UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Ruy Azevedo Cota Vasconcelos

Regressão Bessel Bayesiana com Efeito Espaço-Temporal para Dados Contínuos Limitados

Belo Horizonte
2023

Ruy Azevedo Cota Vasconcelos

Regressão Bessel Bayesiana com Efeito Espaço-Temporal para Dados Contínuos Limitados

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Vinícius Diniz Mayrink

Belo Horizonte
2023

Vasconcelos, Ruy Azevedo Cota.

V331r Regressão bessel bayesiana com efeito espaço-temporal para dados contínuos limitados [recurso eletrônico] / Ruy Azevedo Cota Vasconcelos – 2023.
87 f. il.

Orientador: Vinícius Diniz Mayrink
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística
Referências: f.75-78.

1. Estatística – Teses. 2. Modelos lineares (Estatística)– Teses. 3. Democracia – Índice – Teses.4. Modelo espaço-temporal – Teses. I. Mayrink, Vinícius Diniz. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais – ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

FOLHA DE APROVAÇÃO

"Regressão Bessel Bayesiana com Efeito Espaço-Temporal para Dados Contínuos Limitados"

RUY AZEVEDO COTA VASCONCELOS

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

Aprovada em 31 de outubro de 2023, pela banca constituída pelos membros:

Prof. Vinícius Diniz Mayrink - Orientador
DEST/UFMG

Prof. Fabio Nogueira Demarqui
DEST/UFMG

Prof. Wagner Hugo Bonat
DEST/UFPR

Belo Horizonte, 31 de outubro de 2023.

Agradecimentos

Ao meu orientador, Vinicius Diniz Mayrink, pelo apoio constante no desenvolvimento do trabalho e pela excelência da orientação em todos os aspectos, desde a pesquisa bibliográfica até a correção final do texto. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) pela concessão da bolsa de estudos durante todo o período do curso de Mestrado. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) pelo fornecimento de insumos e recursos computacionais aos laboratórios da pós-graduação do Departamento de Estatística.

Resumo

Este trabalho contribuiu com uma implementação Bayesiana do modelo de regressão Bessel para dados limitados com estrutura de correlação espaço-temporal e apresentou comparações abrangentes entre diferentes modelos. Introduzimos três modelos - M_1, M_2, e M_3 - que diferem principalmente na inclusão de covariáveis e efeitos temporais. O M_1 pode ser considerado uma simplificação do M_2 na qual sumarizamos as covariáveis com medições ao longo do tempo (M_2 usa todas as medições de tempos distintos). Ambos M_1 e M_2 incorporam os efeitos aleatórios (espacial e temporal) de forma aditiva no preditor linear explicando a média. Por outro lado, em M_3, apenas o efeito espacial é aditivo e o efeito temporal é introduzido por meio de coeficientes variando no tempo. Realizamos simulações em vários cenários de má-especificação dos efeitos e comparamos os Vícios Relativos (VRs) das estimativas dos principais parâmetros. Em casos onde não há informação sobre a estrutura dos dados, M_2 se mostrou o mais indicado, pois apresentou ajustes com menores VRs em situações de má-especificação. No entanto, quando o analista conhece a forma do modelo gerador dos dados, a utilização do modelo bem-especificado é sempre a melhor opção, pois obtemos estimativas mais confiáveis, com VRs mais próximos de zero. Na aplicação real, utilizamos um índice de democracia eleitoral como variável resposta e cinco índices socioeconômicos, ambientais e geográficos como covariáveis que explicam a média do modelo Bessel. Aplicamos os três modelos e analisamos as estimativas *a posteriori*. Entre as três propostas, o M_2 com três covariáveis apresentou os resultados mais interessantes. As variáveis selecionadas neste ajuste foram *"Índice de Poluentes do Ar por Ano"*, *"Índice de Tratamento de Resíduos"* e *"Prevalência do Sexo Feminino"*. A primeira apresentou uma relação negativa com a variável resposta, enquanto as duas últimas apresentaram uma relação positiva, o que está de acordo com o esperado. Além disso, observamos um padrão de decrescimento nos valores do efeito temporal a partir de 2013, o que pode estar relacionado a instabilidades geopolíticas ocorridas nesse período.

Palavras-chave: Modelos Lineares Generalizados. Regressão Bessel. Índice de Democracia. Modelo Espaço-Temporal.

Abstract

This work contributed to a Bayesian implementation of the Bessel regression model for limited data with a spatiotemporal correlation structure and provided comprehensive comparisons among different models. We introduced three models - M_1, M_2, and M_3 - primarily differing in the inclusion of covariates and temporal effects. M_1 can be considered a simplification of M_2, where we summarize covariates with measurements over time (M_2 uses all measurements from distinct time points). Both M_1 and M_2 incorporate random effects (spatial and temporal) additively in the linear predictor explaining the mean. On the other hand, in M_3, only the spatial effect is additive, and the temporal effect is introduced through time-varying coefficients. We conducted simulations in various misspecification scenarios of the effects and compared the Relative Biases (RBs) of the estimates of the main parameters. In cases where there is no information about the data structure, M_2 proved to be the most suitable as it showed adjustments with smaller RBs in misspecification situations. However, when the analyst knows the form of the data-generating model, using the well-specified model is always the best option, as we obtain more reliable estimates with RBs closer to zero. In the real-world application, we used an electoral democracy index as the response variable and five socioeconomic, environmental, and geographical indices as covariates explaining the mean of the Bessel model. We applied all three models and analyzed the posterior estimates. Among the three proposals, M_2 with three covariates yielded the most interesting results. The variables selected in this adjustment were "Annual Air Pollution Index", "Waste Treatment Index" and "Female Prevalence". The first exhibited a negative relationship with the response variable, while the latter two showed a positive relationship, which aligns with expectations. Additionally, we observed a decreasing pattern in the values of the temporal effect from 2013 onwards, which may be related to geopolitical instabilities that occurred during that period.

Keywords: Generalized Linear Models. Bessel Regression. Democracy Index. Spatial-Temporal Model.

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

O modelo de regressão beta, introduzido em Ferrari e Cribari-Neto (2004), é o modelo dominante para analisar dados limitados. Sua predominância é reforçada pela facilidade de reparametrização do modelo em função da média e da dispersão e pela estimação via algoritmo de maximização de expectativa (EM) (Barreto-Souza e Simas, 2017). O trabalho de Simas et al. (2010) traz uma possibilidade interessante de inserir variáveis explicativas para avaliar o impacto sobre a média e também sobre a dispersão.

Desde a introdução da extensão multivariada da distribuição beta por Ferguson (1973) - também conhecida como distribuição de Dirichlet -, foram propostos diversas alternativas para a regressão beta, como a regressão Kumaraswamy (Kumaraswamy, 1980) e o modelo simplex (Barndorff-Nielsen e Jørgensen, 1991), contudo, essas opções perdem características importantes quando comparados ao beta, como uma representação estocástica simples, parametrização média-dispersão e estimação de parâmetros via algoritmo EM.

Mais recentemente, diversas propostas para regressão de resposta limitada foram introduzidas na literatura. Alguns exemplos que modelam a média incluem a distribuição log-Lindley introduzida em Gómez-Déniz et al. (2014), a distribuição Lindley-unitária, utilizada para modelagem de proporções em Mazucheli et al. (2019), a distribuição inversa-Gaussiana-unitária de Ghitany et al. (2019) e a distribuição log-Bilal apresentada em Altun et al. (2021). A distribuição beta retangular apresentada em Bayes et al. (2012) introduz covariáveis tanto para a média quanto para a precisão, enquanto a nova classe de distribuições Johnson SB generalizadas (GJS) apresentada em Lemonte e Bazán (2016) e a distribuição *power logit* de Queiroz e Ferrari (2023) modelam a mediana e a dispersão. Além dessas, aplicações da distribuição Birnbaum-Saunders-unitária, que modela os quantis, são expostas em Mazucheli et al. (2018).

Esta dissertação é focada na regressão Bessel, que foi introduzida em Barreto-Souza et al. (2021) como mais uma alternativa ao modelo beta no tratamento de variáveis limitadas. Segundo os autores, essa é uma distribuição particularmente atraente na

modelagem desse tipo de dado, devido a facilidade de reparametrização para média-dispersão e representação estocástica simples, além disso demonstra maior robustez sob má especificação, quando comparada à beta. O objetivo dessa dissertação é fazer uma extensão do modelo Bessel, aproveitando todas as suas características e vantagens para modelar uma variável limitada.

A distribuição Bessel tem em seu núcleo uma função Bessel do segundo tipo de ordem 1, o que foi um elemento determinante para a escolha do nome dessa distribuição. Similarmente, a função beta está definida na construção da distribuição beta. A construção da distribuição Bessel é definida a partir de uma representação estocástica envolvendo duas variáveis aleatórias com distribuição inversa-Gaussiana. Se Y_1 tem distribuição inversa-Gaussiana com parâmetro de escala 1 e de forma $\alpha > 0$ e Y_2 tem distribuição inversa-Gaussiana com parâmetro de escala 1 e de forma $\beta > 0$, podemos escrever

$$Z \stackrel{d}{=} \frac{Y_1}{Y_1 + Y_2}, \tag{1.1}$$

em que Z será uma variável aleatória com distribuição Bessel, tendo parâmetros α e β. Essa representação estocástica simples é análoga à obtida para a distribuição beta. No caso da beta, a distribuição de Y_1 e Y_2 seriam gamas com escala 1 e forma α e β. Assim, a regressão Bessel retém todas as principais propriedades da regressão beta que são determinadas pela construção estocástica. Além disso, a distribuição Bessel se mostra mais robusta que a beta em termos de má especificação do modelo e em aplicações reais (Barreto-Souza et al., 2021). Esses pontos tornam a Bessel uma alternativa atraente para o tratamento de dados limitados.

Este trabalho tem como objetivo mostrar o desempenho do modelo de regressão Bessel em problemas com respostas e covariáveis correlacionadas no espaço e no tempo. Este tipo de análise envolvendo estrutura espaço-temporal ainda não foi explorado na literatura sobre regressão Bessel. Os autores em Barreto-Souza et al. (2021) focaram no tratamento clássico da regressão Bessel com a implementação de estimação por algoritmo EM. A literatura de Modelos Lineares Generalizados indica que é analiticamente e computacionalmente difícil tratar, sob o ponto de vista frequentista, a estimação de parâmetros em uma abordagem espaço-temporal. Problemas de natureza espacial foram abordados em Kalhori e Mohhamadzadeh (2017) utilizando o modelo de regressão beta, porém assumindo o paradigma Bayesiano para estimar. Aqui, utilizaremos modelos Bayesianos com distribuições *a priori* pouco informativas ajustados via `RStan`, que utiliza um método Monte-Carlo Hamiltoniano (HMC) chamado de NUTS (No-U-Turn Sampling, Hoffman et al. (2014)) para realizar a amostragem da distribuição *a posteriori*.

O HMC é uma instância do algoritmo Metrópolis-Hastings (MH) com uma evolução do processo segundo a dinâmica Hamiltoniana (Carpenter et al., 2017). A simulação usa um integrador reversível e que preserva o volume, chamado de integrador *leapfrog*, para propor a mudança para um novo espaço de estados. Essa forma mais complexa de amostragem tem uma grande vantagem em relação ao MH convencional, já que os saltos no espaço paramétrico podem ser distantes do valor atual da cadeia, ao contrário do MH, tornando o HMC mais eficiente que as alternativas. O algoritmo irá funcionar com uma especificação do tamanho e do número de saltos, porém esta escolha pode trazer problemas se for feita de forma inadequada. Basicamente, o tamanho e o número de saltos não podem ser grandes, pois isso acarretaria ao amostrador que retornasse ao ponto gerado na iteração anterior. A solução deste problema veio por meio do NUTS que impede esse "retorno à origem"de acontecer.

Os métodos Bayesianos, que formam o arcabouço teórico deste trabalho, utilizam o teorema de Bayes - em formato simples $P(\theta|\text{Dados}) \propto P(\text{Dados}|\theta)\ P(\theta)$ - para calcular a probabilidade de um evento condicionado aos dados e utilizando informações que expressam o conhecimento prévio ou incerteza do analista sobre o parâmetro alvo na forma de uma distribuição *a priori* $P(\theta)$ (McElreath, 2020).

Dentro do paradigma Bayesiano serão utilizadas técnicas de estatística espacial para analisar dados com correlação geográfica e ao longo do tempo. O modelo condicional autoregressivo (CAR) foi introduzido por Besag (1974) para o tratamento de dados de área com correlação espacial. No contexto de séries temporais (processo de uma única dimensão), o CAR pode ser escrito como uma representação do modelo AR(1), ou seja, autoregressivo de ordem 1. Esse fato será útil, já que utilizaremos a mesma especificação para o efeito aleatório espacial e para o temporal, mudando apenas a forma da matriz de covariâncias em cada caso (espacial e temporal).

O foco de uma regressão Bessel ou beta é explicar uma variável resposta limitada. Dados limitados são encontrados na forma de dados composicionais, índices, proporções, taxas, entre outros. Índices, em particular, tem a propriedade de resumir informações de diversas variáveis em um único número e são muito importantes para quantificar informações qualitativas, principalmente quando tratamos de dados socio-econômicos (Hauser e Warren, 1997) ou não mensuráveis, como ocorrem em pesquisas da área de psicologia (Liu, 1974).

O trabalho desenvolvido nesta dissertação de mestrado tem como motivação uma aplicação real relacionada ao banco de dados do projeto *Varieties of Democracy* (*V-Dem*) (Coppedge et al., 2022) (Hegedüs, 2020). Este projeto desenvolve uma nova metodologia para medir a qualidade da democracia e seus aspectos associados em mais de 200 países.

Os dados se estendem desde os anos 1900, com base em informações históricas, até a atualidade. Os índices de democracia calculados pelo projeto se mostram um excelente objeto de estudo para a aplicação do modelo de regressão Bessel espaço-temporal. Esta conclusão foi tirada a partir de um teste avaliativo do pacote `bbreg` (Barreto-Souza et al., 2021) que compara as regressões Bessel e beta. Os detalhes sobre os elementos deste teste serão dados mais adiante.

O banco de dados V-Dem é composto por mais de 470 indicadores, 82 índices intermediários e 5 índices chamados de "alto-nível" (que são variáveis limitadas). Neste trabalho, utilizaremos o índice de alto-nível *Democracia Eleitoral* como a variável resposta das modelagens a serem exploradas. Em termos de covariáveis, serão utilizadas aquelas relacionadas a dados socioeconômicos e geográficos de cada país e em cada ano disponível. Alguns exemplos destes tipos de informações são: a população, área, índice de desenvolvimento humano, produto interno bruto, emissões de carbono, taxas de natalidade e mortalidade, taxas de desemprego, índice de escolaridade, índice de desenvolvimento industrial, entre outros. Inspirados pela configuração e pela magnitude das variáveis disponíveis neste banco de dados motivador, o procedimento de geração de dados artificiais (para estudos simulados) será especificado; detalhes serão dados adiante no texto.

As contribuições do presente trabalho para a área de estatística são as seguintes:

- Análise inédita do modelo de regressão Bessel com estrutura espaço-temporal e baseada na inferência Bayesiana.
- Avaliação de diferentes formas de inserir a estrutura temporal, que poderá ser por meio de efeito aleatório aditivo ou com estrutura de correlação nos coeficientes do modelo.
- Análise de robustez do modelo quanto à presença ou a ausência de efeito espaço-temporal.
- Desenvolvimento de uma aplicação real com os dados *V-Dem* nunca explorada via regressão Bessel.

Organização da Dissertação

Esta dissertação de mestrado está organizada como segue. No Capítulo 2 trataremos dos principais conceitos metodológicos utilizados. Dando foco à definição teórica da distribuição Bessel e especificação do modelo de regressão Bessel, assim como descrito em Barreto-Souza et al. (2021), especificação do modelo espacial autorregressivo (CAR) e dos modelos de regressão espaço-temporais estudados, com a apresentação das distribuições *a priori*.

No Capítulo 3 apresentamos os resultados dos estudos de simulação. Especificamos os esquemas de simulação de dados utilizados, as especificações *a priori* utilizadas na estimação dos parâmetros, descrevemos em detalhes os modelos geradores de dados e as respectivas estimativas obtidas. Além disso, utilizamos gráficos de intervalos HPD e o Vício Relativo para avaliar a adequação do ajuste obtido.

A aplicação do modelo em dados reais é feita no Capítulo 4. Ajustamos os modelos apresentados anteriormente a dados extraídos de dois bancos, *V-Dem* e *EPI*, que fornecem, respectivamente, a variável resposta, Índice de Democracia Eleitoral, e 5 covariáveis ambientais e socioeconômicas utilizadas para explicar a média do modelo Bessel. A seleção do modelo mais apropriado para estes dados foi feita a partir da análise dos intervalos HPD e da coerência das estimativas obtidas.

No último capítulo, Capítulo 5, expomos as conclusões obtidas e fazemos um resumo dos principais pontos do trabalho, assim como uma breve perspectiva sobre possíveis trabalhos futuros a serem desenvolvidos.

Capítulo 2

Metodologia

O propósito deste capítulo é apresentar alguns conceitos básicos que serão necessários para entender os modelos de regressão e a inferência relacionada a eles no contexto desta dissertação. É importante salientar que o paradigma Bayesiano será usado para estimação, sendo assim, a função de verossimilhança do modelo proposto e as distribuições *a priori* dos parâmetros envolvidos deverão ser especificadas. Na construção deste texto, iniciaremos com a apresentação da distribuição Bessel, que é elemento chave para definir posteriormente a regressão. Em seguida, discutiremos alguns aspectos relevantes da modelagem condicional autoregressiva (CAR) para dados de área, que irá especificar a estrutura de dependência espaço-temporal presente nos dados deste estudo. Finalmente, iremos escrever a especificação completa da regressão Bessel espaço-temporal Bayesiana.

2.1 Distribuição Bessel

A representação da equação (1.1) foi utilizada em Barreto-Souza et al. (2021) para a construção da distribuição Bessel a partir de duas distribuições inversas-Gaussianas com parâmetros de escala 1 e parâmetros de forma maiores que zero. A definição formal da distribuição Inversa-Gaussiana, necessária para entender sua parametrização, é apresentada a seguir.

Definição 1. *Seja Y uma variável aleatória (v.a.) que segue a distribuição inversa-Gaussiana, denotamos $Y \sim IG(\alpha)$. Sua função densidade de probabilidade é dada por:*

$$h(y) = \frac{\alpha}{\sqrt{2\pi}} y^{-3/2} exp\left\{-\frac{1}{2}\left(\frac{\alpha^2}{y} + y\right) + \alpha\right\}, \quad para \quad y > 0.$$

Na representação (1.1), sejam Y_1 e Y_2 v.a.'s independentes, cada uma com distribuição inversa-Gaussiana tendo parâmetros de escala 1 e parâmetros de forma $\alpha > 0$ e

$\beta > 0$, respectivamente. Como consequência desta suposição, a variável Z terá distribuição Bessel univariada. A definição formal desta distribuição é dada a seguir.

Definição 2. *Se Z tem distribuição Bessel, sua densidade é dada por:*

$$f(z) = \frac{\alpha\beta e^{\alpha+\beta}}{\pi z(1-z)}(\alpha^2 z + (1-z)\beta^2)^{-1/2} K_1\left(\sqrt{\frac{\alpha^2}{1-z} + \frac{\beta^2}{z}}\right), \quad z \in (0,1),$$

em que $K_1(\cdot)$ é a função Bessel modificada do terceiro tipo e de ordem 1. Nesta parametrização, escrevemos $Z \sim Bessel(\alpha, \beta)$.

A presença da função $K_1(\cdot)$ motivou os autores em Barreto-Souza et al. (2021) a adotarem o nome "Bessel"para esta distribuição. Outro nome relacionado a este modelo é inversa-Gaussiana normalizada (N-IG) univariada. Em termos de média e variância, temos as seguintes expressões: $E(Z) = \mu = \alpha/(\alpha+\beta)$ e $Var(Z) = \mu(1-\mu)(1-\phi+\phi^2 e^{\phi} E_i(\phi))/2$. Note que $E_i(\phi)$ é a função exponencial integral - definida em Erdélyi (1953) e Abramowitz e Stegun (1968) - avaliada em ϕ e é dada por $E_i(\phi) = \int_1^{\infty} u^{-1} e^{-\phi u} du$. Essa função já está implementada no `R` pelo pacote `expint` (Goulet, 2016).

Uma das grandes qualidades da distribuição beta é a facilidade de reparametrização da função de densidade em termos da média e da dispersão, muito útil na construção de uma regressão pela facilidade de interpretação dos parâmetros. Assim como a beta, a função Bessel também é facilmente reparametrizável considerando explicitamente a média e a dispersão. Neste caso, escrevemos $Z \sim Bessel(\mu, \phi)$ e a densidade associada é

$$f(z) = \frac{\mu(1-\mu)\phi e^{\phi}}{\pi(z(1-z))^{3/2}} \frac{K_1(\phi\zeta_\mu(z))}{\zeta_\mu(z)}, \ z \in (0,1), \tag{2.1}$$

sendo $\zeta_\mu(z) = \sqrt{1 + \frac{(z-\mu)^2}{z(1-z)}}$, para $z \in (0,1)$.

Até o momento nós apenas definimos a distribuição Bessel e suas parametrizações. Na sequência, iremos explicar como a estrutura de regressão é incorporada neste tipo de modelo linear parametrizado por μ e ϕ.

Regressão Bessel

A construção desta modelagem é similar ao que é feito em Modelos Lineares Generalizados (Ravishanker et al., 2021). Definimos o modelo de regressão Bessel utilizando a função de ligação para a média (logit). O objetivo da função de ligação é estabelecer a conexão entre as covariáveis de interesse e o parâmetro alvo.

Definição 3. *Seja $Z_1, Z_2, ..., Z_n$ uma amostra aleatória tal que $Z_i \sim Bessel(\mu_i, \phi)$. Neste caso, adote que:*

$$logit\ \mu_i = \log(\frac{\mu_i}{1-\mu_i}) = \mathbf{X}_i^\top \boldsymbol{\kappa}, \tag{2.2}$$

sendo $\boldsymbol{\kappa} = (\kappa_1, ..., \kappa_p)^\top \in \mathbb{R}^p$ vetor de coeficientes desconhecidos. Ainda em termos de notação, assuma que $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ip})^\top$ são observações de p covariáveis conhecidas. Finalmente, adote que $\mathbf{X}$ representa uma matriz $n \times p$ cujo (i,j)-ésimo elemento é x_{ij}.

A log-verossimilhança do modelo, em função de μ_i e ϕ, é apresentada a seguir.

Definição 4. *Seja $\boldsymbol{\theta} = (\boldsymbol{\kappa}^\top, \phi)$ o vetor de parâmetros. A log-verossimilhança do modelo Bessel é dada por*

$$l(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \{\log \mu_i + \log(1-\mu_i) + \log \phi + \phi - \log \zeta_{\mu_i}(z_i) + \log K_1(\phi \zeta_{\mu_i}(z_i))\},$$

em que z_i é o valor observado de Z_i, para $i = 1, ..., n$.

Note que na expressão da log-verossimilhança, os coeficientes de regressão estão embutidos nos termos μ_i por meio da função de ligação em (2.2). Além disso, veja que o modelo apresentado até aqui é uma abordagem com efeitos fixos. A modelagem principal apresentada neste trabalho envolverá uma estrutura espaço-temporal que será estabelecida pelo modelo CAR, que é muito aplicado em Estatística Espacial para dados de área. A seguir, iremos apresentar alguns detalhes sobre a modelagem CAR.

2.2 Modelo CAR

O modelo autorregressivo condicional - CAR (Besag, 1974; Banerjee et al., 2003; Cressie, 2015) - é uma extensão do modelo autorregressivo (AR). O AR é comumente utilizado para explorar dados de séries temporais, enquanto o CAR proporciona uma estrutura que serve para estudar dados coletados em um espaço com divisão não homogênea e fixa de vizinhança com possibilidade de diferentes quantidades de vizinhos por área.

Considere uma região particionada em n sub-regiões indexadas por $1, 2, ..., n$. Essa coleção de sítios possui um sistema de vizinhança, $\{\partial_i : i = 1, ..., n\}$, em que ∂_i é a coleção de sub-regiões que são ditas vizinhas da região i. O sistema de vizinhanças satisfaz:

$$\forall\ i, j = 1, ..., n, \text{ temos que } \quad j \in \partial_i \iff i \in \partial_j \text{ e } i \notin \partial_i.$$

Podemos escrever o modelo CAR em termos de uma matriz binária de vizinhança W, em que $W_{ij} = 1$, se as regiões i e j são vizinhas e $W_{ij} = 0$, caso contrário. Escrevemos a matriz de covariâncias como $\mathbf{\Sigma} = (D_W - \rho W)^{-1}$, em que $D_W = \text{diag}\{w_{1+}, w_{2+}, \cdots, w_{n+}\}$ é uma matriz diagonal contendo o número de vizinhos de cada região (w_{i+} é o número de vizinhos da região i). Finalmente, ρ é um escalar incluído para garantir que $\mathbf{\Sigma}$ seja inversível.

A estrutura do CAR é usualmente definida pela distribuição Normal Multivariada. Importante ressaltar que o caso Gaussiano não é o único tipo de aplicação, sendo que a abordagem também pode ser definida para outras distribuições da família exponencial (Banerjee et al., 2003). Neste trabalho, iremos assumir o modelo CAR na versão Gaussiana apenas. Desta forma, considere que $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_L)^\top$ é um vetor contendo variáveis aleatórias coletadas em diferentes regiões de um espaço. O modelo CAR Gaussiano é estabelecido assumindo que $\boldsymbol{\xi} \sim N_L(\boldsymbol{\mu}, \tau\Sigma)$, sendo $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_L)^\top$ o vetor de médias, $\mathbf{\Sigma}$ a matriz $L \times L$ de covariâncias, conforme definida anteriormente nesta seção, e τ é um parâmetro que controla a magnitude da variabilidade.

Ressalta-se, ainda, que o modelo CAR também pode ser definido com $\rho = 1$, porém esta especificação gera uma versão em que a distribuição conjunta não está bem definida, visto que a matriz $\mathbf{\Sigma}$ não teria inversa e isso não permite concluir que haverá uma Normal Multivariada para $\boldsymbol{\xi}$. Adote que ξ_{-i} é o vetor ξ sem a variável ξ_i. O modelo sempre fornece as distribuições condicionais de $\xi_i|\xi_{-i}$, porém a versão com $\rho = 1$ determinaria apenas o núcleo da conjunta, sem a constante normalizadora que estabelece uma distribuição de probabilidade própria. Não há prejuízo em assumir a versão imprópria, porém a opção própria (com $\rho \neq 1$) permite uma implementação da conjunta no programa `Stan`, o que é atrativo do ponto de vista computacional para este trabalho.

