible to conclude that the predicted  $\mathbf{y}$  value is given by  $\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \hat{\boldsymbol{\psi}}) = \mathbf{P}(\mathbf{y} - \hat{\boldsymbol{\psi}}) = \mathbf{P}\mathbf{y} - \mathbf{P}\hat{\boldsymbol{\psi}}$ , a projection of  $\mathbf{y}$  onto the space of  $\mathbf{X}$  minus a projection of the latent effect  $\hat{\boldsymbol{\psi}}$  onto the space of  $\mathbf{X}$ . To make it clear, the projection matrix given by  $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  is useful to project any vector, of appropriate dimension, onto the space of  $\mathbf{X}$ . For example, in linear regression we know that  $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ , thus  $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{P}\mathbf{y}$ . In other words,  $\hat{\mathbf{y}}$  is the projection of  $\mathbf{y}$  onto the space of  $\mathbf{X}$ .

When the spatial confounding is present it indicates the existence of duplicated information in the model. One way to alleviate this problem is by using a projection-based model. That is, by decomposing the spatial effect into a projection onto the space of the covariates and a projection onto the orthogonal space of the covariates as in

Equation (1.2.5).

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\psi}, \tau_\epsilon &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\psi} + \mathbf{P}^\perp\boldsymbol{\psi}, \tau_\epsilon\mathbf{I}_n), \\ \boldsymbol{\psi}|\tau_\psi &\sim \text{Normal}(0, \tau_\psi\mathbf{Q}), \end{aligned} \quad (1.2.5)$$

in which,  $\mathbf{P}^\perp = (\mathbf{I} - \mathbf{P})$  is the projection matrix onto orthogonal space of  $\mathbf{X}$ .

Therefore, in the case of Equation (1.2.5),  $\mathbf{P}\boldsymbol{\psi}$  corresponds to the information of  $\boldsymbol{\psi}$  on the space of  $\mathbf{X}$  which in other words represents the duplicated information. Thus, one way to alleviate the spatial confounding is by removing  $\mathbf{P}\boldsymbol{\psi}$  from the model.

This approach is known as restricted spatial regression (RSR) and is also applicable for GLMM. However, for GLMMs, it is not possible to analytically evaluate the impacts of the latent effect on the coefficients estimates as the analytical solution of the involved integrals are not available.

To take some computational advantage of the restricted model, Reich, Hodges and Zadnik (2006) notice that  $\mathbf{P}$  is a rank  $q$  matrix while  $\mathbf{P}^\perp$  is a rank  $n - q$  matrix. It implies that  $\mathbf{P}$  and  $\mathbf{P}^\perp$  have  $q$  and  $n - q$  nonzero eigenvalues respectively.

Because  $\mathbf{P}$  is a square matrix, one can take  $\mathbf{K}$  as the  $p$  eigenvectors rows related to the nonzero eigenvalues of  $\mathbf{P}$  and  $\mathbf{L}$  as the  $n - p$  eigenvectors rows related to the nonzero eigenvalues of  $\mathbf{P}^\perp$ . Thus we can define  $\boldsymbol{\theta}_1 = \mathbf{K}\boldsymbol{\psi}$  and  $\boldsymbol{\theta}_2 = \mathbf{L}\boldsymbol{\psi}$  and rewrite the model in the Equation (1.2.5) as

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \tau_\epsilon &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^T\boldsymbol{\theta}_1 + \mathbf{L}^T\boldsymbol{\theta}_2, \tau_\epsilon\mathbf{I}_n), \\ \boldsymbol{\theta}|\tau_s &\sim \text{Normal}(0, \tau_s\tilde{\mathbf{Q}}), \end{aligned} \quad (1.2.6)$$

where  $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T]^T$  and  $\tilde{\mathbf{Q}} = (\mathbf{K}^T\mathbf{L}^T)^T\mathbf{Q}(\mathbf{K}^T\mathbf{L}^T)$ .

The quantity  $\mathbf{K}^T\boldsymbol{\theta}_1$  is the combinations of the latent effect into the span of  $\mathbf{X}$ . So a solution to alleviate spatial confounding is to remove this quantity from the model

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \tau_\epsilon &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta} + \mathbf{L}^T\boldsymbol{\theta}_2, \tau_\epsilon\mathbf{I}_n), \\ \boldsymbol{\theta}_2|\tau_s &\sim \text{Normal}(0, \tau_s\mathbf{L}\mathbf{Q}\mathbf{L}^T). \end{aligned} \quad (1.2.7)$$

This approach is a better computational solution since it evolves  $\mathbf{L}^T$  a  $(n - p) \times n$  matrix instead of  $\mathbf{P}^\perp$  a  $n \times n$  matrix. However, because it is common to have the number of covariates  $q$  much smaller than areas  $n$ , this simplification does not represent a big computational advantage. In the rest of this work, we will refer to this approach as the RHZ model.

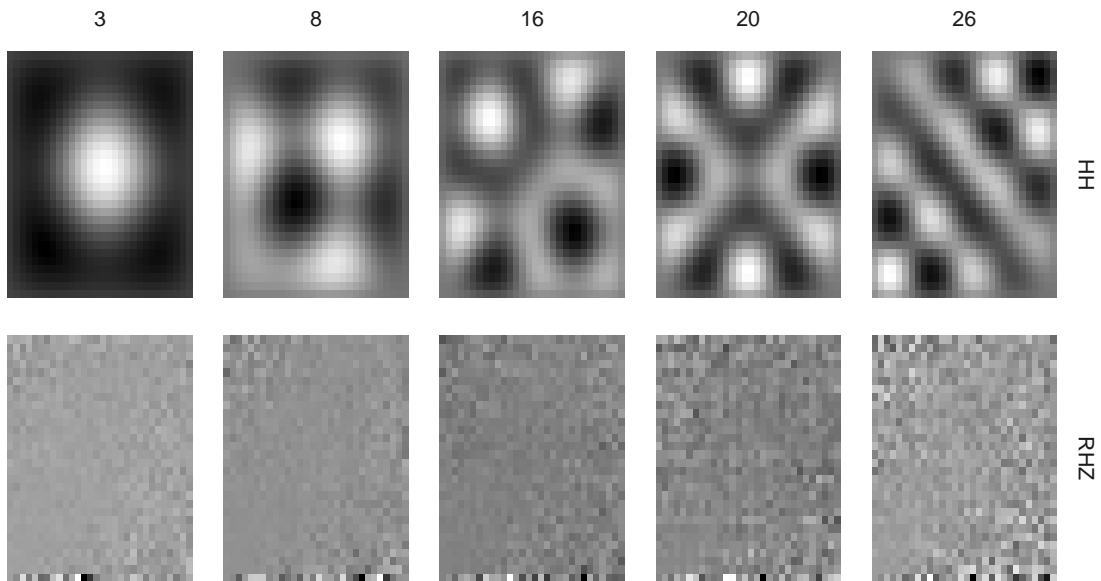
### 1.2.2.2 Hughes and Haran (2013)

Although Reich, Hodges and Zadnik (2006) proposal alleviates the spatial confounding, this method is computationally inefficient. Hughes and Haran (2013) noticed that

$LQL^T$  is a  $(n - q) \times (n - q)$  matrix and  $(n - q) \approx n$ . Thus, [Hughes and Haran \(2013\)](#) proposed a sparse reparametrization of the RHZ model.

The authors noticed that the decomposition of  $\boldsymbol{\psi}$  into  $\mathbf{P}\boldsymbol{\psi} + \mathbf{P}^\perp\boldsymbol{\psi}$  and later into  $\mathbf{K}^T\boldsymbol{\theta}_1 + \mathbf{L}^T\boldsymbol{\theta}_2$  is mathematically correct but it does not consider the spatial structure in the model. Based on the Moran I statistic ([MORAN, 1950](#)) they proposed an alternative operator that has two main advantages: 1) it separates attractive (positive) and repulsive (negative) spatial dependence patterns; 2) it simplifies the problem to a dimension much smaller than  $n + 1$ .

The Moran operator is defined as  $\mathbf{M} = \mathbf{P}^\perp\mathbf{W}\mathbf{P}^\perp$ , where  $\mathbf{W}$  is the adjacency matrix defined in Section 1.2.1. This approach is interesting in the sense that the eigenvectors related to  $\mathbf{M}$  present spatial structure. To illustrate the first advantage of the Moran basis, we followed the [Hughes and Haran \(2013\)](#) experiment simulating a  $30 \times 30$  lattice. Each cell represents a region in this map. We took  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ , where  $\mathbf{X}_i = (x_{1i}, x_{2i})$  contains the  $x$ - and  $y$ -coordinates of the  $i$ -th lattice point, therefore it is an artificial example. Figure 3 shows eigenvectors 3, 8, 16, 20 and 26 related to each cell of the RHZ model basis and of the Moran basis (HH).



**Figure 3** – Eigenvectors 3, 8, 16, 20 and 26 from the RHZ and HH basis disposed in a  $30 \times 30$  lattice for a simulated example.

As one can notice, the selected eigenvectors related to the HH approach have positive spatial patterns and therefore the eigenvectors are taking into account the spatial structure

of the problem. The RHZ eigenvectors appear random for both cases because they are not considering the spatial structure.

The second advantage is that we can use only the first  $m \ll n$  Moran eigenvectors. There are several ways to select which eigenvectors are going to compose the basis of the HH model. The positive and negative eigenvalues correspond to variations of positive and negative spatial dependence. One could choose, for example, the eigenvectors related to the eigenvalues greater than 0 (just attractive dependence) or the  $m$  biggest eigenvalues in absolute value, or even perform a study to select the best configuration using some fit measure as DIC (SPIEGELHALTER et al., 2002) or WAIC (WATANABE, 2010).

To obtain the HH model we just need to replace  $\mathbf{L}$  by  $\mathbf{M}$  in Equation (1.2.7):

$$\begin{aligned} \mathbf{Y} | \boldsymbol{\beta}, \boldsymbol{\theta}, \tau_\epsilon &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta} + \mathbf{M}^T \boldsymbol{\theta}_2, \tau_\epsilon \mathbf{I}_n), \\ \boldsymbol{\theta}_2 | \tau_s &\sim \text{Normal}(0, \tau_s \mathbf{M} \mathbf{Q} \mathbf{M}^T), \end{aligned} \quad (1.2.8)$$

where  $\boldsymbol{\theta}_2 = \mathbf{M}\boldsymbol{\psi}$ . The HH model has  $m + q + 1$  unknown parameters instead of  $n + 1$  in the RHZ model, being computationally interesting.

### 1.2.2.3 Hanks et al. (2015)

Hanks et al. (2015) focused their effort into geostatistical data instead of areal data as the previous works. Also, the authors reported several simulation studies about inference under model misspecification.

For geostatistical data, one must assume that the correlation structure may be a function of the distance between points in the space and some parameters might govern the spatial relationship between areas. One of the most common covariance structure for continuous spatial correlation is the Matérn structure (CRESSIE, 1992). In this model each covariance matrix entry of  $\boldsymbol{\Sigma}$  is defined as

$$\Sigma_{ij} = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left( \sqrt{2\nu} \frac{d_{ij}}{\phi} \right)^\nu K_\nu \left( \sqrt{2\nu} \frac{d_{ij}}{\phi} \right)$$

where  $d_{ij}$  is the Euclidean distance between the  $i$ -th and the  $j$ -th observations,  $\sigma^2$  is the partial sill parameter,  $\nu$  is a smoothness parameter,  $\phi$  is the range parameter, and  $K_\nu$  is the modified Bessel function of the second kind.

In the RHZ and HH models,  $\boldsymbol{\Sigma}$  is fixed, then it is possible to calculate  $\mathbf{L} \mathbf{Q} \mathbf{L}^T$  and  $\mathbf{M} \mathbf{Q} \mathbf{M}^T$  just once. However, in the continuous case, the matrix  $\boldsymbol{\Sigma}$  may vary given the parameters  $\{\sigma^2, \nu, \phi\}$ . Thus, an approach similar to those previously mentioned may not be feasible because each step of MCMC (or the step in a numerical optimization routine) would require the matrix  $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$  to evaluate the likelihood.

To obtain an efficient algorithm Hanks et al. (2015) suggest the use of the conditioning by kriging technique (RUE; KNORR-HELD, 2005). The idea is to sample from the

unrestricted model  $\boldsymbol{\psi}^* \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$  and then to have a sample, under the restriction  $\mathbf{P}\boldsymbol{\psi} = \mathbf{0}$ , take  $\boldsymbol{\psi} = \boldsymbol{\psi}^* - \boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\psi}^*$ .

Through simulations, the authors showed that using the conditioning by kriging technique the inference for fixed effects is more appropriate by comparing the Type-S error (GELMAN et al., 2000). A Type-S error occurs when the regression parameters are equal to zero and the 95% posterior credibility interval does not contain the zero value.

Another important contribution of this work is the possibility to get a sample from both models (restricted and unrestricted) concurrently, along the MCMC. This is possible because there is an equivalence between the restricted model proposed by Reich, Hodges and Zadnik (2006) and the unrestricted model:

$$\begin{aligned} E(Y_i|\boldsymbol{\beta}) &= \mathbf{X}\boldsymbol{\beta}_{rsr} + \boldsymbol{\psi}_{rsr} \\ &= \mathbf{X}\boldsymbol{\beta}_{rsr} + (\mathbf{I} - \mathbf{P})\boldsymbol{\psi}_{sr} \\ &= \mathbf{X}\boldsymbol{\beta}_{rsr} + \boldsymbol{\psi}_{sr} - \mathbf{P}\boldsymbol{\psi}_{sr} \\ &= \mathbf{X}\boldsymbol{\beta}_{rsr} + \boldsymbol{\psi}_{sr} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\psi}_{sr} \\ &= \mathbf{X}(\boldsymbol{\beta}_{rsr} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\psi}_{sr}) + \boldsymbol{\psi}_{sr} \\ &= \mathbf{X}\boldsymbol{\beta}_{sr} + \boldsymbol{\psi}_{sr}, \end{aligned}$$

where  $\boldsymbol{\beta}_{rsr}$  refers to the coefficients of the restricted model proposed by Reich, Hodges and Zadnik (2006),  $\boldsymbol{\beta}_{sr}$  corresponds to the unrestricted model,  $\boldsymbol{\psi}_{rsr}$  are the latent effects of the restricted model and  $\boldsymbol{\psi}_{sr}$  the latent effects of the unrestricted model.

#### 1.2.2.4 Prates, Assunção and Rodrigues (2019)

A convenient alternative to the restricted models is the *SPatial Orthogonal Centroid "K" orrection* model (SPOCK) (PRATES; ASSUNÇÃO; RODRIGUES, 2019). The main idea of this methodology is to create a new spatial neighborhood misplacing the original centroids. The new structure does not lead to shared information with the fixed effects, however, the new structure is still capable of recovering the spatial trends in the model.

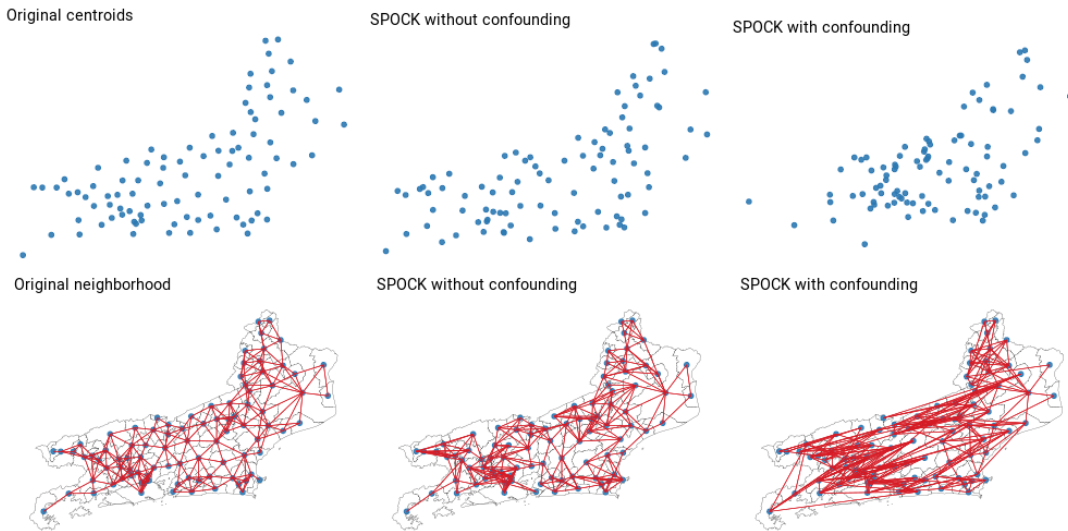
Let  $\mathbf{W}$  be the adjacency matrix of Section 1.2.1 linked to a set of original centroids  $\mathbf{c}_i = [c_{1i}, c_{2i}] \forall i \in \{1, \dots, \}$ . The goal of this method is to calculate  $\mathbf{W}^*$ , a new adjacency matrix, that conducts to a model where the spatial confounding is less intense. Given the set of original centroids, the idea is to obtain a new set of centroids pairs, let's say  $\mathbf{c}^*$  in which fixed and latent effects no longer share information in its correspondent neighborhood structure. Therefore, alleviating the spatial confounding.

To achieve the aim, the first step is to calculate the new set of centroids  $\mathbf{c}^*$  by projecting  $\mathbf{c}$  onto the orthogonal space of  $\mathbf{X}$ :

$$\mathbf{c}^* = \mathbf{P}^\perp \mathbf{c}.$$



After calculating the new set of centroids, one can find the  $k$ -neighbors of each region based on  $\mathbf{c}^*$  where  $k$  is the original number of neighbors of the region  $i$ . If for a given dataset, the model does not suffer from spatial confounding then the new set of centroids are going to be similar to the original one and as a consequence small modifications on the neighborhood structure are made as shown in Figure 4.



**Figure 4** – Centroids and neighborhood structure. Left: Original centroids and respective neighborhood; Center: SPOCK correction without spatial confounding; Right: SPOCK correction under spatial confounding

It is particularly useful because with the new neighborhood structure it is possible to use any desired and developed models or software since the SPOCK method only changes the neighborhood structure.

### 1.2.2.5 Measures of spatial confounding

A common confounding measure for spatial models is the spatial VIF as proposed by Reich, Hodges and Zadnik (2006). This measure is equivalent, for each  $\beta_j$ , to the ratio between the variance of  $\beta_j$  for the spatial model and the variance for the model without spatial component as in Equation (1.2.9). This measure reflects the increment in the variance after adding the spatial component.

For linear regression models, it is possible to calculate the exact value of these two quantities (Equation (1.2.4)). Reich, Hodges and Zadnik (2006) also note that this measure depends only on  $r = \frac{\tau_\psi}{\tau_\epsilon}$  where  $\tau_\epsilon$  is the precision of the Gaussian response, being the scale of the latent effect important in such kind of study (PACIOREK, 2010).

The SVIF is defined as:

$$SVIF(\beta_j|r, \tau_e) = \frac{Var(\beta_j)_{slr}}{Var(\beta_j)_{lr}}, \quad (1.2.9)$$

where  $Var(\beta_j)_{slr}$  is the variance of the coefficient for the spatial linear regression and  $Var(\beta_j)_{lr}$  is the variance for the linear regression.

Because under a GLMM it is not possible to derive a closed form for  $Var(\beta_j)$ , the solution is to approximate it by the Fisher information. The same occurs for frailty models and shared component models and in those cases one may use the Fisher information or the empirical variance obtained in a posterior sample. Thus one can calculate the SVIF as

$$SVIF(\beta_j) = \frac{Var(\beta_j)_{sm}}{Var(\beta_j)_m}, \quad (1.2.10)$$

where  $Var(\beta_j)_{sm}$  is the sample variance of the coefficient for the spatial model and  $Var(\beta_j)_m$  is the sample variance for the model without the spatial component.

With the  $SVIF(\beta_j)$ , one can compare two models and investigate if the variance is inflated after the latent effect inclusion. However, we are interested in evaluating the effectiveness of the restricted model. An equivalent way to measure the variance's impact is by using the variance retraction factor defined as:

$$SVRF(\beta_j) = \frac{Var(\beta_j)_u - Var(\beta_j)_r}{Var(\beta_j)_u}, \quad (1.2.11)$$

where  $Var(\beta_j)_u$  is the variance of the coefficient for the unrestricted model and  $Var(\beta_j)_r$  is the variance for the restricted model.

This measure is zero if  $Var(\beta_j)_u = Var(\beta_j)_r$ , greater than zero if  $Var(\beta_j)_u > Var(\beta_j)_r$  and less than zero otherwise. A  $SVRF(\beta_j) = 0.4$  can be interpreted as a 40% retraction in the coefficient variance under the restricted model in comparison with the unrestricted one.

## 1.3 Shared component model

### 1.3.1 Disease mapping

Disease mapping is used by professionals to obtain spatial information on the number or rate of individuals with a certain disease in specific places. Having this information it is possible to investigate local factors that can influence the highest rates. Also, it can support new policies to contain the disease growth rate.

The most common way to analyze such kind of data is using the well known Poisson model. Let  $\mathbf{Y} = [Y_1, \dots, Y_n]^T$  represent new cases counts of a specific disease in each of the  $n$  counties. Consider that  $\mathbf{T} = [T_1, \dots, T_n]^T$  is the population at risk in each county. The conventional Poisson model is given by:

$$\begin{aligned} Y_i | \theta_i &\sim \text{Poisson}(E_i \theta_i), \\ \log(\theta_i) &= \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + \psi_i, \\ \boldsymbol{\psi} &\sim \text{ICAR}(\mathbf{W}, \tau_\psi), \end{aligned}$$

where  $E_i = T_i \times r$  is the expected number of cases in region  $i$ ,  $r = \frac{\sum_i Y_i}{\sum_i T_i}$  is the overall rate and  $\theta_i$  is called relative risk. The parameter  $\beta_0$  is the intercept,  $\mathbf{X}_i$  is the vector of covariates for region  $i$ ,  $\boldsymbol{\beta}$  is the set of coefficients and  $\psi_i$  is a spatial effect related to region  $i$ . For the vector  $\boldsymbol{\psi}$  a common used model is the ICAR described in Section 1.2.1.

Although it is a useful model, in practice, multiple count responses pertaining to multiple diseases, may be recorded at the same spatial locations and then multivariate models are necessary. In a multivariate context, the natural extension of the CAR model is the multivariate conditional autoregressive model (MCAR). Several works try to extend the CAR model to MCAR setup (GELFAND; VOUNATSOU, 2003; CARLIN; BANERJEE, 2003; JIN; CARLIN; BANERJEE, 2005; JIN; BANERJEE; CARLIN, 2007; SAIN; FURRER; CRESSIE, 2011; RODRIGUES, 2012). However, these models are not intuitive and, up to now, there is no preferable option.

Also, sometimes a non-observed/latent spatially structured covariate may be important for several outcomes. A multivariate model that uses shared spatial patterns as the case of shared component model (KNORR-HELD; BEST, 2001) appears as an alternative for MCAR models.

