FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA

ISAQUE VIEIRA MACHADO PIM

SELEÇÃO DE VARIÁVEIS POR BUSCA ESTOCÁSTICA PARA EPIDEMIOLOGIA ESPACIAL

Rio de Janeiro
2022

ISAQUE VIEIRA MACHADO PIM

SELEÇÃO DE VARIÁVEIS POR BUSCA ESTOCÁSTICA PARA EPIDEMIOLOGIA ESPACIAL

Trabalho de conclusão de curso apresentada para a Escola de Matemática Aplicada (FGV/EMAp) como requisito para o grau de bacharel em Matemática Aplicada.

Área de estudo: Estatística Bayesiana Espacial.

Orientador: Luiz Max Carvalho

Rio de Janeiro

2022

Ficha catalográfica elaborada pela BMHS/FGV

V. M. Pim, Isaque
Seleção de variáveis por busca estocástica para epidemiologia espacial/ Isaque Vieira Machado Pim. – 2022.
48f.

Trabalho de Conclusão de Curso – Escola de Matemática Aplicada.

Advisor: Luiz Max Carvalho.
Includes bibliography.

1. Bioestatística 2. Estatística Espacial 2. Seleção de variáveis I. Carvalho, Luiz Max II. Escola de Matemática Aplicada III. Seleção de variáveis por busca estocástica para epidemiologia espacial

ISAQUE VIEIRA MACHADO PIM

SELEÇÃO DE VARIÁVEIS POR BUSCA ESTOCÁSTICA PARA EPIDEMIOLOGIA ESPACIAL

Trabalho de conclusão de curso apresentada para a Escola de Matemática Aplicada (FGV/EMAp) como requisito para o grau de bacharel em Matemática Aplicada.

Área de estudo: Estatística Bayesiana Espacial.

E aprovado em 09/12/2022
Pela comissão organizadora

Luiz Max Carvalho
Escola de Matemática Aplicada

Dr. Helton Graziadei de Carvalho
FGV EMAp

Dr. Oswaldo Gonçalves Cruz
PROCC/Fiocruz

Dedico essa dissertação aos meus pais: Marcelo e Munira.

Agradecimentos

Agradeço meus pais, Marcelo e Munira, por razões intermináveis e que não caberiam no escopo desse agradecimento, agradeço por todo amor que nunca me faltou. E meu irmão, Miguel, companheiro e amigo à longa distância.

Agradeço meus companheiros de turma por toda jornada até aqui. Em especial, meus companheiros do CDMC, que fizeram parte ativa de minha vida no Rio de Janeiro. E meus colegas de tempos passados, que ou ficaram no Espírito Santo, ou também tomaram outros rumos, mas ainda mantém contato

Agradeço o Centro para o Desenvolvimento de Ciências e Matemática (CDMC) da FGV por acreditarem no meu potencial, e custearem todo processo de graduação.

Agradeço meus professores da graduação em Matemática Aplicada da FGV por toda disposição e contribuição com meu crescimento ao longo do curso.

Destes, agradecimentos especiais ao meu orientador, Luiz Max Carvalho, por não ter faltado em momento algum no desenvolvimento deste trabalho, por todo direcionamento e gama de referências, que certamente me fizeram um acadêmico melhor. E claro, agradeço sua eficiência invejável na comunicação por e-mail.

"Eu vi uma flor
pendida de um galho,
banhada de orvalho
das noites de abril.

Se eu fosse um pintor,
artista de fato,
faria um retrato
de flor tão gentil.

Estava no meio,
por entre a folhagem
da verde ramagem
um tanto escondida.

Talvez com receio
que mãos criminosas,
almas venenosas,
tirassem- lhe a vida."

Albércio Vieira Machado

Resumo

O mapeamento de doenças tem um longo histórico em vigilância sanitária. Mapas provêm um resumo visual rápido da informação espacial, e permitem encontrar padrões que não apareceriam na forma tabular. Com o aumento da disponibilidade de dados georreferenciados, faz-se necessário o desenvolvimento de técnicas para a análise deste tipo de dados. Do ponto de vista epidemiológico, a modelagem dos padrões e estruturas de correlação, estimação dos parâmetros relevantes para o problema e a comparação de diferentes cenários e a predição para regiões sem observações são essenciais para compreensão do cenário e melhor alocação de recursos para reduzir os impactos de um possível surto.

Neste trabalho, utilizo técnicas modernas, na forma de modelos hierárquicos Bayesianos, para estudar a distribuição do risco para uma doença e análise dos fatores relevantes para a propagação desta doença. O uso de modelos hierárquicos permite de forma robusta e flexível introduzir informações de covariáveis para o modelo, acomodar correlação espacial além de prover uma noção formal da incerteza associada às estimativas de risco. Em especial, consegue conciliar esquemas de seleção de variáveis junto com a estimação dos parâmetros de interesse do modelo.

No Capítulo 2 reviso as duas principais técnicas utilizadas ao longo do texto: modelos condicionais autorregressivos (CAR) e seleção bayesiana de variáveis por busca estocástica (BSSVS). Modelos CAR são responsáveis por acomodar estrutura espacial dentro do modelo, enquanto a seleção de variáveis é feita por um esquema de busca estocástica acoplada ao modelo hierárquico. No Capítulo 3, aplico os métodos desenvolvidos para o caso de câncer labial na Escócia, comparando diferentes modelos e produzindo mapas de risco e quantidades de interesse. No Capítulo 4, aplico os métodos desenvolvidos para a epidemia de Ebola na África Ocidental que ocorreu nos anos de 2013-2016. Os dados do Ebola foram enriquecidos com informações filogenéticas como variáveis explanatórias. Os resultados são utilizados para produzir mapas de risco e realizar a predição de áreas sem observações

Palavras-chave: mapeamento de doenças, modelos hierárquicos bayesianos, seleção de variáveis

Abstract

Disease mapping has a long history in health surveillance. Maps provide a quick summary of spatial information that otherwise would not pop up in a tabular format. With the increasing quantity of georeferenced data, it is necessary to develop methods and techniques to analyze such data. From an epidemiological point of view, modeling patterns and correlation structures, estimation of the relevant parameters, comparison of different scenarios and prediction of missing data are essential to understand the situation and better allocate resources, reducing damage from a possible disease outbreak.

In this text, I make use of modern methods in the form of Bayesian hierarchical models to study disease burden and analyze the risk factors that are relevant to the spreading of a disease. Hierarchical modeling allows for the introduction of covariate information and adjustment for spatial correlation in a robust and flexible manner. Moreover, it provides a formal notion of uncertainty associated with the risk estimates and allows variable selection schemes along with parameter estimation.

In Capítulo 2 I go over the two main methods used in the text: conditional autoregressive (CAR) models and Bayesian Stochastic Search Variable Selection (BSSVS). CAR models are responsible for accommodating spatial information contained in data, and variable selection is done by a stochastic search scheme attached as part of the hierarchical model. In Capítulo 3, I apply those methods to model the Scottish Lip Cancer case, making comparisons of different modeling options and producing risk maps and quantities of interest. In Capítulo 3, I apply those methods to model the 2013-2016 West African Ebola epidemic. The data was augmented with phylogenetic information as covariates. The results are then used to produce risk maps and prediction of missing case counts.

Keywords: disease mapping, hierarchical modelling, variable selection

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nos anos recentes, estudos de mapeamento de doenças se tornaram aplicação rotineira de epidemiologia geográfica, e tipicamente esses estudos são feitos dentro de uma formulação hierárquica Bayesiana (RIEBLER et al., 2016). Mapas provêm um resumo visual rápido da informação espacial, e permitem encontrar padrões que não apareceriam na forma tabular. A alocação de recursos para combate a surtos de doenças deve ter como guia básico mapas de zonas de risco para uma doença, e uma lista de possíveis fatores relevantes para a propagação da doença. Neste texto, reviso e aplico métodos comumente sugeridos para a análise de dados de casos de doença agregados por alguma unidade de área no contexto de mapeamento de doenças e regressão espacial, assim como métodos para seleção de variáveis dentro do esquema de regressão.

Devemos ter cuidado com regressões espaciais por dois motivos: Primeiro, sempre que tratamos de dados agregados podemos cair na possibilidade de falácia ecológica, e isso só pode ser amenizado com a inclusão de dados a nível de indivíduo. Segundo, quando existe dependência espacial no resíduo, termo que defino no corpo do texto, e quando existe estrutura espacial na variável resposta, então as estimativas de parâmetros vão mudar quando comparadas a um cenário de independência, e os dados por si só não podem acomodar todo tipo de forma e extensão de correlação espacial (WAKEFIELD, 2007). Para isso, o uso de modelos hierárquicos permite de forma robusta e flexível introduzir informações de covariáveis para o modelo, acomodar correlação espacial além de prover uma noção formal da incerteza associada às estimativas de risco. Em especial, consegue conciliar esquemas de seleção de variáveis junto com a estimação dos parâmetros de interesse do modelo.

A divisão do texto é a que segue: no Capítulo 2 reviso as duas principais técnicas utilizadas ao longo do texto: modelos condicionais autorregressivos (CAR) e seleção bayesiana de variáveis por busca estocástica (BSSVS). Modelos CAR são responsáveis por acomodar estrutura espacial dentro do modelo, enquanto a seleção de variáveis é feita por um esquema de busca estocástica acoplada ao modelo hierárquico. No Capítulo 3, aplico os métodos desenvolvidos para o caso de câncer labial na Escócia, comparando diferentes modelos e produzindo mapas de risco e quantidades de interesse. No Capítulo 4, aplico os métodos desenvolvidos para a epidemia de Ebola na África Ocidental que ocorreu nos anos de 2013-2016. Os dados do Ebola foram enriquecidos com informações filogenéticas como variáveis explanatórias. Os resultados são utilizados para produzir mapas de risco e realizar a predição de áreas sem observações.

2 Modelagem

Neste capítulo, detalho os modelos estatísticos empregados neste trabalho, incluindo os módulos espacial e de seleção de variáveis, respectivamente.

2.1 Modelos Espaciais

> *"Geography isn't just dynamic, it's a narrative — it shows you a place and tells a story about that place, what's happening there now, and what will happen next."*
> ***Jack Dangermond***

A estatística espacial é o ramo da estatística que lida com dados geográficos. Dados geográficos são definidos pela ISO/TC 211 (ISO/TC. . . , 1994) como dados contendo de forma implícita ou explícita sua localização relativa à Terra. Para fins de modelagem (CÂMARA et al., 2004), classifico conjuntos de dados espaciais em três tipos básicos:

- *Eventos ou Padrões Pontuais* - eventos pontuais aleatórios ocorrendo dentro de um domínio. Exemplos são: localização de crimes, últimos locais onde uma espécie animal rara foi avistada, dados de *IOT (Internet-Of-Things)*.
- *Superfícies Contínuas* - também conhecido como dados Geoestatísticos, são referenciados por pontos fixos a partir de amostras coletadas em campo, podendo estar regularmente ou irregularmente distribuídas. Usualmente, este tipo de dado deriva do levantamento de recursos naturais, e que incluem mapas geológicos, topográficos, etc.
- *Áreas com Contagens e Taxas Agregadas* - ou dados em área, ou dados em reticulado. Os dados são agregados em unidades de análise, usualmente delimitadas por polígonos que formam uma partição finita de um determinado domínio. Unidades de análise podem ser, por exemplo, estados, municípios, setores censitários. Exemplos de dados em área são população por município, contagem de casos de uma doença por país, votos por setor censitário.

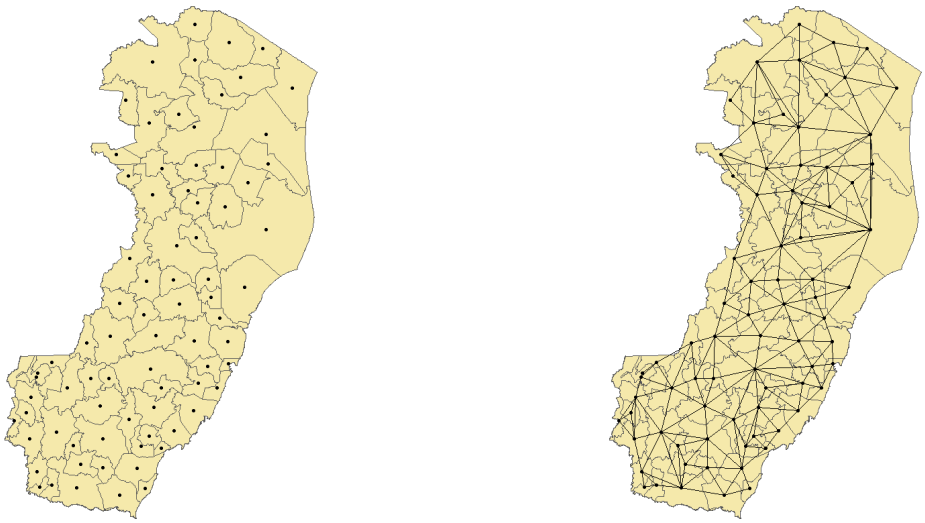
Neste trabalho, daremos enfoque para os dados em área. A análise espacial tem como mantra e principal motivador a *Primeira Lei da Geografia* de (TOBLER, 1970):

"Everything is related to everything else, but near things are more related than distant things"

Este é o princípio básico de conceitos como a dependência espacial e a autocorrelação espacial. Tudo está relacionado, porém as coisas mais próximas estão mais relacionadas que as distantes. Para fazer sentido analítico da Primeira Lei da Geografia, precisamos definir o que queremos dizer com proximidade, e o que queremos dizer com estar relacionado.

