Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação em Estatística

Danna Lesley Cruz Reyes

Spatial models with random covariance structure

Belo Horizonte
2021

Danna Lesley Cruz Reyes

Spatial models with random covariance structure

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

Orientador: Dra. Rosângela Helena Loschi

Coorientador: Dr. Renato Martins Assunção

Belo Horizonte

2021

Cruz Reyes, Danna Lesley.

C957s Spatial models with random covariance structure [manuscrito]. / Danna Lesley Cruz Reyes — 2021.
112 f. il.

Orientadora: Rosângela Helena Loschi.
Coorientador: Renato Martins Assunção

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 94-96; 98

1. Estatística - Teses 2. Análise espacial (Estatística) – teses. 3. Autoregressão (Estatística) – Teses. 4. Câncer – Região Sul – Métodos estatísticos. -Teses. I. Loschi, Rosângela Helena. II. Assunção, Renato Martins. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatísitca .IV.Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz

UNIVERSIDADE FEDERAL DE MINAS GERAIS

UFMG

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ATA DA DEFESA DE TESE DE DOUTORADO DA DANNA LESLEY CRUZ REYES , MATRICULADO, SOB O Nº 2017.667.123, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 28 DE JULHO DE 2021.

Aos vinte e oito (28) dias do mês de Julho de 2021, às 10h00, em reunião pública virtual nº. 69, realizada conforme orientações da Portaria PRPG nº 1819, via Microsoft Teams (link: `https://teams.microsoft.com/l/team/19%3aP5HvHUJq_yMpxX0Bxz8BOpW5llYV3hA1J9ajaI29MD81%40thread.tacv2/conversations?groupId=9da0e259-1265-4f14-b7a8-023dda71eef8&tenantId=64126139-4352-4cd7-b1fb-2a971c6f69a6`), como requisito final para obtenção do Grau de doutor em Estatística, reuniu-se a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística formada pelos professores abaixo relacionados, para julgar a defesa de tese da aluna DANNA LESLEY CRUZ REYES, nº matrícula 2017.667.123, intitulada: "*Spatial models random covariance structure"*. Abrindo a sessão, a Senhora Presidente da Comissão, Profa. Rosangela Helena Loschi, passou a palavra à aluna para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da aluna. Após a defesa, os membros da banca examinadora reuniram-se reservadamente, sem a presença da aluna e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

RHLoschi
Profa. Rosangela Helena Loschi
(Orientadora -EST/UFMG)

Renat Assunp
Prof. Renato Martins Assunção
(Co-Orientador-DCC/UFMG)

Prof. Vinícius Diniz Mayrink
(EST/UFMG)

Guilherme V. Nunes Ludwig
Prof. Guilherme Vieira Nunes Ludwig
(IME/Unicamp)

João Batista de Morais Per
Prof. João Batista de Morais Pereira
(DME/UFRJ)

Giovani Loiola da Silva
Prof. Giovani Loiola da Silva
(IST/ULisboa)

O resultado final foi comunicado publicamente à aluna pela Senhora Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 28 de julho de 2021.

Observações:
1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

FOLHA DE APROVAÇÃO

"Spatial models with random covariance structure"

DANNA LESLEY CRUZ REYES

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Doutor em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 28 de julho de 2021, pela banca constituída pelos membros:

RHLoschi
Prof(a). Rosangela Helena Loschi - Orientador
DEST/UFMG

Renat Assunção
Prof(a). Renato Martins Assunção -Coorientador
DCC/UFMG

GMayrink
Prof(a). Vinícius Diniz Mayrink
DEST/UFMG

Guilherme V. Nunes Ludwig
Prof(a). Guilherme Vieira Nunes Ludwig
IME/UNICAMP

João Batista de Morais Per
Prof(a). João Batista de Morais Pereira
DME/UFRJ

Giovani Loiola da Silva
Prof(a). Giovani Loiola da Silva
IST/ULisboa

Belo Horizonte, 28 de julho de 2021.

Agradecimentos

Inicialmente, agradeço às instituições que me apoiaram financeiramente, pois sem elas minha permanência, bem como minha tranquilidade teriam sido quase impossíveis. Refiro-me à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e á Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), por seu apoio através do pagamento de bolsas e auxílios a estudantes de mestrado e doutorado.

Agradeço à Universidade Federal de Minas Gerais (UFMG), Instituição pública que me permitiu estudar e morar no Brasil.

Agradeço ao Departamento de Estatística da UFMG (DEST-UFMG), formado pelo corpo docente e pelo corpo administrativo. Espero que continuem com sua universidade, a mesma qualidade e serviço. Levam a ciência além do limite.

Este trabajo está dedicado a todos los efectos fijos que hacen parte de mi vida, mi hermano Dizzy, mi Guillermito y mi querida Alison, mi mamita, mi hermana, mis amigas y toda mi familia Reyes.
A mi constante, Alejandro, que desde que lo conocí me enseño a enamorarme de la ciencia, a él le debo mi camino y tiene mi corazón por siempre.
A mis efectos aleatorios, la UFMG la mejor universidad, mis compañeros y administrativos. Pero especialmente al profesor Renato Assunção, no puedo sentirme más honrada de ser su estudiante y a la profesora Rosangela Loschi, "você é a melhor mulher, cientista, professora que já conheci e também tem uma paciência infinita". Finalmente mi amada Cantelli, ninguna ecuación estaría completa sin el amor de un perro.

Resumo

O modelo autorregressivo condicional (modelo CAR) é a distribuição mais popular para conjuntamente modelar a incerteza *a priori* sobre dados espacialmente correlacionados. Em geral, é utilizado em modelos espaciais hierárquicos onde modela a incerteza sobre os efeitos aleatórios espaciais. Uma limitação do modelo CAR é sua incapacidade de produzir correlações altas entre áreas vizinhas. Propomos um modelo robusto para dados de área que ameniza esse problema. Representamos o mapa por um grafo não direcionado onde os nós representam as áreas e as arestas conectam nós vizinhos no mapa. Atribuímos às arestas pesos distintos e aleatórios. O modelo é baseado em uma distribuição multivariada $t-$ Student, espacialmente estruturada, em que a matriz de precisão é indiretamente construída assumindo-se uma distribuição multivariada para os pesos aleatórios das arestas. Tal distribuição $t-$ Student correlaciona espacialmente os pesos das arestas e induz um outro modelo t-Student para o efeitos espaciais das áreas que os correlaciona e é capaz de acomodar *outliers* e comportamento de cauda pesada para estes efeitos. Mais importante, o modelo proposto pode produzir uma correlação marginal mais alta entre os efeitos espaciais do que o modelo CAR, superando uma das principais limitações deste modelo. Ajustamos o modelo proposto para mapear a incidência de alguns tipos câncer na região sul do Brazil e comparamos seu desempenho com vários modelos alternativos propostos na literatura. Os resultados mostram que o modelo proposto é competitivo e fornece resultados similares e, em alguns casos, melhores que os obtidos ajustando modelos comumente usados para analisar este tipo de dados. Na segunda proposta, abordamos o problema de redução de dimensionalidade em modelos de regressão. Um dos métodos mais utilizados para evitar sobreajuste e selecionar variáveis relevantes em modelos de regressão com muitos preditores é a técnica de regressão penalizada. Sob tais abordagens, a seleção de variáveis é realizada de forma não probabilística utilizando algum critério de otimização. Abordagens Bayesianas para a regressão penalizada têm sido proposta assumindo uma distribuição *a priori* para os coeficientes de regressão que desempenha um papel semelhante ao termo de penalidade nas estatísticas clássicas: comprimir em direção a zero coeficientes não significativos e colocar uma massa de probabilidade significativa em coeficientes que podem ser agrupados. Geralmente, tais distribuições *a priori*, chamadas *shrinkage priors* (ditribuições *a priori* de encolhimento), assumem independência entre os efeitos das covariáveis, o que pode não ser uma suposição apropriada em muitos casos. Neste trabalho, focamos na redução de dimensionalidade de variáveis categóricas com muitos níveis. Estas vaiáveis são incluídas no modelo através de variáveis *dummy* induzindo esparsidade na matrix de delineamento, o que pode gerar sobreajuste e dificuldades na interpretação dos resultados. O efeito dos níveis destas variáveis categóricas são naturalmente correlacionados. Para lidarmos com este problema, propomos duas distribuições *a priori* de encolhimento para os coeficientes associados aos níveis de variáveis categóricas, correlacionando-os. As distribuições propostas são próprias e, além de esparsidade, têm a propriedade de agrupar efeitos similares. Ilustrarmos o uso destas distribuições aplicando-as na redução de dimensionalidade em um regressão linear. Seus desempenhos são analisados e comparados a métodos pré-existentes por meio de estudos de dados simulados e considerando dados de preços de habitação disponíveis no Airbnb.

Palavras-chave: Estatística espacial, modelo CAR, grafo de arestas, distribuições *a priori* de regularização, modelo espacial robusto.

Abstract

The conditional autoregressive model (CAR model) is the most popular distribution for jointly modeling the *a priori* uncertainty over spatially correlated data. In general, it is used in hierarchical spatial models where it models the uncertainty about random spatial effects. A limitation of the CAR model is its inability to produce high correlations between neighboring areas. We propose a robust model for area data that alleviates this problem.

We represent the map by an undirected graph where nodes represent areas and edges connect neighboring nodes on the map. We assign distinct and random weights to the edges. The model is based on a spatially structured $t-$Student multivariate distribution, in which the precision matrix is indirectly constructed assuming a multivariate distribution for the random weights of the edges.

Such $t-$ Student distribution spatially correlates the edge weights and induces another t-Student model for the spatial effects of the areas that correlates them and is able to accommodate *outliers* and heavy tail behavior for these effects . More importantly, the proposed model can produce a higher marginal correlation between spatial effects than the CAR model, overcoming one of the main limitations of this model. We adjusted the proposed model to map the incidence of some types of cancer in southern Brazil and compared its performance with several alternative models proposed in the literature. The results show that the proposed model is competitive and provides similar and, in some cases, better results than those obtained by fitting models commonly used to analyze this type of data. In the second proposal, we approach the problem of dimensionality reduction in regression models. One of the most used methods to avoid overfitting and to select relevant variables in regression models with many predictors is the penalized regression technique. Under such approaches, variable selection is performed in a non-probabilistic way using some optimization criterion. Bayesian approaches to penalized regression have been proposed assuming an *a priori* distribution for the regression coefficients that plays a role similar to the penalty term in classical statistics: compressing towards zero non-significant coefficients and putting a probability mass significant in coefficients that can be grouped. Generally, such *a priori* distributions, called *shrinkage priors* (shrinkage *a priori* distributions), assume independence between the effects of the covariates, which may not be an appropriate assumption in many cases. In this work, we focus on the dimensionality reduction of categorical variables with many levels. These variables are included in the model through variables *dummy* inducing sparsity in the design matrix, which can generate overfitting and difficulties in interpreting the results. The effect of the levels of these categorical variables are naturally correlated. To deal with this problem, we propose two *a priori* shrinkage distributions for the coefficients associated with the levels of categorical variables, correlating them. The proposed distributions are proper and, in addition to sparsity, they have the property of grouping similar effects. We illustrate the use of these distributions by applying them to dimensionality reduction in a linear regression. Their performances are analyzed and compared to pre-existing methods through simulated data studies and considering housing price data available on Airbnb.

Keywords: Spatial statistic, CAR model, graph of edges, shrinkage prior, robust spatial models.

Sumário

1 Introdução

Os modelos estatísticos para dados espaciais são divididos por Cressie [1993] em duas classes amplas: modelos geoestatísticos com suporte espacial contínuo e modelos em um lattice, também chamados de modelos de área [Banerjee et al., 2014]. Nesta tese, trabalhamos modelos espaciais para dados de área, especificamente com a estrutura de covariância responsável por capturar a correlação imposta pela natureza espacial dos dados. Analisar esta correlação é essencial nas estatísticas, visto que os dados de área são aplicados em muitas situaçoes, incluindo o mapeamento das taxas de doenças, [Elliott and Wartenberg, 2004], agricultura [Besag and Higdon, 1999], econometria [Lesage and Pace, 2009], ecologia [Arslan and Akyurek, 2018] e análise de imagens, [Besag, 1986].

A abordagem desses problemas envolve a construção de um modelo espacial baseado nos campos aleatórios de Markov (GRMF por suas siglas em inglês). Um campo aleatório descreve a associação espacial entre um conjunto de variáveis univariadas através de distribuições condicionais bivariadas, e como seu nome implica, a distribuição conjunta de todas essas variáveis segue uma distribuição Gaussiana multivariada [Rue and Knorr-Held, 2005]. Um dos modelos mais utilizados, propõe que cada variável seja representada por um efeito espacial e uma distribuição Gaussiana com uma média ponderando observações vizinhas e a variância inversamente proporcional ao número de áreas vizinhas. A definição da vizinhança é dada por um grafo, onde os nós são cada uma das áreas e as arestas são as conexões entre as áreas que compartilham uma borda geográfica. Este modelo é chamado de modelo autorregressivo condicional (CAR) e foi proposto por Besag [1974].

O modelo CAR depende de um único parâmetro ρ de ponderação da média condicional. Este parâmetro visa medir uma força de dependência espacial de todo o mapa. No entanto, na literatura há trabalhos que chamam a atenção para resultados não intuitivos. Especificamente, em Wall [2004], há resultados contraditórios, por exemplo, o sinal do valor estimado do parâmetro é positivo e a correlação entre as áreas pode ser negativa. Alguns desses resultados são explicados em Assunção and Krainski [2009]. No entanto, existem problemas mais sérios, em alguns casos, embora o valor de ρ esteja próximo a seu valor máximo, ele gera correlações marginais muito baixas, mesmo na versão imprópria do modelo CAR, onde o valor do parâmetro é 1, seu máximo valor.

Além disso, o fato de ser função de em um único parâmetro, o modelo CAR força os efeitos aleatórios a exibirem um nível global único de autocorrelação espacial, que varia da independência a uma suavização espacial forte. Um nível uniforme de suavidade espacial para toda a região não é realista, já que provavelmente exitem sub-áreas de autocorrelação espacial separadas por descontinuidades. Tal alisamento espacial localizado pode ocorrer onde comunidades ricas e pobres vivem lado a lado e, neste contexto, é provável que a variável resposta evolua suavemente dentro de cada comunidade com uma mudança repentina em seu valor na fronteira em que as duas comunidades são encontradas. Existem várias propostas para lidar com a suavização localizada. Lawson and Clark [2002] combina o modelo intrínseco com um componente de "salto" para descontinuidades, Brewer and Nolan [2007] suaviza a variável por medio de uma variância espacial aleatoria, Lu et al. [2007] modela a estrutura de adjacência das unidades de área usando regressão logística, Reich et al. [2006] suaviza a variável por meio de uma variância espacial em um ambiente espaço-temporal e Lee and Mitchell [2011] modela a correlação parcial entre efeitos aleatórios em unidades de áreas adjacentes. Porém, esses modelos podem exigir um certo grau de programação para implementá-los [Lee, 2013].

Para resolver esses problemas, propomos duas abordagens diferentes para modelos espaciais. A primeira abordagem segue a mesma estrutura hierárquica do CAR, onde os efeitos espaciais relacionados às áreas representam os nós de um grafo. Este grafo representa a estrutura de vizinhança definida pela região geográfica. No entanto, assumimos que os pesos das arestas deste grafo são distintos e aleatórios. Consideramos um novo grafo, *um grafo de arestas*, o qual estabelece a estrutura de vizinhança entre as arestas do grafo original. Duas arestas são consideradas vizinhas se incidem em um mesmo nó. Propomos um modelo conjunto para o peso das arestas que compõem o grafo original o qual impõe uma estrutura de correlação entre o peso das arestas similar à estrutura presente no CAR. Tal modelo inclui um segundo efeito espacial relacionado à estrutura de vizinhança entre as arestas. Assumimos uma distribuição *a priori* normal multivariada para o vetor dos pesos das arestas, condicional em sua matriz de covariância, e uma distribuição Wishart-Inversa para tal matriz. Como consequência, a distribuição conjunta do peso das arestas áreas têm uma distribuição t-Student cuja matriz de covariâncias tem estrutura similar a um modelo CAR. Considerando propriedades da matriz de incidência do grafo e alguns resultados de álgebra provamos que o efeito aleatório representado por cada nó no grafo original (ou seja, entre os efeitos aleatórios espaciais) é uma combinação linear dos efeitos aleatórios de suas arestas incidentes. Usando propriedades da distribuição t-Student provamos que a distribuição conjunta dos efeitos aleatórios (nós) representados no grafo original também é uma distribuição t-Student.

Este modelo induz uma correlação marginal maior do que a fornecida pelo modelo

CAR entre os efeitos aleatórios. Como a distribuição t-Student tem cauda pesada, o modelo proposto impede a suavização extrema do mapa. Apresentamos as propriedades da correlação marginal, condicional e correlação parcial deste modelo, simulando dados reais.

Em uma segunda contribuição deste trabalho propomos outra maneira de minimizar os problemas do modelo CAR. Para contornar sua dependência em um único parâmetro, abordaremos o problema de modelar o comportamento conjunto dos efeitos espaciais de uma perspectiva diferente em que tais efeitos espaciais são vistos como níveis de uma variável categórica. Um problema recorrente de difícil solução neste tipo de modelagem é a presença de um grande número de níveis ou categorias, como seria o caso, dado que o número de níveis depende do número de áreas. Criscuolo [2019] oferece uma solução Bayesiana para este tipo de problema considerando um modelo de partição aleatória para os níveis da variável categórica. Inspirado nesta ideia de agrupamento, neste trabalho definimos distribuições *a priori* de encolhimento (*shrinkage prior*) para os níveis da variável categórica (efeitos espaciais), as quais impõem uma correlação espacial entre tais níveis. Tais distribuições são inspiradas pelo modelo CAR assumindo que parâmetro ρ é substituido por um vetor $\boldsymbol{\rho}$ cujas coordenadas representam os peso das arestas. Também propomos uma nova distribuição *a priori* para o vetor dos pesos das arestas, a qual também impõe uma correlação espacial por meio do grafo de arestas. A distribuição proposta permite a conjugação simplificando a implementação computacional. A distribuição marginal *a priori* dos efeitos espacias indica que nossa proposta favorece ambos, esparsidade e agrupamento. Considerando tal *shrinkage prior* simultaneamente estimamos os pesos das arestas e o efeito de cada nível da variável categórica além de identificar possíveis “clusters” dos efeitos espaciais sem gasto computacional adicional uma vez que todas as distribuições condicionais *a posteriori* são conhecidas. Através de simulação, comparamos o desempenho das distribuições propostas com outras distribuições de encolhimento recentemente propostas na literatura. Mostramos que tais distribuições têm um desempenho preditivo similar ao de outras abordagens e resultados mais interpretáveis. As distribuições propostas também são consideradas na análise dos dados do Airbnb visando estimar os efeitos da localização do imóvel no preço de seu aluguel e identificar regioes para as quais tais efeitos são similares.

Na próxima seção, apresentamos algumas definições e resultados necessários para o desenvolvimento da tese.

1.1 Campos Aleatorios de Markov Gaussianos

Campos aleatórios são distribuições multivariadas que são, em geral, utilizadas para descrever a associação espacial entre variáveis $\boldsymbol{Y} = (Y_1, ..., Y_n)$. Um campo aleatório Markoviano estende o conceito de cadeia de Markov para um contexto espacial e assume que tal distribuição conjunta de $\boldsymbol{X}$ satisfaz a seguinte condição:

$$f(Y_i|\boldsymbol{Y}_{-i}) = f(Y_i|\boldsymbol{Y}_{j\sim i}),$$

em que, $\boldsymbol{Y}_{j\sim i}$ é o vetor formado por todas as componentes de $\boldsymbol{Y}$ que são vizinhos de i. Neste capítulo discutiremos, brevemente, os campos aleatórios de Markov Gaussianos que são frequentemente utilizados como distribuição *a priori* para os efeitos espaciais.

Um Campo aleatório de Markov Gaussiano (GMRF) é um campo de Markov onde a distribuição do vetor aleatório $\boldsymbol{Y}$ de dimensão finita é Gaussiana, satisfazendo as suposições de independência condicional. Uma discussão detalhada sobre GMRF pode ser encontrada em (Rue and Knorr-Held [2005]) e (Assunção and Krainski [2009]).

Todos os resultados válidos para a distribuição normal, também serão válidos para um GMRF. Na seguinte seção apresentamos os resultados mais relevantes da distribuição normal multivariada. Após definir formalmente um GMRF com todas as propriedades herdadas da distribuição normal, apresentaremos a conexão entre o grafo $\mathcal{G}$ e os parâmetros da distribuição normal multivariada $\boldsymbol{\mu}$ e Σ. Será mostrado que toda a informação do grafo está condensada na matriz de covariâncias Σ por meio da matriz de precisão $Q = \Sigma^{-1}$, e que o vetor de médias $\boldsymbol{\mu}$ não terá influência na estrutura de vizinhança do grafo.

1.1.1 A distribuição normal multivariada

Para facilitar o entendimento dos campos aleatorios Markovianos Gaussianos revisamos a distribuição normal multivariada e algumas das suas propriedades básicas.

Um vetor aleatório $n-$dimensional $\boldsymbol{y}_{n\times 1} = (y_1, y_2, \ldots, y_n)^\top$, $n < \ \infty$ tem distribuição Normal $n-$variada com vetor de média $\boldsymbol{\mu}_{n\times 1}$ e matriz de covariâncias $\boldsymbol{\Sigma}_{n\times n}$, se sua função de densidade de probabilidade (f.d.p) assume a seguinte forma:

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\}, \quad \boldsymbol{y} \in \mathbb{R}^n. \tag{1.1}$$

Esta distribuição será denotada por $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ onde $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são tais que $\mu_i = E(y_i)$, $\Sigma_{ij} = Cov(y_i, y_j)$, $\Sigma_{ii} = Var(y_i)$ e $Corr(y_i, y_j) = \Sigma_{ij}(\Sigma_{ii}\Sigma_{jj})^{-1/2}$.

Para apresentar algumas propiedades da distribução em (1.1), considere a seguinte partição: $\boldsymbol{y} = (\boldsymbol{y}_A, \boldsymbol{y}_B)^\top$, $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)^\top$,

$$\begin{bmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{AB} & \boldsymbol{\Sigma}_{BB} \end{bmatrix}$$

assumindo tal partição, algumas propriedades básicas da distribuição normal são:

- $\boldsymbol{y}_A \sim N(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA})$ é a distribuição marginal do vetor $\boldsymbol{y}_A$ de ordem $A \times 1$;

- $\boldsymbol{\Sigma}_{AB} = \boldsymbol{0}$ se e somente se $\boldsymbol{y}_A$ e $\boldsymbol{y}_B$ são independentes;

- A distribuição condicional de $\boldsymbol{y}_A$ dado $\boldsymbol{y}_B$ é $N(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B})$ onde,

$$\begin{aligned} \boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\boldsymbol{y}_A - \boldsymbol{\mu}_B) \quad \text{e} \qquad (1.2) \\ \boldsymbol{\Sigma}_{A|B} &= \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}. \qquad (1.3) \end{aligned}$$

1.1.2 Definição e propriedades básicas do GMRF

Para construir um GMRF consideramos um grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, com $\mathcal{V} = \{\nu_1, \ldots, \nu_n\}$ os n vértices, onde cada vértice representa uma das componentes do vetor $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ e o conjunto $\mathcal{E}$ as arestas conectam nós que têm algum tipo de associação. Um GMRF assume que $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e que as arestas do grafo conectam os nós i e j se e somente se $y_i \perp y_j | \boldsymbol{y}_{-ij}$, isto é, se y_i é independente de y_j, dados os componentes de $\boldsymbol{y}$ exceto y_i e y_j.

Em um GMRF, a matriz de covariância carrega a informação das conexões entre os nós através da *matriz de precisão* $\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}$ que é uma matrix simétrica e definida positiva.

Teorema 1. *Se* $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{Q})$, *então para* $i \neq j$, $y_i \perp y_j | \boldsymbol{y}_{-ij} \Leftrightarrow Q_{ij} = 0$.

A demonstração deste teorema pode ser encontrada em Rue and Knorr-Held [2005]. Este resultado estabelece que os componentes não nulos de $\boldsymbol{Q}$ determinam a relação de vizinhança presente em $\mathcal{G}$. Isto implica que qualquer distribuição normal com matriz de covariância definida positiva é também um GMRF e vice-versa. Formalmente, um GMRF é definido como segue:

Definição 1. *Um vetor aleatório $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top \in \mathcal{R}^n$ é chamado GMRF correspondente a um grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ com média $\boldsymbol{\mu}$ e matriz de precisão $\boldsymbol{Q} > 0$, se e somente se a f.d.p. de $\boldsymbol{y}$ tem a seguinte forma:*

$$\pi(\boldsymbol{y}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{y}-\boldsymbol{\mu})\right),$$

em que a matriz $\boldsymbol{Q}$ satisfaz a condição:

$$Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in \mathcal{E}, \forall i \neq j.$$

Se $\boldsymbol{Q}$ for uma matriz completamente densa, então $\mathcal{G}$ está totalmente conectado, isto é, o vértice esta conectado a todos os outros vértices do grafo. Vamos nos concentrar no caso em que $\boldsymbol{Q}$ é esparsa.

Teorema 2. *Seja o grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ que representa um GMRF para $\boldsymbol{y}$, com média $\boldsymbol{\mu}$ e matriz de precisão $\boldsymbol{Q}$ simétrica e definida positiva. Então a distribução de cada componente y_i de $\boldsymbol{y}$, dado o vetor $\boldsymbol{y}_{-i}$ formado por todas as componentes de $\boldsymbol{y}$, exceto y_i é uma distribução normal tal que:*

$$\begin{aligned} E(y_i|y_{-i}) &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(y_j - \mu_j), \\ Prec(y_i|y_{-i}) &= Q_{ii}, \\ Corr(y_i, y_j|\boldsymbol{y}_{-ij}) &= -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j, \end{aligned}$$

onde $i \sim j$ denota que o nó j é vizinho do nó i.

1.2 Modelos hierárquicos Bayesianos para dados espaciais de área

Considere uma região geográfica particionada em n sub-regiões indexadas por inteiros $1, 2, \ldots, n$. Assuma que esta coleção de sub-regiões é dotada de um sistema de vizinhança $\{V_i : i : 1, \ldots, n\}$, onde V_i denota a coleção de sub-regiões que, em um sentido bem definido, são vizinhos da subregião i. Em termos geográficos,

$$V_i = \{j : \text{as subregiões } i \text{ e } j \text{ compartilham fronteira}\}, \quad \text{para} \;\; i \in \{1, 2, \ldots, n\},$$

Para este caso, os grafos que subsidiam a construção dos GMRF serão aqueles que expressam essas estruturas de vizinhança. Neste contexto, as arestas $\mathcal{E}$ no grafo $\mathcal{B} = (\mathcal{G}, \mathcal{E})$, representam as conexões na estrutura geográfica e, consequentemente definem os vizinhos que são usados para modelar a dependência espacial. Os componentes do vetor $\boldsymbol{y}$ são nós do grafo. Sejam $y_1, ..., y_n$ as observações feitas nas áreas $1, \ldots, n$. Denotemos por $j \sim i$ que o nó j é um vizinho do nó i. O padrão espacial na resposta é modelado por uma matriz de covariáveis $\boldsymbol{X} = (x_1, ..., x_n)^\top$. Um conjunto de efeitos aleatórios $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, são incluídos para modelar qualquer autocorrelação espacial que permaneça nos dados após os efeitos da covariável terem sido contabilizados.

O vetor de covariáveis para a unidade de área V_i são denotados por $\boldsymbol{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$. O modelo hierárquico Bayesiano pode ser escrito de forma geral como:

$$y_i|\mu_i \sim f(y_i|\mu_i, \tau), \;\text{ para }\; i = 1, \ldots, n, \tag{1.4}$$
$$g(\mu_k) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \theta_i + \epsilon_i.$$

As respostas y_i vêm de uma família exponencial de distribuições $f(Y_i|\mu_i, \sigma^2)$. O valor esperado de Y_i é denotado por $E(Y_i) = \mu_i$, enquanto σ^2 é um parâmetro de escala adicional necessário se a família Gaussiana for usada. Os valores esperados das respostas estão relacionados ao preditor linear através de uma função de ligação invertível $g(.)$, por exemplo, a função logit (família binomial), identidade (família Gaussiana) ou log natural (família de Poisson). O vetor de parâmetros de regressão são denotados por $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$.

1.2.1 Distribuições a priori

Para cada parâmetro de regressão β_j, com $j \in 1, \ldots, p$, é atribuída uma distribuição *a priori* Gaussiana independente não informativa. Para o parâmetro τ da verossimilhança Gaussiana é atribuída uma distribuição *a priori* Gamma (para detalhes, ver Gelman et al. [2004]). Para o vetor $\boldsymbol{\theta}$ é possível implementar uma série de modelos de efeitos aleatórios diferentes, sendo o mais simples assumindo independência *a priori*:

$$\begin{aligned} \theta_i &\sim N(0, \tau_\theta), \\ \tau_\theta &\sim U(0, M_{\tau_\theta}). \end{aligned}$$

Neste caso $U(0, M_{\tau_\theta})$ é uma distribuição uniforme no intervalo $(0, M_{\tau_\theta})$, com, $M_{\tau_\theta} = 1000$. Esta especificação é apropriada se as covariáveis incluídas no modelo (1.4) eliminaram toda a estrutura espacial da resposta. No entanto, para a maioria dos conjuntos de dados, é provável que exista uma autocorrelação espacial residual, nestes casos, pode ser necessária alguma distribuição *a priori* que impõe autocorrelação espacial.

Por exemplo, o modelo CAR proposto por Besag [1974], os modelos intrínsecos e BYM [Besag et al., 1991] e as alternativas desenvolvidas por Leroux et al. [2000]. Cada modelo é um caso especial de um campo aleatório de Markov Gaussiano (GMRF), e pode ser escrito na forma geral $\boldsymbol{\theta} \sim N(\mathbf{0}, \tau_\theta^2 \boldsymbol{Q})$, onde $\boldsymbol{Q}$ é a matriz de precisão. Esta matriz controla a estrutura de autocorrelação espacial dos efeitos aleatórios e é baseada na vizinhança dada pela matriz da adjacência $\boldsymbol{A}$.

A matriz da adjacência $\boldsymbol{A}$ é definida com $a_{ii} = 0$, $a_{ij} = 1$ se i e j $a_{ij} = 0$ se $i \nsim j$ e a matriz do vizinhança $\boldsymbol{M}$, com $\boldsymbol{M} = diag\{d_1, d_2, \ldots, d_n\}$ do grafo definido $\mathcal{G}$. Essas distribuições *a priori* são comumente especificadas como um conjunto de n distribuições condicionais univariadas $f(\theta_i|\boldsymbol{\theta_{-i}})$ para $i \in 1, \ldots, n$, em que, $\boldsymbol{\theta_{-i}} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$.

O modelo CAR é definido pelas condicionais:

$$\theta_i|\theta_{-i} \sim N\left(\rho \frac{\sum_{j \sim i} \theta_i}{n_i}, \tau_\theta n_i\right). \tag{1.5}$$

em que n_i denota o número de vizinhos para a $i-$ésima região e ρ é um parâmetro de autocorrelação espacial. Assim,

$$\boldsymbol{\theta} \sim N(\mathbf{0}, \tau_\theta^2 \boldsymbol{Q}), \tag{1.6}$$

onde, $\boldsymbol{Q} = (\boldsymbol{M} - \rho \boldsymbol{A})$. Para que a matriz de precisão $\boldsymbol{Q}$ seja invertível, se tem que definir uma faixa de valores para o parâmetro ρ. Seja $\{\lambda_i\}$ o conjunto de autovalores da matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$. Se λ_1 é o menor valor próprio da matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$ e λ_p é o maior valor próprio da matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$. Se $1/\lambda_1 < \rho < 1/\lambda_p$, então $(\boldsymbol{M} - \gamma \boldsymbol{A})$ é definida positiva. Existe uma definição semelhante, mas usando a matriz de peso $\boldsymbol{W}$, tal que $w_{ii} = 0$, $w_{ij} = 1/d_{ij}$ se i e j $w_{ij} = 0$ se $i \nsim j$. No entanto, ambas as definições são equivalentes, isso é provado no apêndice A.

Neste modelo, a esperança condicional é a média dos efeitos aleatórios em áreas vizinhas, enquanto a variância condicional é inversamente proporcional ao número de vizinhos. O último é apropriado porque, se os efeitos aleatórios forem espacialmente correlacionados, quanto mais vizinhos uma área tiver, mais informação haverá provediente de seus vizinhos sobre o valor do efeito aleatório. Embora popular, existem muitos resultados intrigantes em relação ao modelo CAR. Conforme discutido por Wall [2004] e Assunção et al. [2010], o modelo CAR produz uma pequena correlação marginal que depende de um parâmetro espacial e da topologia estática de vizinhança assumida *a priori*. Além disso, a modelagem CAR também fornece variância não constante em cada área e a correlação entre pares de áreas separadas pelo mesmo número de vizinhos não são necessariamente iguais [Rue and Knorr-Held, 2005, Besag and Kooperberg, 1995, Bernadinelli et al., 1995, Banerjee et al., 2003]. Consequentemente, o modelo CAR tende a suavizar descontinuidades, tornando-o inadequado para algumas reconstruções de imagens [Aykroyd, 1998]. Alguns desses resultados contra-intuitivos são explicados detalhadamente por Assunção et al. [2010]. Portanto, os mesmos autores propuseram uma extensão para permitir a autocorrelação espacial fraca e forte, substituindo θ_i em 1.4 com $\theta_i + \phi_i$, que são respectivamente

$$\begin{aligned} \theta_i|\boldsymbol{\theta}_{-i} &\sim N\left(\rho\frac{\sum_{j\sim i}\theta_i}{n_i}, \tau_\theta n_i\right), \\ \phi_i &\sim N(0,\tau_\phi), \\ \tau_\theta &\sim U(0, M_{\tau_\theta}). \end{aligned}$$

Este modelo é conhecido como BYM ou modelo de convolução, proposto por Besag et al. [1991]. Ele, requer que dois efeitos aleatórios sejam estimados para cada ponto de dados, enquanto apenas sua soma é identificável a partir dos dados. Logo, Leroux et al. [2000] propôs distribuições *a priori* alternativas para modelar intensidades variáveis de autocorrelação espacial, usando apenas um único conjunto de efeitos aleatórios. O modelo de **Leroux** é dado por

$$\theta_i|\boldsymbol{\theta}_{-i} \sim N\left(\rho\frac{\sum_{j=1}^n a_{ij}\theta_i}{\rho\sum_{j=1}^n a_{ij} + 1 - \rho}, \frac{\tau_\theta^2}{\rho\sum_{j=1}^n a_{ij} + 1 - \rho}\right), \tag{1.7}$$

em que ρ é um parâmetro de autocorrelação espacial. O caso $\rho = 0$ corresponde à independência, enquanto $\rho \approx 1$ corresponde a uma forte autocorrelação espacial. Quando $\rho = 1$ no modelo 1.5 e 1.7 se obtem o modelo impróprio **ICAR**, proposto por Besag et al. [1995]. É chamado impróprio, porque a matriz de precisão que gera não possui inversa.