O parâmetro ρ deve ser escolhido em um intervalo que envolve o menor e o maior autovalor de uma matriz diretamente relacionada a $\mathbf{\Sigma}$. O limite inferior deste intervalo é negativo e o limite superior é positivo. Entretanto, sob o ponto de vista prático, valores negativos de ρ, apesar de possíveis, não proporcionam uma interpretação razoável para o problema espacial. Temos que $\xi_i|\xi_{-i} \sim N(\rho \sum_{j \in \partial_i}^{n} \xi_j / w_{i+}, \tau / w_{i+})$, em que $w_{i+} = \sum_j w_{ij}$. Perceba que a média da distribuição condicional é a média dos vizinhos multiplicada por ρ. Se $\rho < 0$, estaríamos dizendo que a observação na unidade i é negativamente proporcional à média dos seus vizinhos. Esta ideia não faz sentido em muitos problemas práticos, incluindo aqueles investigados nesta dissertação. Esta linha de pensamento nos leva a escolher $\rho > 0$ e diferente de 1. A escolha de $\rho \in (0, 1)$ é bastante usada na literatura e será adotada aqui. Existem trabalhos Bayesianos (Areal et al., 2012; Krisztin e Piribauer, 2021) que estimam ρ especificando uma distribuição *a priori* (e.g.

a distribuição beta). Importante ressaltar que o parâmetro ρ não é fácil de estimar. A inferência neste caso pode estar relacionada com uma grande incerteza no intervalo $(0, 1)$. Além disso, valores de ρ pouco acima de 0.5 podem ser obtidos no caso de dados sem dependência espacial. O modelo CAR é bastante sensível ao valor de ρ. Valores baixos determinam uma modelagem em que a parte espacial importa pouco. O ideal é assumir ρ próximo de 1 para garantir força da estrutura espacial, principalmente na situação em que o pesquisador deseja impor o efeito espacial no modelo. Neste trabalho, não iremos colocar incerteza sobre ρ. Este parâmetro será fixado (perto de 1) para impor a presença do efeito espacial. Dessa forma, estamos supondo um contexto em que o pesquisador tem certeza sobre a presença da dependência espacial.

2.3 Modelo de Regressão Bessel Espaço-Temporal

A modelagem de regressão, a ser estudada aqui, assume que a variável resposta, $\boldsymbol{Z}$, varia no tempo. As covariáveis em $\boldsymbol{X}$ podem ser observadas para um único instante de tempo ou podem ser registradas em tempos diferentes. Além disso, as unidades amostrais são coletadas em um espaço com estrutura de vizinhança conhecida. A possibilidade de termos uma aplicação com covariáveis variando ou não no tempo motiva a construção de dois tipos de modelos considerando a estrutura temporal.

Nesta seção iremos apresentar duas versões de uma modelagem espaço-temporal para a regressão Bessel. A primeira versão é considerada principal em nosso estudo. Ela estabelece a estrutura espacial e temporal por meio de dois efeitos aleatórios independentes. A segunda versão também possui um efeito aleatório espacial, porém ela estabelece a dependência temporal nos coeficientes de regressão que variam no tempo; esta opção só poderá ser usada na situação em que as covariáveis são observadas em tempos diferentes. Iniciamos a seção com a apresentação da estrutura dos dados e seguimos para a definição dos modelos.

Estrutura dos Dados

Os dados serão compostos por uma variável resposta, Z, e p covariáveis explicativas para a média. Z é medida para cada ponto no espaço e em cada tempo, assim, denotamos uma medida dessa variável no sítio l e tempo t por Z_{lt}. Assuma, também, que o banco de dados possua p_1 covariáveis observadas para cada local e tempo, que

serão usadas para explicar μ_{lt}. Adote que $\boldsymbol{X}^{*}_{\bullet lt} = (X^{*}_{1lt}, X^{*}_{2lt}, \cdots, X^{*}_{p_1 lt})^{\top}$ é o vetor dessas covariáveis para o local l e tempo t. Admita que a base de dados possua p_2 covariáveis que são medidas para cada local l, mas não variam no tempo. Neste sentido escreva que $\boldsymbol{X}_{\bullet l} = (X_{1l}, X_{2l}, \cdots, X_{p_2 l})^{\top}$ é o vetor de covariáveis não-temporais para o local l.

Modelo Espaço-Temporal Principal

Dentro do contexto Bayesiano, a introdução de efeitos aleatórios em modelos de regressão é feita de forma trivial por meio da função de ligação. Sejam $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_L)^{\top}$ e $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_T)^{\top}$ dois vetores de efeitos aleatórios espacial e temporal, respectivamente.

A partir da Definição (3), assuma:

$$Z_{lt} \sim Bessel(\mu_{lt}, \phi) \quad \text{para} \quad l \in \{1, \cdots, L\} \quad \text{e} \quad t \in \{1, \cdots, T\}. \tag{2.3}$$

O parâmetro de dispersão, ϕ, terá distribuição *a priori* da forma:

$$\phi \sim Ga(a_\phi, b_\phi), \tag{2.4}$$

enquanto as covariáveis e os efeitos aleatórios explicam a média como segue:

$$\text{logit}\ \mu_{lt} = \kappa_0 + \boldsymbol{X}^{*\top}_{\bullet lt}\boldsymbol{\kappa}^{*}_{\bullet t} + \boldsymbol{X}^{\top}_{\bullet l}\boldsymbol{\kappa}_{\bullet} + \gamma_l + \delta_t, \tag{2.5}$$

em que $\boldsymbol{\kappa}^{*}_{\bullet t} = (\kappa^{*}_{1t}, \kappa^{*}_{2t}, \cdots, \kappa^{*}_{p_1 t})^{\top}$, $\boldsymbol{\kappa}_{\bullet} = (\kappa_1, \kappa_2, \cdots, \kappa_{p_2})^{\top}$. Assuma que $\boldsymbol{\kappa} = (\boldsymbol{\kappa}^{*}_{\bullet t}, \boldsymbol{\kappa}_{\bullet})$.

O modelo Bayesiano hierárquico terá a seguinte distribuição *a priori* para os coeficientes de regressão:

$$\boldsymbol{\kappa} \sim N_p[\boldsymbol{m}_{\boldsymbol{\kappa}}, v_\kappa \boldsymbol{I}_{\boldsymbol{p}}], \tag{2.6}$$

sendo $\boldsymbol{\kappa} = (\kappa_0, \boldsymbol{\kappa}^{*}_{\bullet t}, \boldsymbol{\kappa}_{\bullet})$ um vetor de tamanho p, $\boldsymbol{I}_{\boldsymbol{p}}$ a matriz identidade de tamanho $p \times p$ e $p = 1 + p_1\ t + p_2$ um escalar. Note que estamos assumindo independência *a priori* entre os coeficientes, incluindo aqueles que são referentes a tempos distintos, ou seja, $\boldsymbol{\kappa}^{*}_{t}$'s.

Destacamos que o modelo não está inserindo uma estrutura de dependência temporal "duplicada" pela presença de uma associação entre os efeitos δ_t's e a presença de uma associação entre os coeficientes que variam no tempo. Consideramos que assumir para cada regressor variante no tempo um único coeficiente para todo t determinaria um modelo pouco flexível que assume um impacto constante desse tipo de regressor ao longo tempo. Perceba que estamos dando liberdade ao modelo para estimar impactos distintos, porém a associação temporal será feita por meio da distribuição dos δ_t's.

Conforme estabelecido na Seção 2.2, vamos utilizar a modelagem CAR para o efeito espacial $\boldsymbol{\gamma}$ e o AR para o efeito temporal $\boldsymbol{\delta}$. No contexto temporal, iremos escrever

o AR em uma versão estabelecida pelo CAR, sendo que a matriz W apresenta $w_{ij} = 1$ para $|i - j| = 1$ e $w_{ij} = 0$, caso contrário. Esta relação entre o AR e o CAR é também usada em outros trabalhos como em Mayrink e Gamerman (2009). Tal opção de apresentação da parte temporal do modelo visa manter a consistência notacional em relação ao que é feito para $\boldsymbol{\gamma}$.

Levando em conta a estrutura do CAR Gaussiano, os efeitos aleatórios espacial e temporal serão assumidos com a seguinte configuração *a priori*:

$$\begin{aligned} \boldsymbol{\gamma} &\sim N_L[\mathbf{0_L}, \tau_\gamma(D_{W_\gamma} - \rho W_\gamma)^{-1}], \\ \boldsymbol{\delta} &\sim N_T[\mathbf{0_T}, \tau_\delta(D_{W_\delta} - \rho W_\delta)^{-1}]. \end{aligned} \tag{2.7}$$

Assuma que $\mathbf{0_L}$ é um vetor de zeros de tamanho L. Finalmente, a construção do modelo hierárquico é concluída com a especificação das distribuições *a priori* dos parâmetros de variabilidade a seguir:

$$\begin{aligned} \tau_\gamma &\sim Ga(a_\gamma, b_\gamma), \\ \tau_\delta &\sim Ga(a_\delta, b_\delta). \end{aligned} \tag{2.8}$$

A apresentação completa da modelagem espaço-temporal principal está finalizada neste ponto. Destacamos que poderemos explorar versões deste modelo em que $\boldsymbol{\gamma}$ e/ou $\boldsymbol{\delta}$ são excluídos. Nestes casos de exclusão, simplesmente ignore as distribuições *a priori* indicadas aqui. A seguir, iremos apresentar a segunda versão do modelo na qual assumimos coeficientes com associação temporal.

Modelo Espaço-Temporal com Efeito Aleatório nos Coeficientes da Regressão

A diferença fundamental deste modelo para o principal é a ausência dos efeitos aleatórios temporais aditivos na equação que relaciona a média com as covariáveis. Nessa versão, temos que $Z_{lt} \sim \text{Bessel}(\mu_{lt}, \phi)$, como anteriormente, mas agora as covariáveis e os efeitos aleatórios explicam a média através de uma nova relação. Sendo assim, escrevemos:

$$\text{logit } \mu_{lt} = \kappa_0 + \boldsymbol{X}_{\bullet lt}^{*\top} \boldsymbol{\kappa}_{\bullet t}^{*} + \boldsymbol{X}_{\bullet l}^{\top} \boldsymbol{\kappa}_{\bullet} + \gamma_l, \tag{2.9}$$

novamente escrevemos $\boldsymbol{\kappa}_{\bullet t}^{*} = (\kappa_{1t}^{*}, \kappa_{2t}^{*}, \cdots, \kappa_{p_1 t}^{*})^\top$, $\boldsymbol{\kappa}_{\bullet} = (\kappa_1, \kappa_2, \cdots, \kappa_{p_2})^\top$, para os coeficientes da média. Perceba que o efeito aleatório δ_t foi removido nesta representação em comparação com o modelo principal descrito anteriormente.

A inclusão da dependência temporal será realizada pela especificação *a priori* baseada na normal multivariada. Os coeficientes das covariáveis coletadas ao longo do tempo têm a estrutura temporal dada pelo modelo CAR como segue:

$$\boldsymbol{\kappa}_{j\bullet}^{*} = (\kappa_{j1}^{*}, \kappa_{j2}^{*}, \cdots, \kappa_{jT}^{*})^\top \sim N_T[\mathbf{0_T}, \tau_\kappa(D_{W_\kappa} - \rho W_\kappa)^{-1}], \quad \text{para} \quad j = 1, 2, \cdots, p_1. \tag{2.10}$$

Os demais coeficientes, aqueles que não variam no tempo, são modelados por especificações *a priori* independentes. Assuma que $\kappa_j \sim N[m_\kappa, v_\kappa]$ para $j = 1, 2, \cdots, p_2$.

As distribuições *a priori* para o efeito aleatório espacial é idêntica àquela apresentada na Equação (2.7). Por fim, novamente completamos a construção do modelo com as distribuições *a priori* dos parâmetros de variabilidade que aparecem em (2.7) e (2.10). Considere:

$$\begin{aligned} \tau_\gamma &\sim Ga(a_\gamma, b_\gamma), \\ \tau_\kappa &\sim Ga(a_\kappa, b_\kappa). \end{aligned} \tag{2.11}$$

Concluímos aqui a apresentação do modelo com estrutura temporal para os coeficientes de regressão. Podemos perceber que este modelo admite mais de uma covariável com medidas para diferentes tempos. No entanto, no próximo capítulo, vamos explorar apenas o caso mais simples com $\boldsymbol{X}^*_{lt} = X^*_{1lt}$, ou seja, apenas uma covariável tem caráter temporal. O estudo mostrado na sequência deste trabalho é baseado em dados artificiais, portanto, vamos definir o esquema de simulação dos dados e apresentar os resultados comparativos dos ajustes dos modelos descritos aqui.

Capítulo 3

Resultados dos Estudos de Simulação

O propósito deste capítulo é apresentar um estudo de avaliação dos modelos descritos no capítulo anterior levando em conta dados artificiais. Trabalhar com dados artificiais é importante para uma avaliação de desempenho da modelagem, pois simular dados permite a vantagem de sabermos os valores reais dos parâmetros e, consequentemente, calcular o vício das estimativas. As seções deste capítulo irão explicar em um primeiro momento a maneira com que os dados são gerados. As análises de desempenho serão feitas em seguida para diferentes cenários de geração.

3.1 Esquema de Simulação de Dados

Para estudar as propriedades do modelo Bessel espaço-temporal fizemos estudos de simulação utilizando as distribuições *a priori* descritas anteriormente. O primeiro passo foi gerar as covariáveis. Iremos trabalhar com três tipos de covariáveis explicando μ_{lt}. Seja $\boldsymbol{X^*_{1l\bullet}} = (X^*_{1l1}, X^*_{1l2}, \cdots, X^*_{1lT})^\top$ um vetor contendo as repetições para cada tempo do valor do primeiro tipo de covariável, para o local $l = 1, 2, \cdots, L$. Cada entrada deste vetor é gerada da $U(-1, 1)$, para todo l, de forma independente. O segundo tipo de covariável X_{2l} é também obtido independentemente da $U(-1, 1)$, para cada local l, porém sem medições para diferentes tempos. Finalmente, o terceiro tipo de covariável X_{3l} é binário, sendo gerado independentemente da Bernoulli(0.5), para cada l e sem medições para diferentes tempos. Neste estudo, iremos fixar $T = 20$ e $L = 200$.

A geração dos dados sintéticos levará em conta três modelos geradores diferentes, os quais serão denominados M_1, M_2 e M_3. Os modelos geradores M_1 e M_2 seguem a es-

pecificação espaço-temporal principal, enquanto que o modelo M_3 considera a estrutura espaço-temporal com efeito temporal nos coeficientes de regressão. Em M_1 optamos por utilizar um modelo simplificado considerando o segundo tipo de covariável em substituição do primeiro tipo, ou seja, não há medidas ao longo do tempo. Sendo assim, foram geradas 2 covariáveis $U(-1, 1)$. Perceba que em uma situação prática, o modelo M_1 seria usado perante uma sumarização das covariáveis do primeiro tipo. Os demais tipos de covariáveis permanecem sem alteração. Por outro lado, o M_2 leva em conta os dados completos, ou seja, não aplicamos a sumarização de $\boldsymbol{X}^*_{\mathbf{1}l\bullet}$.

Foram gerados nove conjuntos de dados simulados a partir dos modelos geradores (Tabela 3.1). Os modelos completos são denotados como $M_1^{\oplus}$, $M_2^{\oplus}$ e $M_3^{\oplus}$, incluindo os efeitos aleatórios espacial e temporal na estrutura dos dados gerados. Por outro lado, os geradores $M_1^{\delta\ominus}$ e $M_2^{\delta\ominus}$ produzem dados apenas com o efeito aleatório espacial γ. Similarmente, os geradores $M_1^{\gamma\ominus}$, $M_2^{\gamma\ominus}$ e $M_3^{\gamma\ominus}$ geram dados com foco exclusivamente nos efeitos temporais, ou seja, são modelos sem o efeito espacial introduzido por γ para M_1, M_2 e M_3. Ainda temos o modelo $M_3^{CAR\ominus}$, no qual não há estrutura de correlação CAR na matriz de variâncias e covariâncias associada aos coeficientes do modelo, estes são gerados a partir de distribuições normais independentes e identicamente distribuídas com média 0 e variância 10. Essa abordagem permite uma análise abrangente dos efeitos individuais e combinados dos componentes espaciais e temporais. É possível explorar o ajuste de um modelo com um efeito para dados gerados sem esse efeito e vice-versa. A intenção desta estratégia é examinar as propriedades do modelo principal sob má especificação dos efeitos aleatórios de tempo e espaço.

Tabela 3.1: Cenários de geração de dados com e sem efeitos aleatórios.

Geração de dados	**com** γ	**sem** γ
com δ	$M_1^{\oplus}$, $M_2^{\oplus}$	$M_1^{\gamma\ominus}$, $M_2^{\gamma\ominus}$
sem δ	$M_1^{\delta\ominus}$, $M_2^{\delta\ominus}$	-
com CAR	$M_3^{\oplus}$	$M_3^{\gamma\ominus}$
sem CAR	$M_3^{CAR\ominus}$	-

Para gerar as matrizes de vizinhanças (W), utilizamos a função `band` do pacote `Matrix` (Bates et al., 2022) do `R` (R Core Team, 2023). Com esta função é possível criar matrizes binárias banda diagonais necessárias para estabelecer a estrutura de vizinhança do modelo CAR. Para esclarecer o que é chamado de matriz banda diagonal neste trabalho, considere os exemplos simplificados mostrados na Figura 3.1. A cor vermelha indica

as entradas da matriz sinalizadas com 1 (ou seja, presença de vizinhança). A cor branca indica as entradas sinalizadas com 0 (ausência de vizinhança). A estrutura de vizinhança temporal está representada no Painel (b), correspondendo ao que é estabelecido no processo AR(1). O Painel (a) indica que estamos adotando uma estrutura de vizinhança em que a maioria das regiões possui 4 vizinhos.

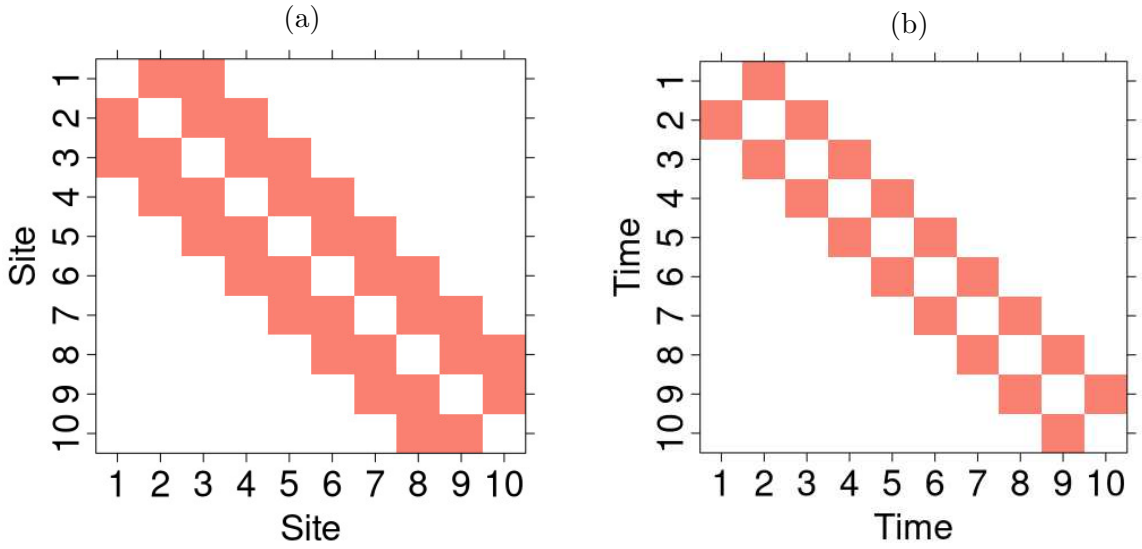


Figura 3.1: Exemplo de matrizes de vizinhança de tamanho 10 com 4 (a) e 1 (b) vizinhos. O painel (b) representa o modelo equivalente ao processo AR(1).

Os efeitos aleatórios γ e δ serão gerados de suas respectivas distribuições *a priori*, estabelecidas no capítulo anterior, com a estrutura de modelagem CAR e parâmetros $\tau_\gamma = \tau_\delta = \tau_\kappa = 2$ e $\rho = 0.9$. Neste estudo, iremos considerar a presença do efeito aleatório espacial e temporal (geradores M_1 e M_2) apenas na estrutura explicativa da média μ_{lt}. A possibilidade de estudo considerando o caso em que covariáveis explicativas estão associadas à dispersão dada por ϕ_{lt}, assim como o impacto de efeitos aleatórios em ϕ_{lt}, é indicado ao fim desta dissertação como um possível caminho futuro.

Para gerar a variável resposta Z, utilizamos a Equação (2.5), com os parâmetros de regressão $\boldsymbol{\kappa}_0 = 2$, o termo $\boldsymbol{X}^*_{\bullet lt}\boldsymbol{\kappa}^*_{\bullet t}$ é nulo e $\boldsymbol{\kappa}_\bullet = (1.2, -1.5, -1)$ para o modelo gerador M_1. A primeira covariável de M_1 é única, sem considerar medições para diferentes tempos, e gerada da uniforme conforme dito anteriormente. Para M_2 fazemos $\boldsymbol{\kappa}_0 = 2$, $\boldsymbol{\kappa}^*_{\bullet t} = (3.0, -2.1, -1.2, -3.0, -2.7, 1.8, -1.2, 2.4, 1.5, -2.7, 2.4, -2.4, -3.0, -1.8, -1.8, 2.7, -1.5, 1.5, 2.7, -1.2)$ e $\boldsymbol{\kappa}_\bullet = (-1.5, -1)$.

No gerador M_3 o efeito aleatório γ é obtido exatamente como foi descrito para M_1 e M_2. Lembre que neste gerador não existe o efeito δ. A estrutura de dependência

temporal será estabelecida em M_3 pela associação dos coeficientes de regressão. Adote a Equação (2.10) para gerar $\boldsymbol{\kappa}^*_{\mathbf{1}\bullet}$. Os demais coeficientes são $\kappa_0 = 2$, $\kappa_1 = -1.5$, $\kappa_2 = -1$.

3.2 Especificações *a priori*

Em termos de especificações *a priori* para descrever nossa incerteza inicial sobre os parâmetros de cada modelo, considere as escolhas estabelecidas aqui. Em geral, iremos adotar distribuições com variabilidade mais alta e que determinam baixo grau de certeza do pesquisador.

Nos modelos M_1 e M_2, assuma que $a_{\tau_\delta} = a_{\tau_\gamma} = 0.1$ e $b_{\tau_\delta} = b_{\tau_\gamma} = 0.1$ determinando $E(\tau_\delta^\mu) = E(\tau_\gamma^\mu) = 1$ e $Var(\tau_\delta) = Var(\tau_\gamma) = 10$ na distribuição gama indicada na Equação (2.8). Note que esta escolha tem variância 10, sugerindo um nível alto de variabilidade, o que configura uma distribuição vaga e de incerteza mais alta. Nas especificações da Equação (2.7) todos os componentes definidos nas distribuições normais multivariadas são fixos de acordo com a estrutura espaço-temporal. Conforme discutido anteriormente na Seção 2.2, iremos adotar $\rho = 0.9$ fixado em todos os modelos investigados. A incerteza *a priori* sobre os coeficientes de regressão será descrita pela especificações Gaussianas na Equação (2.6). Neste caso, faça $\boldsymbol{m}_{\boldsymbol{\kappa}} = \mathbf{0}_p$ e $v_\kappa = 10$. Novamente, perceba que a alta variabilidade é indicada aqui para induzir uma especificação vaga e de maior incerteza *a priori*.

No modelo M_3, considere $a_{\tau_\kappa} = 0.1$ e $b_{\tau_\kappa} = 0.1$, ou seja, novamente média 1 e variância 10 (*priori* vaga) na informação inicial para τ_κ. Assuma também que $\boldsymbol{m}_{\boldsymbol{\kappa}} = \mathbf{0}_{p_2}$ e $v_\kappa = 10$ nas distribuições *a priori* dos coeficientes não-variantes no tempo; veja o parágrafo logo abaixo da expressão (2.10). Todas as demais especificações *a priori* de parâmetros existentes em M_1, M_2 e M_3, serão exatamente as mesmas conforme descrito no parágrafo anterior.

3.3 Ajuste dos Modelos

Para cada banco de dados gerado foram ajustados o modelo completo - com a presença de efeito aleatório espacial (γ) e efeito aleatório temporal (δ no caso de M_1 e M_2 e κ para M_3) - e modelos sem cada um dos efeitos aleatórios. No segundo cenário, ora era ajustado o modelo com efeito temporal e sem efeito espacial, ora o contrário.

Estes modelos de regressão foram ajustados no programa `Stan` (Carpenter et al., 2017) pela interface com o `R` - `RStan` (Stan Development Team, 2023). Em termo de configuração do MCMC (*Markov Chain Monte Carlo)*, devido a dificuldades computacionais para executar o método considerando amostras grandes, realizamos apenas 50 replicações dos dados. Cada réplica foi gerada sob as mesmas condições - mesmas covariáveis e valores reais dos parâmetros - para cada cenário. Em cada réplica foram feitas 5 mil iterações totais do amostrador NUTS, descartando as 2500 primeiras iterações obtidas na fase de adaptação do algoritmo. A análise de convergência do método MCMC é feita pela inspeção visual da trajetória das cadeias (Apêndice C[1]).

Uma medida avaliativa que será empregada neste estudo para verificação de desempenho dos modelos é o Vício Relativo (VR) de cada parâmetro. No caso genérico, dizemos que $\text{VR}(\xi) = 100\ (\hat{\xi} - \xi_{\text{real}})/|\xi_{\text{real}}|$ é o VR do parâmetro ξ, sendo ξ_{real} o valor verdadeiro, escolhido na geração, e $\hat{\xi}$ é a estimativa *a posteriori*. A multiplicação por 100 é feita para expressar o resultado em termos de porcentagem. Em outras palavras, o VR indica o quanto a amplitude da diferença entre o valor real e o estimado representa em relação à magnitude absoluta do valor real. Note que o ideal é obter um VR próximo de zero. Valores positivos sugerem que o parâmetro ξ está sendo sobrestimado. Naturalmente, valores negativos indicam que ξ está sendo subestimado.

A seguir iremos iniciar as análises de cada modelo, começando pelo modelo M_2, que é tido como o modelo principal do estudo. Em seguida tratamos do modelo M_3, uma alternativa para evitar o confundimento do efeito temporal com o intercepto do modelo. Por último ajustamos o modelo M_1, que é uma versão simplificada do M_2, com o intuito de reduzir o número de parâmetros a serem estimados.

[1]Serão apresentados apenas os gráficos das cadeias para M_2, os demais cenários apresentaram padrões de convergência similares.

Modelo Gerador $M_2^{\oplus}$: Ajuste com δ e γ

Iniciaremos pela análise do modelo espaço-temporal completo e igual ao modelo gerador dos dados. Nesse contexto, ajustamos o M_2 considerando tanto o efeito aleatório espacial quanto o temporal em conjuntos de dados gerados a partir do $M_2^{\oplus}$.

Na Figura 3.2(a) temos os *boxplots* dos VRs do efeito aleatório temporal para a unidade de tempo definida nos dados. Não há grandes diferenças entre os VRs dos tempos sob avaliação. Todos os *boxplots* estão centrados em zero e apresentam o segundo e terceiro quartis entre -200% e 200%. Obtivemos um resultado semelhante para o efeito aleatório espacial, Painel (c). Contudo, no caso de γ, houveram alguns sítios cujos vícios relativos alcançaram valores da ordem de 10^4, mas, por se tratarem de *outliers*, não estão representados na Figura 3.2.