## 1.3.2 Shared component model

Disease mapping models are commonly used in epidemiology for detecting important fixed effects as well as spatial rate patterns. In most researches, the outcome is a unique disease and then univariate models are employed. However, multivariate models are more realistic and can provide a best and more trustworthy answers for the desired questions. [Knorr-Held and Best \(2001\)](#) presented the shared component model in which two diseases are jointly modeled as Poisson distributions. The goal of this method is to understand the common (shared) spatial pattern as well as the disease-specific patterns.

To motivate the model let  $Y_{di}$  be the observed number of new cases of the disease  $d = 1, 2$  in region  $i = 1, \dots, n$ . Also let  $E_{di}$  be the expected number of cases for disease  $d$  in region  $i$ .

Then the model assumes that:

$$Y_{di} | \theta_{di} \sim \text{Poisson}(E_{di}\theta_{di}),$$

$$\log(\theta_{di}) = \begin{cases} \phi_{1i} + \delta\psi_i, & \text{if } d = 1 \\ \phi_{2i} + \frac{1}{\delta}\psi_i, & \text{if } d = 2 \end{cases},$$

where  $\phi_1$  and  $\phi_2$  are specific disease spatial effects,  $\psi$  is the spatial shared component effect and  $\delta$  is a scale parameter to allow different levels of dependence on the shared component and to guarantee the model identifiability ([DABNEY; WAKEFIELD, 2005](#)).

Important covariates can be easily inserted in the model by adding  $\mathbf{X}_d\boldsymbol{\beta}_d$  on the linear predictor as in the Equation (1.3.1).

$$\log(\theta_{di}) = \begin{cases} \beta_{10} + \mathbf{X}_{1i}\boldsymbol{\beta}_1 + \delta\psi_i + \phi_{1i}, & \text{if } d = 1 \\ \beta_{20} + \mathbf{X}_{2i}\boldsymbol{\beta}_2 + \frac{1}{\delta}\psi_i + \phi_{2i}, & \text{if } d = 2 \end{cases}, \quad (1.3.1)$$

where  $\beta_{10}$  and  $\beta_{20}$  are the intercepts for each disease and,  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$  are two set of coefficients related to the covariate matrices  $\mathbf{X}_1$  and  $\mathbf{X}_2$ .

[Knorr-Held and Best \(2001\)](#) assumed a cluster model ([KNORR-HELD; RASSER, 2000](#)) for  $\phi_1$ ,  $\phi_2$  and  $\psi$ . However, another option is to use the well known ICAR model to fit the shared component as well as specific disease spatial effects. This is interesting because the ICAR model is the most used spatial approach for areal data and it is implemented in the majority of statistical softwares.

When more than two outcomes are available [Knorr-Held et al. \(2005\)](#) suggests the following model:

$$\log(\theta_{di}) = \beta_{d0} + \mathbf{X}_{di}\boldsymbol{\beta}_d + \phi_{di} + \sum_{k=1}^K \delta_{kd}\psi_{ki}, \quad d = 1, \dots, D, \quad (1.3.2)$$

where  $K$  is the number of shared components  $\psi_k$ ,  $D$  is the number of diseases,  $\delta_k = \{\delta_{kd_1}, \dots, \delta_{kd_{n_k}}\}$  is the set of scale parameters related to the  $n_k$  relevant diseases for  $\psi_k$ ,  $\delta_{kd}$  represents the dependence of the disease  $d$  with the shared component  $\psi_k$ .

Again for model identifiability a constraint is necessary. In this case [Knorr-Held et al. \(2005\)](#) suggests a restriction in log scale

$$\sum_{l=1}^{n_k} \log(\delta_{kd_l}) = 0,$$

that is the same as before if  $d = 2$ .

## 1.4 Spatial frailty model

### 1.4.1 Survival models

Survival models are an important tool in several branches of science mainly in health data analysis. In general, the researcher is interested in using survival models to answer questions about phenomena that can be measured in units of time. Since the response is observed in units of time, it is assumed for it, distributions with support in the positive real numbers. The most commonly employed models make use of simple probability distributions such as exponential, gamma, lognormal or Weibull. More complex models rely on less conventional distributions such as the Birnbaum-Saunders ([BIRNBAUM; SAUNDERS, 1969](#)) or semi-parametric approaches as in the case of the piecewise exponential model ([FRIEDMAN, 1982](#)).

A great differential of survival models is that in practice the phenomenon of interest is not always observed. To deal with this situation without loss of information, it is necessary to insert some censoring schemes into the modeling. There are several censoring schemes in the literature being useful for a variety of real problems.

To motivate right, left and interval censoring schemes let's suppose we are conducting a study on mortality by lung cancer. The participants are interviewed once a month and individual information is collected in each checkpoint. The status (alive or dead) represents the response variable under study. The right censoring scheme is useful when the event of interest will happen above a certain registered time. For example, in one of the checkpoints some people have already died, but others are still alive. However, the study does not have more funds and the researcher must end data collection. In this case, it is not possible to register the time until death for some participants. The information you have is that the time is greater than the time at the end of the study, characterizing a right censoring scheme. Left censoring happens when the event of interest occurs below a certain registered time. For example, in the first checkpoint, some of the invited participants have already died. Then, the researcher does not know exactly when it happened. In this case, we just know that the event of interest happened before the first checkpoint, characterizing a left censoring scheme. Interval censoring happens when the event of interest occurs between two checkpoints. For example, at the end of the study, all participants have died but again the researcher does not know exactly when it happened. However, by having monthly checkpoints, the researcher can inform a time interval in which the death occurred, characterizing an interval censoring scheme.

Besides that, the censoring may happen according to some mechanism being the

most famous one the type I and type II censoring mechanisms. The former happens when the study has a fixed endpoint and then people who have not yet experienced the event are censored. The latter occurs when a number of events, defined a priori, is achieved. Thus, all other individuals are censored. For a review of survival models see [Hosmer, Lemeshow and May \(2008\)](#).

In several cases, as the cited ones (right, left and interval censoring schemes), the likelihood fits in Equation (1.4.1).

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; t) = & \prod_{d \in \mathbf{D}} f_{\boldsymbol{\theta}}(t_d) \\ & \prod_{r \in \mathbf{R}} S_{\boldsymbol{\theta}}(t_r) \prod_{l \in \mathbf{L}} (1 - S_{\boldsymbol{\theta}}(t_l)) \\ & \prod_{k \in \mathbf{K}} (S_{\boldsymbol{\theta}}(t_{k1}) - S_{\boldsymbol{\theta}}(t_{k2})), \end{aligned} \quad (1.4.1)$$

where  $\mathbf{D}$  is the set of observed failure times,  $\mathbf{R}$  is the set of right-censored sample units,  $\mathbf{L}$  is the set of left-censored sample units and  $\mathbf{K}$  is the set of interval-censored sample units. The time until failure or the censoring time is denoted by  $t$  for left and right censoring schemes. For interval censoring, two times are provided and then we denote the lower bound of this interval as  $t_{k1}$  and the upper bound as  $t_{k2}$  for a sample unit  $k$ . The distribution assumed for the phenomenon of interest is represented by  $f_{\boldsymbol{\theta}}(\mathbf{t})$  (in parametric models) and  $S_{\boldsymbol{\theta}}(\mathbf{t})$  is the survival function.

As common choices for  $f_{\boldsymbol{\theta}}(\mathbf{t})$  we can cite exponential, gamma, lognormal and Weibull distributions. The function choice leads to different forms of the  $h_{\boldsymbol{\theta}}(\mathbf{t})$  called hazard function. It measures the instantaneous risk of occurrence of an event.  $S_{\boldsymbol{\theta}}(\mathbf{t})$  is the survival function which indicates the probability of occurrence of the event at a time  $T > t$ . Any other parameters of  $f_{\boldsymbol{\theta}}(t)$  are represented by the vector  $\boldsymbol{\theta}$ .

The functions  $f_{\boldsymbol{\theta}}(t)$ ,  $S_{\boldsymbol{\theta}}(t)$  and  $h_{\boldsymbol{\theta}}(t)$  are linked through the following identities:

$$\begin{aligned} S_{\boldsymbol{\theta}}(t) &= \Pr(T > t) = \int_t^{\infty} f_{\boldsymbol{\theta}}(u) du = 1 - F_{\boldsymbol{\theta}}(t), \\ h_{\boldsymbol{\theta}}(t) &= \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt)}{dt} \times \frac{1}{S_{\boldsymbol{\theta}}(t)} = \frac{f_{\boldsymbol{\theta}}(t)}{S_{\boldsymbol{\theta}}(t)} = -\frac{S'_{\boldsymbol{\theta}}(t)}{S_{\boldsymbol{\theta}}(t)}, \\ f_{\boldsymbol{\theta}}(t) &= h_{\boldsymbol{\theta}}(t) \times S_{\boldsymbol{\theta}}(t). \end{aligned} \quad (1.4.2)$$

The survival function,  $S_{\boldsymbol{\theta}}(t)$ , has the property that  $S_{\boldsymbol{\theta}}(0) = 1$  and  $S_{\boldsymbol{\theta}}(\infty) = 0$  what means that at the beginning of the study the survival probability is 1 while if it was possible to observe an individual for an infinite time, at the end of the study the survival probability would be 0.

However, in some cases, it is possible to observe that some individuals will not experience the event of interest. In those cases one can use a cure fraction ([BOAG, 1949](#)) model where there is a proportion of the individuals which will never experience the event

of interest. The simpler way to introduce a cure fraction in the modeling is by a mixture model. In this case, the survival function is a mixture of a proper survival function and a point mass at a constant  $c$ , called cure fraction as above

$$S_{\theta}(t) = c + (1 - c)S_{\theta}^*(t).$$

Other approaches can be found in the literature as, for example, [Tsodikov, Ibrahim and Yakovlev \(2003\)](#), [Lambert \(2007\)](#), [Scudilio et al. \(2019\)](#).

In survival analysis, the interest is to model the hazard function to understand factors that impact the risk of an event. Therefore, covariates may be included into the model to measure their impact. Several parametric models are described in the literature as the cases in [Table 1](#).

**Table 1** – Distributions and its respective set of parameters ( $\theta$ ), probability density function ( $f_{\theta}$ ), survival function ( $S_{\theta}$ ) and hazard function ( $h_{\theta}$ ). The term  $\Phi(t)$  represents the cumulative distribution function of a standard Normal distribution.

Distribution	$\theta$	$f_{\theta}(t)$	$S_{\theta}(t)$	$h_{\theta}(t)$
Exponential	$\{\lambda\}$	$\lambda \exp\{-\lambda t\}$	$\exp\{\lambda t\}$	$\lambda$
Lognormal	$\{\mu, \sigma\}$	$\frac{1}{(2\pi)^{1/2}\sigma t} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}$	$1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$	$\frac{f_{\theta}(t)}{S_{\theta}(t)}$
Gamma	$\{\alpha, \lambda\}$	$\frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}$	$1 - \frac{1}{\Gamma(\alpha)} \int_0^{\lambda t} u^{\alpha-1} e^{-u} du$	$\frac{f_{\theta}(t)}{S_{\theta}(t)}$
Weibull	$\{\alpha, \lambda\}$	$\alpha \lambda t^{\alpha-1} \exp\{\lambda t^{\alpha}\}$	$\exp\{-\lambda t^{\alpha}\}$	$\alpha \lambda t^{\alpha-1}$

A well studied method to include covariates in the modeling is the Cox proportional hazards model ([COX, 1972](#)). Its idea is to insert the covariates on the hazard function in a multiplicative way ensuring that the hazard is never negative. This model assumes proportional hazards meaning that the hazard ratio for two individuals is constant over time. Next equation shows the hazard function under the Cox proportional hazard model:

$$h_{\theta}(t_i | \mathbf{X}_i) = h_{\theta}^*(t_i) \exp\{\mathbf{X}_i \boldsymbol{\beta}\}, \quad (1.4.3)$$

where  $h_{\theta}^*(t_i)$  is called baseline hazard function.

However, one can use the partial likelihood technique which makes the baseline hazard specification unnecessary ([COX, 1972](#)). Another alternative, is to create a fully parametric proportional hazard model by replacing  $h_{\theta}^*(\cdot)$  by a parametric baseline hazard function as the functions listed on [Table 1](#) ([LAWLESS, 2011](#)).

In many cases, the introduction of covariates is not enough for an appropriate fit. This is explained by the fact that often, important covariates are not observed or they are impossible to measure. In this case, one can introduce a latent effect giving rise to a frailty model.



## 1.4.2 Frailty models

Similarly to GLM models, one can introduce latent effects to take the non-observed covariates and/or clusters effects into consideration. This model family is known as frailty models (WIENKE, 2010). In general, the easiest way to introduce these effects is in a multiplicative way. Because the hazard is a positive quantity it is necessary to guarantee that the multiplicative effect will ensure that the hazard is still positive. One way is to assume that  $\gamma$ , the frailty term, is drawn from a positive probability distribution.

$$h_{\theta}(t_{ij}|\mathbf{X}_{ij}) = h_{\theta}^*(t_{ij})\gamma_j \exp\{\mathbf{X}_{ij}\boldsymbol{\beta}\},$$

where  $\gamma_j$  is called frailty (related to cluster  $j$ ) and a common choice for its distribution is the gamma distribution, given rise to the gamma frailty model. However, under this distribution, it is difficult to insert dependence structures between clusters and then they are, in general, considered independent.

## 1.4.3 Spatial frailty models

One of the interests in spatial statistics is to insert dependence between geographically close locations. Using the gamma frailty model as in Section 1.4.2, the insertion of spatial structure is not trivial. Banerjee, Wall and Carlin (2003) proposed a frailty model that allows the insertion of already known structures of spatial models. For this, the spatial variable enters the model in an additive way, but within the exponential term, which gives rise to the model presented in Equation (1.4.4). Take  $j = 1, \dots, n_i$ , indices of  $n_i$  sample units observed in the location  $i$  for  $i = 1, \dots, n$ ,  $n$  locations. The hazard function of the spatial frailty model is given by:

$$\begin{aligned} h_{\theta}(t_{ij}|\mathbf{X}_{ij}) &= h_{\theta}^*(t_{ij}) \times \gamma_i \times e^{\mathbf{X}_{ij}\boldsymbol{\beta}} \\ &= h_{\theta}^*(t_{ij}) \times e^{\mathbf{X}_{ij}\boldsymbol{\beta} + \log(\gamma_i)} \\ &= h_{\theta}^*(t_{ij}) \times e^{\mathbf{X}_{ij}\boldsymbol{\beta} + \boldsymbol{\psi}_i}, \end{aligned} \tag{1.4.4}$$

where  $\boldsymbol{\psi}_i$  is Gaussian and consequently the vector  $\boldsymbol{\psi}$  is a multivariate normal distributed. This setting is convenient since several spatial models use the multivariate normal distribution. The ICAR described in Section 1.2.1 is one example.

## 1.5 Bayesian inference

### 1.5.1 Introduction

Under the Bayesian paradigm, the lack of knowledge is measured through a probability distribution. Therefore, the main aim is to obtain the posterior distribution of the parameters in the model. Thus, data and prior information are combined to obtain the so-called posterior distribution. The prior information of specialists is described through some probability distribution and then using Bayes theorem one can obtain the posterior distribution as in Equation (1.5.1).

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto \mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \quad (1.5.1)$$

where  $\boldsymbol{\theta}$  is a set of parameters of interest,  $\pi(\boldsymbol{\theta}|\mathbf{y})$  is the  $\boldsymbol{\theta}$  joint posterior distribution,  $\mathcal{L}(\mathbf{y}; \boldsymbol{\theta})$  is the likelihood that depends either on  $\boldsymbol{\theta}$  and the data  $\mathbf{y}$ ,  $\pi(\boldsymbol{\theta})$  is the prior distribution that describes the information about  $\boldsymbol{\theta}$  before looking at the data. The posterior predictive function,  $\int_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ , does not depend on  $\boldsymbol{\theta}$ . Therefore, we say that the posterior distribution is proportional to  $\mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})$  and inference about  $\boldsymbol{\theta}$  is possible using this quantity.

However, in practice  $\boldsymbol{\theta}$  is a length  $p$  vector. Then, to obtain the marginal posterior distribution of  $\theta_j$  one must calculate

$$\pi(\theta_j|\mathbf{y}) \propto \int_{\boldsymbol{\theta}_{-j}} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}_{-j}. \quad (1.5.2)$$

Rarely the analytical solution to these integrals is available. Therefore, one must employ a computational procedure to obtain the posterior distribution. Several approaches are available being the most famous the Markov chain Monte Carlo (MCMC) methods, see (GAMERMAN; LOPES, 2006) for a review. In this method, a sample from the joint posterior distribution is obtained and inference is made based on a posterior sample. However, MCMC methods are computationally intensive and, in general, it takes a long time to achieve convergence. An attractive alternative for MCMC methods is the integrated nested Laplace approximation (RUE; MARTINO; CHOPIN, 2009).

### 1.5.2 Integrated nested Laplace approximation - INLA

Integrated nested Laplace approximation (INLA) (RUE; MARTINO; CHOPIN, 2009) is a powerful methodology that allows the user to fit a huge variety of Bayesian models. A model can be fitted in INLA if for a random variable  $\mathbf{Y}$  its mean  $\boldsymbol{\mu}$  can be

modeled through a link function  $g(\cdot)$  in an additive way as:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{n_\xi} \xi^{(j)}(z_{ji}) + \sum_{k=1}^{n_\beta} \beta_k X_{ki} + \epsilon_i, \quad (1.5.3)$$

in which  $\xi^{(j)}(z_{ji})$  are unknown functions of the covariates  $z_{ij}$ ,  $\beta_0$  is an intercept,  $\beta_k$  is a set of coefficients related to the fixed effects  $X_{ki}$  and  $\epsilon_i$  are unstructured terms. INLA assumes Gaussian priors to the vector  $\mathbf{u} = \{\beta_0, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\epsilon}\}$  giving raise to a Gaussian Markov random field (GMRF) (RUE; KNORR-HELD, 2005). If a model can be written as a GMRF, it is possible to apply the INLA methodology. Most of the common models can be fitted in this framework as for example the GLMM family described in Section 1.1.

The vector  $\mathbf{u} = \{\beta_0, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\epsilon}\}$  may depend on some hyperparameters  $\boldsymbol{\theta}$ , for example variances and correlation parameters that obey, in general,  $\dim(\mathbf{u}) \gg \dim(\boldsymbol{\theta})$ . This way, one must provide the prior distribution for the vector  $\{\mathbf{u}, \boldsymbol{\theta}\}$ . INLA assign priors  $\pi(\mathbf{u}, \boldsymbol{\theta}) = \pi(\mathbf{u}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$  where  $\pi(\mathbf{u}|\boldsymbol{\theta})$  is a GMRF and  $\pi(\boldsymbol{\theta})$  may be decomposed as  $\prod_{j=1}^{n_\theta} \pi(\boldsymbol{\theta}_j)$ . The marginal posterior distributions for the set of parameters are given by:

$$\begin{aligned} \pi(u_j|\mathbf{y}) &= \int \pi(u_j, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int \pi(u_j|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \\ \pi(\theta_k|\mathbf{y}) &= \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-k}. \end{aligned}$$

Because there is no analytical solution for these integrals, numerical approximations are necessary to obtain  $\tilde{\pi}(u_j|\mathbf{y})$  and  $\tilde{\pi}(\theta_k|\mathbf{y})$  in which  $\tilde{\pi}(\cdot)$  denotes an approximate function for  $\pi(\cdot)$ .

### 1.5.2.1 Marginal distribution for $\theta_k$

We can rewrite  $\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{u}, \boldsymbol{\theta}|\mathbf{y})}{\pi(\mathbf{u}|\boldsymbol{\theta}, \mathbf{y})}$  and, to approximate this quantity, Rue, Martino and Chopin (2009) suggest a Gaussian approximation for the denominator becoming

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{u}, \boldsymbol{\theta}, \mathbf{y})}{\pi_G(\mathbf{u}|\boldsymbol{\theta}, \mathbf{y})} \Bigg|_{\mathbf{u}=\mathbf{u}^*(\boldsymbol{\theta})},$$

where  $\pi_G(\cdot)$  is the Gaussian approximation of a density,  $\mathbf{u}^*(\boldsymbol{\theta})$  is the mode of  $\pi(\mathbf{u}|\boldsymbol{\theta}, \mathbf{y})$  to a given  $\boldsymbol{\theta}$ . The better  $\mathbf{u}$  approximates a Gaussian distribution, the better INLA works.

Now, to obtain the marginal distribution  $\tilde{\pi}(\theta_k|\mathbf{y})$  a numerical integration is made. Thus, a grid of  $\theta_k$  values is taken and the marginal is obtained via

$$\pi(\theta_k|\mathbf{y}) = \sum_{h=1}^H \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\Delta_{kh}.$$

### 1.5.2.2 Marginal distribution for $u_j$

Rue, Martino and Chopin (2009) propose three different approximations for this quantity: 1) Gaussian approximation; 2) Laplace approximation, and; 3) simplified Laplace approximation. The Gaussian approximation is the easiest to be obtained but it provides poorer results. At the cost of being computationally expensive, Laplace approximation produces better results. The simplified Laplace approximation is a simplification of the last approach and it gives satisfactory results with a good computational time. Taking one of them as approximation for  $\tilde{\pi}(u_j|\boldsymbol{\theta}, y)$ , one can calculate the posterior marginal distribution as

$$\tilde{\pi}(u_j|\mathbf{y}) \approx \sum_{h=1}^H \tilde{\pi}(u_j|\theta_h^*, \mathbf{y}) \tilde{\pi}(\theta_h^*|\mathbf{y}) \Delta_h.$$

In this part, we have presented important contributions that will be useful in the rest of this work. We presented some approaches to alleviate the spatial confounding issue in GLMM as well as two important models that do not fit in the GLMM framework. The next part aims to describe and summarise the data sets in use in our work.

## Part 2

### Data sources

## 2.1 Data sources

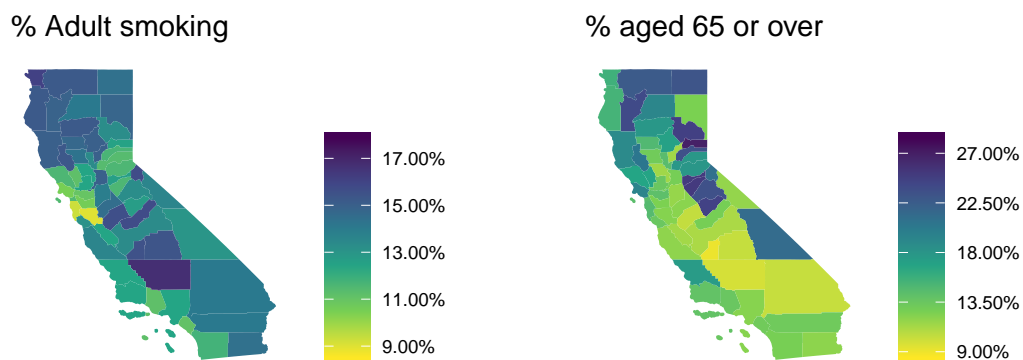
To motivate our contributions, two applications are presented being one of them for the shared component model and another one for the spatial frailty model. For the shared component model, we have an areal level analysis and therefore we may investigate the effects of some areal level covariates. On the other hand, for survival analysis, we will use both individual-level and areal-level covariates to characterize individuals and regions. Therefore, two main data sources are in use and are described below, in addition to the two applications. In both cases, the data were collected in the 58 counties of the California state (US).