Central para a análise espacial, o conceito de proximidade é poderoso e flexível. A forma mais comum de representar a noção de proximidade é com relações de distância e conectividade em um plano Euclidiano vazio. Mas não precisa ser o caso de o espaço ser plano, nem Euclidiano e nem vazio. O interesse de geógrafos está nos geoespaços (MILLER, 2004), espaços que podem representar fenômenos na superfície da terra com noções de caminho de menor custo entre pares de objetos bem definidas. (MILLER, 2004) (MILLER; WENTZ, 2003) descrevem diferentes formas de definir proximidade em diferentes espaços.

Para dados em área, a noção de proximidade comumente usada é a de conectividade, em especial, determinar que duas regiões são vizinhas se compartilham uma fronteira. Para formalizar esta definição considere um conjunto de regiões $B = \{1, \ldots, n\}$. Denoto a relação de vizinhança por $i \sim j$, uma relação simétrica mas não reflexiva, pois uma área não é vizinha de si mesma. O conjunto de vizinhos de uma região é denotado por δi, e o número de vizinhos por n_{δ_i}. A relação de vizinhança também define um grafo em suas regiões (por isso o dado também é referido como reticulado), onde o conjunto de vértices é o próprio B, e a aresta $\{i, j\}$ pertence ao grafo se, e só se, $i \sim j$. A Figura 1 ilustra um mapa e suas relações de vizinhança por meio de um grafo.



(a) Mapa apenas com os municípios e seus respectivos centroides.

(b) Mapa com as relações de vizinhança representadas por um grafo conexo.

Figura 1 – Mapa dos municípios do Espírito Santo.

A matriz de adjacências deste grafo é um exemplo de matriz de proximidades. Uma matriz de proximidades W possui entradas w_{ij} que de alguma forma representam a conexão espacial entre duas regiões i e j (é de costume que $w_{ij} = 0$ se $i = j$). Para a matriz de vizinhança, temos $w_{ij} = 1$ se $i \sim j$, e $w_{ij} = 0$ caso contrário. Há aqui diversas

possibilidades para as escolhas de w_{ij}, que não necessariamente representam conectividade, como por exemplo w_{ij} ser a distância entre os centroides de i e j. Podemos tomar $w_{ij} = 1$ se i e j se a distância mínima entre suas fronteias está dentro de um limite pré-estabelecido. Poderíamos padronizar as linhas de W para torná-la uma matriz estocástica. O último exemplo acabaria com uma propriedade excelente para uma matriz de proximidades que é ser simétrica. Matrizes de proximidade terão um papel importante para definir a estrutura espacial de nossos modelos.

E por relacionados para duas entidades geográficas, o que queremos dizer? No mínimo, esperamos alguma correlação, positiva ou negativa. Mais do que isso, estarem espacialmente correlacionadas. Para isso precisamos de técnicas quantitativas para analisar a correlação entre duas variáveis relativa à distância entre as duas ou a conectividade. Falhar ao levar em conta efeitos espaciais, ou até mesmo ignorá-los pode acarretar em sérios erros na interpretação de modelos como notam (ANSELIN; GRIFFITH, 1988) e (ARBIA, 1989). Duas estatísticas padrão para medir a associação espacial entre dados em área são o I de Moran (MORAN, 1950) e o C de Geary (GEARY, 1954). São um único valor tentando sumarizar toda associação espacial presente nas áreas. Para áreas $B = \{1, \ldots, n\}$ com observações Y_i para a área i, e w_{ij} as entradas da matriz de vizinhança, o I de Moran é dado por

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_{i=1}^{n} (Y_i - \bar{Y})^2}, \tag{2.1}$$

e o C de Geary por

$$C = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Y_i - Y_j)^2}{2(\sum_{i \neq j} w_{ij}) \sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{2.2}$$

Estas estatísticas podem ser utilizadas para realizar testes estatísticos. O I de Moran, sob a hipótese nula que os dados são variáveis i.i.d, é assintoticamente normal com média $-1/(n-1)$. O C de Geary caso assuma valores baixos entre $(0, 1)$ indica a presença de associação espacial. Assim como o I é assintoticamente normal caso as variáveis sejam i.i.d mas o uso dessas métricas é sugerido apenas como medida exploratória (LI; CALDER; CRESSIE, 2007).

Uma forma mais moderna de analisar a dependência espacial de um conjunto de dados é entender que cada unidade de área possui uma quantidade intrínseca de dependência espacial devido à sua situação relativa ao resto do sistema – ao contrário do I de Moran e do C de Geary, que fornecem uma única medida crua para o sistema inteiro. A exemplo, temos as estatísticas G_i e G_i^* de (GETIS; ORD, 1992), sobre as quais não entrarei em detalhes.

Para exemplificar o uso do I de Moran como ferramenta de detecção de dependên-

Figura 2 – Resultados das eleições de 2010 no Brasil. Em vermelho, estados onde o PT conquistou maioria dos votos. Em azul, maioria do PSDB.

cia espacial, veja na Figura 2 o mapa do segundo turno das eleições de 2010 no Brasil. Calculando o I de Moran para a porcentagem de votos úteis do PT (Partido dos Trabalhadores) e métrica de distância a matriz de vizinhanças, obtemos o valor de 0.1456 contra o esperado sobre hipótese nula de $-0,0385$, indicando associação espacial entre os estados que elegeram o PT.

(CLIFF, 1981) detalha o cálculo de $\mathbb{E}[I]$ e $Var(I)$ sob hipótese de normalidade assintótica e de aleatorização. Aleatorização acontece pois sob hipótese nula não há associação espacial, então permutar as variáveis não modifica a distribuição do I de Moran. É possível então calcular o *Z-score* e realizar um teste de hipótese. O teste é implementado em R no pacote **spdep** como a função *moran.test*.

2.1.1 Modelando contagem de casos de uma doença

Para o restante do capítulo, considere a notação $B = \{1, \ldots, n\}$ um conjunto de regiões disjuntas, δ_i o conjunto de vizinhos da região i e n_{δ_i} o tamanho de δ_i e w_{ij} como as entradas da matriz de vizinhança W. Dados em área são objetos de forte interesse da Bioestatística e da Epidemiologia. Tipicamente nos deparamos com dados de contagem de casos por região da seguinte forma:

$$Y_i \quad = \quad \text{casos observados na região } i, \quad i = 1, \ldots, n$$

$$E_i \quad = \quad \text{valor esperado de casos na região } i, \quad i = 1, \ldots, n$$

Y_i variáveis aleatórias e E_i função do número de pessoas em risco na região i.

Caso a doença não seja rara (BANERJEE; CARLIN; GELFAND, 2015), o modelo mais usual é o de Poisson:

$$Y_i|\theta_i \sim \text{Poisson}(E_i\theta_i), \tag{2.3}$$

onde θ_i é o risco relativo da região i. Para efeticamente mapear uma superfície de risco, podemos pensar em um modelo de efeito aleatório para os θ_i. Além de poder assumir que os riscos de cada região vêm de uma mesma distribuição, podemos agregar informação de todas as regiões nas estimativas de risco para uma única região. Também podemos introduzir informação de covariáveis modelando este risco relativo

Considere, então, o seguinte aprimoramento do modelo simples de Poisson, que introduz uma distribuição comum para os θ_i.

$$\begin{aligned} Y_i|\theta_i &\sim \text{Poisson}(E_i\theta_i), \\ \log(\theta_i) &= \sum_{j=1}^{P} X_{ij}\beta_j + \varepsilon_i, \\ \boldsymbol{\varepsilon} &\sim \text{Normal}(0, \Sigma). \end{aligned} \tag{2.4}$$

A (log) taxa relativa é modelada como uma componente linear, que introduz informação de covariáveis para o modelo, acrescida de ruído. Aqui $\mathbf{X}$ é uma matriz $N \times P$ com informação de P covariáveis para cada uma das N. O ruído modela a sobredispersão, fenômeno muito comum em dados ecológicos e biologia evolucionária (HARRISON, 2014). Esta sobredispersão pode ocorrer devido a covariáveis que não foram incluídas no modelo, excesso de zeros e dados não independentes. Esse rúido é chamado de Efeito Aleatório a Nível Observacional, ou ORLE (*Observational Level Random Effects*), e ameniza os problemas da sobredispersão descritos.

Devido à típica incerteza vinda da amostragem, não é recomendado que se analise as taxas diretamente, mas sim tomar "emprestado" informação de regiões vizinhas. Espera-se que regiões próximas apresentem mais similaridades que regiões distantes, e é interessante explorar essa informação para extrair estimativas de risco fidedignas. O efeito é similar a introduzir covariáveis que apresentem tal comportamento: assumem valores similares em locais próximos. Surge então uma dificuldade para modelar esta dependência espacial, que é conseguir compensar estas covariáveis não observadas e tentar simular seus comportamentos.

Uma das formas mais comuns de introduzir correlação espacial para um modelo como a Eq. (2.4) é com um *Campo de Gaussiano de Markov Intrínseco* (IGMRF), que aqui refiro como modelo CAR (ou termo CAR). Especificamos a distribuição de um modelo CAR $\boldsymbol{\nu}$ através de suas condicionais completas

$$\nu_i|\nu_{-i} \sim \text{Normal}\left(\frac{1}{n_{\delta_i}}\sum_{j\in\delta_i}\nu_j, \frac{1}{n_{\delta_i}\kappa_i}\right),$$

onde $\kappa_i \in (0, +\infty)$ é um parâmetro de precisão. A partir do Lema de Brook (Lema 1), e assumindo $\kappa_i = \kappa$ comum parar todas as regiões, podemos derivar a distribuição conjunta de ν:

$$\pi(\boldsymbol{\nu}|\kappa) \sim \text{Normal}\left(-\frac{1}{2\kappa^2}\sum_{i\sim j}(\nu_i - \nu_j)^2\right) = \text{Normal}\left(\frac{1}{2\kappa^2}\nu^T(D_W - W)\nu\right), \tag{2.5}$$

onde D_W é uma matriz diagonal com entradas n_{δ_i}. O modelo é intrínseco pois $(D_W - W)1 = 0$, portanto a matriz de precisão é singular e a deficiência de posto é 1 caso o grafo seja conexo, vide Apêndice A. Note também que a densidade é invariante à adição de uma constante, portanto para que não haja confusão com o intercepto, impomos a restrição $\sum_i Y_i = 0$. Por ser impróprio, vide Apêndice A, os modelos CAR intríscos (ICAR) não podem ser usados para modelar dados, e portanto são designadas como prioris. (BESAG; YORK; MOLLIÉ, 1991) mostra que apesar do modelo ser impróprio, a posteriori é própria.

Uma forma de lidar com a condição imprópria do modelo é adicionar um parâmetro ρ escalando a componente W de $(D_W - W)$, controlando então a contribuição espacial deste termo (W é a matriz de vizinhanças). Veja que com $\rho = 0$ temos uma matriz diagonal e, portanto, entradas independentes. Com $\rho = 1$ temos o modelo ICAR.

$$\pi(\boldsymbol{\nu}|\kappa) \sim \text{Normal}\left(\frac{1}{2\kappa^2}\nu^T(D_W - \rho W)\nu\right), \tag{2.6}$$

Note que agora tomando $\rho \in (\frac{1}{\lambda_{min}}, \frac{1}{\lambda_{max}})$ obtemos um modelo próprio, onde $\lambda_{min}, \lambda_{max}$ são o menor autovalor e o maior autovalor de $(D_W - \rho W)$, respectivamente Note que $\rho = 0$ implica em variáveis ν_i independentes, assim como $\rho = 1$ implica no modelo ICAR. Apesar de tornar o modelo próprio, permitindo práticas como simulação *a priori* do modelo, o modelo próprio, como alertam (BANERJEE; CARLIN; GELFAND, 2015), não traz a maleabilidade desejada para introduzir padrões espaciais significativos.