O Painel (b) mostra os *boxplots* dos coeficientes κ. Neste caso, coeficientes associados à variável medida no tempo são representados pelos Segmentos V_1 ao V_{20}. O coeficiente da variável uniforme com medida única é representado pelo Segmento V_{21}. Finalmente, o coeficiente da variável binária é indicado pelo Segmento V_{22} e o intercepto (κ_0) é relacionado a V_{23}. Aqui notamos VRs muito menores em relação aos VRs dos efeitos aleatórios, entre -20% e 20% para todos os coeficientes (exceto κ_0 entre -40% e 40%). Notavelmente, κ_0 possui uma dificuldade intrínseca para ser estimado devido à natureza aditiva dos efeitos aleatórios, o que pode acarretar confundimento entre estes e o intercepto do modelo. A variável binária também apresentou VR ligeiramente superior às demais, o que é esperado devido a sua natureza discreta assumindo apenas dois tipos de valores.

A existência de confundimento entre o intercepto e um efeito aleatório é comum em modelos mistos de regressão, como o proposto nesta dissertação. Este confundimento é expresso por uma sobrestimação do intercepto e a consequente subestimação do efeito aleatório. Alguns autores trabalhando nesta área preferem um ajuste sem intercepto, considerando que esse elemento será incorporado pelo efeito aleatório; ver, por exemplo, Neuhaus e McCulloch (2006), Schielzeth (2010), Neuhaus e McCulloch (2011) e Barr et al. (2013). Na presente dissertação, iremos incluir o intercepto e recomendar ao leitor que tenha cuidado com a possibilidade de haver maior vício na sua estimação. Este vício não prejudica a estimação dos demais coeficientes atrelados a covariáveis (principalmente se forem contínuas). Estes aspectos serão confirmados em nosso estudo simulado.

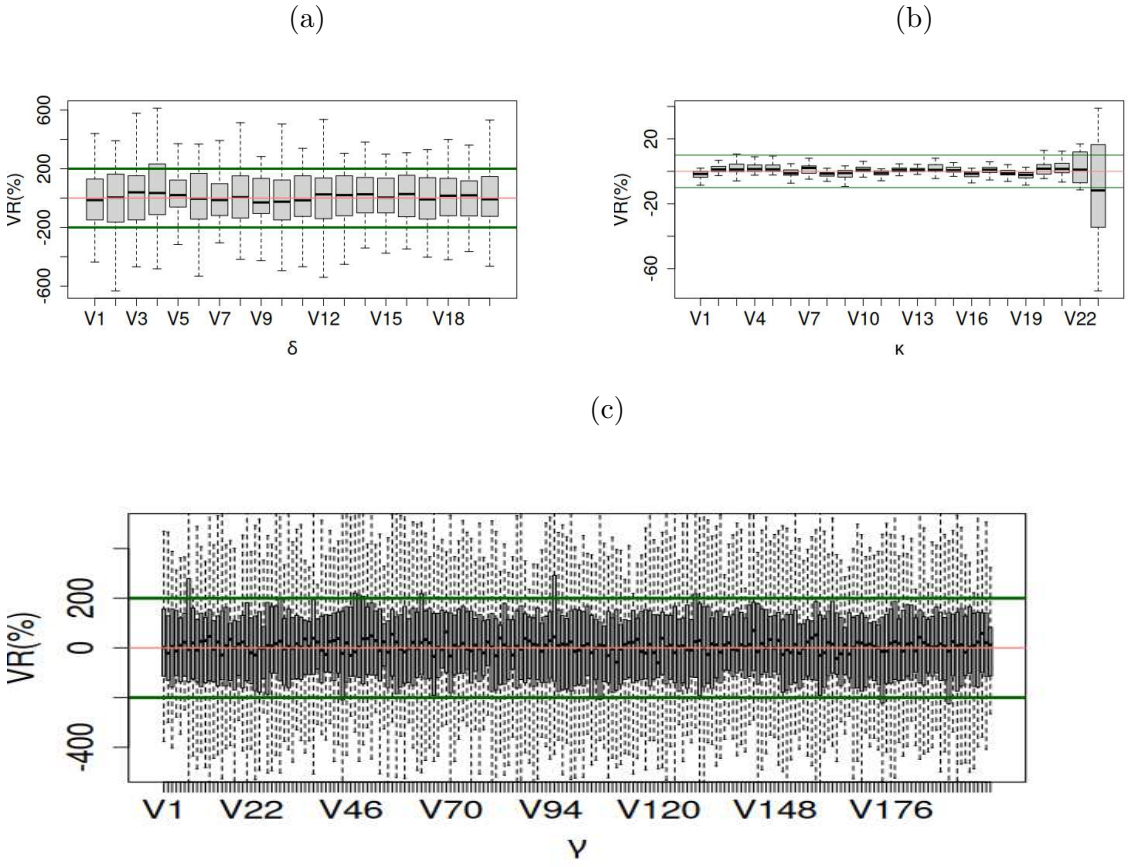


Figura 3.2: *Boxplots* dos VRs das estimativas para o gerador $M_2^{\oplus}$ e ajuste com δ e γ. Painel (a): *boxplots* dos VRs das medianas *a posteriori* do efeito aleatório temporal. Segmentos $V1 - V20$ correspondem aos tempos. Painel (b): *boxplots* dos coeficientes, sendo que $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas no tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto. Painel (c): *boxplots* do efeito espacial, $V1 - V200$ representa os 200 sítios.

Os VRs das medianas *a posteriori* dos parâmetros de variância da estrutura CAR para os efeitos aleatórios, τ_δ e τ_γ, bem como para o parâmetro de dispersão, ϕ, são exibidos na Figura 3.3, Painéis (a),(b) e (c), respectivamente. Observa-se que o VR de τ_δ apresenta um histograma assimétrico que varia de -50% a 100%, com maior número de medidas negativas. Isso sugere uma tendência à subestimação desse parâmetro, o que pode ser resultado, mais uma vez, dos desafios encontrados na estimação do intercepto do modelo. Em contrapartida, o VR de τ_γ varia entre -20% e 20%, com distribuição simétrica em torno de zero. No Painel (c), o VR de ϕ flutua entre -10% e 5%, com moda próxima a zero, mas maior número de medidas positivas, demonstrando uma leve sobrestimação do parâmetro. Contudo, com um VR absoluto inferior a 10%, é possível dizer que na

comparação com outros parâmetros da modelagem, temos maior aproximação para o valor real de ϕ.

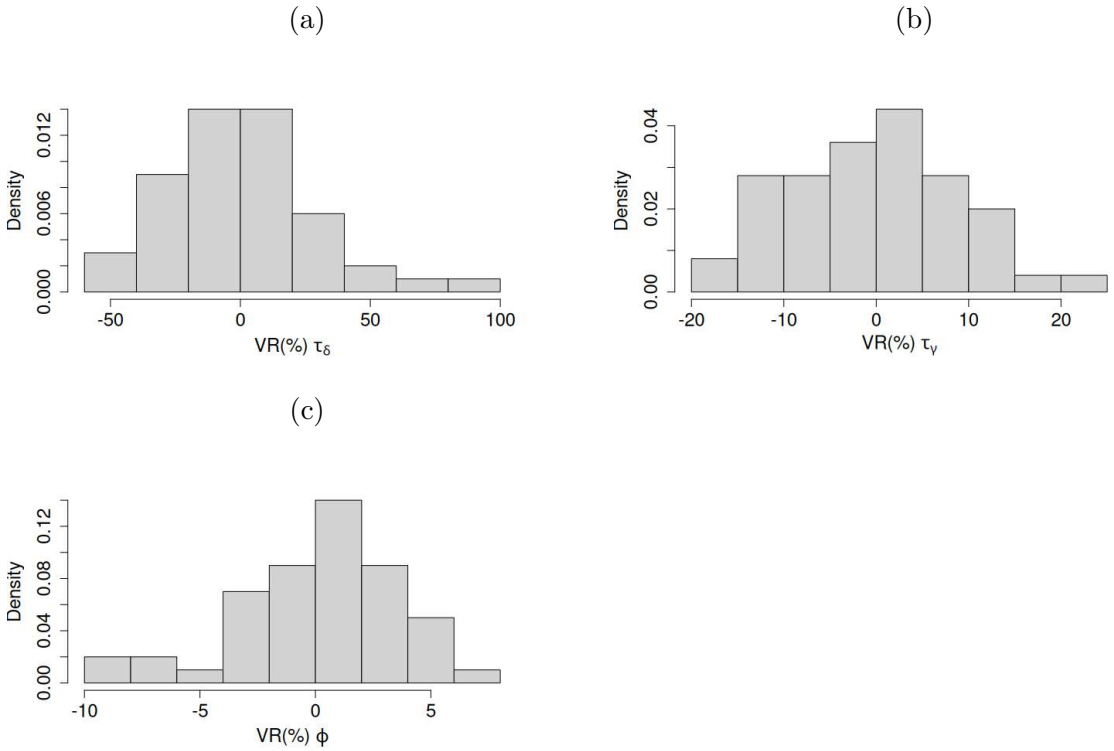


Figura 3.3: Histogramas dos VRs das medianas *a posteriori* dos parâmetros de variância do CAR para os efeitos aleatórios temporal, Painel (a), e espacial, Painel (b). O parâmetro de dispersão da regressão Bessel, ϕ, também é explorada no Painel (c). Gerador $M_2^{\oplus}$ e ajuste com δ e γ.

Esse ajuste completo apresentou VRs absolutos elevados para γ e δ, acima de 500% algumas vezes, mas, no geral, o consideramos um ajuste adequado, já que os parâmetros de regressão mostram uma boa aproximação ao valor real. Perceba que os VRs absolutos são menores que 10% para a grande maioria dos κ's. Temos, também, VRs abaixo de 40% para os parâmetros que são notoriamente mais difíceis de serem estimados, o intercepto do modelo e a variável binária, Segmentos V_{23} e V_{22}, respectivamente. Nas próximas seções vamos analisar situações de modelos mal-especificados e comparar os resultados com o ajuste obtido aqui para a modelagem bem-especificada. Essa estratégia é essencial para entendermos a importância de estimar os efeitos aleatórios no modelo de regressão Bessel e justificar sua introdução na modelagem.

Modelo Gerador $M_2^{\delta\ominus}$: Ajuste com δ e γ

Nessa seção faremos a análise de um modelo que leva em conta ambos os efeitos aleatórios ajustado a dados gerados por $M_2^{\delta\ominus}$. A Figura 3.4(a) mostra as medianas *a posteriori* do efeito temporal, δ, já que não é possível calcular o VR para esse parâmetro dado que os valores reais são zero. Observando as medianas *a posteriori* de δ notamos valores absolutos de no máximo 2×10^{-4}, o que é claramente perto de 0. Podemos dizer que o modelo consegue capturar bem a ausência do efeito temporal nos dados. Considerando agora o efeito aleatório espacial, γ, temos resultados parecidos com o caso da seção anterior, com VRs entre -500% e 500% e os segundo e terceiro quartis entre -200% e 200%, como visto na Figura 3.4(b). É possível concluir que o modelo M_2 não teve prejuízo quanto a estimação de γ ao tentar ajustar dados com ausência de δ. Em termos dos coeficientes κ, Painel (c), percebemos que neste ajuste o VR do intercepto (Segmento V_{23}) foi consideravelmente menor do que aquele na Figura 3.2(b), por volta do intervalo entre -20% e 20%. Este resultado é uma implicação da remoção da fonte de confundimento dada pela conexão existente entre κ_0 e o efeito δ. Em outras palavras, o M_2 é capaz de fornecer melhor aproximação para o intercepto quando δ não está presente. Os demais coeficientes de regressão tiveram VRs muito semelhantes aos exibidos na Figura 3.2(b), logo a ausência do efeito δ nos dados não parece ter impacto sobre coeficientes que multiplicam covariáveis.

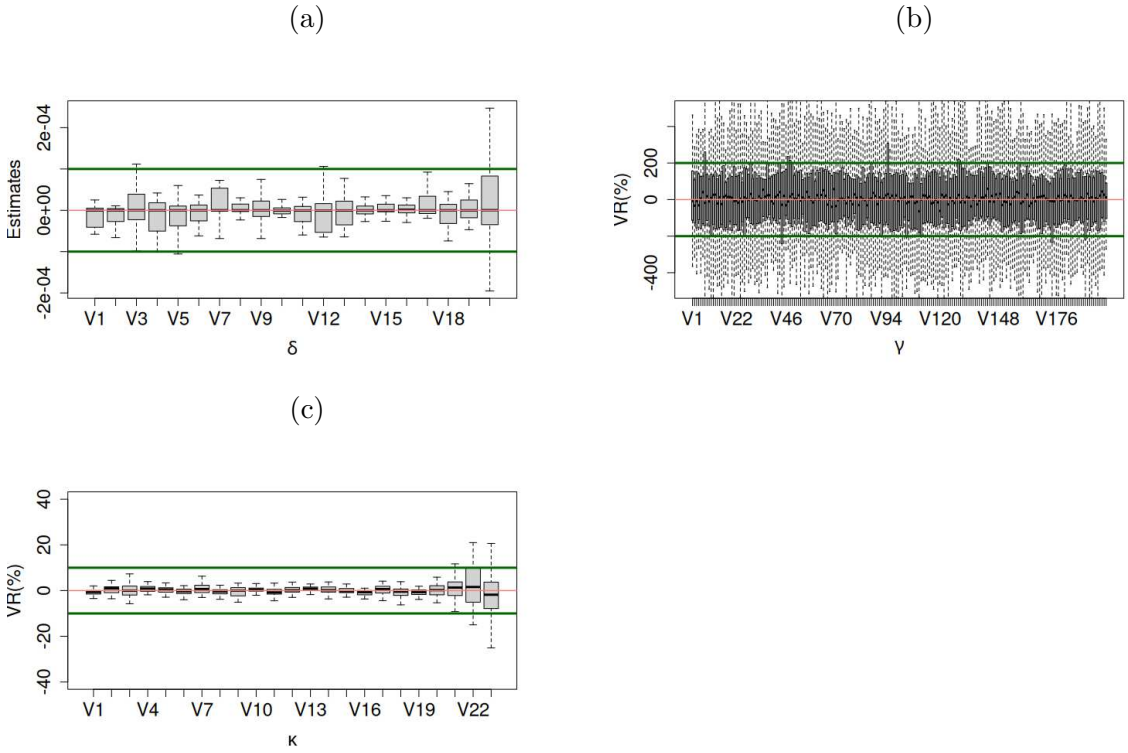


Figura 3.4: *Boxplots* para o gerador $M_2^{\delta\ominus}$ e ajuste com δ e γ. Painel (a): *boxplots* dos VRs das medianas *a posteriori* do efeito aleatório temporal. Segmentos $V1 - V20$ correspondem aos tempos. Painel (b): *boxplots* do efeito espacial, $V1 - V200$ representam os 200 sítios. Painel (c): *boxplots* dos coeficientes, sendo que $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas no tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo.

No que diz respeito aos parâmetros τ_γ e ϕ, observamos na Figura 3.5 que ambos tem VR menor que no modelo bem-especificado, sendo que o VR da mediana *a posteriori* de τ_γ, Painel (a), varia entre -20% e 20% e o de ϕ entre -4% e 4%. Essa melhora nas estimativas desses parâmetros de variância está relacionada à exclusão do efeito temporal na geração dos dados simulados. A ausência do efeito δ contribui para uma menor "distorção" em termos de variabilidade introduzida aos dados. O M_2 consegue detectar bem a ausência de δ e perante um banco simulado com menos fontes de variação, esse modelo mostra uma boa capacidade de adaptação atingindo melhores estimativas para τ_γ e ϕ. Por outro lado, o parâmetro de variância τ_δ não mostrou estimativas absolutas superiores a 1.5×10^{-4}, o que era de se esperar, já que o modelo reconheceu corretamente a ausência de δ nos dados.

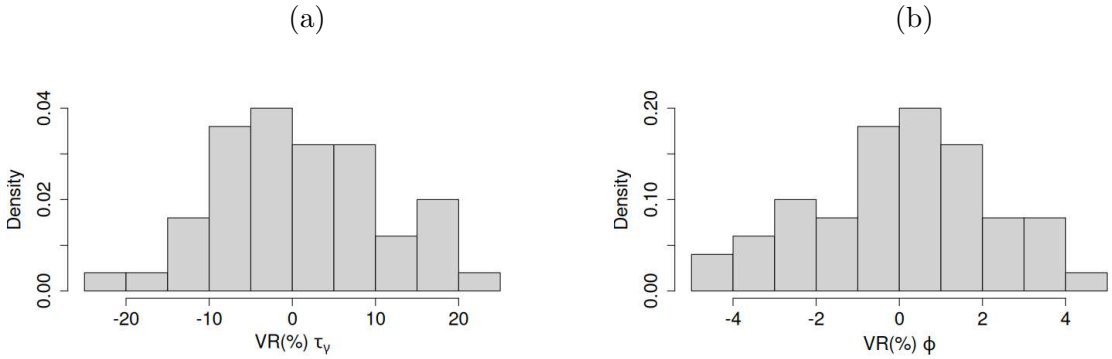


Figura 3.5: Histogramas dos VRs das medianas *a posteriori* dos parâmetros de variância do modelo CAR do efeito aleatório espacial, Painel (a), e do parâmetro de dispersão da regressão Bessel, ϕ, Painel (b), para o gerador $M_2^{\delta\ominus}$ e ajuste com δ e γ.

Esta análise do ajuste M_2 mal-especificado indica que não há perda em ajustar o modelo completo, isto é, que estima γ e δ. É possível notar que se não houver a presença do efeito aleatório temporal, o modelo vai estimar este efeito próximo de zero, tendo ainda a vantagem de conseguir atingir melhores estimativas para ϕ. Além disso, nesse ajuste mal-especificado obtivemos VRs absolutos menores do que no modelo bem-especificado, o que pode ser devido à menor variância presente nos dados do primeiro em relação ao segundo.

Modelo Gerador $M_2^{\oplus}$: Ajuste sem δ e com γ

Nessa seção vamos ajustar o M_2 apenas com efeito aleatório espacial, γ, em dados gerados a partir do gerador $M_2^{\oplus}$. Neste caso, o modelo para ajuste também é dito mal-especificado, por não levar em conta o efeito temporal presente nos dados.

Na Figura 3.6 temos os *boxplots* dos VRs do efeito aleatório espacial, Painel (a), e dos coeficientes de regressão, Painel (b). Não é possível notar grandes diferenças nos VRs de γ em relação ao que foi exibido nas Figuras 3.2(c) e 3.4(b). Contudo fica claro que o modelo tem dificuldades de estimar os coeficientes. A maioria dos *boxplots* do Painel (b) não abrange o zero e possui mediana próxima de -50% ou 50%, o que denota grande vício na estimação. Em conclusão, este resultado mostra claramente que há grande prejuízo em ajustar o modelo sem o efeito δ perante dados que apresentam estrutura de dependência temporal. O leitor deve ficar, então, atento ao fato de que o estudo desenvolvido aqui sugere que o modelo M_2 completo (com δ e γ) deve ser a primeira opção para a modelagem de dados, pois há grande risco de vício elevado quando a versão mais simples (sem δ) é aplicada.

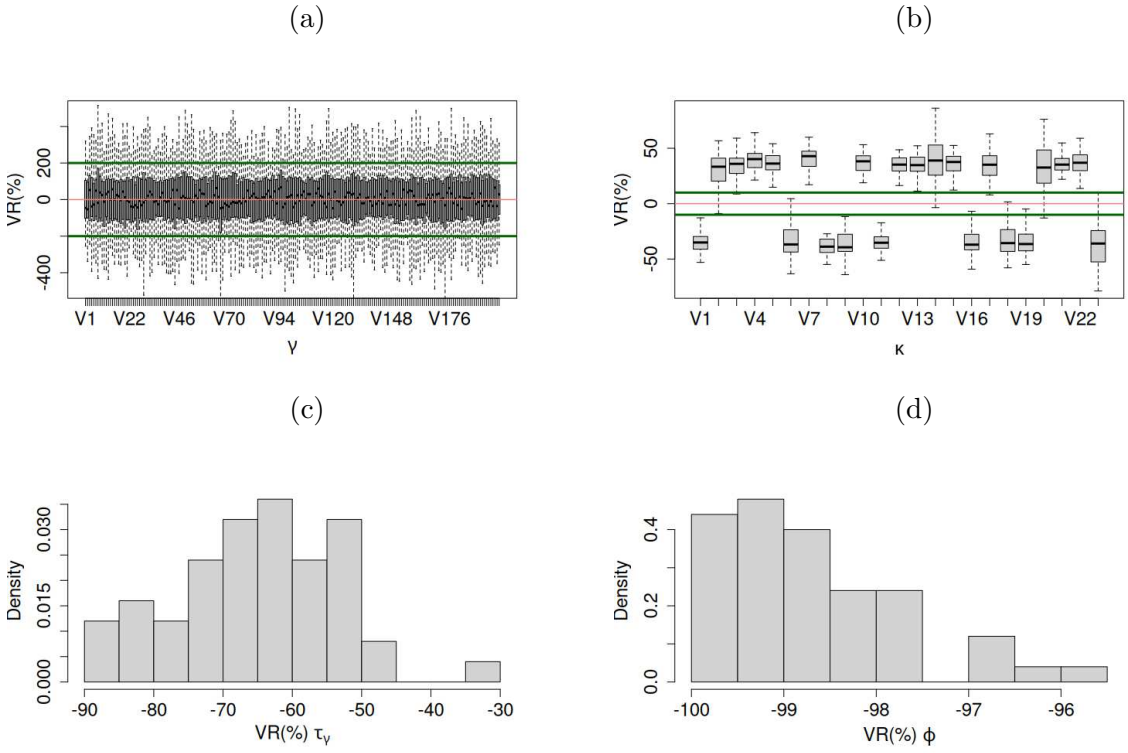


Figura 3.6: VRs das medianas *a posteriori* do gerador $M_2^{\oplus}$ e ajuste sem δ e com γ. Painel (a): *boxplots* do efeito espacial, $V1 - V200$ representam os 200 sítios. Painel (b): *boxplots* dos coeficientes de regressão, segmentos $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo. Histogramas dos VRs dos parâmetros de variância do modelo CAR do efeito aleatório espacial (τ_γ), Painel (c), e do parâmetro de dispersão da regressão Bessel, ϕ, Painel (d).

Na Figura 3.6(c) os VRs do parâmetro τ_γ estão consistentemente negativos, variando entre -90% e -30%, enquanto na Figura 3.6(d), os VRs do parâmetro de dispersão, ϕ, variam no intervalo de -100% a -95%. Em ambos os casos, observa-se uma considerável subestimação desses parâmetros. Esses resultados eram previstos, uma vez que esse ajuste não leva em consideração a importância da correlação temporal presente nos dados. A conclusão é que ajustar o M_2 sem δ para dados contendo o efeito temporal não é recomendado. Reiteramos que o analista estará mais seguro em termos de estimação quando trabalhar com o modelo completo contendo δ.

Modelo Gerador $M_2^{\gamma\ominus}$: Ajuste com δ e γ

Nessa seção analisaremos o caso em que os dados são gerados do $M_2^{\gamma\ominus}$, sem efeito espacial γ, mas ajustamos o modelo completo, contendo efeito espacial e temporal. Assim como no primeiro ajuste realizado para o $M_2^{\oplus}$, Figura 3.2, os VRs de δ, Figura 3.7(a),

têm os quartis centrais na faixa entre -200% e 200%. Já os VRs dos coeficientes κ, Figura 3.7(b), apresentam VRs muito próximos de 0, inclusive aqueles associados às covariáveis com apenas uma medida, Segmentos V_{21} e V_{22}, que possuíam VRs mais altos no caso completo. Contudo, os VRs do coeficiente associado ao intercepto ficaram entre -60% e 50%, que é semelhante ao caso completo bem-especificado. Em termos de τ_δ e ϕ, Figuras 3.7(c) e 3.7(d), também observamos uma redução no intervalo dos VRs em relação ao modelo bem-especificado. Além disso, os parâmetros associados à variação espacial, γ e τ_γ tiveram estimativas muito próximas de 0, menores que 10^{-3}. Assim como no caso em que ajustamos o modelo completo em dados gerados sem efeito temporal, observamos uma redução expressiva nos VRs de quase todos os parâmetros estimados. Isso pode ser novamente atribuído ao fato de termos excluído uma forte fonte de variação na geração dos dados. Portanto, o ajuste completo se mostrou competente em recuperar os parâmetros do modelo mesmo na ausência de um dos efeitos aleatórios na geração dos dados. Neste caso, foi estimada uma correlação espacial como zero e os demais parâmetros com menor incerteza em relação ao caso gerado com δ e γ, Figura 3.2. Em resumo, o M_2 completo tem bom comportamento perante dados sem o efeito espacial.

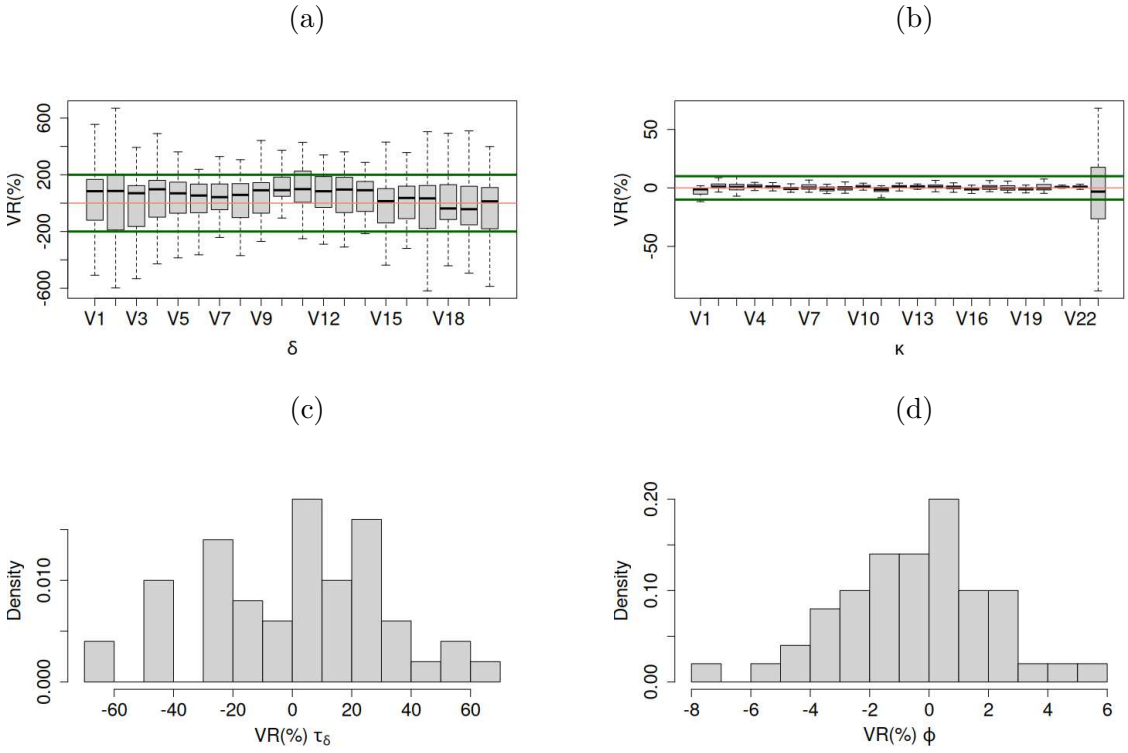


Figura 3.7: VRs das medianas *a posteriori* do gerador $M_2^{\gamma\ominus}$ e ajuste com δ e γ. Painel (a): *boxplots* do efeito aleatório temporal. Segmentos $V1 - V20$ correspondem aos tempos. Painel (b): *boxplots* dos coeficientes, sendo que $V1 - V20$ correspondem aos VRs dos 20 coeficientes de regressão associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo. Histogramas dos VRs dos parâmetros de variância do modelo CAR do efeito aleatório espacial (τ_δ), Painel (c), e do parâmetro de dispersão da regressão Bessel, ϕ, Painel (d).