### 2.1.1 CHRR dataset

For both analyses, we would like to describe each California's county by important covariates that represent each county. Said that, to add county-level information we are using the County Health Rankings & Roadmaps (CHRR) ([RANKINGS; CHRR, 2019](#)) which provides several important indices collected from different sources in the US. The CHRR is provided by the University of Wisconsin, Population Health Institute.

We took two areal level information considered important covariates for both analyses between 2010 and 2016. The areal level covariates are the percentage of adults that smoke every day or most days and the percentage of the population aged 65 or over. Additional information about covariates as well as the dataset can be obtained in the CHRR website <https://www.countyhealthrankings.org/>.

Figure 5 shows the spatial pattern of these covariates for the California state in 2016, year considered in our shared component model application. It is possible to observe



**Figure 5** – Covariates from CHRR in California (US) in 2016.

that these covariates seem to have a spatial pattern which is a factor that can lead to a

spatial confounding problem. Table 2 presents some summary information about these covariates. The covariates are presented as proportions varying from 0.09 to 0.17 for the

**Table 2** – Summary statistics of CHRR covariates (SD: Standard deviation).

Index	Mean	Minimum	Quantile 0.025	Median	Quantile 0.975	Maximum	SD
% Adult smoking	0.13	0.09	0.10	0.13	0.16	0.17	0.02
% Ages 65 and above	0.16	0.09	0.10	0.14	0.25	0.27	0.05

percentage of adult smokers and from 0.09 to 0.27 for the percentage of people aged 65 or over. The break-point at 65 years old is supported by the literature that frequently divides the age into groups, being the most critical the elderly patients aged 65 or older (BARANOVSKY; MYERS, 1986; YANCIK; KESSLER; YATES, 1988; YANCIK; RIES, 1991).

## 2.1.2 SEER dataset

The main database in this work is provided by the Surveillance, Epidemiology, and End Results Program (SEER) (SEER, 2019). The SEER program collects data on cancer cases from several locations and sources since 1973 in the United States of America. The data is provided by the National Cancer Institute (NHI) the American national leader in cancer research. The SEER datasets are not publicly available for free download, it can be requested following NIH/NCI SEER data access options via the link: <https://seer.cancer.gov/data/options.html>.

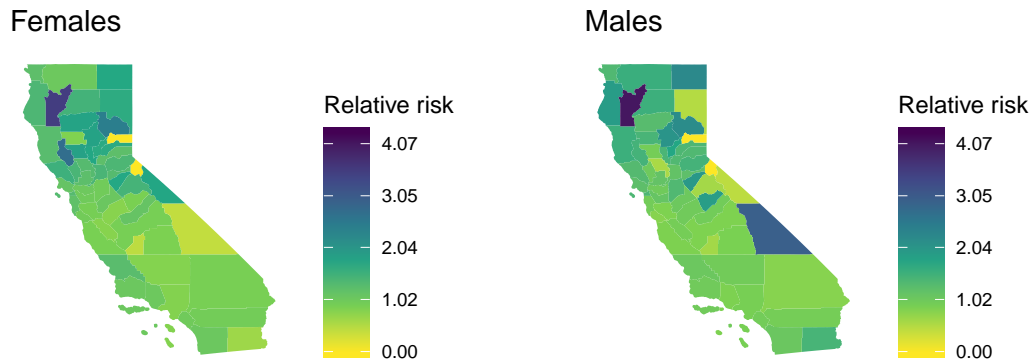
### 2.1.2.1 Incidence of bronchus and lung cancer

SEER provides for each subject information about the cancer status and individual characteristics. Two important characteristics provided are the gender and the county code of each individual. Using the gender information we can calculate, for each gender, the number of new cases of bronchus and lung cancer for each county. These quantities are our outcomes and we would like to understand the characteristics that possibly affect the incidence of this kind of cancer for men and women. Also, we would like to observe whether the spatial pattern is shared or not for men and women and whether there is a specific spatial pattern for each gender or not.

It is well known in the literature that the incidence of bronchus and lung cancer is bigger in men than in women (WHO, 2004; FU et al., 2005). Following the literature, the risk factors are in general similar for both being age and tobacco consumption the most important ones (BOLOKER; WANG; ZHANG, 2018; SIEGEL; MILLER; JEMAL, 2019). Thus, we are using county-level indices that represent an average pattern of these characteristics for each county. In this case, the variables were the percentage of the

population aged 65 or over and the percentage of adults that smoke every day or most days from the CHRR dataset.

For 2016 a total of 5,143 new cases of lung and bronchus cancer were registered for women while for men we observed 5,640 new cases. Figure 6 shows the spatial pattern of the relative risk of lung and bronchus cancer for the California state. A common spatial



**Figure 6** – Relative risk of lung and bronchus cancer in California (US). Source: SEER.

pattern is readily observed. However, our task is to adjust the relative risk by important covariates and observe if the remaining effect is spatially shared between men and women.

### 2.1.2.2 Time until death by bronchus and lung cancer

This data set provided by SEER has cases of lung and bronchus cancer for 72,612 individuals (after cleaning the data between 2010 and 2016) in the California state. In this data set, important covariates are present as gender and the disease stage for each individual. However, important covariates are missing as the case of tobacco consumption. To work around this problem, we are using the areal level covariate that indicates the percentage of adults that smoke every day or most days. Because people start in the program in different years (2010 - 2016), we picked the corresponding statistic in the CHRR dataset.

Table 3 presents some summary information about the individual covariates used from the SEER program as well as the areal level covariate. Continuous covariates are represented by median (quantiles 25% and 75%) and categorical variables are represented by its observed proportion.

The time collected is measured in months and the median is 10 months. The time variable was scaled ( $\text{time}/\max\{\text{time}\}$ ) in our models to avoid computational instabilities.

For each individual, we assign a status of 1 if the individual died by lung or bronchus cancer and zero if the individual died by other causes or is still alive. Therefore, it is a right censoring problem with about 42% censored cases. Given the current dataset structure, it can be seen as a type I censoring scheme.



**Table 3** – Summary statistics of SEER covariates. For categorical variables the sample size and the percentage. For continuous variables median and quantiles 25% and 75%.

Variable	N = 72612
Time until death	10.0 [4.00; 25.0]
Status	
0	31013 (42.7%)
1	41599 (57.3%)
Gender	
Female	34625 (47.7%)
Male	37987 (52.3%)
Race	
Non-black	66723 (91.9%)
Black	5889 (8.11%)
Cancer stage	
In situ	519 (0.71%)
Localized	15870 (21.9%)
Regional	16792 (23.1%)
Distant	39431 (54.3%)
Age at diagnosis	69.0 [62.0; 77.0]
% Smokers	0.14 [0.12; 0.15]

Also, our sample has more cases of lung and bronchus cancer for men than for women as expected. For simplicity, we considered just black and non-black people.

In situ refers to abnormal cells that are present but have not spread to nearby. Located corresponds to the stage where the cancer is limited to where it started and has not spread. In the regional phase, cancer has spread to nearby lymph nodes or organs. The last and more severe phase is the distant stage. In the distant stage, cancer has spread to distant parts of the body. Therefore, we expect an increase in the risk of death following this order. Finally, the median age is 69, which corresponds to an elderly population. The county-level percentage of smokers is around 14%.

The next part will introduce our first contribution providing a methodology to alleviate the spatial confounding in shared component models. Therefore, we provide the method, a simulation study, and one application.

## Part 3

# Spatial confounding in shared component models

## 3.1 Method

As presented in Section 1.2.2, several approaches are available to alleviate the spatial confounding in univariate GLMM models. However, few or no alternatives are available for models that differ from GLMM's conventional approach.

In Section 1.3.2 we introduced the shared component model (SCM) as a good alternative for spatial multivariate models where two or more outcomes share the same spatial structure. Calling Equation (1.3.2), the linear predictor is given by

$$\log(\theta_{di}) = \beta_{d0} + \mathbf{X}_d \boldsymbol{\beta}_d + \phi_{1i} + \sum_{k=1}^K \delta_{kd} \psi_{ki}, d = 1, \dots, D,$$

with the constraint

$$\sum_{l=1}^{n_k} \log(\delta_{kl}) = 0,$$

where  $K$  is the number of shared components  $\boldsymbol{\psi}_k$ ,  $D$  is the number of diseases,  $\boldsymbol{\delta}_k = \{\delta_{kd_1}, \dots, \delta_{kd_{n_k}}\}$  is the set of scale parameters related to the  $n_k$  relevant diseases for  $\boldsymbol{\psi}_k$  and  $\delta_{kd}$  represents the dependence of the disease  $d$  with the shared component  $\boldsymbol{\psi}_k$ . In this case, there are  $D$  sets of covariates, one for each disease. There are also  $K + D$  latent effects that make it difficult to restrict spatial effects to be orthogonal to fixed effects as proposed by Reich, Hodges and Zadnik (2006) or Hughes and Haran (2013).

Therefore, our proposal is to create  $K + D$  different adjacency matrices using SPOCK. Each adjacency matrix is going to be responsible to alleviate the spatial confounding of a latent effect  $\phi_i$  or  $\boldsymbol{\psi}_k$ . The  $K$  adjacency matrices  $\mathbf{W}_{\boldsymbol{\psi}_k}$ ,  $k = 1, \dots, K$  are going to be created for each shared component based on the covariates that can affect this quantity (the union of  $\mathbf{X}_d$  related to the  $n_k$  relevant diseases). For the specific spatial components  $\phi_d$ ,  $d = 1, \dots, D$ ,  $D$  adjacency matrices  $\mathbf{W}_{\phi_d}$  are going to be created based on the specific covariates  $\mathbf{X}_d$  of each disease. We named this approach as restricted shared component model (RSCM). To illustrate let's assume  $D = 2$ . In this case, we have two covariate matrices  $\mathbf{X}_1, \mathbf{X}_2$ , and three spatial effects  $\phi_1, \phi_2$  and  $\boldsymbol{\psi}$ . Let  $\mathbf{W}$  and  $\mathbf{c}$  be the original adjacency matrix and centroids, respectively. Let's suppose that  $\mathbf{X}_1$  and  $\mathbf{X}_2$  have two columns, being the first column shared but not the second. In this case,  $\mathbf{X}_{11}$  and  $\mathbf{X}_{21}$  are the same variable, then the adjacency matrix  $\mathbf{W}_{\boldsymbol{\psi}}$  for the shared component  $\boldsymbol{\psi}$  is created assuming that

$$\begin{aligned} \mathbf{X}_{\boldsymbol{\psi}} &= [\mathbf{1}^T, \mathbf{X}_{11}, \mathbf{X}_{12}, \mathbf{X}_{22}] \\ \mathbf{P}_{\boldsymbol{\psi}} &= \mathbf{X}_{\boldsymbol{\psi}} (\mathbf{X}_{\boldsymbol{\psi}}^T \mathbf{X}_{\boldsymbol{\psi}})^{-1} \mathbf{X}_{\boldsymbol{\psi}}^T \\ \mathbf{c}_{\boldsymbol{\psi}} &= \mathbf{P}_{\boldsymbol{\psi}}^{\perp} \mathbf{c}. \end{aligned}$$

Based on  $\mathbf{c}_{\boldsymbol{\psi}}$  we can create  $\mathbf{W}_{\boldsymbol{\psi}}$  using the k-nearest neighbors algorithm.

Also, matrices  $\mathbf{X}_1, \mathbf{X}_2$ , specific for each disease, are going to be used to create  $\mathbf{W}_{\phi_1}$  and  $\mathbf{W}_{\phi_2}$  based on  $c_{\phi_1}$  and  $c_{\phi_2}$  respectively as

$$\begin{aligned} \mathbf{P}_{\phi_d} &= \mathbf{X}_{\phi_d} (\mathbf{X}_{\phi_d}^T \mathbf{X}_{\phi_d})^{-1} \mathbf{X}_{\phi_d}^T \\ c_{\phi_d} &= \mathbf{P}_{\phi_d}^\perp c. \end{aligned}$$

Thus, the model becomes

$$\begin{aligned} Y_{di} | \theta_{di} &\sim \text{Poisson}(E_{di} \theta_{di}), \\ \log(\theta_{di}) &= \begin{cases} \beta_{10} + \mathbf{X}_1 \boldsymbol{\beta}_1 + \delta \psi_i + \phi_{1i}, & \text{if } d = 1 \\ \beta_{20} + \mathbf{X}_2 \boldsymbol{\beta}_2 + \frac{1}{\delta} \psi_i + \phi_{2i}, & \text{if } d = 2 \end{cases}, \end{aligned} \quad (3.1.1)$$

$$\boldsymbol{\psi} \sim \text{ICAR}(\mathbf{W}_\psi, \tau_\psi); \quad \boldsymbol{\phi}_1 \sim \text{ICAR}(\mathbf{W}_{\phi_1}, \tau_{\phi_1}); \quad \boldsymbol{\phi}_2 \sim \text{ICAR}(\mathbf{W}_{\phi_2}, \tau_{\phi_2}).$$

At this point, one can use its preferred software to fit the RSCM model which makes the use of our approach advantageous (inherited from SPOCK). For fast computation, we are going to employ the INLA methodology to get posterior estimates of the parameters. This way, we are using the R package here named R-INLA (LINDGREN; RUE et al., 2015) to make a difference to the INLA methodology.

Although INLA is a fast alternative for Bayesian models, it is not possible to fit the shared component model directly using the R-INLA package. The reason is that, for each disease, we have unknown weights  $\delta$  and  $\frac{1}{\delta}$  that allow different levels of dependence from the shared effect. To overcome this problem, Gómez-Rubio and Palmí-Perales (2019) suggest the use of INLA within the Metropolis-Hastings algorithm in which the dependence parameter  $\delta$  can be obtained iteratively using an appropriated MCMC technique while other parameters are obtained via INLA methodology. Nevertheless, using the `copy` feature in the R-INLA package, it is possible to write a latent model as linear combination of another as in Equation (3.1.2):

$$\log(\theta_{1i}) = \begin{cases} \beta_{10} + \mathbf{X}_1 \boldsymbol{\beta}_1 + \gamma \psi_i^* + \phi_{1i}, & \text{if } d = 1 \\ \beta_{20} + \mathbf{X}_2 \boldsymbol{\beta}_2 + \psi_i^* + \phi_{2i}, & \text{if } d = 2 \end{cases}, \quad (3.1.2)$$

in which  $\gamma$  works as a coefficient for the latent effect  $\boldsymbol{\psi}$ .

Equation (3.1.2) is not as (3.1.1). However, Vargas (2013) showed that they are equivalent. Following Equation (3.1.1),  $\delta \psi \sim \text{ICAR}(\mathbf{W}_\psi, \frac{\tau_\psi}{\delta^2})$  and  $\frac{1}{\delta} \psi \sim \text{ICAR}(\mathbf{W}_\psi, \tau_\psi \delta^2)$ . Also, by Equation (3.1.2), we have  $\gamma \psi^* \sim \text{ICAR}(\mathbf{W}_\psi, \frac{\tau_\psi^*}{\gamma^2})$  and  $\psi^* \sim \text{ICAR}(\mathbf{W}_\psi, \tau_\psi^*)$ . We would like to have an equality of distributions, then we need to make  $\frac{\tau_\psi}{\delta^2} = \frac{\tau_\psi^*}{\gamma^2}$  and  $\tau_\psi \delta^2 = \tau_\psi^*$ . These equations imply, for a positive  $\gamma$ , that  $\delta = \sqrt{\gamma}$  and also  $\tau_\psi = \frac{\tau_\psi^*}{\gamma}$ . This way, it is possible to recover the  $\delta$  parameter of Equation (3.1.1) using the R-INLA package

---

in the case of two diseases. To get estimates of  $\delta$  it is possible to access the marginal transformation,  $\delta = \sqrt{\gamma}$ , easily using a marginal transformation (`inla.tmarginal` function in R-INLA package for example).

## 3.2 Simulation

To evaluate the model ability to recover parameters we performed a simulation study. Data were generated from the shared component model:

$$Y_{di}|\theta_{di} \sim \text{Poisson}(E_{di}\theta_{di}), \quad (3.2.1)$$

$$\log(\theta_{di}) = \begin{cases} \beta_{10} + \mathbf{X}_1\boldsymbol{\beta}_1 + \delta\psi_i + \phi_{1i}, & \text{if } d = 1 \\ \beta_{20} + \mathbf{X}_2\boldsymbol{\beta}_2 + \frac{1}{\delta}\psi_i + \phi_{2i}, & \text{if } d = 2 \end{cases},$$

$$\boldsymbol{\psi} \sim \text{ICAR}(\mathbf{W}_\psi, \tau_\psi), \quad \boldsymbol{\phi}_i \sim \text{ICAR}(\mathbf{W}_{\phi_i}, \tau_{\phi_i}) \text{ for } i \in \{1, 2\},$$

where  $\beta_{10} = 0.5$ ,  $\beta_{20} = 0.1$ ,  $\boldsymbol{\beta}_1 = [-0.5, -0.2]$ ,  $\boldsymbol{\beta}_2 = [-0.8, -0.4]$ ,  $\tau_\psi = 1$ ,  $\tau_{\phi_1} = 10$ ,  $\tau_{\phi_2} = 10$ . For the  $\delta$  parameter we considered the following grid  $\delta = \{1.00, 1.50, 1.75\}$ .

We generated 1000 datasets considering 4 scenarios. For all cases,  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are composed by two covariates being the first column ( $\mathbf{X}_{11}$  and  $\mathbf{X}_{21}$ ) the same (shared) covariate generated without any spatial structure (drawn from a Gaussian distribution). In the first scenario,  $\mathbf{X}_{12}$  and  $\mathbf{X}_{22}$  are independent random variables (drawn from a Gaussian distribution). In the second case,  $\mathbf{X}_{12_i}$  is the first coordinate of  $i$ -th area centroid and  $\mathbf{X}_{22}$  is a random covariate (drawn from a Gaussian distribution). In the third scenario,  $\mathbf{X}_{22_i}$  is the first coordinate of  $i$ -th area centroid and  $\mathbf{X}_{12}$  is a random covariate (drawn from a Gaussian distribution). Finally, in the fourth scenario,  $\mathbf{X}_{12_i}$  and  $\mathbf{X}_{22_i}$  are the first coordinate of the  $i$ -th area centroid.

The set of areal coordinates is spatially structured and therefore we expect to suffer from spatial confounding when using it. The first scenario should not present any spatial confounding behavior as it does not have any spatially structured covariate. It is going to be our baseline model called here as S1. The second scenario (S2) presents a spatially correlated covariate only for the first disease and then we may suffer from spatial confounding for this disease. Similarly, in the third scenario (S3) we may suffer from spatial confounding for the second disease. In the fourth scenario (S4) the confounded covariate is also shared and then we should suffer from spatial confounding for both diseases.

In our simulations we are using the California spatial structure and the set of weakly informative priors was taken as follow:

$$\beta_{dj} \sim \text{Normal}(0, 0.001), \quad d = 1, 2; \quad j = 0, 1, 2,$$

$$\log(\gamma) \sim \text{Normal}(0, 0.1) \quad (\delta = \sqrt{\gamma})$$

$$\tau_\psi \sim \Gamma(0.5, 0.05); \quad \tau_{\phi_1} \sim \Gamma(0.5, 0.05) \quad \tau_{\phi_2} \sim \Gamma(0.5, 0.05)$$

Table 4 presents the mean of the estimated values (Mean), the mean of the standard deviations (SD), the coverage rate for a nominal rate of 95% (Cov) and the mean squared

error (MSE) for the first and the fourth scenarios. The entire simulation results can be seen in Appendix A. It is possible to observe that without spatial confounding, SCM and

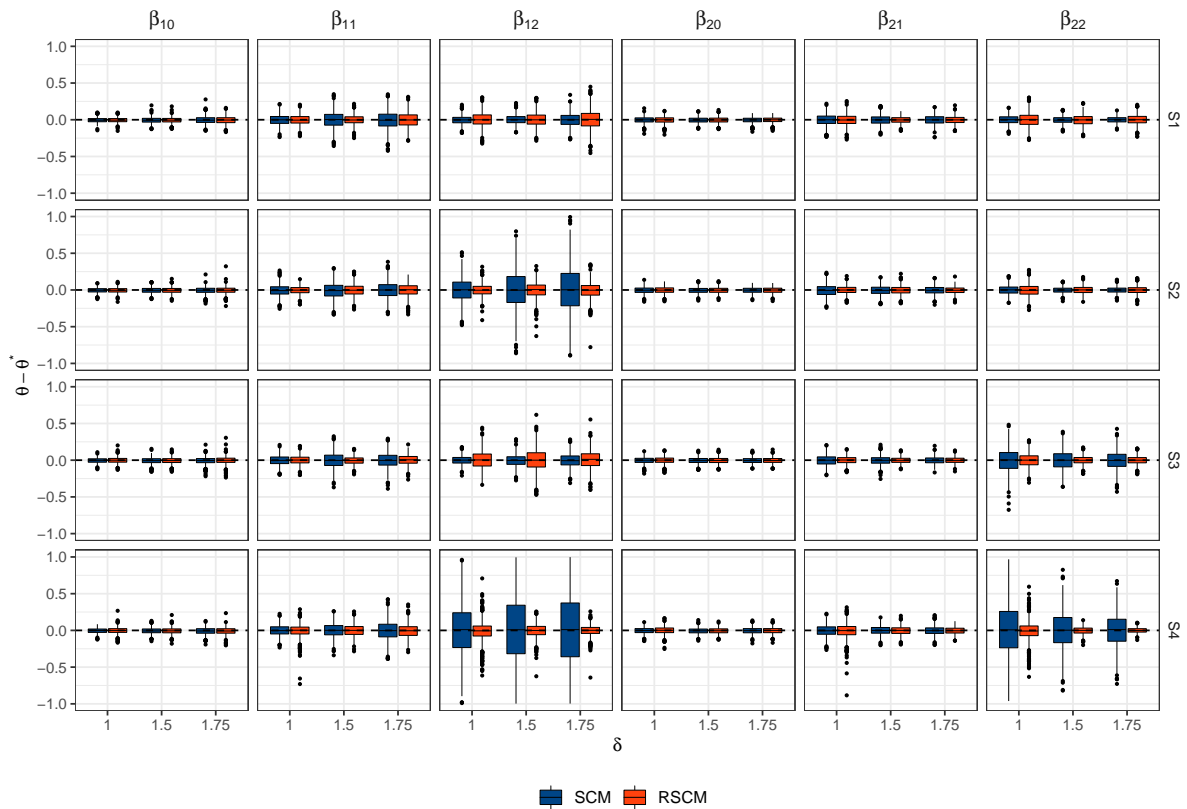
**Table 4** – Simulation results for scenarios 1 and 4 for the shared component model experiment. The results are shown by mean, standard deviation (SD), coverage rate for a nominal rate of 95 % (Cov) and mean square error (MSE).