Besag, York e Mollié (1991) introduziram de forma pioneira regressões de Poisson com erros tanto não-estruturados como com erros estruturados de natureza espacial. O modelo é conhecido como modelo de BYM (Besag-York-Mollié). Sua formulação é a que segue:

$$\begin{aligned} Y_i|\theta_i &\sim \text{Poisson}(E_i\theta_i), \\ \log(\theta_i) &= \sum_{j=1}^{P} X_{ij}\beta_j + \xi_i, \end{aligned} \tag{2.7}$$

$$\xi_i = \varepsilon_i + \nu_i,$$

$$\varepsilon_i \sim \text{Normal}(0, \tau_u),$$

$$\nu_i | \nu_{-i} \sim \text{Normal}\left(\frac{1}{n_{\delta_i}} \sum_{j \in \delta_i} \nu_j, \frac{1}{n_{\delta_i} \tau_s}\right).$$

τ_s e τ_u parâmetros de precisão. Note que estes parâmetros de precisão não podem ser vistos de forma independente. Temos uma observação por área para tentar estimar dois parâmetros!

(LEROUX; LEI; BRESLOW, 2000) e (DEAN; UGARTE; MILITINO, 2001) trouxeram reparametrizações do modelo BYM para lidar com estes problemas. Considere a notação $Q = (D_W - W)$ O modelo de Leroux propõe um compromisso entre a componente estruturada e a componente não estruturada por um parâmetro de mistura $\phi \in [0, 1]$. O modelo é o BYM mas a componente $\boldsymbol{\xi}$ segue uma distribuição normal com média zero e matriz de covariância

$$\text{Var}(\boldsymbol{\xi} | \tau_b, \phi) = \tau_b^{-1} \left((1 - \phi) I + \phi Q\right)^{-1}. \tag{2.8}$$

A variância condicional do modelo de Leroux é dado pela média ponderada por ϕ de $1/\tau_b$ e $1/(\tau_b n_{\delta_i})$. Como consequência, a decomposição aditiva da variância acontece na log escala do risco relativo condicional à vizinhança da região analisada (LEROUX; LEI; BRESLOW, 2000). O modelo de Dean propõe uma reparametrização do termo ξ do modelo BYM como:

$$\boldsymbol{\xi} = \frac{1}{\sqrt{\tau_b}} \left(\sqrt{1 - \phi} \cdot \boldsymbol{\varepsilon} + \sqrt{\phi} \cdot \boldsymbol{\nu}\right), \tag{2.9}$$

tendo matriz de covariância

$$Var(\boldsymbol{\xi} | \tau_b, \phi) = \tau_b^{-1} \left((1 - \phi) I + \phi Q^{-}\right) \tag{2.10}$$

Onde Q^{-} denota a inversa generalizada. Esta reparametrização é o modelo BYM com $\tau_s^{-1} = \tau_b^{-1} \phi$ e $\tau_u^{-1} = \tau_b^{-1}(1 - \phi)$. A decomposição aditiva da variância está na log escala do risco relativo.

Apesar dos modelos de Leroux e Dean lidarem com a identificação de como a sobredispersão se distribui entre componente estruturada e não estruturada, ambos possuem o problema da componente espacial não estar escalada. Como nota (SØRBYE; RUE, 2014), escalar a componente espacial é essencial para a escolha das prioris, e garantir que as escolhas de priori de uma aplicação possam ser utilizadas em outras aplicações.

(RIEBLER et al., 2016) detalha o desenvolvimento de um novo modelo BYM, que se atenta à escala. Por simplicidade, considere ξ composto apenas pela componente

espacial. Geralmente, as variâncias marginais $\tau_b[Q^-]_{ii}$ dependem da estrutura do grafo analisado. Isso pode ser ilustrado calculando uma variância generalizada, como a média geométrica das variâncias marginais

$$\sigma^2_{GV} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{\tau_b}[Q^-]_{ii}\right)\right). \tag{2.11}$$

Para unificar a interpretação de τ_b da componente estruturada e da não estruturada, e torná-la transferível entre aplicações, o efeito espacial precisa ser escalado para que $\sigma^2_{GV} = 1/\tau_b$. Isso faz com que o parâmetro τ_b represente o desvio de um nível constante para qualquer que seja o grafo por trás do modelo. (RUE; HELD, 2005) apresenta um esquema para de forma eficiente calcular as entradas da diagonal de Q^-. Após extraídos estes componentes, σ^2_{GV} é calculado como na Equação 2.11 e é usado para escalar a matriz Q. A partir disso, (SIMPSON et al., 2017) propõe uma nova parametrização do modelo BYM, que recebe o nome de BYM2. Essa nova parametrização utiliza a componente espacial escalada com precisão 1. O efeito aleatório agora é formulado como

$$\boldsymbol{\xi} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\phi}\cdot\boldsymbol{\varepsilon}\sqrt{\phi}\cdot\boldsymbol{\nu}^*\right) \tag{2.12}$$

A nova variância é dada por:

$$\text{Var}(\boldsymbol{\xi}|\tau_b,\phi) = \tau_b^{-1}((1-\phi)I + \phi Q_*^-) \tag{2.13}$$

Veja que agora os hiper-parâmetros τ_b e ϕ tem interpretação clara e não são mais confundidos.

2.1.1.1 Escolha das prioris

Uma questão importante em modelagem bayesiana em geral, e em modelagem espacial em particular é a escolha das distribuições *a priori*. Modelos GMRF em particular podem ser particularmente sensíveis à especificação de prioris. Wakefield (2007) trás uma ótima revisão para o caso de regressões espaciais. Par um link log-linear, como é o caso dos modelos estudados nesta seção, é conveniente especificar prioris lognormais para os parâmetros positivos $\exp(\beta_j)$, e fica bem idreto especificar dois quantis e encontrar os parâmetros associados da distribuição lognormal. Denote por $LN(\mu,\sigma)$ a distribuição da lognormal para um parâmetro genérico θ, com $\mathbb{E}[\log(\theta)] = \mu$ e $Var(\log(\theta)) = \sigma^2$. Por exemplo, supondo que o risco relativo de uma variável β_j tem 50% de chance de ser menor que 1, podemos assumir *a priori* $\mu = 0$. E uma alta chance, de por exemplo 95%, de ser menor que 5. Daí $\sigma = \log 5/1.645 = 0.98 \approx 1$. Uma priori $e^{\beta_j} \sim LN(0,1)$ pode ser uma boa escolha. Caso existam muitas covariáveis, ou existe alta correlação entre as covariáveis,

vale a pena tornar a priori mais informativa. Como exemplo, veja (LI; ZHANG et al., 2015) que usa prioris de Ising para encorajar a clusterização de variáveis correlacionadas.

Prioris para os parâmetros de precisão do risco relativo residual já não são tão diretas de se derivar. A escolha da distribuição $Gama(a, b)$ para a precisão tanto da componente estruturada, como da não estruturada, ou do risco relativo residual escaldo como no BYM2, é conveniente pois produz uma distribuição marginal em forma fechada. Em especial, sabemos que o modelo de dois níveis

$$\xi_i|\tau \sim \text{Normal}(0, 1/\tau), \quad \tau \sim \text{Gama}(a, b)$$

produz distribuição marginal para ξ_i como uma t-student generalizada $t_{2a}(0, b/a)$ com $2a$ graus de liberdade, locação zero e escala b/a. Uma forma para determinar a e b é especificar um alcance $\exp(\pm R)$ para que o risco residual fique dentro com probabilidade q, e usamos a simetria da t para obter que $\pm t^{2a}_{q/2}\sqrt{b/a} = \pm R$, Onde $\pm t^{2a}_{q/2}$ é o quantil q da t generalizada. Por exemplo, suponha que o risco relativo fique dentro do intervalo (0.5, 2.0) com probabilidade de 95% , então obtemos que $\tau \sim \text{Gama}(1, 0.026)$.

É importante garantir que a priori acesse todos os possíveis níveis de variabilidade no resíduo. Por esta razão Wakefield (2007) não recomenda o uso de prioris vagas como $\text{Gama}(0.001, 0.001)$.

2.1.1.2 Simulações

Os termos CAR conseguem compensar por covariáveis espacialmente associadas e controlar pela correlação espacial do ruído. Para ilustrar o fato, considere o seguinte exemplo artificial. Gerei uma covariável x^* com clara dependência espacial para o mapa do Espírito Santo como mostra a Figura 3. Em adição foram geradas duas covariáveis x_1, x_2 de uma normal padrão para cada região. A quantidade de interesse Y é então obtida pondo $Y_i \sim \text{Poisson}(\theta_i)$, e $\log(\theta) = 1.5 + 0.5x_1 - 0.5x_2 + 0.3x^*$. Vamos amostrar de dois cenários diferentes, desconsiderando a variável x^*: o primeiro um modelo Poisson log-normal simples, e o segundo o mesmo modelo com um termo CAR. Para medir a associação espacial no resíduo utilizo o teste de Moran. Para o modelo tradicional o I de Moran no resíduo vale 0.194, contra o esperado sobre hipótese nula de -0.013 e um p-valor de 0.0025. Para o modelo com termos CAR o I de Moran no resíduo vale -0.04 contra um esperado de -0.13 e um p-valor de 0.70.

Um segundo experimento é verificar se as reparametrizações de Dean, Leroux e o BYM2 reduzem aos modelos básicos, i.e, num cenário de alta correlação espacial o parâmetro de mistura ϕ se aproximar de 1, no caso de efeitos regionais aleatórios ϕ se aproximar de zero. Para isso, simulo do modelo de Poisson com $\log(\theta_i) = 0.4 + 0.5\xi_i$. Suponho três cenários: ξ_i vindo de uma normal padrão, ξ tendo distribuição $N(0, Q_*^-)$, onde Q_* é a matriz Q escalada para o grafo de adjacências do Espírito Santo, e um terceiro

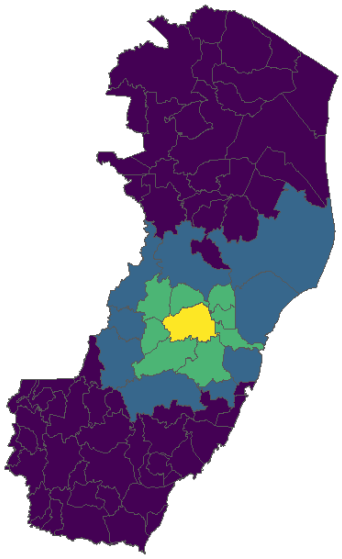


Figura 3 – Covariável espacialmente correlacionada gerada artificialmente para o mapa do Espírito Santo.

	β_0		ρ		$\frac{1}{\sqrt{\tau}}$	
Cenário 1	0.32	(0.10)	0.51	(0.29)	0.28	(0.10)
Cenário 2	0.45	(0.10)	0.36	(0.29)	0.40	(0.16)
Cenário 3	0.43	(0.10)	0.68	(0.14)	0.54	(0.14)

Tabela 1 – Sumários à posteriori para os 3 cenários descritos: superfície de risco constante, superfície de risco com efeito aleatório regional e superfície de risco com efeito aleatório espacial. Para cada parâmetro temos a mediana à posteriori junto com o desvio-padrão.

que é uma superfície de risco constante. O modelo avaliado é o BYM2. A priori para o parâmetro ϕ é a da indiferença $U(0,1)$. Os resultados estão apresentados na Tabela 1. Veja que para o modelo de superfície constante o BYM2 a posteriori de ρ não se distancia muito da priori. Já para os cenários de 2 e 3 o parâmetro ρ já apresenta um desvio da priori imposta. Não houve uma redução completa para $\phi = 0$ ou $\phi = 1$. (SIMPSON et al., 2017) propõe o uso de prioris penalizadoras de complexidade (prioris PC) para o BYM2 que conseguem reduzir o parâmetro de mistura para zero. Prioris PC são baseadas na Navalha de Ockham - modelos mais simples devem ser preferidos até que se tenha evidência o suficiente para um modelo mais complexo.