Modelo Gerador $M_2^{\oplus}$: Ajuste com δ e sem γ

A Figura 3.8 mostra as estimativas do modelo mal-especificado em que ajustamos M_2 sem a presença do efeito espacial em dados gerados por $M_2^{\oplus}$. O efeito temporal está no modelo gerador e no modelo de ajuste. Nesse caso, obtivemos resultados praticamente idênticos aos observados na situação em que o M_2 não estima a parte temporal δ, Figura 3.6. Perceba a alta amplitude dos VRs no Painel (a) e os *boxplots* em patamares longe de zero no Painel (b). Os parâmetros τ_δ e ϕ estão claramente subestimados, conforme indicam os VRs negativos nos Painéis (c) e (d). Ressaltamos, mais uma vez, que o modelo sem a presença de um dos efeitos aleatórios não tem um ajuste satisfatório. Logo é recomendada a utilização do modelo completo nas análises.

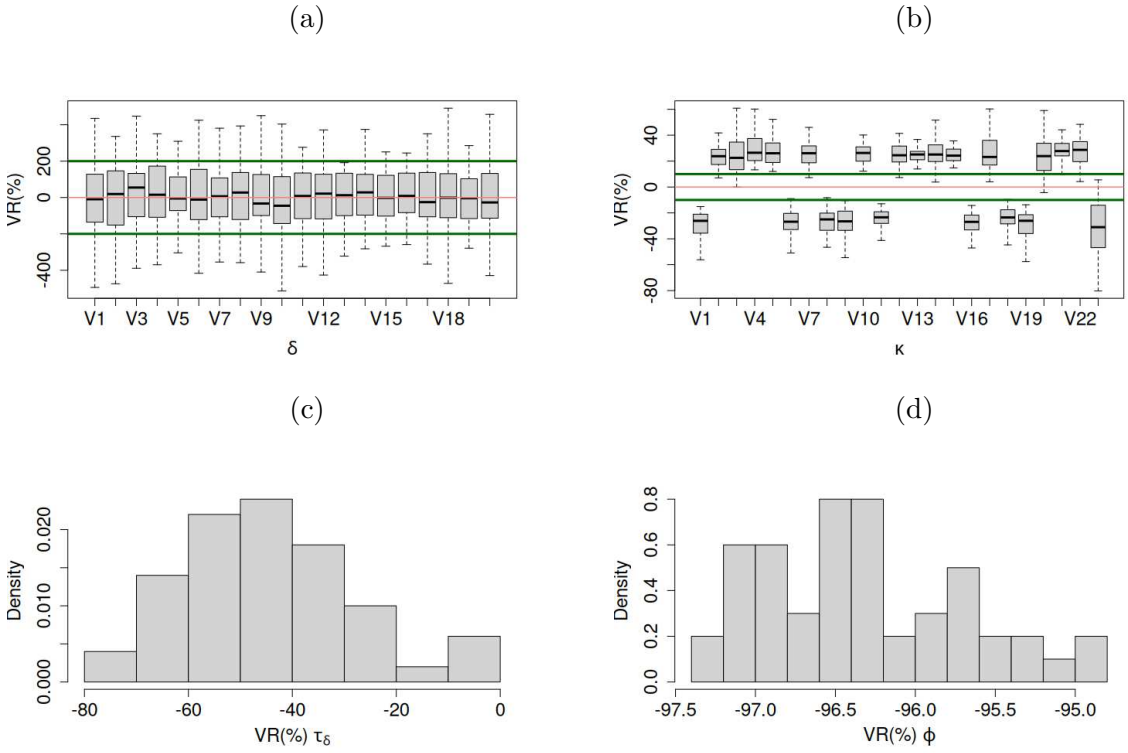


Figura 3.8: VRs das medianas *a posteriori* do gerador $M_2^{\oplus}$ e ajuste com δ e sem γ. Painel (a): *boxplots* do efeito aleatório temporal. Segmentos $V1 - V20$ corresponde aos tempos. Painel (b): *boxplots* dos coeficientes. Segmentos $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo. Histogramas dos VRs dos parâmetros de variância do modelo CAR do efeito aleatório espacial (τ_δ), Painel (c), e do parâmetro de dispersão da regressão Bessel, ϕ, Painel (d).

Terminamos aqui a análise dos cenários envolvendo dados gerados por M_2 e ajustados também por M_2. A seguir trataremos dos ajustes de M_3 a dados provenientes do gerador M_3.

Modelo Gerador $M_3^{\oplus}$: Ajuste com γ e Dependência Temporal em κ

Nessa seção, trataremos da análise do modelo ajustado em dados gerados por $M_3^{\oplus}$. Esse modelo não possui o efeito aleatório temporal δ, em contrapartida, introduzimos a dependência no tempo a partir de uma estrutura de correlação aplicada aos coeficientes da regressão, κ.

Observamos na Figura 3.9 os *boxplots* do efeito espacial, γ, e dos coeficientes κ, Painéis (a) e (b), respectivamente. Assim como no caso em que aplicamos o M_2 completo em dados gerados por $M_2^{\oplus}$, Figura 3.2, os dois quartis centrais dos VRs de γ estão entre -200% e 200%, contudo, há uma mudança drástica nos VRs dos elementos

de κ. Os coeficientes associados à variável medida ao longo do tempo, Segmentos V_1 ao V_{20}, apresentam VRs entre -600% e 600% com os quartis centrais entre -200% e 200%, enquanto os coeficientes de variáveis com medidas únicas, Segmentos V_{21}, V_{22} e V_{23}, tiveram estimativas entre -20% e 20%, semelhante ao que ocorreu para M_2. É importante frisar que este modelo não apresenta as mesmas dificuldades na estimativa do intercepto que M_2, já que o efeito aleatório temporal não é introduzido de forma aditiva, evitando que haja confundimento com o intercepto. O aumento dos VRs das estimativas de κ (comparando com M_2) era esperado, já que esses coeficientes agora contemplam a estrutura de correlação temporal do modelo.

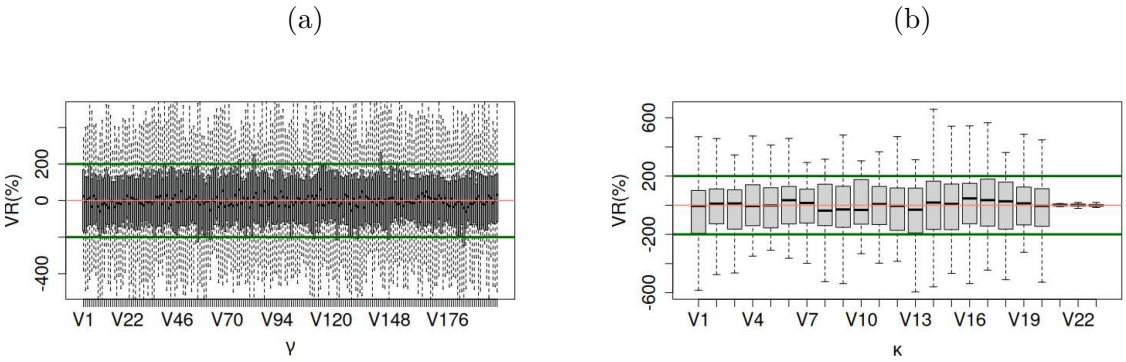


Figura 3.9: VRs do gerador $M_3^{\oplus}$ e ajuste com γ e dependência temporal em κ. Painel (a): *boxplots* do efeito aleatório espacial, segmentos $V1 - V200$ representando os 200 sítios. Painel (b): *boxplots* dos coeficientes, onde segmentos $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo.

Na Figura 3.10(a) os VRs do parâmetro τ_κ estão entre -60% e 60% com moda por volta de -30%, valores semelhantes àqueles encontrados para τ_δ em M_2. Já τ_γ, Figura 3.10(b), apresenta os VRs entre -25% e 20%, que também é semelhante aos valores do M_2 bem-especificado, Figura 3.3(b). Já na Figura 3.10(c), os VRs do parâmetro ϕ estão centrados em 0 e dentro do intervalo entre -10% e 10%. Portanto, não há grandes diferenças entre os modelos M_2 e M_3 bem-especificados no que diz respeito aos VRs dos parâmetros de variância dos efeitos aleatórios e do parâmetro de dispersão.

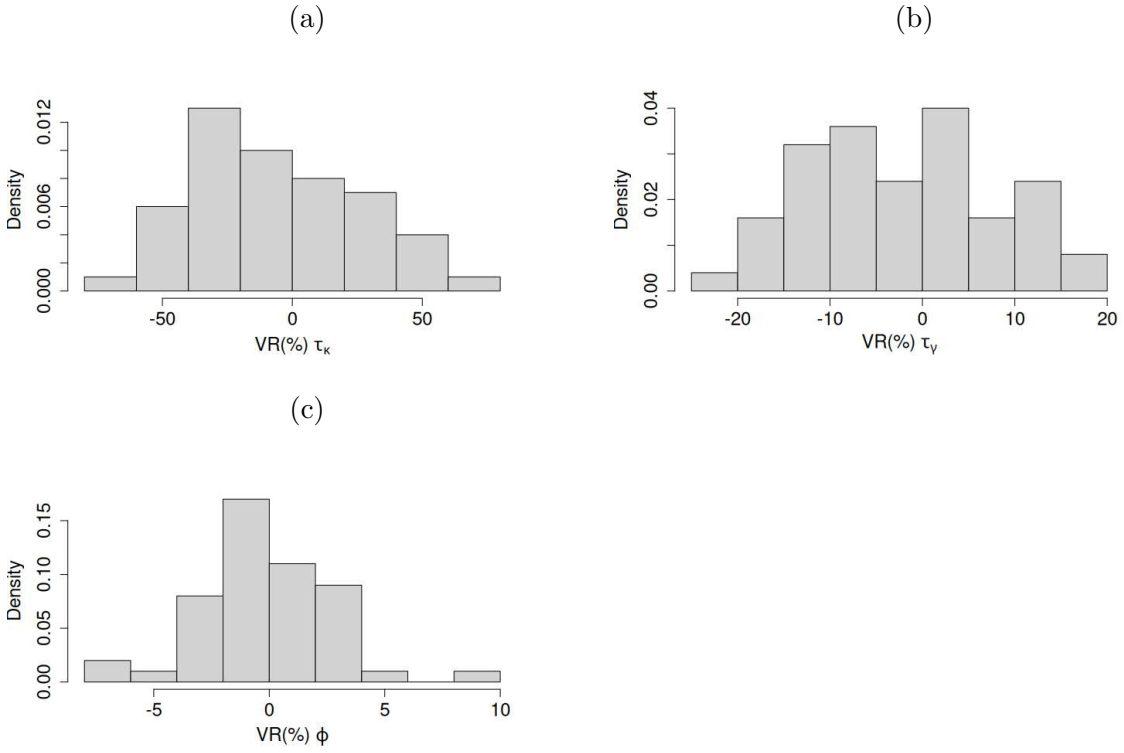


Figura 3.10: Histogramas da distribuição dos VRs para os parâmetros de variabilidade τ_κ, τ_γ e ϕ, Painéis (a), (b) e (c), respectivamente. Gerador $M_3^{\oplus}$ e ajuste com γ e dependência temporal em κ.

Finalizamos aqui a análise de desempenho do M_3 bem-especificado. A próxima seção deste estudo considera os resultados da modelagem M_3 com má-especificação ocorrendo no sentido de ignorar a dependência temporal existente nos dados.

Modelo Gerador $M_3^{\oplus}$: Ajuste com γ e sem Dependência Temporal em κ

Nessa seção, examinaremos o cenário em que os dados são gerados de $M_3^{\oplus}$, mas o modelo estima os coeficientes κ assumindo que eles são independentes. Os VRs desse modelo mal-especificado são apresentados na Figura 3.11. No Painel (a), os VRs de γ exibem uma ampla variação, com o primeiro e o segundo quartil entre -500% e 500%. Além disso, as medianas dos VRs apresentam valores absolutos acima de 200% em várias instâncias. Já no Painel (b), é notável a sobrestimação dos coeficientes de regressão, κ, com medianas dos VRs em torno de 200%, com exceção do intercepto, cuja mediana está em torno de -100%. A subestimação do intercepto ocorreu diante de um ajuste assumindo independência no tempo para dados correlacionados no tempo.

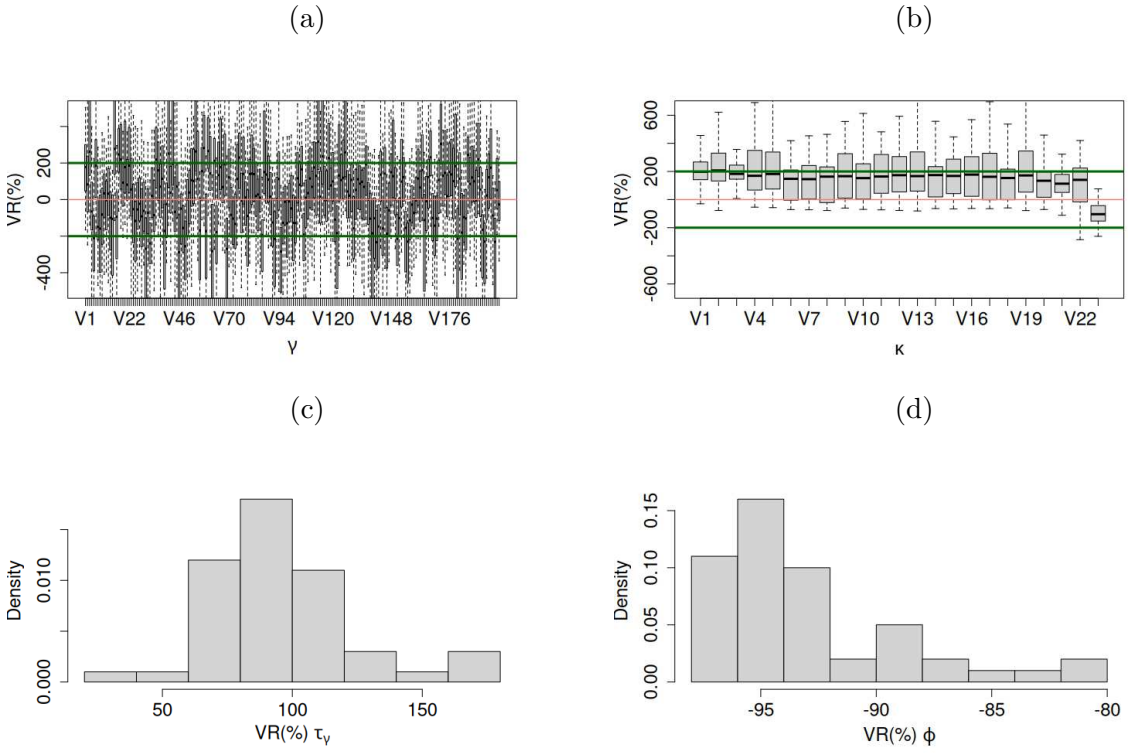


Figura 3.11: VRs do gerador $M_3^{\oplus}$ e ajuste com γ e sem dependência temporal em κ. Painel (a): *boxplots* do efeito aleatório espacial. Segmentos $V1-V200$ representam os 200 sítios. Painel (b): *boxplots* dos coeficientes, onde segmentos $V1-V20$ correspondem aos VRs dos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo. Histogramas dos VRs do parâmetro de variância do efeito espacial (τ_γ), Painel (c), e do parâmetro de dispersão do modelo Bessel (ϕ), Painel (d).

No que diz respeito ao parâmetro de variância de γ, τ^γ, observamos VRs centrados em 100% variando entre cerca de 20% e 180% (ver Painel (c)). Já os VRs de ϕ são negativos e variam entre -100% e -80%, com moda em -95% (Painel (d)). Em ambos os casos os VRs estão muito grandes, em valores absolutos, na comparação com o que observamos no modelo bem especificado, conforme a Figura 3.10 (b) e (c).

Mais uma vez, torna-se evidente que ocorrem perdas significativas na estimação quando não consideramos a presença da correlação temporal do modelo, principalmente no caso $M_3^{\oplus}$. Nesse cenário, a falta da informação temporal acarretou em VRs muito maiores do parâmetro κ e até mesmo do efeito aleatório espacial, γ, em relação aos VRs dos casos mal-especificados de $M_2^{\oplus}$.

Modelo Gerador $M_3^{CAR\ominus}$: Ajuste com γ e Dependência Temporal em κ

Tratamos agora do cenário $M_3^{CAR\ominus}$, em que geramos dados sem qualquer correlação temporal, mas ainda com γ e ajustamos o modelo completo, que estima a correlação temporal via modelo CAR nos coeficientes κ e a correlação espacial por meio de γ. Nesse contexto podemos avaliar se o M_3 com CAR (com dependência temporal entre coeficientes) é robusto mesmo na ausência de correlação temporal nas covariáveis.

Notamos, pela Figura 3.12, que os VRs de γ e κ, Painéis (a) e (b), respectivamente, são muito semelhantes àqueles do caso bem especificado, Figura 3.9. Isso demonstra que o modelo estima corretamente os coeficientes mesmo que estes não possuam correlação temporal.

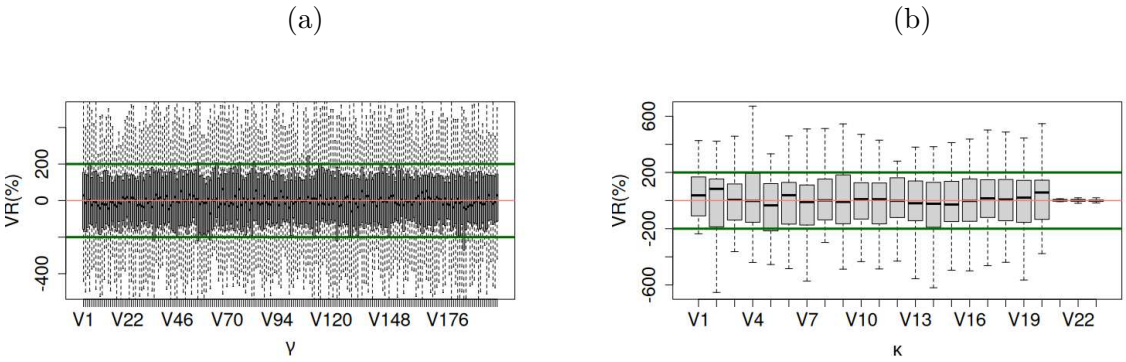


Figura 3.12: VRs para o gerador $M_3^{CAR\ominus}$ e ajuste com γ e dependência temporal em κ. Painel (a): *boxplots* do efeito aleatório espacial, onde segmentos $V1-V200$ representam os 200 sítios. Painel (b): *boxplots* dos coeficientes, os segmentos de $V1-V20$ correspondem aos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo.

Assim como γ e κ, os demais parâmetros também obtiveram VRs muito similares ao cenário bem-especificado. O parâmetro τ_γ estimado aqui tem uma distribuição entre -25% e 20%, praticamente o mesmo histograma do caso bem-especificado. Contudo, obtivemos uma dispersão maior dos VRs de τ_κ quando comparamos à Figura 3.10(a). Além disso, os VRs de ϕ estão entre -6% e 7%, um intervalo ligeiramente inferior ao caso bem-especificado.

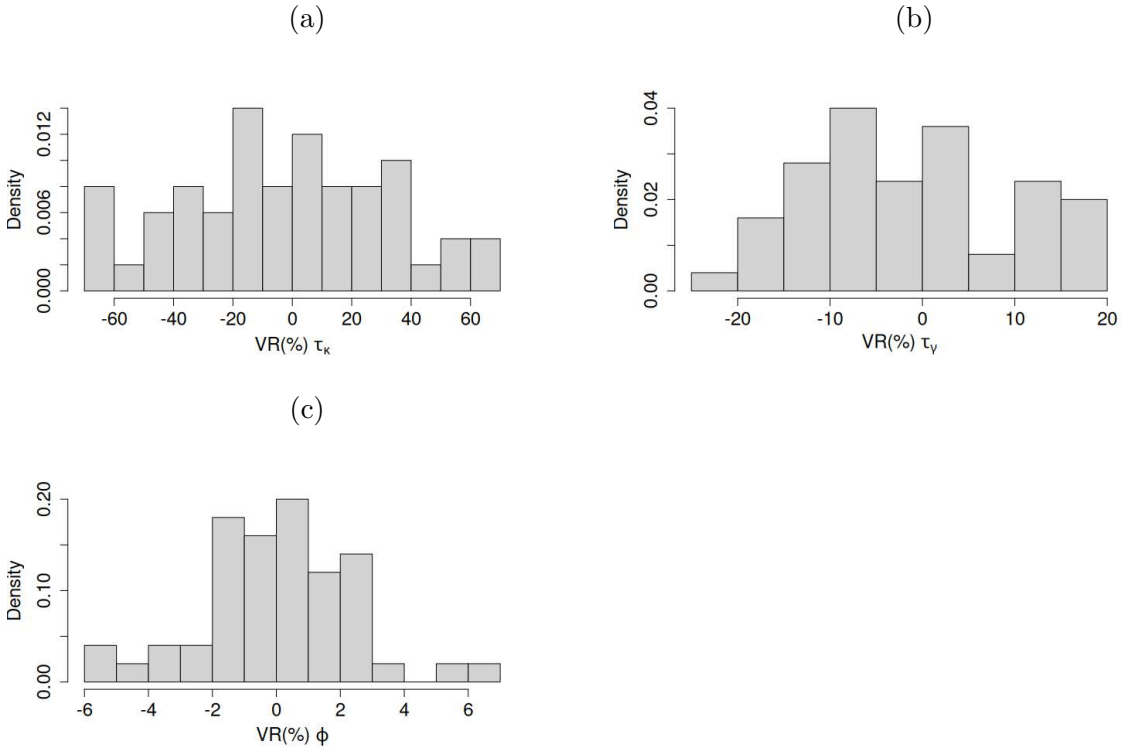


Figura 3.13: Histogramas dos VRs das medianas *a posteriori* dos parâmetros de variância do CAR para o efeito temporal, Painel (a), para o efeito espacial, Painel (b), e para o parâmetro de dispersão da regressão Bessel, ϕ, Painel (c). Gerador $M_3^{CAR\ominus}$ e ajuste com γ e dependência temporal em κ.

Uma vez que a variabilidade dos dados foi reduzida, poderíamos antecipar uma diminuição mais acentuada nos VRs desse cenário, de forma semelhante à redução observada ao passar do caso bem-especificado para aqueles em que ajustamos M_2 completo em dados gerados de $M_2^{\delta\ominus}$ ou $M_2^{\gamma\ominus}$. Entretanto, nos cenários envolvendo M_3, não fixamos os coeficientes κ. Mesmo quando não há correlação entre esses coeficientes, eles são gerados aleatoriamente a partir de distribuições normais independentes. Esse fator representa uma fonte significativa de aleatoriedade quando comparado com os dados gerados a partir de M_2, em que os elementos κ são fixos.

Para os casos sem a presença do efeito aleatório espacial, obtivemos resultados semelhantes àqueles dos cenários ajustando M_2. Os resultados desses ajustes estão no Apêndice A. Na próxima seção, faremos uma avaliação de má-especificação entre M_2 e M_3, assumindo um deles como gerador e o outro para ajuste. Em outras palavras, iremos gerar dados via $M_2^{\oplus}$ e ajustar o M_3 completo e vice-versa. A expectativa é que essa comparação demonstre qual modelo de ajuste incorre em VRs mais elevados quando a

modelagem geradora é desconhecida (sendo a abordagem oposta). Desta forma podemos decidir qual modelo é o mais "seguro"ao ser aplicado de forma ingênua.

Estudo de Má-especificação com Gerador Diferente do Modelo de Ajuste

Como descrito anteriormente, esse estudo cruzado visa definir qual modelo, M_2 ou M_3, seria o mais recomendado para uma aplicação ingênua, em que o analista não consegue ou não sabe definir qual tipo de gerador melhor se aplica aos dados e comete o erro de escolher o modelo errado para ajustar. Nessa situação, consideramos que o modelo cujos VRs tem distribuição mais próxima de zero, em geral, seria o mais adequado. Comparamos aqui apenas os parâmetros comuns a ambos os ajustes, isto é, κ, ϕ, γ e τ_γ.

A Figura 3.14 mostra o ajuste de M_3 em dados gerados por $M_2^{\oplus}$ na Coluna 1 (Painéis a, c, e, g). Por simplificação de linguagem, iremos mencionar este caso como M_3G_2. O caso de ajuste com M_2 para dados gerados via $M_3^{\oplus}$ será denominado M_2G_3, o qual está na Coluna 2 da Figura 3.14 (Painéis b, d, f, h).

No que diz respeito aos coeficientes κ, o caso M_2G_3 apresenta VRs centrados em zero, mas com variabilidade muito mais alta que M_3G_2, com intervalo 10 vezes maior. Os VRs de M_3G_2, por sua vez, apesar de indicarem *boxplots* com amplitudes menores (comparando com o Painel b), possuem centro próximo a -40% ou 40% (linhas verdes nos Painéis a e b). Em relação ao parâmetro de dispersão, ϕ, M_3G_2 demonstra uma forte subestimação, com VRs entre -100% e -95%, se compararmos com aqueles obtidos em M_2G_3, que estão centrados em zero e entre -4% e 4%. Os VRs para o parâmetro γ foram praticamente idênticos em ambos os modelos, contudo, aqueles obtidos para a variância de γ, τ_γ, espelham o que foi descrito para ϕ, com VRs negativos entre -90% e -30% no caso M_3G_2 e com valores entre -30% e 30% (centro zero) no M_2G_3.

Os resultados da Figura 3.14 sugerem que o M_2 apresenta estimação com viés absoluto menor (em relação ao M_3) para os parâmetros de variância. Além disso, apesar de existir a possibilidade de VRs absolutos da ordem de 10^2 para κ no M_2, os dois quartis centrais estão entre -150% e 150% com existência de VRs próximos de zero para algumas das réplicas no esquema MC. Já no M_3, apesar dos VRs absolutos de κ não ultrapassarem 60%, perceba que quase não existe réplica MC associada a um ajuste com VR de κ perto de zero. Dessa forma, esta avaliação permite concluir que o M_2 parece ser o mais seguro de ser aplicado na ausência de informações sobre a forma da correlação temporal do gerador dos dados.