Parameter	Real	Without spatial confounding						With spatial confounding						
		SCM			RSCM			SCM			RSCM			
		Mean (SD)	Cov	MSE	Mean (SD)	Cov	MSE	Mean (SD)	Cov	MSE	Mean (SD)	Cov	MSE	
$\delta = 1.00$	$\beta_{10}$	0.50	0.49 (0.03)	95.90%	0.0010	0.49 (0.03)	95.70%	0.0010	0.50 (0.03)	93.70%	0.0011	0.50 (0.03)	91.70%	0.0016
	$\beta_{20}$	0.10	0.10 (0.04)	94.20%	0.0018	0.10 (0.04)	94.30%	0.0019	0.10 (0.04)	95.30%	0.0018	0.10 (0.04)	90.60%	0.0024
	$\beta_{11}$	-0.50	-0.50 (0.07)	95.30%	0.0048	-0.50 (0.08)	97.00%	0.0045	-0.50 (0.07)	94.30%	0.0052	-0.50 (0.09)	97.30%	0.0170
	$\beta_{21}$	-0.80	-0.80 (0.07)	93.30%	0.0056	-0.80 (0.08)	96.10%	0.0058	-0.80 (0.08)	94.50%	0.0057	-0.80 (0.09)	96.00%	0.0139
	$\beta_{12}$	-0.20	-0.20 (0.05)	94.50%	0.0031	-0.20 (0.06)	78.30%	0.0088	-0.19 (0.35)	93.80%	0.1234	-0.21 (0.10)	86.00%	0.0917
	$\beta_{22}$	-0.40	-0.40 (0.06)	95.40%	0.0033	-0.40 (0.06)	78.60%	0.0087	-0.40 (0.34)	93.20%	0.1243	-0.41 (0.10)	88.00%	0.0532
	$\delta$	1.00	1.02 (0.06)	92.20%	0.0044	1.03 (0.06)	88.60%	0.0060	1.02 (0.05)	91.20%	0.0041	1.03 (0.05)	76.60%	0.0070
$\delta = 1.50$	$\beta_{10}$	0.50	0.49 (0.04)	95.20%	0.0015	0.49 (0.04)	95.40%	0.0014	0.49 (0.04)	94.70%	0.0016	0.49 (0.04)	96.10%	0.0017
	$\beta_{20}$	0.10	0.09 (0.04)	93.80%	0.0016	0.10 (0.04)	93.80%	0.0016	0.09 (0.04)	93.80%	0.0015	0.09 (0.04)	94.10%	0.0016
	$\beta_{11}$	-0.50	-0.50 (0.11)	94.30%	0.0119	-0.50 (0.10)	99.60%	0.0040	-0.50 (0.09)	95.20%	0.0084	-0.50 (0.13)	99.70%	0.0067
	$\beta_{21}$	-0.80	-0.80 (0.07)	95.80%	0.0040	-0.80 (0.06)	99.10%	0.0019	-0.80 (0.06)	94.10%	0.0034	-0.80 (0.07)	98.10%	0.0031
	$\beta_{12}$	-0.20	-0.20 (0.06)	93.10%	0.0042	-0.20 (0.07)	83.80%	0.0083	-0.19 (0.50)	93.10%	0.2604	-0.19 (0.13)	99.70%	0.0084
	$\beta_{22}$	-0.40	-0.40 (0.05)	94.40%	0.0023	-0.40 (0.05)	82.20%	0.0051	-0.40 (0.25)	93.60%	0.0632	-0.40 (0.07)	99.10%	0.0026
	$\delta$	1.50	1.52 (0.10)	93.50%	0.0129	1.53 (0.10)	89.30%	0.0149	1.52 (0.11)	92.00%	0.0136	1.54 (0.10)	83.40%	0.0238
$\delta = 1.75$	$\beta_{10}$	0.50	0.50 (0.05)	94.50%	0.0024	0.50 (0.05)	94.60%	0.0026	0.49 (0.05)	94.10%	0.0022	0.49 (0.05)	96.50%	0.0023
	$\beta_{20}$	0.10	0.10 (0.04)	94.20%	0.0015	0.10 (0.04)	94.50%	0.0014	0.10 (0.04)	94.30%	0.0017	0.10 (0.04)	94.10%	0.0018
	$\beta_{11}$	-0.50	-0.50 (0.12)	94.30%	0.0139	-0.50 (0.13)	99.50%	0.0095	-0.50 (0.13)	94.70%	0.0170	-0.50 (0.16)	100.00%	0.0083
	$\beta_{21}$	-0.80	-0.80 (0.06)	94.90%	0.0037	-0.80 (0.06)	98.20%	0.0028	-0.80 (0.06)	94.60%	0.0034	-0.80 (0.07)	98.80%	0.0022
	$\beta_{12}$	-0.20	-0.20 (0.09)	94.40%	0.0083	-0.20 (0.10)	84.80%	0.0180	-0.19 (0.58)	93.90%	0.3437	-0.19 (0.14)	99.80%	0.0046
	$\beta_{22}$	-0.40	-0.40 (0.05)	95.60%	0.0018	-0.40 (0.05)	80.90%	0.0046	-0.40 (0.22)	94.70%	0.0477	-0.40 (0.05)	99.70%	0.0011
	$\delta$	1.75	1.79 (0.14)	94.10%	0.0234	1.78 (0.13)	87.90%	0.0295	1.78 (0.13)	91.80%	0.0208	1.79 (0.12)	82.60%	0.0341

RSCM results are similar in terms of mean, standard deviation and mean squared error. These results are expected because without confounding the adjacency matrices under RSCM are almost the same as the original ones. Then, the fitted models should be almost the same.

Under spatial confounding, it is possible to observe that the point estimates (mean) are accurate for all parameters and both models. However, the standard deviations of  $\beta_{21}$  and  $\beta_{22}$  are, on average, greater for the SCM model than for the RSCM model. Also, it is possible to observe greater MSEs for SCM than RSCM approaches. The result shows how the spatial confounding can affect the model fit by increasing variance and bias.

Because the SPOCK method estimates  $\beta^* = \beta + P^\perp \psi$  (PRATES; ASSUNÇÃO; RODRIGUES, 2019), RSCM also does. In this case, we reported all results in the Figure 7 for  $(\theta - \theta^*)$  where  $\theta = \{\beta_{10}, \beta_{20}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$ . We expect all the values to be around 0, which means that the estimate is unbiased. We can see, for  $\beta_{10}$ ,  $\beta_{20}$ ,  $\beta_{11}$  and  $\beta_{21}$ , that the estimates are similar for both SCM and RSCM models in all scenarios. However, for  $\beta_{12}$  and  $\beta_{22}$ , the behavior changes according to each scenario. For S1, as expected, it behaves similarly to the other parameters. For S2, S3 and S4, scenarios we expect to suffer from spatial confounding. Then, we can observe that the difference  $(\theta - \theta^*)$  tends to be around 0 for SCM as well as for the RSCM model. Although they are centered at 0, the dispersion of SCM is bigger than the dispersion of the RSCM model, which indicates that there is variance inflation in the estimates ( $\beta_{12}$  for S2 and S4 and  $\beta_{22}$  for S3 and S4).



**Figure 7** – Boxplot of  $(\theta - \theta^*)$  for  $\theta = \{\beta_{10}, \beta_{20}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$  in the shared component model. Dashed line represents the value 0.

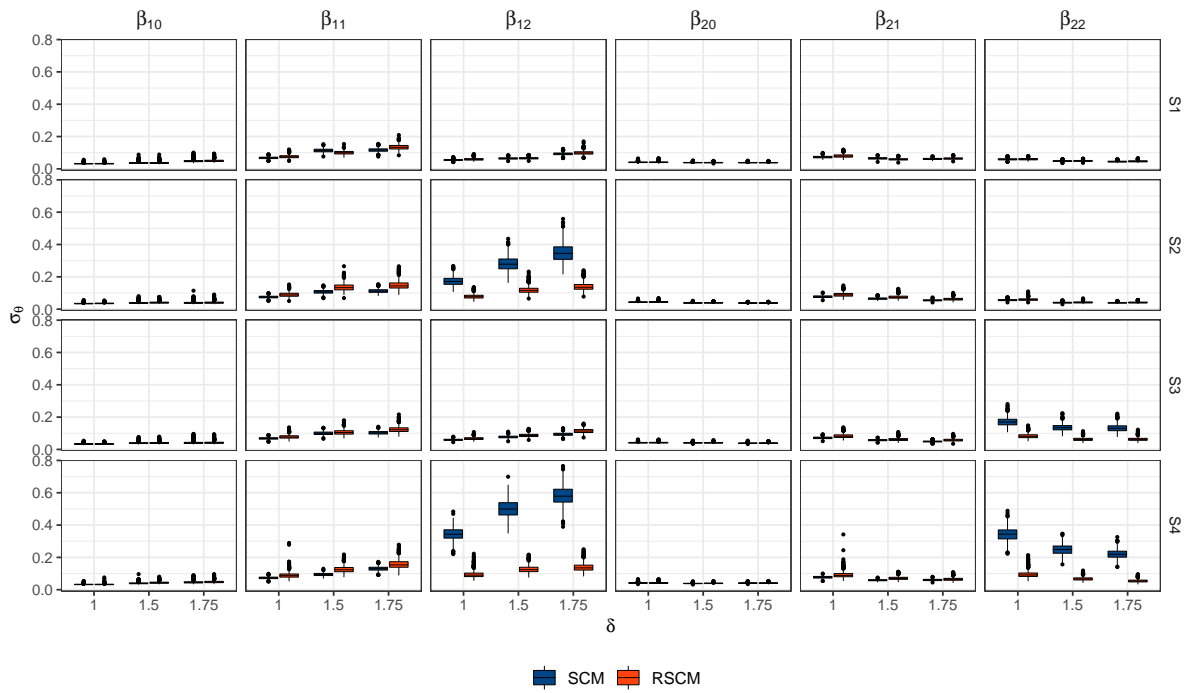
The RSCM approach appears as a good alternative to deal with the spatial confounding problem, since the model estimates are similar when the spatial confounding is not present (S1) and the standard deviation and MSE are smaller for the confounded covariates in scenarios S2, S3, and S4. Besides that, using the RSCM approach, the user is free to choose its preferred software due to the fact that RSCM changes the neighborhood structures before fitting the model.

Also, the  $\delta$  parameter was well recovered, which means that we are able to estimate the shared component model, for two diseases, using the R-INLA package. For more than two diseases other approaches may be considered. One can employ a pure MCMC approach that may take a long time due to the complexity of the model or the approach provided by [Gómez-Rubio and Palmí-Perales \(2019\)](#) in which they use INLA within MCMC, for example.

Figure 8 shows the coefficients' standard deviation. As one can notice, the standard deviations are similar when comparing the SCM and the RSCM for all parameters except those which we have spatial confounding ( $\beta_{12}$  for S2 and S4 and  $\beta_{22}$  for S3 and S4). Combining Figures 7 and 8 we can observe that we have, for  $\beta_{12}$  (S2 and S4) and  $\beta_{22}$  (S3 and S4), bias and variance inflation, the effects of spatial confounding.

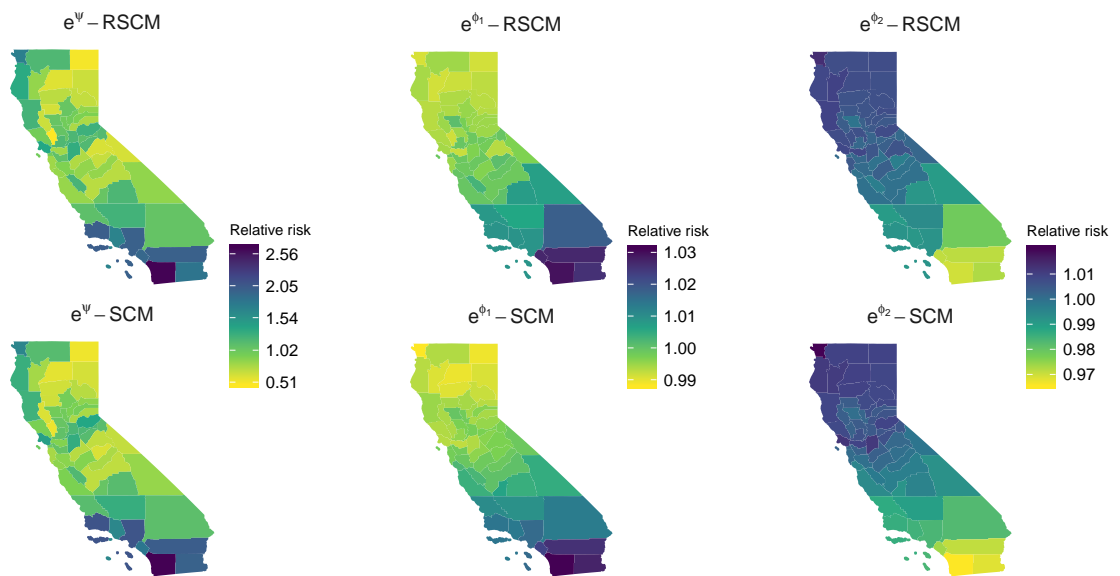
For a simulated data set, we can also compare the estimated spatial effects of the





**Figure 8** – Boxplot of  $\sigma_\theta$  for  $\theta = \{\beta_{10}, \beta_{20}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$  in the shared component model where  $\sigma_\theta$  represents the standard deviation of  $\theta$ .

SCM and RSCM models. Figure 9 presents these spatial patterns for the set of relative risks  $\{\exp\{\psi\}, \exp\{\phi_1\}, \exp\{\phi_2\}\}$ . As it can be seen, even changing the neighborhood structure, the spatial pattern remains very similar for both approaches. It is another motivation for using the RSCM correction since it is still capable to recover the spatial pattern without inflating the variances of the coefficients.



**Figure 9** – Estimated spatial patterns of a simulated dataset without spatial confounding for the shared component model.

### 3.3 Male vs female lung and bronchus cancer incidence in California

In our application, we are interested in finding the relevance of some fixed effects as well as checking whether the two outcomes (new cases of lung and bronchus cancer for men and women) share the same spatial pattern or not. As we are also interested in investigating the occurrence of spatial confounding in this application, we are going to fit spatial and non-spatial models as well as univariate and multivariate models. We suggest to fit the following models:

$\mathcal{M}_1$  - Univariate non-spatial model:

$$\begin{aligned} Y_{di} | \theta_{di} &\sim P(E_{di} \theta_{di}), \\ \log(\theta_{di}) &= \beta_{d0} + \mathbf{X}_{di} \boldsymbol{\beta}, \\ \mathbf{Y}_1 &\perp\!\!\!\perp \mathbf{Y}_2. \end{aligned}$$

$\mathcal{M}_2$  - Univariate spatial models:

$$\begin{aligned} Y_{di} | \theta_{di} &\sim P(E_{di} \theta_{di}), \\ \log(\theta_{di}) &= \beta_{d0} + \mathbf{X}_{di} \boldsymbol{\beta} + \psi_{di}, \\ \psi_1 &\sim \text{ICAR}(\mathbf{W}, \tau_{\psi_1}); \quad \psi_2 \sim \text{ICAR}(\mathbf{W}, \tau_{\psi_2}), \\ \mathbf{Y}_1 &\perp\!\!\!\perp \mathbf{Y}_2. \end{aligned}$$

$\mathcal{M}_3$  - Shared Component model without specific spatial term:

$$\begin{aligned} Y_{di} | \theta_{di} &\sim P(E_{di} \theta_{di}), \\ \log(\theta_{1i}) &= \beta_{10} + \mathbf{X}_{1i} \boldsymbol{\beta} + \delta \psi_i; \quad \log(\theta_{2i}) = \beta_{20} + \mathbf{X}_{2i} \boldsymbol{\beta} + \frac{1}{\delta} \psi_i, \\ \psi &\sim \text{ICAR}(\mathbf{W}, \tau_{\psi}). \end{aligned}$$

$\mathcal{M}_4$  - Shared Component model with specific spatial term:

$$\begin{aligned} Y_{id} | \theta_{di} &\sim P(E_{id} \theta_{id}), \\ \log(\theta_{i1}) &= \beta_{10} + \mathbf{X}_{i1} \boldsymbol{\beta} + \delta \psi_i + \phi_{1i}; \quad \log(\theta_{i2}) = \beta_{20} + \mathbf{X}_{i2} \boldsymbol{\beta} + \frac{1}{\delta} \psi_i + \phi_{2i}, \\ \psi &\sim \text{ICAR}(\mathbf{W}, \tau_{\psi}); \quad \phi_1 \sim \text{ICAR}(\mathbf{W}, \tau_{\phi_1}); \quad \phi_2 \sim \text{ICAR}(\mathbf{W}, \tau_{\phi_2}). \end{aligned}$$

In each case,  $\mathbf{W}$  is assigned as the original adjacency matrix or the matrices created with the RSCM approach. The set of priors is the same as in the simulation study in Chapter 3.2.

**Table 5** – Analysis of the incidence of lung and bronchus cancer for men and women in California (US). Results are presented as mean, standard deviation (SD) and the 95% credibility interval (ICr).

Model	Parameter	SCM						RSCM					
		Women			Men			Women			Men		
		Mean	SD	ICr	Mean	SD	ICr	Mean	SD	ICr	Mean	SD	ICr
$\mathcal{M}_1$	$\beta_0$	-1.15	0.14	[-1.44; -0.87]	-1.15	0.14	[-1.42; -0.88]						
	% Smokers	2.23	0.86	[0.55; 3.91]	2.91	0.81	[1.31; 4.50]						
	% 65 or over	6.73	0.60	[5.55; 7.90]	6.07	0.58	[4.92; 7.19]						
	WAIC	452.34			452.03								
$\mathcal{M}_2$	$\beta_0$	-0.57	0.26	[-1.06; -0.05]	-0.90	0.27	[-1.43; -0.37]	-1.04	0.20	[-1.44; -0.63]	-1.28	0.22	[-1.73; -0.85]
	% Smokers	1.57	1.35	[-1.11; 4.20]	3.55	1.43	[0.75; 6.42]	3.05	1.23	[0.65; 5.50]	4.81	1.37	[2.20; 7.59]
	% 65 or over	3.64	0.94	[1.73; 5.44]	3.92	0.93	[2.06; 5.73]	5.36	0.77	[3.82; 6.84]	5.26	0.79	[3.70; 6.79]
	WAIC	366.57			390.44			377.30			392.10		
$\mathcal{M}_3$	$\beta_0$	-0.60	0.26	[-1.09; -0.08]	-0.81	0.20	[-1.18; -0.42]	-1.14	0.21	[-1.56; -0.72]	-1.15	0.17	[-1.49; -0.81]
	% Smokers	2.03	1.36	[-0.65; 4.71]	2.84	1.05	[0.78; 4.91]	3.84	1.29	[1.33; 6.42]	4.00	1.05	[1.98; 6.11]
	% 65 or over	3.40	0.93	[1.53; 5.18]	4.01	0.78	[2.43; 5.49]	5.30	0.79	[3.73; 6.82]	5.16	0.68	[3.81; 6.47]
	$\delta$	1.26	0.12	[1.05; 1.51]				1.24	0.12	[1.03; 1.49]			
$\mathcal{M}_4$	WAIC	776.94			783.72			783.72			783.72		
	$\beta_0$	-0.57	0.26	[-1.07; -0.06]	-0.87	0.26	[-1.39; -0.36]	-1.07	0.21	[-1.48; -0.66]	-1.26	0.22	[-1.70; -0.85]
	% Smokers	1.70	1.36	[-1.00; 4.36]	3.44	1.38	[0.77; 6.25]	3.30	1.27	[0.84; 5.84]	4.76	1.32	[2.25; 7.47]
	% 65 or over	3.56	0.94	[1.66; 5.37]	3.84	0.91	[2.02; 5.60]	5.33	0.78	[3.77; 6.83]	5.20	0.77	[3.67; 6.70]
$\delta$	1.67	0.26	[1.21; 2.22]				1.62	0.26	[1.17; 2.18]				
WAIC	758.00			768.44			768.44			768.44			

In our application, the main aim is to model new cases of bronchus and lung cancer in California in 2016 (most recent year available in January 2020). SEER provides information for each subject being two of them the gender and the county code. Using this individual information, we are able to calculate, for each gender, the number of new cases of bronchus and lung cancer by county. These two quantities are our response variables and we would like to understand the characteristics that possibly affect the number of new cases of such disease and also whether the spatial pattern is shared for men and women or not. The covariates in this application are the percentage of adults that smoke every day or most days (% Smokers) and the percentage of the population aged 65 or over (% 65 or over). To compare the results we are using the WAIC criterion (WATANABE, 2010). Table 5 shows the estimates of each parameter for this set of models. Note that  $\mathcal{M}_1$  is agreeing with the literature, since smoking and age are well-known factors that increase the risk of lung and bronchus cancer for both genders. This way the areal information should also show the augmented incidence of this cancer. All coefficients are positive and the credibility interval does not contain the value zero, which is a clue of the importance of these covariates.

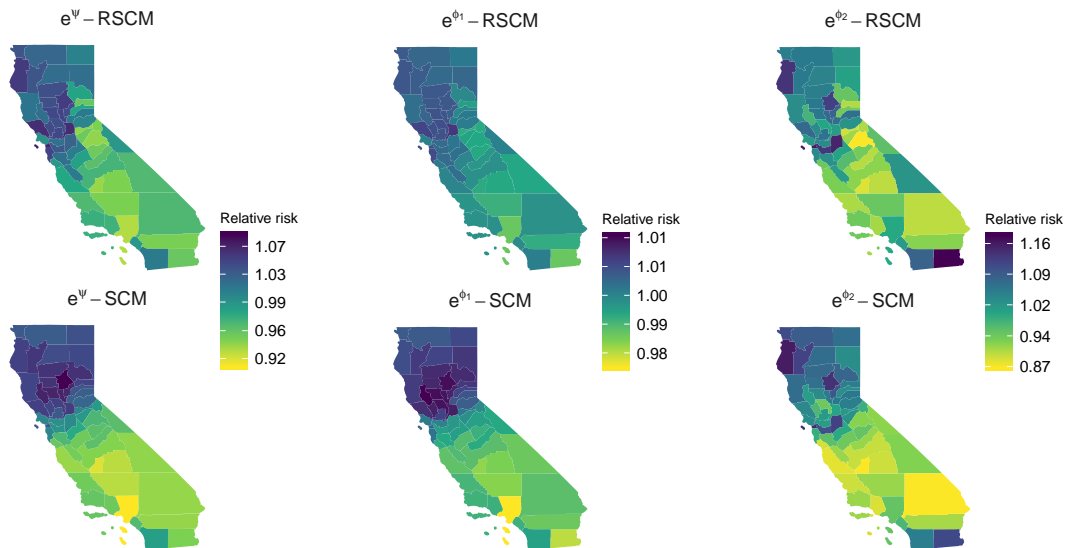
The difference between  $\mathcal{M}_1$  and  $\mathcal{M}_2$  is that  $\mathcal{M}_2$  also contains a spatial effect. We can observe, comparing  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , that the coefficients' standard deviation increased and also the point estimates changed for the SCM model. RSCM correction also changed the point estimate, however, the standard deviations are smaller than in the SCM case. Comparing the credibility intervals, we can notice that the covariate “% Smokers” became not important for females under the SCM model which is contradictory with the literature. RSCM remains pointing that this covariate is important to the model.