2.2 Seleção de Variáveis

Ao lidar com um conjunto de preditores num termo linear, junto com a estimação dos coeficientes é comum também determinar quais destes preditores estão mais associados com a variável a ser predita. Para o caso de P preditores, queremos determinar um "bom"

submodelo dentre os 2^P submodelos possíveis, baseado em algum critério apropriado . Para modelos Bayesianos, uma métrica comum para a comparação de modelos é o WAIC (Widely Applicable Information Criterion) Watanabe (2010). Usando a mesma notação da seção 2.1 para modelos de contagem de casos, considere uma variável de exposição Y para N regiões e um conjunto de P preditores $X = (X_1, \ldots, X_p)$ e θ todos os parâmetros de interesse. O WAIC do modelo é calculado como

$$\begin{aligned} \text{WAIC} &= \mathbb{E}[D(\theta)] + 2\sum_{i=1}^{N} \log(\mathbb{E}[p(Y_i|\theta)]) - \mathbb{E}[\log p(Y_i|\theta)], \\ \mathbb{E}[D(\theta)] &= -2\log(p(Y|\theta)) \end{aligned} \tag{2.14}$$

.

Tem a vantagem de poder ser calculado sem assumir nenhuma distribuição "verdadeira", e funciona como um *cross-validation* Bayesiano medindo o poder preditivo do modelo fora da amostra. No entanto, para o trabalho de seleção de variáveis, onde o espaço de modelos cresce exponencialmente com o número de covariáveis, é inviável executar essa computação para todos os modelos.

Do ponto de vista Bayesiano, este problema é contornado ao considerar a seleção de variáveis como uma forma de estimação de parâmetros. A ideia básica é definir as variáveis latentes $\boldsymbol{\gamma} = \{\gamma_i, 1 \leq i \leq P\}$ onde γ_i é uma variável indicadora da inclusão do i-ésimo preditor X_i no modelo. Em muitos casos, a decisão de incluir ou não uma variável recai na estimação da probabilidade marginal à posteriori de incluir uma variável no modelo (probabilidade marginal à posteriori dos parâmetros γ_i). (O'HARA; SILLANPÄÄ, 2009) traz uma revisão de diversas formas de incluir esta seleção de variáveis para o problema de regressão Bayesiana. Mantendo a notação da seção 2.1, considere agora a seguinte versão aumentada do termo linear da Equação 2.4

$$\begin{aligned} \log(\theta_i) &= \beta_0 + \sum_{j=1}^{P} X_{ij}\nu_j + \varepsilon_i, \\ \nu_j|\gamma_j, \beta_j &\sim G. \end{aligned} \tag{2.15}$$

A tarefa de seleção de variáveis então se simplifica a decidir quais dos parâmetros ν_j valem zero. A distribuição condicional G seguem, então, a forma de uma mistura "spike-and-slab" (pico e platô), possuindo um pico de densidade em zero e um platô de densidade nos outros pontos. Aqui considero dois casos para a densidade G. A primeira, de (KUO; MALLICK, 1998) e que me refiro por seleção de variáveis por indicadoras, de forma simples e direta atribui $\nu_j = \gamma_j\beta_j$, e prioris independentes para $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$. A segunda, inicialmente proposta por (GEORGE; MCCULLOCH, 1993) e que me refiro como SSVS (Stochastic Search Variable Selection), atribui $\nu_j = \beta_j$ e altera a distribuição de β_j baseado nas indicadoras da seguinte forma

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\sigma^2) + \gamma_j N(0, c_j^2\sigma^2). \tag{2.16}$$

Aqui o pico é uma distribuição centrada em zero com baixa variância, e o parâmetro de σ^2 é calibrado para que isso aconteça. c_j é calibrado para garantir o platô da distribuição.

Para todos os casos a escolha da priori para γ é parte essencial do problema. Essa escolha deve incorporar qualquer informação prévia sobre os modelos mais plausíveis. No entanto, com 2^P modelos isto se torna uma tarefa complexa. Para γ_i's independentes com distribuição marginal Bernoulli com $P(\gamma_i = 1) = p_i$ a densidade conjunta de γ vale:

$$f(\boldsymbol{\gamma}) = \prod_{i=1}^{P} p_i^{\gamma_i}(1-p_i)^{1-\gamma_i}. \tag{2.17}$$

Algumas hipóteses de simetria podem facilitar nossa atribuição da priori. A priori uniforme, ou priori da indiferença, é dada por:

$$f(\boldsymbol{\gamma}) = 2^{-p}. \tag{2.18}$$

Com $p_i = \frac{1}{2}$ para todo $i \in \{1, \ldots, P\}$. Podemos também penalizar o tamanho do modelo $M_\gamma = \sum_{i=1}^{p} \gamma_i$ pondo mais densidade a priori em modelos menores, favorecendo a parcimônia. Umas das sugestões de (GEORGE; MCCULLOCH, 1993) é colocar

$$f(\boldsymbol{\gamma}) = w_{M_\gamma} \binom{P}{M_\gamma}^{-1}, \tag{2.19}$$

onde w_{M_γ} é a probabilidade a priori de obter um modelo de tamanho M_γ. Para controlar o tamanho do modelo, (CARVALHO, 2019) sugere a parametrização de p_i pela maleabilidade do modelo. Um modelo nada maleável não inclui nenhuma covariável. Seja $w = P(M_\gamma = 0)$ nosso parâmetro de maleabilidade. Considerando que $p_i = q, i \in \{1, \ldots, P\}$, é fácil ver que $w = (1-q)^P$, e então $q = 1 - w^{1/P}$.

Uma vantagem desta formulação é poder rapidamente avaliar a importância de uma variável por meio do Fator de Bayes de γ. Podemos escrever o Fator de Bayes para a i-ésima covariável como a razão da razão de chances à posteriori com as chances à priori:

$$BF_i = \frac{\hat{\gamma}_i}{1-\hat{\gamma}_i} / \frac{p_i}{1-p_i}, \tag{2.20}$$

onde $\hat{\gamma_i}$) é um estimador da probabilidade à posteriori γ_i. Para o caso da SSVS, precisamos também definir prioris para σ. σ deve ser tal que $\beta_j \sim N(0, \sigma^2)$ pode ser seguramente substituído por zero. Note que o parâmetro c_i representa a razão das alturas de $N(0, \sigma^2)$ e $N(0, c_i^2\sigma^2)$ em 0. Portanto, pode ser interpretado como a razão de chances de

β_i ser excluído quando assume valores próximos de 0. (GEORGE; MCCULLOCH, 1993) propõe uma alternativa semi-automática, utilizando informação do estimador de mínimos quadrados para σ. Prefiro não utilizar tal abordagem, adotando prioris mais generalistas para não comprometer a característica Bayesiana completa dos modelos.

Por fim, precisamos detalhar a escolha de priori para $\beta|\gamma$. Priori para o termo ORLE segue as mesmas recomendações da seção 2.1. Para isso, uso uma normal multivariada

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(0, D_\gamma R D_\gamma), \tag{2.21}$$

onde R é a matriz de correlação a priori e

$$D_\gamma = \text{diag}[(1-\gamma_1)\tau_1 + \gamma_1 c_1 \tau_1, \ldots, (1-\gamma_p)\tau_p + \gamma_p c_p \tau_p]. \tag{2.22}$$

Para a escolha de R é interessante observar o efeito dela na matriz de covariância à posteriori de β, a saber

$$(\sigma^{-2} X^T X + D_\gamma^{-1} R^{-1} D_\gamma^{-1})^{-1}. \tag{2.23}$$

Casos notáveis são $R = I$ e $R = (X^T X)^{-1}$. Para o primeiro, os betas ão independentes à priori e as correlações à posteriori serão menores que as correlações da matriz de design. Para o último, correlações à priori e à posteriori serão as mesmas da matriz de design. (GEORGE; MCCULLOCH, 1993) observa considerável diferença entre ambas as escolhas. Para a escolha $R = I$, modelos menores foram escolhidos com uma maior frequência.

2.2.1 Simulações

Para avaliar as técnicas de seleção de variáveis vou conduzir algumas simulações artificiais aos moldes de van Erp, Oberski e Mulder (2019), testando ambos os modelos de Kao e Mallick e o de McCulloch, que me refiro como modelos KM e MC, respectivamente. As simulações são feitas do modelo $y = X^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon \sim N(0, \sigma^2 I)$ em R com o pacote NIMBLE. Para garantir convergência foi analisado o $\hat{R}$ de Gelman-Rubin, avaliando se o valor para cada parâmetro é menor que 1.05 na última amostra. Todas as cadeias foram executadas com 10.000 iterações, sendo 2.00 usadas como *burn-in*. Como priori para σ^2 é usado para ambos os modelos uma gama inversa de parâmetros (5,5). Os parâmtros c e σ para o modelo MC são fixados como $c = 500$ e $\sigma = 0.01$. Considero que o modelo incluiu uma variável se sua média à posteriori é maior que 0.5

- Cenário 1 - $\beta = (3, 0, 0, -1, 0, 2, 0)$, $\sigma^2 = 9$, X vindo de uma normal multivariada com vetor de média zero, variâncias 1 e correlação entre as variáveis igual a 0.5. São observadas 200 amostras.

- Cenário 2 - $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$. As outras configurações iguais ao cenário 1.

- Cenário 3 - $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{15})$. $\sigma^2 = 225$. $x_j = Z_1 + w_j, \quad j = 1, \ldots, 5$, $x_j = Z_2 + w_j, \quad j = 6, \ldots, 10$, $x_j = Z_3 + w_j, \quad j = 11, \ldots, 15$, $x_j \sim N(0, 1), \quad j = 16, \ldots 30$. $Z_1, Z_2, Z_3 \sim N(0, 1)$ e $w_j \sim N(0, 0.01)$. São observadas 200 amostras.

- Cenário 4 - Cenário 3 com 400 amostras.

- Cenário 5 - Cenário 3 com $\beta = (\underbrace{3, \ldots, 3}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{3, \ldots, 3}_{10})$ e 40 observações.

Para o cenário 1 ambos os modelos geram bons estimadores para os parâmetros da regressão. Já se nota algumas diferenças. O modelo KM ainda tem os betas definidos como zero sendo escolhidos, enquanto o modelo MC rapidamente entende que esses betas não devem ser selecionados. Algo que era esperado, o ESS do modelo MC é menor do que o modelo KM. Para o cenário 2, o modelo KM manteve o mesmo comportamento, aceitando todas as variáveis e fornecendo estimadores razoáveis. Já o modelo MC performou pior, e foi necessário recalibrar os parâmetros τ e c para obter resultados melhores. Para o cenário 3, o modelo KM incluiu de forma errada 2 variáveis, enquanto do modelo MC incluiu 4 variáveis de forma errada. Para o cenário 4, tivemos4 falsas inclusões para o modelo KM e 6 falsas inclusões para o modelo de MC. Para o cenário 5, 15 falsas inclusões para o modelo de MC e 11 falsas inclusões para o modelo de KM.

Os resultados das simulações indicam que o modelo de MC precisa de um trabalho mais cuidadoso para a escolha do hiper-parâmetros do problema. Por este motivo, para as simulações praticas utilizo o modelo de KM como método de SSVS.

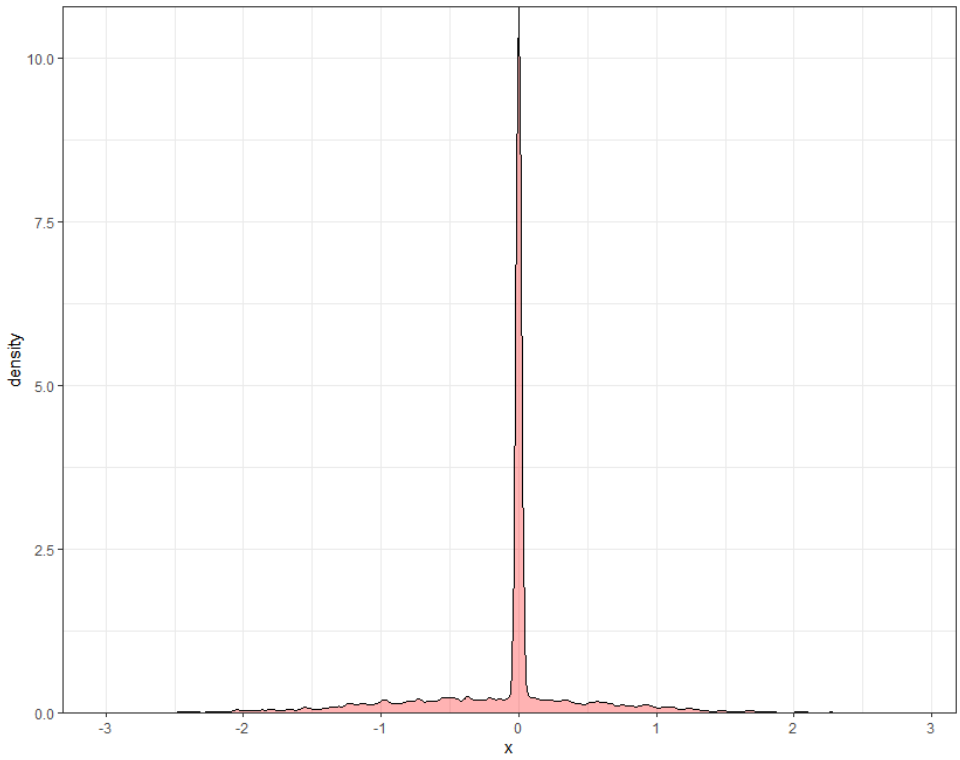


Figura 4 – Ilustração da densidade *à posteriori* de uma excluída pra o cenário de simulação 3 para o modelo de Kuo e Mallick. Veja como a concentração de densidade fica em zero, mas o modelo permitiu que ela explorasse outros possíveis valores

3 Aplicação 1: Câncer Labial na Escócia

(CLAYTON; KALDOR, 1987) apresentam dados de casos registrados durante 6 anos (1975-1980) de câncer labial na Escócia. Para cada um dos 56 condados da Escócia, temos como informação a contagem de casos observados, assim como o "valor esperado"de casos naquela região. Este valor esperado é obtido através de características demográficas e foram calculados de acordo com (MANTEL; STARK, 1968) e aqui são tratados como constantes dadas.