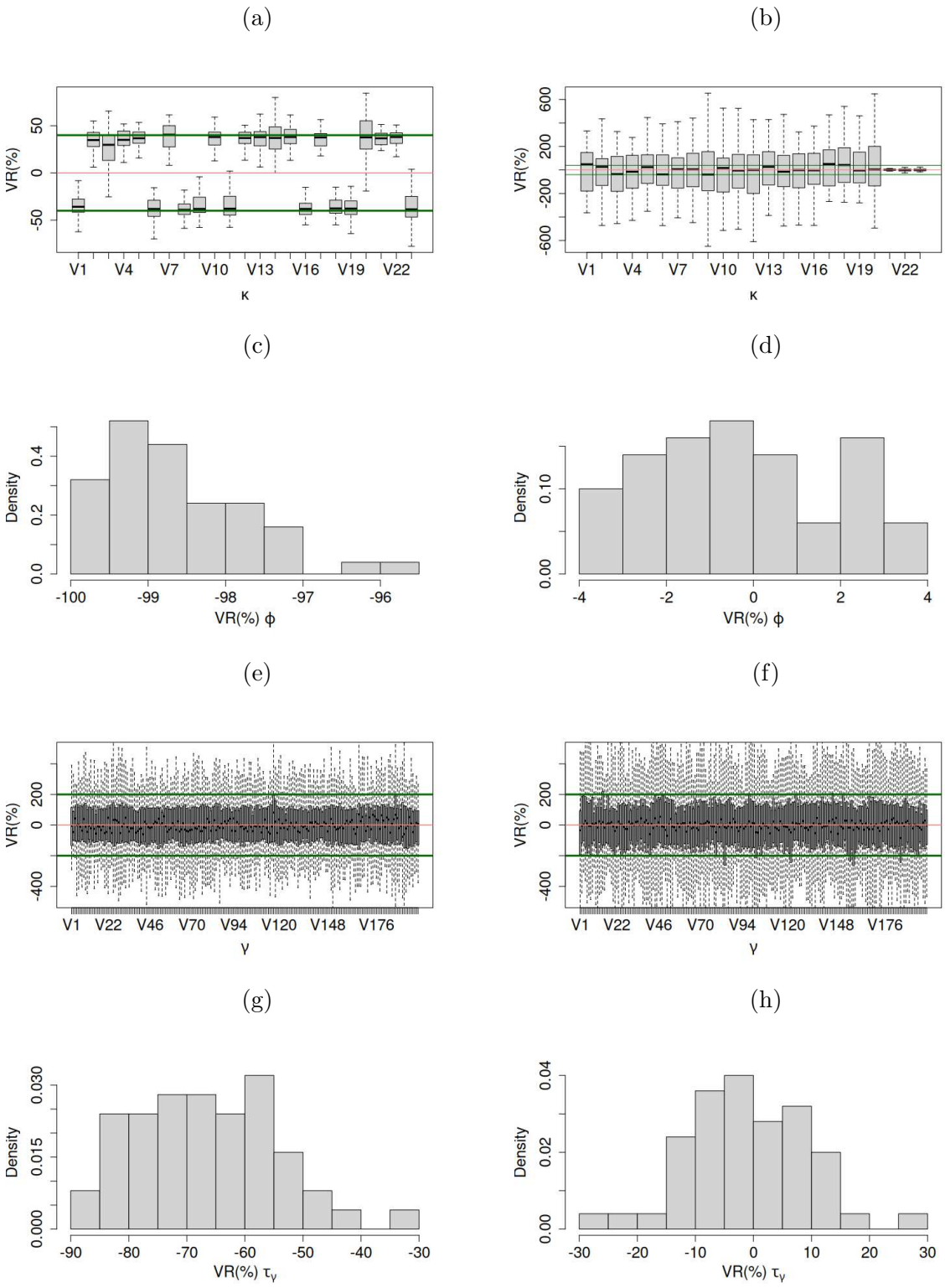


Figura 3.14: VRs em situação de má-especificação. Ajuste de M_3 com gerador $M_2^{\oplus}$ (Painéis a, c, e, g) e ajuste de M_2 com gerador $M_3^{\oplus}$ (Painéis b, d, f, h). Linha 1: *boxplots* dos coeficientes, κ. Segmentos $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto. Linha 2: histogramas do parâmetro de dispersão ϕ. Linha 3: *boxplots* do efeito espacial, γ, em que $V1 - V200$ representam os 200 sítios. Linha 4: histogramas do parâmetro de variância do efeito espacial, τ_γ.

A seguir vamos avaliar o modelo simplificado M_1. O M_1 é mais parcimonioso no sentido de resumir a informação das covariáveis observadas para diferentes tempos e adotar um único coeficiente para essa medida resumida. Diante desta maior simplificação, que de certa forma acarreta perda de informação, optamos por analisar o M_1 por último na sequência do estudo.

Modelo Gerador $M_1^{\oplus}$: Ajuste com δ e γ

Além dos geradores M_2 e M_3 temos um terceiro modelo, M_1. Ele pode ser visto como uma simplificação do M_2, em que não consideramos todos os tempos medidos de cada variável, mas criamos uma variável única que é a média das variáveis medidas em diferentes tempos e para cada sítio. Dessa forma mantemos o modelo espaço-temporal, porém, com menos covariáveis. A ideia dessa seção é avaliar o efeito da simplificação na estimação dos efeitos aleatórios.

Iremos tratar aqui do ajuste bem-especificado, isto é, o modelo completo com γ e δ sendo ajustado em dados gerados de $M_1^{\oplus}$. Na Figura 3.15 observamos os *boxplots* e histogramas dos VRs dos parâmetros estimados. Quando comparamos com os VRs obtidos no caso M_2 bem-especificado, Figuras 3.2 e 3.3, nota-se que, à exceção de κ, os demais parâmetros (Painéis a, c, d, e, f) apresentam valores muito parecidos nos dois ajustes, indicando que M_1 seja uma opção razoável (em relação a M_2) para explorar esses tipos de parâmetros (que não são coeficientes). Em relação aos κ's, Figura 3.15(b), os VRs das variáveis que não variam no tempo, Segmentos V_2 e V_3, se mantiveram entre -10% e 10%. Já o intercepto do modelo, Segmento V_4, agora varia entre -40% e 60% com mediana centrada em 0. Esse fato pode ser considerado uma melhora na estimação do intercepto em relação ao M_2. O Segmento V_1 está associado à variável que antes seria medida no tempo, mas que sofreria uma transformação de sumarização para adaptar ao M_1. Nesse caso, ela é gerada de uma distribuição uniforme entre -1 e 1. Portanto, os coeficientes representados em V_1 e V_2 estão associados a variáveis idênticas e devem ter os VRs semelhantes, o que é observado no Painel (b).

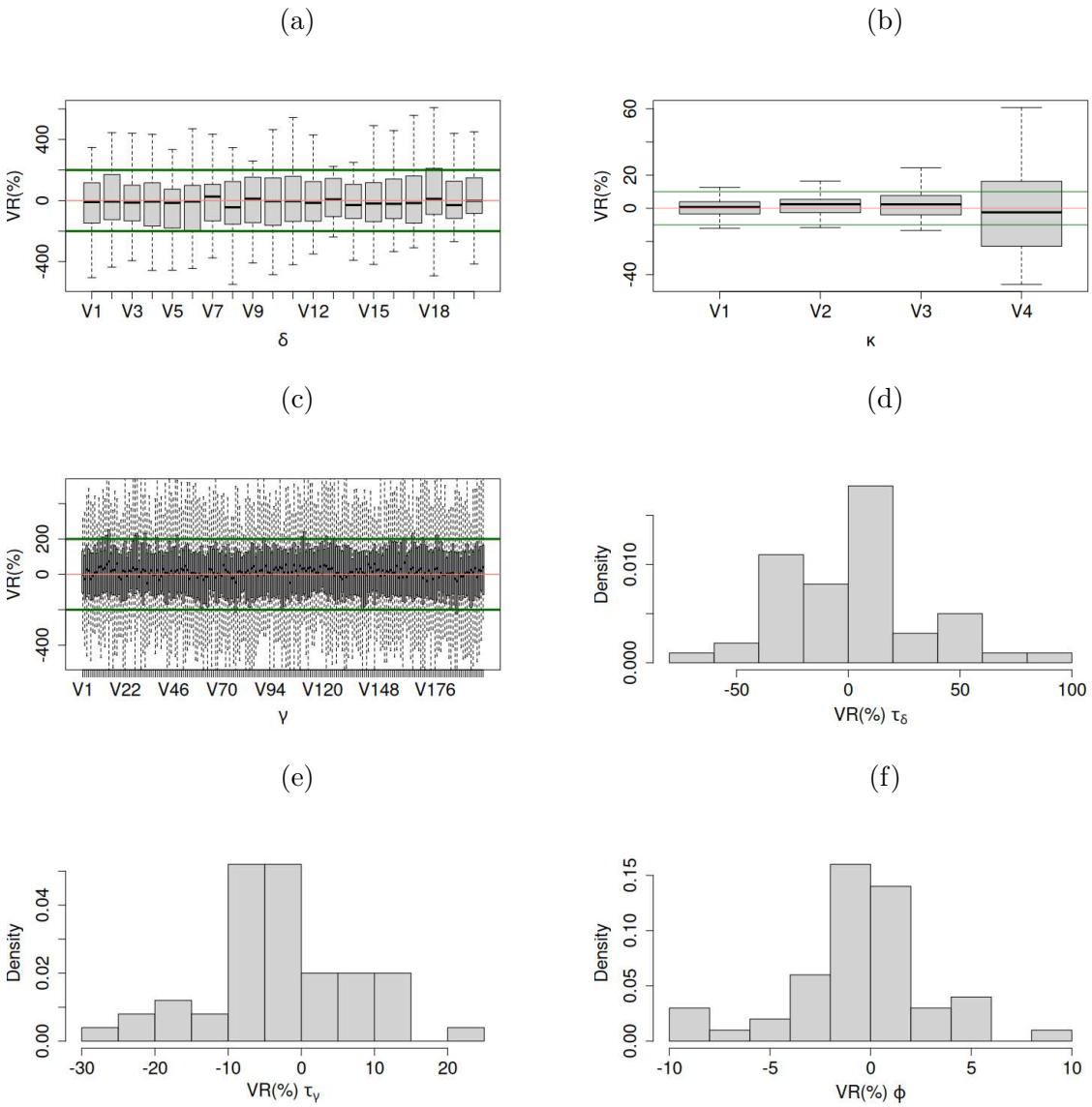


Figura 3.15: VRs das medianas *a posteriori* para gerador $M_1^{\oplus}$ e ajuste com δ e γ. Painel (a): *boxplots* do efeito temporal. Segmentos $V1 - V20$ correspondem aos tempos. Painel (b): *boxplots* dos coeficientes, em que os segmentos $V1 - V20$ correspondem aos VRs dos 20 coeficientes associados às medidas ao longo do tempo, $V21$ à variável uniforme de medida única, $V22$ à variável binária e $V23$ ao intercepto do modelo. Painel (c): *boxplots* do efeito espacial, segmentos $V1 - V200$ representam os 200 sítios. Histogramas dos parâmetros de variância do CAR para os efeitos temporal, Painel (d), e espacial, Painel (e). O parâmetro de dispersão da regressão Bessel, ϕ, é explorado no Painel (f).

Os demais casos em que há má-especificação dos modelos aplicados a dados gerados de $M_1^{\oplus}$, $M_1^{\delta\ominus}$ e $M_1^{\gamma\ominus}$ estão no Apêndice B. A escolha de não colocá-los no texto se deve ao fato dos resultados obtidos serem muito parecidos com os casos equivalentes de M_2 avaliados anteriormente. Todos os parâmetros do modelo são estimados com distribuição

de VRs muito semelhantes àqueles obtidos para M_2. Além disso, em todos os casos de M_1 há uma melhora nos VRs do coeficiente associado ao intercepto (em relação ao M_2), demonstrando que a simplificação imposta no M_1 não gera grandes perdas na estimação de diversos parâmetros comuns com M_2. Intuitivamente, isso seria especialmente verdade na situação em que as correlações entre as covariáveis medidas no tempo não são fortes.

Conclusão do Capítulo

Neste capítulo, foram feitas simulações, com esquema MC, de diversos cenários para cada modelo. Esses cenários incluíram ajustes bem-especificados e várias formas de má-especificação na tentativa de observar a robustez dos modelos sugeridos. Nenhum dos modelos é perfeito, enquanto M_2 apresenta um ajuste simples e eficaz, há certa dificuldade na estimação do intercepto, que interpretamos como proveniente de um confundimento com os efeitos aleatórios. Na tentativa de melhorar esse efeito indesejado presente em M_2, M_3 trás uma estrutura de correlação temporal entre os coeficientes da regressão, que aumenta o nível de complexidade da modelagem e dificulta a estimação em casos de má-especificação, mas ajuda a resolver o problema de identificação do intercepto. O M_1 foi introduzido como uma forma de simplificar o M_2 e se mostrou capaz de fornecer boas estimativas para os parâmetros comuns com M_2, mesmo diante de perda de informação a respeito das covariáveis. Vamos explorar os efeitos de cada um desses ajustes em dados reais no capítulo seguinte, onde faremos a análise dos dados de democracia que motivaram esse estudo.

Capítulo 4

Aplicação Real

A democracia é descrita por Kriesi et al. (2013) como o sistema político mais legítimo e desejável, que promove a igualdade, a liberdade e a participação política dos cidadãos. Contudo, nas últimas décadas, os regimes democráticos vêm encarando diversos desafios que enfraquecem sua legitimidade. Portanto, avaliar a qualidade da democracia permite entender o grau em que esses ideais democráticos estão sendo alcançados ao redor do mundo. Medir a qualidade da democracia ajuda a identificar dificuldades enfrentadas pelos sistemas políticos, assim como possíveis áreas de melhorias. Isso permite que governos, organizações da sociedade civil e pesquisadores identifiquem as lacunas existentes e adotem medidas para fortalecer e aprimorar as instituições democráticas.

A qualidade da democracia está intrinsecamente ligada a fatores socioeconômicos, geográficos e ambientais que podem influenciar o funcionamento e a sustentabilidade dos sistemas democráticos. Por exemplo, altos níveis de desigualdade econômica podem minar a inclusão política e social, comprometendo a qualidade da democracia. Da mesma forma, conflitos étnicos ou religiosos, instabilidade econômica, acesso limitado à educação e recursos naturais escassos podem representar desafios para o desenvolvimento e a manutenção de instituições democráticas robustas.

A análise desses fatores permite compreender como contextos específicos afetam a qualidade da democracia e como diferentes países podem enfrentar desafios distintos em sua jornada democrática, sendo fundamental para entendermos os problemas e as oportunidades que os sistemas democráticos enfrentam, auxiliando na busca por sociedades mais justas, inclusivas e participativas. Com essa motivação em mente, descrevemos nas seções subsequentes as variáveis selecionadas para a aplicação do modelo de regressão Bessel aos dados reais.

4.1 O Banco de Dados V-Dem

O banco de dados *V-Dem* (Coppedge et al., 2022), abreviação de *Varieties of Democracy*, é uma fonte de informações e indicadores sobre democracia ao redor do mundo. Desenvolvido por um consórcio de pesquisadores de várias instituições acadêmicas, o *V-Dem* reúne dados de mais de 200 países, abrangendo várias décadas. Contudo, neste trabalho selecionamos para análise apenas 17 anos, entre 2003 e 2019. Este é um período com dados mais completos das covariáveis e que abrange mudanças geopolíticas no cenário internacional que poderiam ser capturados na modelagem espaço-temporal.

O objetivo principal deste consórcio é medir e analisar diferentes aspectos da democracia, incluindo instituições políticas, liberdades civis, participação popular e direitos humanos. O banco de dados, portanto, oferece uma ampla gama de medidas e variáveis que permitem estudar as variações e os padrões da democracia ao longo do tempo e entre diferentes nações, sendo amplamente utilizadas em estudos acadêmicos, pesquisas comparativas e análises políticas.

Dentre os cinco principais índices de democracia disponíveis no banco de dados, selecionamos o *Índice de Democracia Eleitoral* como nossa variável de interesse. Esta é uma variável agregada que captura a presença e a qualidade dos elementos essenciais de uma democracia representativa, combinando diversos indicadores que abrangem diferentes dimensões da democracia. Segundo Lindberg et al. (2014), uma democracia eleitoral, ou poliarquia, é um sistema político caracterizado por múltiplos atores e instituições que participam no processo de tomada de decisões via processo eleitoral periódico. Dessa forma, esse índice serve como um ponto de referência fundamental para a construção das outras variáveis, permitindo uma avaliação mais precisa e abrangente da qualidade da democracia em um determinado contexto. Ela ajuda a sintetizar e contextualizar os demais indicadores, fornecendo uma visão geral da saúde democrática de uma nação e facilitando análises comparativas entre diferentes países ao longo do tempo.

A variável resposta mencionada é construída a partir de uma combinação de indicadores que capturam diferentes dimensões da democracia, como eleições livres e justas, pluralismo político, liberdade de expressão, participação popular e Estado de direito. Por meio da análise desses indicadores, o *V-Dem* atribui um valor entre 0 e 1 para cada país e ano em que os dados estão disponíveis.

Valores mais altos indicam uma maior presença e qualidade dos elementos democráticos. Por exemplo, um país com uma pontuação próxima de 1 em determinado ano seria considerado uma democracia robusta, com eleições livres, instituições políticas

estáveis e respeito aos direitos e liberdades individuais. Por outro lado, um país com uma pontuação mais próxima de 0 indicaria a presença de restrições significativas à participação política e à liberdade de expressão. Dessa forma, ela fornece uma medida importante para avaliar o nível de democratização e o grau de respeito às instituições democráticas. No banco de dados a ser analisado, valores iguais a 0 ou 1 não são observados.

4.2 O Banco de Dados EPI

O *Environmental Performance Index* (EPI) (Wolf et al., 2022) é um banco de dados construído com medições de desempenho ambiental dos países ao redor do mundo. Ele é uma ferramenta criada pelo Centro de Direito e Política Ambiental da Universidade de Yale (EUA) e pelo Centro de Informações sobre Ciência da Terra Internacional da Universidade de Columbia (EUA) em parceria com o Fórum Econômico Mundial.

O EPI avalia e compara o desempenho ambiental de cerca de 180 países em uma ampla gama de indicadores e métricas relacionadas à saúde ambiental e ao bem-estar humano. Os indicadores abrangem várias áreas, como qualidade do ar, recursos hídricos, biodiversidade, mudanças climáticas, recursos naturais, poluição e saúde ambiental. O objetivo do EPI é fornecer uma avaliação abrangente e comparativa do desempenho ambiental dos países, a fim de incentivar e informar ações políticas e práticas que promovam a sustentabilidade ambiental e o desenvolvimento sustentável.

Neste trabalho não utilizamos os índices finais calculados no EPI. Apenas foram utilizadas algumas variáveis ambientais combinadas em um único índice via análise de componentes principais (PCA) exploratória, selecionando a primeira componente como a variável de interesse. Dessa forma, desenvolvemos um índice próprio para este trabalho, construído como uma combinação linear de variáveis originais do conjunto de dados.

No caso das variáveis que estão relacionadas à qualidade do ar, essa seleção resultou na utilização apenas de poluentes atmosféricos: dióxido de enxofre (SO_2), óxidos de nitrogênio (NOx), ozônio troposférico (O_3) e monóxido de carbono (CO). Esses poluentes são medidos em partes-por-milhão (ppm) ao longo de vários anos, criando, portanto, um índice anual.

Além da qualidade do ar, também utilizamos variáveis relacionadas com o gerenciamento de resíduos. As duas variáveis são: resíduos sólidos controlados e taxas de

reciclagem. Elas foram medidas apenas uma vez no ciclo deste banco de dados, portanto, foi criado (via PCA) um índice único que leva em consideração ambas as variáveis.

Além das variáveis descritas anteriormente, utilizamos o logaritmo da densidade demográfica de cada país por ano - a transformação "*log*" foi aplicada em função da presença de países com valores muito altos de densidade demográfica, enquanto outros têm valores próximos de zero. Também está na análise o Produto Interno Bruto *per capita* por hora (PIB *per capita*/hora), calculado com base no PIB de cada país por ano. Por fim, incluímos uma variável binária de medida única que indica prevalência do sexo feminino nos países durante o período avaliado, tomando 1 se a população feminina foi maior que a masculina na maioria dos anos e 0 caso contrário.

4.3 Análise Exploratória dos Dados

Agora, vamos explorar as variáveis selecionadas para integrar o modelo de regressão, com o objetivo de compreender suas características, distribuições e relações com a variável de resposta. Essa análise nos permitirá identificar padrões, tendências e peculiaridades que podem influenciar o desempenho do modelo. Ao final deste capítulo teremos uma visão abrangente das variáveis que farão parte da regressão, estabelecendo uma base sólida para a etapa de escolha e ajuste do modelo.

Os gráficos na Figura 4.1 são o histograma, Painel (a), e *boxplot*, Painel (b), da variável resposta, *Índice de Democracia Eleitoral*. No histograma podemos observar uma leve bimodalidade na distribuição desta variável, com maior concentração de dados em torno dos valores 0.25 e 0.9. Essa característica sugere que a regressão beta não seria o modelo mais adequado para esses dados, sendo mais indicado o ajuste do modelo Bessel. Para confirmar essa suposição utilizamos a função `dbb` do pacote `bbreg` que validou o modelo Bessel como o mais adequado para o ajuste nesse conjunto de dados, quando comparado ao modelo beta. Pela Figura 4.1(b), podemos observar que a variável em questão não possui valores iguais a 0 ou 1, o que permite a viabilidade e adequação da utilização do modelo Bessel.

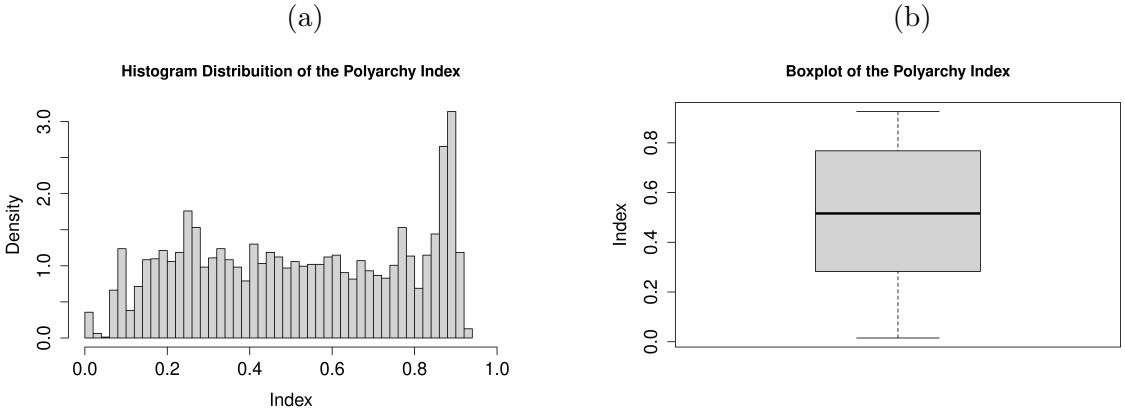


Figura 4.1: Histograma, Painel (a), e *boxplot*, Painel (b), da variável resposta, *Índice de Democracia Eleitoral*.

Calculamos, também, a estatística I de Moran para cada ano da variável resposta. A estatística I de Moran (Moran, 1950) é uma medida utilizada em análise de autocorrelação espacial para avaliar se existe padrão espacial em um conjunto de dados. A ideia básica é examinar se valores semelhantes estão próximos uns dos outros (autocorrelação espacial positiva) ou se valores diferentes estão próximos uns dos outros (autocorrelação espacial negativa).

Essa estatística compara a covariância ponderada dos valores entre locais com a variância total dos valores. Essencialmente, ela mede o grau de associação entre os valores em diferentes locais, levando em consideração a estrutura espacial definida pelos pesos w_{ij}. A fórmula geral para a estatística I de Moran é:

$$\frac{n}{\sum_{i,j} w_{ij}} \frac{\sum_{i,j} w_{ij}(\xi_i - \bar{\xi})(\xi_j - \bar{\xi})}{\sum_i (\xi_i - \bar{\xi})^2},$$

sendo que n é o número de sítios, w_{ij} são os elementos da matriz de pesos espaciais entre os sítios i e j, ξ_i e ξ_j são os valores observados nas localidades i e j e $\bar{\xi}$ é a média.

O resultado da estatística I de Moran é uma medida que varia entre -1 e 1. Se I é próximo de 1, indica autocorrelação espacial positiva (valores similares próximos uns dos outros). Por outro lado, se é próximo de -1, indica autocorrelação espacial negativa (valores diferentes próximos uns dos outros). Se I é próximo de 0, sugere uma distribuição espacial aleatória. No caso da variável resposta, obtivemos valores de I entre 0.49 e 0.56, com *p-valores* menores que 0.001, o que indica a presença de correlação espacial positiva nos índices de democracia medidos no mundo. Tal correlação sugere que países com altos (ou baixos) índices de democracia tendem a ser vizinhos de outros países com padrões semelhantes. Essa estrutura espacial nos dados é estatisticamente significativa, conforme

indicado pelos baixos *p-valores* obtidos, e, portanto, ressalta a importância de incorporar a informação espacial na modelagem para uma representação mais precisa da dinâmica dos índices de democracia ao longo do tempo e espaço.

No que diz respeito à variável "Predominância do Sexo Feminino", constatamos que aproximadamente 60% dos países exibem uma maior proporção de população feminina no período entre os anos de 2003 e 2019.

Temos na Figura 4.2 os histogramas de duas covariáveis utilizadas no ajuste do modelo de regressão para o ano de 2003. O Painel (a) apresenta a distribuição do PIB *per capita*/hora. Notamos que a grande maioria dos países está na faixa de 0 a 2 dólares americanos e que apenas uma pequena minoria está acima de 6 dólares. O Painel (b) exibe o histograma da covariável "logaritmo da densidade populacional". Observamos valores entre -2 e 10, com maior concentração entre 1 e 7 e moda em torno de 4. Os dois gráficos indicados na Figura 4.2 sugerem uma escala razoável para utilização em um modelo de regressão, permitindo estimação de coeficientes sem a dificuldade de lidar com escalas muito grandes ou pequenas.

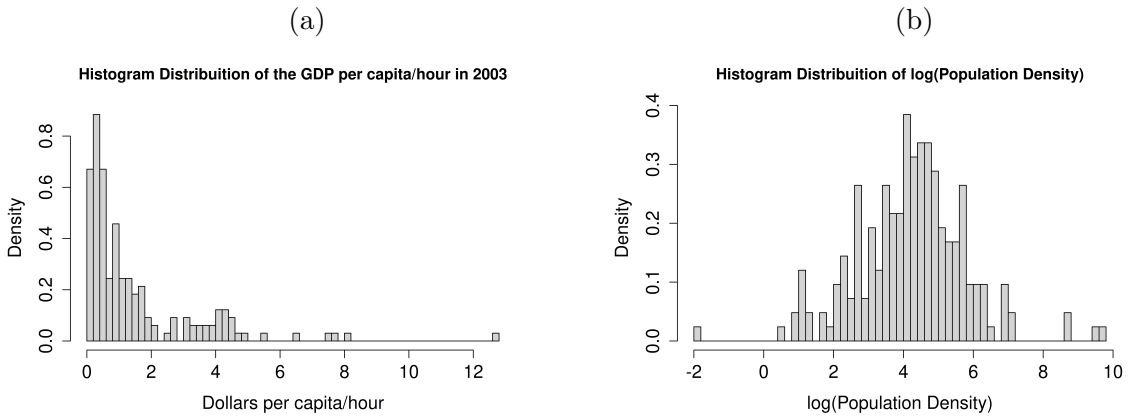


Figura 4.2: Histogramas das covariáveis PIB *per capita*/hora, Painel (a), e log(densidade populacional), Painel (b), para o ano de 2003.

Iremos seguir com o ajuste do modelo de regressão Bessel aos dados reais relacionados à aplicação proposta neste capítulo.

