

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Modelagem estatística da impunidade: métodos inferenciais e preditivos para dados de crimes no estado de São Paulo

Fernanda Bartolo dos Santos Saran

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGes)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Fernanda Bartolo dos Santos Saran

Modelagem estatística da impunidade: métodos inferenciais e preditivos para dados de crimes no estado de São Paulo

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Cibele Maria Russo

USP – São Carlos
Maio de 2026

Folha de Aprovação

Defesa de Tese de Mestrado do(a) candidato(a) Fernanda Bartolo dos Santos Saran, realizada em 20/03/2026.

Comissão Julgadora:

Prof(a). Dr(a). Cibele Maria Russo Novelli (USP)

Prof(a). Dr(a). Daiane de Souza Santos (USP)

Prof(a). Dr(a). Rosângela Helena Loschi (UFMG)

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S243m Saran, Fernanda Bartolo dos Santos
Modelagem estatística da impunidade: métodos
inferenciais e preditivos para dados de crimes no
estado de São Paulo / Fernanda Bartolo dos Santos
Saran; orientadora Cibele Maria Russo. -- São
Carlos, 2026.
76 p.

Dissertação (Mestrado - Programa
Interinstitucional de Pós-graduação em Estatística)
-- Instituto de Ciências Matemáticas e de
Computação, Universidade de São Paulo, 2026.

1. Criminalidade. 2. Impunidade. 3. Análise de
sobrevivência. 4. Dados espaciais. I. Russo, Cibele
Maria, orient. II. Título.

Fernanda Bartolo dos Santos Saran

**Statistical modeling of impunity: inferential and predictive
methods for crime data in the state of São Paulo**

Master dissertation submitted to the Institute of
Mathematics and Computer Sciences – ICMC-USP
and to the Department of Statistics – DEs-UFSCar, in
partial fulfillment of the requirements for the degree of
the Master Interagency Program Graduate in Statistics.
FINAL VERSION

Concentration Area: Statistics

Advisor: Profa. Dra. Cibele Maria Russo

USP – São Carlos
May 2026

*Aos meus falecidos avós maternos, Neusa Maria Bartolo e Lourival dos Santos, a quem devo
minha criação, educação e coragem para seguir.*

AGRADECIMENTOS

A todos os que, de alguma forma, me fizeram ver sentido em continuar: os colegas de pós-graduação Riquelme, Lucas Amaral, Lucas Rezende, Guilherme, Rafaela, Rayza e Abderraman - pelos momentos de trégua em meio à tempestade; os meus grandes amigos Isabelle de Freitas e Giuliano Lourençon, pelo incondicional apoio em diversos sentidos; o marido da Isabelle, Renan Santos, pela generosidade; a minha família, em especial, minha mãe Alessandra Bartolo, minha tia-avó Ângela Bartolo e minha prima Roberta Bartolo; o Vitor Mateus Zenatte, por ter onde repousar o coração de vez em quando; a minha psicóloga Daiana Carolina Gagliardi, pela fundamental assistência; e a prof. Cibele Maria Russo Novelli, pela dedicação ao me orientar e confiança depositada em mim durante todo o processo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, e faz parte do projeto temático “Criminalidade, Insegurança e Legitimidade: uma abordagem transdisciplinar”, processo nº 2024/03817-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

“A injustiça num lugar qualquer é uma ameaça à justiça em todo o lugar.”
(Martin Luther King)

RESUMO

SARAN, F. B. S. **Modelagem estatística da impunidade: métodos inferenciais e preditivos para dados de crimes no estado de São Paulo**. 2026. 76 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Cientistas políticos interpretam os altos índices de violência na sociedade brasileira como uma crise de legitimidade, ou seja, uma difundida falta de confiança dos cidadãos no estado. Este cenário é agravado pela impunidade, que se configura como uma problemática social relevante, dada a quebra do princípio da igualdade dos indivíduos perante a lei. Considerando a complexidade associada à quantificação da impunidade e a fim de investigar a eficiência do sistema judiciário no país, o presente estudo utiliza o recorte do cenário paulista e propõe um novo índice para essa problemática: o tempo de conversão de boletins de ocorrência em inquéritos policiais. Dessa forma, os dados utilizados são referentes a casos de homicídios dolosos no estado de São Paulo entre 2017 e 2022. Aplica-se um modelo de sobrevivência com fragilidade espacial modelada por meio da *priori Intrinsic Conditional Autoregressive* (ICAR) considerando variáveis socioeconômicas como o Índice de Desenvolvimento Humano do município (IDH-M) em relação à educação, renda per capita e índice de Gini de cada cidade. Também foi realizado um estudo de simulação, que concluiu que o modelo utilizado é capaz de acessar adequadamente os verdadeiros parâmetros. O estudo é, então, relevante enquanto enriquecimento da literatura acerca do tema, ao propor uma nova abordagem para o contexto de análise da impunidade no Brasil.