Ao invés de mapear os casos de câncer, mapeio a SIR (Standard Infection Ratio) em porcentagem para cada condado. A SIR é dada por:

$$SIR_i = 100 \times \frac{y_i}{E_i}, \quad i = 1, \ldots, n \tag{3.1}$$

onde y_i é o número de casos observados e E_i é o valor esperado de casos para o condado i. Temos também como covariável a porcentagem a população envolvida em atividades de agricultura, caça e pesca por condado. Ambas as informações estão mapeadas pelas Figura 5 e Figura 6.

Modelo o risco relativo θ_i em cada condado como

$$\log(\theta_i) = \beta_0 + \beta_1 \cdot AFF_i + \xi_i, \tag{3.2}$$

onde β_1 é o efeito da covariável AFF e ξ_i termo de erro para capturar a sobre-

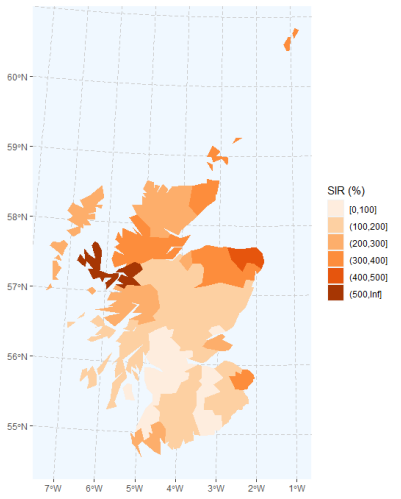


Figura 5 – Mapa da SIR em porcentagem para os condados da Escócia.

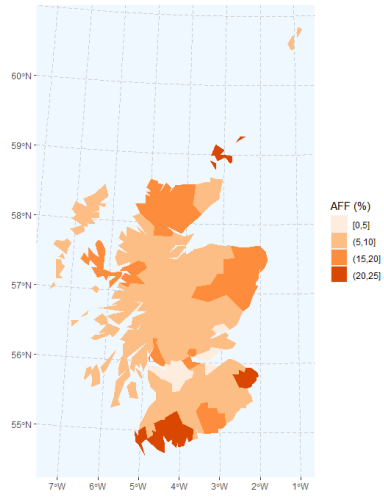


Figura 6 – Mapa da porcentagem da população envolvida em atividades de agricultura, caça e pesca (AFF).

	β_0		β_1		WAIC
Poisson lognormal	-0.49	(-0.80,-0.19)	6.88	(3.97, 9.64)	307.7
CAR	-0.31	(-0.55, -0.08)	4.38	(1.84, 6.82)	300.2
BYM	-0.33	(-0.59, -0.07)	4.94	(2.22, 7.50)	303.3
BYM2	-0.32	(-0.57, -0.06)	4.60	(1.74, 7.17)	298.8

Tabela 2 – Sumários à posteriori para os modelos de Câncer Labial na Escócia. Para os parâmetros beta a média a posteriori junto com os intervalos de credibilidade superiores e inferiores de 95%.

dispersão da Poisson. Serão usadas 4 formulações para ξ_i revisadas no Capítulo 2: erro não-estruturado, termo CAR, o modelo de BYM e o modelo BYM2. As simulações da posteriori são feitas utilizando Cadeias de Markov e Monte Carlo. A linguagem utilizada foi R na versão 4.1.3. Para rodar as cadeias, a biblioteca Nimble, versão 0.12.2 e em especial para o calculo do fator de escala do BYM2 o pacote INLA versão 22.05.03. Todas as cadeias foram rodadas para um número fixo de 10.000 iterações com 2.000 iterações de *burn-in*, 4 cadeias em paralelo para cada modelo. Todas as cadeias apresentaram no diagnóstico de Gelman-Rubin $\hat{R} < 1.05$.

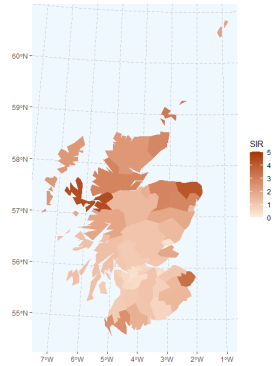
Para todos os modelos foram escolhidas prioris vagas para β_0 e β_1, normais de média zero e precisão 10^{-3}. Todos os modelos também seguem a recomendação de (STERRANTINO; VENTRUCCI; RUE, 2017) para modelar casos onde existem ilhas desconectadas (no caso da Escócia, existem 3 ilhas), e modelam a taxa relativa nas ilhas sem componente estruturada.

Os parâmetros de precisão para serem calibrados são τ_s, τ_u e τ_r, precisão da componente estruturada, precisão da componente não estruturada e precisão conjunta, respectivamente. Seguindo a mesma prática que (WAKEFIELD, 2007), atribuo prioris $Gamma(1, 0.1)$ para esses parâmetros. Para o parâmetro de mistura ρ do BYM2, a priori não-informativa $Beta(1, 1)$. A Tabela 2 mostra o sumário da mediana à posteriori junto com intervalos de credibilidade inferiores e superiores para o intercepto e o coeficiente a covariável AFF. O WAIC de cada modelo também é exibido na tabela. Note que entre os modelos que possuem componente espacial, quase não há diferença entre as estimativas dos parâmetros. A principal diferença entre eles é a interpretabilidade de cada hiperparâmetro, como nota também (RIEBLER et al., 2016).

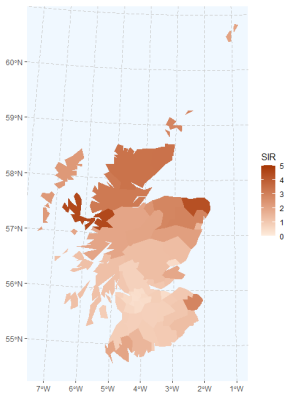
Para cada modelo temos também a SIR à posteriori como mostra a Figura 7. Observe como a componente espacial suaviza o mapa de risco da doença. Além disso, usando o teste de Moran no resíduo destes modelos, obtive O I de Moran com valor 0.21, contra um esperado de -0.019 (p-valor 0.006), alto indício de dependência espacial no resíduo. Já no modelo de BYM a estatística de Moran cai para -0.004 (p-valor de 0.86), ou seja, evidenciando uma diminuição da dependência espacial do resíduo, o que melhora

a interpretação dos parâmetros do modelo (WAKEFIELD, 2007).

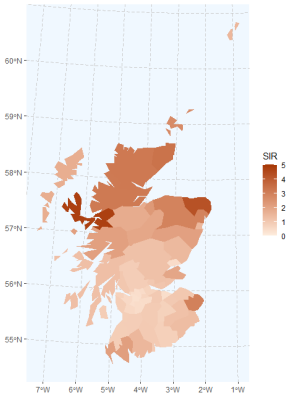
Vale notar o parâmetro ρ do modelo BYM2, que controla a mistura entre erro estruturado e não estruturado. A priori supuz indiferença para este parâmetro, e comecei as cadeias com valores de ρ diferentes, variando a contribuição de cada tipo de erro. A posteriori converge para uma concentração maior de erro estruturado. A mediana à posteriori de ρ foi $0.83(0.36, 0.99)$, o que pode explicar o Modelo de Besag ter tido um WAIC menor (afinal, ele possui erro estruturado puro). A vantagem do BYM2 é poder agora analisar o parâmetro τ_r e entender quanto da SIR à posteriori foi explicada pelo erro. Lembrando que o erro tem média constante zero, a mediana do desvio padrão à posteriori vale $0.47(0.34, 0.64)$.



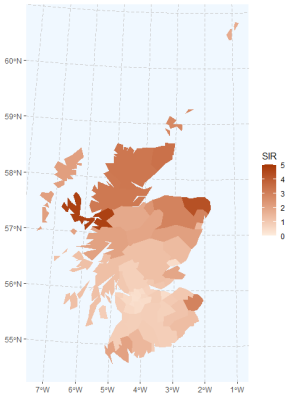
(a) Modelo Poisson log-normal puro



(b) Modelo de Besag



(c) Modelo BYM



(d) Modelo BYM2

Figura 7 – SIR à posteriori média para cada um dos modelos analisados. Observe como o modelo sem componente espacial possui o mapa da SIR bem similar ao da taxa crua, e como os modelos espaciais distribuem o efeito da doença para as vizinhanças, promovendo suavização.

Uma última análise interessante é verificar a probabilidade à posteriori da SIR em um determinado condado ser maior que 1. A figura trás essa visualizacão para o modelo BYM2.

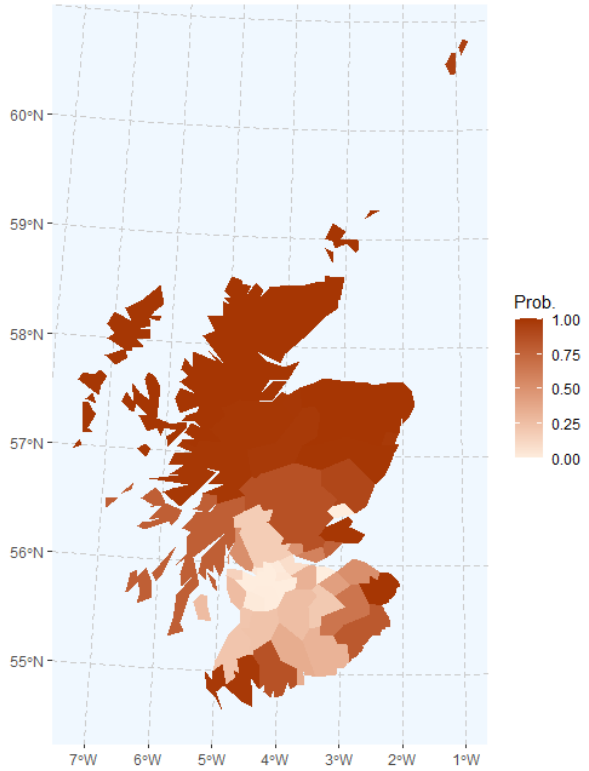


Figura 8 – Probabilidade à posteriori da SIR ser maior que 1 (modelo BYM2)

4 Aplicação 2: Epidemia de Ebola na África Ocidental

Em meados de março de 2014 os primeiros casos de Ebola foram detectados na Guiné, e futuras investigações epidemiológicas iriam sugerir que os primeiros casos ocorreram por volta de Dezembro de 2013. Em março de 2016, já se somavam 28.816 suspeitas, possibilidades e casos confirmados de Ebola em Guiné, Libéria e Serra Leoa, com 11.310 mortes. Isso marca a epidemia de Ebola do período de 2013-2016 como uma das piores na história (CARVALHO, 2019). Isso pois o Ebola já tem casos registrados desde 1976, o primeiro deles na República Democrática do Congo. A pergunta científica que se faz então é: quais foram os fatores relevantes para o considerável aumento repentino do número de casos e de sua propagação geográfica. Para isso, vamos utilizar as técnicas desenvolvidas ao longo do texto para lidar com este problema.

Seguindo os estudos de (CARVALHO, 2019), um rico conjunto de dados socioeconômicos, climáticos e genéticos foi disponibilizado. Os dados vem agregados em área para os distritos da Serra Leoa, prefeituras da Guiné e condados da Libéria. Das 81 localidades resultantes, 63 reportaram 1 ou mais casos de Ebola. Essas regiões estão dispostas na Figura 9.