O problema é que a estrutura de vizinhança desses modelos determina o grau de suavização e é usada para estimar o risco relativo. Assim, o modelo tende a mitigar muito os riscos quando se utiliza a estrutura usual de vizinhança adjacente. Rodrigues and Assunção [2012] investigou estruturas de vizinhança mais flexíveis para modelos autorregressivos condicionais espaciais e propôs o modelo HND (higher-neigbourhood dependence), no qual a estrutura de vizinhança faz parte do espaço de parâmetros. Então, mantendo o modelo definido em 1.4, o vetor $\boldsymbol{\theta}$ tem uma distribuição normal multivariada com média vetor de zeros e matriz de precisão dada por

$$\boldsymbol{Q} = \tau_\theta(\lambda_1\boldsymbol{I} + \lambda_2\boldsymbol{R}^{(1)} \ldots, +\lambda_k\boldsymbol{R}^{(k)}) \tag{1.8}$$

onde $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$ e $\lambda_i \geq 0$. O inteiro k é o diâmetro do grafo que é o caminho mais longo entre todos os caminhos mais curtos que conectam dois sites. Em outras palavras, o diâmetro conta o número mínimo de etapas necessárias para sair de um site e ir para qualquer outro site do grafo. O $\boldsymbol{R}$ é o grafo Laplaciano que inclui as vizinhaça até a ordem $l > 1$. O modelo permite a definição múltipla de uma vizinhança de suavização e pode ser especialmente útil na situação em que o risco subjacente é praticamente constante. O modelo estende os modelos BYM e Leroux, considerando uma dependência de vizinhaça superior. No entanto, Rodrigues and Assunção [2012] descreve que é recomendável que seja aplicado a outros tipos de dados espaciais requerendo a especificação de estruturas de vizinhança, como problemas de espaço-tempo ou análise de dados de sobrevivência espacial.

Em alguns casos, propõe-se alterar a estrutura do grafo, como no caso do modelo proposto por Datta et al. [2019] chamado DAGAR (Directed Acyclic Graph Autoregressive) é determinado pelas distribuições condicionais θ_i

$$\begin{aligned}
\theta_1 &= \epsilon_1 \\
\theta_2 &= b_{21}\theta_1 + \epsilon_2 \\
\theta_3 &= b_{31}w_1 + b_{32}\theta_2 + \epsilon_3 \\
&\vdots \\
\theta_k &= b_{k1}\theta_1 + b_{k2}\theta_2 + \cdots + b_{k,2k-1}\theta_{k-1} + \epsilon_k,
\end{aligned}$$

em que $\epsilon_i \sim N(0, \tau_i)$ independente. Então, se $B = (b_{ij})$, se tem

$$\boldsymbol{\theta} \sim N(0, L^\top F L), \quad \text{sendo}, \qquad L = I - B.$$

Outro modelo alternativo é o modelo GSN (generalized skew-normal), proposto por Prates et al. [2012]. Este artigo desenvolve um novo processo espacial usando distribuições generalizadas enviesadas normais e independentes quando as suposições usuais do processo Gaussiano são questionáveis e a transformação para um campo aleatório Gaussiano não é apropriada. O modelo proposto fornece flexibilidade na captura dos efeitos de assimetria e comportamento de cauda pesada dos dados. Isso é feito mantendo a dependência espacial através do uso uma estrutura autorregressiva condicional. Da mesma forma anterior, no o modelo definido em 1.4, o vetor $\boldsymbol{\theta}$ agora tem una distribuição *a priori* Gaussiana generalizada ou **GSGSF** (por suas siglas em inglês, Generalized skew-Gaussian spatial field),

$$\boldsymbol{\theta} \sim GSGSF_n\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, H_n(\nu)\right) \tag{1.9}$$

onde $\boldsymbol{\Sigma}$ tem uma dependência espacial gerada por uma estrutura CAR e $H_n(\nu)$ é uma das distribuições enviesadas definidas em Prates et al. [2012].

Na segunda parte desta tese, construímos as distribuições *a priori* de encolhimento (*shrinkage prior*) para a modelagem de dados categóricos. Na próxima seção, é feita uma revisão de alguns métodos de regularização utilizados na redução de dimensionalidade em modelos de regressão.

1.3 Visão geral: penalização do tipo L1

Considere o modelo de regressão linear normal

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n), \tag{1.10}$$

onde $\boldsymbol{Y}$ é um vetor $n \times 1$ das variáveis dependentes, $\boldsymbol{X}$ é a matriz de delineamento de dimensão $n \times p$ cuja entrada x_{ij} representa o valor observado da j-ésima covariável para o i-ésimo indivíduo, $i = 1, \dots, n$ e $j = 1, \dots, p$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ é o vetor dos coeficientes de regressão e $\boldsymbol{I}_n$ é a matriz identidade de ordem $n \times n$.

Em problemas de regressão envolvendo um grande número de preditores, o problema de super-dispersão é um ponto crítico. Para amenizar esse tipo de problema, métodos de redução de dimensionalidade têm sido propostos.

Os métodos clássicos de regularização minimizam a soma dos quadrados dos resíduos sujeitos a um termo de penalização que impõe uma suposição de esparsidade ou agrupamento. O termo para induzir esparsidade visa reduzir para zero efeitos de covariáveis não significativas (próximos de zero). O termo para induzir o agrupamento surge quando se deseja detectar variáveis cujos efeitos βs são similares e diferente de zero [Liu et al., 2014].

Os métodos com penalização do tipo $L1-$ agrupam as estimativas dos coeficientes de regressão em direção a zero [Tibshirani, 1996]. Diferentes penalidades do tipo $L1-$ são introduzidas para efeitos de covariáveis em modelos lineares, por exemplo, o Lasso convencional [Tibshirani, 1996], Lasso para agrupamento [She, 2010] e Lasso para fusão [Tibshirani et al., 2005] introduzido para tratar covariáveis categóricas com muitos níveis. Em particular, as penalidades envolvidas nos métodos Lasso para agrupamento e fusão foram introduzidas para tratar preditores categóricos nominais e ordinais, abordando características específicas. Para obter uma solução esparsa, o Lasso utiliza a norma $L1-$ como termo de penalidade.

Os problemas de esparsidade e o agrupamento dos efeitos de covariáveis podem ser abordados usando o paradigma Bayesiano, onde o efeito da penalização é introduzido através de distribuições *a priori* de encolhimento ou *shrinkage priors*. *Shrinkage priors* para os métodos Lasso bayesiano e os métodos bayesianos Lasso para agrupamento e fusão foram propostos por Hans [2009], Kyung et al. [2010] e Li and Lin [2010]). Nos métodos bayesianos, as penalidades são substituidas por distribuições *a priori* sobre os βs, as quais são construídas hieraquicamente como uma mistura da distribução normal multivariada e uma escolha adequada da distribuição *a priori* para a matriz de covariância.

O método **Lasso convencional** proposto Tibshirani [1996] assume para a estimação dos coeficiente de regressão β_j a seguinte penalização

$$\lambda \sum_{j=1}^{p} |\beta_j|,$$

em que o parâmetro λ é um parâmetro de ajuste que controla a força geral da penalidade. Na versão bayesiana deste método, o **Lasso Bayesiano**, a penalização e substituida pela seguinte estrutura hierárquica para a distribuição *a priori* dos βs

$$\pi(\boldsymbol{\beta}|\sigma,\ \tau_1, \ldots, \tau_p) \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{D}_\tau),$$

onde $\boldsymbol{D}_\tau = diag(\tau_1^2, \ldots, \tau_p^2)$ e a distribuição *a priori* para $(\tau_1, \ldots, \tau_p)$ é

$$\pi(\tau_1, \ldots, \tau_p) \propto \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\{-\lambda^2 \tau_j^2/2\}.$$

Esta estrutura de penalidade reduz o coeficiente de regressão para zero. Se o efeito for suficientemente pequeno, o coeficiente de regressão pode até mesmo ser definido com precisão para zero. Assumindo-se grandes valores de λ, apenas os efeitos βs mais influentes são mantidos no modelo e todos os outros efeitos são reduzidos a zero Groll et al. [2019].

O método **Lasso para agrupamento** foi proposto por She [2010] para reduzir dimensionalidade de uma covariável categórica. Neste método, utiliza-se a norma L2 de β_j para construir a função de penalização a qual é dada por

$$\lambda \sum_{j=1}^{p} |\beta_j| + \lambda \sum_{i<j}^{p} |\beta_i - \beta_j|,$$

No Lasso para agrupamento Bayesiano, a estrutura hierárquica é definida por

$$\pi(\boldsymbol{\beta}|\sigma,\ \tau_1,\ldots,\tau_p) \sim N_{m_k}(\mathbf{0},\sigma^2\tau_k^2 I_{mk}),$$

em que m_k é o número pré-especificado de variáveis no $k-$ésimo grupo do Lasso de agrupamento Bayesiano, e

$$\pi(\tau_1,\ldots,\tau_p) \propto \prod_{j=1}^{p} Gamma((m_k+1)/2, \lambda^2/2).$$

Em alguns casos, a característica de interesse pode ter ordenação natural como, por exemplo, uma estrutura espacial ou temporal a qual deve ser considerada durante a análise. Tibshirani et al. [2005]) propõem uma metodologia para redução de dimensionalidade nestas situações conhecida como **Lasso para fusão** (*fused Lasso*). A função objetivo neste método é

$$\min_{\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i^t\beta\right)^2\right\} \tag{1.11}$$

$$\text{sujeito a } \sum_{j=1}^{p}|\beta_j| \leq t_1 \ \text{ e } \ \sum_{\mathrm{j=2}}^{\mathrm{p}}|\beta_{\mathrm{j}} - \beta_{\mathrm{j-1}}| \leq \mathrm{t_2}. \tag{1.12}$$

A primeira restrição em (1.12) é a restrição assumida no método Lasso e tem a mesma função "encolhendo"para zero os coeficientes não significativos. A segunda restrição em (1.12) penaliza diretamente grandes mudanças em relação à estrutura temporal ou espacial, o que força os coeficientes a variarem suavemente para refletir a lógica subjacente do sistema. O método Lasso para agrupamento é uma generalização do método Lasso para Fusão que identifica e agrupa covariáveis relevantes com base em seus efeitos (coeficientes). A ideia básica é penalizar as diferenças entre os coeficientes de forma que coeficientes diferentes de zero e similares sejam agrupados. Isso pode ser feito usando a seguinte regularização: $\sum_{i<j}^{p}|\beta_i - \beta_j| \leq t_2$.

1.4 Contribuiçoes deste trabalho

Como mencionamos, existem várias distribuições *a priori* propostas na literatura para modelar o comportamento de efeitos espaciais como, por exemplo, os modelos propostos por Knorr-Held and Best [2001], MacNab and Dean [2000], Martínez-Beneito et al. [2008], Silva et al. [2008], Carlin and Banerjee [2003] e Jin and Carlin [2005] para citar alguns. Nosso objetivo principal neste trabalho é construir novos modelos que sejam capazes de amenizar os problemas mencionados na introdução deste capítulo. Este objetivo é alcançado ao propormos duas distribuições *a priori* para o efeito espacial $\boldsymbol{\theta}$. No primeiro modelo discutido no Capítulo 3, propomos uma distribuição de cauda pesada induzido pelas correções entre os pesos das arestas que conectam os nós (efeitos espaciais) em nosso grafo original. Na segunda abordagem discutida no no Capítulo 3 propomos uma distribuição tipo *shrinkage prior* para tais efeitos visando a redução de dimensionalidade em nosso problema de predição.

O modelo proposto no Capítulo 2 traz a novidade na forma como construímos a matriz de precisão. Representamos o mapa através de um grafo em que os nós são os efeitos espaciais e as arestas são fornecidas pelos mapa geográfico. Os pesos $\boldsymbol{\rho}$ das arestas são aleatórios. A esses pesos, atribuímos uma distribuição $t-$ Student multivariada onde a matriz de covariâncias tem uma estrutura semelhante à do modelo CAR, mas definida sob grafo de arestas. Então, o componente θ_i do vetor $\boldsymbol{\theta}$ é definido como a soma de suas arestas incidentes. Alcançamos isso empregando a matriz de correlação dos pesos das arestas e a matriz de incidência do grafo. O modelo espacial resultante para $\boldsymbol{\theta}$ tem uma distribuição multivariada de $t-$ Student e herda as correlações impostas pelo grafo de arestas e as propriedades essenciais da distribuição t, como sua capacidade para dados atípicos e comportamento de cauda pesada. Também induz uma correlação marginal mais alta do que o modelo CAR, fornecendo uma solução para uma das principais limitações do CAR. Ajustamos o modelo proposto para analisar mapas reais de câncer e comparamos seu desempenho com vários modelos propostos na literatura. Nosso modelo proposto se ajusta melhor em quase todos os casos. O modelo é chamado RENeGe (do inglês, *Randomly edge-weighted neighborhood graphs model*).

No Capítulo 3 consideramos um problema de variáveis categóricas com muitos níveis. Estas variáveis são representadas usando uma variável indicadora para cada nível possível da variável categórica inflando o número de parâmetros do modelo. Como conseqüência, temos instabilidade nas estimativas dos coeficientes, gerando os resultados difíceis de interpretar (Criscuolo [2019]). Além da questão da interpretação, a inclusão de variáveis categóricas com muitos níveis em um modelo preditivo facilmente leva a uma matriz esparsa (Bateni et al. [2019]). A consequência pode ser um problema de otimização mal condicionado que resulta em um modelo sobreajustado. Propomos uma abordagem via *shrinkage prior* que é atribuída ao preditor categórico $\boldsymbol{\theta}$ onde seus níveis são correlacionados. Os efeitos de cada nível da variável categórica corresponde aos nós de grafo. As arestas deste grafo conectam níveis associados. Se a variável é nominal temos um grafo completo com todos os nós conectados entre si. Se a variável é ordinal, as arestas conectam níveis subsequentes na ordenação e no caso de uma variável categórica espacial as arestas conectam níveis vizinhos cuja vizinhança é estabelecida pelo mapa geográfico original. Nosso método caracteriza a estrutura de dependência entre variáveis através da matriz de precisão da distribuição *shrinkage prior*. A matriz de precisão proposta expande a matriz do CAR, com dependência de um único parâmetro para uma classe que heterogeneiza a matriz de covariância, convertendo o parâmetro ρ do CAR em um vetor $\boldsymbol{\rho}$ aleatório. A distribuição *a priori* para o vetor $\boldsymbol{\rho}$, carrega a dependência espacial dada pelo grafo de arestas, mas a imposição desta distribuição não gera um gasto computacional, pois implica cálculos algébricos simples, produzindo resultados teóricos e distribuições condicionais *a posteriori* conhecidas. A distribuição marginal do vetor $\boldsymbol{\theta}$ tem uma forma fechada e favorece a esparsidade e o agrupamento.

Ambos os trabalhos contribuem para o desenvolvimento da análise de dados na área e consideram um modelo probabilístico flexível e de fácil extensão.

1.5 Referências

Ozan Arslan and Ozer Akyurek. Spatial modelling of air pollution from pm10 and so2 concentrations during winter season in marmara region. *International Journal of Environment and Geoinformatics*, pages 1 – 16, 2018. doi: 10.30897/ijegeo.412391.

Renato M. Assunção and Elias Krainski. Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, 51(5):851–869, 2009.

Renato M. Assunção, Erica C. Rodrigues, and Elias Krainski. Campos aleatórios de markov e distribuções condicionalmente especificadas. *XIX SINAPE, Simpósio Nacional de Probabilidade e Estatística*, 2010.

Robert G. Aykroyd. Bayesian estimation for homogeneous and inhomogeneous gaussian random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):533–539, 1998.

Sudipto Banerjee, Bradley P. Carlin, and Alan E Gelfand. Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics*, 7(7):45–63, 2003.

Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical Modeling and Analysis of Spatial Data*, volume 101. Routledge, 01 2014. doi: 10.1201/9780203487808.

Mohammad Hossein Bateni, Lin Chen, Hossein Esfandiari, Thomas Fu, Vahab S. Mirrokni, and Afshin Rostamizadeh. Categorical feature compression via submodular optimization. *International Conference on Machine Learning*, 2019.

Luisa Bernadinelli, David Clayton, and Cristina Montomoli. Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, 14:2411–2431, 1995.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36(2):192–236, 1974. ISSN 00359246.

Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B*, 48(3):259–302, 1986.

Julian Besag and D. Higdon. Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B*, 61(4):691–746, 1999. doi: 10.1111/1467-9868.00201.

Julian Besag and Charles Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746, 1995.

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43 (1):1–20, 1991.

Julian Besag, Peter Green, David Higdon, and Kerrie Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, 02 1995.

Mark J. Brewer and Andrew J. Nolan. Variable smoothing in bayesian intrinsic autoregressions. *Environmetrics*, 18(8):841–857, 2007.

Bradley P. Carlin and Sudipto Banerjee. Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics*, 7:45–63, 2003.

Noel A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley Classics Library, 1993. ISBN 9780471002550.

Tulio L. Criscuolo. *Modelo partição produto para atributos categóricos*. PhD thesis, Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, 2019.

Abhirup Datta, Sudipto Banerjee, and James S. Hodges. Spatial disease mapping using directed acyclic graph auto-regressive (dagar) models. *Bayesian Analysis*, 14(8):1221–1244, 2019.

Paul Elliott and Daniel E. Wartenberg. Spatial epidemiology: Current approaches and future challenges. In *Environmental health perspectives*, 2004.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

Andreas Groll, Julien Hambuckers, Thomas Kneib, and Nikolaus Umlauf. LASSO-type penalization in the framework of generalized additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 140:59–73, 2019.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 09 2009. ISSN 0006-3444.

Xiaoping Jin and Bradley P. Carlin. Multivariate parametric spatiotemporal models for county level breast cancer survival data. *Lifetime data analysis*, 11(1):5—27, March 2005. ISSN 1380-7870.

Leonhard Knorr-Held and Nicola G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A*, 164:73–85, 2001.

Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 06 2010.

Andrew B. Lawson and Allan Clark. Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21(3):359–370, 2002.

Duncan Lee. Carbayes: An r package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013. ISSN 1548-7660.

Duncan Lee and Richard Mitchell. Boundary detection in disease mapping studies. *Biostatistics (Oxford, England)*, 13:415–26, 10 2011.

Brian G. Leroux, Xingye Lei, and Norman Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. pages 179–191, 2000.

James P. Lesage and Kelly R. Pace. Introduction to spatial econometrics. crc press, boca raton, fl. *Introduction to Spatial Econometrics*, 1, 2009. doi: 10.1201/9781420064254.

Qing Li and Nan Lin. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 03 2010.

Fei Liu, Sounak Chakraborty, Fan Li, Yan Liu, and Aurelie C. Lozano. Bayesian regularization via graph laplacian. *Bayesian Analysis*, 9(2):449–474, 06 2014.

H. Lu, C. Reilly, Bradley P Carlin, and Sudipto Banerjee. Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14:433–452, 2007.

Ying MacNab and C. B. Dean. Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in Medicine*, 19:2421–2435, 2000.

Miguel A. Martínez-Beneito, Antonio López-Quilez, and Paloma Botella-Rocamora. An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27 (15):2874–2889, 2008.

Marcos O. Prates, D. Kumar Dey, and Victor H. Lachos. A dengue fever study in the state of rio de janeiro with the use of generalized skew-normal/independent spatial fields. *Chilean Journal of Statistics.*, 3(2):143–155, 09 2012.

Brian J Reich, James S Hodges, and Vesna Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

Erica C. Rodrigues and Renato M. Assunção. Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, 109:88 – 102, 2012. ISSN 0047-259X.

Håvard. Rue and Leonhard. Knorr-Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 2005.

Yiyuan She. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4: 1055–1096, 2010.

Giovani L. Silva, C. B. Dean, Théophile Niyonsenga, and Alain Vanasse. Hierarchical bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine*, 27:2381–2401, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, February 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00490.x.

Melanie M. Wall. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121(2):311–324, 2004. ISSN 0378-3758.

2 Inducing high spatial correlation with randomly edge-weighted neighborhood graphs

2.1 Abstract

Traditional models for areal data assume a hierarchical structure where one of the components the random effects that spatially correlate the areas. The conditional autoregressive (CAR) model is the most popular distribution to jointly model the prior uncertainty about these spatial random effects. One limitation of the CAR distribution is the inability of producing high correlations between neighboring areas. We propose a robust model for areal data that alleviates this problem. We represent the map by an undirected graph where the nodes are the areas and randomly-weighted edges connect nodes that are neighbors. The model is based on a multivariate Student-t distribution, spatially structured, in which the precision matrix is indirectly built assuming a multivariate distribution for the random edges. The weights' joint distribution is a spatial multivariate Student-t that induces another t distribution for the areas' spatial effects, which inherit its capacity to accommodate outliers and heavy-tail behavior. Most importantly, it can produce a higher marginal correlation between the spatial effects than the CAR model overcoming one of the main limitations to this model. We fit the proposed model to analyze real cancer maps and compared its performance with several state-of-art competitors. Our proposed model provides better fitting in almost all cases.

Keywords: Graph of edges, Spatial correlation, Student-t distribution.

2.2 Introduction

The conditional autoregressive (CAR) model introduced by Besag [1974] has been one of the main drivers of spatial models for area or lattice data. It appeared again in Besag et al. [1991] in its intrinsic version (ICAR), gaining visibility and importance as the main framework to specify joint distributions through the set of the conditional distribution of each area given its neighbors [Martínez-Beneito and Botella-Rocamora, 2019]. The spatially motivated Markov property enjoyed by this model, coupled with its computational ease for Bayesian analysis and the availability of fast computation [Gelfand

and Vounatsou, 2003], are responsible for its large appeal. Typically, the CAR and ICAR models are used to describe the joint behavior of random effects associated with each small area on a map. Such effects are latent factors representing the spatial dependence beyond the small area geographical boundary. A key point regarding the construction of CAR or ICAR models is the specification of an appropriate neighborhood structure. The usual is to take as neighbors any pair of areas sharing boundaries. This approach is popular because it can be easily calculated using GIS (Geographic Information System) routines. The Markov property is defined in terms of this neighborhood structure inducing a sparse precision matrix. This strategy facilitates Bayesian computational approaches.

Spatial statisticians extended the CAR and ICAR models in many different directions. Classes of space–time generalized linear models are proposed by Knorr-Held and Best [2001], MacNab and Dean [2000], Martínez-Beneito et al. [2008] and Silva et al. [2008]. Carlin and Banerjee [2003] and Jin and Carlin [2005] extend the idea to model spatial survival data. Spatially-varying parameters models are introduced by Assunção [2003], Assunção et al. [2002] and Gelfand et al. [2003] and generalized additive models can be found in Fahrmeir and Lang [2001]. Extensions incorporating two correlated sets of spatial effects are proposed by Jin and Carlin [2005], [Gelfand and Vounatsou, 2003] and Knorr-Held et al. [2006].

Notwithstanding its popularity, there exist several critical points related to CAR and ICAR models [Martínez-Beneito and Botella-Rocamora, 2019, p. 134]. In one line of criticism, Wall [2004] showed that there are many puzzling results involving the CAR model. For example, the correlation between any pair of neighboring areas is negatively associated with the number of neighbors of each region but this is not sufficient to explain the dependence structure. Besides is that sites with equal numbers of neighbors have different variances. Even more puzzling, the spatial structure depends on the CAR spatial parameter ρ in an unexpected way: a pair of areas more correlated than another one if, for instance, $\rho = 0.5$ may become less correlated for some other value of ρ different of 0.5. The ICAR model may exacerbate these problems. Wall [2004] concluded that the spatial correlation induced by the CAR model is not intuitive and does not follow a practical scheme. All these counterintuitive results were explained away by Assunção and Krainski [2009]. They showed that the correlation structure between two areas depends on the entire neighborhood graph structure, and not only their immediate neighborhood. A crucial role in how these puzzling results appear is played by the second largest eigenvalue modulus of the neighborhood matrix used in the CAR or ICAR models. A more serious concern is the lack of ability of CAR or ICAR spatial effects models to produce high pairwise spatial correlation even when the parameter ρ is near or equal to 1, as in the ICAR (Gelfand and Vounatsou, 2003).

Another approach to model area data can be found in Leroux et al. [2000] and MacNab and Dean [2000]. These models assume that the precision matrix is a linear combination of a diagonal matrix and the precision matrix of the ICAR model. They accommodate over-dispersion but inherit the lack of interpretability issue of the CAR model. More recent developments in this topic include Prates et al. [2012], Rodrigues and Assunção [2012], and Datta et al. [2019]. The CAR structure may not be appropriate to describe the spatial correlation if some areas experience atypical spatial effects. In Prates et al. [2012], the normal distribution involved in the CAR model is replaced by distributions in the generalized skew-normal/independent class, a robust class of distributions that is able to simultaneously accommodate heavy-tail and asymmetry. Rodrigues and Assunção [2012] proposes a spatial model in which the neighborhood structure is a parameter that must be estimated. This model preserves the Markov property as it assumes that, given the neighborhood graph, the areal parameters follow a conditional autoregressive model. The directed acyclic graph autoregressive (DAGAR) model [Datta et al., 2019] constructs the spatial precision matrix considering a directed acyclic graph derived from the original undirected graph associated with the map. DAGAR model provides a different approach to model multivariate Gaussian data always providing a positive definite covariance matrix. Its use is particularly appealing to analyze large spatial datasets due to the sparsity induced in such matrix.

We will focus on spatial data that may be represented by an undirected graph. Our goal is to build a model (Section 2.4) able to alleviate one of the main constraints of CAR-based models: its incapacity of generating high marginal correlations. As usual, we represent the map with a graph where the nodes stand for the small areas and edges link geographically neighboring areas. The novelty in our approach is that we assign spatial random effects to the *edges* of this neighborhood graph. The spatial random effect of each area is the sum of its incident edges' effects. The joint distribution for the incident edges' effects is a multivariate Student-t distribution where the spatial covariance matrix has a CAR-like structure. The resulting spatial model for the area effects (randomly edge-weighted neighborhood graphs (RENeGe model) is a multivariate Student-t distribution inheriting the heavy-tail behavior and robustness to outliers. More importantly, it induces a higher marginal correlation than the CAR model, overcoming one of the main limitations of the CAR and ICAR models. We study the marginal and conditional correlation properties of this model and a deep look at the partial correlation is provided in Section 2.5. The proposed model is applied to the analysis of many cancer maps, and its performance is compared to six other alternative spatial models (Section 2.6). This comparison shows that it is a competitive model, providing the best fit in almost all cases. We start presenting some preliminar concepts that are useful for the model definition (Section 2.3).

2.3 Preliminary Definitions

Consider a map with n contiguous geographical regions. The map is identified with an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of vertices or nodes representing the areas and $\mathcal{E}$ is the set of p edges connecting unordered pairs of distinct vertices and representing the adjacency relationship among regions. The edge connecting $v_i \in \mathcal{V}$ and $v_j \in \mathcal{V}$ is alternatively represented by $[ij]$ or (v_i, v_j). We assume that the edges are undirected, implying that $[ij] = [ji]$. If two nodes v_i and $v_j \in \mathcal{V}$ are connected, this will be denoted by $v_i \sim v_j$. When $v_i \in \mathcal{V}$ is a node in the edge $[ij] \in \mathcal{E}$ we say that the edge is incident on v_i. Usually, the graph is visualized by plotting each vertex on a typical spatial location inside the corresponding area, such as its geographical centroid. The spatial neighbourhood structure is represented by the set $\mathcal{E}$ of edges. These edges will determine the stochastic dependence between the areas. The most common choice is to have an edge $[ij] = (v_i, v_j)$ when areas i and j share boundaries but other neighbourhood structures may be assumed. A *path* from node v_1 to node v_m is a set of different nodes $v_1, v_2, \ldots, v_m$ in $\mathcal{V}$ which are connected by edges $(v_1, v_2), \ldots, (v_{m-1}, v_m)$ where $v_i \neq v_{i+1}$, for $i = 1, \ldots, n$, except, possibly, by the initial and final vertices. A path with $v_1 = v_m$, is called a *circuit*. A graph is said to be *connected* if, for any pair of nodes v_i and v_j, there is at least one path connecting them. Although it is not strictly necessary, we assume that the graph is connected. We also assume the most common situation in practice, that the number p of edges is larger than the number n of nodes. The *adjacency matrix* $\boldsymbol{A}_v$ of $\mathcal{G}$ is a $n \times n$ binary matrix representing the neighborhood structure. That is, $\boldsymbol{A}_v(i, j) = a_{ij} = 1$, if $v_i \sim v_j$ (or, equivalently, if $[ij] \in \mathcal{E}$), and it is 0, otherwise. We assume that $a_{ii} = 0$ for all $i \in \mathcal{V}$, with $\mathcal{V} = \{\nu_1, \ldots, \nu_n\}$.

Associated with the original graph $\mathcal{G}$, we define the *graph of edges* $\mathcal{L}(\mathcal{G})$ that is a fundamental tool for our purposes. The graph of edges represents the adjacency relationship among the edges of the original graph $\mathcal{G}$. The nodes in $\mathcal{L}(\mathcal{G})$ are the edges $[ij] \in \mathcal{E}$ connecting the nodes v_i and v_j, with $i \neq j$. The edges in $\mathcal{L}(\mathcal{G})$ are also determined by the topology of $\mathcal{G}$. Two nodes $[ij]$ and $[kl]$ in $\mathcal{L}(\mathcal{G})$ are adjacent if, and only if, the edges $[ij]$ and $[kl]$ are incident on a common vertex. This means that the pair of neighbouring edges must be of the form $[ij]$ and $[jk]$ for some $v_j \in \mathcal{V}$. Let $\mathcal{I}_i = \{[ik] \in \mathcal{E}, v_k \in \mathcal{V}\}$ be the set of edges incident on area i. The adjacency matrix $\boldsymbol{A}_e$ associated with the graph of edges $\mathcal{L}(\mathcal{G})$ is a $p \times p$ matrix defined based on the edges' neighborhood structure in $\mathcal{L}(\mathcal{G})$. That is, $\boldsymbol{A}_e([ij],[jk]) = a_{[ij][jk]} = 1$, if $[ij]$ and $[jk]$ belong to $\mathcal{I}_j$. Otherwise, $\boldsymbol{A}_e([ij],[jk]) = 0$. Finally, the *incidence matrix* $\boldsymbol{C}$ associated with $\mathcal{G}$ is a $n \times p$ binary matrix such that $c_{ie} = 1$ if edge e is incident on node i, and $c_{ie} = 0$, otherwise. Figures 1-3 present a toy example of the graph of edges $\mathcal{L}(\mathcal{G})$.

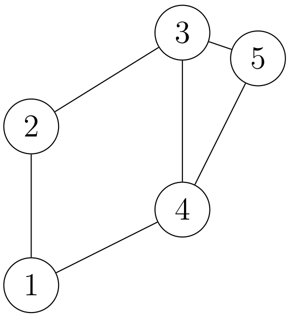


Figure 1 – $\mathcal{G}$

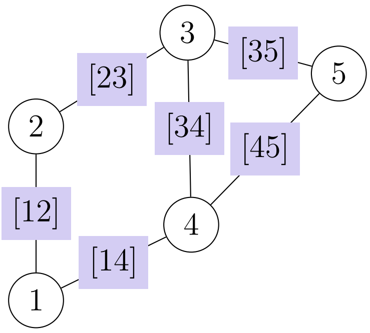


Figure 2 – $\mathcal{E}$

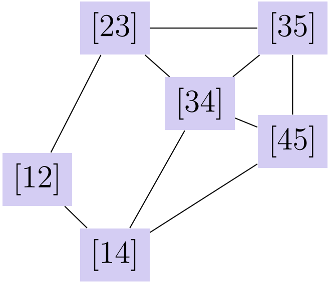


Figure 3 – $\mathcal{L}(\mathcal{G})$ edges.

2.4 Randomly edge-weighted neighborhood graphs model RENeGe for spatial random effects

In a map partitioned into n contiguous regions, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a random vector where θ_i is a random variable associated with the i-th region. Assume that the coordinates in $\boldsymbol{\theta}$ are spatially correlated. The vector $\boldsymbol{\theta}$ may represent the spatial effects in a hierarchical model. The goal is to model the uncertainty about $\boldsymbol{\theta}$ accounting for this spatial correlation. Differently from the CAR-related models, we build the prior distribution for $\boldsymbol{\theta}$ in such a way that their correlation is induced by the prior correlation in spatial effects assigned to the neighbourhood graph edges. The novelty of our approach is the use of a distribution over the pairs of neighbouring areas to induce a distribution over individual areas.

We look for a model able to generate a stronger marginal correlation between neighboring areas, overcoming one of the main limitations of the CAR model.

2.4.1 Modeling the edges' random effects

Assume that our map is represented by the connected graph $\mathcal{G}$ such that each component of $\boldsymbol{\theta}$ is a node in $\mathcal{G}$. There are p undirected edges connecting pairs of neighboring nodes in $\mathcal{G}$. Let $\rho_{[ij]} \in \mathbb{R}$ be a random variable or weight associated with the edge $[ij]$ connecting nodes θ_i and θ_j. The vector of such weights is $\boldsymbol{\rho} \in \mathbb{R}^p$.