Palavras-chave: Criminalidade, Impunidade, Análise de sobrevivência, Dados espaciais.

ABSTRACT

SARAN, F. B. S. **Statistical modeling of impunity: inferential and predictive methods for crime data in the state of São Paulo.** 2026. 76 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

High levels of violence in Brazilian society have been interpreted by political scientists as indicative of a legitimacy crisis - namely, a widespread lack of public trust in the State. This context is exacerbated by impunity, a critical social issue that undermines the principle of equality before the law. Due to the inherent challenges in quantifying impunity and in order to investigate the efficiency of the judicial system in the country, the analysis focuses on the state of São Paulo and proposes a novel indicator: the time elapsed between the filing of a police report and the initiation of a formal criminal investigation. The analysis focuses on intentional homicide cases in the state of São Paulo from 2017 to 2022. A survival model with Intrinsic Conditional Autoregressive (ICAR) spatial frailty is employed, incorporating municipal-level socioeconomic variables such as the Human Development Index (HDI) for education, per capita income, and the Gini index. Also, a simulation study was conducted and concluded that the model is capable of accessing the true parameters appropriately. The study is thus relevant as it contributes to the literature on the subject by proposing a novel approach to the analysis of impunity in Brazil.

Keywords: Criminality, Impunity, Survival analysis, Spatial data.

SUMÁRIO

1	INTRODUÇÃO	17
2	MOTIVAÇÃO	21
2.1	Boletins de ocorrência	21
2.1.1	<i>Informações acerca das vítimas</i>	21
2.1.2	<i>Informações acerca dos crimes por região</i>	23
2.2	Inquéritos	31
3	METODOLOGIA	33
3.1	Dados espaciais	33
3.2	Modelo CAR	34
3.2.1	<i>Aspectos teóricos para a definição do modelo CAR</i>	34
3.2.1.1	<i>Definição de vizinhança</i>	34
3.2.1.2	<i>Lema de Brook</i>	34
3.2.1.3	<i>Campo aleatório de Markov e distribuição de Gibbs</i>	35
3.2.2	<i>Especificação dos modelos CAR e ICAR</i>	35
3.3	Modelos de sobrevivência	36
3.3.1	<i>Análise do gráfico de sobrevivência</i>	37
3.3.2	<i>Modelo semi-paramétrico de riscos proporcionais de Cox</i>	39
3.3.3	<i>Modelo semi-paramétrico de chances proporcionais</i>	39
3.3.4	<i>Fragilidade espacial</i>	40
3.3.5	<i>Modelo semi-paramétrico com fragilidade espacial ICAR</i>	41
3.3.5.1	<i>Riscos proporcionais</i>	41
3.3.5.2	<i>Chances proporcionais</i>	41
3.3.5.3	<i>Distribuições a priori dos parâmetros</i>	41
3.3.5.4	<i>Implementação do método Monte Carlo Markov Chain (MCMC)</i>	42
3.3.5.5	<i>Métricas avaliativas do modelo</i>	46
4	SIMULAÇÃO	47
4.0.1	<i>Aspectos teóricos</i>	47
4.0.2	<i>Resultados</i>	48
5	APLICAÇÃO	59
5.1	Tratamento dos dados	59

5.2	Resultados	61
5.2.1	<i>Análise visual das curvas de sobrevivência</i>	61
5.2.2	<i>Análise visual do índice de impunidade considerado</i>	62
5.2.3	<i>Medidas avaliativas dos modelos ajustados</i>	63
5.2.4	<i>Resíduos do modelo escolhido</i>	65
5.2.5	<i>Convergência do método MCMC</i>	65
5.2.6	<i>Estimativas do modelo escolhido</i>	67
6	DISCUSSÃO E TRABALHOS FUTUROS	71
	REFERÊNCIAS	73