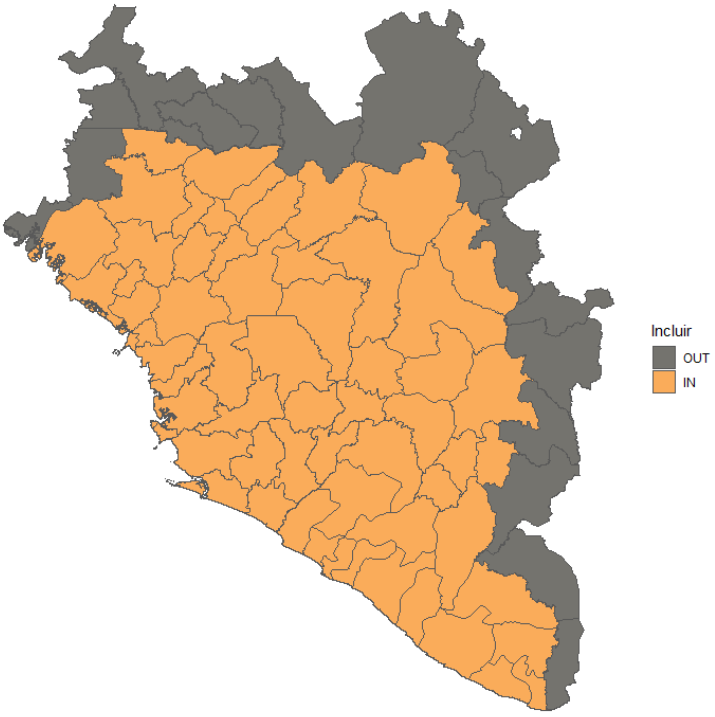


Figura 9 – Regiões da África Ocidental consideradas no modelo. Regiões com a marcação IN registraram 1 ou mais casos de Ebola.

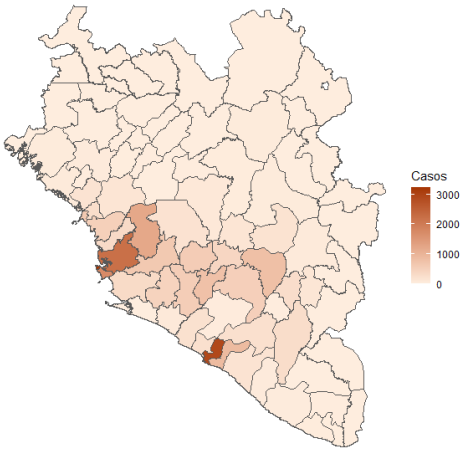


Figura 10 – Mapa dos casos brutos de ebola.

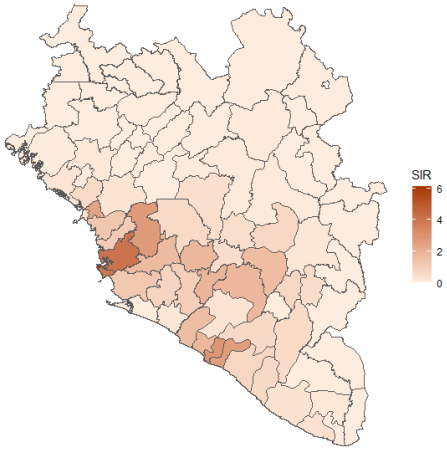


Figura 11 – Mapa da SIR para o caso do Ebola. A população suscetível foi considerada a população total.

Os preditores utilizados estão disponíveis na Tabela 3.

Para a modelar a contagem de casos utilizarei 6 modelos diferentes: Poisson Log-normal simples, Poisson Log-normal com ORLE e o modelo BYM2, e todos estes modelos novamente mas com a inclusão do esquema de seleção de variáveis SSVS de Kuo e Mallick. Para cada modelo também teremos o WAIC como comparativo, a eficiência em (ESS/s) e o erro médio quadrático à posteriori, disposta na Tabela 5. Foi rodada uma única cadeia com 300.000 iterações, 200.000 de *burnin* com parâmetro de thinning igual a 10 para cada modelo.

Os diagnósticos de cadeia não foram muito satisfatórios. A quantidade de amostras efetivas produzidas por rodada é bem baixa. Os parâmetros β apresentaram alta autocorrelação, o que sinaliza convergência lenta. Apesar disso, os resultados em termos de WAIC e RMSE foram bons. Antes de comparara relevância das variáveis, um resultado interessante apresentado foi o BYM2 no cenário de seleção de variáveis. Aplicando o teste de Moran para cada covariável, apenas uma não apresenta forte estrutura espacial que foi a de densidade populacional. Todas as outras tiveram p-valor menor que 0.0001. O BYM2 acabou absorvendo toda a informação espacial, e nenhuma variável foi selecionada mais de 20 % das vezes. No modelo com a inclusão de todas as variáveis o parâmetro ρ se comportou de forma parecida com a priori uniforme, comportamento similar aos testes da seção 2.1 de quando não havia resíduo espacial. Na seleção de variáveis, a média a posteriori de ρ salta para 0.8, indicando forte associação espacial, e nenhuma variável foi selecionada.

Por este motivo, utilizarei o modelo com ORLE para analisar a relevância de variáveis. Um ponto a se notar aqui é a alta correlação entre as covariáveis (Figura 12), que afeta a interpretação dos resultados. outro problema está na convergência: a cadeia visita

Tipo de preditor	Abreviação	Descrição
Demográfico	pdensMN	Densidade populacional (hab/km^2)
Demográfico	geconMN	Produção econômica média
Demográfico	geconMIN	Mínimo da produção econômica
Demográfico	geconMAX	Máximo da produção econômica
Demográfico	geconSTD	Desvio padrão da produção econômica
Demográfico	ttXkMN	Tempo de viagem média estimado para a cidade mais próxima com uma população de pelo menos $X \times 10.000$ habitantes, para $X = 50, 100$ e 500
Climático	altMN	Altitude média
Climático	tempMN	Temperatura anual média
Climático	tempssMN	Sazonalidade média da temperatura
Climático	precMN	Precipitação anual média
Climático	precssMN	Sazonalidade média da precipitação
Filogenético	Introadmin	Número médio de introduções virais preditas usando fronteiras administrativas
Filogenético	Introdista	Número médio de introduções virais preditas usando fronteiras administrativas

Tabela 3 – Preditores considerados para a modelagem da epidemia de Ebola. Todos os preditores contínuos foram padronizados subtraindo-se a média e dividindo pelo desvio-padrão.

o espaço de modelos de forma bem lenta. A tendência observada foi de terem variáveis que sempre são selecionadas, e as outras variáveis sendo selecionadas esporadicamente. A cadeia consegue se desvencilhar do estado inicial. Para avaliar a situação melhor, rodei 400 cadeias em paralelo, observando as frequências de modelo escolhidas em cada uma. As cadeias começam todas de pontos diferentes: 100 delas começando de $\gamma = (0, \ldots, 0)$, 100 delas começando em $\gamma = (1, \ldots, 1)$, e o resto de algum ponto aleatório uniformemente escolhido em $\{0, 1\}^p$. Apresento como resultado as médias à posteriori da indicadora de cada covariável:

As análises do fator de Bayes não serão interessantes nesse cenário de fraca convergência da cadeia. Mas podemos gerar mapas da probabilidade *à posteriori* da SIR ser maior do que 1, assim como realizar a predição das regiões sem casos. Aqui o modelo BYM2 mostra sua vantagem em relação ao BYM. A posteriori para τ_b é compatível com

Covariável	**Média**
geconMN	0.09
geconMIN	0.36
geconMAX	0.14
geconSTD	0.04
pdensMN	0.07
tt50kMN	0.20
tt100kMN	0.38
tt500kMN	0.18
altMN	0.21
tempMN	0.33
tmpssMN	0.42
precMN	0.16
precssMN	0.56
Introadmin	0.04
Introdista	0.03

Tabela 4 – Médias *à posteriori* das indicadoras das covariáveis para as múltiplas cadeias do modelo com ORLE.

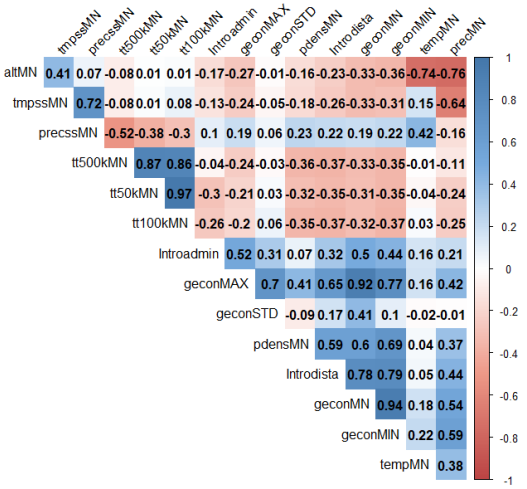
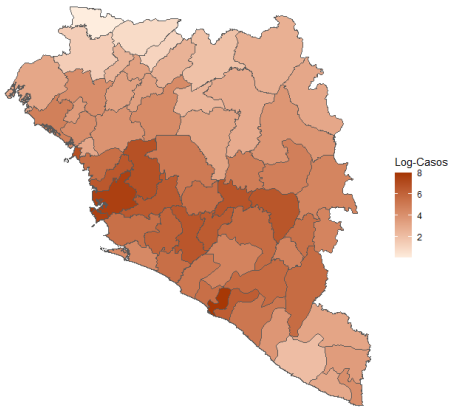
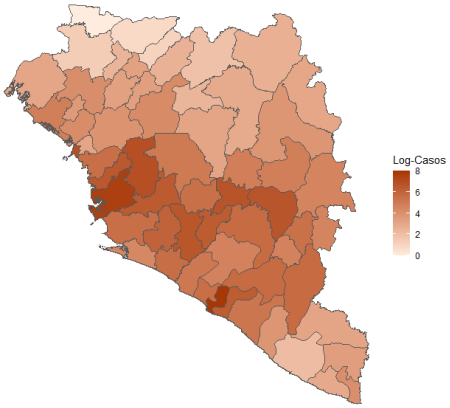


Figura 12 – Correlação entre as variáveis disponíveis para o estudo do Ebola. Destas, apenas a densidade populacional (pdensMN) não apresenta forte estrutura espacial.

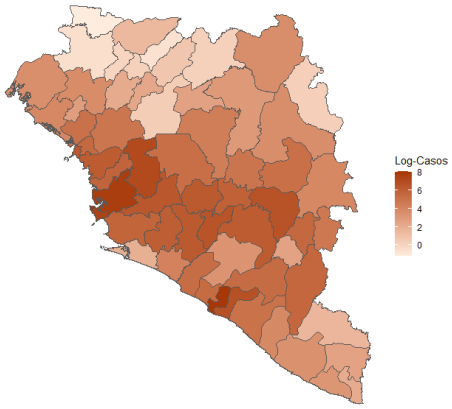
ambos os grafos de adjacências. Os resultados estão dispostos na Figura 14



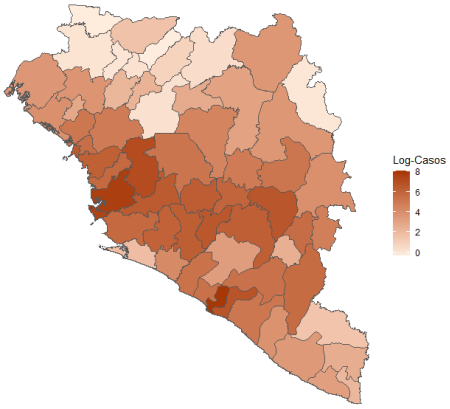
(a) Modelo simples



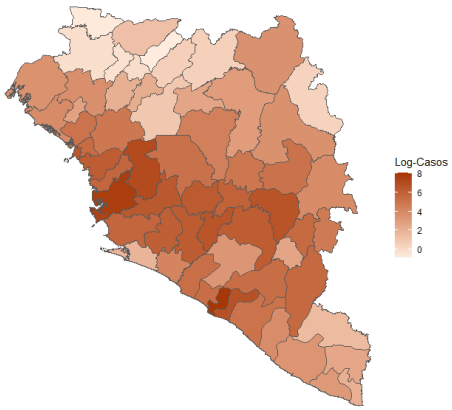
(b) Modelo simples com seleção de variáveis



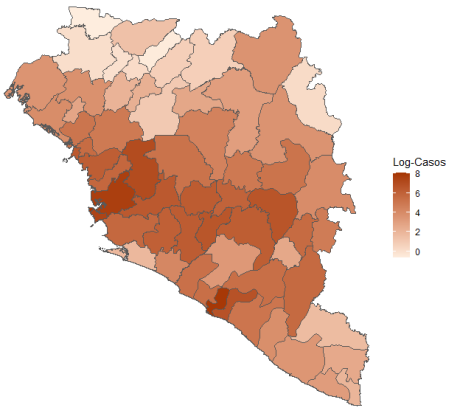
(c) Modelo com ORLE



(d) Modelo com ORLE e seleção de variáveis



(e) Modelo BYM2



(f) Modelo BYM2 com seleção de variáveis

Figura 13 – Mapa do log da média a posteriori de casos para os seis modelos de contagem de casos utilizados

	SSVS	WAIC	Eficiência	Eficiência mínima	RMSE
Simples	Não	5856.5	126.2	30.4	139.9
ORLE	Não	459.7	145.1	0.03	0.68
BYM2	Não	699.4	30.25	0.03	0.77
Simples	Sim	5670.6	6.26	0.165	141.0
ORLE	Sim	475.5	37.0	0.01	0.72
BYM2	Sim	711.9	39.05	0.01	1.24

Tabela 5 – Comparativo dos 6 modelos usados para modelagem dos casos de Ebola. Eficiência (ESS por segundo) representa a eficiência média de todos os parâmetros da cadeia. Eficiência mínima é a menor eficiência encontrada. RMSE é a raiz do Erro Médio Quadrático.