4.4 Definição das Vizinhanças dos Países Selecionados

Para introduzir a correlação espacial no modelo de regressão é preciso definir, primeiramente, as relações de vizinhança. Dentre todos os países, apenas 168 possuíam medidas de todas as covariáveis nos 17 anos entre 2003 e 2019. Com o intuito de definir uma matriz de vizinhança representativa das zonas de influência de cada país, consideramos a extensão territorial e a quantidade de vizinhos de cada um deles. A matriz de vizinhança foi construída levando em conta os 5 vizinhos mais próximos - utilizando a distância geodésica entre centroides - e a existência de fronteira entre dois territórios. A escolha do número de vizinhos mais próximos (k) foi feita de maneira empírica. A seleção de $k = 5$ gerou um grafo mais conexo, quando comparado com $k < 5$. Enquanto nos casos em que $k > 5$ houve a ocorrência de países com vizinhanças muito distantes, que, dado o contexto geográfico, não aparentavam ser relevantes. Entendemos que a matriz de vizinhança gerada a partir da combinação dos critérios captura de maneira mais satisfatória a influência geográfica de cada país nos demais, mesmo com as grandes variabilidades observadas em termos de extensão territorial e quantidade de vizinhos de fronteira. Assim, obtivemos uma matriz de vizinhança cujo o número de vizinhos de cada país está entre 5 e 16, sendo a China o país com mais relações de vizinhança, seguido da Rússia com 15.

A Figura 4.3 apresenta o mapa com a divisão política do mundo (Esri—HERE, 2022). Os pontos correspondem aos centroides de cada país, enquanto as linhas representam a relação de vizinhança criada a partir do esquema descrito anteriormente. O grafo gerado é desconexo, com dois subgrafos, o menor contém os países do continente americano e o maior contém os demais. As cores são representativas do *Índice de Democracia Eleitoral* do ano 2003. Tonalidades mais escuras representam índices mais próximos de 1, já aquelas mais claras denotam índices mais perto de 0.

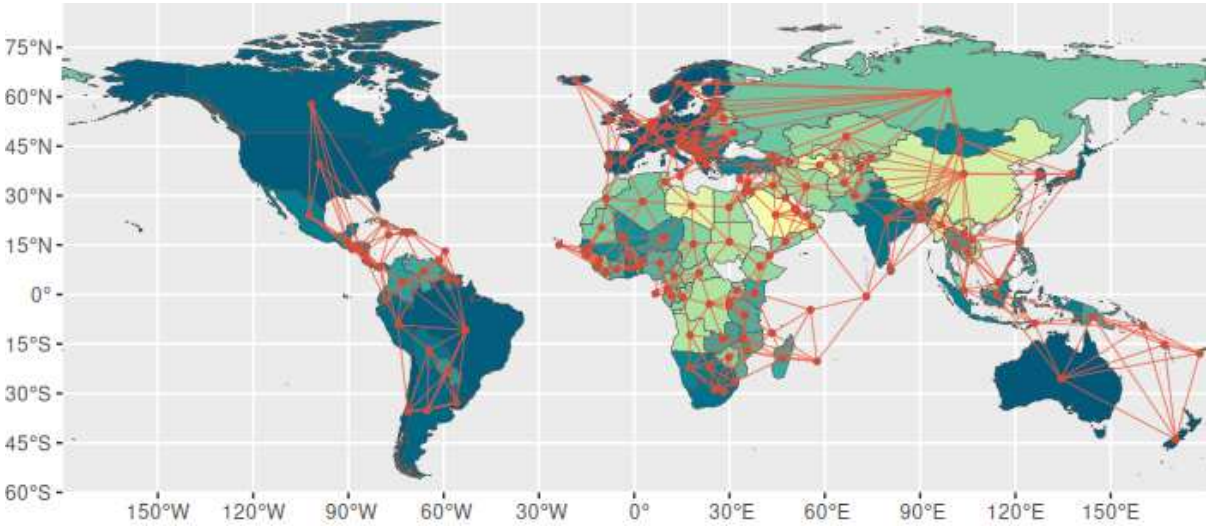


Figura 4.3: Mapa político global destacando os centroides dos países e suas relações de vizinhança. As cores refletem a variável resposta, *Índice de Democracia Eleitoral*, no ano de 2003. Observa-se uma tendência visual de países com tons mais escuros estarem frequentemente próximos, assim como países com tons mais claros, sugerindo uma relação espacial.

4.5 Ajuste do Modelo de Regressão Bessel

Nesta seção, apresentaremos os resultados obtidos a partir dos modelos de regressão Bessel ajustados ao banco de dados mencionado anteriormente. Os três modelos estudados nesta dissertação são identificados como M_1, M_2 e M_3 e apresentados no Capítulo 2. Optamos por analisar os resultados começando pelo modelo mais complexo até o mais simples, de modo que a apresentação dos modelos M_2 e M_3 será realizada conjuntamente, seguida pelo modelo M_1.

A principal diferença entre M_2 e M_3 é a forma como o efeito aleatório temporal é introduzido na estrutura de regressão. No M_2, esse efeito aleatório, representado por δ, é inserido de forma aditiva à estrutura de regressão. Por outro lado, o modelo M_3 incorpora uma estrutura de correlação na matriz de covariâncias dos coeficientes κ.

As Figuras 4.4 e 4.5 apresentam as estimativas obtidas (modelos M_2 e M_3) para os parâmetros γ e κ, respectivamente. Essas figuras fornecem uma visualização das estimativas obtidas, permitindo uma análise comparativa das diferentes estruturas de regressão utilizadas.

A Figura 4.4 tem quatro painéis, sendo que (a) e (b) representam os modelos M_2 e M_3, respectivamente. Estes gráficos indicam segmentos representando intervalos HPD de 95% para o efeito aleatório espacial. Uma linha cinza identificando o patamar zero foi incluída para orientação na avaliação. O círculo localizado dentro de cada intervalo é a mediana *a posteriori*. Perceba que vários intervalos não incluem o valor zero, sugerindo significância do parâmetro de efeito espacial de cada país no estudo. Visando obter uma análise comparativa entre estes resultados de M_2 e M_3, o Painel (c) apresenta a mesma informação exibida em (a) e (b), porém, os intervalos HPDs foram ordenados (decrescente) em relação à mediana *a posteriori* obtida para o M_2. Intervalos azuis são mostrados para M_2 e vermelhos para M_3. A similaridade entre os tamanhos e posicionamento dos segmentos sugere que ambos os modelos fornecem estimativas parecidas para o efeito espacial. Finalmente, no Painel (d) temos o mapa *mundi* onde as cores representam a distância absoluta entre os valores de γ estimados nos dois modelos, podemos notar que uma fração muito pequena dos países tem valores próximas de 1.4[1] (países com cores claras). Esse mapa demonstra, novamente, que a diferença na estimação de γ nos dois modelos é pequena.

Ao analisar a Figura 4.5, notamos que ambos os modelos apresentam uma relação positiva entre a variável resposta, o índice de tratamento de resíduos (intervalo 52) e a predominância do sexo feminino na população (intervalo 53). Os segmentos vermelhos representam M_2 e os azuis indicam M_3. O último intervalo exibido na Figura 4.5 representa o intercepto, κ_0. Note que a incerteza *a posteriori* é relativamente grande para este parâmetro em comparação com os demais. Essa incerteza é ligeiramente menor para M_3. É importante lembrar que no estudo simulado do Capítulo 3 foi detectada uma certa dificuldade para estimar o intercepto devido a uma maior comunicação entre este parâmetro e os efeitos aleatórios presentes na estrutura aditiva do preditor linear. Destacamos que apesar da dificuldade para estimar κ_0, os demais coeficientes não sofrem do mesmo problema, portanto, o pesquisador precisa estar mais cauteloso apenas na interpretação do intercepto. No entanto, há diferenças significativas nas estimativas dos demais coeficientes κ entre M_2 e M_3. Enquanto o M_3 estima $\kappa_1, \cdots, \kappa_{51}$ muito próximos de zero, com intervalos HPD (em azul) muito estreitos, o M_2 mostra uma relação negativa entre a variável resposta e os índices de poluentes do ar por ano (Segmentos 18 a 34) e uma relação positiva com a densidade demográfica (Segmentos 35 a 51). Além disso, os coeficientes relacionados ao PIB *per capita*/hora (Segmentos 1 a 17) são estimados muito próximos de zero, com intervalos HPD ligeiramente maiores do que aqueles em M_3.

[1]Este é o maior valor de distância obtido em todas as análises. Fixamos o limite superior em 1.4 de forma que todas as comparações fossem feitas na mesma escala.

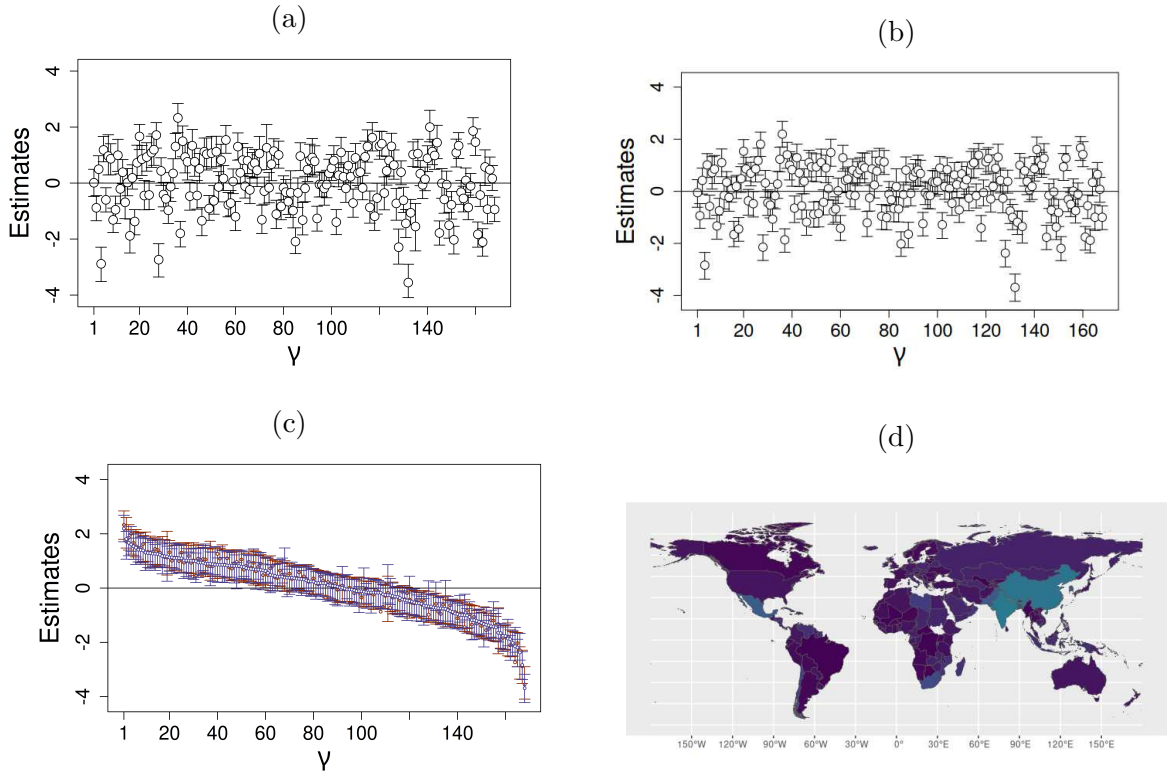


Figura 4.4: Medianas *a posteriori* (círculos) e intervalos HPDs (95%) do efeito aleatório espacial, γ, para M_2, Painel (a), e M_3, Painel (b). Este é um ajuste completo com as 5 covariáveis selecionadas. No Painel (c), todas as estimativas estão ordenadas pela ordem decrescente da mediana *a posteriori* de M_2; vermelho = M_2 e azul = M_3. O Painel (d) apresenta o mapa do mundo, as cores são a distância absoluta entre os valores de γ estimados nos dois ajustes, valores próximos de 0 estão com cor escura, valores próximos de $1,4$ estão com cor clara/esverdeada.

Com base na observação de que os coeficientes relacionados às covariáveis PIB *per capita*/hora e densidade populacional no modelo M_2 não são significativos, já que abrangem o zero em todos os anos analisados, é adequado ajustar novamente o M_2 sem a presença dessas covariáveis. Isso pode ajudar a simplificar o modelo e melhorar sua interpretação, concentrando-se nas covariáveis mais relevantes. Após remover as covariáveis não significativas, o novo M_2 será ajustado apenas com as variáveis restantes, aquelas consideradas significativas no ajuste anterior. Isso resulta em um modelo mais parcimonioso, de mais fácil interpretação e possivelmente mais adequado para a análise dos dados. Além disso, prosseguiremos com o ajuste do M_3 utilizando as mesmas variáveis selecionadas, permitindo uma comparação direta entre os modelos.

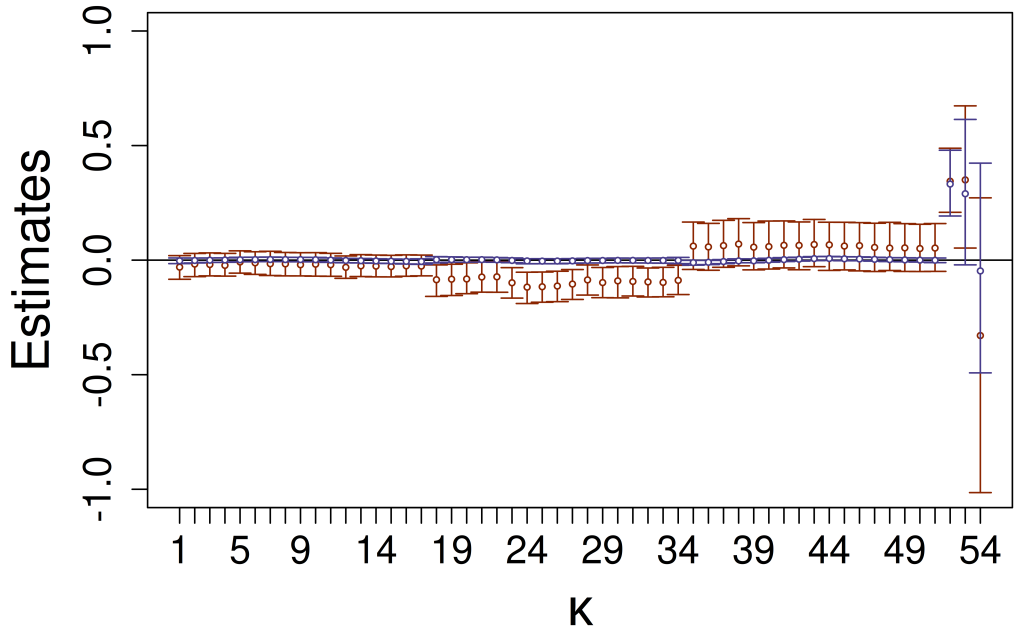


Figura 4.5: Medianas *a posteriori* e intervalos HPDs (95%) dos coeficientes de regressão, κ, para M_2 e M_3. Este é um ajuste completo com as 5 covariáveis selecionadas. M_2 = vermelho e M_3 = azul. "PIB *per capita*/hora" (Segmentos 1 a 17), "Índices de Poluentes do Ar por Ano" (Segmentos 18 a 34), "Densidade Demográfica" (Segmentos 35 a 51), "Índice de Tratamento de Resíduos" (Segmento 52), "Prevalência do Sexo Feminino" (Segmento 53) e κ_0 (Segmento 54).

Analisando a Figura 4.6, os ajustes feitos considerando apenas as três covariáveis significativas no modelo M_2 - "Índice de Qualidade do Ar por Ano", "Índice de Tratamento de Resíduos Sólidos" e "Predominância do Sexo Feminino" - resultaram em estimativas similares àquelas do M_3 com 3 covariáveis para o efeito aleatório espacial. As mesmas interpretações que foram realizadas comparando M_2 e M_3 na Figura 4.4(c) podem ser feitas também para os Painéis (c) e (d) da Figura 4.6. Em resumo, temos grande similaridade nas magnitudes e posição dos intervalos HPDs.

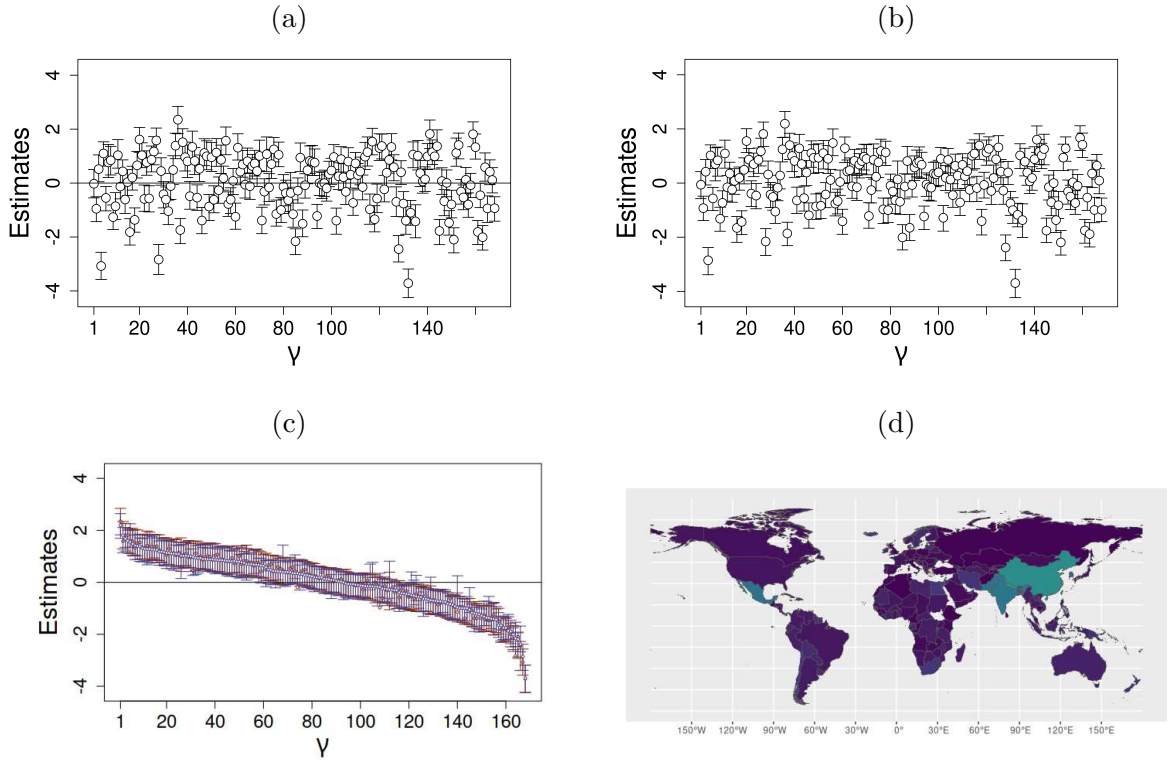


Figura 4.6: Medianas a *posteriori* (círculos) e intervalos HPDs (95%) do efeito aleatório espacial, γ, para M_2, Painel (a), e M_3, Painel (b). Este é um ajuste com as 3 covariáveis selecionadas. No Painel (c), todas as estimativas estão ordenadas pela ordem decrescente da mediana *a posteriori* de M_2; vermelho = M_2 e azul = M_3. O Painel (d) apresenta o mapa do mundo, as cores são a distância absoluta entre os valores de γ estimados nos dois ajustes, valores próximos de 0 estão com cor escura, valores próximos de $1,4$ estão com cor clara/esverdeada.

O mapa com os valores estimados de γ no M_2 ajustado com 3 covariáveis é apresentado na Figura 4.7. Como já havia sido indicado pelo teste da estatística I de Moran, percebemos a presença do efeito espacial no modelo. Países com valores mais altos de γ, cores escuras, tendem a ser próximos. O mesmo vale para países com valores de γ relativamente pequenos (cores claras). Podemos perceber, principalmente, que os países do oeste europeu têm cores escuras, assim como a maior parte da América. Há uma divisão clara entre os países do leste europeu e aqueles do oeste. O leste da Europa apresenta tonalidades mais claras, parecidos com os países do Oriente Médio, África e Ásia Central. Essas estimativas de γ são parecidas com o que temos para os valores da variável resposta para cada ano, com a vantagem de que esse é um valor que leva em consideração os dados para todo o período analisado. Portanto o γ cumpre o seu papel de capturar a estrutura espacial do modelo, conforme esperado.

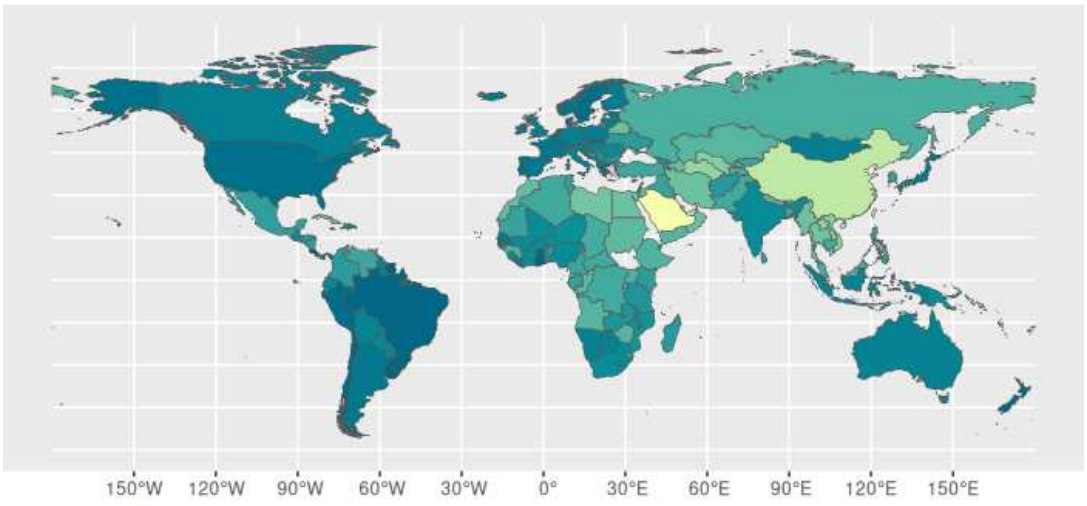


Figura 4.7: Mapa do mundo, as cores são a distância absoluta entre os valores de γ estimados no ajuste M_2 com 3 covariáveis, valores próximos de 0 estão com cor clara, valores mais altos estão com cor escura.

A Figura 4.8 exibe as estimativas dos coeficientes κ resultantes do ajuste com três covariáveis. Os resultados confirmam que as covariáveis selecionadas são estatisticamente significativas no modelo M_2, porém, não no modelo M_3, como destacado na própria Figura 4.8. É importante observar que os intervalos HPD para o modelo M_3 são notavelmente mais estreitos e tendem a se centralizar em torno do valor zero.

Além disso, nota-se que apesar da incerteza na estimativa de κ_0 (Segmento 20) ser mais acentuada, o modelo M_2 com três covariáveis oferece um intervalo de confiança menor do que aquele obtido no modelo M_2 com cinco covariáveis. É importante ressaltar que a interpretação das covariáveis "Índice de Tratamento de Resíduos"(Segmento 18) e "Predominância do Sexo Feminino"(Segmento 19) não sofreu alterações.

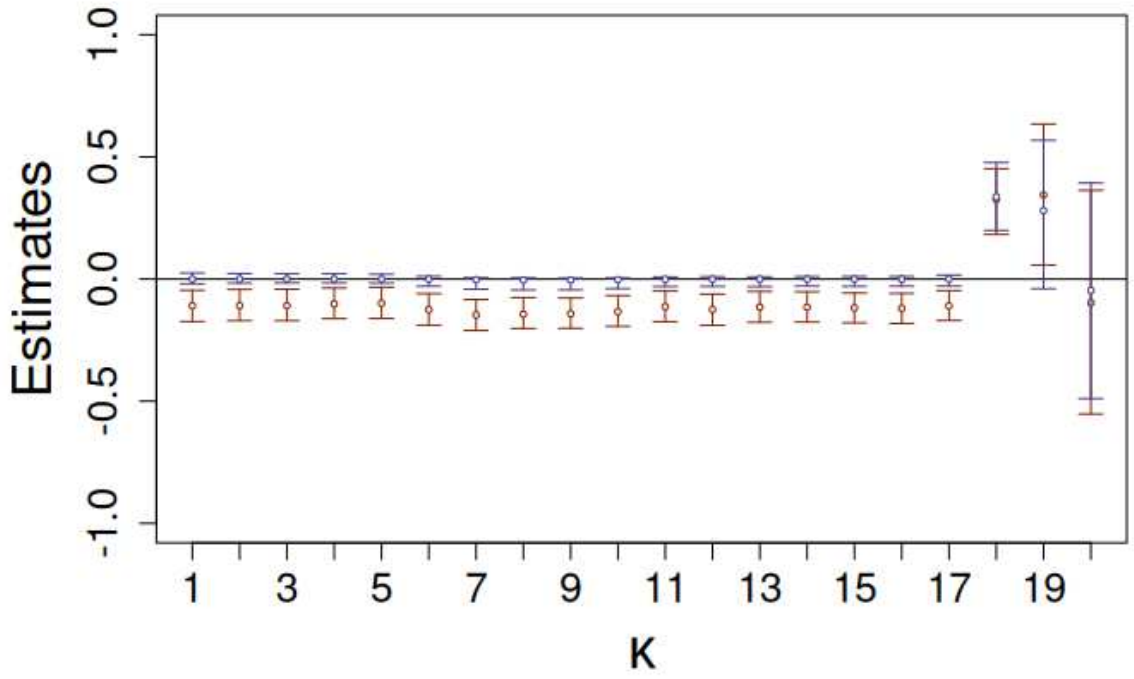


Figura 4.8: Medianas *a posteriori* e intervalos HPDs (95%) dos coeficientes de regressão, κ, para M_2 e M_3. Este é um ajuste com as 3 covariáveis selecionadas. M_2 = vermelho e M_3 = azul. "Índices de Poluentes do Ar por Ano" (Segmentos 1 a 17), "Índice de Tratamento de Resíduos" (Segmento 18), "Prevalência do Sexo Feminino" (Segmento 19) e κ_0 (Segmento 20).

No modelo M_2 também consideramos o efeito aleatório temporal, representado por δ. Os resultados desse efeito, tanto para o modelo completo quanto para aquele contendo apenas as três covariáveis significativas, são apresentados na Figura 4.9. Observamos que ambos os modelos exibem um padrão ascendente nos valores estimados, seguido por uma queda. No entanto, há uma diferença significativa no ponto de ápice desse padrão entre os Painéis (a) e (b): no modelo completo, ocorre no 14^o ano da série, enquanto que no modelo com as covariáveis selecionadas, ocorre no 10^o ano da série. Esse padrão pode ser atribuído a fatores geopolíticos que não foram considerados na análise. O período após o 10^o ano da série abrange os anos de 2013 a 2019, em que crises internacionais como os protestos no Oriente Médio em 2011 e guerras no leste europeu causaram um volume grande de migrações nos anos seguintes, além da intensificação da crise econômica mundial a partir de 2012. Não podemos afirmar com total certeza, mas esses processos poderiam ter uma relação que ajudaria explicar a tendência de queda nos valores de δ no período final da série.

Além disso, é notável que no modelo com três covariáveis, os intervalos HPD de 95% dos dois primeiros anos não abrangem o zero, enquanto no modelo completo, todos os intervalos contêm o zero. Essa discrepância é resultado do fato de que, no modelo com menos covariáveis, os intervalos HPD são consideravelmente menores. Esse aspecto é relevante, pois indica uma menor incerteza nas estimativas com menos covariáveis, tornando a modelagem mais decisiva na identificação de padrões temporais.