INTRODUÇÃO

Em 2015, a América Latina apresentou um índice de 20,6 homicídios a cada 100 mil habitantes, em comparação com 0,6 para a Europa Ocidental, ou Oeste Europeu ([SANTOS; TESTA, 2018](#)). De 2000 a 2007, o número de homicídios per capita no Brasil foi de 23 por 100 mil habitantes. Entre os países latino-americanos, esse índice foi menor apenas do que os registrados na Venezuela (37), Honduras (42,9) e Colômbia (61,6) ([BRICEÑO-LEÓN; VILLAVECES; CONCHA-EASTMAN, 2008](#)). Além disso, a violência raramente é uniforme mesmo dentro de um mesmo país: em 2010, as estatísticas brasileiras a nível municipal variaram entre 12,9 e 66,8 por 100 mil habitantes ([HOELSCHER, 2015](#)).

Sobre as possíveis causas desse tipo de criminalidade, um estudo propôs o isolamento do efeito exógeno, ou seja, externo, da desigualdade de renda na taxa de homicídios intra e entre 39 países no período de 1965 a 1995. Concluiu-se que a desigualdade, retratada por meio do dobro da área entre a reta identidade e a curva de Lorenz ([GASTWIRTH, 1972](#)) - o chamado índice de Gini, como será explicado posteriormente -, estava associada positivamente a essa taxa tanto dentro de cada país quando entre as nações ([HOELSCHER, 2015](#)). Outro trabalho utilizou dados de 5.000 municípios brasileiros entre 1997 e 2010 a fim de relacionar a violência a ferramentas políticas e institucionais em países sub-desenvolvidos, propondo que a competição política, a concentração do poder político e as restrições em cooperação entre setores da administração pública podem criar oportunidades estruturais para a ocorrência da violência ([FAJNZYLBBER; LEDERMAN; LOAYZA, 2002](#)).

Ainda sobre as estatísticas brasileiras, cientistas políticos afirmam que o país enfrenta uma crise de legitimidade das instituições, que afeta o monopólio estatal da violência, um dos pilares fundamentais para a formação das sociedades modernas ([ADORNO; PASINATO, 2010](#)). Ele assegura que nenhum grupo particular ou comunidade humana possa utilizar a violência como recurso para a resolução de conflitos em relações interpessoais ou entre cidadãos e o Estado. Esse direito estaria reservado a um grupo autorizado, que o faz em obediência às leis

vigentes, de forma considerada legítima (ADORNO; PASINATO, 2007). Nesse sentido, a crise de legitimidade gera revolta na população, o uso da força de forma privativa e, em última instância, aumenta os sentimentos de medo e insegurança coletivos.

Outro agravante dessa situação é a impunidade. De acordo a definição do *Economic and Social Council of the United Nations* (Conselho Econômico e Social das Nações Unidas), ela é caracterizada pela impossibilidade, *de jure* ou *de facto*, isto é, pela lei ou na prática, de responsabilizar autores de violações - de processos criminais, civis, administrativos ou disciplinares - uma vez que não estão sujeitos a qualquer investigação que possa levar à sua acusação, prisão, julgamento e, se considerados culpados, à aplicação de sanções apropriadas, bem como à reparação às suas vítimas (CLERCQ; LARA, 2021). Visto que a falta de punição pode estar relacionada a fatores socioeconômicos, ela está diretamente ligada à existência de desigualdade na sociedade e, assim sendo, também é capaz de gerar quebra de confiança entre os cidadãos e o Estado, o que dificulta a cooperação social e fere o princípio da isonomia, isto é, o conceito de que todos os indivíduos são iguais perante a lei (LOPES, 2000).