Spatial effects observed in area data are surrogates for unknown or unobserved factors that vary on a scale extending beyond the geographical boundaries of the small areas. As these effects spread throughout the surrounding neighborhood of an area, they can be viewed acting on the edges that connect neighboring areas. Instead of directly modeling the θ_i effects in each small area, we can obtain them as a result of a random effects model for the neighborhood graph edges. Two edges, $[ij]$ $[ik]$, incident on the same i-th node, should be correlated due to the extrapolating scale of the hidden factors. By defining θ_i as a function of the correlated weights $\rho_{[ij]}$, we will induce spatial correlations between the areas. Differently from CAR, this definition will allow for i and j neighboring areas to be more strongly correlated than in the CAR or ICAR models. This stronger correlation will affect also pairs of areas that are not directly connected in $\mathcal{G}$.

Assume that the edges weights vector $\boldsymbol{\rho} \in \mathbb{R}^p$ has the centered p-variate normal distribution,

$$\boldsymbol{\rho}|\boldsymbol{S} \sim N(\mathbf{0}, (\nu - p - 1)\boldsymbol{S}), \tag{2.1}$$

where $(\nu - p - 1)\boldsymbol{S}$ is an unknown $p \times p$ covariance matrix representing the correlation among the edges weights and $\nu > p + 1$. The uncertainty about $\boldsymbol{S}$ is modeled with an Inverse-Wishart (IW) distribution. Besides the flexibility to model positive-definite matrices, the Inverse-Wishart distribution has the additional advantage of its conjugacy with the normal distribution. We represent the dependence structure of the components in $\boldsymbol{\rho}$ through the graph of edges $\mathcal{L}(\mathcal{G})$ and its associated adjacency matrix $\boldsymbol{A}_e \in \mathbb{R}^p \times \mathbb{R}^p$. Let $\boldsymbol{M}_e = \text{diag}(m_1, \ldots, m_p)$ be the diagonal matrix where m_k is the number of edges in $\mathcal{L}(\mathcal{G})$ that are neighbors of edge k or, equivalently, the sum of the k-th row elements of $\boldsymbol{A}_e$. We center the distribution of the precision matrix $\boldsymbol{S}^{-1}$ around a sparse matrix to induce conditional independence between unlinked edges, that is, we assume that $\mathbb{E}(\boldsymbol{S}^{-1})$ is proportional to $(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)$. More specifically, we let

$$\boldsymbol{S} \sim IW_p(\nu, \tau_{\boldsymbol{\theta}}^{-1}(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)^{-1}), \tag{2.2}$$

where $\tau_{\boldsymbol{\theta}} > 0$ is precision parameter, γ is a constant assuming values in the interval $(1/\lambda_p, 1/\lambda_1)$ and λ_1 and λ_p are, respectively, the minimum and the maximum eigenvalues of $\boldsymbol{M}_e^{-1/2}\boldsymbol{A}_e\boldsymbol{M}_e^{-1/2}$. This constraint over γ guarantee that the matrix $(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)$ is a positive-definite matrix. It follows from (2.1) and (2.2) that $\boldsymbol{\rho}$ has a centered p-variate Student-t distribution with escale parameter $\tau_{\boldsymbol{\theta}}^{-1}(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)^{-1}(\nu - p - 1)/(\nu - p + 1)$ and

$\nu - p + 1$ degree of freedom and with density

$$\begin{aligned} p(\boldsymbol{\rho}) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu-p+1}{2}\right)\pi^{p/2}(\nu-p+1)^{p/2}} \mid \tau_{\boldsymbol{\theta}}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e) \mid^{1/2} \\ &\times \left[1 + \boldsymbol{\rho}^t \frac{\tau_{\boldsymbol{\theta}}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)}{\nu - p + 1} \boldsymbol{\rho}\right]^{-(\nu+1)/2}. \end{aligned} \tag{2.3}$$

This distribution is denoted by $\boldsymbol{\rho} \sim \mathrm{t}_p(\mathbf{0}, \tau_{\boldsymbol{\theta}}^{-1}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}(\nu - p - 1)/(\nu - p + 1), \nu - p + 1)$. As a consequence, the correlation structure between the edges weights follows the CAR structure:

$$\mathrm{Cov}[\boldsymbol{\rho}] = \tau_{\boldsymbol{\theta}}^{-1}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}. \tag{2.4}$$

An alternative proof for (2.3) is given by results obtained by Iranmanesh et al. [2012]. As the Inverse-Wishart distribution in (2.2) is the inverse matrix variate gamma distribution with parameters $\nu/2$, 2 and $\tau_{\boldsymbol{\theta}}^{-1}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}$, Theorems 2.1 and 3.1 in Iranmanesh et al. [2012] provide that $\boldsymbol{\rho}$ has a p-dimensional generalized multivariate t-distribution with parameters $(\mathbf{0}, \tau_{\boldsymbol{\theta}}^{-1}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}, \nu/2, 2)$, recovering the distribution in (2.3).

2.4.2 The induced distribution for the areas' random effects

The proposed model assumes that the random effect θ_i associated with area i is a linear combination of the effects of edges incident on area i, that is,

$$\theta_i = \sqrt{\frac{\nu - n - 1}{\nu - p - 1}} \sum_{[ik] \in \mathcal{I}_i} \rho_{[ik]} \;\Rightarrow\; \boldsymbol{\theta} = \sqrt{\frac{\nu - n - 1}{\nu - p - 1}}\, \boldsymbol{C}\, \boldsymbol{\rho},$$

where $\boldsymbol{C}$ is the $n \times p$ incidence matrix and $\mathcal{I}_i$ is the set of edges incident on area i. Given $\boldsymbol{S}$, we have from (2.1) that

$$\boldsymbol{\theta} | \boldsymbol{S} \sim N_n(\mathbf{0}, (\nu - n - 1)\boldsymbol{\Sigma}), \tag{2.5}$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^t$. The choice of ν may inflate or deflate the variance of θ_i, being handful to establish more or less informative priors for $\boldsymbol{\theta}$. The prior distribution for $\boldsymbol{\Sigma}$ is thus obtained from the prior distribution in (2.2) and is given by

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^t \sim IW_n(\nu, \tau_{\boldsymbol{\theta}}^{-1}\boldsymbol{C}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}\boldsymbol{C}^t), \tag{2.6}$$

as long as the scale matrix $\tau_{\boldsymbol{\theta}}^{-1}\boldsymbol{C}(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)^{-1}\boldsymbol{C}^t$ is positive definite. As $\gamma \in (1/\lambda_p, 1/\lambda_1)$, this can be proved by assuming Theorem 4.2.1 in Golub and Van Loan [1996]. In this proof, we invoke Lemma 2.17 in Bapat [2014], that guarantees that, as $\mathcal{G}$ is not a bipartite graph, the rank of $\boldsymbol{C}$ is n.

Mixing (2.5) and (2.6), it follows that $\boldsymbol{\theta} \sim \mathrm{t}_n\left(\boldsymbol{0}, \tau_{\boldsymbol{\theta}}^{-1}\boldsymbol{K}\frac{(\nu-n-1)}{(\nu-n+1)}, \nu-n+1\right)$, where $\boldsymbol{K} = \boldsymbol{C}(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)^{-1}\boldsymbol{C}^t$, with density

$$p(\boldsymbol{\theta}) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu-n+1}{2}\right)} \frac{\tau_{\boldsymbol{\theta}}^{n/2}}{\pi^{n/2}(\nu-n-1)^{n/2}} |\boldsymbol{K}|^{-1/2} \left(1 + \boldsymbol{\theta}^T \frac{\boldsymbol{K}^{-1}\tau_{\boldsymbol{\theta}}}{\nu-n-1}\boldsymbol{\theta}\right)^{-(\nu+1)/2}. \tag{2.7}$$

The covariance structure of $\boldsymbol{\theta}$ inherits the CAR-type correlations between the edges' weights $\rho_{[ij]_k}$ in $\mathcal{L}(\mathcal{G})$ transformed by the incidence matrix $\mathbf{C}$ ending up with the symmetric, positive definite matrix given by

$$\mathrm{Cov}[\boldsymbol{\theta}] = \boldsymbol{C}\ \mathrm{Cov}(\boldsymbol{\rho})\ \boldsymbol{C}^t = \tau_{\boldsymbol{\theta}}^{-1}\boldsymbol{C}(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)^{-1}\boldsymbol{C}^t. \tag{2.8}$$

The distribution in (2.7) is known as the generalized n-variate Student-t distribution with location $\mathbf{0}$, scale $\boldsymbol{K}$ and shape parameters $\nu-n+1 > 0$ and $(\nu-n-1)\tau_{\boldsymbol{\theta}}^{-1} > 0$, which is denoted by $\boldsymbol{\theta} \sim \mathrm{T}_n\left(\boldsymbol{0}, \boldsymbol{K}, \tau_{\boldsymbol{\theta}}^{-1}(\nu-n-1), \nu-n+1\right)$. Consequently, we have [Kotz and Nadarajah, 2004, Arellano-Valle and Bolfarine, 1995] that

(i) The marginal distribution of each component θ_i of $\boldsymbol{\theta}$ is $\theta_i \sim T(0, \tau_{\boldsymbol{\theta}}^{-1}(\nu-n-1)\boldsymbol{K}_{ii}, \nu)$, a univariate centered t distribution where $\boldsymbol{K}_{ii}$ is entry that lies in ith line and ith column of matrix $\boldsymbol{K} = \boldsymbol{C}(\boldsymbol{M}_e - \gamma\boldsymbol{A}_e)^{-1}\boldsymbol{C}^t$.

(ii) Let the column vectors $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ define a partition of $\boldsymbol{\theta}$ where $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ have dimensions n_1 and $n_2 = n - n_1$, respectively. Let the matrices $\boldsymbol{K}_{AA}$, $\boldsymbol{K}_{BB}$ and $\boldsymbol{K}_{AB}$ of order $n_1 \times n_1$, $n_2 \times n_2$ and $n_1 \times n_2$, respectively, be the partition of $\boldsymbol{K}$ induced by $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$. The conditional distribution of $\boldsymbol{\theta}_A|\boldsymbol{\theta}_B$ is the generalize n_1-variate Student-t distribution $\boldsymbol{\theta}_A \mid \boldsymbol{\theta}_B \sim \mathrm{T}_{n_1}\left(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star, \tau_{\boldsymbol{\theta}}^{-1}(\nu-n-1) + \boldsymbol{\theta}_B^t\boldsymbol{K}_{BB}^{-1}\boldsymbol{\theta}_B,\ \nu - n_1 + 1\right)$, where $\boldsymbol{\mu}^\star = \boldsymbol{K}_{AB}\boldsymbol{K}_{BB}^{-1}\boldsymbol{\theta}_B$ and $\boldsymbol{\Sigma}^\star = \boldsymbol{K}_{AA} - \boldsymbol{K}_{AB}\boldsymbol{K}_{BB}^{-1}\boldsymbol{K}_{BA}$.

2.4.3 The covariance structure of θ

By assuming a CAR-type structure for $\boldsymbol{\rho}$, the correlation among the edges weights inherits the counterintuitive results pointed out by Wall [2004] and Assunção and Krainski [2009]. As $\mathrm{Cov}(\boldsymbol{\theta}) = \boldsymbol{C}\ \mathrm{Cov}(\boldsymbol{\rho})\ \boldsymbol{C}^t$, these results are passed on to the covariance structure of $\boldsymbol{\theta}$. Our goal in this section is to investigate the impact of the covariance of $\boldsymbol{\rho}$ in the marginal $\mathrm{Cov}(\theta_i, \theta_j)$. For simplification, we set $\tau_{\boldsymbol{\theta}}^{-1} = 1$ in this section. It is not possible to obtain the exact value for $\mathrm{Cov}(\theta_i, \theta_j)$ as it requires the calculation of the inverse matrix $(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}$. To investigate such an impact, we consider the relationship

$$(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1} = (\boldsymbol{I} - \gamma \boldsymbol{W}_e)^{-1} \boldsymbol{M}_e^{-1} \tag{2.9}$$

where $\boldsymbol{W}_e$ is a $p \times p$ matrix whose entries are given by $\boldsymbol{W}_{e[ik][jl]} = a_{[ik][jl]}/d_{[ik]}$. The binary variable $a_{[ik][kj]}$ is equal to 1 if, and only if, edges $[ik]$ and $[kj]$ are edges incident on a same node k, $a_{[ik][ik]} = 0$ for all edges $[ik]$, and $d_{[ik]}$ denotes the total number of edges in $\mathcal{L}(\mathcal{G})$ that are neighbors of edge $[ik]$.

The matrix $\boldsymbol{W}_e$ is a transition matrix of a random walk defined on the graph of edges $\mathcal{L}(\mathcal{G})$. Its elements are non-negative, its rows sum up to 1, and the diagonal elements are equal to zero. A particle sitting in an edge $[ij]$ at time t moves to a different edge randomly selected among those in $\mathcal{I}_i$, the neighbors of $[ij]$ in $\mathcal{L}(\mathcal{G})$, with equal probability. The power matrix $\boldsymbol{W}_e^s$ represents the transition probabilities in s steps for this Markov chain. The non-zero elements in the $[ij]$-th row of $\boldsymbol{W}_e^s$ indicate the other edges to which one may visit s steps ahead. That is, $\boldsymbol{W}_{e[ik][jl]}^s > 0$ if there is at least one path composed by s edges in $\mathcal{L}(\mathcal{G})$ linking the initial edge $[ik]$ to the final edge $[jl]$.

For the matrix $\boldsymbol{W}_e$ to be ergodic and aperiodic, $\mathcal{L}(\mathcal{G})$ must be a connected graph. As we assume that $\mathcal{G}$ is a connected graph, the graph of edges $\mathcal{L}(\mathcal{G})$ is also connected. Furthermore, $\boldsymbol{W}_e$ and $\boldsymbol{M}_e^{-1/2} \boldsymbol{A}_e \boldsymbol{M}_e^{-1/2}$ are similar matrices and, therefore, they share the same eigenvalues. Thus, for $\gamma \in (1/\lambda_p, 1/\lambda_1)$, the matrix $(\boldsymbol{I} - \gamma \boldsymbol{W}_e)$ is non-singular and, from results in Assunção and Krainski [2009], its inverse is given by

$$(\boldsymbol{I} - \gamma \boldsymbol{W}_e)^{-1} = \boldsymbol{I} + \gamma \boldsymbol{W}_e + \gamma^2 \boldsymbol{W}_e^2 + \gamma^3 \boldsymbol{W}_e^3 + \cdots . \tag{2.10}$$

Replacing the result in (2.10) in equation (2.8), we have that

$$\mathrm{Cov}[\boldsymbol{\theta}] = \boldsymbol{C}(\boldsymbol{I} + \gamma \boldsymbol{W}_e + \gamma^2 \boldsymbol{W}_e^2 + \gamma^3 \boldsymbol{W}_e^3 + \cdots) \boldsymbol{M}_e^{-1} \boldsymbol{C}^t.$$

For $i \sim j$, the element $\mathrm{Cov}(\theta_i, \theta_j)$ in this matrix is given by

$$\frac{1}{d_{[ij]}} + \gamma \sum_{[jl] \in \mathcal{I}_j} \frac{1}{d_{[jl]}} \left(\sum_{\substack{[ik] \in \mathcal{I}_i \\ [ik] \sim [jl]}} \frac{1}{d_{[ik]}} \right) + \gamma^2 \sum_{[jl] \in \mathcal{I}_j} \frac{1}{d_{[jl]}} \left(\sum_{[ik] \in \mathcal{I}_i} \frac{1}{d_{[ik]}} \sum_{\substack{[lr] \sim [ik] \\ [lr] \sim [jl]}} \frac{1}{d_{[lr]}} \right) + \ldots \tag{2.11}$$

while, if $i = j$, we have $\mathbb{V}(\theta_i) = \mathrm{Cov}(\theta_i, \theta_i)$ given by

$$\sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} + \gamma \sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} \left(\sum_{\substack{[ik]\in\mathcal{I}_i \\ [ik]\sim[ij]}} \frac{1}{d_{[ik]}} \right) + \gamma^2 \sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} \left(\sum_{[ik]\in\mathcal{I}_i} \frac{1}{d_{[ik]}} \sum_{\substack{[lr]\sim[ik] \\ [lr]\sim[ij]}} \frac{1}{d_{[lr]}} \right) + \dots$$

For pairs of areas that are not adjacent, the formula is more convoluted, summing over all the paths connecting the two areas. The proof can be found in the supplementary material B.1.

The expansion in expression (2.11) shows that the correlation structure between θ_i and θ_j cannot be explained considering only the interaction between first-order neighbors. The $\mathrm{Cov}(\theta_i, \theta_j)$ is a polynomial in γ where the k-th order coefficient is a weighted sum of all paths of order k connecting edges $[ik]$ and $[jl]$. If we consider only a first order approximation, $\mathrm{Cov}(\theta_i, \theta_j) \approx d_{[ij]}^{-1}$, inversely proportional to the number $d_{[ij]}$ of edges in $\mathcal{L}(\mathcal{G})$ that are neighbors of edge $[ij]$. Consequently, it is inversely proportional to the total number of neighbors in $\mathcal{G}$ of nodes θ_i and θ_j since $d_{[ij]} = d_i + d_j - 2$ where d_s is the number of neighbors of node θ_s, $s = i, j$. Including, for instance, the third term, the product $a_{[ik][uv]}a_{[uv][jl]}$ is equal to 1 only if the edges $[ik]$ and $[jl]$ are 2nd-order neighbors in $\mathcal{L}(\mathcal{G})$. That implies that the edge $[uv]$ connects nodes i and j in $\mathcal{G}$, establishing a more complex dependence structure between the related nodes in the original graph $\mathcal{G}$ by imposing that (i) nodes k and j are 1st-order neighbor, (ii) nodes k and l and nodes i and j are 2nd-order neighbors and (iii) and nodes i and l are 3rd-order neighbors.

2.4.4 The regular lattice case

We obtain additional properties of our model by considering that the graph $\mathcal{G}$ is a regular lattice or rectangular grid with n^2 nodes symmetrically wrapped into a torus. In this simpler structure, the number of first-order neighbors of each node θ_i is constant and equal to four (see Figure 4). To simplify the expressions, we take $\tau_{\boldsymbol{\theta}} = 1$. Denote by $\mathrm{Cov}(\rho_{[i*]} \overset{n}{\sim} \rho_{[j*]})$, for all $[i*] \in \mathcal{I}_i$ and $[j*] \in \mathcal{I}_j$, the covariance between edges weights $\rho_{[i*]}$ and $\rho_{[j*]}$ whenever the edges $[i*]$ and $[j*]$ are nth-order neighbor in $\mathcal{L}(\mathcal{G})$. The index n is omitted from the notation for the first-order neighbor case.

If $\mathcal{G}$ is a regular lattice, then the graph of edges $\mathcal{L}(\mathcal{G})$ is also a regular lattice. The covariance between any pair of nodes θ_i and θ_j in $\mathcal{G}$ depends on the weights related to the edges belonging to $\mathcal{I}_i$ and $\mathcal{I}_j$:

$$\mathrm{Cov}(\theta_i,\theta_j)=\begin{cases}\mathrm{Var}(\rho_{[ij]})+6\ \mathrm{Cov}(\rho_{[i*]}\sim\rho_{[j*]})+9\ \mathrm{Cov}(\rho_{[i*]}\overset{2}{\sim}\rho_{[j*]}), & \text{if } i\sim j,\\ \mathrm{Cov}(\rho_{[i*]}\overset{n}{\sim}\rho_{[j*]})+6\mathrm{Cov}(\rho_{[i*]}\overset{n+1}{\sim}\rho_{[j*]})+9\mathrm{Cov}(\rho_{[i*]}\overset{n+2}{\sim}\rho_{[j*]}), & \text{if } i\overset{n}{\sim}j.\end{cases} \tag{2.12}$$

As $\boldsymbol{\rho}$ follows a CAR model, $\mathrm{Cov}(\rho_{[i*]}\overset{n}{\sim}\rho_{[j*]})$ decreases with the neighboring order n. Hence, we also have $\mathrm{Cov}(\theta_i,\theta_j)$ decreasing in our model. A direct consequence of (2.12), the marginal correlation between the first-order neighbor θ_i and θ_j is given by

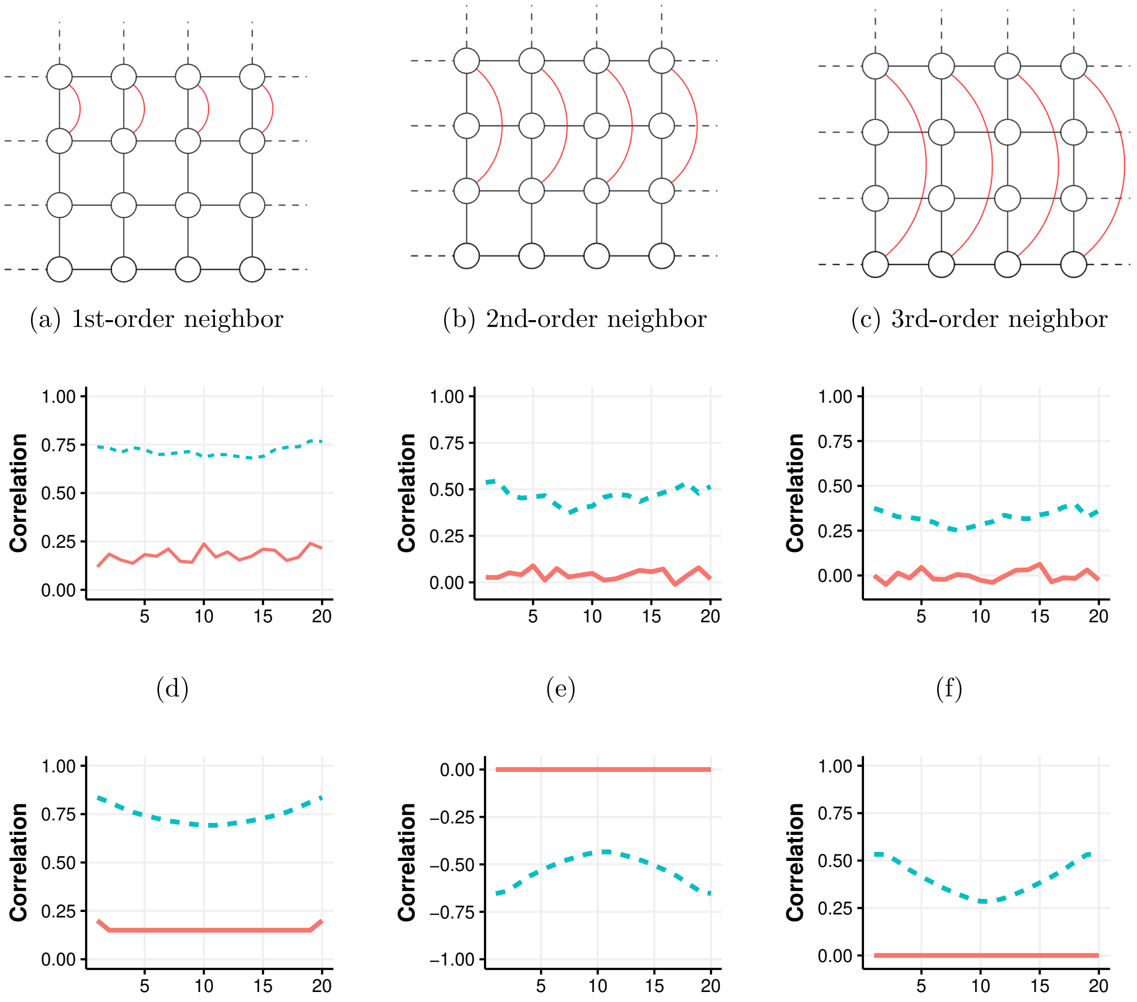
$$\mathrm{Corr}(\theta_i,\theta_j)=\frac{\mathrm{Var}(\rho_{[ij]})+6\mathrm{Cov}(\rho_{[i*]}\sim\rho_{[j*]})+9\mathrm{Cov}(\rho_{[i*]}\sim\rho_{[j*]})}{4\mathrm{Var}(\rho_{[ij]})+12\mathrm{Cov}(\rho_{[i*]}\sim\rho_{[j*]})}. \tag{2.13}$$

The marginal correlations differ from that produced by the CAR model. Figure 4 compare such correlations for first, second and third-order neighbors taking an inner line in a regular lattice with 400 nodes. The correlations are obtained assuming the covariance matrix $\boldsymbol{C}(\boldsymbol{M}_e-\gamma\boldsymbol{A}_e)^{-1}\boldsymbol{C}^t$ with $\gamma=0.8$ under the proposed model. In the CAR model, we assume $(\boldsymbol{D}_v-\rho\boldsymbol{A}_v)^{-1}$ with $\rho=0.8$.

The proposed model produces a higher (in absolute value) marginal and conditional correlation between θ_i and θ_j than the CAR model irrespective of their order of neighboring. For first-order neighbors, the marginal correlation is around 0.8 while, under the CAR model, it is below 0.45. As expected, under both models, the correlation is high for nodes near each other and decreases as the distance between the nodes increases. For instance, for 3rd-order neighbors, the CAR model produces correlation approximately null while this marginal correlation is above 0.25 under the proposed model.

Although the marginal correlations under both models are positive for all neighboring order, the proposed model produces a conditional correlation structure between θ_i and θ_j that is negative if these nodes are neighbors of even order (Figure 5). This behavior for the conditional correlation seems puzzling. In order to gain an empirical view, we used a one-dimensional temporal grid in Figure 6. We simulated θ_i following our model where each θ_i is the sum of the antecedent and the subsequent edges' effects, that is, $\theta_i=\rho_{[i-1,i]}+\rho_{[i,i+1]}$. We assume the parameter values $\gamma=0.8$ and $\tau_\theta=1$.

The typical realization is seen in the top graph of Figure 6. We deleted the pair (θ_i,θ_j) from this particular realization and kept all the other values fixed. Then we simulated a large number of (θ_i,θ_j) values conditioned on all the other values $\boldsymbol{\theta}_{-ij}$ in this time series. In turn, we removed nodes θ_i and θ_j that are neighbors of first, second and third orders ($j=i+1,i+2$, or $i+3$). From the bottom plots in Figure 6, for 2nd-order neighbors, if the generated value θ_i is higher than its conditional mean, then the generated value of θ_{i+2}



(a) 1st-order neighbor

(b) 2nd-order neighbor

(c) 3rd-order neighbor

(d)

(e)

(f)

(g)

(h)

(i)

Figure 4 – Marginal correlations (plots (d), (e), and (f)) and conditional correlations (plots (g), (h), (i)) between θ_i and θ_j assuming first (d,g), second (e,h) and third (f,i) neighboring order, under the proposed (blue dashed line) and the CAR (red solid line) models.

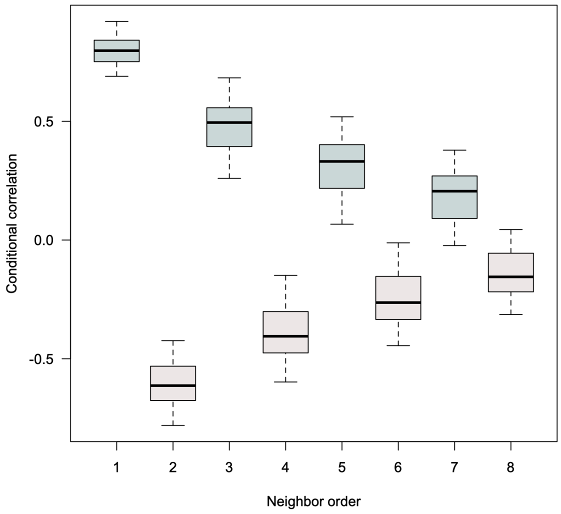


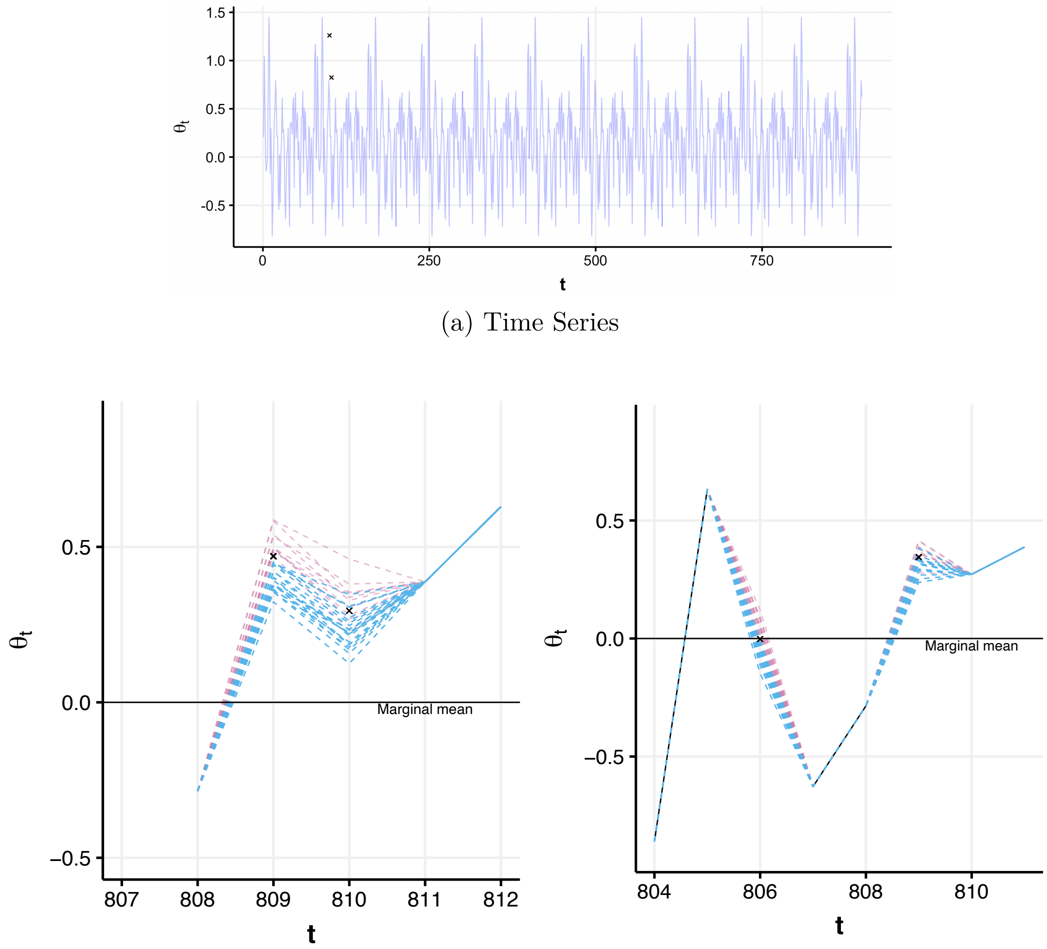
Figure 5 – Conditional correlations for neighboring of order 1 to 9.

tends to be smaller than its conditional mean. For first $(j = i + 1)$ and third $(j = i + 3)$ order neighbors, when the generated value θ_i is higher than its conditional mean, the generated value of θ_j tends to be higher than its conditional mean.

As γ is positive, another interesting characteristic of the proposed model given by expression (2.11), is that the correlation increases monotonically with the growth of γ. However, the increasing rates differ depending on the path connecting the nodes. Let $i \to j$ denote that nodes i and j are separated by a common neighbor and are in the same horizontal or vertical straight line in the grid. Denote by $i \neg j$ a pair of nodes separated by a single common neighbor but with a connecting path that is not a straight line.

Figure 7 shows the evolution of the terms $\gamma^k \boldsymbol{C}\boldsymbol{W}^k\boldsymbol{D}^{-1}\boldsymbol{C}^t$ in (2.4.3) and its cumulative sums $\sum_{j=1}^{k} \gamma^j \boldsymbol{C}\boldsymbol{W}^j\boldsymbol{D}^{-1}\boldsymbol{C}^t$ for paths of different orders k and taking $\gamma = 0.4$ and $\gamma = 0.9$. As this is a convergent series, the first terms in (2.11) are more relevant to obtain a good approximation for $\mathrm{Cov}(\theta_i, \theta_j)$. The relevance degree depends on the type of paths connecting the nodes. The influence of the higher order neighbors in $\mathrm{Cov}(\theta_i, \theta_j)$ increases with γ. Their influence decays more slowly, requiring more terms from (2.11) to get a better approximation the true covariance.