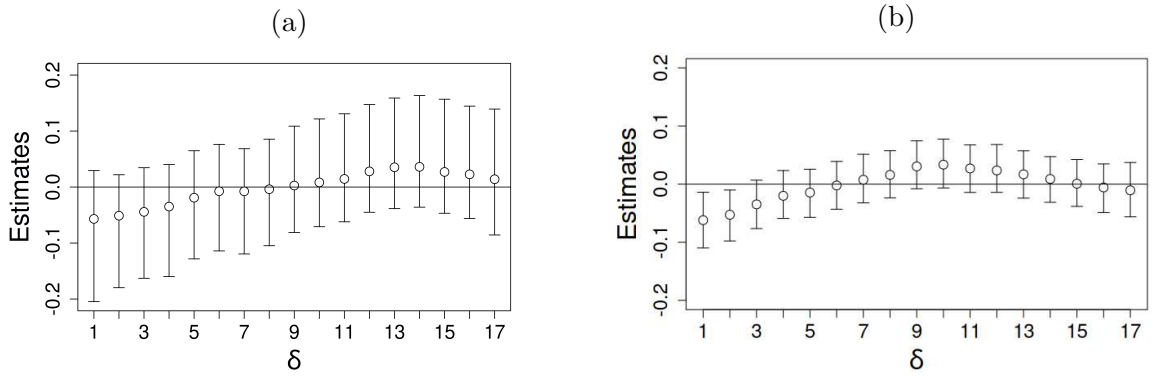


Figura 4.9: Segmentos representando intervalos HPD de 95% e medianas *a posteriori* (pontos) para o efeito aleatório temporal, δ. Painel (a) refere-se ao M_2 com 5 covariáveis. Painel (b) indica M_2 com 3 covariáveis.

Em termos das estimativas de ϕ, τ_δ (apenas para M_2), τ_κ (apenas para M_3) e τ_γ, a Tabela 4.1 apresenta os valores estimados de cada uma dessas variáveis em cada modelo. Notamos estimativas muito semelhantes em todos os modelos. Os parâmetros de variabilidade τ_δ e τ_κ foram estimados com valores positivos pequenos e próximos de zero, o que sugere a ausência do efeito temporal tanto em M_2 quanto em M_3, independente das covariáveis utilizadas. Por outro lado, a estimativa do parâmetro τ_γ é positiva e muito semelhante nos quatro ajustes, sendo maior no M_2 completo (4.88) e menor no M_3 com três covariáveis (4.55). No que diz respeito ao parâmetro de dispersão ϕ, novamente obtivemos resultados semelhantes em todos os modelos. Os ajustes realizados com o M_2 sempre fornecem estimativas maiores do que aquelas obtidas via M_3. A diferença de magnitude é bastante leve, portanto, não há evidências de que um modelo está se comportamento de forma muito distinta na inferência destes parâmetros.

Tabela 4.1: Mediana *a posteriori* dos parâmetros ϕ, τ_δ, τ_κ e τ_γ para M_2 e M_3. As estimativas de τ_δ e τ_κ são valores pequenos, abaixo de 0.01.

	ϕ		τ_δ	τ_κ	τ_γ	
Modelo	M_2	M_3	M_2	M_3	M_2	M_3
5 covariáveis	46.71	46.35	< 0.01	< 0.01	4.88	4.62
3 covariáveis	46.97	45.91	< 0.01	< 0.01	4.85	4.55

Agora, avançaremos com a análise do modelo simplificado M_1. Como mencionado anteriormente, o M_1 simplifica as covariáveis com medições ao longo do tempo usando suas médias como representação única no modelo de regressão. Dessa forma, as variáveis "Índice de Contaminação do Ar por Ano", "PIB *per capita*/hora" e "Densidade Populacional" serão representadas por um único valor para cada sítio neste modelo.

A Figura 4.10 demonstra que as estimativas de γ obtidas via M_1 são consistentes com aquelas obtidas via M_2 e M_3 até o momento. Por outro lado, as estimativas de κ, Painel (b), são bastante similares àquelas obtidas para M_2 com cinco covariáveis, em que apenas a covariável "Tratamento de Resíduos" (Segmento 4) é significativa. Também notamos a mesma dificuldade na estimativa do intercepto do modelo (Segmento 6), que apresenta um intervalo HPD de 95% mais amplo que os demais coeficientes. No entanto, todos os coeficientes κ indicam uma correlação positiva entre as covariáveis e a resposta "Índice de Democracia Eleitoral". Isto não ocorre nos demais modelos, em que as covariáveis "PIB *per capita*/hora" (Segmento 1) e "Índice de Poluentes do Ar por Ano" (Segmento 2) apresentavam medianas *a posteriori* com valores negativos, sugerindo uma relação negativa entre elas e a resposta. Diante destes resultados, podemos perceber que há uma modificação na interpretação do impacto existente entre covariáveis e a resposta quando o M_1 é usado. Entretanto, perceba que as estimativas não se mostraram significativas indicando que o sinal não tem forte importância. Utilizamos o critério de que uma covariável é significativa quando o intervalo HPD de 95% não abrange o zero. Neste caso, apenas a covariável "Índice de Tratamento de Resíduos" é dita significativa na modelagem simplificada dada por M_1. O mapa com apresentado no Painel (d) também é consistente com os resultados que obtivemos nos demais modelos.

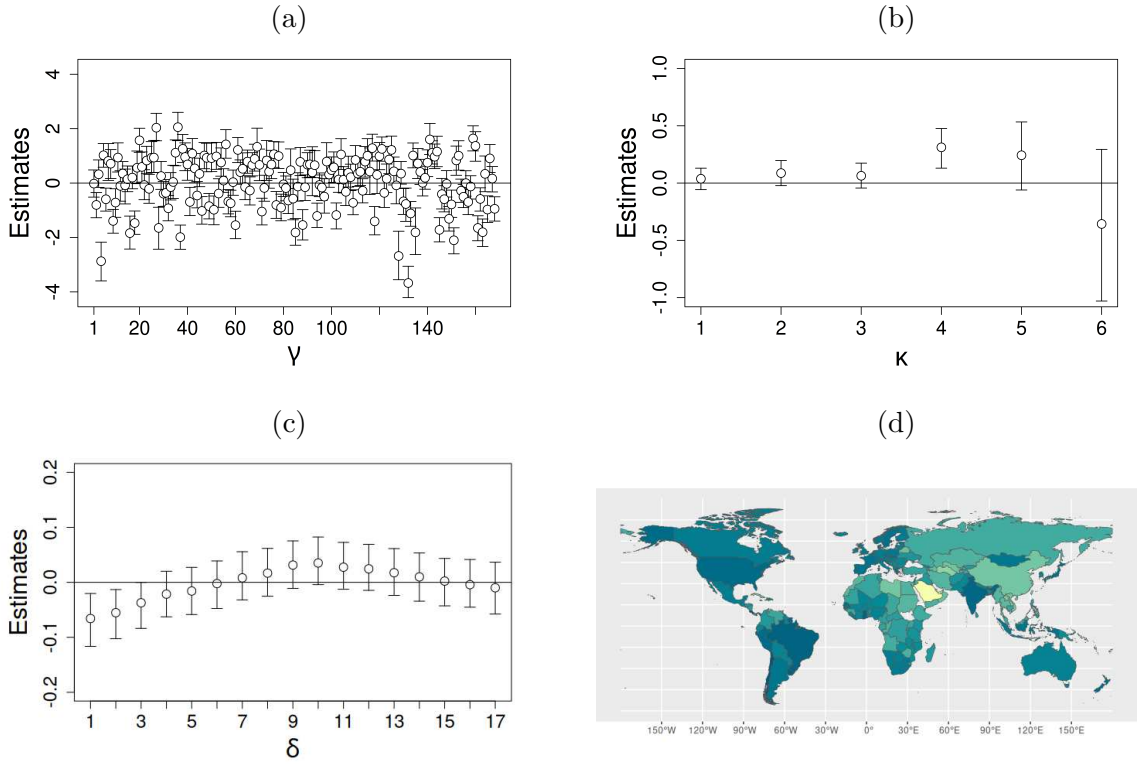


Figura 4.10: Medianas *a posteriori* e intervalos HPDs (95%) para os parâmetros do M_1 com 5 covariáveis. Painel (a): Efeito aleatório espacial, γ. Painel (b): Coeficientes de regressão, κ, "PIB *per capita*/hora" (Segmento 1), "Índice de Poluentes do Ar por Ano" (Segmento 2), "Densidade Demográfica" (Segmento 3), "Índice de Tratamento de Resíduos" (Segmento 4), "Prevalência do Sexo Feminino" (Segmento 5) e κ_0 (Segmento 6). Painel (c): Efeito aleatório temporal δ. O Painel (d) apresenta o mapa do mundo, as cores são os valores estimados de γ para esse ajuste, cores claras representam γ's menores enquanto valores mais altos estão associados a cores escuras.

A Figura 4.11 apresenta o mapa com a distância absoluta entre os valores de γ estimados nos dois ajustes. Podemos notar que que os países onde há maior similaridade na estimação (cores escuras) predominam, assim como nas demais comparações feitas. Contudo, nesse caso temos um maior número de tons claros, e países com coloração verde/amarela, o que indica uma maior diferença na estimação do parâmetro. Podemos concluir, portanto, que apesar de obtermos um ajuste aparentemente adequado com o resumo das covariáveis feito no M_1, há diferenças significativas na estimativa do γ associado a alguns países, principalmente aqueles onde o índice de democracia é relativamente baixo.

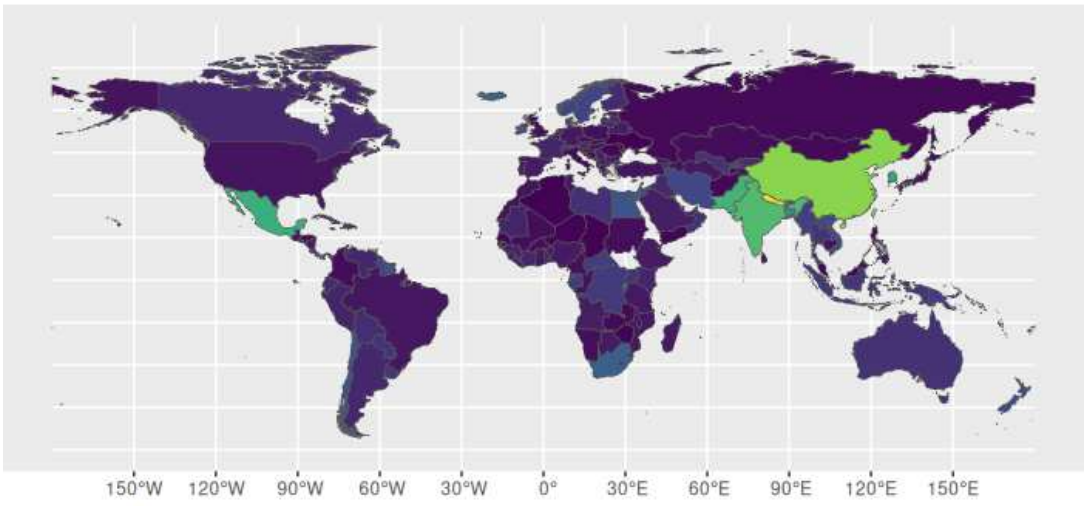


Figura 4.11: Mapa do mundo, as cores são a distância absoluta entre os valores de γ estimados nos dois ajustes, M_2 com 3 covariáveis e M_1 com 5 covariáveis, valores próximos de 0 estão com cor escura, valores próximos de $1,4$ estão com cor clara/esverdeada.

Os demais parâmetros também foram estimados com valores próximos àqueles apontados na Tabela 4.1, com $\phi = 46.34$, $\tau_\delta < 0.01$ e $\tau_\gamma = 4.45$. Portanto, a estimativa de τ_γ obtida via M_1 é a menor dentre todos os modelos ajustados. Isto implica que as associações entre as observações de regiões adjacentes tem sua força mais reduzida no M_1. Lembre que o parâmetro τ_γ multiplica a matriz $[D_{w_\gamma} - \rho_\gamma W_\gamma]^{-1}$, a qual contém a informação de covariâncias (entre vizinhos) e variâncias (de cada região) no espaço.

Dessa forma, embora o modelo M_1 seja mais simples, o M_2 com três variáveis parece trazer um ajuste com informações mais interessantes e tendo menos coeficientes "indecisos" no sentido de apresentarem HPD contendo o zero. De certa forma, podemos falar mais sobre evidências de impactos na resposta "Índice de Democracia Eleitoral" quando avaliamos o M_2 com 3 covariáveis. O M_3 não parece ser o mais adequado para esses dados, pois não consegue capturar de forma decisiva em termos de incerteza (HPD incluindo zero) a informação temporal das covariáveis na estrutura de correlação dos coeficientes κ. Em resumo, esta análise sugere usar o M_2 com três covariáveis como aquele que traz interpretações mais enfáticas sobre impactos.

Pela Equação 3, podemos analisar o ajuste do M_2 com três covariáveis em termos da razão de chances de o país ser democracia baseando-nos na média. A maioria das covariáveis do modelo toma valores entre -6 e 2, com valores predominantemente em

torno de 0 entre -2 e 2. Sendo assim, avaliamos uma mudança de 0.1 em cada covariável do modelo, exceto para a variável "Prevalência do Sexo Feminino"onde a mudança foi de 1, os resultados são apresentados na Tabela 4.2. Podemos perceber que, individualmente, cada um dos κ's associados com a variável "Índice de Poluentes do Ar por Ano"tem influência entre 1% e 1.5% de redução sobre a *odds* de democracia, fixadas as demais variáveis. A influência da covariável binária é de cerca de 41% de aumento sobre a *odds*, quando passamos da categoria base, onde há prevalência do sexo masculino na população, para o caso onde o sexo feminino predomina. Já a variável uniforme de medida única (κ_{18}) apresenta um aumento de cerca de 3.3% sobre a *odds*.

Tabela 4.2: Porcentagem de aumento ou redução sobre a *Odds* de democracia, fixando as demais variáveis, aumentando a variável correspondente em 0.1 unidade. No caso de κ_{19} o valor representa a porcentagem de aumento quando mudamos da categoria base para a categoria de interesse.

κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7	κ_8	κ_9	κ_{10}
-1.1%	-1.1%	-1.1%	-1.0%	-1.0%	-1.3%	-1.5%	-1.4%	-1.4%	-1.3%
κ_{11}	κ_{12}	κ_{13}	κ_{14}	κ_{15}	κ_{16}	κ_{17}	κ_{18}	κ_{19}	-
-1.1%	-1.2%	-1.1%	-1.2%	-1.2%	-1.2%	-1.1%	3.3%	41.1%	-

Conclusão do Capítulo

Neste capítulo, foram apresentados os dois bancos de dados para os ajustes de M_1, M_2 e M_3. O M_2 (com 3 covariáveis) é aquele que mostrou resultados mais interessantes. Esse ajuste nos fornece informações relevantes sobre a variação da correlação temporal da resposta no período, além de identificar de forma coerente a relação da resposta com covariáveis ambientais e sociais. O ajuste M_2 apresenta valores negativos dos coeficientes associados à covariável *"Índices de Poluentes do Ar por Ano"*. Com a função de ligação *logit* temos uma relação negativa entre esta variável e a resposta, ou seja, quanto maior o valor da covariável, mais próximo de zero será a média da resposta. Os coeficientes das variáveis *"Índice de Tratamento de Resíduos"* e *"Prevalência do Sexo Feminino"* são positivos, implicando que quanto maior o valor dessas variáveis mais próximo de 1 será a média da resposta. Para o coeficiente do intercepto obtivemos uma estimativa muito próxima de zero, mas ligeiramente negativa. Em termos dos efeitos temporais, observamos uma tendência crescente entre os anos de 2003 e 2012, seguido de um período de decrescimento entre 2013 e 2019, o que pode estar relacionado com instabilidades geopolíticas ocorridas no mundo a partir de 2012. Essa análise indica coerência das estimativas do efeito aleatório temporal em M_2 com 3 covariáveis.

Capítulo 5

Conclusões

Este trabalho é focado na avaliação do modelo de regressão Bessel para resposta contínua, limitada e observada para diferentes locais e tempo. A regressão Bessel é uma alternativa ao modelo beta. Os autores em Barreto-Souza et al. (2021) argumentam que o modelo Bessel é uma alternativa mais flexível para este tipo de resposta. A grande diferença entre o que foi feito por esses autores e a proposta desta dissertação é o desenvolvimento da modelagem espaço-temporal sob o ponto de vista Bayesiano que apresentamos. Quando lidamos com este tipo de modelagem, que inclui efeitos aleatórios correlacionando tempo e espaço, utilizar o paradigma Bayesiano para inferência é mais acessível do que desenvolver uma análise frequentista, baseada em máxima verossimilhança.

O trabalho foi motivado por um banco de dados reais envolvendo o índice de democracia eleitoral para diferentes países do mundo. Ao longo deste trabalho descrevemos com detalhes os modelos propostos para avaliação de dados dessa natureza. Uma diferença crucial entre esses modelos é a forma na qual o efeito temporal é introduzido, que pode ser via efeito aleatório δ_t, modelos M_1 e M_2, ou estabelecido por meio de uma associação entre coeficientes de regressão (κ), modelo M_3. Tivemos o cuidado de não misturar estes dois formatos em uma mesma modelagem, pois isto implicaria na situação de inclusão de dependência temporal duplicada por partes distintas do modelo.

Após essa apresentação, o texto explicou como é feita a geração de dados artificiais e estabeleceu cenários de análises considerando combinações de modelos geradores e de ajuste. Resumidamente, o M_2 (com efeito aleatório espacial e efeito temporal) apresentou maior robustez em termos de má-especificação com Vícios Relativos (VRs) menores do que o M_3 (com efeito aleatório espacial e coeficientes correlacionados no tempo) para a grande maioria dos parâmetros em casos nos quais o gerador e o modelo ajustado não eram iguais. Contudo, o M_2 apresenta maior viés na estimação do intercepto, chamado aqui de κ_0. Já M_1 é uma boa alternativa para M_2, por ser mais parcimonioso ao resumir a informação das covariáveis medidas ao longo do tempo. Esse resumo não acarretou em grandes perdas na estimativa dos demais parâmetros comuns em relação ao M_2. Pelo

contrário, apresentou uma ligeira melhora no VR associado ao intercepto. Entretanto, perceba que aplicar o M_1 implica em uma simplificação que reduz o nível de detalhes e informação que o M_2 é capaz de fornecer com seus coeficientes atrelados a cada covariável de tempos diferentes.

No que diz respeito à estimação do modelo completo, quando o gerador não contém o efeito temporal ou o espacial, as três abordagens se mostraram capazes de recuperar os demais parâmetros sem dificuldade na estimação, muitas vezes com VRs menores que no caso bem-especificado. Em contrapartida, o caso reverso, em que o modelo ajustado não estima um dos efeitos mas o gerador é completo, há grandes perdas. Temos aqui um aumento dos VRs e até mesmo casos em que eles não estão centrados em zero, como era de se esperar. Dessa forma, sugerimos sempre ajustar o modelo completo, mesmo com a possibilidade de que não exista correlação espacial ou temporal nos dados.

Após o estudo das propriedades dos modelos por meio de simulações com réplicas MC, eles foram aplicados aos dados reais que motivaram este trabalho. As variáveis foram obtidas de dois bancos de dados distintos, sendo a variável resposta o *"Índice de Democracia Eleitoral"* extraído do banco de dados *V-Dem*. As cinco variáveis explicativas foram obtidas a partir da fonte de dados *"Environmental Performance Index"* (EPI). O período selecionado para a análise abrangeu os anos de 2003 a 2019 e 168 países.

Dentre os três modelos, o M_2 mostrou resultados interessantes com maior quantidade de coeficientes significativos (referentes a 3 das 5 variáveis explicativas) e identificação de efeito temporal também significativo. O "Índice de Poluentes do Ar por Ano" apresenta uma correlação negativa com a variável resposta, enquanto o "Índice de Tratamento de Resíduos" e a "Prevalência do Sexo Feminino" estão positivamente relacionados com a resposta. Assim como na simulação, observa-se novamente no caso real a dificuldade do M_2 em ajustar o intercepto, indicada pela grande amplitude do intervalo HPD de 95%. Em relação ao efeito δ, o M_2 revela uma tendência crescente até o 10^o ano da série, em 2012, seguida por um declínio até o último ano, em 2019. Essa tendência parece ser coerente quando consideramos as diversas instabilidades políticas e econômicas que surgiram no mundo entre 2013 e 2019, como a intensificação dos conflitos no Oriente Médio que culminaram em uma onda migratória de refugiados e os desdobramentos da crise econômica mundial a partir de 2012. No geral, o M_2 permite realizar inferências coerentes (dentro do esperado) sobre a evolução da democracia nos países entre 2003 e 2019.

Concluímos, portanto, que os três modelos de regressão Bessel espaço-temporais apresentados aqui tem aplicações únicas, dependendo da intenção do analista. Contudo, o M_2 aparenta ser o mais robusto (em termos de vício) sob má-especificação dos efeitos

aleatórios, o que o coloca como o modelo preferível, enquanto M_1 e M_3 são alternativas úteis quando há apoio de informação *a priori* sobre a estrutura dos dados.

A seguir, apresentamos algumas alternativas para trabalhos futuros que podem representar um caminho viável na continuação dos estudos da regressão Bessel com efeito espaço-temporal.

Trabalhos Futuros

- Comparação com a regressão beta em diferentes cenários. Note que o modelo beta Bayesiano também terá que ser implementado. Não existe um pacote que faça este ajuste na versão espaço-temporal que estamos propondo aqui.
- Os modelos apresentados podem ser estendidos por meio da introdução de covariáveis explicativas também conectadas ao parâmetro ϕ, de maneira semelhante ao que foi apresentado para μ.
- Determinação de um tipo de resíduo que seja apropriado para a regressão Bessel e assim permita avaliação de qualidade de ajuste dos modelos.
- Inclusão também dependência espacial na estrutura de covariância dos coeficientes do M_3.
- Avaliação de robustez dos modelos em dados com mais e menos tempos e sítios.

Bibliografia

Abramowitz, M. e Stegun, I. A. (1968), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55, US Government printing office.

Altun, E., El-Morshedy, M., e Eliwa, M. (2021), "A new regression model for bounded response variable: An alternative to the beta and unit-Lindley regression models." *Plos one*, 16, e0245627.

Areal, F. J., Balcombe, K., e Tiffin, R. (2012), "Integrating Spatial Dependence Into Stochastic Frontier Analysis." *Australian Journal of Agricultural and Resource Economics.*, 56, 521–541.

Banerjee, S., Carlin, B. P., e Gelfand, A. E. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC, Boca Raton.

Barndorff-Nielsen, O. E. e Jørgensen, B. (1991), "Some Pametric Models on the Simplex." *Journal of Multivariate Analysis*, 39, 106–116.

Barr, D. J., Levy, R., Scheepers, C., e Tily, H. J. (2013), "Random effects structure for confirmatory hypothesis testing: Keep it maximal." *Journal of Memory and Language*, 68, 255–278.

Barreto-Souza, W. e Simas, A. B. (2017), "Improving Estimation for Beta Regression Models Via EM-Algorithm and Related Diagnostic Tools." *Journal of Statistical Computation and Simulation*, 87, 2847–2867.

Barreto-Souza, W., Mayrink, V. D., e Simas, A. B. (2021), "Bessel Regression and bbreg Package to Analyse Bounded Data." *Australian & New Zealand Journal of Statistics*, 63, 685–706.

Bates, D., Maechler, M., e Jagan, M. (2022), *Matrix: Sparse and Dense Matrix Classes and Methods*, R package version 1.5-1, `https://CRAN.R-project.org/package=Matrix`.

Bayes, C. L., Bazán, J. L., e García, C. (2012), "A new robust regression model for proportions." *Baysian Analysis*, 7 (4), 841–866.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society: Series B*, 36, 192–225.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., e Riddell, A. (2017), "Stan: A probabilistic Programming Language." *Journal of Statistical Software*, 76, 1–32.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Alizada, N., Altman, D., Bernhard, M., Cornell, A., Fish, M. S., et al. (2022), "V-Dem Country-Year Dataset v12." .

Cressie, N. (2015), *Statistics for Spatial Data*, John Wiley & Sons, Hoboken.

Erdélyi, A. (1953), "Higher transcendental functions." *Higher transcendental functions*, p. 59.

Esri—HERE (2022), "World Countries Generalized." `https://hub.arcgis.com/datasets/esri::world-countries-generalized/about`, [Acessado em 16 de junho de 2023].

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1, 209–230.

Ferrari, S. e Cribari-Neto, F. (2004), "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics*, 31, 799–815.

Ghitany, M., Mazucheli, J., Menezes, A., e Alqallaf, F. (2019), "The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval." *Communications in Statistics-Theory and methods*, 48, 3423–3438.

Gómez-Déniz, E., Sordo, M. A., e Calderín-Ojeda, E. (2014), "The Log–Lindley distribution as an alternative to the beta regression model with applications in insurance." *Insurance: Mathematics and Economics*, 54, 49–57.

Goulet, V. (2016), *expint: Exponential Integral and Incomplete Gamma Function*, R package `https://cran.r-project.org/package=expint`.

Hauser, R. M. e Warren, J. R. (1997), "Socioeconomic Indexes for Occupations: A Review, Update, and Critique." *Sociological Methodology*, 27, 177–298.

Hegedüs, D. (2020), "Varieties of Democracy: Measuring Two Centuries of Political Change. By Michael Coppedge, John Gerring, Adam Glynn, Carl Henrik Knutsen,

Staffan I. Lindberg, Daniel Pemstein, Brigitte Seim, Svend-Erik Skaaning, and Jan Teorell." *Perspectives on Politics*, 18, 1258–1260.

Hoffman, M. D., Gelman, A., et al. (2014), "The No-U-Turn sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15, 1593–1623.

Kalhori, L. e Mohhamadzadeh, M. (2017), "Spatial Beta Regression Model with Random Effect." *Journal of Statistical Research of Iran*, 13, 215–230.

Kriesi, H., Bochsler, D., Matthes, J., Lavenex, S., Bühlmann, M., e Esser, F. (2013), *Democracy in the Age of Globalization and Mediatization*, Challenges to Democracy in the 21st Century, Palgrave Macmillan UK, London.

Krisztin, T. e Piribauer, P. (2021), "A Bayesian spatial autoregressive logit model with an empirical application to European regional FDI flows." *Empirical Economics*, 61, 231–257.

Kumaraswamy, P. (1980), "A Generalized Probability Density Function for Double-Bounded Random Processes." *Journal of Hydrology*, 46, 79–88.

Lemonte, A. J. e Bazán, J. L. (2016), "New class of Johnson distributions and its associated regression model for rates and proportions." *Biometrical Journal*, 58, 727–746.

Lindberg, S. I., Coppedge, M., Gerring, J., e Teorell, J. (2014), "V-Dem: A new way to measure democracy." *Journal of Democracy*, 25, 159–169.

Liu, B.-c. (1974), "Variations in the Quality of Life in the United States by State, 1970." *Review of Social Economy*, 32, 131–147.

Mayrink, V. D. e Gamerman, D. (2009), "On Computational Aspects of Bayesian Spatial Models: Influence of the Neighboring Structure in the Efficiency of MCMC Algorithms." *Computational Statistics*, 24, 641–669.

Mazucheli, J., Menezes, A. F., e Dey, S. (2018), "The unit-Birnbaum-Saunders distribution with applications." *Chilean Journal of Statistics*, 9, 47–57.