Nesse sentido, o *Global Impunity Index* (GII), ou Índice Global de Impunidade, não apresentou informações acerca do Brasil em 2020, pois foi considerado um caso de “impunidade estatística”, isto é, quando não é possível quantificar, em termos estatísticos, a capacidade e operação dos sistemas de segurança e justiça. Isso pode acontecer por três fatores: falta de capacidade institucional acerca das estatísticas nacionais, falta de interesse político para obtenção de tais informações ou manipulação intencional das estatísticas oficiais (CLERCQ; LARA, 2021). Vale ressaltar que os requisitos fundamentais para o desenvolvimento de um sistema sólido de estatísticas de crimes e justiça criminal são a viabilidade de métodos de coleta de dados adaptados ao contexto local, a disponibilidade de experiência técnica e equipamentos para análise e coleta de dados e, por fim, o comprometimento por parte do governo para abordar estrategicamente a coleta e análise das estatísticas (HARRENDORF; HEISKANEN; MALBY, 2010).

Dessa forma, dada a complexidade do tema, desafios são esperados ao lidar com dados relacionados a crimes. De acordo com Harrendorf, Heiskanen e Malby (2010), uma das dificuldades em relação à análise desse tipo de dados é a falta de consistência das informações, que diferem a depender da fonte, além dos registros representarem apenas a parcela dos casos que veio à tona. Apesar disso, ainda segundo o trabalho, a falta de treinamento, comprometimento por parte do governo, legislação adequada e informações suficientes em boas práticas pode representar um obstáculo ainda maior.

Apesar disso, esforços ainda são feitos para contornar essa realidade. Um estudo brasileiro concluiu que o país apresentou cerca de 250 mil mais casos de homicídios do que presos por esse crime entre 2006 e 2014 (NERY; NADANOVSKY, 2020). Sendo essa estatística alarmante, é necessário o estudo mais aprofundado sobre o tema, cuja complexidade se dá devido à relação direta com fenômenos sociais como desigualdade e violência. No cenário do presente estudo, a impunidade é representada pela demora na conversão de um boletim de ocorrência (BO) em

inquérito policial (IP) em casos de homicídios dolosos, ou até, em cenários mais graves, quando ela não ocorre.

Na literatura, outro índice de impunidade é proposto, a partir do número total de homicídios em cinco anos dividido pelo número total de indivíduos presos dois anos depois por este crime ou qualquer outra causa (NERY; NADANOVSKY, 2020). Foram utilizadas informações dos 26 estados brasileiros e também do Distrito Federal (DF) entre 2006 e 2016 e uma regressão linear de Prais-Winsten foi aplicada a fim de estimar a tendência temporal dos índices de impunidade. Em outro trabalho, considera-se o nível global de impunidade baseado no *World Justice Project Rule of Law Index* (Índice sobre Estado de Direito do Projeto de Justiça Mundial). Como o índice está limitado ao intervalo $[0, 1]$, utiliza-se uma regressão beta para a modelagem, considerando variáveis socioeconômicas como explicativas (CRIBARI-NETO; SANTOS, 2024). Em relação à criminalidade, a abordagem espacial é inserida em contextos como o estudo da recidiva de indivíduos no crime a partir de um modelo hierárquico para análise de dados zero inflados sobre eventos espacialmente correlacionados na região de Belo Horizonte, Minas Gerais (SILVA *et al.*, 2025).

O objetivo do presente trabalho é, então, complementar a literatura acerca da impunidade no Brasil ao propor um novo índice: o tempo de conversão de boletins de ocorrência (BOs) em inquéritos policiais (IPs). Com isso, pretende-se compreender a morosidade do processo desde o registro do caso até a instauração do processo administrativo na justiça brasileira e, assim, investigar a eficiência do sistema judiciário do país em relação à devida punição de crimes de homicídios. Para isso, considera-se os casos de homicídios dolosos ocorridos no estado de São Paulo entre 2017 e 2022. Esses dados são analisados por meio de estatísticas como o índice de Moran (MORAN, 1950) para a verificação da dependência espacial das observações regionais. Além disso, utiliza-se o modelo *Intrinsic Conditional Autoregressive* (ICAR) (BESAG, 1974) para a abordagem espacial, incorporada no modelo de sobrevivência semi-paramétrico (MITRA; MÜLLER *et al.*, 2015) por meio da fragilidade. Utiliza-se como variáveis explicativas a população estimada (por cem mil habitantes), o Índice de Desenvolvimento Humano do município (IDH-M) em relação à educação, a renda per capita, o índice de vulnerabilidade (IVS) e o índice de Gini do município. O pacote *spBayesSurv* (ZHOU; HANSON; ZHANG, 2020) foi utilizado para a implementação da modelagem, realizada em *software R*.