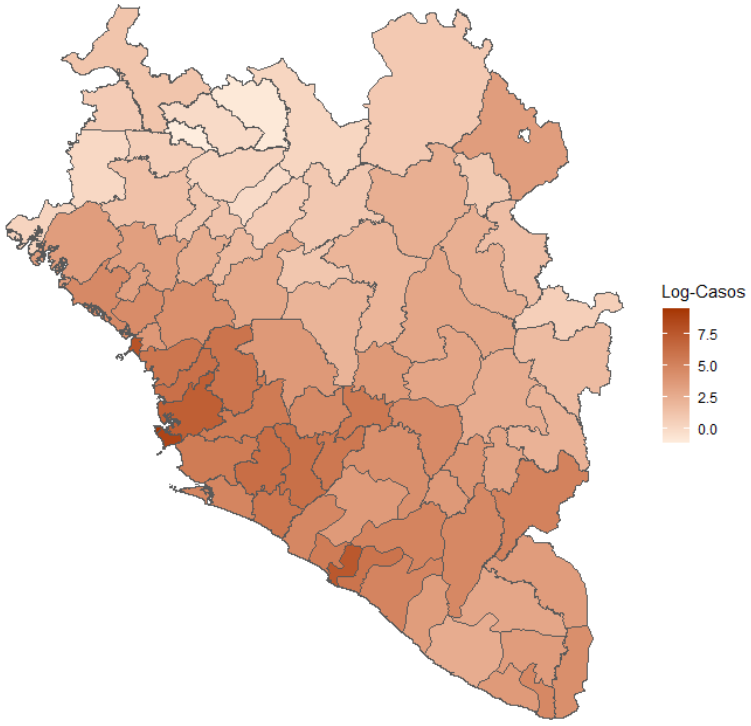


Figura 14 – Preditiva à posteriori, incluindo regiões agora que não tiveram casos registrados.

5 Conclusão

Modelos hierárquicos Bayesianos são um forte aliado de qualquer pesquisador da área de epidemiologia. Uma de suas grandes vantagens é ser altamente interpretável, e poder acomodar todo tipo de hipótese dentro de sua formulação. Em especial, para epidemiologia espacial conseguimos acoplar a necessidade por dependência espacial, que está presente em todas as coisas, como diz a Primeira Lei da Geografia de (TOBLER, 1970). Conseguimos acoplar também uma análise de regressão, uma poderosa ferramente estatística, mas que deve ser tratada com cautela. Neste trabalho tivemos complicações com o modelo do Ebola, não obtendo a desejada convergência da cadeia. Apesar do modelo ter tido bons resultados em termos de RMSE, resultados estatísticos não podem ser traçados.

Possíveis correções e sugestões para futuros trabalhos são: analisar diferentes amostradores e sua eficiência comparada aos amostrados base do NIMBLE. O próprio NIMBLE já possui em versão beta amostradores por HMC (Hamiltonian Monte Carlo). Outra opção é o INLA (INtegrated Nested Laplace Approximation). Mesmo que tivessemos conseguido um bom amostrador, temos o problema da alta correlação entre as covariáveis do ebola, que pode distribuir a importância dessas variáveis e levar à interpretação errada de parâmetros. Algumas sugestões são adicionar ao modelo através das prioris, estruturas que permitam a agregação de variáveis correlacionadas, como em (LI; ZHANG et al., 2015). Uma terceira indagação para o futuro é sobre quais seriam os melhores diagnósticos de cadeia para este cenário. Veja que no conjunto de dados do Ebola, apesar de não termos atingido convergência da cadeia, obtivemos bons resultados de RMSE, o que poderia levar a conclusões erradas sobre a eficácia do modelo.

Referências

ANSELIN, L.; GRIFFITH, D. A. Do spatial effects really matter in regression analysis? In: PAPERS - Regional Science Association. [S.l.: s.n.], 1988. v. 65, p. 11–34.

ARBIA, Giuseppe. **Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems**. [S.l.]: Springer Dordrecht, 1989. 1st ed.

BANERJEE, Sudipto; CARLIN, Bradley P.; GELFAND, Alan E. **Hierarchical Modeling and Analysis for Spatial Data**. [S.l.]: CRC Press, 2015. 2nd ed.

BESAG, Julian; YORK, Jeremy; MOLLIÉ, Annie. Bayesian Image Restoration, With Two Applications in Spatial Statistics. **Biometrics**, IBS, n. 4, p. 997–1005, 1991.

CÂMARA, G. et al. **Análise Espacial de Dados Geográficos**. [S.l.]: Embrapa, 2004. 1st ed.

CARVALHO, Luiz Max De Fagundes. **Statistical approaches to viral phylodynamics**. 2019. Tese (Doutorado) – University of Edinburgh.

CLAYTON, David; KALDOR, John. Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping. **Biometrics**, IBS, v. 43, n. 3, p. 671–681, 1987.

CLIFF, A. D. (Andrew David). **Spatial processes : models & applications / A.D. Cliff & J.K. Ord.** London: Pion, 1981. ISBN 0850860814.

DEAN, C. B.; UGARTE, M. D.; MILITINO, A. F. Detecting Interaction between Random Region and Fixed Age Effects in Disease Mapping. **Biometrics**, [Wiley, International Biometric Society], v. 57, n. 1, p. 197–202, 2001. ISSN 0006341X, 15410420. Disponível em: <http://www.jstor.org/stable/2676860>. Acesso em: 20 nov. 2022.

GEARY, R. C. The Contiguity Ratio and Statistical Mapping. English (US). **Royal Statistical Society**, Wiley, v. 5, n. 3, p. 115-127+129–146, nov. 1954.

GEORGE, Edward I.; MCCULLOCH, Robert E. Variable Selection via Gibbs Sampling. **Journal of the American Statistical Association**, Taylor & Francis, v. 88, n. 423, p. 881–889, 1993. DOI: 10.1080/01621459.1993.10476353. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1993.10476353. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353>.

GETIS, Arthur; ORD, J. K. The Analysis of Spatial Association by Use of Distance Statistics. **Geographical Analysis**, v. 24, n. 3, p. 189–206, 1992. DOI: https://doi.org/10.1111/j.1538-4632.1992.tb00261.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1992.tb00261.x. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1992.tb00261.x>.

HARRISON, Xavier A. Using observation-level random effects to model overdispersion in count data in ecology and evolution. en. **PeerJ**, United States, v. 2, e616, out. 2014.

ISO/TC 211 Geographic information/Geomatics. 1994. Disponível em: <https://www.iso.org/committee/54904.html>. Acesso em: 25 out. 2022.

KUO, Lynn; MALLICK, Bani. Variable Selection for Regression Models. **Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)**, Springer, v. 60, n. 1, p. 65–81, 1998. ISSN 05815738. Disponível em: <http://www.jstor.org/stable/25053023>. Acesso em: 25 nov. 2022.

LEROUX, Brian G.; LEI, Xingye; BRESLOW, Norman. Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. In_____. **Statistical Models in Epidemiology, the Environment, and Clinical Trials**. New York, NY: Springer New York, 2000. P. 179–191.

LI, Fan; ZHANG, Tingting et al. Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. **The Annals of Applied Statistics**, Institute of Mathematical Statistics, v. 9, n. 2, p. 687–713, 2015. DOI: 10.1214/15-AOAS818. Disponível em: <https://doi.org/10.1214/15-AOAS818>.

LI, Hongfei; CALDER, Catherine A.; CRESSIE, Noel. Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. **Geographical Analysis**, v. 39, n. 4, p. 357–375, 2007. DOI: https://doi.org/10.1111/j.1538-4632.2007.00708.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.2007.00708.x. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.2007.00708.x>.

MANTEL, Nathan; STARK, Charles R. Computation of Indirect-Adjusted Rates in the Presence of Confounding. **Biometrics**, IBS, n. 4, p. 997–1005, 1968.

MILLER, Harvey J. Tobler's First Law and Spatial Analysis. **Annals of the Association of American Geographers**, Taylor & Francis, Ltd., n. 2, p. 284–289, 2004.

MILLER, Harvey J.; WENTZ, Elizabeth. Representation and spatial analysis in geographic information systems. English (US). **Annals of the American Association of Geographers**, Taylor & Francis Ltd., v. 93, n. 3, p. 574–594, set. 2003. ISSN 2469-4452. DOI: 10.1111/1467-8306.9303004.

MORAN, P. A. P. Notes on Continuous Stochastic Phenomena. English (US). **Biometrika**, Oxford University Press, v. 37, n. 1/2, p. 17–23, jun. 1950.

O'HARA, R. B.; SILLANPÄÄ, M. J. A review of Bayesian variable selection methods: what, how and which. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 4, n. 1, p. 85–117, 2009. DOI: 10.1214/09-BA403. Disponível em: <https://doi.org/10.1214/09-BA403>.

RIEBLER, Andrea et al. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. **Statistical Methods in Medical Research**, v. 25, n. 4, p. 1145–1165, 2016. PMID: 27566770. DOI: 10.1177/0962280216660421. eprint: https://doi.org/10.1177/0962280216660421. Disponível em: <https://doi.org/10.1177/0962280216660421>.

RUE, Havard; HELD, Leonhard. **Gaussian Markov Random Fields: Theory and Applications**. [S.l.]: Chapman e Hall/CRC, 2005. P. 280. 1st ed.

SIMPSON, Daniel et al. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. **Statistical Science**, Institute of Mathematical Statistics, v. 32, n. 1, p. 1–28, 2017. DOI: 10.1214/16-STS576. Disponível em: <https://doi.org/10.1214/16-STS576>.

SØRBYE, Sigrunn Holbek; RUE, Håvard. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. **Spatial Statistics**, v. 8, p. 39–51, 2014. Spatial Statistics Miami. ISSN 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2013.06.004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2211675313000407>.

STERRANTINO, Anna; VENTRUCCI, Massimo; RUE, Håvard. A note on intrinsic Conditional Autoregressive models for disconnected graphs. **Spatial and Spatio-temporal Epidemiology**, v. 26, mai. 2017. DOI: 10.1016/j.sste.2018.04.002.

TOBLER, W. R. A Computer Movie Simulating Urban Growth in the Detroit Region. **Economic Geography**, Clark University, p. 234–240, 1970.

VAN ERP, Sara; OBERSKI, Daniel L.; MULDER, Joris. Shrinkage priors for Bayesian penalized regression. **Journal of Mathematical Psychology**, v. 89, p. 31–50, 2019. ISSN 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2018.12.004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022249618300567>.

WAKEFIELD, Jon. Disease mapping and spatial regression with count data. **Biostatistics (Oxford, England)**, v. 8, p. 158–83, mai. 2007. DOI: 10.1093/biostatistics/kxl008.

WATANABE, Sumio. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. **Journal of Machine Learning Research**, v. 11, n. 116, p. 3571–3594, 2010. Disponível em: <http://jmlr.org/papers/v11/watanabe10a.html>.

Apêndices

APÊNDICE A – Gaussian Markov Random Fields

Gaussian Markov Random Fields (GMRF's) são objetos simples: vetores aleatórios (de dimensão finita) seguindo uma distribuição normal multivariada. Apesar disso, nosso interesse está em versões mais restritas das GMRF's, que satisfazem certas condições de independência (daí o cunho *Markov*).