Considering a third-degree approximation under the proposed model, it follows



(a) Time Series

(b) 1st order

(c) 3rd order

Figure 6 – Generate time series (a), the conditional expectations (black dot) and the generate values (dashed lines) removing observations that are neighbors of first (b), second (c) and third (d) orders.

from (2.11) that

$$\mathrm{Cov}(\theta_i, \theta_j) \approx \begin{cases} d_{[ij]}^{-1} + \gamma \sum_{[lm]\sim[ij]} (d_{[ij]} d_{[lm]})^{-1} + \gamma^2 \sum_{\mathcal{C}_{ij}} (d_{[ij]} d_{[lm]} d_{[nk]})^{-1} & \text{if } \ i \sim j, \\ \gamma \sum_{[ik]\sim[kj]} (d_{[ik]} d_{[kj]}) + \gamma^2 \sum_{\mathcal{C}_{ij}} (d_{[ij]} d_{[lm]} d_{[nk]})^{-1} & \text{if } \ i \neg j \\ \gamma \sum_{[ik]\sim[kj]} (d_{[ik]} d_{[kj]}) + \gamma^2 \sum_{\mathcal{C}_{ij}} (d_{[ij]} d_{[lm]} d_{[nk]})^{-1} & \text{if } \ i \to j, \end{cases}$$

where $\mathcal{C}_{ij}$ denotes all paths with three edges connecting nodes i and j. Figure 8 shows how the correlation between θ_i and θ_j evolves for different values for γ and compares it with the CAR(ρ) model for different values for ρ and different neighboring structures. In the proposed model, the relationship of the correlation with the spatial parameter γ has a similar role as ρ in the CAR model. However, under our model, the correlation between

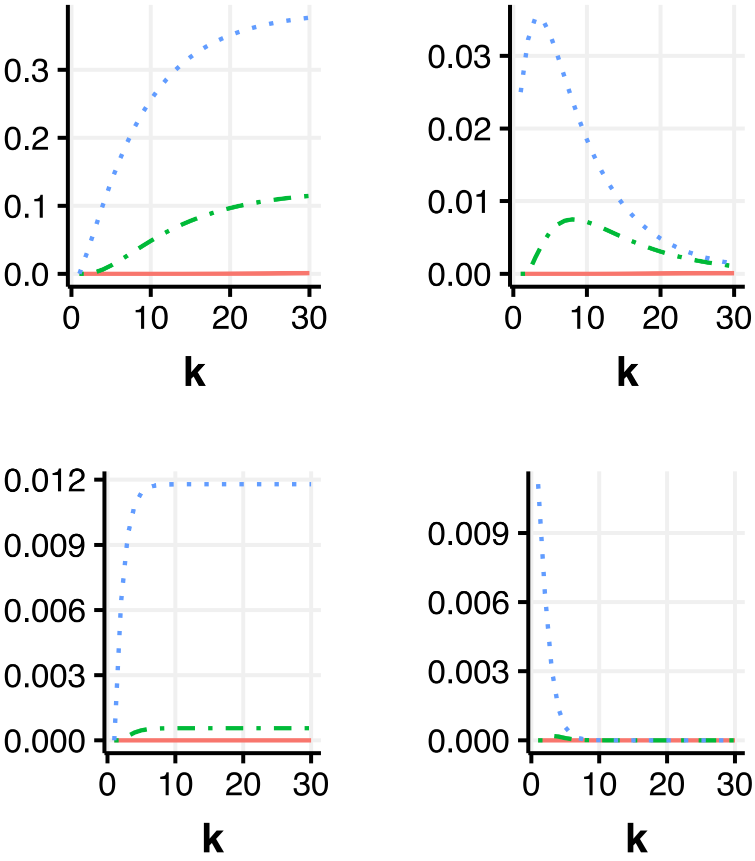


Figure 7 – Cumulative summation $\sum_{j=1}^{k} \gamma^j \boldsymbol{C}\boldsymbol{W}^j\boldsymbol{D}^{-1}\boldsymbol{C}^t$ and the values of $\boldsymbol{C}\gamma^k\boldsymbol{W}^k\boldsymbol{D}^{-1}\boldsymbol{C}^t$ for different neighboring orders k and path types, $\gamma = 0.9$ and 0.4 ($i \sim j$ red solid line, $\rightarrow j$ green dot-dashed line and $i \neg j$ blue dotted line).

neighboring areas is higher for any value of γ. This behavior is explained by the stronger influence of the higher order neighbors in our model. This influence is more noticeable for first-order neighbors nodes.

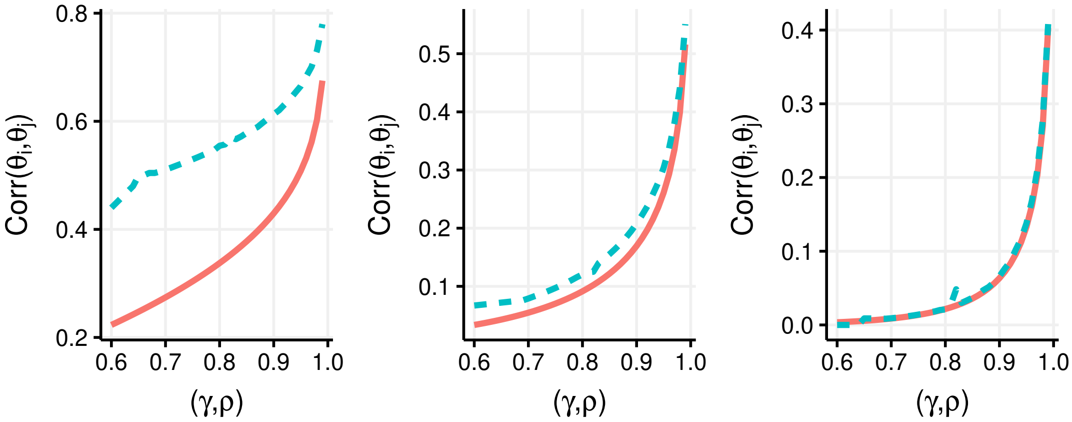


Figure 8 – Correlation between first order (left), second order (middle) and third order (right) neighbors under the proposed (blue dashed line) and CAR (red solid line) models for different values of γ and ρ.

2.5 A close look at the negative partial correlations

The sign changing phenomenon experienced by the correlation of $(\theta_i, \theta_{i+2k})$, $k = 1, 2, \ldots$, given $\boldsymbol{\theta}_{(-i,i+2k)}$, discussed in the previous section has a mathematical justification if we look at it conditionally on $\boldsymbol{S}$. This requires further explanation, which we provide next. We return to general graphs, not necessarily regular grids. We have $\boldsymbol{\theta} \propto \boldsymbol{C}\boldsymbol{\rho}$ implying that $\theta_i \propto \boldsymbol{c}_i^t \boldsymbol{\rho}$ where $\boldsymbol{c}_i$ is a p dimensional column-vector whose coordinates are the entries in the i-th row of matrix $\boldsymbol{C}$. Let $\mathcal{T}$ be the vector space spanned by the set of vectors $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$. The dimension of the vector space $\mathcal{T}$ is p because the rows of an incidence matrix C are linearly independent if and only if no connected component is bipartite in the graph [Bapat, 2014, p.22]. Define the inner product $< \boldsymbol{c}_i, \boldsymbol{c}_j >= \boldsymbol{c}_i^t \boldsymbol{S}\ \boldsymbol{c}_j$ with the consequent angle definition

$$\beta = \cos^{-1}\left(\frac{< \boldsymbol{c}_i, \boldsymbol{c}_j >}{||\boldsymbol{c}_i||||\boldsymbol{c}_j||}\right).$$

Let $\boldsymbol{C}_{-(i,j)}$ be the $(n-2) \times p$ matrix obtained from $\boldsymbol{C}$ after deleting rows i and j. Without loss of generality, assume that $(\nu - p - 1) = 1$. We denote by $\mathcal{S}_{-(i,j)}$ the subspace spanned by the $n-2$ rows of $\boldsymbol{C}_{-(i,j)}$ and by $\mathcal{S}^{\perp}_{-(i,j)}$ its orthogonal subspace, that is,

$$\mathcal{S}^{\perp}_{-(i,j)} = \left\{\boldsymbol{c} = (c_1, \ldots, c_p)^t :< \boldsymbol{c}, \boldsymbol{s} >= 0, \boldsymbol{s} \in \mathcal{S}_{-(i,j)}, \boldsymbol{c} \in \mathcal{V}\right\}.$$

We have the following theorem:

Teorema 3. *Let $\mathcal{S}_{\boldsymbol{c}_{-(i,j)}}$ be the sub-space spanned by the rows of $\boldsymbol{C}_{-(i,j)}$. Assume $\boldsymbol{\rho}|\boldsymbol{S} \sim N_p(\boldsymbol{0}, \boldsymbol{S})$, $\theta_i = \boldsymbol{c}_i\boldsymbol{\rho}$, $\theta_j = \boldsymbol{c}_j\boldsymbol{\rho}$ and $\boldsymbol{\theta}_{-(ij)} = \boldsymbol{C}_{-(i,j)}\boldsymbol{\rho}$. Then,*

$$Cov(\theta_i, \theta_j|\boldsymbol{\theta}_{-(ij)}, \boldsymbol{S}) = \begin{cases} > 0, & \text{if } \beta_{\boldsymbol{c}_i^\perp, \boldsymbol{c}_j^\perp} < \pi/2 \\ < 0, & \text{if } \beta_{\boldsymbol{c}_i^\perp, \boldsymbol{c}_j^\perp} > \pi/2, \end{cases} \tag{2.14}$$

where $\beta_{\boldsymbol{c}_i^\perp, \boldsymbol{c}_j^\perp}$ is the angle between the orthogonal projections $\boldsymbol{c}_i^\perp$ and $\boldsymbol{c}_j^\perp$ in the sub-space $\mathcal{S}^\perp_{-(i,j)}$.

Proof: Decompose $\mathcal{V}$ into $\mathcal{V} = \mathcal{S}_{-(i,j)} \oplus \mathcal{S}^\perp_{-(i,j)}$. Consider the unique representation of $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ in their orthogonal components:

$$\boldsymbol{c}_i = \boldsymbol{C}_{-(i,j)}\boldsymbol{a}_1 + \boldsymbol{c}_i^\perp \quad \text{and} \quad \boldsymbol{c}_j = \boldsymbol{C}_{-(i,j)}\boldsymbol{a}_2 + \boldsymbol{c}_j^\perp, \tag{2.15}$$

where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are $(n-2) \times 1$ vectors of constants. Under the assumptions for $\boldsymbol{\rho}$, it follows that $\boldsymbol{\theta} \mid \boldsymbol{S} \sim N_n(\boldsymbol{0}, \boldsymbol{CSC}^t)$. Consequently, we have

$$\begin{aligned} \text{Cov}(\theta_i, \theta_j|\boldsymbol{\theta}_{-(ij)}, \boldsymbol{S}) &= \text{Cov}(\boldsymbol{c}_i^t\boldsymbol{\rho}, \boldsymbol{c}_j^t\boldsymbol{\rho}|\boldsymbol{C}_{-(i,j)}\boldsymbol{\rho}, \boldsymbol{S}) \\ &= \boldsymbol{c}_i^t\boldsymbol{S}\boldsymbol{c}_j - \frac{(\boldsymbol{c}_i^t\boldsymbol{S}\boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_1)(\boldsymbol{c}_j^t\boldsymbol{S}\boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_2)}{\boldsymbol{a}_1^t\boldsymbol{C}_{-(i,j)}\boldsymbol{S}\boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_2} \\ &= < \boldsymbol{c}_i, \boldsymbol{c}_j > - \frac{< \boldsymbol{c}_i, \boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_1 >< \boldsymbol{c}_j, \boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_2 >}{< \boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_1, \boldsymbol{C}^t_{-(i,j)}\boldsymbol{a}_2 >}. \end{aligned} \tag{2.16}$$

Considering the results in (2.15), we obtain that

$$\text{Cov}(\boldsymbol{c}_i\boldsymbol{\rho}, \boldsymbol{c}_j\boldsymbol{\rho}|\boldsymbol{C}_{-(i,j)}\boldsymbol{\rho}, \boldsymbol{S}) \;= < \boldsymbol{c}_i^\perp, \boldsymbol{c}_j^\perp \;> = \; ||\boldsymbol{c}_j^\perp||\; ||\boldsymbol{c}_j^\perp|| \; \cos(\beta_{\boldsymbol{c}_i^\perp, \boldsymbol{c}_j^\perp})\,.$$

Therefore, the covariance sign is determined by the angle $\beta_{\boldsymbol{c}_i^\perp, \boldsymbol{c}_j^\perp}$. □

A geometric interpretation of the result in Theorem 3 is the following. Let $\mathcal{S}_{\boldsymbol{c}_{-(i,j)}}$ be a $(p-1)-$dimensional hyperplane. For $p = 3$, $\mathcal{S}_{\boldsymbol{c}_{-(i,j)}}$ is a plane in $\mathbb{R}$ and $\mathcal{S}^\perp_{\boldsymbol{c}_{-(i,j)}}$ is its perpendicular plane. Consider a graph with 4 nodes connected in the form of a line $\theta_1 - \theta_2 - \theta_3 - \theta_4$, that is, the nodes θ_i and θ_{i+1} are first-order neighborn for all i. To determine the sign of $\text{Cov}(\theta_1, \theta_2|\theta_{-(1,2)})$, in the Figure 9 (a), we geometrically represent the vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ related to nodes θ_1 and θ_2, respectively, the plan $\mathcal{S}_{\boldsymbol{c}_{-(1,2)}}$ generated by the vectors $\boldsymbol{c}_{-(1,2)}$ and its perpendicular plan. Figure 9(b) shows that the vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are on the same side of the plan generated by $\boldsymbol{c}_{-(1,2)}$ and the angle $\beta_{\boldsymbol{c}_1^\perp, \boldsymbol{c}_2^\perp}$ between vectors $\boldsymbol{c}_1^\perp$ and $\boldsymbol{c}_2^\perp$ is smaller than $\pi/2$. Consequently, it follows that $\text{Cov}(\theta_1, \theta_2|\theta_{-(1,2)}) > 0$. Figures 9 (d) and (c) show that the vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_3$ related to the second-order neighbor nodes θ_1 and θ_3 are on opposite sides of the generated plane $\boldsymbol{c}_{-(1,3)}$ and that the angle $\beta_{\boldsymbol{c}_1^\perp,\boldsymbol{c}_3^\perp} > \pi/2$. Thus, as opposed to what was observed for neighbors nodes of first order, for these neighbors nodes of even order we have that $\mathrm{Cov}(\theta_1, \theta_3|\theta_{-(1,3)}) < 0$.

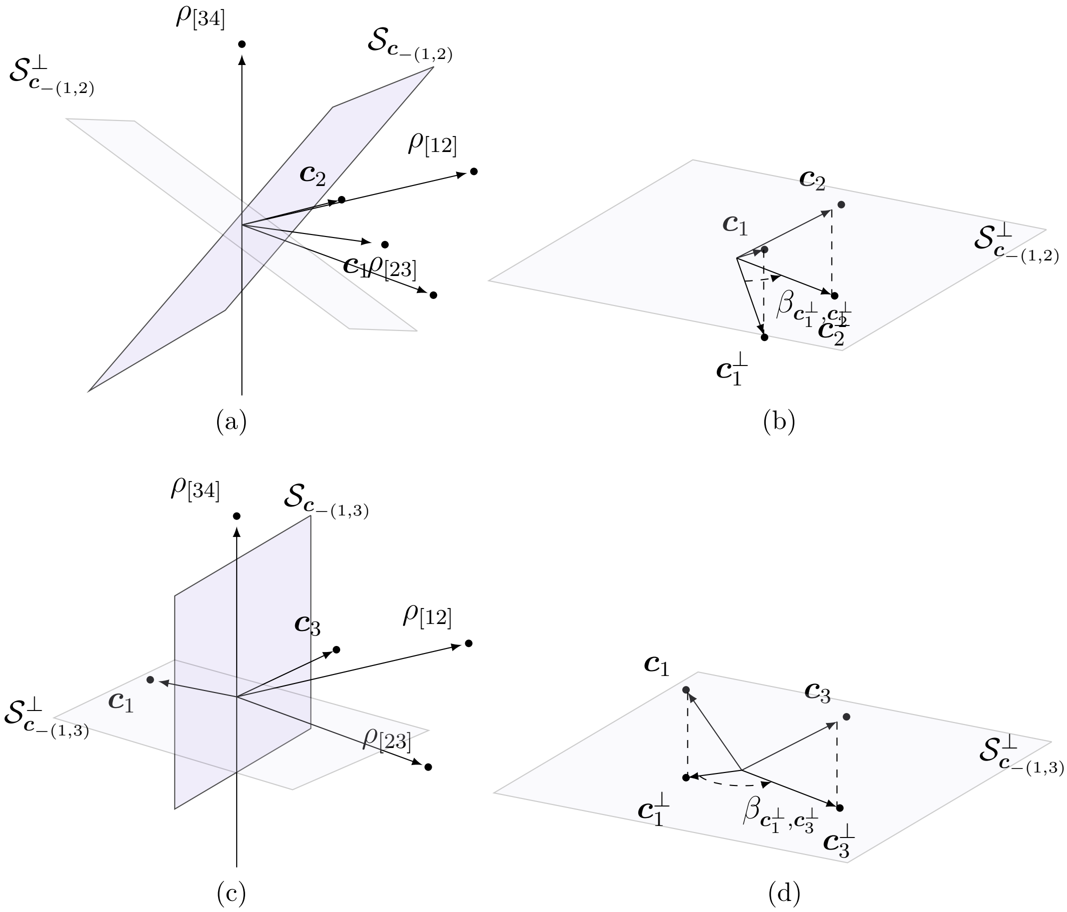

$\rho_{[34]}$
$\mathcal{S}_{\boldsymbol{c}_{-(1,2)}}$
$\mathcal{S}^\perp_{\boldsymbol{c}_{-(1,2)}}$
$\rho_{[12]}$
$\boldsymbol{c}_2$
$\boldsymbol{c}_1$
$\rho_{[23]}$
(a)
$\boldsymbol{c}_2$
$\boldsymbol{c}_1$
$\mathcal{S}^\perp_{\boldsymbol{c}_{-(1,2)}}$
$\beta_{\boldsymbol{c}_1^\perp,\boldsymbol{c}_2^\perp}$
$\boldsymbol{c}_2^\perp$
$\boldsymbol{c}_1^\perp$
(b)
$\rho_{[34]}$
$\mathcal{S}_{\boldsymbol{c}_{-(1,3)}}$
$\boldsymbol{c}_3$
$\rho_{[12]}$
$\boldsymbol{c}_1$
$\mathcal{S}^\perp_{\boldsymbol{c}_{-(1,3)}}$
$\rho_{[23]}$
(c)
$\boldsymbol{c}_1$
$\boldsymbol{c}_3$
$\mathcal{S}^\perp_{\boldsymbol{c}_{-(1,3)}}$
$\boldsymbol{c}_1^\perp$
$\beta_{\boldsymbol{c}_1^\perp,\boldsymbol{c}_3^\perp}$
$\boldsymbol{c}_3^\perp$
(d)

Figure 9 – Geometric representation for the sign of $\mathrm{Cov}(\theta_i, \theta_j|\theta_{-(i,j)})$. Geometric representation of vectors $\boldsymbol{c}_1, \boldsymbol{c}_2$ and the plane $\mathcal{S}_{\boldsymbol{c}_{-(1,2)}}$ (a) and $\boldsymbol{c}_1, \boldsymbol{c}_3$ and the plane $\mathcal{S}_{\boldsymbol{c}_{-(1,3)}}$ (c). Angle geometric representation $\beta_{\boldsymbol{c}_1^\perp,\boldsymbol{c}_2^\perp}$ between the vectors $\boldsymbol{c}_1^\perp, \boldsymbol{c}_2^\perp$ and the plan $\mathcal{S}^\perp_{\boldsymbol{c}_{-(1,2)}}$ rotated (b) and $\beta_{\boldsymbol{c}_1^\perp,\boldsymbol{c}_3^\perp}$ between the vectors $\boldsymbol{c}_1^\perp, \boldsymbol{c}_3^\perp$ and the plan $\mathcal{S}^\perp_{\boldsymbol{c}_{-(1,3)}}$ rotated (d).

2.5.1 Comparing the exact and the second order approximation correlation matrices

Given the convergent series in (2.11), we can approximate the correlation matrix of $\boldsymbol{\theta}$ by taking only the first k terms. If this approximation is appropriate, we can see the correlation structure as one approximately determined only by the node local neighborhood up to the k-th order. For the CAR and ICAR models, Assunção and Krainski [2009] showed that the quality of such approximation depends on how close to 1 is $|\lambda_2|$, the second largest eigenvalue of $\boldsymbol{W}_v$, which is the neighborhood matrix of $\mathcal{G}$. The parameter $|\lambda_2|$ is the spectral radius and it is a high-level summary of the graph neighborhood structure.

We are interested in knowing how sensitive is the approximation quality when the neighborhood structure changes. We can make the graph denser and more complex by adding edges or, in the opposite direction, we can thin and simplify the graph by pruning some of its existing edges. The presence of $\boldsymbol{C}$ in (2.11) makes difficult an analytic approach. Hence, we carry out an empirical evaluation of the impact of thinning or enlarging the neighborhood graph in the approximation quality based on truncating expression (2.11). We show the results using a second-order approximation for the correlation matrix. We change the neighborhood structure of some real maps by randomly adding or removing some edges. More specifically, to thin the graph, we randomly select an edge to be removed until only $n-1$ edges remain in the original graph. The removal is carried out under the constraint that the graph remains connected. To make it denser, we enlarge the neighborhood structure by randomly including edges between second-order neighbors in the original graph. This experiment is repeated 100 times for each addition or deletion.

Figure 10 shows the maximum absolute difference between the entries of the correlation matrix and its second order approximation. Four different neighborhood structures are considered: the USA, two american states (Wyoming and Iowa) and one the Brazilian state (Minas Gerais), with the neighborhood graphs shown in the left-hand side column. The horizontal axis in the plots shows the number of removed (negative values) or added edges (positive values) in the original neighborhood graph. The second column of plots show five selected simulations while the third columns shows the 95% pointwise confidence bands based on 100 simulations. For sparse neighbourhood structures, such as those from the USA and Wyoming, the differences between the matrices are larger. In this type of sparse neighborhood , we need to take a higher-order to obtain better approximation of the correlation matrix. In the inverse direction, graphs that are dense can have their correlation matrix well represented by considering only the local neighborhood of each pair. In other words, the more dense is the graph, the smaller the effect of far away neighbors. This result is similar to that obtained in Assunção and Krainski [2009] using the CAR and ICAR models.

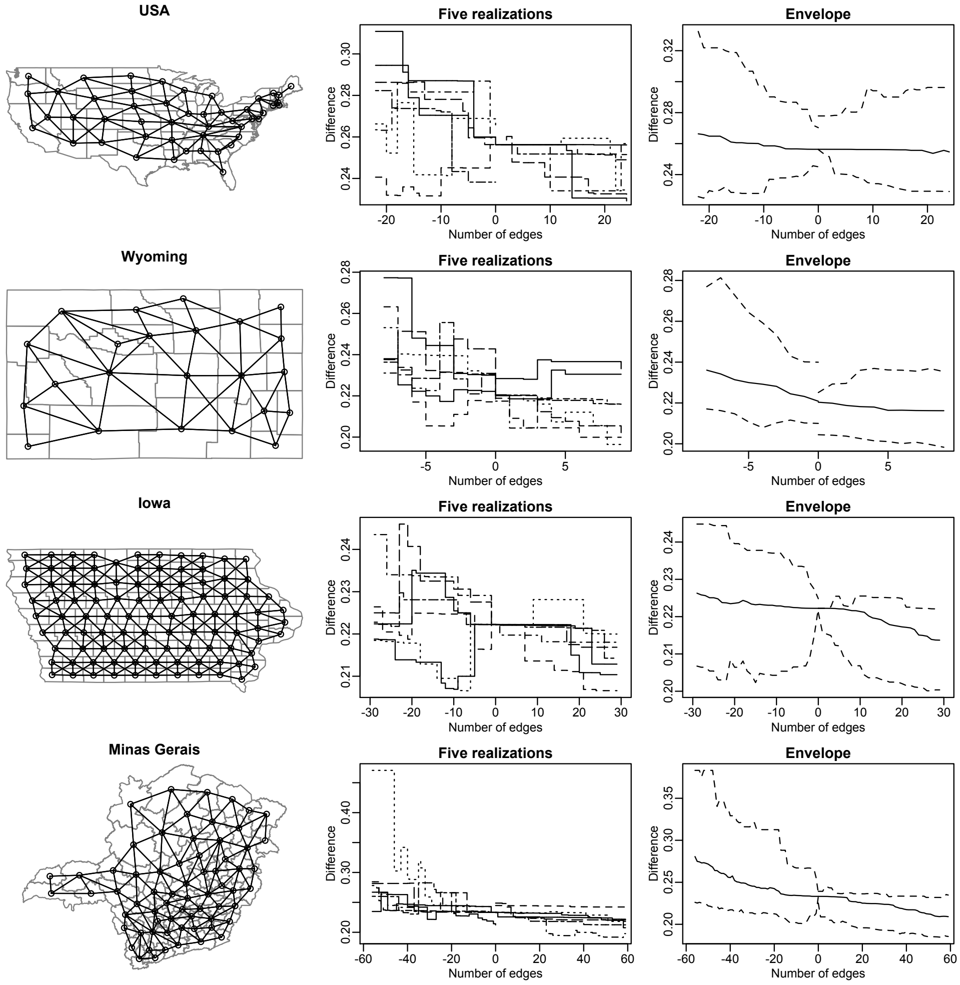


Figure 10 – Neighboring structures (left), the difference for five selected samples (middle) and the average and interval with 95% of the generated difference between the exact correlations and its second order approximation.

2.6 Modeling spatial data using RENeGe

In this section, we empirically explore different features of RENeGe model. First, its ability to induce a higher marginal correlation between neighboring areas than CAR models is presented. Next, we show that RENeGe does not oversmooth when recovering random effects. Finally, we make a comparative analysis of five types of cancer mortality data using our model and a set of alternative models.

2.6.1 Inducing correlation between neighbors

One great limitation of the CAR model is the small correlation this model induces between neighboring areas, even when a high value for the spatial parameter ρ is assumed

[Ver Hoef et al., 2017]. To compare the marginal correlation induced by RENeGe and CAR models, we consider simulated datasets assuming that the areas are organized in a regular square lattice witn $n = 30$ nodes. Wrapping the grid into a torus, we have four neighbors for each area.

Data are generated assuming that independently $(Y_i \mid \boldsymbol{\theta}, \tau_y) \sim N(\theta_i, \tau_y)$, where $\tau_y \in \mathbb{R}_+$ is the precision parameter and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$.

For RENeGe, we assume that $\boldsymbol{\theta} \sim T_n\left(\mathbf{0}; \frac{\boldsymbol{K}(\nu-n-1)}{\tau_{\boldsymbol{\theta}}(\nu-n+1)}; \nu - n + 1\right)$ where $\boldsymbol{K} = \boldsymbol{C}(\boldsymbol{M}_e - \gamma \boldsymbol{A}_e)^{-1}\boldsymbol{C}^t$. For the CAR model, we consider $\boldsymbol{\theta} \sim N_n\left(\mathbf{0}, ((\tau_c \boldsymbol{D}_{\mathcal{G}} - \rho \boldsymbol{A}_{\mathcal{G}}))^{-1}\right)$. We take $\tau_y = \tau_\theta = \tau_c = 1$ and $\nu = 32$. We select three high values for the spatial parameters γ and ρ of the two models: $0.800, 0.900$, and 0.999. We calculated the Moran's index I in each simulated map:

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n A_{ij}(y_i - \bar{y})(y_j - \bar{y})}{s_y^2 \sum_{i=1}^n \sum_{j=1}^n A_{ij}},$$

where s_y is the sample standard deviation and A_{ij} is the adjacency matrix defined by the torus-wrapped regular square lattice.

Table 1 shows the average value of the Moran's index taken over 100 simulated datasets. It also presents the theoretical bounds limiting the possible values of the Moran I index in the last column. For any value of the spatial parameters, RENeGe has much higher correlation between neighboring areas than the CAR model. Taking $\gamma = 0.9$ in RENeGe, we have $I = 0.357$ while $I = 0.22$ even when $\rho = 0.999$ in the CAR model. Although it is clear that RENeGe induces higher correlation between neighboring values than CAR, it is disappointing that the index does not come close to its maximum value even when the spatial parameter γ has a value very close to limit 1. The pairwise correlation between neighbors reaches only 0.384 when $\gamma = 0.999$. It is possible that conditionally specified spatial models may have some kind of intrinsic limitation to reach very high values for this marginal correlation.

Parameter	RENeGe	CAR	Theoretical
$(\gamma,\ \rho)$			Bounds
0.8	0.170	0.062	$(-1.03, 1.02)$
0.9	0.357	0.081	$(-1.03, 1.02)$
0.999	0.384	0.222	$(-1.03, 1.02)$

Table 1 – Average Moran I index under the RENeGe and CAR models for different values of γ and ρ.

2.6.2 RENeGe does not oversmooth

Markov Random Fields have often been used for image reconstruction [Cross and Jain, 1983, Qian and Titterington, 1991, Chaur-Chin and Chung-Ling, 1993, Aykroyd, 1998]. However, such models tend to oversmooth discontinuities. To contrast this aspect with our proposed RENeGe, we consider a toy image reconstruction example. Our focus is not on image processing, which is a specialized topic with a huge literature. We are using an image because it highlights the differences between the methods as we can visually contrast the smoothing properties of the CAR and RENeGe models.

The grayscale image is represented by a two-dimensional regular lattice with n pixels or cells. Denote by $\theta_i \in \mathbb{R}$ the shading variable at pixel i. Let y_i be the observed image at pixel i, which is a degraded copy of θ_i. That is, $Y_i = \beta + \theta_i + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \tau_y)$ and τ_y is the precision parameter. To model the uncertainty about the true shading pixel variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, for comparison purpose, we assume the spatial models RENeGe or CAR. To complete the model specification, we assume that $\beta \sim N(\mu_\beta, \tau_\beta^{-1})$. For all precision parameters, we consider flat Gamma distributions.

Consider the 8836 × 8836 tiger image in Figure 11a, implying in a graph with $p = 17484$ edges. The observed image is given in Figure 11b and it is generated by adding independent Gaussian errors to the original pixels with precision 10. To analyse the data, we assume that $\beta \sim N(0, 0.1)$, and τ_y, τ_θ, and τ_C follow a Gamma$(10^{-2}, 10^{-2})$ distribution, and $\gamma \sim$ Uniforme$(0, 1)$. For RENeGe model, we fix $\nu = 8838$. Both models were fit by collecting 500 MCMC iterations after discarding the first 100 as burn-in.

Figure 11 shows the images estimated using RENeGe (Figure 11c) and the CAR (Figure 11d) models. We obtain a better image recovery using RENeGe. For example, the regions around the mouth and the nose are less blurred in the image estimated using the proposed model. The image using the CAR model is over smoothed, not showing the contrasts that exists in the original image.

2.6.3 Case Study: cancer maps

In this section, we analyse the spatial pattern of deaths caused by five types of cancer: lung/bronchial, colon/rectal, stomach, female breast cancer, and male prostrate cancer.

They are selected due to the fact that thet have higher letality among the different cancers. We collect the total number of deaths occurred in the 2008-2019 periods in

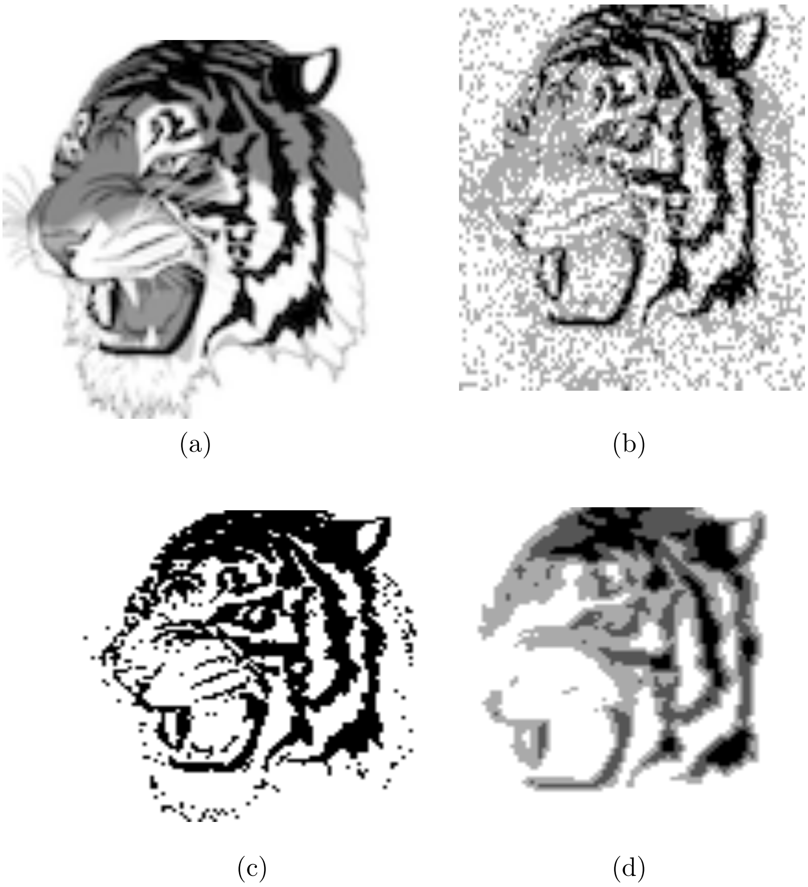


Figure 11 – Original (a) and noised images(b) and the images estimated using RENeGe (c) and the CAR (d) models.

areas from four states located in the south of Brazil: Rio Grande do Sul, Santa Catarina, Paraná and São Paulo. This region has 73 million inhabitants and it is partitioned into $n = 159$ administrative areas called micro-regions. Cancer and population data were collected from the DATASUS website (http://datasus.saude.gov.br/), the official Brazilian Health Department data repository. This region has been selected because it contains approximately 35% of the Brazilian population and its data has good quality, having little under reporting or cause of death misreporting problems.

As covariate, we consider the the Municipal Human Development Index (MHDI), a composite index measuring social deprivation and calculated by the United Nations. These data were obtained from `https://www.br.undp.org/`. MHDI is a combination of three indicators capturing different dimensions of human development: longevity, education, and economic well-being. The index varies from 0 to 1. The closer to 1, the greater the human development. Figure 12 shows the spatial distribution of the MHDI in the southern region of Brasil. There is a clustered spatial pattern in the MHDI distribution.

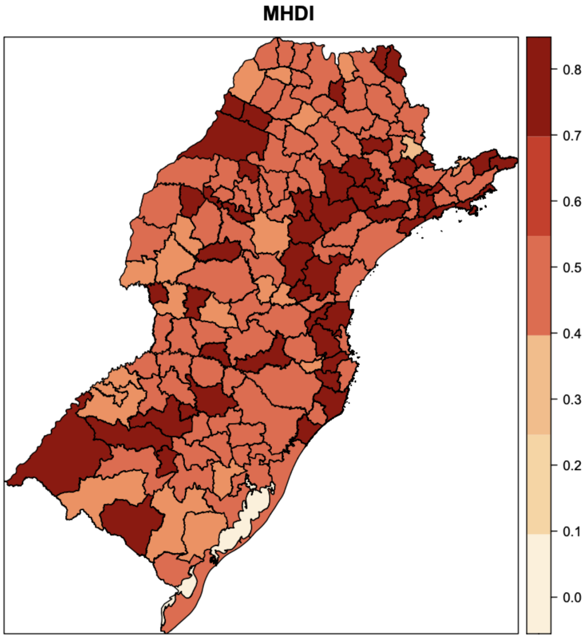


Figure 12 – Thematic map of the Municipal Human Development Index (MHDI) in the micro-regions of the Brazilian Southern region.