At this point, we can see that by adding the spatial effect to the univariate model causes a variance inflation on coefficients estimates, which is an indication of spatial confounding. Thus, when employing the RSCM approach, the model is less affected by the spatial confounding.

Models  $\mathcal{M}_3$  and  $\mathcal{M}_4$  are both multivariate with a shared component. It is possible to see that the  $\delta$  estimation is greater than 1 which means that men and women share a spatial pattern with different dependence levels. Also, we can observe that, because of the variance inflation, the credibility interval of “% Smokers” is including the value 0 for the SCM model and this behavior is not observed under the RSCM approach. This means that the conclusion of the importance of this covariate is changing by adding the spatial effects, as observed in model  $\mathcal{M}_2$ .

Comparing the WAIC of  $\mathcal{M}_3$  and  $\mathcal{M}_4$  we conclude that the model with specific disease components has a better fit when compared to that model without specific components. Also, in all cases, the SCM model has a better fit to the data than the RSCM model. However, because of the interpretability, we would prefer to select the RSCM model which is the one that does not change drastically the model’s conclusions.

Figure 10 shows the shared spatial effect and also the specific gender spatial effect for model  $\mathcal{M}_4$ . As expected we can see that the patterns are similar when we compare

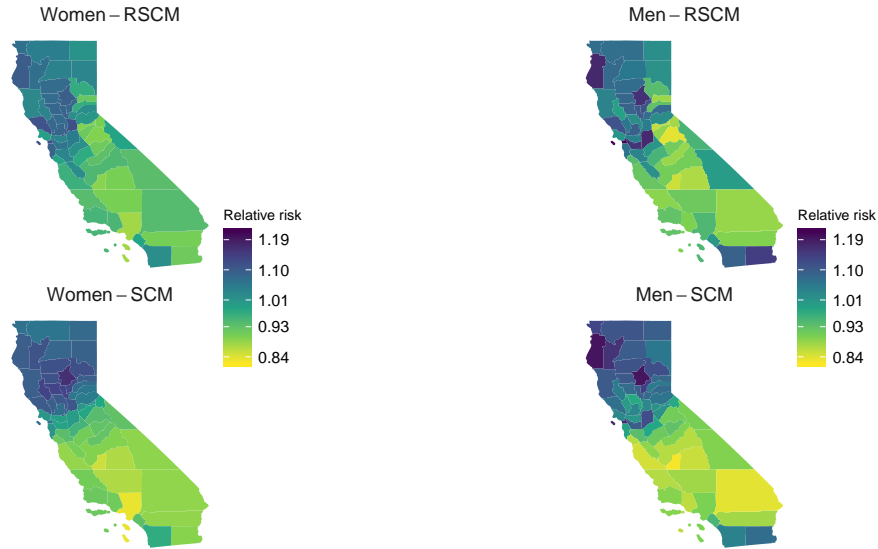


**Figure 10** – Shared and specific spatial patterns estimates for the incidence of lung and bronchus cancer in California (US).

SCM and RSCM approaches. The specific spatial patterns are similar for men and women. However, one can see that the risk scale differs for men and women. We can observe, relatively, stronger risk in the south for women when compared with the spatial effect for men. Both effects are higher in the north being more homogeneous for women.

Figure 11 show the aggregated spatial effects for women ( $\exp\{\delta\psi + \phi_1\}$ ) and for

men ( $\exp\{\frac{1}{\delta}\psi + \phi_2\}$ ). The combined effects are similar for men and women while the specific spatial patterns differ. This motivates the use of both shared and specific spatial components.



**Figure 11** – Aggregated spatial pattern estimates for the incidence of lung and bronchus cancer in California (US).

Briefly, the spatial confounding effect changed conclusions about one important covariate (% Smokers) in the model. The spatial models  $\mathcal{M}_2$ ,  $\mathcal{M}_3$  and  $\mathcal{M}_4$  introduced bias and variance inflation in such coefficient. Also, the conclusion under such models was not supported by the literature. Because of these characteristics, we decided to correct for possible spatial confounding.

Currently, the models capable to alleviate the spatial confounding in the literature do not cover the shared component model. Therefore, we proposed a method to alleviate the spatial confounding in this model family by creating several different spatial structures one for each spatial effect aiming to alleviate the spatial confounding effects. The methodology proved to be interesting and effective, alleviating the spatial confounding drawbacks and bringing practical sense to the models  $\mathcal{M}_2$ ,  $\mathcal{M}_3$  and  $\mathcal{M}_4$ .

The next part will introduce our second contribution providing a methodology to alleviate the spatial confounding in spatial frailty models. Therefore, we provide the method, a simulation study, and one application.

## Part 4

### Spatial confounding in spatial frailty models

## 4.1 Method

The likelihood of the spatial frailty model depends on the hazard function which is related to the baseline hazard function, covariates, and latent effects. As a consequence, this likelihood can be written according to Equation (4.1.1). Let  $h_{\theta}^0$  and  $H_{\theta}^0$  be the baseline hazard function and the cumulative baseline hazard function, respectively. Let  $\boldsymbol{\psi} = [\psi_1, \dots, \psi_n]^T$  be a vector of latent effects related to each location. Define  $\boldsymbol{\epsilon}$  as a vector with entries  $\epsilon_{ij}$   $i = 1, \dots, n; j = 1, \dots, n_i$  where  $\epsilon_{ij}$  is an unstructured latent effect related to the sample unit  $j$  at location  $i$ . The likelihood is

$$\begin{aligned} \mathcal{L}(t, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\epsilon}) &= \prod_{i=1}^n \prod_{j=1}^{n_i} \left[ [h_{\theta}(t_{ij}) S_{\theta}(t_{ij})]^{\Delta_D} S_{\theta}(t_{ij})^{\Delta_R} (1 - S_{\theta}(t_{ij}))^{\Delta_L} (S_{\theta}(t_{ij1}) - S_{\theta}(t_{ij2}))^{\Delta_K} \right], \\ h_{\theta}(t_{ij}) &= h_{\theta}^0(t_{ij}) \exp \{ \mathbf{X}_{ij} \boldsymbol{\beta} + \psi_i + \epsilon_{ij} \}, \\ S_{\theta}(t_{ij}) &= \exp \left\{ -H_{\theta}^0(t_{ij}) \exp \{ \mathbf{X}_{ij} \boldsymbol{\beta} + \psi_i + \epsilon_{ij} \} \right\}, \end{aligned} \quad (4.1.1)$$

and  $\boldsymbol{\Delta} = \{\Delta_D, \Delta_R, \Delta_L, \Delta_K\}$  are indicator functions of events, right-censored, left-censored and interval-censored sample units, respectively, and represent, for each individual, which term will contribute in the likelihood.

In this model there is more sample units than locations which implies in different supports for  $\mathbf{X}_{N \times p}$  and  $\boldsymbol{\psi}_{n \times 1}$ , where  $N = \sum_{i=1}^n n_i$ . As mentioned by Hanks et al. (2015), the projection-based approach is intuitive when the support of the observations is identical to spatial support, but we might be careful when this is not true as in the case of spatial frailty models. That said, in the conventional projection-based approach, the projection matrix is given by  $\mathbf{P}_{(N \times N)} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  and, therefore, it is not possible to make the projection of  $\boldsymbol{\psi}_{n \times 1}$  onto the orthogonal space of  $\mathbf{X}_{N \times p}$  directly. The simpler solution is to create a new vector of the same length as  $\mathbf{X}$  by repeating the spatial effects according to the areas where  $\mathbf{X}$  were collected. Define  $\boldsymbol{\Psi} = [\psi_1 \times \mathbf{1}_{n_1}, \dots, \psi_n \times \mathbf{1}_{n_n}]^T$  where  $\mathbf{1}_m$  is a length  $m$  row vector of ones. Thus,  $\boldsymbol{\Psi}$  is represented in Equation (4.1.2)

$$\boldsymbol{\psi}_{n \times 1} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix}; \quad \boldsymbol{\Psi}_{N \times 1} = \left. \begin{array}{c} \left[ \begin{array}{c} \psi_1 \\ \vdots \\ \psi_1 \end{array} \right] \\ \left[ \begin{array}{c} \vdots \\ \psi_n \\ \vdots \\ \psi_n \end{array} \right] \end{array} \right\} \begin{array}{l} n_1 \text{ times} \\ n_n \text{ times} \end{array} . \quad (4.1.2)$$

Then we can rewrite the hazard function in terms of the new vector.

$$h_{\theta}(t) = h_{\theta}^0(t) \exp \{ \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\Psi} + \boldsymbol{\epsilon} \}. \quad (4.1.3)$$

Given the vector  $\Psi$ , we can apply a projection-based approach and decompose the vector into  $\mathbf{P}\Psi$  and  $\mathbf{P}^\perp\Psi$  where  $\mathbf{P}^\perp = (\mathbf{I} - \mathbf{P})$

$$h_\theta(t) = h_\theta^0(t) \exp \left\{ \mathbf{X}\beta + \mathbf{P}\Psi + \mathbf{P}^\perp\Psi + \epsilon \right\}. \quad (4.1.4)$$

The duplicated information in Equation (4.1.4) is the vector  $\mathbf{P}\Psi$  and may promote the bias and variance inflation. To alleviate it, a convenient solution is to remove this quantity giving rise to the following model

$$h_\theta(t) = h_\theta^0(t) \exp \left\{ \mathbf{X}\beta_{rsf} + \mathbf{P}^\perp\Psi + \epsilon \right\}, \quad (4.1.5)$$

as  $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ , the expected value for the coefficients are given by  $\beta_{rsf} = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Psi$  where ‘‘rsf’’ means ‘‘restricted spatial frailty’’. However, this solution implies in a new limitation to the model. The spatial effect ( $\mathbf{P}^\perp\Psi$ ), free of spatial confounding, is a  $N \times 1$  vector which does not have a meaning as we just have  $n$  locations. Therefore, we propose a solution that summarises the information creating two vectors. The first vector contains the mean by locations of the orthogonal quantity  $\mathbf{P}^\perp\Psi$  and the second one contains the deviations from these means.

Define  $\psi_{rsf} = [\psi_{rsf_1}, \dots, \psi_{rsf_n}]^T$  the vector containing the  $n$  means of  $\mathbf{P}^\perp\Psi$ , one for each region  $i = 1, \dots, n$ , and  $\tilde{\psi} = [\tilde{\psi}_{11}, \dots, \tilde{\psi}_{nn}]^T$ , where  $\tilde{\psi}_{ij}$  represents the individual distance of each element of  $\mathbf{P}^\perp\Psi$  to its respective mean  $\psi_{rsf_i}$ . In this case,  $\Psi_{rsf} = [\psi_{rsf_1} \times \mathbf{1}_{n_1}, \dots, \psi_{rsf_n} \times \mathbf{1}_{n_n}]^T$  is a vector of remaining mean effects of each location and  $\tilde{\psi}$  represents, for each sample unit, an individual distance from the mean as in Equation (4.1.6). In this case, both  $\Psi_{rsf}$  and  $\tilde{\psi}$  are  $N \times 1$  vectors

$$\Psi_{rsf} = \left[ \begin{array}{c} \psi_{rsf_1} \\ \vdots \\ \psi_{rsf_1} \\ \vdots \\ \psi_{rsf_n} \\ \vdots \\ \psi_{rsf_n} \end{array} \right] \left. \begin{array}{l} \left. \vphantom{\begin{array}{c} \psi_{rsf_1} \\ \vdots \\ \psi_{rsf_1} \\ \vdots \\ \psi_{rsf_n} \\ \vdots \\ \psi_{rsf_n} \end{array}} \right\} n_1 \text{ times} \\ \left. \vphantom{\begin{array}{c} \psi_{rsf_1} \\ \vdots \\ \psi_{rsf_1} \\ \vdots \\ \psi_{rsf_n} \\ \vdots \\ \psi_{rsf_n} \end{array}} \right\} n_n \text{ times} \end{array} \right\} ; \quad \tilde{\psi} = \left[ \begin{array}{c} \tilde{\psi}_{11} \\ \vdots \\ \tilde{\psi}_{11} \\ \vdots \\ \tilde{\psi}_{nn} \end{array} \right], \quad (4.1.6)$$

and then we can rewrite the model as in Equation (4.1.7)

$$h_\theta(t) = h_\theta^0(t) \exp \left\{ \mathbf{X}\beta_{rsf} + \Psi_{rsf} + \tilde{\psi} + \epsilon \right\}. \quad (4.1.7)$$

Once  $\tilde{\psi}$  is a vector of the same length of  $\epsilon$ , it is not possible to estimate both of them but just the sum. Let's call  $\tilde{\psi} + \epsilon$  as  $\epsilon_{rsf}$  and finally our final model is given by Equation (4.1.8).

$$h_\theta(t) = h_\theta^0(t) \exp \left\{ \mathbf{X}\beta_{rsf} + \Psi_{rsf} + \epsilon_{rsf} \right\}. \quad (4.1.8)$$



Our main objective is to fit the restricted spatial model. However, we would like to have estimates of the unrestricted model as well as the restricted model estimates. Therefore, we need to find equivalences between the restricted quantities and the unrestricted ones. With this equivalence, it is possible to have samples from both models. Equation (4.1.9) presents this equivalence.

$$\begin{aligned}
h_{\theta}(t) &= h_{\theta}^0(t) \exp \{ \mathbf{X} \boldsymbol{\beta}_{rsf} + \boldsymbol{\Psi}_{rsf} + \boldsymbol{\epsilon}_{rsf} \} \\
&= h_{\theta}^0(t) \exp \{ \mathbf{X} \boldsymbol{\beta}_{rsf} + \boldsymbol{\Psi}_{rsf} + \tilde{\boldsymbol{\psi}} + \boldsymbol{\epsilon}_{sf} \} \\
&= h_{\theta}^0(t) \exp \{ \mathbf{X} \boldsymbol{\beta}_{rsf} + \mathbf{P}^{\perp} \boldsymbol{\Psi}_{sf} + \boldsymbol{\epsilon}_{sf} \} \\
&= h_{\theta}^0(t) \exp \{ \mathbf{X} (\boldsymbol{\beta}_{rsf} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Psi}_{sf}) + \boldsymbol{\Psi}_{sf} + \boldsymbol{\epsilon}_{sf} \} \\
&= h_{\theta}^0(t) \exp \{ \mathbf{X} \boldsymbol{\beta}_{sf} + \boldsymbol{\Psi}_{sf} + \boldsymbol{\epsilon}_{sf} \},
\end{aligned} \tag{4.1.9}$$

where “sf” means “spatial frailty” and represent the estimates of the conventional spatial method and “rsf” means “restricted spatial frailty” and represents the model referred in Equation (4.1.8).

Given the unrestricted model, we can calculate the restricted quantities since  $\boldsymbol{\beta}_{rsf} = \boldsymbol{\beta}_{sf} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Psi}_{sf}$ ,  $\mathbf{P}^{\perp} \boldsymbol{\Psi}_{sf} = \boldsymbol{\Psi}_{rsf} + \tilde{\boldsymbol{\psi}}$  and  $\boldsymbol{\epsilon}_{rsf} = \boldsymbol{\epsilon}_{sf} + \tilde{\boldsymbol{\psi}}$ . With these equivalences it is possible to have estimates of all parameters of the restricted model. The general formulation in Equation (4.1.9) shows how to obtain the restricted models estimates for the proportional hazard family including the Cox model (when  $h_{\theta}^0(t)$  is not defined). In other words, we just need a sample from the unrestricted model to get estimates from both unrestricted and restricted models. These results are applied for the entire family of proportional hazards models. It is important to notice that even if we fit a model without the  $\boldsymbol{\epsilon}$  (independent) term, under the restricted model, the component  $\boldsymbol{\epsilon}_{rsf}$  will appear.

### 4.1.1 Reduction operator

Although enlarging the spatial effect vector is a straightforward solution in Equation (4.1.4), the projection approach requires, for each element of a posterior sample, the calculations:

- $\boldsymbol{\beta}_{rsf} = \boldsymbol{\beta}_{sf} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Psi}_{sf}$ ,
- $\mathbf{P}^{\perp} \boldsymbol{\Psi}_{sf} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \boldsymbol{\Psi}_{sf} = \boldsymbol{\Psi}_{rsf} + \tilde{\boldsymbol{\psi}}$ ,

which requires products of matrices with lengths equal to the sample size ( $N$ ).

It is not unusual to work with data sets in which, for each area, several individuals are observed. As this number increases, the total sample size  $N$  also increases, but not the

number of areas,  $n$ , that remains fixed. Said that, the computation of the restricted model increases as  $N$  increases. However, because  $\Psi_{sf}$  is constant by area, it is possible to get the same desired results but computing it with a reduced version of  $\mathbf{P}$  and  $\mathbf{P}^\perp$  matrices in which the new matrices are  $(n \times n)$ -dimensional instead of  $(N \times N)$ -dimensional.

Let's define an operator that will help us to achieve the computational improvement. Let  $\mathbf{X}_{N \times p}$  be a matrix with entries  $X_{ijk}$  for an index  $i$ , an element  $j$  and column  $k$ , and  $\mathbf{G}_{N \times 1}$  is a vector of indices indicating, for each row of  $\mathbf{X}_{N \times p}$ , an index  $i$  in a set of indices starting from 1 until  $n$  ( $n \ll N$ ). Then the reduction operator  $\textcircled{\mathbb{T}}$  is defined by:

$$\mathbf{X}_{N \times p} \textcircled{\mathbb{T}} \mathbf{G} = \mathbf{x}_{n \times p}, \quad (4.1.10)$$

in which  $x_{ik} = \sum_{j=1}^{n_i} X_{ijk}$ , and  $n_i$  is the number of elements related with index  $i$ . This operator has several properties that allow us to simplify the computational procedure. Let  $c$  be a constant,  $\mathbf{r}_{n \times 1}$  is a column vector,  $\mathbf{R} = [r_{G_1}, \dots, r_{G_N}]^T$  is a  $N \times 1$  vector with repeated entries for each index of  $\mathbf{G}$  (constant by indices),  $\mathbf{P}_{p \times p}$  is a squared matrix and,  $\mathbf{Q}_{m \times p}$  is a matrix. Therefore, the following properties are true:

1.  $(\mathbf{X}_1 + \mathbf{X}_2) \textcircled{\mathbb{T}} \mathbf{G} = (\mathbf{X}_1 \textcircled{\mathbb{T}} \mathbf{G}) + (\mathbf{X}_2 \textcircled{\mathbb{T}} \mathbf{G})$ ,
2.  $(c\mathbf{X}) \textcircled{\mathbb{T}} \mathbf{G} = c(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})$ ,
3.  $\mathbf{X}^T \mathbf{R} = (\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})^T \mathbf{r}$ ,
4.  $(\mathbf{Q}\mathbf{X}^T) \textcircled{\mathbb{T}} \mathbf{G}^T = \mathbf{Q}(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})^T$ ,
5.  $(\mathbf{X}\mathbf{P}\mathbf{X}^T) \textcircled{\mathbb{T}} \mathbf{G} = (\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})\mathbf{P}\mathbf{X}^T$ ,
6.  $((\mathbf{X}\mathbf{P}\mathbf{X}^T) \textcircled{\mathbb{T}} \mathbf{G}) \textcircled{\mathbb{T}} \mathbf{G}^T = (\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})\mathbf{P}(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})^T$ ,
7.  $(\mathbf{X}\mathbf{P}\mathbf{X}^T \mathbf{R}) \textcircled{\mathbb{T}} \mathbf{G} = (\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})\mathbf{P}(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})^T \mathbf{r}$ .

The proofs of these properties are in Appendix B. Using the reduction operator it is possible to compute  $\beta_{rsf}$  efficiently (by property 4):

$$\begin{aligned} \beta_{rsf} &= \beta_{sf} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Psi_{sf} \\ &= \beta_{sf} + (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})^T \psi_{sf}, \end{aligned} \quad (4.1.11)$$

that is a product on a smaller dimension because  $(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})^T$  is a  $p \times n$  matrix.

Also, to compute  $\psi_{rsf}$ , using properties 1, 3 and 4, and defining  $\mathbf{N}_{N \times N}$  as a diagonal matrix with  $N_{ii} = n_{G_i}$  being the number of elements in each area, and  $\mathbf{n}_{n \times n}$  being a

diagonal matrix with  $n_{jj} = n_{G_j}$  it is the same as

$$\begin{aligned}
\psi_{rsf} &= \mathbf{N}^{-1}(\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \Psi_{sf} \oplus \mathbf{G} & (4.1.12) \\
&= (\mathbf{N}^{-1} \Psi_{sf} - \mathbf{N}^{-1} \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Psi_{sf}) \oplus \mathbf{G} \\
&= (\mathbf{N}^{-1} \Psi_{sf}) \oplus \mathbf{G} - \mathbf{N}^{-1}(\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Psi_{sf}) \oplus \mathbf{G} \\
&= (\mathbf{N}^{-1} \oplus \mathbf{G})^T \psi_{sf} \oplus \mathbf{G} - \mathbf{N}^{-1} \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X} \oplus \mathbf{G})^T \psi_{sf} \oplus \mathbf{G} \\
&= \psi_{sf} - (\mathbf{N}^{-1} \mathbf{X} \oplus \mathbf{G})(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X} \oplus \mathbf{G})^T \psi_{sf} \\
&= (\mathbf{I}_n - \mathbf{n}^{-1}(\mathbf{X} \oplus \mathbf{G})(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \oplus \mathbf{G})) \psi_{sf}.
\end{aligned}$$

Note that  $(\mathbf{N}^{-1} \mathbf{P}^\perp \Psi_{sf}) \oplus \mathbf{G} = \psi_{rsf}$ . Then, for both  $\beta_{rsf}$  and  $\psi_{rsf}$  it is possible to calculate their values using small length matrices which is computationally attractive.

#### 4.1.1.1 HH model

A similar approach can be applied for the before-mentioned HH model (HUGHES; HARAN, 2013). To apply their ideas we can first create the Moran operator replacing  $\mathbf{P}^\perp$  in the Equation (1.2.8) by  $((\mathbf{N}^{-1} \mathbf{P}^\perp \oplus \mathbf{G}^T) \oplus \mathbf{G}) \mathbf{W} ((\mathbf{N}^{-1} \mathbf{P}^\perp \oplus \mathbf{G}^T) \oplus \mathbf{G})$  in the original formulation. After that, one can perform the spectral decomposition and take the relevant eigenvalues to create a low-rank model as mentioned in Section 1.2.2.2. It is important to notice that, in this case, it is not possible to find equivalence between restricted and unrestricted models and then it is necessary to fit the restricted model directly.

#### 4.1.1.2 SPOCK model

Similarly, to apply the SPOCK method in models where the support of observations is not identical to spatial support, caution is needed. In this case, one solution is to get the new set of centroids by projecting the original set onto the orthogonal space of covariates, after the application of the reduction operator.

Let's call by  $\mathbf{W}$  the adjacency matrix of Section 1.2.1 linked to the set of original centroids  $\mathbf{c}_i = \{c_{1i}, c_{2i}\} \forall i \in [1, \dots, n]$ . The new set of centroids  $\mathbf{c}^*$  by projecting  $\mathbf{c}$  onto the orthogonal space of  $\mathbf{X}$  is given by

$$\mathbf{c}^* = ((\mathbf{N}^{-1} \mathbf{P}^\perp \oplus \mathbf{G}^T) \oplus \mathbf{G}) \mathbf{c}.$$

Then, as showed in Section 1.2.2.4, one can choose its preferable software to implement the solution.