Definição A.0.1. Seja $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ um vetor aleatório com distribuição normal de média μ e matriz de covariância $\boldsymbol{\Sigma}$. Defina o grafo $G = (V, E)$, onde $V = \{1, \ldots, n\}$, tal que não há aresta conectando dois nós i e j se, e somente se, $x_i \perp x_j | \boldsymbol{x}_{-ij}$. Dizemos que $\boldsymbol{x}$ é um GMRF com respeito ao grafo G.

A matriz de precisão da normal multivariada explicita algumas características de independência condicional. Em especial, temos o teorema abaixo:

Teorema 1. *Seja $\boldsymbol{x}$ um vetor com distribuição normal de média μ e matriz de precisão $Q \succ 0$. Então, para $i \neq j$,*

$$x_i \perp x_j | \boldsymbol{x}_{-ij} \iff Q_{ij} = 0. \tag{A.1}$$

Prova. Para isso, particionamos o vetor $\boldsymbol{x} = (x_i, x_j, \boldsymbol{x}_{-ij})$ e utilizamos do critério da fatorização. Sem perda de generalidade, assuma $\mu = 0$. A densidade conjunta de $\boldsymbol{x}$ pode ser expressa como:

$$\begin{aligned} \pi(\boldsymbol{x}) = \pi(x_i, x_j, \boldsymbol{x}_{-ij}) &\propto \exp(-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x}) \\ &\propto \exp\left(-\frac{1}{2}\sum_{k,l} x_k Q_{k,l} x_l\right) \\ &\propto \exp\left(\underbrace{-\frac{1}{2}x_i x_j (Q_{ij} + Q_{ji})}_{\text{Termo 1}} \underbrace{-\frac{1}{2}\sum_{\{k,l\}\neq\{i,j\}} x_k Q_{kl} x_l}_{\text{Termo 2}}\right). \end{aligned} \tag{A.2}$$

Veja que o Termo 2 não envolve o produto $x_i x_j$, enquanto o Termo 1 envolve $x_i x_j$ se, e somente se, $Q_{ij} \neq 0$. Então, se vale que $Q_{ij} = 0$, é possível fatorar $\pi(\boldsymbol{x})$ como:

$$\pi(x_i, x_j, \boldsymbol{x}_{-ij}) = f(x_i, \boldsymbol{x}_{-ij}) g(x_j, \boldsymbol{x}_{-ij}). \tag{A.3}$$

Pelo critério de fatorização, x_i e x_j são independentes se, e somente se, $Q_{ij} = 0$

□

Faz sentido então caracterizar os GMRF's através de sua matriz de precisão. Considere então a seguinte reformulação de A.0.1:

Definição A.0.2. Um vetor aleatório $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ é um GMRF com respeito ao grafo $G = (V, E)$ com média μ e matriz de precisão $\mathbf{Q} \succ 0$ se, e somente se, sua densidade é da forma:

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\mathbf{Q}|^{1/2} \exp(-\frac{1}{2}\left(\boldsymbol{x} - \mu)^T\mathbf{Q}(\boldsymbol{x} - \mu)\right), \tag{A.4}$$

e

$$Q_{ij} \neq 0 \iff \{i, j\} \in E \quad \forall i \neq j.$$

Uma forma alternativa de caracterizar um GMRF é através das condicionais completas. Essa abordagem foi desenvolvida por Besag e os modelos são abreviados como CAR (*conditional autoregressions*). Então suponha que para o vetor de dados $\mathbf{x} = (x_1, \ldots, x_n)$ especificamos as condicionais completas como normais de parâmetros:

$$\mathbb{E}(x_i|\boldsymbol{x}_{-i}) = \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j), \tag{A.5}$$

$$\text{Var}(x_i|\boldsymbol{x}_{-i}) = \frac{1}{\kappa_i} > 0. \tag{A.6}$$

Aqui, $\sim$ [e uma relação de simetria definida implicitamente pelos termos não-nulos de $\boldsymbol{\beta}$. Essas condicionais completas devem ser consistentes de modo a gerar a densidade conjunta $\pi(\boldsymbol{x})$. De fato, veja que se escolhermos as entradas da matriz de precisão como:

$$Q_{ii} = \kappa_i \quad \text{and} \quad Q_{ij} = \kappa_i\beta_{ij}$$

com a condição de Q simétrica, ou seja,

$$\kappa_i\beta_{ij} = \kappa_j\beta_{ji}$$

temos um candidato a densidade conjunta. Este candidato a densidade é único, e para mostrar isso precisamos do seguinte lema.

Lema 1 (Lema de Brook)**.** *Seja $\pi(\boldsymbol{x})$ a densidade de $\boldsymbol{x} \in \mathbb{R}^n$ e defina o suporte dessa densidade como o conjunto $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \pi(\boldsymbol{x}) > 0\}$. Se $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$, então:*

$$\frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}')} = \prod_{i=1}^{n} \frac{\pi(x_i|x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)}{\pi(x'_i|x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)} \tag{A.7}$$

$$= \prod_{i=1}^{n} \frac{\pi(x_i|x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_n)}{\pi(x'_i|x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_n} \tag{A.8}$$

Demostração. Como estamos dentro do suporte da distribuição, podemos usar as definições de densidade condicional para obter a seguinte relação:

$$\frac{\pi(x_n|x_1,\ldots,x_{n-1})\pi(x_1,\ldots,x_{n-1})}{\pi(x'_n|x_1,\ldots,x_{n-1})\pi(x_1,\ldots,x_{n-1})} = \frac{\pi(x_1,\ldots,x_n)}{\pi(x_1,\ldots,x'_n)}$$

que nos permite expressar a densidade conjunta como:

$$\pi(x_1,\ldots,x_n) = \frac{\pi(x_n|x_1,\ldots,x_{n-1})}{\pi(x'_n|x_1,\ldots,x_{n-1})} \cdot \pi(x_1,\ldots,x_{n-1},x'_n).$$

Podemos da mesma forma expressar o último termo da equação acima como

$$\pi(x_1,\ldots,x_{n-1},x'_n) = \frac{\pi(x_{n-1}|x_1,\ldots,x_{n-2},x'_n)}{\pi(x'_{n-1}|x_1,\ldots,x_{n-2},x'_n)} \cdot \pi(x_1,\ldots,x_{n-2},x'_{n-1},x'_n).$$

Substituindo na densidade conjunta

$$\pi(x_1,\ldots,x_n) = \frac{\pi(x_n|x_1,\ldots,x_{n-1})}{\pi(x'_n|x_1,\ldots,x_{n-1})} \cdot \frac{\pi(x_{n-1}|x_1,\ldots,x_{n-2},x'_n)}{\pi(x'_{n-1}|x_1,\ldots,x_{n-2},x'_n)} \cdot \pi(x_1,\ldots,x_{n-2},x'_{n-1},x'_n).$$

Basta então repetir o procedimento até o termo à direita se tornar a densidade conjunta de $\boldsymbol{x}'$ $\square$.

Com isso, se fixarmos $\boldsymbol{x}'$, obtemos a densidade de $\boldsymbol{x}$ a menos de uma constante de proporcionalidade.

Teorema 2. *Dadas as n condicionais completas como em A.5, então $\boldsymbol{x}$ é um GMRF com média μ e matriz de precisão $Q = (Q_{ij})$, onde*

$$Q_{ij} = \begin{cases} \kappa_i\beta_{ij} & i \neq j \\ \kappa_i & i = j \end{cases}$$

com $\kappa_i\beta_{ij} = \kappa_j\beta_{ij}$ e $Q \succ 0$.

Demonstração. Sem perda de generalidade, assuma $\mu = 0$ e fixe $\boldsymbol{x}' = 0$ no Lema de Brook. Então o log de A.7 pode ser simplificado para:

$$\log\frac{\pi(\boldsymbol{x})}{\pi(0)} = -\frac{1}{2}\sum_{i=1}^{n}\kappa_i x_i^2 - \sum_{i=2}^{n}\sum_{j=1}^{i-1}\kappa_i\beta_{ij}x_i x_j, \tag{A.9}$$

e A.8 para

$$\log\frac{\pi(\boldsymbol{x})}{\pi(0)} = -\frac{1}{2}\sum_{i=1}^{n}\kappa_i x_i^2 - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\kappa_i\beta_{ij}x_i x_j. \tag{A.10}$$

Para isso, basta ver como cada termo do produtório se comporta

$$\prod_{i=1}^{n}\frac{\pi(x_i|x'_1,\ldots,x'_{i-1},x_{i+1},\ldots,x_n)}{\pi(x'_i|x'_1,\ldots,x'_{i-1},x_{i+1},\ldots,x_n)} = \exp\left(-\frac{\kappa_i}{2}(x_i^2 - 2x_i[\sum_{\substack{j:i\sim j\\ j<i}}\beta_{ij}x'_j + \sum_{\substack{j:i\sim j\\ j>i}}\beta_{ij}x_j])\right).$$

$$\prod_{i=1}^{n} \frac{\pi(x_i|x_1,\ldots,x_{i-1},x'_{i+1},\ldots,x'_n)}{\pi(x'_i|x_1,\ldots,x_{i-1},x'_{i+1},\ldots,x'_n)} = \exp\left(-\frac{\kappa_i}{2}(x_i^2 - 2x_i[\sum_{\substack{j:i\sim j\\ j>i}} \beta_{ij}x'_j + \sum_{\substack{j:i\sim j\\ j<i}} \beta_{ij}x_j])\right).$$

Como ambas devem ser iguais, segue que $\kappa_i\beta_{ij} = \kappa_j\beta_{ij}$, para $i \neq j$. A densidade de $\boldsymbol{x}$ então pode ser expressa como

$$\log \pi(\boldsymbol{x}) = \text{const} - \frac{1}{2}\sum_{i=1}^{n} \kappa_i x_i^2 - \frac{1}{2}\sum_{i\neq j} \kappa_i\beta_{ij}x_i x_j. \qquad \square$$

APÊNDICE B – Intrinsic Gaussian Markov Random Fields

Neste apêndice introduzo um tipo especial de GMRF chamadas de IGMRF, Campos Gaussianos Aleatórios de Markov Intrínsecos. Antes de começar a falar sobre IGMRF's, precisamos enunciar algumas definições de álgebra linear.

Primeiramente, o *núcleo* ou espaço nulo de uma matriz $\mathbf{A}$ é o conjunto de todos os vetores $\mathbf{x}$ tais que $\mathbf{Ax} = \mathbf{0}$. A *nulidade* é a dimensão do núcleo. Para uma matriz $n \times m$ o *posto* é definido como $\min(m, n) = k$, onde k é a nulidade de $\mathbf{A}$. Para uma matriz singular, i.e, não inversível, com nulidade k, denotamos $|\mathbf{A}|^*$ o produto dos $n - k$ autovalores não nulos de $\mathbf{A}$.

Com isso, podemos definir o que é uma IGMRF de primeira ordem.

Definição B.0.1. Seja $\mathbf{Q}$ uma matriz $n \times n$ semi-positiva definida com posto $n - k > 0$. Então $\mathbf{x} = (x_1, \ldots, x_n)^T$ é uma GMRF imprópria de posto $n - k$ com parâmetros $(\mu, \mathbf{Q})$, se sua densidade é

$$\pi(\mathbf{x}) = (2\pi)^{\frac{-(n-k)}{2}} (|\mathbf{Q}|^*)^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}(\mathbf{x} - \mu)\right). \tag{B.1}$$

Uma GMRF intrinseca é uma GMRF imprópria de rank $n - 1$, onde o vetor $\mathbf{1} = (1, \ldots, 1)$ gera o espaço nulo de $\mathbf{Q}$. Ou seja, $\mathbf{Q1} = 0$. Já é possível ver que a densidade de uma IGMRF é invariante a adição de uma constante $c\mathbf{1} = \mathbf{c}$, basta olhar o núcleo da exponencial na densidade.

$$\begin{aligned}(\mathbf{x} - \mu + \mathbf{c})^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}(\mathbf{x} - \mu + \mathbf{c}) &= (\mathbf{x} - \mu)^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}(\mathbf{x} - \mu) + \mathbf{c}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}(\mathbf{x} - \mu) \\ &\quad + (\mathbf{x} - \mu)^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}\mathbf{c} + \mathbf{c}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}\mathbf{c} \\ &= (\mathbf{x} - \mu)\mathbf{Q}^{\mathbf{T}}(\mathbf{x} - \mu)\end{aligned}$$

É importante notar isso para impor restrições para a IGMRF somar zero. Caso contrário, pode haver confusão com o intercepto.

"