Let Y_i and E_i denote the observed and the expected counts of death by cancer at microregion i, respectively. The expected value E_i is calculated using the age-sex population distribution and assuming that the age-specific risk is constant in the entire map. The data are analyzed assuming that

$$Y_i|\theta_i, \tau_y \overset{ind}{\sim} \text{Poisson}(E_i\lambda_i), \quad i = 1, \dots, 159.$$

The λ_i follows the log-linear regression structure $\log(\lambda_i) = X\beta + \theta_i + \epsilon_i + \log(E_i)$ where θ_i is the spatial random effects at area i and $\epsilon_i \sim N(0, \sigma^2)$ is the local random effects.

To fit RENeGe, we assume uninformative prior distributions in the following hierarchical structure: $\beta{\sim}N(0, 10^2)$; $\boldsymbol{\theta}|\boldsymbol{\Sigma} \sim N_n(\mathbf{0}, (\nu{-}n{-}1)\boldsymbol{\Sigma})$; $\boldsymbol{\Sigma} \sim \text{IW}_n(\nu, \sigma^2, \tau_{\boldsymbol{\theta}}^{-1}\boldsymbol{C}(\boldsymbol{M}-\gamma\boldsymbol{A})^{-1}\boldsymbol{C}^t)$; $\gamma \sim \text{Uniform}(0, 1)$ and $\tau_{\boldsymbol{\theta}} \sim \text{Gamma}(10^{-2}, 10^{-2})$.

We compare RENeGe with the plain CAR model [Besag et al., 1995], the generalized skew-normal (GSN) model [Prates et al., 2012], Leroux model [Leroux et al., 2000], the BYM model [Besag et al., 1991], the model introduced by [Rodrigues and Assunção, 2012] that extends BYM and Leroux models by considering a higher neigborhood dependence

(HND) and the DAGAR model Datta et al. [2019]. In all these models, we considered flat prior distributions for all parameters.

To compare the models, we apply some usual model selection criteria: the Deviance Information Criteria (DIC), the Extended Bayesian Information Criterion (EBIC), the extended Akaike information criterion (EAIC), the Watanabe–Akaike information criterion (WAIC) and the Root Mean Squared Error (RMSE). To assess the predictive accuracy of the models, we removed 31 microregions from the map in order to obtain a connected graph representing the remaining areas. RENeGe and the other models are fitted to this subset of regions and the counts for the removed areas are estimated using the posterior averages. This procedure is repeated 5 times and the BIAS and RMSE are respectively calculated using the formulas

$$BIAS = \frac{\sum_{j=1}^{5}\sum_{i=1}^{n}(\hat{Y}_{ij} - Y_i)}{5n} \qquad RMSE = \sqrt{\frac{\sum_{j=1}^{5}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{5n}},$$

where Y_i is the observed count in area i and $\hat{Y}_{ij}$ is its predicted value at replication j.

Although the models have similar performance, all model selection criteria (Table 2) point out that the proposed model outperforms the other models for all datasets. In general, DAGAR, GSN, and HND are the models with the poorest performances.

Figures 13 and 14, respectively, show the relative risk and spatial random effects estimates for the two most prevalent types of cancers, female breast cancer and lung/bronchial cancer, under all fitted models. Results for the other cancers can be found in the Appendix B.3. The estimates are the posterior mean of θ_i in the case of the Bayesian model.

Although Table 2 shows differences between the models, these differences cannot be clearly visualized in Figure 13. CAR, GSN, Leroux, BYM, HND and DAGAR models provide very similar maps for the relative risk of Lung/Bronchi cancer mortality and the estimates are comparable to the naive SMR estimate. Estimates provided by RENeGe are almost homogeneous and very small in all areas. The lung cancer is the most prevalent cancer and its large numbers allow for stable rates that are not much different from the Bayesian smoothed rates for this cancer. The relative risk are grouped in relatively large clusters indicating that neighboring areas tend to have similar risk. CAR, GSN, Leroux, BYM, HND and DAGAR models indicate a very clear North-South gradient with the lung cancer risk increasing as we move towards the south. This gradient is the most distinctive spatial pattern in this map, with small differences across the East-West direction. This is also observed under RENeGe but the differences are not so noticeable. Disturbing the smooth trend along the North-South gradient in this map, we have three high risk spatial

Model	DIC	EBIC	EAIC	WAIC	RMSE	BIAS
			Breast cancer			
RENeGe	**1384.14**	**1918.02**	**1513.35**	**1332.42**	**776.82**	**84.31**
CAR	1388.86	2042.80	1549.58	1335.89	777.87	85.52
GSN	1492.47	2146.58	1653.22	1442.46	879.64	87.52
Leroux	1390.25	2025.94	1546.48	1344.21	777.84	85.53
BYM	1390.10	2053.67	1553.18	1335.38	777.71	85.52
HND	1390.54	2044.65	1551.30	1340.53	777.71	85.60
DAGAR	1483.81	2117.22	1639.48	1437.95	879.73	87.41
			Lung/Bronchial Cancer			
RENeGe	**1497.79**	**2110.66**	**1648.41**	**1456.84**	**452.83**	**85.45**
CAR	1514.37	2204.80	1684.06	1457.24	462.83	86.47
GSN	1663.19	2542.00	1879.18	1563.84	564.87	86.57
Leroux	1509.94	2208.67	1681.66	1440.59	462.78	86.49
BYM	1508.60	2173.84	1672.09	1457.52	462.86	86.46
HND	1561.27	2440.08	1777.25	1461.92	462.94	86.39
DAGAR	1599.71	2212.58	1750.34	1558.77	462.44	86.77
			Colon/rectal cancer			
RENeGe	**1493.61**	**1975.42**	**1612.02**	**1473.48**	**1353.27**	**166.77**
CAR	1535.04	2181.42	1693.90	1481.71	1353.68	166.97
GSN	1644.51	2279.93	1800.68	1606.46	1455.65	168.74
Leroux	1609.60	2459.86	1818.57	1529.40	1353.68	167.02
BYM	1540.23	2207.45	1704.21	1482.29	1353.63	166.95
HND	1542.59	2178.01	1698.75	1504.54	1353.72	166.81
DAGAR	1711.52	2561.79	1920.49	1631.33	1355.61	168.95
			Stomach cancer			
RENeGe	**1502.12**	**2059.86**	**1518.57**	**1429.40**	**461.87**	**85.50**
CAR	1514.37	2204.80	1684.06	1457.24	462.83	86.47
GSN	1663.19	2542.00	1879.18	1563.84	564.87	86.59
Leroux	1509.94	2208.67	1681.66	1440.59	462.78	86.49
BYM	1508.60	2173.84	1672.09	1457.52	462.86	86.46
HND	1561.27	2440.08	1777.25	1461.92	462.94	86.39
DAGAR	1711.52	2561.79	1920.49	1631.33	464.79	88.42
			Prostate cancer			
RENeGe	**1509.60**	**1459.86**	**1418.57**	**1429.40**	**788.31**	**123.32**
CAR	1527.12	2216.26	1696.49	1468.43	798.34	124.29
GSN	2511.65	6729.18	3548.18	1725.84	900.21	226.08
Leroux	1530.53	2296.37	1718.75	1431.95	798.29	124.31
BYM	1526.38	2200.23	1691.99	1476.42	798.07	124.35
HND	2409.72	6627.26	3446.25	1623.92	798.29	124.16
DAGAR	1632.45	2398.29	1820.67	1533.88	800.22	126.23

Table 2 – Model comparison criteria for all fitted models.

clusters on the South. Two of them are on the border between Brazil and Uruguay while the third one contains Porto Alegre. The relative risks in these clusters are around 2, meaning that they have a risk twice as large as the average risk in the entire region. The estimates for the relative mortality risk of breast cancer are spatially more heterogeneous than the lung cancer risk. Except for the GLM, all other models provide visually similar estimates for the mortality risk. The maps do not show any striking differences between the spatial models. In fact, they look identical and differences between the models must be ascertained through the model selection metrics from Table 2. Returning to the breast cancer maps, they point out to the same four spatially unconnected areas with the highest relative risk, around twice the average risk.

Figure 14 shows the random effects estimates $\hat{\theta}_i$ for the Breast (right) and Lung/Bronchial (left) cancer mortality under all fitted spatial models. In these maps, we see that the spatial effects intertwine the effects of the variables for almost all models. The graphics on the left of Figures 13 and 14 show that, for Lung/ Bronchial cancer, the spatial effects are essentially grouped in three clusters and are positive in the South and negative in the North of the region. The most important difference is the risk decrease in most models of the few clusters of high risk in the South. Our RENeGe model also provides similar cluster behavior for the spatial effects but points three areas with more extreme spatial effect: one below -1 (Capão Bonito, São Paulo State) and two around 1 (Campanha Meridional and Litoral Lagunar, extreme south of the region). Similar feature is observed under HND model. Considering the substantial smoothing induced by the CAR, as seen in Section 2.6.2, we may suspect that the higher estimates produced by RENeGe may be closer to the unknown truth.

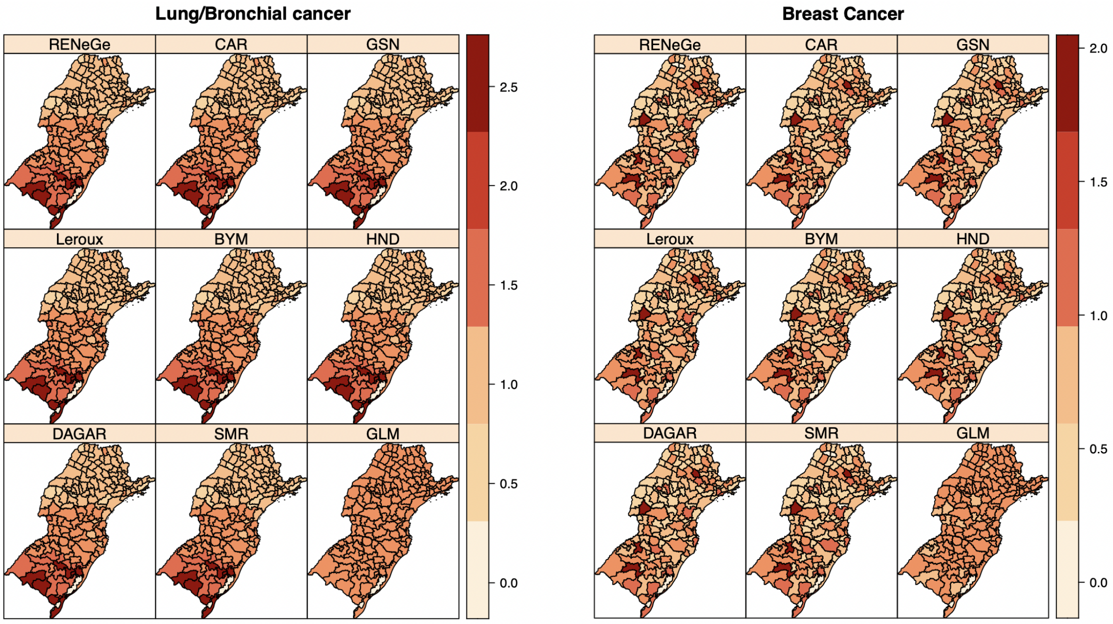


Figure 13 – Posterior means for the relative risk of the Breast (left) and Lung/Bronchial (right) cancer mortality in the southern region of Brazil under all fitted models.

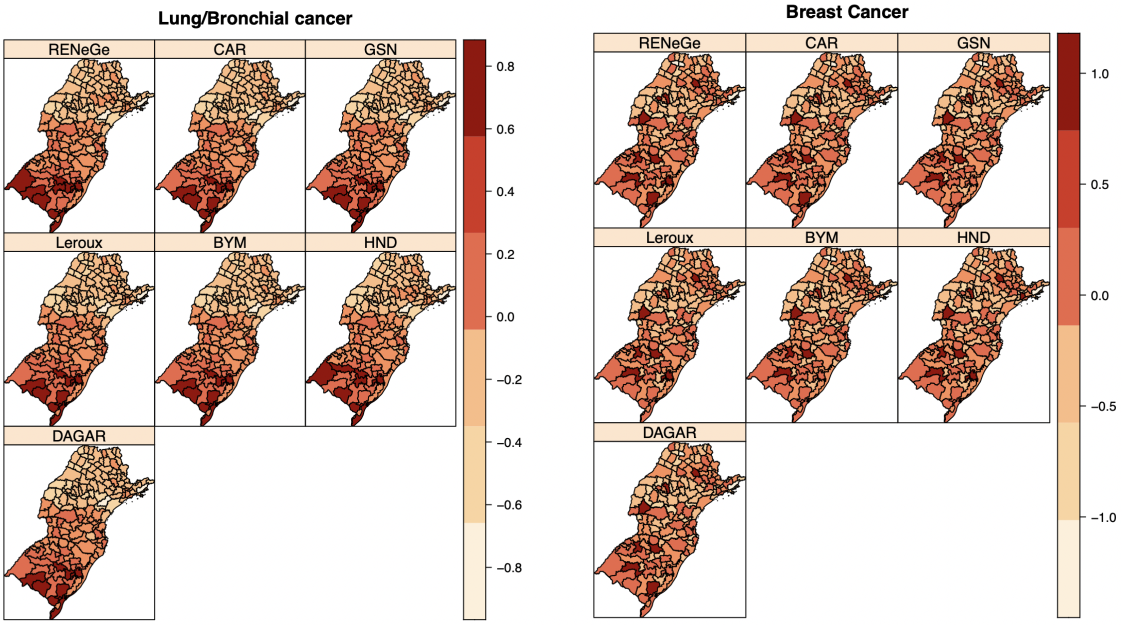


Figure 14 – Posterior means for the spatial effects in the Breast (left) and Lung/Bronchial (right) data in the southern region of Brazil under all fitted models.

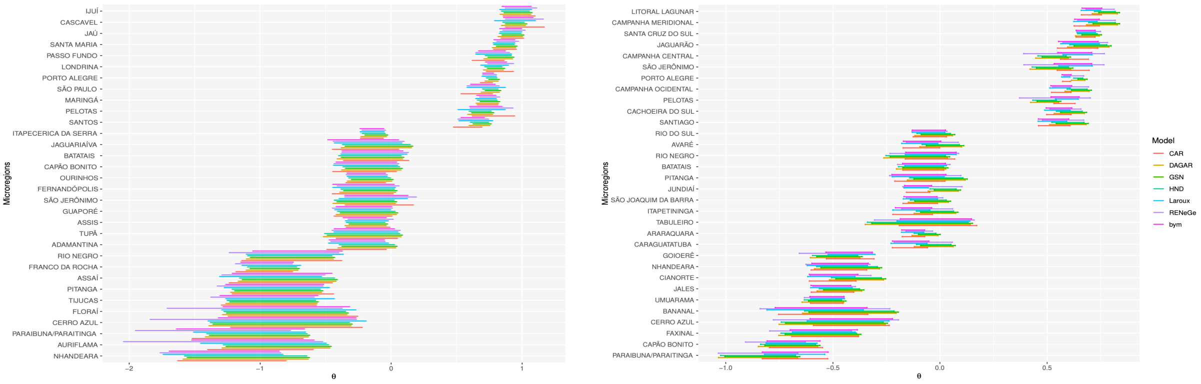
Another important aspect is that the spatial pattern in $\hat{\theta}_i$ (Figure 14a) is very similar to that found in $\hat{\lambda}_i$ (Figure 13a) for the majority of the models. The smooth North-South gradient is almost the same in both maps. This is indicative that the quality-of-life index, as measured by MHDI, has little correlation with the lung/bronchi cancer mortality risk. In fact, this will be reinforced by the results shown later in Table 3. For the breast cancer risk, we also have practically the same spatial pattern when comparing $\hat{\theta}_i$ in Figure 14b is very similar to that found in $\hat{\lambda}_i$ in Figure 13b. Again, the risk for this cancer seems to have little association with the MHDI social-economic index.

Table 3 shows the posterior means and the 95% highest posterior density (HPD) intervals for the parameters in the models analysed. Using a non-spatial GLM Poisson regression model, we find a significant parameter associated with the covariate MHDI. It is positive for breast cancer but negative for lung cancer. There is some possible explanation for this: lung cancer is mainly due to cigarette consumption that used to be higher in Brazil in those areas with greater income. Breast cancer mortality may be related with the lack of preventive examinations which are higher in poorer areas. However, this covariate coefficient is outside the HPD intervals for most spatial models. That is, as soon as we allow for random spatial effects, the covariate is not relevant any longer. This puzzling result may be caused by the confounding between the strongly spatially patterned covariate and the spatial effects. Several recent papers have been dedicated to this thorny issue in spatial statistics [Hodges and Reich, 2010, Hughes and Haran, 2013, Hanks et al., 2015, Prates et al., 2019, Nobre et al., 2020]. Although we do not pursue this further in this paper, some kind to control for spatial confounding may be necessary in the analysis of these cancers in Brazil if the MHDI covariate is used.

Coefficients	Mean	95% HPD	Mean	95% HPD	Mean	95% HPD
			Breast cancer			
	RENeGe		CAR		GSN	
Intercept	-0.21	(-0.60, 0.14)	-0.68	(-1.34, -0.26)	-0.29	(-0.49, -0.08)
MHDI	-0.11	(-0.68, 0.46)	0.60	(-0.05, 1.60)	-0.03	(-0.36, 0.28)
τ_θ	1.99	(1.53, 2.61)	1.39	(1.10, 1.77)	0.00	(0.00, 0.00)
ρ	-1.077	(-2.27, 0.75)	0.28	(0.03, 0.66)	0.06	(0.06, 0.06)
σ^2	1.99	(1.53, 2.78)	-	-	-	-
	Leroux		BYM		HND	
Intercept	-0.45	(-1.06, 0.23)	-0.52	(-0.94, -0.13)	-0.29	(-0.49, -0.08)
MHDI	0.23	(-0.81, 1.18)	0.34	(-0.25, 1.01)	-0.03	(-0.36, 0.28)
τ_θ	0.41	(0.27, 0.71)	0.16	(0.05, 0.35)	0.00	(0.00, 0.01)
ρ	0.13	(0.02, 0.38)	0.21	(0.14, 0.28)	0.06	(0.06, 0.06)
	DAGAR		GLM			
Intercept	-0.52	(-0.94, -0.13)	-0.60	(-0.70, -0.50)		
MHDI	0.34	(-0.25, 1.01)	1.04	(0.94, 1.14)		
τ_θ	0.16	(0.05, 0.35)	-	-		
ρ	0.21	(0.14, 0.28)	-	-		
			Lung/bronchi Cancer			
	RENeGe		CAR		GSN	
Intercept	0.22	(0.04, 0.40)	0.17	(0.03, 0.30)	0.61	(0.45 0.76)
MHDI	-0.001	(-0.27, 0.27)	0.10	(-0.11, 0.31)	-0.62	(-0.85, -0.37)
τ_θ	1.92	(1.43, 2.53)	0.10	(0.08, 0.12)	0.00	(0.00, 0.01)
ρ	-1.06	(-2.21, 0.75)	0.99	(0.98, 1.00)	0.06	(0.06, 0.06)
σ^2	1.91	(1.43, 2.53)	-	-	-	-
	Leroux		BYM		HND	
Intercept	0.1	(-0.04, 0.3)	0.16	(-7e-02, 0.35)	0.610	(0.44, 0.76)
MHDI	0.2	(-0.09, 0.4)	0.09	(-2e-01, 0.46)	-0.62	(-0.84, -0.36)
τ_θ	0.1	(0.07, 0.1)	0.087	(7e-02, 0.11)	0.005	(0.00, 0.00)
ρ	1.0	(0.91, 1.0)	0.002	(6e-04, 0.06)	0.05	(0.05, 0.05)
	DAGAR		GLM			
Intercept	0.17	(-0.07, 0.35)	0.53	(0.43, 0.63)		
MHDI	0.10	(-0.18, 0.47)	-0.35	(-0.45, -0.25)		
τ_θ	0.09	(0.07, 0.12)	-	-		
ρ	0.00	(0.00, 0.01)	-	-		

Table 3 – Posterior means and 95% HPD intervals under all models.

Figure 15 presents the posterior means and the 95% HDP intervals for the spatial random effects of 32 areas selected as follows: 10 areas with the lowest SMR (bottom), 11 areas with the intermediary SMR, and 11 areas with the highest SMR (top). For both types of cancers, the higher the SMR, the higher the spatial effect. Areas with the intermediary and low values for the SMR experience negative spatial effects. However, these areal effects approach zero for areas with intermediary SMR in both types of cancers. Thus, the relative risk in these areas is essentially explained by the global effect, except when fitting the GSN and HND models to analyze Lung/bronchial cancer data, for which the MHDI is a significant factor to explain the mortality rate. In general, the uncertainty about the random effects is higher in areas with the smallest SMR. This is more clearly perceived for Breast cancer data. However, there is no well-defined standard for the estimates of spatial effects in each area when comparing the models. For instance, for Breast cancer, the RENeGe indicates the highest uncertainty about the areal effects in Floraí, Auriflora, Paraibuna/Paraitinga, and Cerro Azul while for São Paulo and Cascavel, there is more uncertainty about this effect when assuming the CAR model.



(a)

(b)

Figure 15 – Posterior estimates of spatial effects for Breast (a) and Lung/Bronchi (b) cancers, under all models.

2.7 Conclusions

The CAR model, the most popular approach to handle data spatially correlated, is unable to take into account the high spatial correlation thus over-smoothing out discontinuities. While this model characteristic is important in many practical situations, it is not appropriate if heterogeneity is an important feature to be captured by the model. This, for instance, occurs in the reconstruction of images where its is important to recover the correct texture or in desease mapping, if the neighboring structure is not properly defined by areas that share borders on the map.

We propose a robust spatial model that accounts, in part, for this limitation of the CAR model. We consider a Student-t distribution for the random effects. The novelty is the way the precision matrix for such a distribution is built. It is built assigning a Student-t distribution for the random weights of the edges connecting the spatial random effects. Several properties of the covariance matrix are discussed and compared to that of CAR model. We prove that the correlation induced by the proposed model is higher.

The proposed model showed that, if compared to CAR model, it better accounts for heterogeneity providing a better reconstruction of the image. For the cancer datasets, the proposed model outperform many other spatial models previously introduced in the literature showing to be a competitive model to account for spatial correlation.

2.8 Bibliography

Arellano-Valle, R. and Bolfarine, H. [1995]. On some characterizations of the t-distribution, *Statistics & Probability Letters* **25**(1): 79–85.

Assunção, R. M. [2003]. Space varying coefficient models for small area data, *Environmetrics* **14**(5): 453–473.

Assunção, R. M. and Krainski, E. [2009]. Neighborhood dependence in bayesian spatial models, *Biometrical Journal* **51**(5): 851–869.

Assunção, R. M., Potter, J. E. and Cavenaghi, S. M. [2002]. A bayesian space varying parameter model applied to estimating fertility schedules, *Statistics in Medicine* **21**(14): 2057–2075.

Aykroyd, R. G. [1998]. Bayesian estimation for homogeneous and inhomogeneous gaussian random fields, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(5): 533–539.

Bapat, R. B. [2014]. *Graphs and Matrices*, Universitext, Springer London.

Besag, J. [1974]. Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B* **36**(2): 192–236.

Besag, J., Green, P., Higdon, D. and Mengersen, K. [1995]. Bayesian computation and stochastic systems, *Statistical Science* **10**(1): 3–41.

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

Carlin, B. P. and Banerjee, S. [2003]. Hierarchical multivariate car models for spatio-temporally correlated survival data, *Bayesian statistics* **7**: 45–63.

Chaur-Chin, C. and Chung-Ling, H. [1993]. Markov random fields for texture classification, *Pattern Recognition Letters* **14**(11): 907 – 914.

Cross, G. R. and Jain, A. K. [1983]. Markov random field texture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-5**: 25–39.

Datta, A., Banerjee, S. and Hodges, J. S. [2019]. Spatial disease mapping using directed acyclic graph auto-regressive (dagar) models, *Bayesian Analysis* **14**(8): 1221–1244.

Fahrmeir, L. and Lang, S. [2001]. Bayesian inference for generalized additive mixed models based on markov random field priors, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**: 201 – 220.

Gelfand, A. E., Hyon-Jung, K., Sirmans, C. F. and Banerjee, S. [2003]. Spatial modeling with spatially varying coefficient processes, *Journal of the American Statistical Association* **98**(462): 387–396.

Gelfand, A. E. and Vounatsou, P. [2003]. Proper multivariate conditional autoregressive models for spatial data analysis, *Biostatistics* **4**(1): 11–15.

Golub, G. H. and Van Loan, C. F. [1996]. *Matrix Computations*, third edn, The Johns Hopkins University Press.

Hanks, E. M., Schliep, E. M., Hooten, M. B. and Hoeting, J. A. [2015]. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification, *Environmetrics* **26**(4): 243–254.

Hodges, J. S. and Reich, B. J. [2010]. Adding spatially-correlated errors can mess up the fixed effect you love, *The American Statistician* **64**(4): 325–334.

Hughes, J. and Haran, M. [2013]. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1): 139–159.

Iranmanesh, A., Arashi, M., Nagar, D., Nadarajah, S. and Tabatabaey, S. [2012]. A new mixture representation for multivariate t, *Journal of Multivariate Analysis* **107**: 227–231.

Jin, X. and Carlin, B. P. [2005]. Multivariate parametric spatiotemporal models for county level breast cancer survival data, *Lifetime data analysis* **11**(1): 5—27.

Knorr-Held, L. and Best, N. G. [2001]. A shared component model for detecting joint and selective clustering of two diseases, *Journal of the Royal Statistical Society, Series A* **164**: 73–85.

Knorr-Held, L., Graziano, G., Frank, C. and Rue, H. [2006]. Joint spatial analysis of gastrointestinal infectious diseases, *Statistical Methods in Medical Research* **15**(5): 465–480. PMID: 17089949.

Kotz, S. and Nadarajah, S. [2004]. *Multivariate T-Distributions and Their Applications*, Cambridge University Press.

Leroux, B. G., Lei, X. and Breslow, N. [2000]. Estimation of disease rates in small areas: A new mixed model for spatial dependence, pp. 179–191.

MacNab, Y. and Dean, C. B. [2000]. Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models, *Statistics in Medicine* **19**: 2421–2435.

Martínez-Beneito, M. A. and Botella-Rocamora, P. [2019]. *Disease Mapping: From Foundations to Multidimensional Modeling*, CRC Press.

Martínez-Beneito, M. A., López-Quilez, A. and Botella-Rocamora, P. [2008]. An autoregressive approach to spatio-temporal disease mapping, *Statistics in Medicine* **27**(15): 2874–2889.

Nobre, W. S., Schmidt, A. M. and Pereira, J. B. [2020]. On the effects of spatial confounding in hierarchical models, *International Statistical Review* .

Prates, M. O., Assunção, R. M., Rodrigues, E. C. et al. [2019]. Alleviating spatial confounding for areal data problems by displacing the geographical centroids, *Bayesian Analysis* **14**(2): 623–647.

Prates, M. O., Kumar Dey, D. and Lachos, V. H. [2012]. A dengue fever study in the state of rio de janeiro with the use of generalized skew-normal/independent spatial fields, *Chilean Journal of Statistics.* **3**(2): 143–155.

Qian, W. and Titterington, M. D. [1991]. Multidimensional markov chain models for image textures, *Journal of the Royal Statistical Society: Series B* **53**(3): 661–674.

Rodrigues, E. C. and Assunção, R. M. [2012]. Bayesian spatial models with a mixture neighborhood structure, *Journal of Multivariate Analysis* **109**: 88 – 102.

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

Ver Hoef, J., Peterson, E., Hooten, M., Hanks, E. and Fortin, M.-J. [2017]. Spatial autoregressive models for statistical inference from ecological data, *Ecological Monographs* .

Wall, M. M. [2004]. A close look at the spatial structure implied by the CAR and SAR models, *Journal of Statistical Planning and Inference* **121**(2): 311–324.

3 Spatial shrinkage prior: A probabilistic approach to model for categorical variables with many levels

3.1 Abstract

One of the most used methods to prevent overfitting and select relevant variables in regression models with many predictors is the penalized regression technique. Under such approaches, variable selection is performed in a non-probabilistic way, using some optimization criterion. A Bayesian approach to penalized regression has been proposed by assuming a prior distribution for the regression coefficients that plays a similar role as the penalty term in the classical statistics: to shrink toward zero non-significant coefficients and to put a significant probability mass to non-despicable coefficients. These prior distributions, called *shrinkage priors*, usually assume independence among the covariates, which may not be an appropriate assumptions in many cases. We propose two shrinkage priors to model the uncertainty about coefficients that are spatially correlated. The proposed priors are assumed as an alternative approach to model the uncertainty about coefficient of categorical variables with many levels. To illustrate their uses, we consider the linear regression model. We evaluate the proposed method through several simulation studies and analyzing the housing prices dataset available from Airbnb.

Keywords: Spatial statistics, Robust spatial model, edge graph, shrinkage prior.

3.2 Introduction

Penalized regression or regularization is a statistical methodology widely used to avoid overfitting if the model includes a large number of predictors. This method allows to select relevant variables from a large set without losing computational efficiency (Derksen and Keselman [1992], Tibshirani [1996]). The regularization technique adds a penalty term to the sum of the model squared residuals grouping the coefficients that are close to zero and penalizing high-valued regression coefficients. Estimates for the regression coefficients are obtained solving the equation

$$min\left\{\frac{1}{2n}||\boldsymbol{Y}-\beta_0\mathbf{1}-\boldsymbol{X}\boldsymbol{\beta}||^2+\lambda||\boldsymbol{\beta}||_q\right\}, \tag{3.1}$$

where $||\boldsymbol{\beta}||_q = \left(\sum_{j=1}^{D}|\beta_j|^q\right)^{\frac{1}{q}}$, $\boldsymbol{Y} = (y_1,\ldots,y_n)^t$ is a n-dimensional response vector, $\mathbf{1}$ denotes the $n\times 1$ vector of ones, $\boldsymbol{X}$ is a $n\times D$ matrix of covariates, $\beta_0 \in \mathbb{R}$ is the intercept, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_D)^t$ is the D-dimensional vector of the regression coefficients and, λ is the penalty parameter. The value of λ is proportional to the level of shrinkage to zero of the coefficients. If $\lambda = 0$ we recover the traditional least squares estimates. There are several types of penalties which are defined by the value of q. If $q = 1$ then we obtain the least absolute shrinkage and selection operator (Lasso) (Tibshirani [1996]). In this case, we add to the model a L_1 penalty equal to the absolute value of coefficients' magnitude. This regularization criterion may provides sparse models as part of the coefficients can be shrunk toward zero and eliminated from the model. If $q = 2$, the Ridge's penalty proposed by Hoerl and Kennard [1970] is obtained. In this case all coefficients are shrunk by the same factor but none is eliminated from the model.

In Bayesian statistics, regularization and variable selection problems have been addressed as a problem of prior specifications. Variable selection procedures have been discussed by several authors and may be found, for instance, in George and McCulloch [1993], Smith and Kohn [1996], Kuo and Mallick [1998], Vannucci et al. [2012] and Li and Zhang [2010]. A commom approach for variable selection is to consider as prior a finite mixture of distributions as in spike-slab method George and McCulloch [1993]. Bayesian methods for regularization consist of building a prior distribution for the regression coefficients that plays a role similar to the penalty term in (3.1). This prior shrinks to zero coefficients that are not significant and puts a significant probability mass to non-despicable coefficients. Bayesian methods rise to several advantages, such as to allow the penalty parameter to be estimated simultaneously with the model, adding more flexibility to the selection process(Van Erp et al. [2019]). Also, the estimation of the parameters can be done using the MCMC methods which, in general, is robust to non-convex or multimodal penaltie, conditions that often generate problems when considering the classic methods. Bayesian shrinkage priors corresponding to the Lasso, grouped Lasso, Elastic nets, and Fused Lasso have been proposed by Park and Casella [2008], Hans [2009], Kyung et al. [2010] and Li and Lin [2010]. In such shrinkage prior formulations, the penalties correspond to the special choices of priors and are expressed as a scale-mixture of normal distribution. For example, in linear regression model, the Ridge's penalty is equivalent to assume a Gaussian prior distribution for $\boldsymbol{\beta}$ centered around zero and with standard deviation defined as a function of λ. Lasso is recovered if the prior distribution for $\boldsymbol{\beta}$ is the Laplace distribution with mean equal to zero and the scale parameter is dependent of λ (James et al. [2013]). For large values of λ, only the most influencial covariates are mantained in the model. These distributions are called *shrinkage priors*. All these methods, however, assume independence among the variables or a dependence structure that are completely known. Both assumptions can be strong in several contexts and, if not accounted for, can lead to poor predictions.

The dependence structure between variables plays an important role in predicting the response and in the presence of a large number of predictors it is desirable to select the ones that significantly affect the response. To account for correlated covariables in normal linear model, Liu et al. [2014] propose a Bayesian regularization approach in which the dependence structure among the covariates is characterized through a graph Laplacian matrix.

As mentioned in Liu et al. [2014], a better predictive capacity of the model can be obtained when the dependence structure among the effects is taken into account. This strategy enables us to borrow information across variables and overcomes colinearity [Storey and Tibshirani, 2003, Kim and Xing, 2009, Li and Li, 2010, Liu et al., 2014]. This is a problem that frequently occurs if the model includes categorical variables with many levels [Gertheiss and Tutz, 2010, Pauger and Wagner, 2017, Criscuolo, 2019]. Such variables are represented using an appropriate dummy coding which requires the introduction of indicator variables representing each level of the categorical feature. If this number of levels is high, the learned coefficients become very unstable, making hard the interpretation of the results. To aggregate such levels is desirable. However, it is not clear how to aggregate them into higher-level categories obtaining an interpretable and statistically efficient model. A response to this problem is provided by [Gertheiss and Tutz, 2010] that proposed a Lasso-constrained regression approach for analysis of variance aiming at collapsing levels of a categorical covariate. Pauger and Wagner [2017] proposed a Bayesian regularization method to aggregate the covariates' levels based on the effect fusion prior. This prior is a modification of the spike-slab distribution as a regularizer considering all level effects as well as their differences. It allows for sparsity and for clustering similar coefficients. A different approach for custering the levels of a categorical variable is considered by [Criscuolo, 2019] which introduced a random partition model to clusterize the coefficients.

Although some of the existents shrinkage priors take into account the correlation among the covariates, none of them consider the correlation induced by the spatial neighbor structure that may, for example, occur in credit risk analysis or when predicting the house rent prices where the geographical location is an important feature.