## 4.2 Simulation

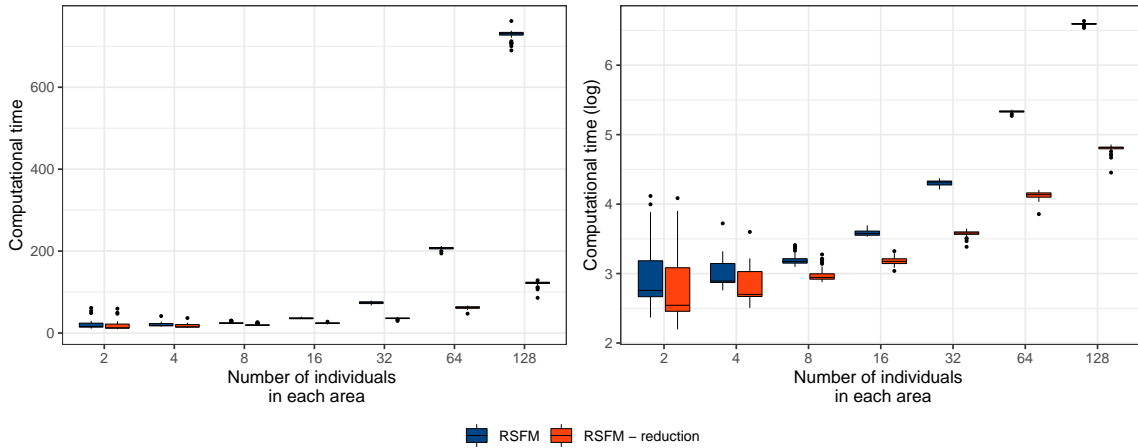
This study is divided into two sections. First, the computational improvement will be presented by a simulation study that shows the reduction operator efficiency. Next, the capacity to recover the model parameters and the efficiency of the proposed correction will be shown. The methodology presented here does not depend on the method used to get samples from the unrestricted model. This way, for computational benefits we are using the INLA method to fit the unrestricted model. Also, we are using R-INLA to generate posterior samples of the parameters involved. The `inla.posterior.sample` allows us to have a sample from the approximated posterior distribution. The hyperparameters are sampled from the grid used in the numerical integration and the latent field is sampled from the Gaussian approximation conditioned on the hyperparameters. Based on a posterior sample of the unrestricted model and using Equation (4.1.9) we obtained the posterior samples from the restricted model. Also, the method does not depend on the parametric model chosen for the baseline hazard. In this case, we choose the widely applied Weibull proportional hazard model.

### 4.2.1 Computational improvement

To show the computational improvement using the reduction operator, we performed a simulation study. The time spent to get samples from the restricted model using the methodology described in Section 4.1 and the time spent applying the reduction operator were recorded.

The data were generated from the Weibull proportional hazard model for a spatial structure (polygons) containing 92 areas. We vary the number of individuals in each area in the following grid: 2, 4, 8, 16, 32, 64 and 128. Therefore, the total sample size  $N$  is  $92 \times 2 = 184$  in the first scenario and  $92 \times 128 = 11,776$  in the last one. For each case, a posterior sample of size 5000 was taken and the correction was made based on it. This is a two-step technique, where first, we get samples from the unrestricted model and then, a posteriori, we get samples from the restricted model. Thus, we are able to record the time to fit the model and also the time to perform the correction. It is interesting to notice that, in both cases, the time spent to fit the unrestricted model is the same and therefore we are not reporting it. Also, using the reduction operator, the correction step is always involving the same length matrix while the matrix using the methodology without the reduction becomes larger at each step.

Figure 12 shows the computational cost for applying these two approaches, varying according to the number of subjects in each area. As one can see, the computational cost



**Figure 12** – Time spent to fit the Weibull proportional hazard model with and without the reduction operator. Right: Original scale in seconds; Left: Logarithmic scale.

is increasing as  $N$  increases for the model without the reduction step. The increment in time for the pure model increases drastically because for each posterior sample we must calculate  $\mathbf{P}^\perp \Psi_{sf}$ . This is a product of a  $N \times N$  matrix by a  $N \times 1$  vector (this product is repeated 5,000 times). Instead, the model with the reduction operator calculates  $((\mathbf{P}^\perp \textcircled{R} \mathbf{G}) \textcircled{R} \mathbf{G}) \psi_{sf}$  which is a product of a  $n \times n$  matrix by a  $n \times 1$  vector.

The time spent to calculate  $(\mathbf{P}^\perp \textcircled{R} \mathbf{G}) \textcircled{R} \mathbf{G}$  also increases as  $N$  increases, but the calculation occurs just once. Also, it is a straightforward calculation that is not strongly affected by the sample size. Thus, since the computational cost to calculate the reduced model is preferable and the results are strictly the same, we will use the reduction operator for the rest of the work.

## 4.2.2 Confounding alleviation

To evaluate the model ability to estimate the parameters, we conducted a simulation study. Data were generated from the Weibull proportional hazard model

$$\begin{aligned} h(t_i) &= \alpha t_i^{\alpha-1} \exp \{ \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_i \}, \\ \psi &\sim \text{ICAR}(\mathbf{W}, \tau_\psi), \end{aligned} \quad (4.2.1)$$

where  $\alpha = 1.2$ ,  $\beta_0 = 0$ ,  $\beta_1 = -0.3$ ,  $\beta_2 = 0.3$  and  $\tau_\psi = 0.75$ . To evaluate the performance in terms of recovering the parameters in this model, we have 4 censoring levels: 0%, 25%, 50% and 75%.

We generated 1,000 datasets under each setup and 2 scenarios: 1)  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are random variables and therefore no spatial confounding is expected; 2)  $\mathbf{X}_1$  is a random variable but  $\mathbf{X}_2$  is the set of centroids' latitudes of each county. The set of weakly informative priors was taken as follow

$$\alpha \sim \Gamma(0.001, 0.001),$$

$$\beta_j \sim \text{Normal}(0, 0.001), \quad j = 0, 1, 2,$$

$$\tau_\psi \sim \Gamma(0.5, 0.0005).$$

Table 6 presents the mean of the estimated values (Mean), the mean of the standard deviations (SD), the coverage rate for a nominal rate of 95% (Cov) and the mean squared error (MSE) for each scenario. It is possible to observe that, without spatial

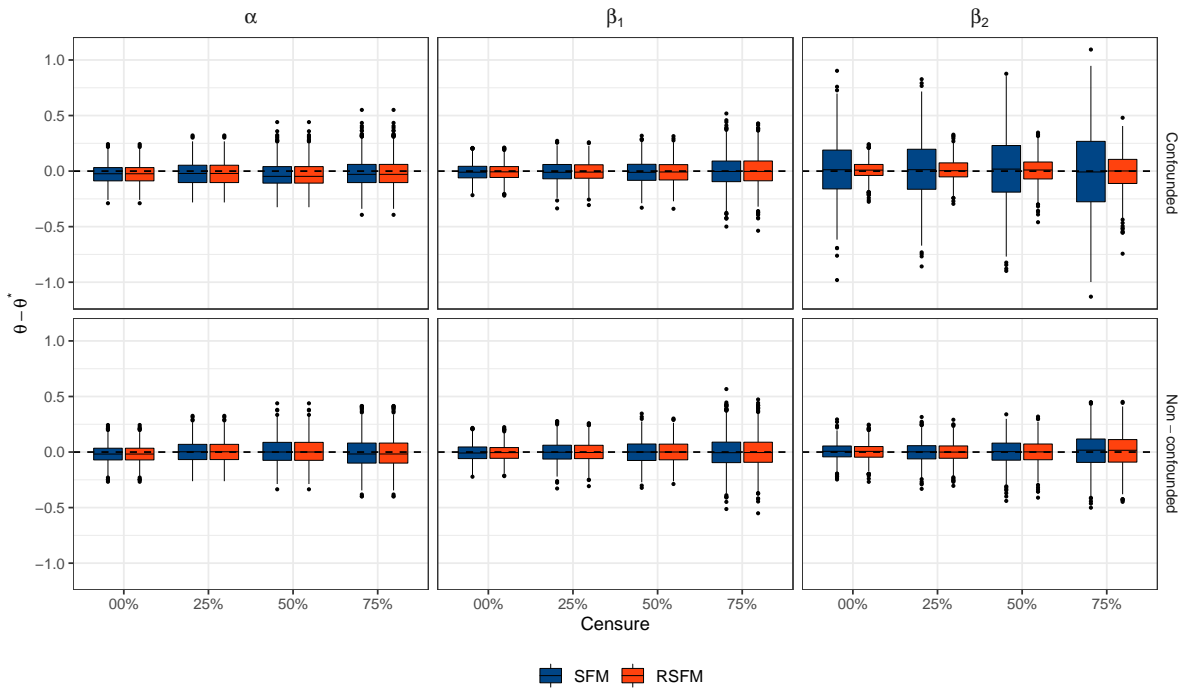
**Table 6** – Simulation results the spatial frailty model experiment. The results are shown by mean, standard deviation (SD), coverage rate for a nominal rate of 95 % (Cov) and mean square error (MSE).

Censure	Parameter	Without Spatial Confounding						With Spatial Confounding					
		SFM			RSFM			SFM			RSFM		
		Mean (SD)	COV	MSE	Mean (SD)	COV	MSE	Mean (SD)	COV	MSE	Mean (SD)	COV	MSE
00.00%	$\alpha$	1.18 (0.07)	83.20%	0.0071	1.18 (0.07)	83.20%	0.0071	1.17 (0.07)	77.80%	0.0087	1.17 (0.07)	77.80%	0.0087
	$\beta_1$	0.29 (0.08)	93.90%	0.0060	0.29 (0.07)	94.80%	0.0053	0.29 (0.08)	93.60%	0.0061	0.29 (0.07)	94.90%	0.0053
	$\beta_2$	-0.29 (0.08)	93.70%	0.0062	-0.30 (0.07)	93.80%	0.0054	-0.28 (0.20)	83.40%	0.0609	-0.29 (0.07)	93.00%	0.0062
25.00%	$\alpha$	1.20 (0.08)	80.40%	0.0097	1.20 (0.08)	80.40%	0.0097	1.18 (0.08)	69.10%	0.0118	1.18 (0.08)	69.10%	0.0118
	$\beta_1$	0.30 (0.09)	93.70%	0.0082	0.30 (0.08)	94.40%	0.0073	0.29 (0.09)	92.80%	0.0083	0.29 (0.08)	94.40%	0.0072
	$\beta_2$	-0.30 (0.09)	92.80%	0.0083	-0.30 (0.08)	93.60%	0.0074	-0.28 (0.19)	77.10%	0.0723	-0.29 (0.09)	93.60%	0.0085
50.00%	$\alpha$	1.21 (0.09)	80.00%	0.0137	1.21 (0.09)	80.00%	0.0137	1.17 (0.08)	72.80%	0.0131	1.17 (0.08)	72.80%	0.0131
	$\beta_1$	0.30 (0.10)	94.30%	0.0116	0.30 (0.10)	93.60%	0.0106	0.29 (0.10)	93.60%	0.0114	0.29 (0.10)	95.10%	0.0104
	$\beta_2$	-0.30 (0.11)	93.20%	0.0129	-0.30 (0.10)	94.10%	0.0112	-0.28 (0.17)	63.80%	0.0936	-0.28 (0.11)	93.10%	0.0131
75.00%	$\alpha$	1.20 (0.11)	82.30%	0.0185	1.20 (0.11)	82.30%	0.0185	1.18 (0.11)	83.30%	0.0158	1.18 (0.11)	83.30%	0.0158
	$\beta_1$	0.30 (0.14)	92.20%	0.0223	0.30 (0.14)	93.30%	0.0211	0.30 (0.14)	93.20%	0.0210	0.30 (0.14)	94.30%	0.0194
	$\beta_2$	-0.29 (0.14)	93.10%	0.0236	-0.29 (0.14)	93.80%	0.0215	-0.30 (0.17)	56.30%	0.1418	-0.30 (0.15)	92.80%	0.0264

confounding, the SFM (Spatial Frailty Model) and the RSFM (Restricted Spatial Frailty Model) approaches present similar values for mean, coverage and mean squared error. Under spatial confounding, it is possible to see that the point estimates (mean) are accurate for all parameters and both models. However, the standard deviation of  $\beta_2$  is, on average, greater for the SFM model than for the RSFM model. Also, it is possible to observe that the MSE of  $\beta_2$  is greater for the SFM than for the RSFM. The coverage rate seems adequate for both models except for the parameter  $\alpha$ . Since this inconsistency happens also for the model without spatial confounding, it is out of our scope to investigate this phenomenon.

Another interesting conclusion is that as much the censor level increases the standard deviations also increase in all cases. It is showing that, in those models with bigger censoring rates, the estimates are less accurate (as expected).

The projection-based approach aims to estimate  $\beta^* = \beta_{sf} + P^\perp \Psi$ . In this case, we reported Figure 13 for  $(\theta - \theta^*)$  where  $\theta = \{\alpha, \beta_1, \beta_2\}$ . We expect all the values to be around 0, which means that the estimate is not biased. We can see, for  $\alpha$  and  $\beta_1$ , that the estimates are similar for both SFM and RSFM models. However, for  $\beta_2$  the behavior changes for the model with and without confounding. The model without spatial confounding, as expected, behaves in the same way for  $\beta_1$ . For the model with spatial confounding, we can observe that the  $(\theta - \theta^*)$  tends to be around 0 for the SFM and also



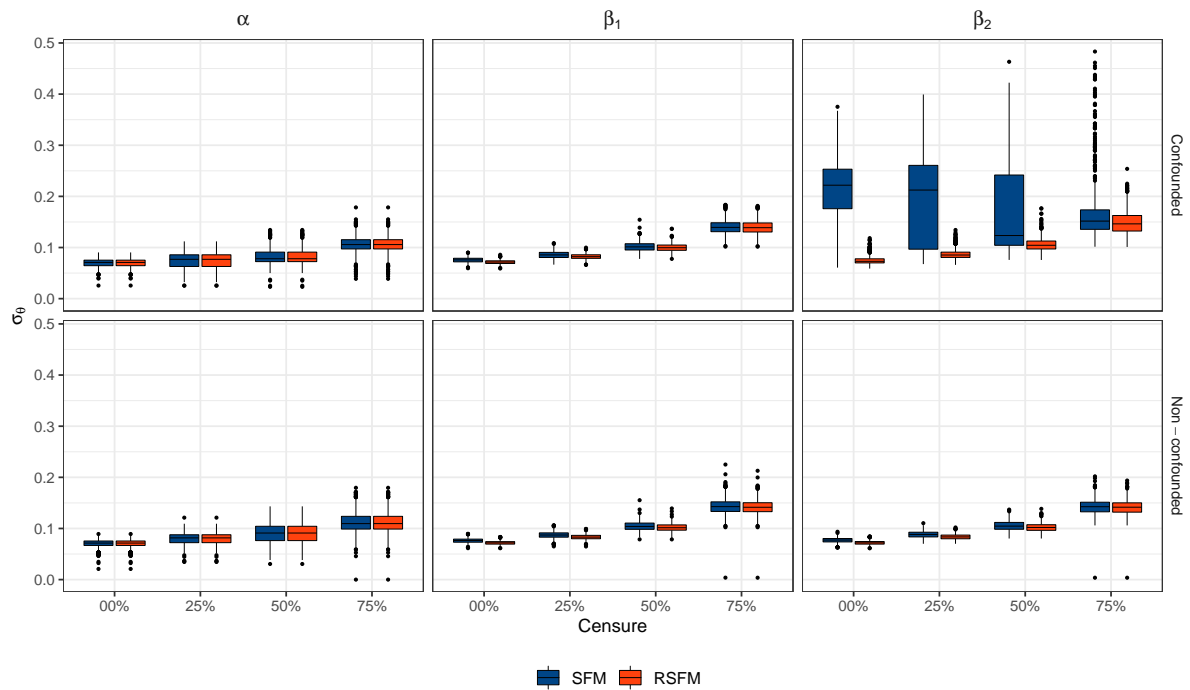
**Figure 13** – Boxplot of  $(\theta - \theta^*)$  for  $\theta = \{\alpha, \beta_1, \beta_2\}$  in the spatial frailty model. Dashed line represents the value 0.

for the RSFM. Although they are centered at 0, the dispersion of SFM seems to be bigger than the dispersion of the RSFM model which, in this case, suggests bias in the estimates.

From the perspective of the level of censorship, we can see a smooth increment in the coefficients' variance for all cases. This result is explained by the fact that, with the increment of the censoring rate, we have less information about the responses.

Figures 14, 15, show for SFM and RSFM, the standard deviations and the SVIF (defined in Section 1.2.2.5) comparing with the non-spatial model. An SVIF equal to 1 indicates that the variances of both models are the same. However, because the spatial model is more complex it is expected an increment in the variance. We can observe in Figure 14 that the standard deviations are similar in all cases except for  $\beta_2$  in the scenario with spatial confounding. In this case, we can see that the higher the level of censorship, the more similar is the standard deviation. We can also notice that even for the model without spatial confounding, the SVIF is bigger for SFM than for the RSFM, which indicates that the RSFM approach alleviates the variance inflation even in the cases we are not expecting it. However, comparing with the dashed line, we note that the variance is almost always increasing for both SFM and RSFM. Also, we can observe a downward trend in the SVIF for both models when the level of censorship. It means that the efficiency of the correction decreases with the increment of the censored individuals, which is in agreement with Figure 13.

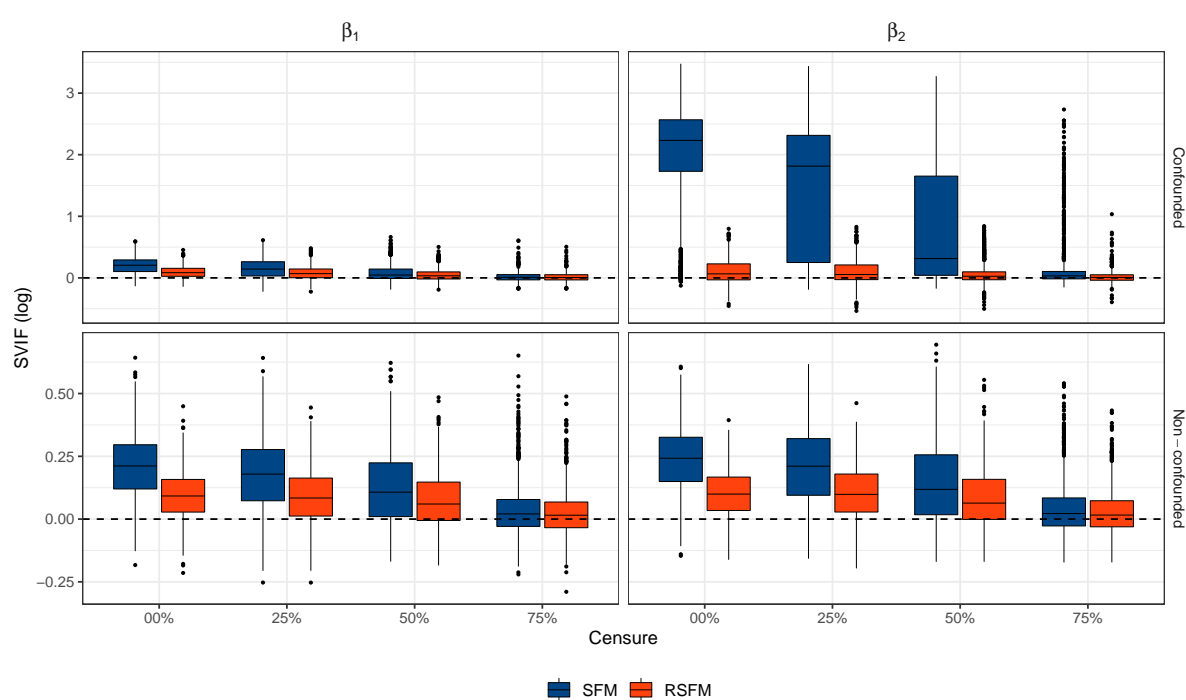
The behavior for  $\beta_1$  and  $\beta_2$  in the model without spatial confounding are similar and



**Figure 14** – Boxplot of  $\sigma_\theta$  for  $\theta = \{\beta_1, \beta_2\}$  for the SFM and RSFM where  $\sigma_\theta$  represents the standard deviation of  $\theta$ .

this is also true for the behavior of  $\beta_1$  under spatial confounding. However, the parameter  $\beta_2$  under spatial confounding presents huge inflation of variances for the model without correction. In some cases, we experienced a variance  $\exp\{3\} \approx 20$  times bigger. In these cases, the restricted model behaves well and it keeps the variance stable.





**Figure 15** – Boxplot of the SVIF (log scale) between spatial models (SFM and RSFM) and the baseline model (Weibull proportional hazard model). Dashed line marks the value 0, which in the log scale represent the equality of variances.

## 4.3 Time until death by lung and bronchus cancer in California

To fit the model, we use a right censoring scheme with the Weibull proportional hazard model. Our baseline model is the Non-spatial (NS) model given by the Weibull proportional hazard model and five covariates: 1) gender; 2) race; 3) disease stage; 4) age at diagnosis; 5) the percentage of people who smoke every day or most days (areal level). The spatial frailty model (SFM) also includes the ICAR spatial term, and the restricted spatial frailty model (RSFM) alleviates the spatial frailty model for possible spatial confounding.

In Table 7,  $\alpha$  is the shape parameter of the Weibull distribution and the estimate was almost the same in the NS and SFM models (RSFM estimate is the same of the SFM for hyperparameters). The parameter  $\tau_w$  represents the precision for the ICAR model. The other parameters are related with the covariates in the modeling.

From the epidemiological point of view, the NS model reflects the theory that patients in a more advanced stage of the disease have a higher risk of death (In situ < Localized < Regional < Distant). Also, males have a higher risk when compared to females. Same way, black people have a higher risk when compared with non-black people. Also, the older the individual the greater is the risk. The coefficient for percentage of smokers in the county indicates an increment in the risk of death due to lung and bronchus cancer. This covariate is our best guess about individual tobacco consumption.

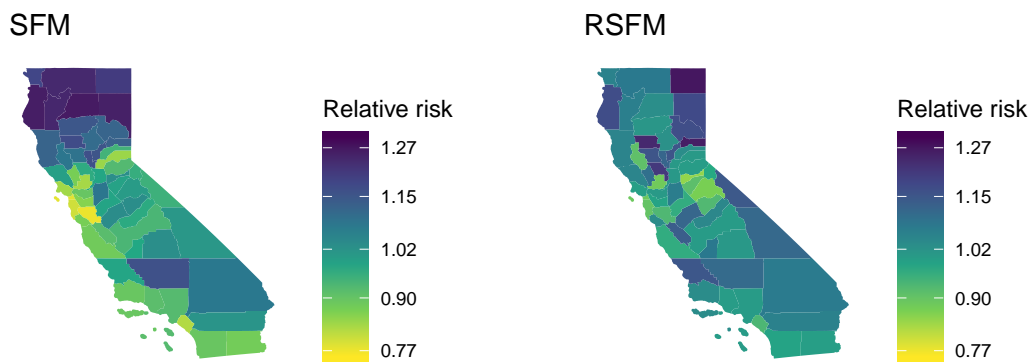
When we compare the results from the NS model with those of the SFM, one can

**Table 7** – Time until death by lung and bronchus cancer in California (US). Results are presented as mean, standard deviation (SD) and the 95 %credibility interval (ICr).