Além disso, no presente trabalho, é feita uma simulação de dados de sobrevivência, a fim de verificar a adequabilidade do modelo aos dados sintéticos e, consequentemente, a capacidade dele em acessar os verdadeiros parâmetros. Para cada um de 4 tamanhos amostrais, consideram-se 1.000 conjuntos de dados simulados, e analisam-se métricas como o erro quadrático médio (EQM), erro-padrão, viés e probabilidade de cobertura de cada parâmetro em função do tamanho amostral. O mesmo processo é feito considerando 5.000 conjuntos de dados. Com isso, é possível examinar a confiabilidade do modelo.

Destaca-se que, ao investigar o processo de morosidade da conversão de boletins de

ocorrência em inquéritos policiais, analisa-se o fluxo entre delegacias, que recebem os casos de ocorrências, e o sistema judiciário brasileiro. Nesse contexto, os dados dos boletins e dos inquéritos não só representam bases diferentes, como também não apresentam correspondência direta, o que é discutido nos [Capítulo 2](#) e [Capítulo 5](#), e reflete os obstáculos para análise de dados já apresentados. Ainda assim, uma aproximação é proposta, e suas limitações são discutidas no [Capítulo 5](#).

Ressalta-se a importância do trabalho na medida em que visa enriquecer a literatura em termos de quantificação da impunidade ao propor um novo índice para medir essa problemática social. Além disso, a metodologia utilizada é nova nesse contexto. Ao considerar a análise de sobrevivência com a componente espacial, a dependência entre as regiões é modelada e a análise torna-se mais completa e informativa. Os casos mais graves de impunidade, ou seja, a falta de conversão do boletim de ocorrência em inquérito, são considerados como censura e os demais cenários são também abordados nessa estratégia.

No [Capítulo 2](#), faz-se a análise descritiva dos dados. No [Capítulo 3](#), apresentam-se os passos teóricos para a aplicação da modelagem, desde as definições necessárias para a análise espacial até os detalhes de implementação do método bayesiano de estimação dos parâmetros do modelo e as respectivas métricas de avaliação deste. Além disso, discorre-se sobre os procedimentos necessários para a simulação de dados. No [Capítulo 5](#), apresentam-se a análise gráfica do índice considerado, as medidas de avaliação de convergência do método de estimação dos parâmetros, as estimativas do modelo ajustado e os respectivos intervalos de credibilidade. Nessa seção, também são apresentados os resultados da simulação. Por fim, no [Capítulo 6](#), as limitações e a relevância do trabalho são discutidas, bem como possíveis próximos passos são propostos.

MOTIVAÇÃO

Neste capítulo, faz-se uma análise descritiva e exploratória dos dados que motivaram o estudo. Isso é feito com o objetivo de compreender possíveis padrões e promover uma visão geral dos dados, que torna possível a compreensão mais ampla do contexto, ou seja, os casos de homicídios dolosos no estado paulista entre 2017 e 2022. Divide-se na [Seção 2.1](#), na qual são descritos os dados dos boletins de ocorrência, e [Seção 2.2](#), na qual os dados dos inquéritos são explorados.

2.1 Boletins de ocorrência

Os dados dos boletins de ocorrência (BOs) foram retirados do site da Secretaria de Segurança Pública de São Paulo (SSP-SP) e consistem em 18.439 observações correspondentes a pessoas envolvidas nos boletins abertos entre janeiro de 2017 e dezembro de 2022 no estado. Há informação acerca da delegacia onde os BOs foram abertos, o número de vítimas do homicídio doloso, a data, hora e localidade do fato, sexo, idade, cor da pele e profissão da pessoa envolvida.

2.1.1 Informações acerca das vítimas

A [Tabela 1](#) apresenta a relação entre o tipo de pessoa envolvida no crime e a frequência com que aparece na base de dados. Categorias como “Indiciado”, “Encontrado”, “Adolescente”, “Adolescente infrator” e “Autor” são tratadas como “Outros”. A fim de analisar somente as 17.817 vítimas em termos de descrição e posterior modelagem, as demais categorias são desconsideradas.