Motivated by this, our goal is to propose shrinkage prior distributions to model categorical predictors which levels are spatially correlated. Besides sparsity, this prior also has the fuzion property clustering the covariate levels that share a similar effect. We represent these levels as spatial random effects $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ which neighboring structures will be represented by a graph, where r is the number of different categories and θ_i is the effect of the ith level of the categorical covariate. The nodes of the graph will represent the random level effect $\boldsymbol{\theta}$, and the edges will connect neighboring level effects. As in Liu et al. [2014], we assume a multivariate normal prior for $\boldsymbol{\theta}$ which covariance matrix depending on the random weights associated to the edges connecting neighbor nodes. The novelty is the way the prior for this random edges weights are built. Our method explicitly characterizes the dependency structure between two levels effects in $\boldsymbol{\theta}$ through the weights of edges that connect such levels. The variability of each level effect is defined by the weights of edges that are incident in such node.

To define the proposed model let us consider the following definitions. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph where $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of vertices or nodes representing the levels of our categorica variable and $\mathcal{E}$ is the set of p edges connecting unordered pairs of distinct vertices and representing the adjacency relationship among levels. The edge connecting $v_i \in \mathcal{V}$ and $v_j \in \mathcal{V}$ is by $[ij]$. We assume that the edges are undirected, implying that $[ij] = [ji]$. If two nodes v_i and $v_j \in \mathcal{V}$ are connected, this will be denoted by $v_i \sim v_j$. When $v_i \in \mathcal{V}$ is a node in the edge $[ij] \in \mathcal{E}$ we say that the edge is incident on v_i. Associated with the original graph $\mathcal{G}$, we define the *graph of edges* $\mathcal{L}(\mathcal{G})$. The graph of edges represents the adjacency relationship among the edges of the original graph $\mathcal{G}$. The nodes in $\mathcal{L}(\mathcal{G})$ are the edges $[ij] \in \mathcal{E}$ connecting the nodes v_i and v_j, with $i \neq j$. The edges in $\mathcal{L}(\mathcal{G})$ are also determined by the topology of $\mathcal{G}$. Two nodes $[ij]$ and $[kl]$ in $\mathcal{L}(\mathcal{G})$ are adjacent if, and only if, the edges $[ij]$ and $[kl]$ are incident on a common vertex. This means that the pair of neighbouring edges must be of the form $[ij]$ and $[jk]$ for some $v_j \in \mathcal{V}$. Let $\mathcal{I}_i = \{[ik] \in \mathcal{E}, v_k \in \mathcal{V}\}$ be the set of edges incident on area i.

3.3 Proposed model

Consider a sample of n subjects independently selected in the population. Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^t$ denotes the vector of dependent continuous variables, where $y_i \in \mathbb{R}$ is the response for ith subject. Assume that D covariates are measured and let $\boldsymbol{X}_i = (x_{i,1}, ..., x_{i,D})$

be the vector of covariates for the ith subject generating a design matrix $\boldsymbol{X} = (\boldsymbol{X}_1^t, ..., \boldsymbol{X}_n^t)^t$ of order $n \times D$. Such covariates might include quantitative and dummy encoding of categorical variables with few levels. Denote by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)^t$ the D-dimensional vector of regression coefficients that are assumed to be the same for all subjects.

Assume a categorical variable $\boldsymbol{Z}$ with a large number r of levels. As in Criscuolo [2019] we assume that this variable is represented by an undirect graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each level of $\boldsymbol{Z}$ is associated to a vertex in $\mathcal{V}$ and the edges in set $\mathcal{E}$ connects pairs of neighbour vertices or levels in $\mathcal{V}$. The neighborhood structure in $\mathcal{G}$ depends on the categorical variable features. The graph $\mathcal{G}$ is complete if $\boldsymbol{Z}$ is a nominal variable. If $\boldsymbol{Z}$ is ordinal variables, the structure of $\mathcal{G}$ is simplified as we can only connect levels following its natural ordination. If our categorical feature is related to some spatial measurement, the graph $\mathcal{G}$ may represent a map where vertices denote the regions in the map and edges connect regions that are spatially neighbors. Denote by $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{ir})$ the one-hot encoding of $\boldsymbol{Z}$ indicating the vertex to which subject i belongs. Thus, for the subject i, the coordinate $z_{ir} = 1$ if it belongs to vertex r, and it is zero otherwise. Let $\tilde{\boldsymbol{Z}}$ be the $n \times r$ matrix $(\boldsymbol{Z}_1^t, \ldots, \boldsymbol{Z}_n^t)^t$. Consider the r-dimensional vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^t$, where θ_r is the regression coefficient associated to the rth level of $\boldsymbol{Z}$ and it is shared by all responses Y_i in vertex r.

We assume that the vertex effect is additive and that for all subject i there is a linear relationship between y_i, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ such that:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\theta} + \epsilon_i, \tag{3.2}$$

for $i = 1, \ldots, n$, where the errors ϵ_i are independent and identically distributed (iid) as $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The matrix representation of (3.2) is given by:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{\theta} + \boldsymbol{e}, \tag{3.3}$$

where $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_y^2\boldsymbol{I}_n)$ and $\boldsymbol{I}_n$ denotes the identity matix of order n. Consequently, the joint distribution for $\boldsymbol{Y}$, given $\boldsymbol{\Psi} = (\boldsymbol{X}, \tilde{\boldsymbol{Z}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$, is the following n-variate normal distribution

$$\boldsymbol{Y} \mid \boldsymbol{\Psi} \sim N_n(\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}). \tag{3.4}$$

To complete the model specification, for the regression coefficients, we *a priori* assume that $\boldsymbol{\beta}|\sigma \sim N_D(\boldsymbol{\mu}_\beta, \sigma\boldsymbol{\Sigma}_\beta)$, where $\mu_\beta \in \mathbb{R}^D$ is the known vector of prior means and Σ_β is a $D \times D$ symmetric, positive definite matrix; the prior distribution of the model variance σ^2 is the inverse-gamma distribution with parameters $a > 0$ and $b > 0$, denoted by $\sigma^2 \sim IG(a, b)$.

Regarding the categorical variable $\boldsymbol{Z}$, our goal is to reduce dimension, shrinking toward zero the effect of non-significant levels and clustering the effects of significant ones that share similar or equal effects. Besides, we should respect the neighbor structure of the levels of $\boldsymbol{Z}$. With this purpose, we introduce a *shrinkage priori* for the effects $\boldsymbol{\theta}$ which enables grouping levels that are strongly correlated. If our categorical covariate is an indicator of a geographical region, we want to detect groups of areas where $\boldsymbol{Z}$ is expected to take the same value. If no group is detected, the categorical variable are expected to take different values over the complete region. This prior is introduced in next section

3.3.1 Spatial Fusion-Shrinkage Prior for θ

The *spatial fusion-shrinkage prior* distribution (SFS-prior) for the vector $\boldsymbol{\theta}$ is inspired by the distribution defined in Liu et al. [2014]. To this end, let $\rho_{[ji]}$ denotes the random weight for the edge $[ij]$ connecting levels i and j of the categorical covariate $\boldsymbol{Z}$ and denote by $\boldsymbol{\rho}$ the vector of all no-null weights related to graph $\mathcal{G}$. The prior for $\boldsymbol{\theta}$ is hierchically built firstly specifying the conditional distribution of $\boldsymbol{\theta}$ given the random weights $\boldsymbol{\rho}$. Taking the advantage that our representation of the categorical variable defines the Laplacian matrix, as in Liu et al. [2014], we assume that

$$\boldsymbol{\theta}|\boldsymbol{Q}, \sigma^2 \sim N_r(\mathbf{0}, \sigma^2\boldsymbol{Q}^{-1}). \tag{3.5}$$

However, we consider a slightly different structure for the $r \times r$ precision matrix $\boldsymbol{Q}$ precision matrix by considering

$$\boldsymbol{Q} = \begin{bmatrix} 1+\rho_{[1]}+\sum_{j\neq 1}|\rho_{[1j]}| & -\rho_{[12]} & \cdots & -\rho_{[1r]} \\ -\rho_{[21]} & 1+\rho_{[2]}+\sum_{j\neq 2}|\rho_{[2j]}| & \cdots & -\rho_{[2r]} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{[r1]} & \cdots & \cdots & 1+\rho_{[r]}+\sum_{j\neq ir}|\rho_{[r]}| \end{bmatrix}, \tag{3.6}$$

where $\rho_{[i]} > 0$ for all $i = 1, \ldots, r$ and the random weights $\rho_{[ij]} \in \mathbb{R}$ are such that $\rho_{[ij]} = \rho_{[ji]} \neq 0$, if $i \sim j$ and $\rho_{[ij]} = 0$, otherwise. The matrix $\boldsymbol{Q}$ given in (3.6) is positive definite because it is a real, symmetric and strictly diagonally dominant matrix. However, it can be a sparse matrix as it depends on the neighborhood structure.

The elements of $\boldsymbol{Q}$ provide a good conditional interpretation for the proposed model. One of the advantages of the *prior* distribution in 3.5 is that the partial correlation is calculated directly from the precision matrix and is provided directly by the values of the $\boldsymbol{\rho}$ vector. For two elements $(\theta_i, \theta_j) \in \boldsymbol{\theta}$ it follows that

$$Corr(\theta_i, \theta_j | \theta_{-ij}, \boldsymbol{\rho}) = \begin{cases} -\frac{\rho_{[ij]}}{\sqrt{(1+\rho_{[i]}+\sum_{l\neq i}|\rho_{[il]}|)(1+\rho_{[j]}+\sum_{l\neq j}|\rho_{[jl]}|)}} & \text{se } i \sim j \\ 0 \ \text{ otherwise.} \end{cases} \tag{3.7}$$

To obtain a prior for $\boldsymbol{\theta}$ that is able to shrink to zero effects θ_i that are non-significant and to cluster the ones with similar effect, the prior for $\boldsymbol{\rho}$ should be appropriately choosen to guarantee that the random matrix $\boldsymbol{Q}$ is positive definite and it can also capture the dependency structure defined by the nature of the data. Besides, it should be such that the marginal distribution of $\boldsymbol{\theta}$ has a connection with some cluster-type penalty. Under the Liu et al. [2014] approach, the dependency structure between the fixed effects is the regression model is characterized through a full Laplacian matrix and the prior distribution for the unknown elements $\boldsymbol{\rho}$ generate a marginal distribution for $\boldsymbol{\theta}$ that has a direct connection with the cluster-type penalty presented by She [2010]. She [2010] proposed a regularization method based on a L_1-type penalty based on the magnitude of the coefficient and the pairwise difference between them. The prior for $\boldsymbol{\rho}$ proposed by Liu et al. [2014] models a general structure of dependence between the covariates but does not specifically accounts for a neighborhood dependence structure as it occurs if, for instance, the categorical variable represents a geographic or spatial location. We modify such prior distribution to take into account this "spatial feature" considering the neighboring structure in the graph of the edges that connect the levels of our categorical covariate.

Inspired by the work of Liu et al. [2014], we propose a new prior distribution for $\boldsymbol{\rho}$ such that the induced prior marginal distribution for $\boldsymbol{\theta}$ takes into account the spatial correlation between the effects θ_i and θ_j induced by the correlation among the incident edges on these nodes. We consider the adjacency relationship among the edges of the original graph $\mathcal{G}$ represented by the *graph of edges* $\mathcal{L}(\mathcal{G})$. Based on this structure, we build the following prior distribution for $\boldsymbol{\rho}$:

$$p(\boldsymbol{\rho}|\gamma) = C_\theta |\boldsymbol{Q}|^{-1/2} \prod_{i=1}^{r} \rho_{[i]}^{-3/2} \exp\left\{-\frac{\gamma^2}{2\rho_{[i]}}\right\} \prod_{i=2}^{r} \prod_{j<i} |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}|\right\}, \tag{3.8}$$

for all $\rho_{[i]} > 0$ and $\rho_{[ij]} \in \mathbb{R}$, where $n_{[ij]}$ is the number of neighbors of the edge $\rho_{[ij]}$ in the edge graph $\mathcal{L}(\mathcal{G})$, C_θ is a normalizing constant and $\gamma \in \mathbb{R}_+$ is a hyperparameter. The summation involved in the second term of the exponential function in (3.8) is over all neighbors of edge $[ij]$ in the graph $\mathcal{L}(\mathcal{G})$.

The prior distribution in (3.8) is proper and, if mixed with the prior in (3.5), we obtain a prior distribution for $\boldsymbol{\theta}$ that accommodates both sparsity and clustering. These properties are discussed in the following theorems.

Teorema 4. *The prior distribution defined in (3.8) is proper.*

Proof. We have to prove that its integral over ther parametric space is finite. From results in Liu et al. [2014], we have to $|\boldsymbol{Q}| \geq 1$. Using Fubini's theorem, it follows that

$$
\begin{aligned}
&\int |\boldsymbol{Q}|^{-1/2} \prod_{i=1}^{r} \rho_{[i]}^{-3/2} \exp\left\{-\frac{\gamma^2}{2\rho_{[i]}}\right\} \prod_{i=2}^{r}\prod_{j<i} |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}|\right\} d\boldsymbol{\rho} \\
\leq\ &\int \prod_{i=1}^{r} \rho_{[i]}^{-3/2} \exp\left\{-\frac{\gamma^2}{2\rho_{[i]}}\right\} \prod_{i=2}^{r}\prod_{j<i} |\rho_{[ij]}|^{-3/2} \exp\left\{\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}|\right\} d\boldsymbol{\rho} \\
=\ &\prod_{i=1}^{r} \left[\int_0^\infty \rho_{[i]}^{-3/2} \exp\left\{-\frac{\gamma^2}{2\rho_{[i]}}\right\} d\rho_{[i]}\right] \\
&\times \left[\prod_{i=2}^{r}\prod_{j<i} \int_{-\infty}^{\infty} |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}|\right\} d\boldsymbol{\rho}\right]
\end{aligned} \tag{3.9}
$$

The first integral in expression (3.9) is finite as we are integrating the kernel of a inverse-gamma distribution. For the second integral, if $n_{ij} > 0$ it follows that

$$
\begin{aligned}
&\prod_{i=2}^{r}\prod_{j<i} \int_{-\infty}^{\infty} |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}|\right\} d\boldsymbol{\rho} \\
=\ &\prod_{i=2}^{r}\prod_{j<i} \left[\int_{-\infty}^{\infty} |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2|\rho_{[ij]}|}\right\} d\rho_{[ij]} \int_{-\infty}^{\infty} \exp\left\{-n_{[ij]}|\rho_{[kl]}|\right\} d\rho_{[kl]}\right]
\end{aligned}
$$

$$
\begin{aligned}
\propto\ &\prod_{i=2}^{r}\prod_{j<i} \left[\int_{-\infty}^{\infty} |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2|\rho_{[ij]}|}\right\} d\rho_{[ij]} \frac{1}{n_{[kl]}}\right] \\
=\ &\prod_{i=2}^{r}\prod_{j<i} \left[\int_{0}^{\infty} \rho_{[ij]}^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2\rho_{[ij]}}\right\} d\rho_{[ij]} + \int_{-\infty}^{0} -\rho_{[ij]}^{-3/2} \exp\left\{-\frac{n_{[ij]}^2}{2(-\rho_{[ij]})}\right\} d\rho_{[ij]}\right] \\
\propto\ &\prod_{i=2}^{r}\prod_{j<i} \left[n_{ij}^{-1/2} + n_{ij}^{-1/2}\right] < \infty,
\end{aligned}
$$

which concludes the proof. □

Teorema 5. *Assume the distributions in expressions (3.8) and (3.5). Let $c_{ij} = sign(\rho_{ij})$, then, the prior distribution of $\boldsymbol{\theta}$, given σ^2, is given by*

$$p(\boldsymbol{\theta}|\sigma) = \frac{1}{\sigma^{r/2}} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{r} \theta_i^2 + \gamma\sigma \sum_{i=1}^{r} |\theta_i| + \sigma \sum_{i=2}^{r} \sum_{\substack{j<i \\ j\sim i}} n_{[ij]}[|\theta_i - \theta_j| 1_{\rho_{ij}>0} + |\theta_i + \theta_j| 1_{\rho_{ij}<0}] \right) \right], \tag{3.10}$$

where $n_{[ij]}$ is the number of neighbors of the edge $\rho_{[ij]}$ in the edge graph $\mathcal{L}(\mathcal{G})$,$\gamma \in \mathbb{R}_+$ is a hyperparameter and 1_A is the indicator function of event A.

Proof. Assuming the distributions in expression (3.5) and (3.8) and integrating out $\boldsymbol{\rho}$, it follows that the distribution of $\boldsymbol{\theta}$, given σ^2 is

$$\begin{aligned} p(\boldsymbol{\theta}|\sigma^2) &= \frac{1}{\sigma^{r/2}} \int_0^\infty \prod_{i=1}^{r} \exp\left\{ -\frac{1}{2\sigma^2}(1+\rho_{[i]})\theta_i^2 \right\} \rho_{[i]}^{-3/2} \exp\left\{ -\frac{\gamma^2}{2\rho_{[i]}} \right\} d\boldsymbol{\rho} \qquad (3.11) \\ &\times \int_{-\infty}^{\infty} \prod_{i=2}^{r} \prod_{j<i} \exp\left\{ -\frac{1}{2\sigma^2} |\rho_{[ij]}| (\theta_i + c_{ij}\theta_j)^2 \right\} |\rho_{[ij]}|^{-3/2} \exp\left\{ -\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}| \right\} d\boldsymbol{\rho}. \end{aligned}$$

As $a^{-1} \exp\{-a|z|\} = \int_0^\infty (2\pi)^{-1/2} t^{-3/2} \exp\left\{-\frac{z^2 t}{2}\right\} \exp\left\{-\frac{a^2}{2t}\right\} dt$, it follows that first integral in (3.11) is

$$\int_0^\infty \exp\left\{ -\frac{1}{2\sigma^2}(1+\rho_{[i]})\theta_i^2 \right\} \rho_{[i]}^{-3/2} \exp\left\{ -\frac{\gamma^2}{2\rho_{[i]}} \right\} d\rho_{[i]} \propto \exp\left\{ -\frac{1}{2\sigma^2}\theta_i^2 - \frac{\gamma}{\sigma}|\theta_i| \right\}.$$

The second integral in (3.11) is

$$\begin{aligned} &\int_{-\infty}^{\infty} \prod_{i=2}^{r} \prod_{j<i} \exp\left\{ -\frac{1}{2\sigma^2} |\rho_{[ij]}| (\theta_i + c_{ij}\theta_j)^2 \right\} |\rho_{[ij]}|^{-3/2} \exp\left\{ -\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}| \right\} d\boldsymbol{\rho} \\ \propto\ & \prod_{i=2}^{r} \prod_{j<i} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2\sigma^2} |\rho_{[ij]}| (\theta_i + c_{ij}\theta_j)^2 \right\} |\rho_{[ij]}|^{-3/2} \exp\left\{ -\frac{n_{[ij]}^2}{2|\rho_{[ij]}|} \right\} d\rho_{[ij]} \times \prod_{[kl]\sim[ij]} \frac{1}{n_{[kl]}} \\ \propto\ & \prod_{i=2}^{r} \prod_{j<i} \exp\left\{ -n_{[ij]} |\theta_i + c_{ij}\theta_j| \frac{1}{\sigma} \right\} \\ =\ & \prod_{i=2}^{r} \prod_{j<i} \exp\left\{ -\frac{n_{[ij]}}{\sigma} [|\theta_i - \theta_j| 1_{\rho_{ij}>0} + |\theta_i + \theta_j| 1_{\rho_{ij}<0}] \right\}, \end{aligned}$$

which concludes the proof. □

In situations where it is reazonable to assume only positive values for all componentes of $\boldsymbol{\rho}$, a less general prior for $\boldsymbol{\theta}$ that also induces sparsity and grouping can be built assuming that

$$p(\boldsymbol{\rho}|\gamma) = C_\beta|\boldsymbol{Q}|^{-1/2}\prod_{i=1}^{r}\rho_{[i]}^{-3/2}\exp\left\{-\frac{\gamma^2}{2\rho_{[i]}}\right\}\prod_{i=2}^{r}\prod_{j<i}\rho_{[ij]}^{-3/2}\exp\left\{-\frac{n_{[ij]}^2}{2\rho_{[ij]}}-\sum_{[kl]\sim[ij]}\rho_{[kl]}\right\}, \tag{3.12}$$

for $\rho_{[i]} > 0$ and $\rho_{[ij]} > 0$, where $n_{[ij]}$ is the number of neighbors of the edge $\rho_{[ij]}$ in the edge graph $\mathcal{L}(\mathcal{G})$, C_β is a normalizing constant and $\gamma \in \mathbb{R}$ is a hyperparameter. The summation involved in the second term of the exponential function in (3.12) is over all neighbors of edge $[ij]$ in the graph $\mathcal{L}(\mathcal{G})$.

The prior distribution defined in (3.12) is also proper and by mixing it with the distribution in (3.5) we obtain that the prior distribution of $\boldsymbol{\theta}$, given σ^2, is given by

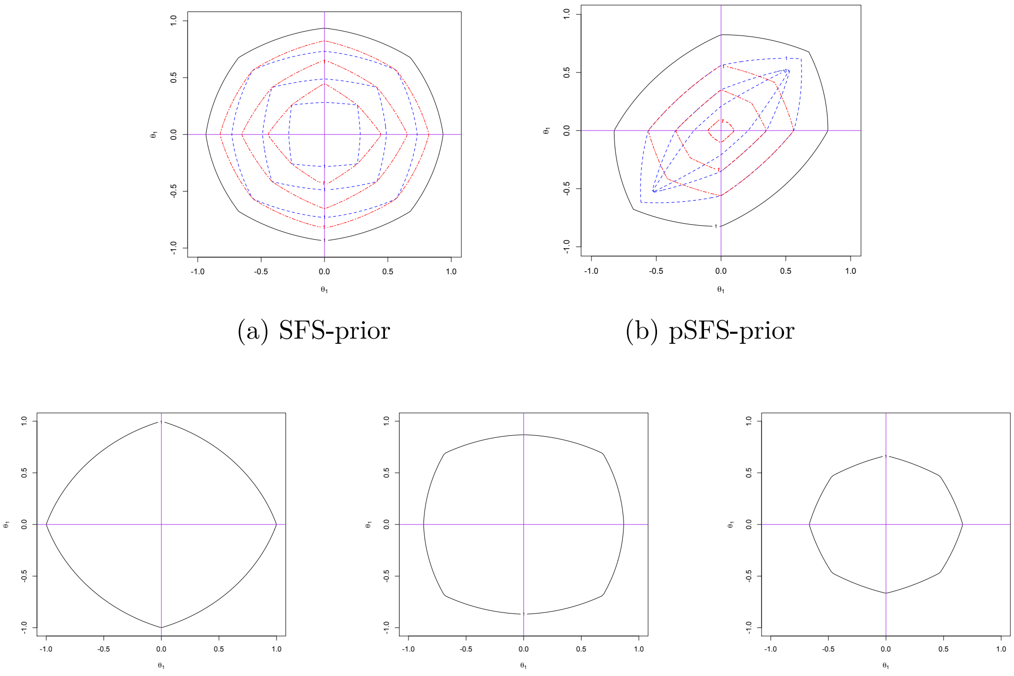
$$p(\boldsymbol{\theta}|\sigma) = \frac{1}{\sigma^{r/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{r}\theta_i^2+\gamma\sigma\sum_{i=1}^{r}|\theta_i|+\sigma\sum_{i=2}^{r}\sum_{\substack{j<i\\ j\sim i}}n_{[ij]}|\theta_i-\theta_j|\right)\right]. \tag{3.13}$$

The positive spatial fusion-shrinkage prior given in 3.13 is denoted by pSFS-Prior.

3.3.1.1 The geometry of the proposed priors for θ

The geometry of the contour curves related to the prior distribution of θ provides evidence about its capacity of clustering similar values and shrink to zero the non-significant ones. Contour curves with a diamond shape, which vertices are located on the horizontal and vertical axes, favor sparsity as the vertices are defined by at least one coordinate equal to zero. The Lasso-prior given in Figure 16 (c) is an example in this family of distributions. If the vertices of this diamond contour curve are over the line $\theta_1 = \pm\theta_2$ as shown in Figure 16 (d), the prior favors the clustering of the effects. The contour curve of the GL-prior distribution proposed by Liu et al. [2014] has an octagonal shape (Figure 16 (e)) meaning that it has both properties.

The prior distributions for $\boldsymbol{\theta}$ given in (3.10) and 3.13 accommodate both sparsity since it includes the term $|\theta_i|$ and grouping by including the terms $|\theta_i - \theta_j|$ and $|\theta_i + \theta_j|$. The constants γ and $n_{[ij]}$ reflect the degree of sparsity and grouping, respectively, induced by such priors. Panels (a) and (b) on Figure 16 show the contour plot for the SFS-prior and pSFS-prior, respectively, providing geometric evidence of these characteristics in the bi-dimensional case. We assume $\sigma^2 = 1$ and varie γ and the number $n_{[ij]}$ of neighbors of the edge $\rho_{[ij]}$ in the edge graph $\mathcal{L}(\mathcal{G})$. The black line in Panel (a) and (b) in Figure 16 respectively show the contour curves of SFS-prior and pSFS-prior if $\gamma = n_{[ij]} = 1$. Assuming this parametrization these priors favor both sparsity and clustering. The dashed (blue) lines (dot-dashed (red) lines, resp.) show the contour curves of SFS-prior and pSFS-prior if $\gamma = 1$ ($\gamma = 2, 4, 6$) and $n_{[ij]} = 2, 4, 6$ ($n_{[ij]} = 1$). From the dashed (blue) lines we notice that, fixing $\gamma = 1$, if $n_{[ij]}$ is large the contour curves of the SFS-prior assume the standard of a grouping prior distribution favoring the clustering of neigbour effects. Fixing $n_{[ij]}$, if γ is large the dot-dashed (red) lines shows it assumes the Lasso-prior shape favoring sparsity.



(a) SFS-prior

(b) pSFS-prior

(c) Lasso

(d) Grouping

(e) GL-prior

Figure 16 – Bidimensional contour plot of $-\log(p(\theta|\sigma^2))$ for the proposed priors, SFS-prior(a) and pSFS-prior (b), Lasso prior (c), Grouping prior (d) and GL-prior (e).

3.4 Posterior Inference

To sample from the posterior distribution, we use a Markov chain Monte Carlo (MCMC) scheme. Thus, we need the full conditional distribution of all parameters. Denote by $\boldsymbol{D} = \{\boldsymbol{X}, \tilde{\boldsymbol{Z}}\}$ the observed design matrix and by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ the independent sample of the response variable. Assuming the model in (3.4) the likelihood function is

$$p(\boldsymbol{Y} \mid \boldsymbol{\Psi}) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{Y} - (\tilde{\boldsymbol{Z}}\boldsymbol{\theta} + \boldsymbol{X}\boldsymbol{\beta})]^t[\boldsymbol{Y} - (\tilde{\boldsymbol{Z}}\boldsymbol{\theta} + \boldsymbol{X}\boldsymbol{\beta})]\right\}.$$

Assuming that, *a priori*, $\boldsymbol{\beta}|\sigma \sim N_D(\boldsymbol{\mu}_\beta, \sigma\boldsymbol{\Sigma}_\beta)$, $\sigma^2 \sim IG(a,b)$ and the distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ given in expressions (3.5) and (3.8), we obtain that

- The posterior full conditional distributions (fcd) of σ^2 is the Inverse-Gamma distribution

  $$\sigma|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{Y}, \boldsymbol{D} \sim IG\left(a + \frac{1}{2}(n + D + r), b + \frac{1}{2}(\tilde{Y} + \tilde{\beta} + \tilde{\theta})\right),$$

  where $\tilde{Y} = (\boldsymbol{Y} - (\tilde{\boldsymbol{Z}}\boldsymbol{\theta} + \boldsymbol{X\beta}))^t(\boldsymbol{Y} - (\tilde{\boldsymbol{Z}}\boldsymbol{\theta} + \boldsymbol{X\beta}))$, $\tilde{\beta} = (\boldsymbol{\beta} - \mu_\beta)^t\Sigma_\beta(\boldsymbol{\beta} - \mu_\beta)$ and $\tilde{\theta} = \boldsymbol{\theta}^t\boldsymbol{Q}^{-1}\boldsymbol{\theta}$.

- The posterior fcd of $\boldsymbol{\beta}$ is the normal distribution

  $$\boldsymbol{\beta}|\sigma, \boldsymbol{Q}, \boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{D} \sim N(\tilde{\boldsymbol{S}}(\boldsymbol{X}^t(\boldsymbol{Y} - \tilde{\boldsymbol{Z}}\boldsymbol{\theta}) + \Sigma_\beta^{-1}\mu_\beta), \sigma\tilde{\boldsymbol{S}})$$

  where $\tilde{\boldsymbol{S}} = (\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{\Sigma}_\beta^{\mathbf{-1}})^{-1}$.

- The posterior fcd of $\boldsymbol{\theta}$ is the normal distribution

  $$\boldsymbol{\theta}|\sigma, \boldsymbol{Q}, \beta, \boldsymbol{Y}, \boldsymbol{D} \sim N_r(\tilde{\boldsymbol{S}}'(\tilde{\boldsymbol{Z}}^t(\boldsymbol{Y} - \boldsymbol{X\beta})), \sigma\tilde{\boldsymbol{S}}')$$

  where $\tilde{\boldsymbol{S}}' = (\tilde{\boldsymbol{Z}}^t\tilde{\boldsymbol{Z}} + \boldsymbol{Q})^{-1}$.

Considering the SFS-prior it is simpler to separately sample from the posterior distribution of $\rho_{[i]}$ and $\rho_{[ij]}$. Let $c_{ij} = sign(\rho_{ij})$. It follows from (3.5) that, given $\boldsymbol{\theta}$ and σ^2, $\rho_{[i]}$ and $\rho_{[ij]}$ are independent. Thus the posterior fcd of $\rho_{[i]}$ is given by

$$p(\rho_{[i]}|\sigma, \boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{D}) \quad \propto \quad \rho_{[i]}^{-3/2} \exp\left\{-\frac{1}{2\sigma^2}\rho_{[i]}\theta_i^2 - \frac{\gamma^2}{2\rho_{[i]}}\right\}$$

which is a inverse-Gaussian distribution with mean $\gamma\sigma|\theta_i|^{-1}$ and shape parameter γ^2 denoted by $\rho_{[i]}|\sigma, \boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{D} \sim IGaussian(\gamma\sigma|\theta_i|^{-1}, \gamma^2)$. This is also the posterior fcd if the pSFS-prior is assumed. The posterior fcd of $\rho_{[ij]}$ assuming the SFS-Prior and pSFS-Prior have unknown closed-form and are given, respectively, by

$$p(\rho_{[ij]}|\sigma, \boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{D}) \quad \propto \quad |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{1}{2\sigma^2}|\rho_{[ij]}|(\theta_i + c_{ij}\theta_j)^2 - \frac{n_{[ij]}^2}{2|\rho_{[ij]}|} - \sum_{[kl]\sim[ij]} |\rho_{[kl]}|\right\};$$

and

$$p(\rho_{[ij]}|\sigma, \boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{D}) \quad \propto \quad |\rho_{[ij]}|^{-3/2} \exp\left\{-\frac{1}{2\sigma^2}|\rho_{[ij]}|(\theta_i + c_{ij}\theta_j)^2 - \frac{n_{[ij]}^2}{2\rho_{[ij]}} - \sum_{[kl]\sim[ij]} \rho_{[kl]}\right\}.$$

To sample from the posterior, we propose the following Gibbs sampler scheme with Metropolis-Hasting step. We choose to use the sampling random draws from the truncated normal with a random walk approach. This is $\rho'_{[ij]} \sim N(\rho_{t-1}, \sigma_p^2)$. For the value σ_p^2, convergence types must be analyzed to find the optimal value.

Algoritmo 1: MCMC scheme to sample from the posterior distribution.

```
Input: $\boldsymbol{D}, \boldsymbol{Y}$ ;
initialization($\sigma^{2(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\rho}^{(0)}$) ;
for $t$=1 to $T$ do
    $\sigma^{2(t)} \sim p(\sigma|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{Q}^{(t-1)}, \boldsymbol{Y}, \boldsymbol{D})$ ;
    $\boldsymbol{\beta}^{(t)} \sim p(\boldsymbol{\beta}|\sigma^{(t)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{Q}^{(t-1)}, \boldsymbol{Y}, \boldsymbol{D})$ ;
    $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta}|\sigma^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{Q}^{(t-1)}, \boldsymbol{Y}, \boldsymbol{D})$ ;
    for $i$=1 to $n$ do
        $\rho_{[i]} \sim p(\rho_{[i]}|, \sigma^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{Y}, \boldsymbol{D})$;
    end
    for [ij]=1 to $p$ do
        $\rho'_{[ij]} \sim N(\rho_{t-1}, \sigma_p^2)$;
        if $\alpha(\rho'_{[ij]}|\rho^{t-1}_{[ij]}) \geq unif(0,1)$ then
            $\rho^t_{[ij]} = \rho'_{[ij]}$ else
                $\rho^t_{[ij]} = \rho^{t-1}_{[ij]}$
            end
            $\boldsymbol{Q}^{(t)} = f(\boldsymbol{\rho}_{[i]}, \boldsymbol{\rho}_{[ij]})$
        end
    end
end
```

3.5 Simulation study

To evaluate the performance of the proposed approaches for clustering and shrink to zero non-significant effects we consider a regression model in (3.4) where the categorical variable is related to a spatial effect. We consider the map of Brazil divided in its 27 States assuming that $\boldsymbol{\theta}$ has dimension $r = 27$. We generate $T = 150$ samples of size $n = 27,000$ $(1,000$ within each State) such that each observation Y_i belonging to the rth cluster is generated from the normal distribution $N(1 + \theta_r, 1)$.

Three different scenarios are assumed. In Scenario 1, we assume that there is no spatial cluster thus 27 different effects θ_r are independently generated from the distribution $Uniform(-10, 10)$. In Scenarios 2 and 3, three clusters are considered. For $r = 1, \ldots, 3$, we fixed $\theta_r = 800, 100, 10$ in Scenario 2 and in $\theta_r = -800, 10, 800$ in Scenario 3. Such clusters are shown in Figure 17(a).