Parameter	NS		SFM		RSFM	
	Mean (SD)	ICr	Mean (SD)	ICr	Mean (SD)	ICr
$\alpha$	0.85 (0.0032)	(0.85; 0.86)	0.86 (0.0033)	(0.85; 0.86)	0.86 (0.0033)	(0.85; 0.86)
$\tau_w$			22.75 (6.8306)	(10.99; 36.49)	22.75 (6.8306)	(10.99; 36.49)
$\beta_0$	-4.07 (0.2048)	(-4.44; -3.65)	-3.58 (0.2091)	(-4.02; -3.20)	-4.06 (0.2001)	(-4.47; -3.69)
Gender						
Female	ref.		ref.		ref.	
Male	0.19 (0.0099)	(0.17; 0.21)	0.19 (0.0098)	(0.18; 0.21)	0.19 (0.0098)	(0.18; 0.21)
Race						
Non-black	ref.		ref.		ref.	
Black	0.16 (0.0177)	(0.13; 0.20)	0.17 (0.0176)	(0.13; 0.20)	0.17 (0.0175)	(0.13; 0.20)
Cancer stage						
In situ						
Localized	1.51 (0.1985)	(1.12; 1.87)	1.51 (0.1940)	(1.13; 1.90)	1.51 (0.1940)	(1.13; 1.90)
Regional	2.60 (0.1984)	(2.20; 2.96)	2.60 (0.1935)	(2.25; 3.01)	2.60 (0.1935)	(2.25; 3.02)
Distant	3.73 (0.1985)	(3.35; 4.11)	3.74 (0.1936)	(3.37; 4.14)	3.74 (0.1936)	(3.37; 4.14)
Age at diagnosis	0.02 (0.0005)	(0.02; 0.02)	0.02 (0.0005)	(0.02; 0.02)	0.02 (0.0005)	(0.02; 0.02)
% Smokers	2.13 (0.1821)	(1.79; 2.49)	-0.90 (0.3848)	(-1.68; -0.21)	2.13 (0.1819)	(1.79; 2.51)

notice that for gender, race, stage of the disease and age at diagnosis, the results are similar with small differences in the estimates. However, for the coefficient of the percentage of smokers, the point estimate changes drastically and there is variance inflation (variance is about 5 times greater for the SFM). Another important point in the SFM model is that the credibility interval changes drastically pointing that the percentage of smokers is a protective factor for cancer death. The restricted spatial frailty model (RSFM) was applied and we can notice that it returns similar estimates to those from the NS model, as expected. The credibility interval now is pointing that the higher the percentage of smokers, higher is the risk for cancer death.

Figure 16 shows the spatial effect  $\exp\{\psi\}$  for the SFM and RSFM. We can see



**Figure 16** – Spatial risk effects for death by lung and bronchus cancer in California (US).

that the patterns are smoother for the RSFM case. However, the pattern remains similar to the SFM model being higher in the north of the state, less intense in the center and again high in the south. This result might be useful to create new policies or new health care centers for lung and bronchus cancer in California.

We can conclude that the employment of the proposed restricted model is important in several ways. The first advantage is that the model conclusions retains the interpretability of the baseline model, keeping important conclusions about the model's covariates. Secondly, the computational improvement provided by the reduction operator appears as an important feature since it performs better than the pure restricted model since the reduction operator induces a not time-consuming model. Also, a user can apply its preferable software to get posterior samples from the restricted model and, a posteriori, correct for possible spatial confounding which gives more freedom. Another point is that under spatial confounding, the variances of coefficients are not inflated as it is for the conventional model. Finally, the spatial pattern is similar when compared with the unrestricted model which shows that the correction does not disorder the spatial patterns.

The next part will introduce our package named **RASCO**. The **RASCO** implements our contributions for shared component models and spatial frailty models as well as some approaches to alleviate the spatial confounding in GLMM.

## Part 5

RASCO: An R package to alleviate spatial  
confounding

## 5.1 RASCO: An R package to alleviate spatial confounding

Although several approaches to tackle spatial confounding are available, there is lack of adequate software for an unified implementation of these methods. The final contribution of this work is developing the R package `RASCO` for easy implementation of the approaches described in this work and also the GLMM alternatives, viz., the RHZ (REICH; HODGES; ZADNIK, 2006), the HH (HUGHES; HARAN, 2013), and the SPOCK (PRATES; ASSUNÇÃO; RODRIGUES, 2019).

For the HH, `RASCO` acts as a wrapper for the `ngspatial` package that implements the HH model (HUGHES; CUI, 2018) for GLMM. For this approach, four families are available: Gaussian, binomial, Poisson and negative binomial. For RSFM in spatial frailty models, RSCM in shared component models, RHZ for GLMM and SPOCK also for GLMM, `RASCO` relies on the `R-INLA` package for a faster INLA implementation. Therefore, all distributions and models from `R-INLA` are inherited.

Next sections show how to install and how to use the `RASCO` package.

### 5.1.1 Installation

To install the most recent version of the `RASCO` package from GitHub, one needs to type the following commands in a R console:

```
install.packages("devtools")
devtools::install_github("douglasmesquita/RASCO")
```

The functions that allow the use of the restricted models are:

- `rsglmm`: generalized linear mixed model wrapper,
- `rscm`: shared component model wrapper,
- `rsfm`: spatial frailty model wrapper,

These functions have three compulsory parameters: `data`, `formula` and `family`. For the shared component models, `family` is a vector of size 2 containing the families for each outcome. Also, `RASCO` has three functions to generate random data:

- `rglmm`: generalized linear mixed model random data,
- `rshared`: shared component model random data,
- `rsurv`: spatial frailty model random data,

The standard models are always unrestricted. To fit a restricted model, it is necessary to specify the column corresponding to the areas by the argument `area`, the neighborhood object via a spatial polygon (PEBESMA; BIVAND, 2005), or a simple feature object (PEBESMA, 2018) and, the spatial model as `"besag"` or its restricted version `"r_besag"`, for example. If the user chooses a restricted model (as `"r_besag"`), one can specify the projection based approach by the `proj` argument. The available projections may vary according to each model as described in the next sections.

For more details and examples access `?rsglmm`, `?rscm`, `?rsfm`.

## 5.1.2 Generalized Linear Mixed models

The `rsglmm` function has three compulsory parameters: `data`, `formula` and `family`. For the `family` parameter, the user is restricted to the Gaussian, Binomial, Poisson and negative binomial (with fixed  $m$ ) families if using the HH model, or any other family implemented in the R-INLA package for RHZ and SPOCK approaches. Some recommended families are: `"gaussian"` (Gaussian distribution), `"poisson"` (Poisson distribution), `"binomial"` (binomial distribution), `"gpoisson"` (generalized Poisson distribution) and `"nbinomial"` (negative binomial distribution). For a full list of available likelihoods use `INLA::inla.list.models()`.

The available projections for the restricted models are `"none"`, `"rhz"`, `"hh"`, or `"spock"`. Other possible parameters are:

- `nsamp`: number of desired samples. Default = 1000.
- `priors`: a list containing `prior_prec` a vector of size two containing shape and scale for the gamma prior distribution applied to  $\tau$ .
- `...`: other parameters used in `?INLA::inla` or `?ngspatial::sparse.sglm`

Below a simulated example using the Rio de Janeiro shapefile:

```
##-- Seed
set.seed(123456)

##-- Spatial structure
```

```

data("neigh_RJ")

##-- Parameters
beta <- c(-0.5, -0.2)
tau <- 1

##-- Data
family <- "poisson"
data <- rglmm(beta = beta, tau = tau, family = family,
              confounding = "none", neigh = neigh_RJ,
              scale = TRUE)

##-- Models
##-- + Non-spatial Poisson model
sglmm_mod <- rsglmm(data = data, formula = Y ~ X1 + X2,
                   family = family,
                   proj = "none", nsamp = 1000)

##-- + Spatial Poisson model
sglmm_mod <- rsglmm(data = data, formula = Y ~ X1 + X2,
                   family = family,
                   area = "reg", model = "besag",
                   neigh = neigh_RJ,
                   proj = "none", nsamp = 1000)

##-- + Restricted Spatial Poisson model - RHZ
rglmm_rhz <- rsglmm(data = data, formula = Y ~ X1 + X2,
                   family = family,
                   area = "reg", model = "r_besag",
                   neigh = neigh_RJ,
                   proj = "rhz", nsamp = 1000)

```

To fit the SPOCK or HH models, the unique change is in the `proj` argument that must be replaced by `"spock"` or `"hh"`. If we replace `"poisson"` by `"nbinomial"`, then a negative binomial model is fitted.

The outputs are standardized, and display the time elapsed (`$time`), the model fitted by R-INLA or `ngspatial` (`$out`) and two lists: `$unrestricted` and `$restricted`. Both of them have four entries; `$unrestricted$sample` corresponds to the sample taken from the

model, `$unrestricted$summary_fixed`, and `$unrestricted$summary_hyperpar`, contains the summaries for fixed effects and hyperparameters, respectively. The random effects summary is contained in `$unrestricted$summary_random`.

### 5.1.3 Shared Component models

The `rscm` function has four compulsory parameters: `data`, `formula1`, `formula2` and `family`. For the `family` parameter, the user should insert a vector of size 2 containing two families. `formula1` and `formula2` contains the fixed effects for disease 1 and 2, respectively. Again one can choose among several families as `"poisson"` (Poisson distribution), `"gpoisson"` (generalized Poisson distribution), `"nbinomial"` (negative binomial distribution), `"zeroinflatedpoisson0"` (zero-inflated poisson distribution) and `"zeroinflatednbinomial0"` (zero-inflated negative binomial distribution). For a full list of available likelihoods use `INLA::inla.list.models()`.

The available projections for the restricted models are `"none"` and `"spock"`. Other possible parameters are:

- `nsamp`: number of desired samples. Default = 1000.
- `priors`: a list containing `prior_gamma` and `prior_prec`. `prior_gamma` is a vector of size two containing mean and precision for the Gaussian prior distribution applied to  $\gamma$  parameter. `prior_prec` is a list containing `tau_s`, `tau_1` and `tau_2`. For each entry a vector of size two containing shape and scale for the gamma prior distribution applied to  $\tau_\psi$ ,  $\tau_{\phi_1}$  and  $\tau_{\phi_2}$  parameters, respectively.
- `random_effects`: a list containing three logical entries: `shared` to fit the shared component in the model, `specific_1` to fit the specific component for disease one and `specific_2` to fit the specific component for disease two.
- `...`: other parameters used in `?INLA::inla`

Below a simulated example using the Rio de Janeiro shapefile:

```
library(spdep)

set.seed(123456)

##-- Spatial structure
data("neigh_RJ")

##-- Parameters
```



```

alpha_1 <- 0.5
alpha_2 <- 0.1
beta_1 <- c(-0.5, -0.2)
beta_2 <- c(-0.8, -0.4)
tau_s <- 1
tau_1 <- tau_2 <- 10
delta <- 1.5

##-- Data
data <- rshared(alpha_1 = alpha_1, alpha_2 = alpha_2,
               beta_1 = beta_1, beta_2 = beta_2,
               delta = delta,
               tau_1 = tau_1, tau_2 = tau_2, tau_s = tau_s,
               confounding = "linear",
               neigh = neigh_RJ)

##-- Models
##-- + Restricted shared component model
scm_inla <- rscm(data = data,
                formula1 = Y1 ~ X11 + X12,
                formula2 = Y2 ~ X21 + X12,
                family = c("nbinomial", "poisson"),
                E1 = E1, E2 = E2,
                area = "reg", neigh = neigh_RJ,
                proj = "none", nsamp = 1000)

##-- + Restricted shared component model
rscm_inla <- rscm(data = data,
                 formula1 = Y1 ~ X11 + X12,
                 formula2 = Y2 ~ X21 + X12,
                 family = c("nbinomial", "poisson"),
                 E1 = E1, E2 = E2,
                 area = "reg", neigh = neigh_RJ,
                 proj = "spock", nsamp = 1000)

```

The outputs are standardized, and display the time elapsed (`$time`), the model fitted by R-INLA (`$out`), `$sample` corresponds to the sample taken from the model, `$summary_fixed`, and `$summary_hyperpar`, contains respectively, the summaries for fixed effects and hyperparameters, and the summary for the random effects is contained in the



```
    spatial = "ICAR",
    neigh = neigh_RJ, tau = tau,
    confounding = "linear", proj = "none")

##-- Models
##-- Spatial frailty model
sfm_inla <- rsfm(data = data,
                formula = surv(time = L, event = status) ~ X1 + X2,
                family = "weibull", model = "none",
                proj = "rhz", nsamp = 1000, approach = "inla")

##-- Restricted spatial frailty model
rsfm_inla <- rsfm(data = data,
                 formula = surv(time = L, event = status) ~ X1 + X2,
                 family = "weibull", area = "reg",
                 model = "r_besag", neigh = neigh_RJ,
                 proj = "rhz", nsamp = 1000, approach = "inla")
```

The outputs are standardized, and display the time elapsed (`$time`), the model fitted by R-INLA (`$out`) and two lists: `$unrestricted` and `$restricted`. Both of them have four entries; `$unrestricted$sample` corresponds to the sample taken from the model, `$unrestricted$summary_fixed`, and `$unrestricted$summary_hyperpar`, contains respectively, the summaries for fixed effects and hyperparameters, and the random effects summary is in `$unrestricted$summary_random` entry.

# Final remarks

The spatial confounding is a limitation of spatial models that needs attention since it can imply in wrong conclusions about important covariates effects. The conventional solution based on projections cannot be directly applicable for the shared component models neither for frailty models. For the first one, there are multiple spatial effects making the orthogonal projection a difficult task. For the second one, the fact that the support of fixed and random effects does not match makes the direct projection impossible.

This work showed alternatives to alleviate the effects of spatial confounding in the shared component models using the RSCM. Our approach creates several adjacency matrices using the SPOCK algorithm, one for each spatial effect in the model. This is preferable as an alternative to projection-based approaches that are confused in this scenario. Our approach appears as a good solution for spatial confounding since it is a prior correction on the original neighborhood structure. This is an important aspect because it enables the user to choose its preferred software to fit the model after creating the new adjacency matrices. We conducted a simulation study that showed the adequacy of our method in alleviating the spatial confounding. Also, we provided an efficient framework to fit the restricted spatial frailty model based on a posterior sample of the unrestricted model. To solve the difference in the lengths of fixed and spatial effects, we proposed a reduction operator that is not only adequate to alleviate the spatial confounding but also has computational benefits. The method adequacy and efficiency were shown by a simulation study that proved its relevance and importance.

We have developed two applications with data provided by the Surveillance, Epidemiology, and End Results (SEER) ([SEER, 2019](#)). Also, we enrich the data set with some county-level information provided by the County Health Rankings & Roadmaps (CHRR) ([RANKINGS; CHRR, 2019](#)). For the shared component model, we studied the new cases of bronchus and lung cancer in California in 2016 (last available year in January 2020). We showed that the percentage of people who smoke every day or most days were confounded with the spatial effect. The method proposed alleviated the effects of the spatial confounding keeping important conclusions about this covariate.

The spatial frailty model was employed to model the time until death by lung and bronchus cancer in California between 2010 and 2016. Our method provided an alleviation of the spatial confounding also keeping the model interpretability.

We developed an R package named `RASCO` aiming to unify the approaches in the literature that deal with spatial confounding. Nowadays it has functions to fit generalized linear mixed models, shared component models (two diseases) and spatial frailty models.

Also, a projection-based approach for these models is available as well as the SPOCK approach. The package is available at <https://github.com/douglasmesquita/RASCO>.

For future work, we may investigate the effects of temporal and spatio-temporal confounding in statistical models. To the best of our knowledge, up to now, little or no attention is paid to these models. Also, the reduction operator seems an easy and applied tool for statistical models. It is directly employed for discrete models in which the math involves products of a matrix and a variable that is constant by groups. Therefore, it is also possible to think in a discretization of continuous variables aiming to reduce the computational effort keeping the desired accuracy.

The SPOCK approach also needs the length equality of fixed and latent effects. Although we provided a way to proceed in this case, we did not investigate the effects of this methodology. Therefore, further studies are necessary to evaluate its capability of alleviating the spatial confounding in spatial frailty models. The same can be made for the HH approach and other projection-based approaches in the literature.

# Appendices

# Appendix A – RSCM - simulation

**Table 8** – Simulation results for the shared component model experiment (Scenarios S2 and S3). The results are shown by mean, standard deviation (SD), coverage rate for a nominal rate of 95 % (Cov) and mean square error (MSE).

Parameter	Real	S2						S3						
		SCM			RSCM			SCM			RSCM			
		Mean (SD)	Cov	MSE	Mean (SD)	Cov	MSE	Mean (SD)	Cov	MSE	Mean (SD)	Cov	MSE	
$\delta = 1.00$	$\beta_{10}$	0.50	0.50 (0.03)	94.70%	0.0012	0.49 (0.04)	96.10%	0.0012	0.50 (0.03)	94.90%	0.0011	0.50 (0.03)	93.20%	0.0014
	$\beta_{20}$	0.10	0.10 (0.04)	95.00%	0.0019	0.10 (0.04)	95.90%	0.0018	0.10 (0.04)	93.60%	0.0019	0.10 (0.04)	93.30%	0.0020
	$\beta_{11}$	-0.50	-0.51 (0.07)	94.20%	0.0057	-0.51 (0.09)	99.90%	0.0026	-0.50 (0.07)	94.60%	0.0043	-0.50 (0.08)	98.80%	0.0034
	$\beta_{21}$	-0.80	-0.81 (0.08)	94.70%	0.0060	-0.80 (0.09)	99.90%	0.0031	-0.80 (0.07)	94.30%	0.0051	-0.80 (0.08)	99.60%	0.0027
	$\beta_{12}$	-0.20	-0.20 (0.17)	95.20%	0.0271	-0.19 (0.08)	95.80%	0.0063	-0.20 (0.06)	94.20%	0.0036	-0.20 (0.07)	74.30%	0.0142
	$\beta_{22}$	-0.40	-0.40 (0.06)	94.70%	0.0032	-0.40 (0.06)	83.90%	0.0066	-0.40 (0.17)	96.40%	0.0262	-0.40 (0.08)	93.60%	0.0085
	$\delta$	1.00	1.02 (0.06)	90.10%	0.0039	1.07 (0.07)	84.20%	0.0116	1.02 (0.06)	89.80%	0.0040	0.95 (0.06)	83.20%	0.0074
$\delta = 1.50$	$\beta_{10}$	0.50	0.50 (0.04)	93.50%	0.0016	0.50 (0.04)	95.70%	0.0016	0.49 (0.04)	92.60%	0.0018	0.49 (0.04)	94.80%	0.0018
	$\beta_{20}$	0.10	0.09 (0.04)	93.20%	0.0016	0.10 (0.04)	93.00%	0.0017	0.10 (0.04)	94.50%	0.0018	0.10 (0.04)	92.80%	0.0019
	$\beta_{11}$	-0.50	-0.51 (0.11)	95.40%	0.0109	-0.51 (0.14)	99.90%	0.0060	-0.50 (0.10)	94.10%	0.0107	-0.50 (0.11)	99.90%	0.0030
	$\beta_{21}$	-0.80	-0.81 (0.07)	95.40%	0.0039	-0.80 (0.07)	98.80%	0.0032	-0.80 (0.06)	93.40%	0.0035	-0.80 (0.06)	99.40%	0.0016
	$\beta_{12}$	-0.20	-0.20 (0.28)	95.40%	0.0710	-0.19 (0.12)	97.80%	0.0111	-0.20 (0.08)	94.60%	0.0059	-0.20 (0.09)	76.50%	0.0209
	$\beta_{22}$	-0.40	-0.40 (0.04)	94.20%	0.0016	-0.40 (0.04)	86.40%	0.0027	-0.40 (0.14)	95.50%	0.0164	-0.39 (0.06)	97.70%	0.0031
	$\delta$	1.50	1.52 (0.11)	94.50%	0.0131	1.58 (0.13)	92.70%	0.0287	1.52 (0.10)	94.40%	0.0107	1.35 (0.10)	65.60%	0.0351
$\delta = 1.75$	$\beta_{10}$	0.50	0.49 (0.04)	94.00%	0.0019	0.49 (0.04)	93.80%	0.0019	0.49 (0.04)	94.40%	0.0020	0.50 (0.04)	93.20%	0.0022
	$\beta_{20}$	0.10	0.10 (0.04)	93.50%	0.0016	0.10 (0.04)	93.40%	0.0016	0.09 (0.04)	95.10%	0.0015	0.10 (0.04)	95.10%	0.0015
	$\beta_{11}$	-0.50	-0.50 (0.11)	94.20%	0.0130	-0.50 (0.15)	99.90%	0.0070	-0.50 (0.10)	94.40%	0.0105	-0.50 (0.12)	100.00%	0.0044
	$\beta_{21}$	-0.80	-0.80 (0.05)	93.10%	0.0032	-0.80 (0.06)	98.70%	0.0021	-0.80 (0.05)	94.20%	0.0024	-0.80 (0.06)	99.00%	0.0017
	$\beta_{12}$	-0.20	-0.20 (0.35)	94.40%	0.1127	-0.19 (0.14)	98.80%	0.0109	-0.20 (0.09)	94.40%	0.0086	-0.20 (0.11)	92.40%	0.0152
	$\beta_{22}$	-0.40	-0.40 (0.04)	93.80%	0.0016	-0.40 (0.04)	89.30%	0.0025	-0.40 (0.13)	93.80%	0.0165	-0.39 (0.06)	96.40%	0.0030
	$\delta$	1.75	1.79 (0.14)	93.90%	0.0239	1.83 (0.15)	90.60%	0.0411	1.78 (0.14)	94.40%	0.0235	1.50 (0.14)	53.90%	0.0870

## Appendix B – Reduction operator proofs

Let  $\mathbf{X}_{N \times p}$  be a matrix with entries  $X_{ijk}$  for an index  $i$ , an element  $j$  and column  $k$ , and  $\mathbf{G}_{N \times 1}$  a vector of indices indicating for each line of  $\mathbf{X}_{N \times p}$  an index  $i$  in a set of indices starting from 1 until  $n$  ( $n \ll N$ ). Then the reduction operator  $\textcircled{\mathbb{T}}$  is defined by:

$$\mathbf{X}_{N \times p} \textcircled{\mathbb{T}} \mathbf{G} = \mathbf{x}_{n \times p}, \quad (\text{B.1})$$

in which  $x_{ik} = \sum_{j=1}^{n_i} X_{ijk}$ , and  $n_i$  is the number of elements associated with index  $i$ .