A [Tabela 2](#) mostra a frequência de cada sexo das vítimas de homicídios dolosos no estado no período de estudo. Os homens representam a grande maioria dos casos, sendo a chance de uma vítima escolhida ao acaso ser homem equivalente a 4,74 vezes maior do que a de ser uma mulher, considerando o grupo de pessoas cujo sexo foi informado.

Tabela 1 – Frequência de cada tipo de envolvido.

Tipo de pessoa	Frequência
Vítima	17.817
Autor/Vítima	526
Encontro de cadáver	32
Adolescente infrator/Vítima	26
Desaparecido	10
Outros	28

Fonte: Elaborada pela autora.

Tabela 2 – Frequência do sexo das vítimas

Sexo da vítima	Frequência
Masculino	14.934
Feminino	2.602
Indefinido	281

Fonte: Elaborada pela autora.

Já a [Tabela 3](#) apresenta a frequência de cada cor da pele das vítimas. Observa-se que a maioria delas eram brancas e pardas, seguidas pelas pretas. Além disso, em 485 observações, a cor da pele da vítima não foi informada, sendo 482 constando como “Ignorada” e 3 como “Nulo”. As demais informações são referentes à cor da pele “Amarela” (36 casos), “Vermelha” (4 casos) e “Outros” (110 casos), completando as 17.817 observações.

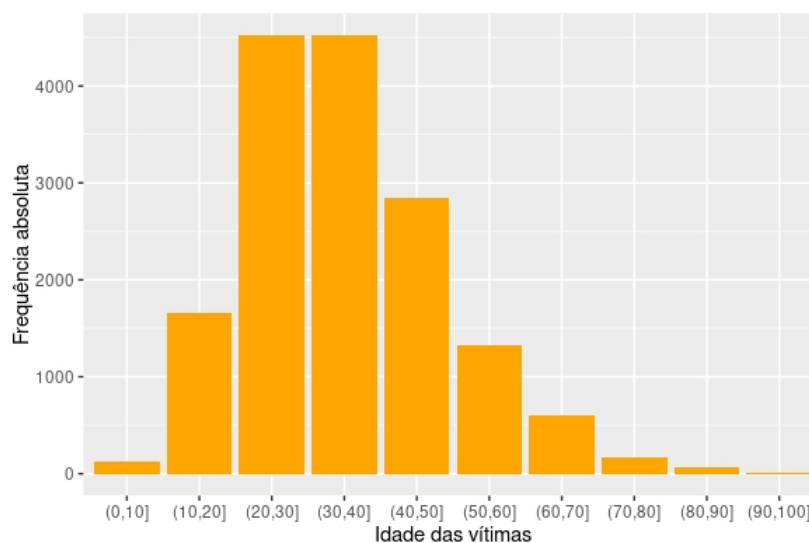
Tabela 3 – Frequência da cor da pele das vítimas

Cor da pele da vítima	Frequência
Branca	8.258
Parda	7.409
Preta	1.515
Ignorada	482
Outros	110
Amarela	36
Vermelha	4
Nulo	3

Fonte: Elaborada pela autora.

A seguir, na [Figura 1](#), apresenta-se o gráfico de barras das idades das vítimas. Foram retiradas 1.993 observações que não continham a informação referente à idade da vítima. Entre as vítimas cuja idade consta na base de dados, 93,22% possuíam entre 10 e 60 anos.

Figura 1 – Gráfico de barras da idade das vítimas.

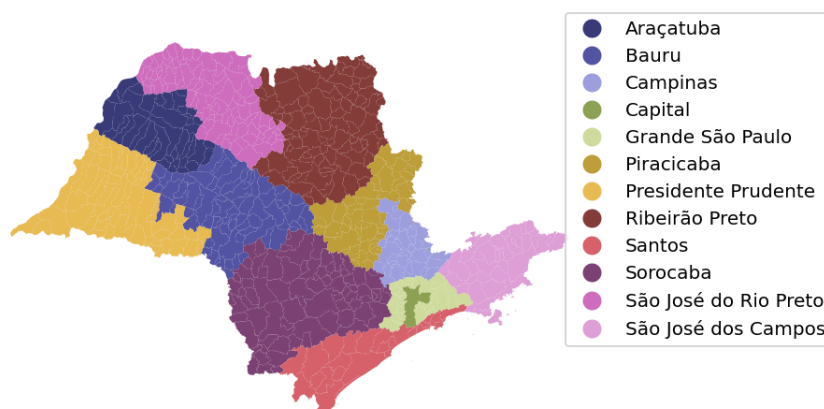


Fonte: Elaborada pela autora.