To analyse the data, we consider the proposed priors given in expressions (3.10) and (3.13) assuming $\gamma = 1$. We aslso assume that, *a priori*, $\beta \sim N(0, 0.1)$, and $\sigma^2 \sim Gamma(10^{-2}, 10^{-2})$. For each data set, we collect 500 MCMC iterations after discarding the first 100 as burn-in.

We compare SFS-prior and pSFS-prior with the unpenalized maximum likelihood (MaxLik), the PPRM [Criscuolo, 2019] model and some models based on different shrinkage priors: Lasso [Park and Casella, 2008, Hans, 2009, Kyung et al., 2010, Li and Lin, 2010], grouping [She, 2010], the EffectFusion [Pauger and Wagner, 2017] and the GL-prior [Liu et al., 2014] methods. The PPRM is a different approach based on random partition that simultaneously carries out the model fitting and the aggregation of similar categorical levels into larger groups.

To compare the models, we consider some usual model selection criteria: the Deviance Information Criteria (DIC), the extended Akaike information criterion (AIC) and the Root Mean Squared Error (RMSE) given by $RMSE = \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{27} (\hat{Y}_{it} - Y_{it})^2 / 27T}$, where Y_{it} and $\hat{Y}_{it}$ are the generated count and its predicte value at area i at replication t. We also calculate the false positives and negatives rates related to the cluster estimation. To evaluate the false positive rate, we estimate the clusters from the posterior estimates of $\theta's$ by testing the differences between all $\theta's$. To that end, we consider the posterior 95% confidence intervals for the differences $\theta_i - \theta_j$ clustering these two effects if zero belongs to such an interval.If the built clusters do not conform to the originals, we have a false positive classification. To calculate the false-negative rate, the posterior estimated $\theta's$ are divided according to the original cluster, then the difference of the estimated values within each cluster is evaluated. If this difference is approximately zero, the methods correctly estimated the cluster; if the difference is significant, the method does not estimate the cluster correctly corresponding to a false negative classification. This procedure is carried out for each sample replication. It is calculated the percentage of false negative and false positive classifications based on the total of replicas.

For MaxLik we use the `R` package `MaxLik` (Umlauf et al. [2018]). For Lasso, Grouping, EffectFusion and GL-prior we use the available codes in the author's webpages. The code for PPRM was grantted by Tulio Criscuolo. Our models were implemented using the software R Team [2015].

Table 4 shows that for the scenario where there is no cluster and no null effect (Scenario 1), MaxLik had a better performance. However, according to DIC, the shrinkage priors which allow for spatial association among the categorical effects (SFS-Prior, pSFS-prior, and GL-Prior) provided comparable results. However, the SFS-Prior is the only model that presents a value greater than zero in the false positive which indicates that, probably, this prior distribution forces too much to cluster when it is not necessary. For Scenario 2 in which the categorical variable has positive effects in all three clusters, the pSFS-prior leads to better model fitting presenting the smallest DIC and AIC. In this case, SFS-prior and GL-prior also have a reazonable performance according to AIC and DIC, respectively. For Scenario 3, the best model fitting is obtained assuming the SFS-Prior. In this scenario the GL-prior that also accounts for spatial association among the effects has the second best performance. In summary, the proposed approaches were competitive even in scenarios that did not favor their features.

Figure 17 shows the clusters induced by the posterior estimates of θ_i in Scenario 3. We calculate the average of the posterior means in each State and put into the same cluster the States for which such averages are not significantly differents. We consider the highest density posterior intervals for $\theta_i - \theta_j$, with probability 0.95 to determine the significant differences. Excepting for Maxlink, Grouping and GL-Prior methods, all other models perfectly recover the original clustering structure. This is partially in agreement with our findings in Table 4 that shows that EffectFusion, PPRM, GL-prior,pSFS-Prior and SFS-prior perform similarly in this scenario.

3.6 Case Study: Inside Airbnb Datasets

In this section, we fit the proposed and all other model considered in Section 3.5 to analyze the dataset related to the rental price per night of full houses or apartment registred at Airbnb platform for Toronto and London cities. Toronto is divided in $R = 136$ regions an a sample of $n = 4340$ obsrvations is colected. In London there is $R = 33$ regions and we observe a sample of size $n = 24573$. The Airbnb rent dataset is available in the R package named catdata (Schauberger and Tutz [2020]).

In this study, the house/apartment location plays an important role in the determination of the rent price [Gertheiss and Tutz, 2010]. This categorical variable Z, which is of fundamental importance in our study, usually has many levels which are the geographical

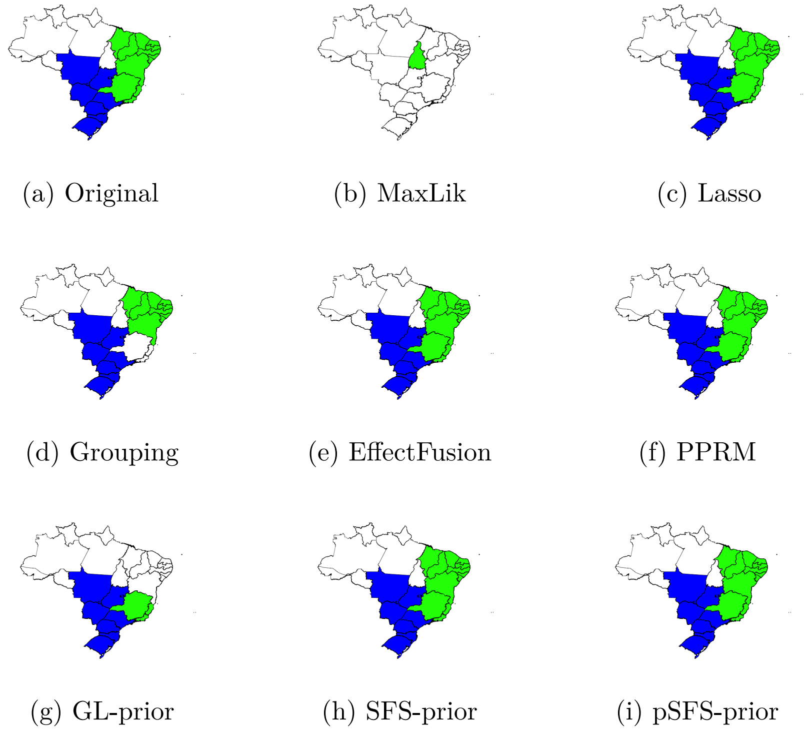
Model	MSE	DIC	AIC	FN	FP
		Scenario 1			
Maxlik	0.30	**4436.02**	**4463.21**		0.00
Lasso	0.35	4728.32	4756.32		0.00
Grouping	0.48	6245.62	6245.66		0.00
EffectFussion	0.52	6529.12	5821.72		0.00
PPRM	0.51	7764.17	7803.02		0.00
GL-prior	0.35	4834.09	7901.60		0.00
pSFS-prior	**0.32**	4992.61	5649.61		0.00
SFS-prior	0.34	4464.35	4914.63		**0.20**
		Scenario 2			
Maxlik	0.67	11014.22	11089.51	1.00	0.00
Lasso	0.29	10604.51	10604.52	0.75	0.00
Grouping	**0.11**	11345.32	11345.53	0.40	0.00
EffectFussion	0.22	11345.21	11346.23	0.40	0.20
PPRM	0.13	**7163.82**	12725.62	0.25	**0.15**
GL-prior	0.12	8733.91	13362.92	0.25	0.25
pSFS-prior	0.12	7279.81	**10089.52**	**0.20**	0.25
SFS-prior	0.13	8940.71	11032.22	0.25	0.25
		Scenario 3			
Maxlik	0.57	17521.82	12522.93	1.00	0.00
Lasso	0.23	12057.67	12557.62	1.00	0.00
Grouping	0.23	11057.68	11157.82	1.00	0.15
EffectFussion	0.11	1284.40	3598.04	1.00	0.15
PPRM	0.11	1264.40	3358.54	0.50	0.25
GL-prior	0.10	1107.71	1799.11	0.50	0.25
pSFS-prior	0.11	1259.74	2893.99	0.25	0.25
SFS-prior	**0.10**	**935.83**	**1658.72**	**0.25**	**0.25**

Table 4 – Model comparison criteria for all fitted models.

areas in a region, such as the ZIP codes in a metropolitan region. As in Criscuolo [2019], we assume that these levels are organized in a planar graph $\mathcal{G}$ representing the geographical adjacency relationship between them, where each node represents a neighborhood, and the edges connect the neighborhood that shares a geographic border.

We consider the pre-processed dataset from Criscuolo [2019] selecting only listings with at least one review and rentals of a full house or apartment. We consider 19 continuous covariates or discrete ones with a small number of levels. The response variable is the logarithm of the price per night of stay. All these variates are described in Table 5.

To analyse the data, we assume *a priori* that $\beta_i \overset{iid}{\sim} N(0, \sigma^2)$, $\theta_i \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 \sim Gamma(10^{-2}, 10^{-2})$, $\mu_\beta = 0$ and $\gamma = 1$. To initialize the MCMC chains we let $\sigma^0 = 1$,



(a) Original (b) MaxLik (c) Lasso

(d) Grouping (e) EffectFusion (f) PPRM

(g) GL-prior (h) SFS-prior (i) pSFS-prior

Figure 17 – Clusters of the spatial effects for each model.

$\boldsymbol{\beta}^0 \sim \boldsymbol{N}(\mathbf{0}, \sigma^0 \boldsymbol{I})$, $\boldsymbol{\theta}^0 \sim \boldsymbol{N}(\mathbf{0}, \sigma^0 \boldsymbol{I})$ and for each $\rho^0_{[ij]} \sim \text{Uniforme}(0,1)$, $i, j \in \{1, \ldots, p\}$ and for each $\rho^0_{[i]} \sim \text{Uniforme}(0,1)$, $i \in \{1, \ldots, n\}$. All models are fitted by collecting 10000 MCMC iterations after discarding the first 500 as burn-in.

Table 6 shows the results of the RMSE, DIC, and AIC for all methods. In general, for all datasets, SFS-Prior, pSFS-Prior, GL-prior, and PPRM are comparable models better fitting the data according to RMSE and DIC criteria. Assuming these criteria, pSFS-Prior is the best model produce the best fit for London. PPRM is the best model for London data if we consider AIC.

	Toronto $n = 4340$ $R = 136$			London $n = 24573$ $R = 33$		
	RMSE	DIC	AIC	RMSE	DIC	AIC
Maxlik	0.58	13580.33	13580.33	0.52	10931.31	10931.31
Lasso	0.32	12895.04	12892.06	0.27	9275.11	9334.06
Grouping	0.35	13500.09	13500.09	0.66	11100.00	11100.03
EffectFussion	0.39	13500.04	13501.02	0.38	9275.37	9334.46
PPRM	0.31	12788.79	12187.71	0.21	8680.03	**8681.31**
GL-prior	0.29	12712.12	12924.57	0.21	8782.02	9536.98
pSFS-prior	0.29	12683.25	129140.75	**0.21**	**8332.82**	9294.63
SFS-prior	**0.21**	**12611.93**	**12029.19**	0.22	8340.58	9831.74

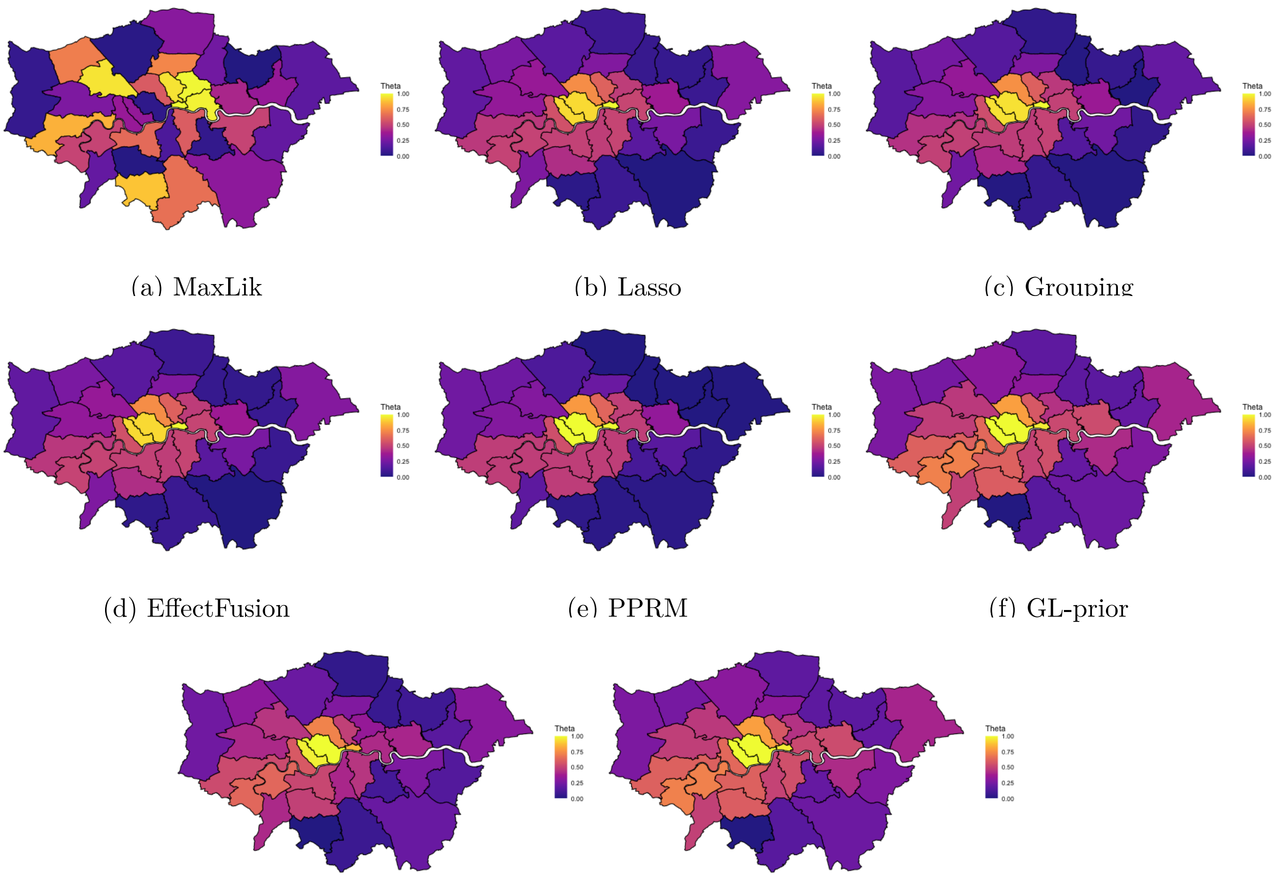
Table 6 – Model comparison criteria for all fitted models.

Variable	Type	Description	Values
fire extinguisher	cat	Amenity indicator	0/1
indoor fireplace	cat	Amenity indicator	0/1
dishwasher	cat	Amenity indicator	0/1
family kid friendly	cat	Amenity indicator	0/1
coffeemaker	cat	Amenity indicator	0/1
freestreet street parking	cat	Amenity indicator	0/1
Iron	cat	Amenity indicator	0/1
Dryer	cat	Amenity indicator	0/1
Airconditioning	cat	Amenity indicator	0/1
TV	cat	Amenity indicator	0/1
is superhost	cat	Indicator if a host is superhost	0/1
guests included	cat	Number of guests included	$(0,1],(1,2],(2,\infty)$
review scores value	cat	Mean of reviews scores value	$[0,9],(9,10]$
review scores cleanliness	cat	Mean of reviews score cleanliness	$[0,9],(9,10]$
bathrooms	cat	Number of bathrooms	$(0,1],(1,2],(2,\infty)$
accommodates	cat	Number of persons that can be accommodates	$(0,2],(2,4],(4,\infty)$
bedrooms	cat	Number of bedroom	$(0,1],(1,2],(2,\infty)$
minimum nights	cat	Minimum nights required to book the listing	$(0,3],(3,7],(7,30],(30,\infty)$
property type	cat	Type of the listed property	House/Apartment
Z	cat	Neighborhood	Neighbourhood name
Y	num	log of the listing price per night	$(0,\infty)$

Table 5 – Explanatory variables for monthly rent per square meter

The Figures 19 and 18 show the posterior means of the effects θ_i of the neighborhood indicator for each area of Toronto and London regions, respectively, provided by the different methods. For the London data set, we learn from PPRM that, with high posterior probability (> 0.95), the spatial diversity can be represented by only three clusters. However, the posterior means for the neighboring effects indicate different effects for a significant number of areas under all methods. With the exception of the MaxLik, all other methods point to the same three central areas as the ones with the highest effects on the rental price. In general, such methods indicate that the neighboring effects are small for areas located far from this central areas. The pSFS-prior and GL-prior tend to over smooth the effects of neighboring areas making challenging to identify the number of clusters of similar areas. Considering these methods we estimated nine clusters, and for SFS-prior, eight clusters. Estimates provided by Maxlik method are not affected by the spatial correlation that may occur among the areas. The PPRM is a method for cluster identifications. Consequently, it identifies well-delimited clusters.

The proposed methods SFS-prior and pSFS-prior do not carry out a precise identification of the clusters. Moreover, the SFS-prior tends to over smooth the clusters' limits because it takes into account two features, clustering, and sparsity, which do not allow the clusters to be fully identified. You may also notice the similarity between the GL-prior and the pSFS-prior because their configuration of the covariance matrix is similar. It is important notice that our model indicates similar behaviors as the PPRM approach,



(a) MaxLik

(b) Lasso

(c) Grouping

(d) EffectFusion

(e) PPRM

(f) GL-prior

(g) SFS-prior

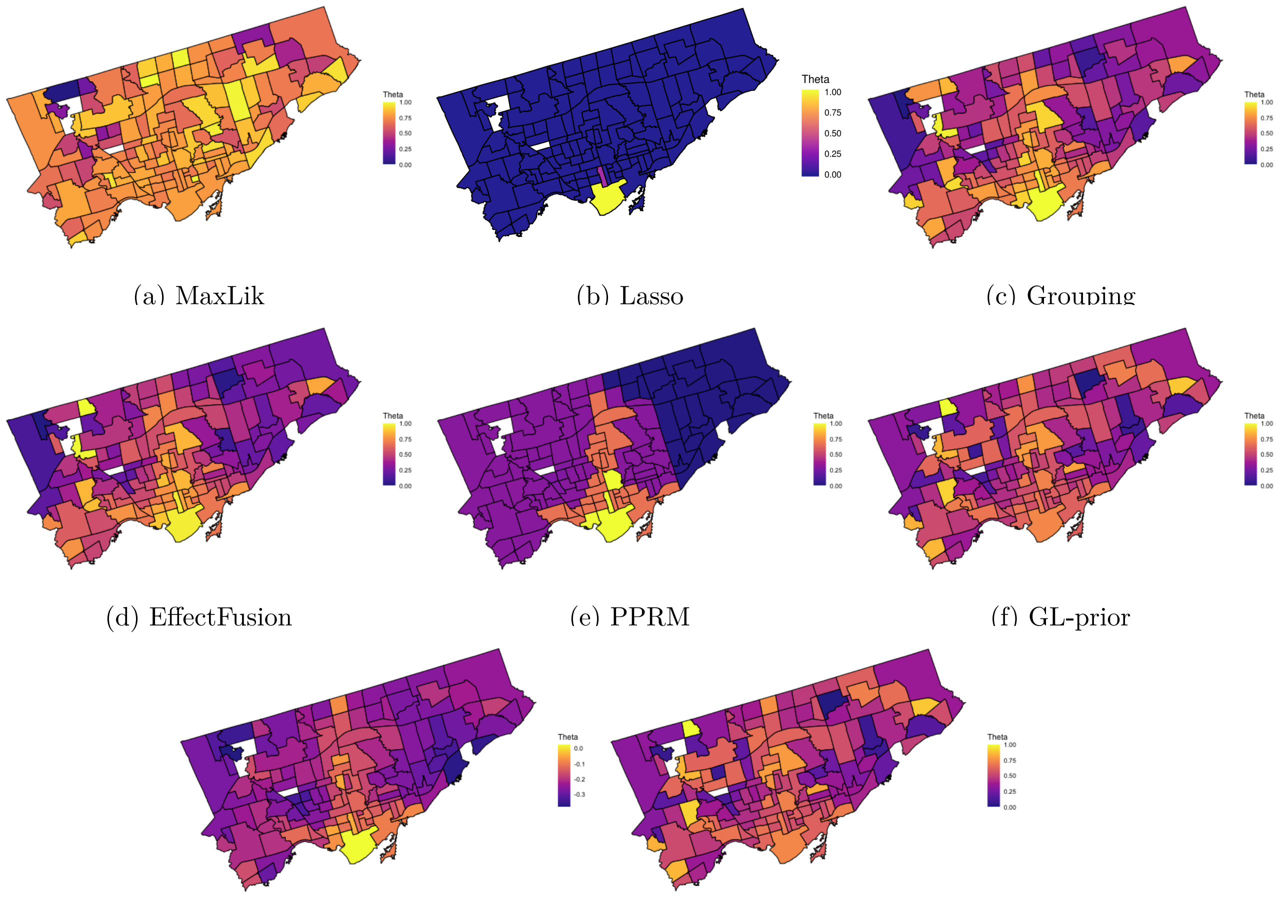
(h) pSFS-prior

Figure 18 – Posterior means (min-max normalized) for the neighborhood-vertices effects $\boldsymbol{\theta}$ in each neighborhood using differents types of penalization, London data

which we use to benchmark its ability to separate districts.

For Toronto data, Lasso points that we have only three different effects for the complete map being two of than related to two neighbohoods. All other areas have the same effect in the rental price. PPRM also indicate four groups of areas, two of them containing a great number of areas which identical neighboring effects. SFS-Prior and pSFS-prior indicates more heterogenety among the neighboring effects. However, under the SFS-Prior, neighbor areas are more prone to experience similar effect. pSFS-prior and GL-Prior provided very similar estimates for the effects.

Figure 20 shows the 95% posterior credible intervals for the neighborhood effects for Toronto (left) and London(right) for model SFS-prior and PPRM. All areas have significant, positive spatial effects according to both methods which points to solid evidence of clearly



(a) MaxLik

(b) Lasso

(c) Grouping

(d) EffectFusion

(e) PPRM

(f) GL-prior

(g) SFS-prior

(h) pSFS-prior

Figure 19 – Posterior means (min-max normalized) for the neighborhood-vertices effects $\boldsymbol{\theta}$ in each neighborhood using differents types of penalization, Toronto data

differentiated geographic price regimes.

We have more uncertainty about the neighboring effects under SFS-prior for both datasets since we obtained higher wide for posterior credible intervals. Besides, for Toronto dataset, the SFS-prior indicate a higher effect of the neighboring in the rental price and also a higher number of areas with similar effects. For London dataset we can not see a well defined delimitation of the clusters, even so, after evaluating the specific features associated with the houses, we conclude that the average price fluctuates depending on geographical aspects.

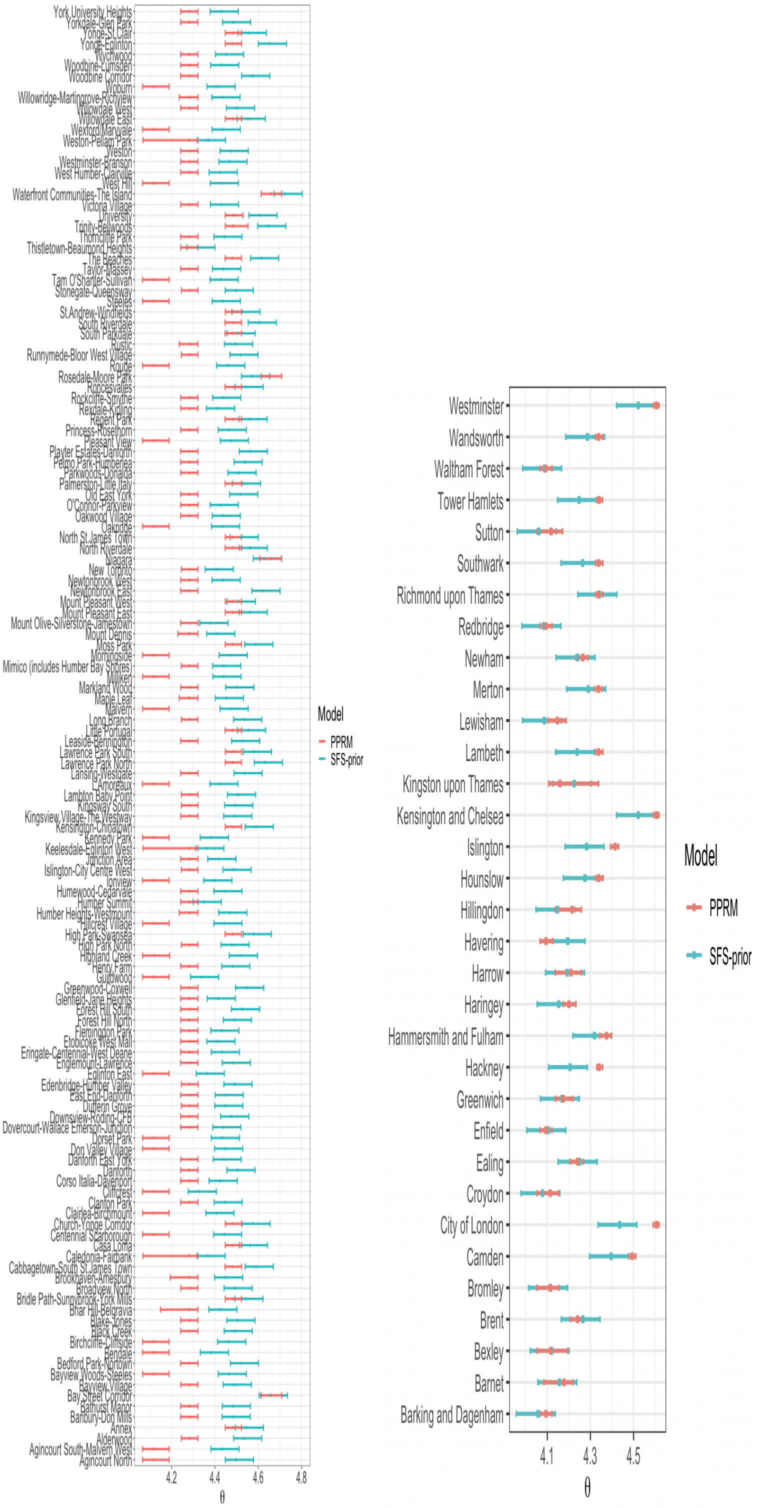


Figure 20 – Centered 95% posterior credible interval for the neighborhood-vertices effects, Toronto (left) and London (right) datasets.

3.7 Conclusion

We present an approach to analyzing categorical variables with a high number of levels and a correlation between these levels. For this, an a priori distribution inspired by group Lasso and GL-prior for geographical predictors called shrinkage prior is proposed.

We investigated the performance of prior shrinkage by comparing it with Lasso, grouping, effect fusion, and the GL-prior type penalties, and not penalized type commonly used in a study of categorical variables. In addition, we compare the performance with a clustering method that served as a guide to evaluating the generation of possible clusters. The proposed shrinkage prior performance was superior to some current methods, like a lasso, groping, and even fussed methods and GL-prior.

We show two different types of a priori distributions, one that only considers clusters of the same sign and the other that feels different signs. It is important to highlight these two distributions since it is convenient to use only correlations of the same signal in many applications in spatial statistics so that the model would be more appropriate.

The proposed method was also applied to real data that corresponds to the data from the Airbnb dataset for 2007. Criscuolo [2019] had already shown that Lasso's behavior was not enough to infer the clusters. However, here it is illustrated that they recover in a very similar way without the extra computational load of the method proposed by Criscuolo [2019] still, the technique does not delimit the clusters and tends to smooth the edges, showing a smoothing behavior and not exclusively clustering.

One of the advantages of the model is that it simultaneously estimates the values of the precision matrix that correspond to the edge weights, however, the behavior of the edge weights is not optimized yet, which requires a more in-depth investigation.

3.8 Bibliography

Tulio L. Criscuolo. *Modelo partição produto para atributos categóricos*. PhD thesis, Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, 2019.

Shelley Derksen and H. J. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.

Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Jan Gertheiss and Gerhard Tutz. Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, 4(4):2150 – 2180, 2010.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 09 2009. ISSN 0006-3444.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

Seyoung Kim and Eric P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLOS Genetics*, 5(8):1–18, 08 2009.

Lynn Kuo and Bani Mallick. Variable selection for regression models. 1998.

Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 06 2010.

Caiyan Li and Hongzhe. Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.*, 4(3):1498–1516, 09 2010. doi: 10.1214/10-AOAS332.

Fan Li and Nancy R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214, 2010.

Qing Li and Nan Lin. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 03 2010.

Fei Liu, Sounak Chakraborty, Fan Li, Yan Liu, and Aurelie C. Lozano. Bayesian regularization via graph laplacian. *Bayesian Analysis*, 9(2):449–474, 06 2014.

Trevor Park and George Casella. BAMLSS: The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–685, 2008.

Daniela Pauger and Helga Wagner. Bayesian effect fusion for categorical predictors. 2017.

Gunther Schauberger and Gerhard Tutz. *catdata: Categorical Data*, 2020. R package version 1.2.2.

Yiyuan She. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4: 1055–1096, 2010.

Michael Smith and Robert Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. ISSN 0027-8424.

R Core Team. R: a language and environment for statistical computing. r foundation for statistical computing. vienna, austria., 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

Nikolaus Umlauf, Nadja Klein, and Achim Zeileis. BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627, 2018.

Sara Van Erp, Daniel L. Oberski, and Joris Mulder. Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50, 2019. ISSN 0022-2496.

Marina Vannucci, Francesco C. Stingo, and Carlo Berzuini. Bayesian models for variable selection that incorporate biological information. 9780199694587, January 2012. Publisher Copyright: © Oxford University Press 2011. All rights reserved. Copyright: Copyright 2018 Elsevier B.V., All rights reserved.

4 Conclusões

Neste trabalho abordamos dois problemas distintos envolvendo dados espaciais. O primeiro é a hipersuavização promovida por alguns modelos espaciais. O segundo envolve a redução de dimensionalidade em problemas de regressão contendo variáveis categóricas de alta dimensão e cujos níveis são espacialmente correlacioados.

No primeiro caso, propomos um modelo espacial robusto que explica a limitação do modelo CAR para levar em consideração a alta correlação espacial, assim suavizando descontinuidades. Consideramos uma distribuição $t-$Student para os efeitos aleatórios. A novidade em nossa abordagem é a forma como a correlação espacial é induzida. Atribuímos efeitos aleatórios espaciais às *arestas* do grafo de vizinhança. O efeito aleatório espacial de cada área é uma combinação linear dos efeitos das arestas incidentes. Para os efeitos das arestas incidentes, atribuímos uma distribuição multivariada $t-$Student onde a matriz de covariância espacial tem uma estrutura semelhante CAR, induzindo um comportamento de cauda pesada para os efeitos aleatórios espaciais. Provamos que o modelo RENeGe proposto induz uma correlação marginal maior do que o modelo CAR, aliviando uma das principais limitações dos modelos CAR e ICAR.

O modelo proposto mostrou que, comparado ao modelo CAR, responde melhor pela heterogeneidade proporcionando uma melhor reconstrução de imagem. Para os conjuntos de dados de câncer, o modelo proposto supera muitos outros modelos espaciais previamente introduzidos na literatura, mostrando ser competitivo para considerar a correlação espacial. Porém, a proposta apresentou um comportamento inesperado com respeito às correlações condicionais, que mudaram de sinal dependendo do grau de vizinhança entre as variáveis.

Na segunda parte, apresentamos uma abordagem para analisar variáveis categóricas com um grande número de níveis e uma correlação entre esses níveis. Para isso, foram propostas duas distribuições *a priori* de encolhimento para preditores inspiradas no Lasso de agrupamento She [2010] e a GL-prior Liu et al. [2014]. As distribuições propostas são denominadas *shrinkage prior* e são chamadas SFS-prior e pSFS-prior. A pSFS-prior assume que os efeitos têm sinal positivo e a SFS-prior considera efeitos reais. A pSFS-prior foi construída pois, em muitas aplicações em estatísticas espaciais, é conveniente usar apenas correlações de mesmo sinal entre os efeitos espaciais para que o modelo seja mais adequado.

Investigamos o desempenho das distribuições SFS-prior e pSFS-prior comparando-o com Lasso, Lasso de agrupamento, fusão Lasso, GL-prior e quando não tem penalização, as quais são comumente usadas em um estudo de variáveis categóricas. Além disso, comparamos o desempenho com um método de clustering que serviu de guia para avaliar a geração de possíveis clusters proposto por Criscuolo [2019]. O desempenho das distribuições SFS-prior e pSFS-prior foi superior a alguns métodos, como um Lasso e GL-prior.

O método proposto também foi aplicado a dados reais que correspondem aos dados do conjunto de dados do Airbnb de 2007. Criscuolo [2019] já havia mostrado que o comportamento do Lasso não era adequado para inferir os clusters de áreas com efeitos iguais. Ficou ilustrado com este exemplo que os métodos propostos recuperam de forma muito semelhante à de métodoss pré-existente a inferência sobre o efeito da variável vizinhança no preço do aluguel. No entanto, estas técnicas também não delimitam os clusters e tende a suavizar as arestas, apresentando um comportamento de suavização e não exclusivamente agrupamento.

Uma das vantagens do modelo é que ele estima simultaneamente os valores da matriz de precisão que correspondem aos pesos das arestas, porém o comportamento dos pesos das arestas ainda não está otimizado, o que requer uma investigação mais aprofundada.

4.1 Bibliography

Tulio L. Criscuolo. *Modelo partição produto para atributos categóricos*. PhD thesis, Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, 2019.

Fei Liu, Sounak Chakraborty, Fan Li, Yan Liu, and Aurelie C. Lozano. Bayesian regularization via graph laplacian. *Bayesian Analysis*, 9(2):449–474, 06 2014.

Yiyuan She. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4: 1055–1096, 2010.

Appendix

APPENDIX A – Prova da matriz definida positiva do modelo CAR

O objetivo desta seção é demostrar a equivalência sob a matrizes de covariancias: $(\boldsymbol{M}-\gamma\boldsymbol{A})^{-1}$ e $(\boldsymbol{I}-\gamma\boldsymbol{W})^{-1}\boldsymbol{M}^{-1}$. Para esto, precisamos da seguente definição tomada de Harville [2012].