**Definition B.1.1.** For  $\mathbf{X}_1$  and  $\mathbf{X}_2$ ,  $N \times p$  matrices with entries  $X_{dijk}$  for  $d = 1, 2$ , it is true that  $(\mathbf{X}_1 + \mathbf{X}_2) \textcircled{\mathbb{T}} \mathbf{G} = (\mathbf{X}_1 \textcircled{\mathbb{T}} \mathbf{G}) + (\mathbf{X}_2 \textcircled{\mathbb{T}} \mathbf{G})$ .

*Proof.* Consider the general term

$$\begin{aligned} [(\mathbf{X}_1 + \mathbf{X}_2) \textcircled{\mathbb{T}} \mathbf{G}]_{ik} &= \sum_{j=1}^{n_i} (X_{1ijk} + X_{2ijk}) \\ &= \sum_{j=1}^{n_i} X_{1ijk} + \sum_{j=1}^{n_i} X_{2ijk} \\ &= [\mathbf{X}_1 \textcircled{\mathbb{T}} \mathbf{G}]_{ik} + [\mathbf{X}_2 \textcircled{\mathbb{T}} \mathbf{G}]_{ik} \\ &= [(\mathbf{X}_1 \textcircled{\mathbb{T}} \mathbf{G}) + (\mathbf{X}_2 \textcircled{\mathbb{T}} \mathbf{G})]_{ik}, \end{aligned}$$

$$\forall i \in \{1, \dots, n\}, \forall k \in \{1, \dots, p\}$$

$$\implies (\mathbf{X}_1 + \mathbf{X}_2) \textcircled{\mathbb{T}} \mathbf{G} = (\mathbf{X}_1 \textcircled{\mathbb{T}} \mathbf{G}) + (\mathbf{X}_2 \textcircled{\mathbb{T}} \mathbf{G}). \quad \square$$

**Definition B.1.2.** For  $\mathbf{X}_{N \times p}$  matrix with entries  $X_{ijk}$  and a constant  $c$ , it is true that  $(c\mathbf{X}) \textcircled{\mathbb{T}} \mathbf{G} = c(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})$ .

*Proof.* Consider the general term

$$\begin{aligned} [(c\mathbf{X}) \textcircled{\mathbb{T}} \mathbf{G}]_{ik} &= \sum_{j=1}^{n_i} (cX_{ijk}) \\ &= c \sum_{j=1}^{n_i} X_{ijk} \\ &= c [\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G}]_{ik} \\ &= [c(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G})]_{ik}, \end{aligned}$$

$$\forall i \in \{1, \dots, n\}, \forall k \in \{1, \dots, p\}$$

$$\implies (c\mathbf{X}) \textcircled{\mathbb{T}} \mathbf{G} = c(\mathbf{X} \textcircled{\mathbb{T}} \mathbf{G}). \quad \square$$



**Definition B.1.3.** For  $\mathbf{X}_{N \times p}$  matrix with entries  $X_{ijk}$ ,  $\mathbf{r}_{n \times 1}$  a column vector,  $\mathbf{R} = [r_{G_1}, \dots, r_{G_N}]^T$  a  $N \times 1$  vector with repeated entries for each index of  $\mathbf{G}$  (constant by indices), it is true that  $\mathbf{X}^T \mathbf{R} = (\mathbf{X} \oplus \mathbf{G})^T \mathbf{r}$ .

*Proof.* Consider the general term

$$\begin{aligned}
[\mathbf{X}^T \mathbf{R}]_k &= \sum_{l=1}^N (X_{lk} R_l) \\
&= \sum_{i=1}^n \sum_{j=1}^{n_i} (X_{ijk} r_i) \\
&= \sum_{i=1}^n r_i \sum_{j=1}^{n_i} X_{ijk} \\
&= \sum_{i=1}^n r_i [\mathbf{X} \oplus \mathbf{G}]_{ik} \\
&= [\mathbf{X} \oplus \mathbf{G}]_{1k} r_1 + \dots + [\mathbf{X} \oplus \mathbf{G}]_{nk} r_n \\
&= [(\mathbf{X} \oplus \mathbf{G})^T]_{.k} \mathbf{r} \\
&= [(\mathbf{X} \oplus \mathbf{G})^T \mathbf{r}]_k,
\end{aligned}$$

$\forall k \in \{1, \dots, p\}$

$$\implies \mathbf{X}^T \mathbf{R} = (\mathbf{X} \oplus \mathbf{G})^T \mathbf{r}. \quad \square$$

**Definition B.1.4.** For  $\mathbf{X}_{N \times p}$  matrix with entries  $X_{ijk}$  and  $\mathbf{Q}_{M \times p}$  a matrix with entries  $Q_{mk}$ , it is true that  $(\mathbf{Q} \mathbf{X}^T) \oplus \mathbf{G}^T = \mathbf{Q} (\mathbf{X} \oplus \mathbf{G})^T$ .

*Proof.* Consider the general term  $[(\mathbf{Q} \mathbf{X}^T)]_{ml} = \sum_{k=1}^p Q_{mk} X_{lk}$ , and

$$\begin{aligned}
[(\mathbf{Q} \mathbf{X}^T) \oplus \mathbf{G}^T]_{mi} &= \sum_{j=1}^{n_i} \sum_{k=1}^p Q_{mk} X_{ijk} \\
&= \sum_{k=1}^p Q_{mk} \sum_{j=1}^{n_i} X_{ijk} \\
&= \sum_{k=1}^p Q_{mk} [(\mathbf{X} \oplus \mathbf{G})]_{ik} \\
&= Q_{m1} [(\mathbf{X} \oplus \mathbf{G})]_{i1} + \dots + Q_{mp} [(\mathbf{X} \oplus \mathbf{G})]_{ip} \\
&= \mathbf{Q}_m \cdot [(\mathbf{X} \oplus \mathbf{G})^T]_{.i} \\
&= [\mathbf{Q} (\mathbf{X} \oplus \mathbf{G})^T]_{mi},
\end{aligned}$$

$\forall m \in \{1, \dots, M\}, \forall k \in \{1, \dots, p\}$

$$\implies (\mathbf{Q} \mathbf{X}^T) \oplus \mathbf{G}^T = \mathbf{Q} (\mathbf{X} \oplus \mathbf{G})^T. \quad \square$$

**Definition B.1.5.** For  $\mathbf{X}_{N \times p}$  matrix with entries  $X_{ijk}$  and  $\mathbf{P}_{p \times p}$  a squared matrix, it is true that  $(\mathbf{X}\mathbf{P}\mathbf{X}^T) \oplus \mathbf{G} = (\mathbf{X} \oplus \mathbf{G})\mathbf{P}\mathbf{X}^T$ .

*Proof.*

$$\begin{aligned}
(\mathbf{X}\mathbf{P}\mathbf{X}^T) \oplus \mathbf{G} &= (\mathbf{X}\mathbf{K}) \oplus \mathbf{G} \\
&= ((\mathbf{K}^T \mathbf{X}^T) \oplus \mathbf{G}^T)^T \\
&\stackrel{1.4}{=} (\mathbf{K}^T (\mathbf{X} \oplus \mathbf{G})^T)^T \\
&= (\mathbf{X} \oplus \mathbf{G}) \mathbf{K} \\
&= (\mathbf{X} \oplus \mathbf{G}) \mathbf{P}\mathbf{X}^T.
\end{aligned}$$

□

**Definition B.1.6.** For  $\mathbf{X}_{N \times p}$  matrix with entries  $X_{ijk}$  and  $\mathbf{P}_{p \times p}$  a squared matrix, it is true that  $((\mathbf{X}\mathbf{P}\mathbf{X}^T) \oplus \mathbf{G}) \oplus \mathbf{G}^T = (\mathbf{X} \oplus \mathbf{G})\mathbf{P}(\mathbf{X} \oplus \mathbf{G})^T$ .

*Proof.*

$$\begin{aligned}
((\mathbf{X}\mathbf{P}\mathbf{X}^T) \oplus \mathbf{G}) \oplus \mathbf{G}^T &\stackrel{1.5}{=} ((\mathbf{X} \oplus \mathbf{G})\mathbf{P}\mathbf{X}^T) \oplus \mathbf{G}^T \\
&= (\mathbf{K}\mathbf{X}^T) \oplus \mathbf{G}^T \\
&\stackrel{1.4}{=} \mathbf{K}(\mathbf{X} \oplus \mathbf{G})^T \\
&= (\mathbf{X} \oplus \mathbf{G})\mathbf{P}(\mathbf{X} \oplus \mathbf{G})^T.
\end{aligned}$$

□

**Definition B.1.7.** For  $\mathbf{X}_{N \times p}$  matrix with entries  $X_{ijk}$ ,  $\mathbf{r}_{n \times 1}$  a column vector,  $\mathbf{R} = [r_{G_1}, \dots, r_{G_N}]^T$  a  $N \times 1$  vector with repeated entries for each index of  $\mathbf{G}$  (constant by indices) and  $\mathbf{P}_{p \times p}$  a squared matrix, it is true that  $(\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{R}) \oplus \mathbf{G} = (\mathbf{X} \oplus \mathbf{G})\mathbf{P}(\mathbf{X} \oplus \mathbf{G})^T \mathbf{r}$ .

*Proof.*

$$\begin{aligned}
(\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{R}) \oplus \mathbf{G} &= (\mathbf{X}\mathbf{K}) \oplus \mathbf{G} \\
&= ((\mathbf{K}^T \mathbf{X}^T) \oplus \mathbf{G}^T)^T \\
&\stackrel{1.4}{=} (\mathbf{K}^T (\mathbf{X} \oplus \mathbf{G})^T)^T \\
&= (\mathbf{X} \oplus \mathbf{G}) \mathbf{K} \\
&= (\mathbf{X} \oplus \mathbf{G}) \mathbf{P}\mathbf{X}^T \mathbf{R} \\
&\stackrel{1.3}{=} (\mathbf{X} \oplus \mathbf{G}) \mathbf{P}(\mathbf{X} \oplus \mathbf{G})^T \mathbf{r}.
\end{aligned}$$

□

## References

- BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. *Hierarchical modeling and analysis for spatial data*. [S.l.]: CRC Press, 2014. Cited 3 times on pages [19](#), [20](#) e [22](#).
- BANERJEE, S.; WALL, M. M.; CARLIN, B. P. Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, Oxford University Press, v. 4, n. 1, p. 123–142, 2003. Cited 2 times on pages [15](#) e [36](#).
- BARANOVSKY, A.; MYERS, M. H. Cancer incidence and survival in patients 65 years of age and older. *CA: a Cancer Journal for Clinicians*, Wiley Online Library, v. 36, n. 1, p. 26–41, 1986. Cited on page [42](#).
- BASTOS, L. S.; GAMERMAN, D. Dynamic survival models with spatial frailty. *Lifetime Data Analysis*, Springer, v. 12, n. 4, p. 441–460, 2006. Cited on page [15](#).
- BESAG, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 36, n. 2, p. 192–225, 1974. Cited 3 times on pages [15](#), [20](#) e [21](#).
- BESAG, J.; YORK, J.; MOLLIÉ, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, Springer, v. 43, n. 1, p. 1–20, 1991. Cited on page [21](#).
- BIRNBAUM, Z. W.; SAUNDERS, S. C. A new family of life distributions. *Journal of Applied Probability*, Cambridge University Press, v. 6, n. 2, p. 319–327, 1969. Cited on page [33](#).
- BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, JSTOR, v. 11, n. 1, p. 15–53, 1949. Cited on page [34](#).
- BOLOKER, G.; WANG, C.; ZHANG, J. Updated statistics of lung and bronchus cancer in united states (2018). *Journal of Thoracic Disease*, AME Publications, v. 10, n. 3, p. 1158, 2018. Cited on page [42](#).
- BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, CRC press, v. 88, n. 421, p. 9–25, 1993. Cited on page [18](#).
- CARLIN, B. P.; BANERJEE, S. Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian Statistics*, Oxford University Press Oxford, v. 7, n. 7, p. 45–63, 2003. Cited on page [30](#).
- CLAYTON, D. G.; BERNARDINELLI, L.; MONTOMOLI, C. Spatial correlation in ecological analysis. *International Journal of Epidemiology*, Oxford University Press, v. 22, n. 6, p. 1193–1202, 1993. Cited on page [14](#).
- COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Cited on page [35](#).

CRESSIE, N. Statistics for spatial data. *Terra Nova*, Wiley Online Library, v. 4, n. 5, p. 613–617, 1992. Cited 2 times on pages 15 e 26.

CZADO, C.; SANTNER, T. J. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, Elsevier, v. 33, n. 2, p. 213–231, 1992. Cited on page 19.

DABNEY, A. R.; WAKEFIELD, J. C. Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 14, n. 1, p. 83–112, 2005. Cited on page 31.

DATTA, A. et al. Spatial disease mapping using directed acyclic graph auto-regressive (dagar) models. *Bayesian Analysis*, International Society for Bayesian Analysis, v. 14, n. 4, p. 1221–1244, 2019. Cited on page 20.

DAVIS, M. L. et al. Addressing geographic confounding through spatial propensity scores: a study of racial disparities in diabetes. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 28, n. 3, p. 734–748, 2019. Cited 2 times on pages 14 e 19.

DIGGLE, P. et al. *Analysis of longitudinal data*. [S.l.]: Oxford University Press, 2002. Cited on page 19.

FRIEDMAN, M. Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 10, n. 1, p. 101–113, 1982. Cited on page 33.

FU, J. B. et al. Lung cancer in women: analysis of the national surveillance, epidemiology, and end results database. *Chest*, Elsevier, v. 127, n. 3, p. 768–777, 2005. Cited on page 42.

GAMERMAN, D.; LOPES, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. [S.l.]: Chapman and Hall/CRC press, 2006. Cited on page 37.

GELFAND, A. E.; VOUNATSOU, P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, Oxford University Press, v. 4, n. 1, p. 11–15, 2003. Cited on page 30.

GELMAN, A. et al. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 49, n. 2, p. 247–268, 2000. Cited on page 27.

GÓMEZ-RUBIO, V.; PALMÍ-PERALES, F. Multivariate posterior inference for spatial models with the integrated nested laplace approximation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 68, n. 1, p. 199–215, 2019. Cited 2 times on pages 47 e 51.

GUAN, Y.; HARAN, M. A computationally efficient projection-based approach for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, CRC press, v. 27, n. 4, p. 701–714, 2018. Cited 3 times on pages 14, 19 e 22.

HANKS, E. M. et al. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, Wiley Online Library, v. 26, n. 4, p. 243–254, 2015. Cited 5 times on pages 11, 14, 22, 26 e 58.

- HEFLEY, T. J. et al. The bayesian group lasso for confounded spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, Springer, v. 22, n. 1, p. 42–59, 2017. Cited 3 times on pages [14](#), [19](#) e [22](#).
- HENDERSON, R.; SHIMAKURA, S.; GORST, D. Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, CRC press, v. 97, n. 460, p. 965–972, 2002. Cited on page [15](#).
- HODGES, J. S.; REICH, B. J. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, CRC press, v. 64, n. 4, p. 325–334, 2010. Cited on page [19](#).
- HOSMER, D. W.; LEMESHOW, S.; MAY, S. *Applied survival analysis: regression modeling of time-to-event data*. [S.l.]: Wiley-Interscience, 2008. v. 618. Cited on page [34](#).
- HUGHES, J.; CUI, X. *ngspatial: Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data*. Denver, CO, 2018. R package version 1.2-1. Cited on page [72](#).
- HUGHES, J.; HARAN, M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 75, n. 1, p. 139–159, 2013. Cited 9 times on pages [11](#), [14](#), [19](#), [22](#), [24](#), [25](#), [46](#), [62](#) e [72](#).
- JIN, X.; BANERJEE, S.; CARLIN, B. P. Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 69, n. 5, p. 817–838, 2007. Cited on page [30](#).
- JIN, X.; CARLIN, B. P.; BANERJEE, S. Generalized hierarchical multivariate car models for areal data. *Biometrics*, Wiley Online Library, v. 61, n. 4, p. 950–961, 2005. Cited on page [30](#).
- KNORR-HELD, L.; BEST, N. G. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 164, n. 1, p. 73–85, 2001. Cited 3 times on pages [15](#), [30](#) e [31](#).
- KNORR-HELD, L. et al. Towards joint disease mapping. *Statistical Methods in Medical Research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 14, n. 1, p. 61–82, 2005. Cited 2 times on pages [31](#) e [32](#).
- KNORR-HELD, L.; RASSER, G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, Wiley Online Library, v. 56, n. 1, p. 13–21, 2000. Cited on page [31](#).
- LAMBERT, P. C. Modeling of the cure fraction in survival studies. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 7, n. 3, p. 351–375, 2007. Cited on page [35](#).
- LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley and Sons, 2011. v. 362. Cited on page [35](#).

- LEROUX, B. G.; LEI, X.; BRESLOW, N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: *Statistical models in epidemiology, the environment, and clinical trials*. [S.l.]: Springer, 1999. p. 179–191. Cited on page 20.
- LI, Y.; RYAN, L. Modeling spatial survival data using semiparametric frailty models. *Biometrics*, Wiley Online Library, v. 58, n. 2, p. 287–297, 2002. Cited on page 15.
- LINDGREN, F.; RUE, H. et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, Foundation for Open Access Statistics, v. 63, n. 19, p. 1–25, 2015. Cited on page 47.
- MORAN, P. A. A test for the serial independence of residuals. *Biometrika*, JSTOR, v. 37, n. 1/2, p. 178–181, 1950. Cited on page 25.
- NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Cited on page 18.
- NETER, J. et al. *Applied linear statistical models*. [S.l.]: Irwin Chicago, 1996. v. 4. Cited on page 18.
- ORD, K. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, CRC press, v. 70, n. 349, p. 120–126, 1975. Cited on page 20.
- OSAMA, M.; ZACHARIAH, D.; SCHÖN, T. Inferring heterogeneous causal effects in presence of spatial confounding. *arXiv preprint arXiv:1901.09919*, 2019. Cited 2 times on pages 14 e 19.
- PACIOREK, C. J. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, NIH Public Access, v. 25, n. 1, p. 107, 2010. Cited on page 28.
- PAPADOGEORGOU, G.; CHOIRAT, C.; ZIGLER, C. M. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*, Oxford University Press, v. 20, n. 2, p. 256–272, 2018. Cited 2 times on pages 14 e 19.
- PEBESMA, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, v. 10, n. 1, p. 439–446, 2018. Disponível em: <<https://doi.org/10.32614/RJ-2018-009>>. Cited on page 73.
- PEBESMA, E. J.; BIVAND, R. S. Classes and methods for spatial data in R. *R News*, v. 5, n. 2, p. 9–13, November 2005. Disponível em: <<https://CRAN.R-project.org/doc/Rnews/>>. Cited on page 73.
- PRATES, M. O.; ASSUNÇÃO, R. M.; RODRIGUES, E. C. Alleviating spatial confounding for areal data problems by displacing the geographical centroids. *Bayesian Analysis*, International Society for Bayesian Analysis, v. 14, n. 2, p. 623–647, 2019. Cited 8 times on pages 11, 14, 16, 19, 22, 27, 50 e 72.
- RANKINGS, C. H.; CHRR, R. *University of Wisconsin Population Health Institute. County Health Rankings and Roadmaps 2019*. 2019. <<http://www.countyhealthrankings.org>>. Accessed: 2020-01-16. Cited 2 times on pages 41 e 79.

- REICH, B. J.; HODGES, J. S.; ZADNIK, V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, Wiley Online Library, v. 62, n. 4, p. 1197–1206, 2006. Cited 11 times on pages 11, 14, 16, 19, 22, 23, 24, 27, 28, 46 e 72.
- RODRIGUES, E. C. *Estruturas de Covariância em Modelos Espaciais Bayesianos*. Tese (Doutorado) — ICEX - Universidade Federal de Minas Gerais, 2012. Disponível em: <[http://www.est.ufmg.br/portal/arquivos/doutorado/teses/erica\\_castilho\\_rodrigues.pdf](http://www.est.ufmg.br/portal/arquivos/doutorado/teses/erica_castilho_rodrigues.pdf)>. Cited on page 30.
- RODRIGUES, E. C.; ASSUNÇÃO, R. Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, Elsevier, v. 109, p. 88–102, 2012. Cited on page 20.
- RUE, H.; KNORR-HELD, L. *Gaussian Markov random fields: theory and applications*. [S.l.]: Chapman and Hall/CRC press, 2005. Cited 2 times on pages 26 e 38.
- RUE, H.; MARTINO, S.; CHOPIN, N. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 71, n. 2, p. 319–392, 2009. Cited 3 times on pages 37, 38 e 39.
- SAIN, S. R.; FURRER, R.; CRESSIE, N. A spatial analysis of multivariate output from regional climate models. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 5, n. 1, p. 150–175, 2011. Cited on page 30.
- SCUDILIO, J. et al. Defective models induced by gamma frailty term for survival data with cured fraction. *Journal of Applied Statistics*, CRC press, v. 46, n. 3, p. 484–507, 2019. Cited on page 35.
- SEER, S. R. P. *National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER\*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2018 Sub (1975-2016)*. 2019. Released April 2019, based on the November 2018 submission. Cited 3 times on pages 16, 42 e 79.
- SIEGEL, R. L.; MILLER, K. D.; JEMAL, A. Cancer statistics, 2019. *CA: a Cancer Journal for Clinicians*, Wiley Online Library, v. 69, n. 1, p. 7–34, 2019. Cited on page 42.
- SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Cited on page 26.
- THADEN, H.; KNEIB, T. Structural equation models for dealing with spatial confounding. *The American Statistician*, CRC press, v. 72, n. 3, p. 239–252, 2018. Cited 3 times on pages 14, 19 e 22.
- TSODIKOV, A.; IBRAHIM, J.; YAKOVLEV, A. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, CRC press, v. 98, n. 464, p. 1063–1078, 2003. Cited on page 35.
- VARGAS, F. R. *Bayesian estimates of the lethality rate of acute myocardial infarction*. Dissertação (Mestrado) — ICEX - Universidade Federal de Minas Gerais (UFMG), 2013. Disponível em: <<http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/BUOS-98KHRK>>. Cited on page 47.



- WATANABE, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, v. 11, n. Dec, p. 3571–3594, 2010. Cited 2 times on pages 26 e 54.
- WHITTLE, P. On stationary processes in the plane. *Biometrika*, Oxford University Press, v. 41, n. 3/4, p. 434–449, 1954. Cited on page 20.
- WHO, W. H. O. Gender in lung cancer and smoking research. Geneva: World Health Organization, 2004. Disponível em: <<https://apps.who.int/iris/bitstream/handle/10665/43086/9241592524.pdf>>. Cited on page 42.
- WIENKE, A. *Frailty models in survival analysis*. [S.l.]: Chapman and Hall/CRC press, 2010. Cited on page 36.
- YANCIK, R.; KESSLER, L.; YATES, J. W. The elderly population opportunities for cancer prevention and detection. *Cancer*, Wiley Online Library, v. 62, n. S1, p. 1823–1828, 1988. Cited on page 42.
- YANCIK, R.; RIES, L. G. Cancer in the aged. an epidemiologic perspective on treatment issues. *Cancer*, Wiley Online Library, v. 68, n. S11, p. 2502–2510, 1991. Cited on page 42.