2.1.2 Informações acerca dos crimes por região

A regionalização utilizada nesse estudo está de acordo com a segmentação utilizada no site da Secretaria de Segurança Pública do estado de SP, cuja representação é apresentada na [Figura 2](#). Na [Tabela 4](#), tem-se o tamanho populacional total de cada região.

Figura 2 – Representação da regionalização utilizada - de acordo com os dados da Secretaria de Segurança Pública do estado de SP (SSP-SP).



Fonte: Elaborada pela autora.

Verifica-se o número absoluto ([Figura 3](#)) e per capita ([Figura 4](#)) de vítimas de homicídios dolosos no estado paulista por mês no período de estudo por região. É verificável que, em termos absolutos, as regiões da grande São Paulo e da capital lideram o número mensal de vítimas de casos de homicídios dolosos entre 2017 e 2022 no estado paulista, seguidas pela região de São José dos Campos. Tanto na região metropolitana quanto na capital, é possível perceber que há

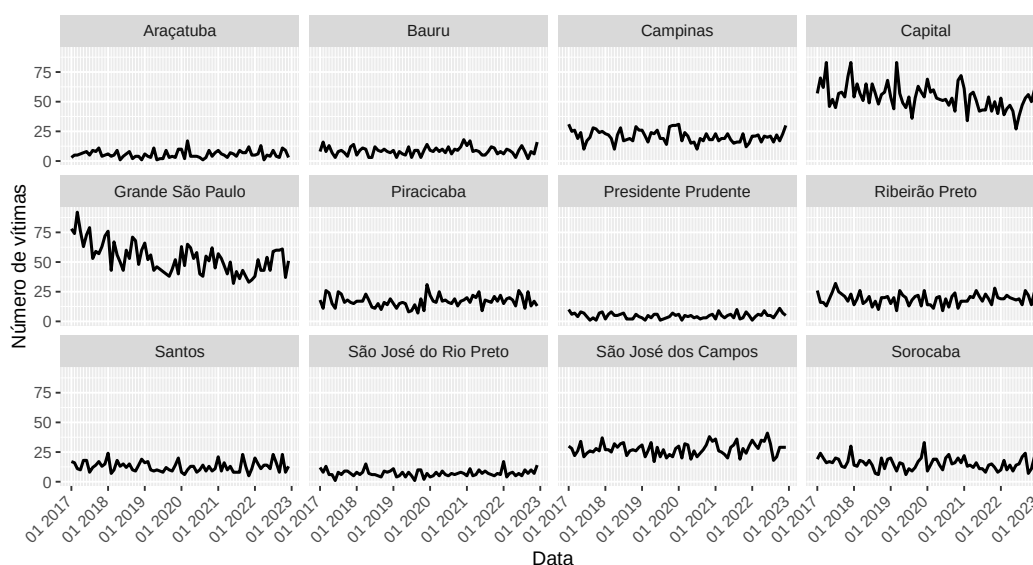
Tabela 4 – Tamanho populacional total de cada região

Região	População total
Araçatuba	819.830
Bauru	1.934.714
Campinas	3.893.806
Capital	12.396.372
Grande São Paulo	9.652.132
Piracicaba	3.378.700
Presidente Prudente	1.170.317
Ribeirão Preto	3.792.726
Santos	2.182.882
Sorocaba	3.212.306
São José do Rio Preto	1.616.129
São José dos Campos	2.599.218

Fonte: Elaborada pela autora.

uma tendência de queda no índice ao longo do tempo, o que não parece ocorrer para as demais regiões.