Definição 2 (Matrizes semelhantes)**.** *Duas matrizes quadradas $\boldsymbol{D}$ e $\boldsymbol{E}$ são semelhantes (ou similares) se existir uma matriz invertível $\boldsymbol{M}$ tal que:*

$$\boldsymbol{D}=\boldsymbol{M}^{-1}\boldsymbol{E}\boldsymbol{M}$$

e as matrizes semelhantes, tem as seguentes propriedades (A prova pode ser encontrada em Harville [2012], Teorema **21.3.1.**):

Proposição 1. *Se $\boldsymbol{D}$ e $\boldsymbol{E}$ são semelhantes então possuem o mesmo polinômio característico e tem os mesmos valores próprios com a mesma multiplicidade.*

Proposição 2. *Seja a matriz $(\boldsymbol{I}-\gamma\boldsymbol{W})^{-1}\boldsymbol{M}^{-1}$ do tamanho $n\times n$ onde $w_{i,i}=0$ e $w_{i,j}\neq 0$ se i e j são vizinhos. Seja $\{\lambda_i\}$ o conjunto de valores próprio da matriz $\boldsymbol{W}$. A matriz $\boldsymbol{W}$ é simetrica, por tanto, seus valores próprios são reais. Seja λ_1 ser o menor valor proprio da matriz $\boldsymbol{W}$ e λ_p ser o maior valor proprio da matriz $\boldsymbol{W}$. Se $1/\lambda_1<\gamma<1/\lambda_p$, então $(\boldsymbol{I}-\gamma\boldsymbol{W})$ é definida positiva.*

demonstração.

Na primeira parte temos que:

$$\begin{aligned}((\boldsymbol{I}-\gamma\boldsymbol{W})^{-1}\boldsymbol{M}^{-1})^{-1} &= \boldsymbol{M}(\boldsymbol{I}-\gamma\boldsymbol{W})\\ &= \boldsymbol{M}^{1/2}(\boldsymbol{I}-\gamma\boldsymbol{M}^{1/2}\boldsymbol{W}\boldsymbol{M}^{-1/2})\boldsymbol{M}^{1/2}\end{aligned}$$

por tanto, $(\boldsymbol{I}-\gamma\boldsymbol{W})^{-1}\boldsymbol{M}^{-1}$ é positiva definida se $(\boldsymbol{I}-\gamma\boldsymbol{M}^{1/2}\boldsymbol{W}\boldsymbol{M}^{-1/2})$ é positiva definida.

Por outro lado,

$$\boldsymbol{M}^{-1/2}(\boldsymbol{I}-\gamma\boldsymbol{M}^{1/2}\boldsymbol{W}\boldsymbol{M}^{-1/2})\boldsymbol{M}^{1/2} = (\boldsymbol{I}-\gamma\boldsymbol{W})$$

o seja, $(\boldsymbol{I} - \gamma \boldsymbol{M}^{1/2}\boldsymbol{W}\boldsymbol{M}^{-1/2})$ tem valores próprios positivos se e somente se $(\boldsymbol{I} - \gamma \boldsymbol{W})$ tem valores próprios positivos já que são matrizes semelhantes.

Então, para provar que $(\boldsymbol{I} - \gamma \boldsymbol{W})^{-1}\boldsymbol{M}^{-1}$ é positivamente definida é suficiente provar que $(\boldsymbol{I} - \gamma \boldsymbol{W})$ é definida positiva.

Seja ω um valor próprio de $(\boldsymbol{I} - \gamma \boldsymbol{W})$, então,

$$(\boldsymbol{I} - \gamma \boldsymbol{W})\boldsymbol{x} = \omega \boldsymbol{x}$$

e seja $\boldsymbol{v}_i$ o vetor próprio correspondente ao valor próprio ω. Então se $\boldsymbol{x} = \boldsymbol{v}_i$,

$$\begin{aligned} \boldsymbol{v}_i - \gamma \boldsymbol{W}\boldsymbol{v}_i &= \omega_i \boldsymbol{v}_i \\ \rightarrow \boldsymbol{v}_i - \gamma \lambda_i \boldsymbol{v}_i &= \omega_i \boldsymbol{v}_i \\ \rightarrow (1 - \gamma \lambda_i)\boldsymbol{v}_i &= \omega_i \boldsymbol{v}_i \\ \rightarrow (1 - \gamma \lambda_i) &= \omega_i \end{aligned}$$

então, para todo $(1 - \gamma\lambda_i) > 0$ vai implicar que $\omega_i > 0$, isto acontece se $1 > \gamma\lambda_i, \forall i$

.

Se $\lambda_i < 0 \rightarrow 1/\lambda_i < \gamma$ e se $\lambda_i > 0 \rightarrow 1/\lambda_i > \gamma$.

Para todo $\lambda_i < 0$, só $1/\lambda_1 < \gamma$ vai garantir $(1 - \gamma\lambda_i) > 0$ e para todo $\lambda_i > 0$ só $1/\lambda_p > \gamma$ vai garantir $(1 - \gamma\lambda_i) > 0$.

Assim, $1/\lambda_1 < \gamma < 1/\lambda_p$ garantirá que todos os valores próprios de $(\boldsymbol{I} - \gamma \boldsymbol{W})$ sejan positivos.

Proposição 3. *Seja a matriz $(\boldsymbol{M} - \gamma \boldsymbol{A})$ do tamanho $n \times n$ onde $a_{i,i} = 0$ e $a_{i,j} \neq 0$ se i e j são vizinhos. Seja $\{\lambda_i\}$ o conjunto de valores próprio da matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$. A matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$ é simetrica, por tanto, seus valores próprios são reais. Seja λ_1 ser o menor valor proprio da matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$ e λ_p ser o maior valor proprio da matriz $\boldsymbol{M^{-1/2}AM^{-1/2}}$. Se $1/\lambda_1 < \gamma < 1/\lambda_p$, então $(\boldsymbol{M} - \gamma \boldsymbol{A})$ é definida positiva.*

demonstração. A prova é análoga ao anterior sempre que:

$$(\boldsymbol{M} - \gamma \boldsymbol{A}) = \boldsymbol{M}^{1/2}(\boldsymbol{I} - \gamma \boldsymbol{M}^{-1/2}\boldsymbol{A}\boldsymbol{M}^{-1/2})\boldsymbol{M}^{1/2}$$

o seja, a matriz $(\boldsymbol{M}-\gamma\boldsymbol{A})$ tem valores próprios positivos se e somente se $(\boldsymbol{I}-\gamma\boldsymbol{M^{-1/2}}\boldsymbol{A}\boldsymbol{M^{-1/2}})$ tem valores próprios positivos.

Aparentemente estamos chegando a dois resultados diferentes, por um lado temos que a condição $1/\lambda_1 < \gamma < 1/\lambda_p$ é valida se λ_i são os valores próprios de $\boldsymbol{W}$ e por outro lado temos que a condição e válida se λ_i são os valores próprios de $\boldsymbol{M}^{-1/2}\boldsymbol{A}\boldsymbol{M}^{-1/2}$. A solução e simple, o que acontece é que as matrizes $\boldsymbol{W}$ e $\boldsymbol{M}^{-1/2}\boldsymbol{A}\boldsymbol{M}^{-1/2}$ são semelhantes, já que

$$\boldsymbol{W} = \boldsymbol{M}^{-1}\boldsymbol{A} = \boldsymbol{M}^{-1/2}(\boldsymbol{M}^{-1/2}\boldsymbol{A}\boldsymbol{M}^{-1/2})\boldsymbol{M}^{1/2}$$

logo, tem os mesmos valores próprios.

APPENDIX B – Further Results for Chapter 2

B.1 Proof of expansion presented in Section 2.4

We need to prove that the expression defined in 2.11. We have,

$$\operatorname{Cov}[\boldsymbol{\theta}] = \boldsymbol{C}(\boldsymbol{I} + \gamma \boldsymbol{W}_e + \gamma^2 \boldsymbol{W}_e^2 + \gamma^3 \boldsymbol{W}_e^3 + \cdots)\boldsymbol{M}_e^{-1}\boldsymbol{C}^t. \tag{B.1}$$

$$\begin{aligned}\operatorname{Cov}[\boldsymbol{\theta}] &= \boldsymbol{C}(\boldsymbol{I} + \gamma \boldsymbol{W}_e + \gamma^2 \boldsymbol{W}_e^2 + \gamma^3 \boldsymbol{W}_e^3 + \cdots)\boldsymbol{M}_e^{-1}\boldsymbol{C}^t. \\ &= \boldsymbol{C}\boldsymbol{I}\boldsymbol{M}_e^{-1}\boldsymbol{C}^t + \gamma \boldsymbol{C}\boldsymbol{W}_e\boldsymbol{M}_e^{-1}\boldsymbol{C}^t + \gamma^2 \boldsymbol{C}\boldsymbol{W}_e^2\boldsymbol{M}_e^{-1}\boldsymbol{C}^t + \cdots\end{aligned}$$

First $\boldsymbol{M}_e^{-1}\boldsymbol{C}^t$ is a $p \times n$ matrix, such that $(\boldsymbol{M}_e^{-1}\boldsymbol{C}^t)_{ei} = 1/d_{[ij]}$ if edge e is incident on node i, $d_{[ij]}$ is the number of neighbors of the edge $[ij]$ on the *graph of edges* $\mathcal{L}(\mathcal{G})$, and $(\boldsymbol{M}_e^{-1}\boldsymbol{C}^t)_{ei} = 0$, otherwise.

Then, to perform the test, we divide into each of the components:

First component: $\boldsymbol{C}\boldsymbol{M}_e^{-1}\boldsymbol{C}^t$

For $i \sim j$, the element (i, j) in this matrix is given by

$$\frac{1}{d_{[ij]}}$$

while, if $i = j$, the element of the diagonal in this matrix is given by

$$\sum_{[ij] \in \mathcal{I}_i} \frac{1}{d_{[ij]}}$$

Second component: $\boldsymbol{C}\boldsymbol{W}_e\boldsymbol{M}_e^{-1}\boldsymbol{C}^t$

$\boldsymbol{C}\boldsymbol{W}_e$ is a $n \times p$ matrix, such that

$$(\boldsymbol{C}\boldsymbol{W}_e)_{ie} = \sum_{[ik] \sim e} \frac{1}{d_{[ik]}}, \quad \text{if edge } [ik] \text{ is incident on node } i$$

$(\boldsymbol{CW}_e)_{ie} = 0$, otherwise. Then, for $i \sim j$, the element (i,j) in this matrix is given by

$$(\boldsymbol{CW}_e)(\boldsymbol{M}_e^{-1}\boldsymbol{C}^t)_{ij} = \sum_{[jl]\in\mathcal{I}_j} \frac{1}{d_{[jl]}} \left(\sum_{[ik]\sim[jl]} \frac{1}{d_{[ik]}} \right), \text{ if edge } [ik] \text{ is incident on node } i$$

while, if $i = j$, the element of the diagonal in this matrix is given by

$$(\boldsymbol{CW}_e)(\boldsymbol{M}_e^{-1}\boldsymbol{C}^t)_{ii} = \sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} \left(\sum_{[ik]\sim[ij]} \frac{1}{d_{[ik]}} \right), \text{ if edge } [ik] \text{ is incident on node } i$$

Third component: $\boldsymbol{CW}_e^2\boldsymbol{M}_e^{-1}\boldsymbol{C}^t$

Here it gets a little more complicated. First, the matrix $\boldsymbol{W}_e^2$ is a $p \times p$ matrix, such that the element (e,w) is defined by:

$$\frac{1}{d_{[e]}} \sum_{\substack{[lr]\sim[e] \\ [lr]\sim[w]}} \frac{1}{d_{[lr]}},$$

then,the matrix $\boldsymbol{CW}_e^2$ is a $n \times p$ matrix, such that the element (i,e) is defined by:

$$\sum_{[ik]\in\mathcal{I}_i} \frac{1}{d_{[ik]}} \left(\sum_{\substack{[lr]\sim[ik] \\ [lr]\sim[e]}} \frac{1}{d_{[lr]}} \right),$$

$(\boldsymbol{CW}_e^2)_{ie} = 0$ if there is no edge $[lr]$ that is neighboring $[e]$ and $[ik]$. Then, for $i \sim j$, the element (i,j) in this matrix is given by

$$(\boldsymbol{CW}_e^2)(\boldsymbol{M}_e^{-1}\boldsymbol{C}^t)_{ij} = \sum_{[jl]\in\mathcal{I}_j} \frac{1}{d_{[jl]}} \left(\sum_{[ik]\in\mathcal{I}_i} \frac{1}{d_{[ik]}} \sum_{\substack{[lr]\sim[ik] \\ [lr]\sim[jl]}} \frac{1}{d_{[lr]}} \right),$$

while, if $i = j$, the element of the diagonal in this matrix is given by

$$(\boldsymbol{CW}_e^2)(\boldsymbol{M}_e^{-1}\boldsymbol{C}^t)_{ij} = \sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} \left(\sum_{[ik]\in\mathcal{I}_i} \frac{1}{d_{[ik]}} \sum_{\substack{[lr]\sim[ik] \\ [lr]\sim[ij]}} \frac{1}{d_{[lr]}} \right).$$

Finally, the expression up to the third order is:

For $i \sim j$, the element (i,j) in this matrix is given by

$$\frac{1}{d_{[ij]}} + \gamma \sum_{[jl]\in\mathcal{I}_j} \frac{1}{d_{[jl]}} \left(\sum_{\substack{[ik]\in\mathcal{I}_i \\ [ik]\sim[jl]}} \frac{1}{d_{[ik]}} \right) + \gamma^2 \sum_{[jl]\in\mathcal{I}_j} \frac{1}{d_{[jl]}} \left(\sum_{[ik]\in\mathcal{I}_i} \frac{1}{d_{[ik]}} \sum_{\substack{[lr]\sim[ik] \\ [lr]\sim[jl]}} \frac{1}{d_{[lr]}} \right) + \ldots,$$

while, if $i = j$, the element of the diagonal in this matrix is given by

$$\sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} + \gamma \sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} \left(\sum_{\substack{[ik]\in\mathcal{I}_i \\ [ik]\sim[ij]}} \frac{1}{d_{[ik]}} \right) + \gamma^2 \sum_{[ij]\in\mathcal{I}_i} \frac{1}{d_{[ij]}} \left(\sum_{[ik]\in\mathcal{I}_i} \frac{1}{d_{[ik]}} \sum_{\substack{[lr]\sim[ik] \\ [lr]\sim[ij]}} \frac{1}{d_{[lr]}} \right) + \ldots .$$

B.2 Posterior full conditional distributions

In this section, we provide details on the calculations of the posterior full conditional distributions (fcd) that are required to sample from the posterior distributions.

To define a distribution *a priori* of γ, it is necessary to keep in mind that $\gamma \in (\lambda_p^{-1}, \lambda_1^{-1})$ where λ_1 be the smallest value of the matrix $\boldsymbol{M^{-1/2}AM^{-1/2}}$ and λ_p be the highest value of the matrix $\boldsymbol{M^{-1/2}AM^{-1/2}}$. (Besag [1974]), as this parameter corresponds to the same parameter defined in CAR model for the edge graph. Thus, we assign the distribution prior to γ as a restricted uniform to these two values, namely, $\gamma \sim Unif(\lambda_p^{-1}, \lambda_1^{-1})$.

For β we adopt an distribution prior Normal $\beta \sim N(\mu_\beta, \tau_\beta^{-1})$; τ_y we consider a prior distribution $Gamma(\alpha, \eta)$, $\alpha > 0$ and $\eta > 0$. For $\tau_{\boldsymbol{\theta}}$ we consider a prior distribution $Gamma(\alpha_0, \eta_0)$ with $\alpha_0 > 0$ and $\eta_0 > 0$.

Let the vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ be an independent and identically distributed sample from a $n-$dimensional multivariate normal distribution, the likelihood for the data is the product of distributions normal:

$$p(y|\beta, \boldsymbol{\theta}, \tau_y) \propto \tau_y^{-n/2} \exp\left\{-\frac{\tau_y}{2}[\boldsymbol{y} - (\beta + \boldsymbol{\theta})]^t[\boldsymbol{y} - (\beta + \boldsymbol{\theta})]\right\}.$$

Assuming independence between the prioris, then we have to

$$\begin{aligned}
\boldsymbol{\theta}, \boldsymbol{\Sigma}, \beta, \tau_y, \tau_{\boldsymbol{\theta}}, \gamma, |\boldsymbol{y}, \quad &\propto \quad \tau_y^{n/2} \exp\left(-\frac{\tau_y}{2}(\boldsymbol{y} - (\beta + \boldsymbol{\theta}))^t (\boldsymbol{y} - (\beta + \boldsymbol{\theta}))\right) \\
&\times \quad |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2(\nu - n - 1)} \boldsymbol{\theta} \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}\right) \\
&\times \quad \tau_{\boldsymbol{\theta}}^n |\boldsymbol{\Sigma}|^{-(\nu+n+1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{K}\Sigma^{-1})\right) \\
&\times \quad \exp\left\{-\frac{1}{2\tau_\beta^{-2}}(\beta - \mu_\beta)^2\right\} \\
&\times \quad \tau_y^{\alpha-1} \exp(-\tau_y \eta) \\
&\times \quad \tau_{\boldsymbol{\theta}}^{-\alpha_0 - 1} \exp(-\tau_{\boldsymbol{\theta}}^{-1} \eta_0) \\
&\times \quad P(\gamma),
\end{aligned}$$

where $\boldsymbol{K} = \tau_{\boldsymbol{\theta}}^{-1} \boldsymbol{C}(\boldsymbol{D} - \gamma \boldsymbol{A})^{-1} \boldsymbol{C}^t$.. We can obtain the conditional distributions for each of the parameters $(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \beta, \tau_y, \tau_{\boldsymbol{\theta}}, \gamma)$.

Posterior fcd of $\boldsymbol{\theta}$

We have to $\boldsymbol{\theta}|\boldsymbol{\Sigma}, \beta, \tau_y, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y}$

$$\begin{aligned}
&\propto \quad \exp\left(-\frac{\tau_y}{2}(\boldsymbol{y} - (\boldsymbol{\beta} + \boldsymbol{\theta}))^t (\boldsymbol{y} - (\boldsymbol{\beta} + \boldsymbol{\theta})) - \frac{1}{2(\nu - n - 1)} \boldsymbol{\theta} \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}\right) \\
&\propto \quad \exp\left(-\frac{1}{2}[\boldsymbol{\theta}^t((\nu - n - 1)\boldsymbol{\Sigma})^{-1} \boldsymbol{\theta} - 2\tau_y (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^t \boldsymbol{\theta} + \boldsymbol{\theta}^t (\tau_y \boldsymbol{I}) \boldsymbol{\theta}]\right) \\
&\propto \quad \exp\left(-\frac{1}{2}[\boldsymbol{\theta}^t[((\nu - n - 1)\boldsymbol{\Sigma})^{-1} + \tau_y \boldsymbol{I}]\boldsymbol{\theta} - 2\tau_y (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^t \boldsymbol{\theta}]\right),
\end{aligned}$$

considering $\boldsymbol{\Phi} = \tau_y \boldsymbol{I} + (\nu - n - 1)^{-1} \boldsymbol{\Sigma}^{-1}$

$$\boldsymbol{\theta}|\boldsymbol{\Sigma}, \beta, \tau_y, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y} \quad \propto \quad \exp\left\{-\frac{1}{2}[(\boldsymbol{\theta} - \tau_y \boldsymbol{\Phi}^{-1}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}))^t \boldsymbol{\Phi} (\boldsymbol{\theta} - \tau_y \boldsymbol{\Phi}^{-1}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}))]\right\},$$

so we have to

$$\boldsymbol{\theta}|\boldsymbol{\Sigma}, \beta, \tau_y, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y} \sim MVN\left(\tau_y \boldsymbol{\Phi}^{-1}(\boldsymbol{y} - \boldsymbol{\beta}), \boldsymbol{\Phi}^{-1}\right).$$

Posterior fcd of $\boldsymbol{\Sigma}$

$$\begin{aligned}\boldsymbol{\Sigma}|\boldsymbol{\theta},\beta,\tau_y,\tau_{\boldsymbol{\theta}},\gamma,\boldsymbol{y} &\propto |\boldsymbol{\Sigma}|^{-n/2}\exp\left(-\frac{\tau_{\boldsymbol{\theta}}}{2(\nu-n-1)}\boldsymbol{\theta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right)\\ &\times |\boldsymbol{\Sigma}|^{-(\nu+n+1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{K}\Sigma^{-1})\right),\end{aligned}$$

taking all the properties of features and determinants into account, with $\boldsymbol{S}$ it is the sample of squared matrix sums of $\boldsymbol{\theta}$, we have,

$$\begin{aligned}&\boldsymbol{\Sigma}|\boldsymbol{\theta},\beta,\tau_y,\tau_{\boldsymbol{\theta}},\gamma,\boldsymbol{y}\\ &\quad= |\boldsymbol{\Sigma}|^{-n/2}|\boldsymbol{\Sigma}|^{-(\nu+n+1)/2}\exp\left(-\frac{1}{2(\nu-n-1)}\mathrm{Tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})\right)\exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{K}\Sigma^{-1})\right)\\ &\quad= |\boldsymbol{\Sigma}|^{-n/2-(\nu+n+1)/2}\exp\left(-\frac{1}{2(\nu-n-1)}\mathrm{Tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})-\frac{1}{2}\mathrm{Tr}(\boldsymbol{K}\Sigma^{-1})\right)\\ &\quad= |\boldsymbol{\Sigma}|^{-(\nu+n+m+1)/2}\exp\left(-\frac{1}{2(\nu-n-1)}\mathrm{Tr}\left(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}+(\nu-n-1)\boldsymbol{K}\Sigma^{-1}\right)\right)\\ &\quad= |\boldsymbol{\Sigma}|^{-((\nu+n)+m+1)/2}\exp\left(-\frac{1}{2(\nu-n-1)}\mathrm{Tr}\left(\left(\boldsymbol{S}+(\nu-n-1)\boldsymbol{K}\right)\boldsymbol{\Sigma}^{-1}\right)\right)\end{aligned}$$

$$\boldsymbol{\Sigma}|\boldsymbol{y},\boldsymbol{\theta},\gamma\sim IW_p\left(\nu+m,\boldsymbol{S}+(\nu-n-1)\boldsymbol{K}\right).$$

Therefore, ν essentially acts as the number of observations we had observed before collecting the data. The estimate of the covariance matrix given by the *a posteriori* distribution is a linear combination between the sample covariance of $\boldsymbol{\theta}$ and the covariance of the original graph which is defined by the covariance of the edge graph.

Posterior fcd of β

$$\begin{aligned}\beta|\boldsymbol{\Sigma},\boldsymbol{\theta},\tau_y,\tau_{\boldsymbol{\theta}},\gamma,\boldsymbol{y}, &\propto \tau_y^{n/2}\exp\left(-\frac{\tau_y}{2}(\boldsymbol{y}-(\beta+\boldsymbol{\theta}))^t(\boldsymbol{y}-(\beta+\boldsymbol{\theta}))\right)\\ &\times \exp\left\{-\frac{1}{2\tau_\beta^{-2}}(\beta-\mu_\beta)^2\right\},\end{aligned}$$

therefore, we have to

$$\beta|\boldsymbol{\Sigma}, \boldsymbol{\theta}, \tau_y, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y}, \quad \sim \quad N\left(\frac{(\boldsymbol{y}-\boldsymbol{\theta})\tau_y}{n\tau_y+\tau_\beta^{-1}}, \frac{1}{\sqrt{n\tau_y+\tau_\beta^{-1}}}\right).$$

This result can be verified Hoff [2009].

Posterior fcd of τ_y

$$\tau_y|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \beta, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y}, \quad \propto \quad \tau_y^{n/2+\alpha+1}\exp\left[-\tau_y\left(\eta+\frac{1}{2}(\boldsymbol{y}-(\beta+\boldsymbol{\theta}))^t(\boldsymbol{y}-(\beta+\boldsymbol{\theta}))\right)\right],$$

that is, that $\tau_y|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \beta, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y}$ has distribution

$$\tau_y|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \beta, \tau_{\boldsymbol{\theta}}, \gamma, \boldsymbol{y}, \quad \sim \quad Gamma(\alpha+n/2, \eta+\frac{1}{2}(\boldsymbol{y}-(\beta+\boldsymbol{\theta}))^t(\boldsymbol{y}-(\beta+\boldsymbol{\theta}))).$$

Posterior fcd of $\tau_{\boldsymbol{\theta}}$

$$\tau_{\boldsymbol{\theta}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \tau_y, \beta, \gamma, \boldsymbol{y} \quad \propto \quad \tau_{\boldsymbol{\theta}}^{-\nu/2-\alpha_0-1}\exp\left[-\tau_{\boldsymbol{\theta}}^{-1}\left(\mathrm{Tr}(\boldsymbol{C}(\boldsymbol{D}-\gamma\boldsymbol{A})^{-1}\boldsymbol{C}^t\boldsymbol{\Sigma}^{-1})/2+\eta_0\right)\right],$$

that is, that $\tau_{\boldsymbol{\theta}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \tau_y, \beta, \gamma, \boldsymbol{y}$ has distribution:

$$\tau_{\boldsymbol{\theta}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \tau_y, \beta, \gamma, \boldsymbol{y} \quad \sim \quad Gamma(\alpha_0+\nu/2, \mathrm{Tr}(\boldsymbol{C}(\boldsymbol{D}-\gamma\boldsymbol{A})^{-1}\boldsymbol{C}^t\boldsymbol{\Sigma}^{-1})/2+\eta_0).$$

Posterior fcd of $\boldsymbol{\gamma}$

$$\gamma|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \tau_y, \tau_{\boldsymbol{\theta}}, \beta, \boldsymbol{y} \propto |\boldsymbol{\Sigma}|^{-(\nu+m+1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{K}\Sigma^{-1})\right)P(\gamma),$$

$$\gamma|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \tau_y, \tau_{\boldsymbol{\theta}}, \beta, \boldsymbol{y} \sim \exp\left(-\frac{1}{2}\mathrm{Tr}(\tau_{\boldsymbol{\theta}}^{-1}\boldsymbol{C}(\boldsymbol{D}-\gamma\boldsymbol{A})^{-1}\boldsymbol{C}^t\boldsymbol{\Sigma}^{-1})\right)P(\gamma). \tag{B.2}$$

The γ parameter does not have a known distribution and a *Metropolis-Hasting* sampling is required. To sample the posterior distribution of γ we make a reparametrizacion $U = logit(\gamma)$. This method of transforming the parameter space through a variable altering method within the Metropolis-Hastings algorithm is useful for problems with constrained parameter spaces, in this case $\gamma \in (\lambda_p^{-1}, \lambda_1^{-1})$, Givens and Hoeting [2012].

B.3 Additional results and graphs of the analysis of colon/rectal, stomach, and prostate cancers(Section 2.6.3)

Coefficients	Mean	95% HPD	Mean	95% HPD	Mean	95% HPD
			Colon/rectal cancer			
	RENeGe		CAR		GSN	
Intercept	-0.10	(-0.38, 0.18)	-0.08	(-0.32, 0.18)	0.08	(-0.22, 0.29)
$IDHM$	0.36	(-0.08, 0.79)	0.34	(-0.07, 0.72)	0.08	(-0.25, 0.53)
τ_θ	0.14	(0.10, 0.19)	0.12	(0.07, 0.19)	0.00	(0.00, 0.00)
ρ	0.75	(0.48, 0.95)	0.01	(0.00, 0.03)	0.06	(0.06, 0.06)
σ^2	1.84	(1.40, 2.3)	-	-	-	-
	Leroux		BYM		HND	
Intercept	-0.10	(-0.38, 0.18)	-0.08	(-0.32, 0.18)	0.08	(-0.22, 0.29)
$IDHM$	0.36	(-0.08, 0.79)	0.34	(-0.07, 0.72)	0.08	(-0.25, 0.53)
τ_θ	0.14	(0.10, 0.19)	0.12	(0.07, 0.19)	0.00	(0.00, 0.00)
ρ	0.75	(0.48, 0.95)	0.01	(0.00, 0.03)	0.06	(0.06, 0.06)
	DAGAR		GLM			
Intercept	0.08	(-0.32, 0.18)	-0.50	(-0.60, -0.40)		
$IDHM$	0.34	(-0.07, 0.72)	1.20	(1.10, 1.30)		
τ_θ	0.12	(0.07, 0.19)	-	-		
ρ	0.00	(0.00, 0.01)	-	-		
			Stomach cancer			
	RENeGe		CAR		GSN	
Intercept	-1.00	(-1.93, -0.16)	-0.80	(-2.09, 0.36)	-1.93	(-2.34,-1.50)
$IDHM$	2.01	(0.70, 3.47)	1.69	(-0.12, 3.71)	3.46	(2.78, 4.09)
τ_θ	1.58	(1.20, 2.18)	9.78	(7.88, 12.39)	0.10	(0.08, 0.13)
ρ	-1.07	(-2.28, 0.76)	0.37	(0.05, 0.71)	0.06	(0.06, 0.06)
σ^2	1.08	(0.73, 2.28)	-	-	-	-
	Leroux		BYM		HND	
Intercept	-2.09	(-2.51, -1.59)	-0.50	(-1.85, 0.67)	-1.93	(-2.34, -1.50)
$IDHM$	3.71	(2.93, 4.36)	1.22	(-0.60, 3.32)	3.46	(2.78, 4.09)
τ_θ	3.01	(2.01, 5.19)	2.50	(1.37, 3.77)	0.10	(0.08, 0.13)
ρ	0.13	(0.02, 0.40)	1.16	(0.87, 1.58)	0.06	(0.06, 0.06)
	DAGAR		GLM			
Intercept	-0.50	(-1.85, 0.67)	-2.75	(-2.85, -2.65)		
$IDHM$	1.22	(-0.60, 3.32)	4.57	(4.47, 4.67)		
τ_θ	2.50	(1.37, 3.77)	-	-		
ρ	0.00	(0.00, 0.01)	-	-		
			Prostate cancer			
	RENeGe		CAR		GSN	
Intercept	-0.10	(-0.57, 0.44)	1.18	(0.73, 1.63)	0.51	(0.15, 0.73)
$IDHM$	-0.94	(-1.77, -0.22)	-2.92	(-3.60, -2.21)	-1.88	(-2.23, -1.27)
τ_θ	3.60	(1.97, 4.59)	14.20	(11.39, 18.03)	0.26	(0.20, 0.33)
ρ	-1.10	(-2.28, 0.76)	0.09	(0.00, 0.32)	0.06	(0.06, 0.06)
σ^2	1.10	(0.65, 2.96)	-	-	-	-
	Leroux		BYM		HND	
Intercept	1.28	(0.13, 2.21)	0.03	(-0.91, 0.98)	0.51	(0.15, 0.73)
$IDHM$	-3.10	(-4.55, -1.29)	-1.16	(-2.61, 0.31)	-1.88	(-2.23, -1.27)
τ_θ	2.96	(2.27, 4.17)	4.31	(2.93, 6.01)	0.26	(0.20, 0.33)
ρ	0.02	(0.00, 0.09)	1.55	(1.18, 2.09)	0.06	(0.06, 0.06)
	DAGAR		GLM			
Intercept	0.03	(-0.91, 0.98)	1.31	(1.21, 1.41)		
$IDHM$	-1.16	(-2.61, 0.31)	-2.89	(-2.99 , -2.79)		
τ_θ	4.31	(2.93, 6.01)	-	-		
ρ	1.55	(1.18 , 2.09)	-	-		

Table 7 – Posterior means and 95% HPD intervals under all models.

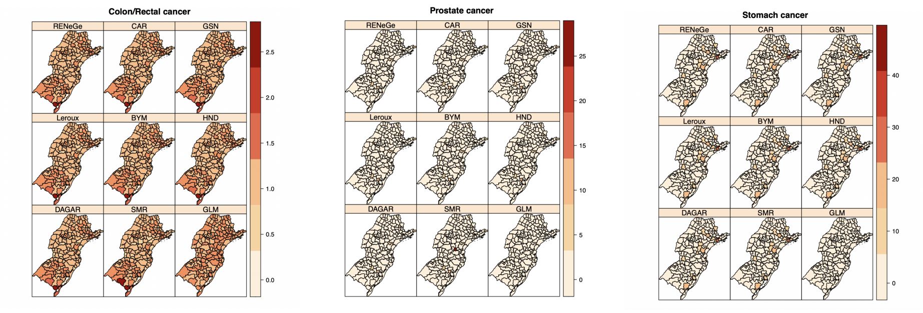


Figure 21 – Posterior estimates for the relative risk of the Colon/Rectal, Prostate cancer and Stomach cancer mortality in the southern region of Brazil under all fitted models.

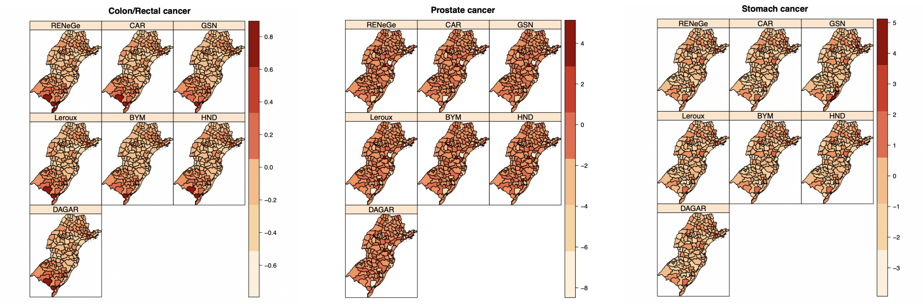


Figure 22 – Posterior estimates for the random effects of the Colon/Rectal, Stomach cancer Prostate cancer mortality in the southern region of Brazil under all fitted models.

B.4 Bibliography

Besag, J. [1974]. Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B* **36**(2): 192–236.

Givens, G. H. and Hoeting, J. A. [2012]. *Computational statistics*, 2 edn, John Wiley & Sons, Hoboken, NJ, USA.

Harville, D. [2012]. *Matrix Algebra From a Statistician's Perspective*, Vol. 40.

Hoff, P. D. [2009]. *A First Course in Bayesian Statistical Methods*, 1st edn, Springer Publishing Company, Incorporated.