Mazucheli, J., Menezes, A. F. B., e Chakraborty, S. (2019), "On the one parameter unit-Lindley distribution and its associated regression model for proportion data." *Journal of Applied Statistics*, 46, 700–714.

McElreath, R. (2020), *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*, Chapman and Hall/CRC, Boca Raton.

Moran, P. A. (1950), "Notes on continuous stochastic phenomena." *Biometrika*, 37, 17–23.

Neuhaus, J. M. e McCulloch, C. E. (2006), "Separating between-and within-cluster covariate effects by using conditional and partitioning methods." *Journal of the Royal Statistical Society Series B*, 68, 859–872.

Neuhaus, J. M. e McCulloch, C. E. (2011), "Estimation of covariate effects in generalized linear mixed models with informative cluster sizes." *Biometrika*, 98, 147–162.

Queiroz, F. F. e Ferrari, S. L. (2023), "Power logit regression for modeling bounded data." *Statistical Modelling*, p. 1471082X221140157.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org/`.

Ravishanker, N., Chi, Z., e Dey, D. K. (2021), *A First Course in Linear Model Theory*, Chapman and Hall/CRC, Boca Raton.

Schielzeth, H. (2010), "Simple means to improve the interpretability of regression coefficients." *Methods in Ecology and Evolution*, 1, 103–113.

Simas, A. B., Barreto-Souza, W., e Rocha, A. V. (2010), "Improved Estimators for a General Class of Beta Regression Models." *Computational Statistics & Data Analysis*, 54, 348–366.

Stan Development Team (2023), "RStan: the R interface to Stan," R package version 2.32.3.

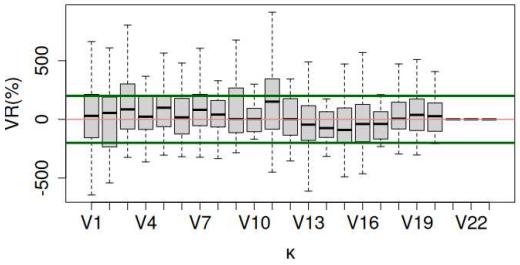
Wolf, M. J., Emerson, J. W., Esty, D. C., Sherbinin, A. d., e Wendling, Z. A. (2022), *2022 Environmental Performance Index (EPI) results*, Technical Report, Yale Center for Environmental Law & Policy, New Haven.

Apêndice A

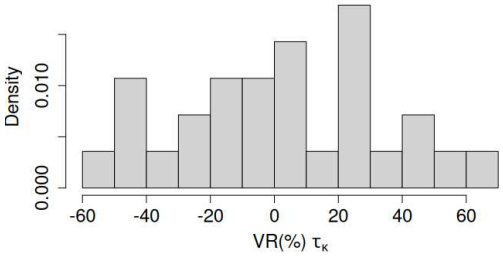
Gráficos dos VRs para Casos de Má-especificação de M_3

Modelo Gerador $M_3^{\oplus}$: Ajuste sem γ e com Dependência Temporal em κ.

(a)



(b)



(c)

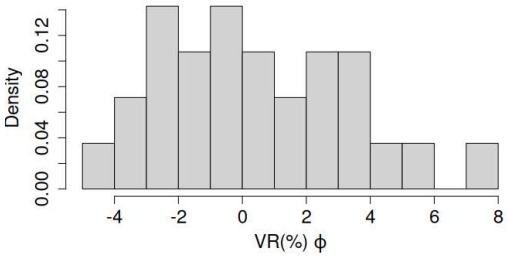


Figura A.1: *Boxplot* dos VRs dos coeficientes de regressão, κ, Painel (a). Histogramas dos VRs para os parâmetros de variabilidade τ_κ e ϕ, Painéis (b) e (c), respectivamente.

Modelo Gerador $M_3^{\gamma\ominus}$: Ajuste com γ e Dependência Temporal em κ.

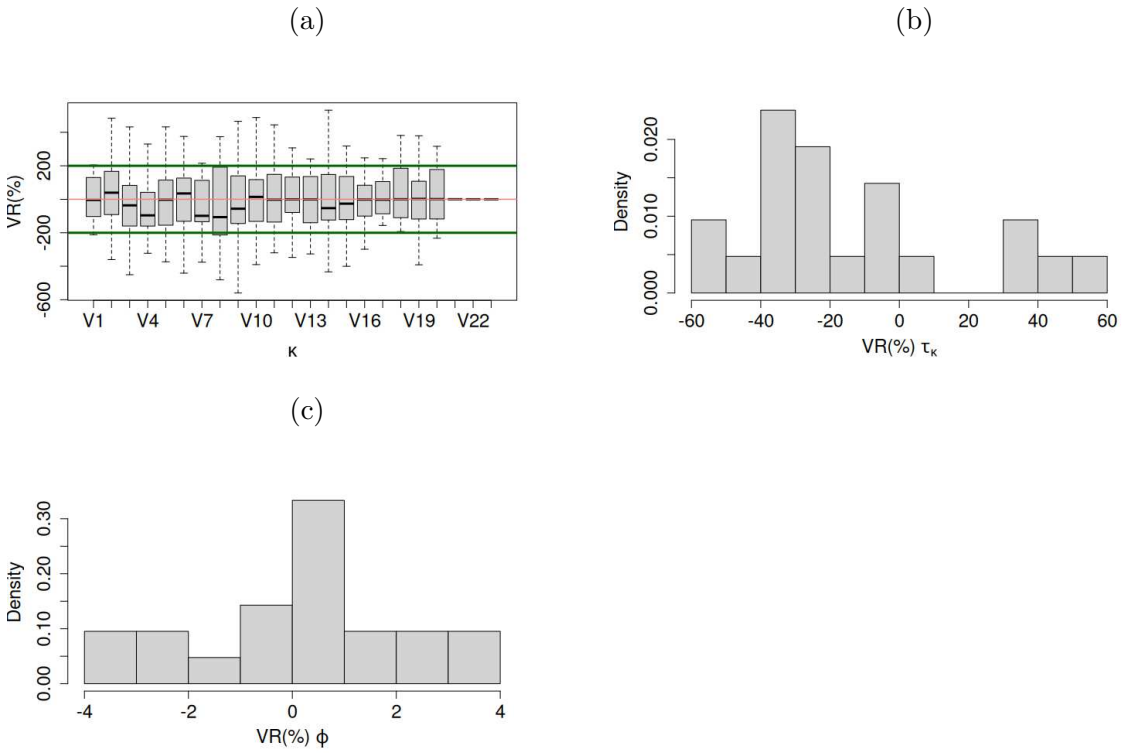


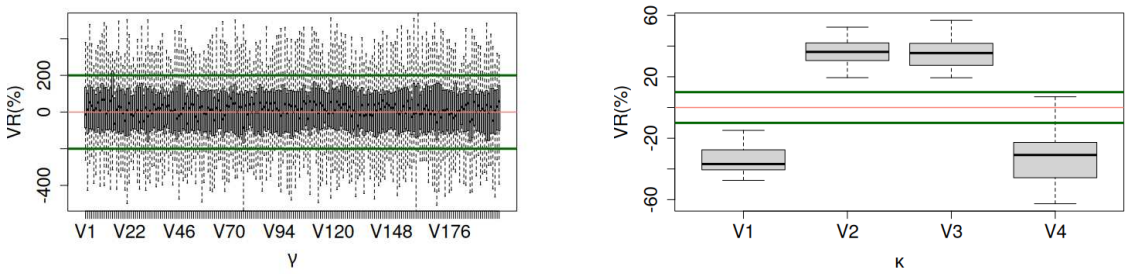
Figura A.2: *Boxplot* dos VRs dos coeficientes de regressão, κ, Painel (a). Histogramas dos VRs para os parâmetros de variabilidade τ_κ e ϕ, Painéis (b) e (c), respectivamente..

Apêndice B

Gráficos dos VRs para Casos de Má-especificação de M_1

Modelo Gerador $M_1^{\oplus}$: Ajuste sem δ e com γ.

(a) (b)



(c) (d)

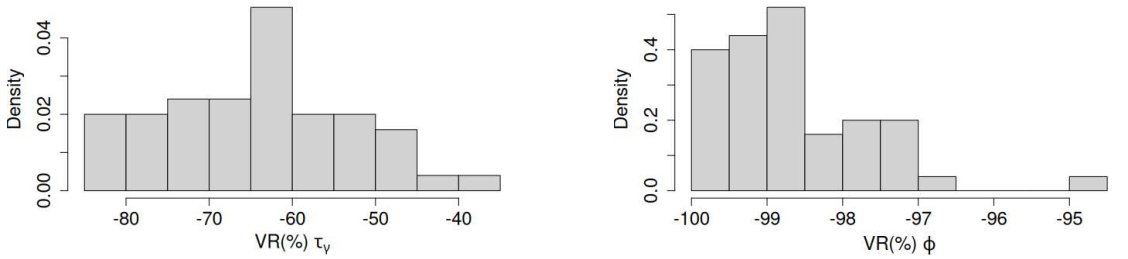


Figura B.1: Painel (a): *boxplots* dos VRs do efeito aleatório espacial, γ. Painel (b): *boxplots* dos VRs dos coeficientes da regressão, κ. Painel (c) e (d): histograma dos VRs de τ_γ e ϕ, respectivamente.

Modelo Gerador $M_1^{\oplus}$: Ajuste com δ e semγ.

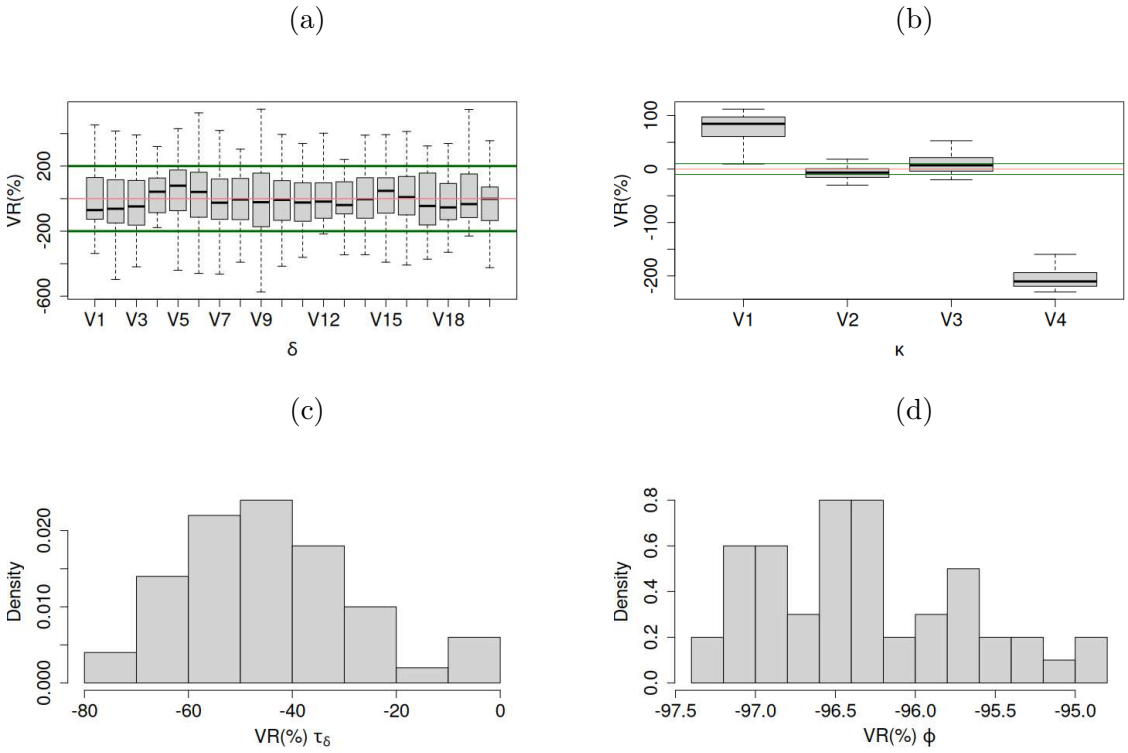


Figura B.2: Painel (a): *boxplots* dos VRs do efeito aleatório temporal, δ. Painel (b): *boxplots* dos VRs dos coeficientes da regressão, κ. Painel (c) e (d): histograma dos VRs de τ_δ e ϕ, respectivamente.

Modelo Gerador $M_1^{\delta\ominus}$: Ajuste com δ e γ.

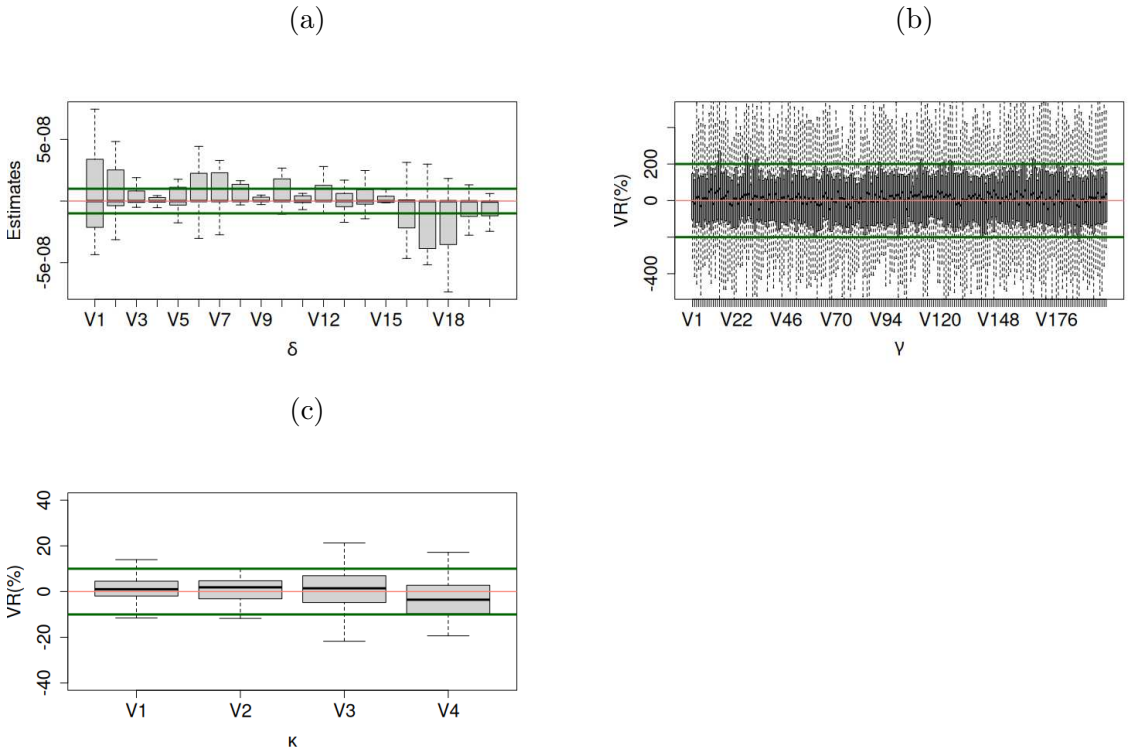


Figura B.3: *Boxplots* das medianas *a posteriori* do efeito aleatório temporal, δ, Painel (a), dos VRs das medianas *a posteriori* do efeito aleatório espacial, Painel (b), e dos coeficientes de regressão, Painel (c).

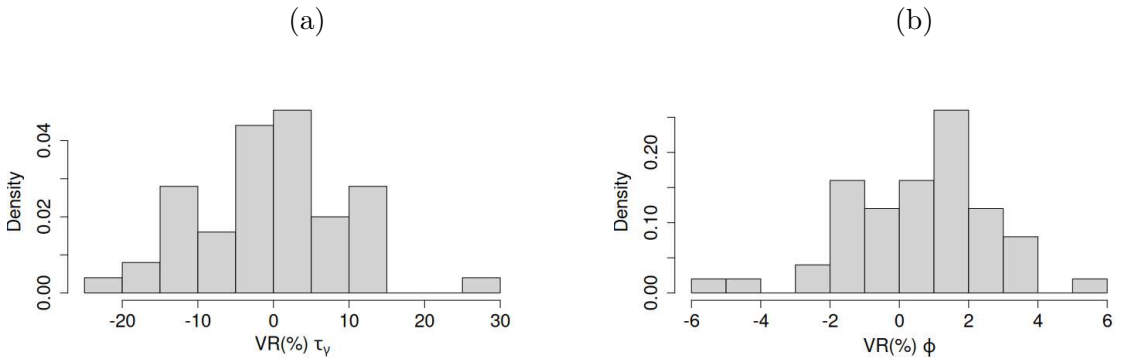


Figura B.4: Histogramas dos VRs das medianas *a posteriori* dos parâmetros de variância do modelo CAR do efeito aleatório espacial, τ_γ, Painel (a), e do parâmetro de dispersão da regressão Bessel, ϕ, Painel (b).

Modelo Gerador $M_1^{\gamma\ominus}$: Ajuste com δ e γ.

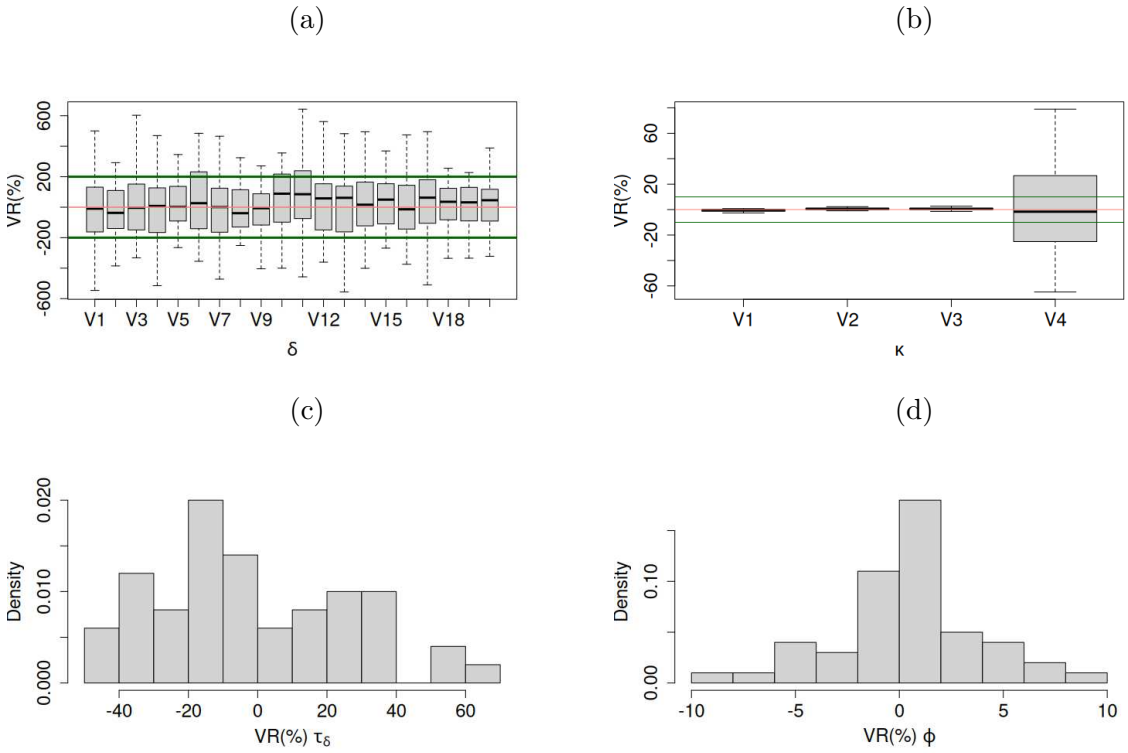


Figura B.5: Painel (a): *boxplots* dos VRs do efeito aleatório temporal, δ. Painel (b): *boxplots* dos VRs dos coeficientes da regressão, κ. Painel (c) e (d): histogramas dos VRs de τ_δ e ϕ, respectivamente.

Apêndice C

Gráficos dos Caminhos das Cadeias *à posteriori* Geradas pelo Stan

Modelo Gerador $M_2^{\oplus}$: Ajuste com δ e γ.

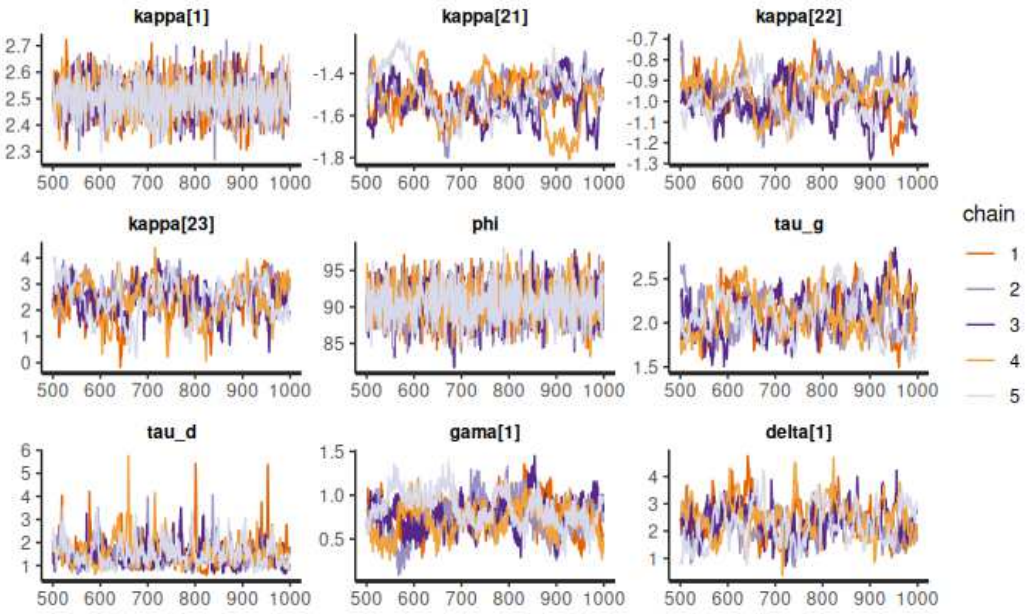


Figura C.1: Gráfico das cadeias amostradas pelo Stan para uma réplica da simulação. Para o modelo gerador $M_2^{\oplus}$ com ajuste com δ e γ. κ_1 corresponde ao primeiro ano da variável medida ao longo do tempo. $\kappa_2 1$, $\kappa_2 2$ e $\kappa_2 3$ estão relacionados, respectivamente, à covariável uniforme com apenas uma medida, à covariável binária e ao intercepto do modelo.

Modelo Gerador $M_2^{\oplus}$: Ajuste sem δ e com γ.

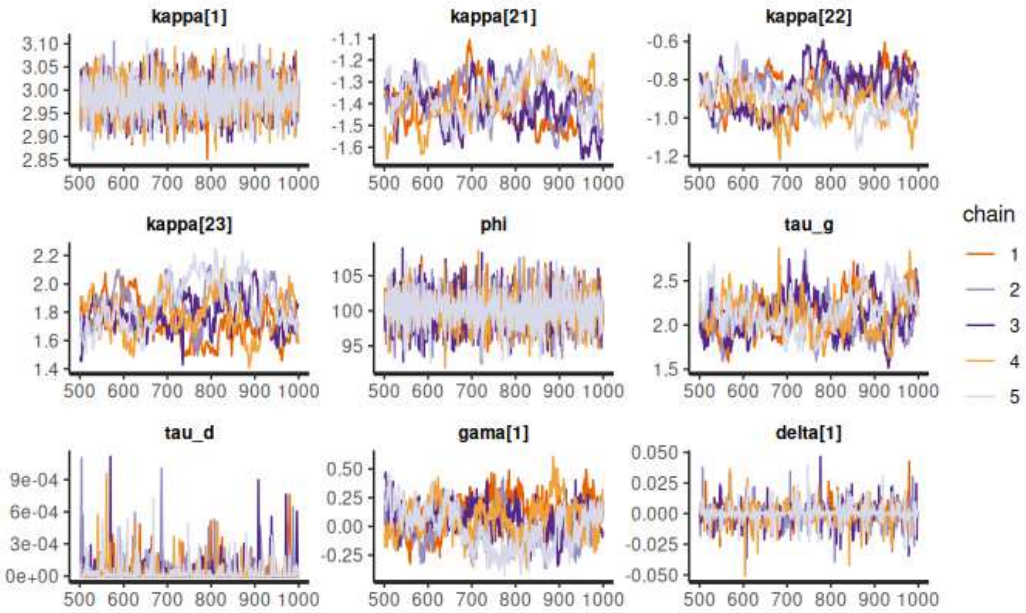


Figura C.2: Gráfico das cadeias amostradas pelo Stan para uma réplica da simulação. Para o modelo gerador $M_2^{\oplus}$ com ajuste sem δ e com γ. κ_1 corresponde ao primeiro ano da variável medida ao longo do tempo. $\kappa_2 1$, $\kappa_2 2$ e $\kappa_2 3$ estão relacionados, respectivamente, à covariável uniforme com apenas uma medida, à covariável binária e ao intercepto do modelo.

Modelo Gerador $M_2^{\delta\ominus}$: Ajuste com δ e γ.

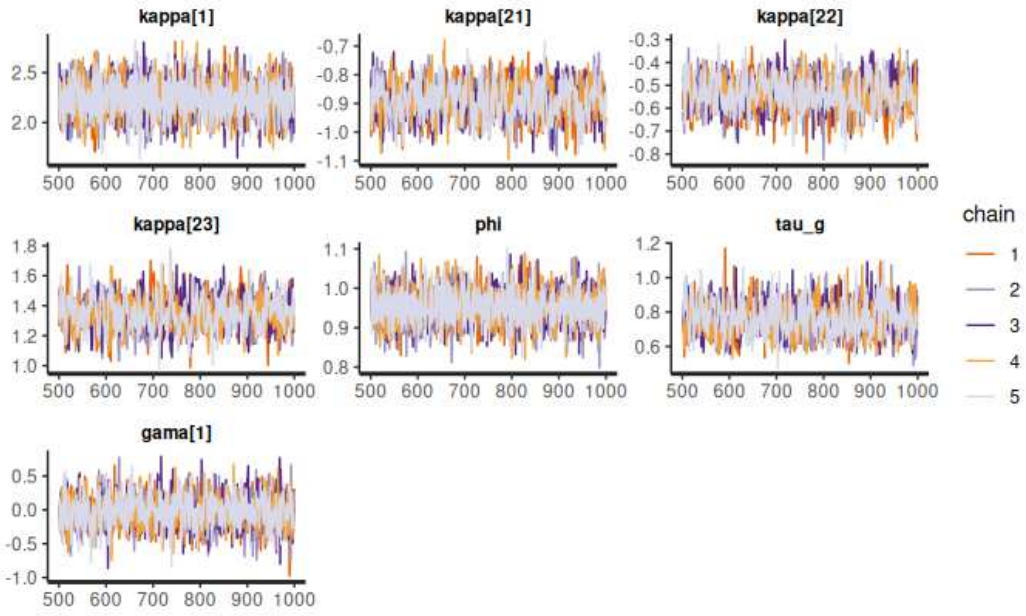


Figura C.3: Gráfico das cadeias amostradas pelo Stan para uma réplica da simulação. Para o modelo gerador $M_2^{\delta\ominus}$ com ajuste com δ e γ. κ_1 corresponde ao primeiro ano da variável medida ao longo do tempo. $\kappa_2 1$, $\kappa_2 2$ e $\kappa_2 3$ estão relacionados, respectivamente, à covariável uniforme com apenas uma medida, à covariável binária e ao intercepto do modelo.