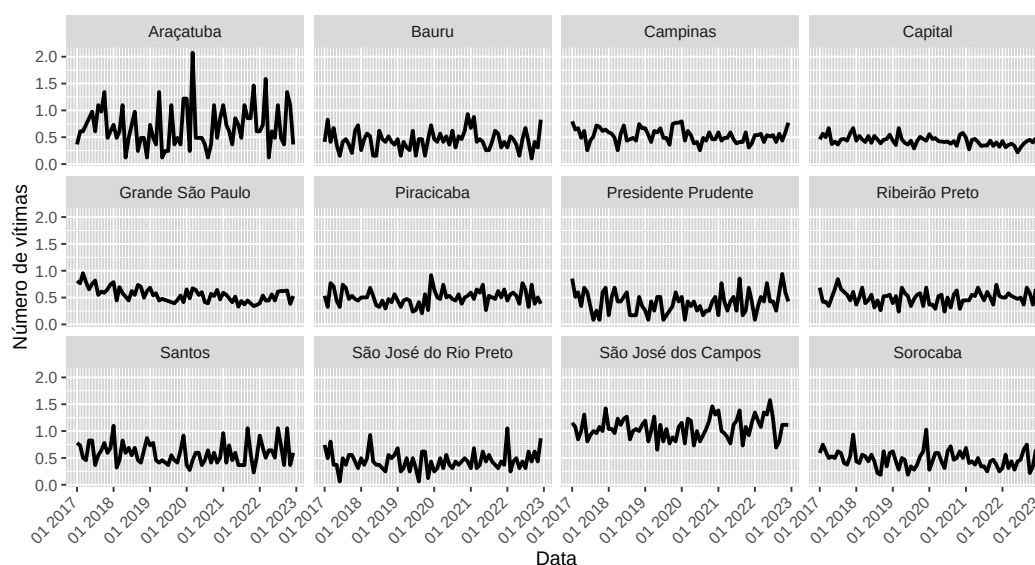
Figura 3 – Número absoluto de vítimas de homicídios dolosos em SP por mês e região.



Fonte: Elaborada pela autora.

Em termos per capita, São José dos Campos se destaca, com valores em torno de 1 homicídio a cada 100 mil habitantes por mês. Já a região de Araçatuba possui variabilidade maior em relação ao índice do que as demais regiões, alternando-se entre picos que chegam até 2,07 homicídios a cada 100 mil habitantes e vales de, no mínimo, 0,12. Isso corresponde a uma amplitude de 1,95 - a maior, seguida por São José do Rio Preto (0,99). O desvio-padrão da métrica para a localidade ficou em 0,39, também a maior entre as demais, seguida por Presidente Prudente e São José dos Campos (0,2).

Figura 4 – Número de vítimas de homicídios dolosos a cada 100 mil habitantes em SP por mês e região.



Fonte: Elaborada pela autora.

Na [Figura 5](#), apresenta-se o mapa das ocorrências relatadas em boletins de ocorrência no estado de São Paulo entre 2017 e 2022. É possível observar que a região metropolitana e capital concentram o número de homicídios dolosos relatados em boletins de ocorrência em todo o período no estado paulista. Há também uma concentração ao leste, na região de São José dos Campos, localizada na mesorregião do Vale do Paraíba e Litoral Norte.

Na [Figura 6](#), apresenta-se o número per capita a cada 100 mil habitantes de vítimas de homicídios dolosos no estado de São Paulo no período analisado, observado por região. Novamente, percebe-se que a região onde os números desse tipo de crime predominam sobre as outras em todo o período de estudo é a de São José dos Campos. Além disso, a região de Araçatuba também se destaca em anos como 2017, com queda no ano seguinte, seguida de progressão ano a ano até 2021, quando novamente chega a valores próximos da região de São José dos Campos, e leve queda no último ano de estudo.

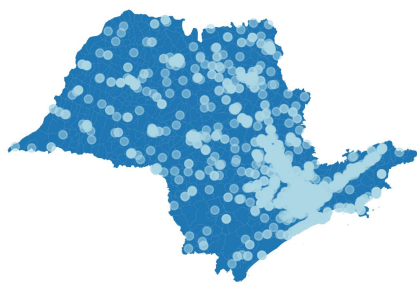
Apresenta-se também, na [Figura 7](#), o número per capita a cada 100 mil habitantes de vítimas de homicídios dolosos no estado de São Paulo no período analisado observado por município. Aqui, uma pergunta de interesse pode ser se há relação espacial do índice analisado, seja positiva, ou seja, regiões com valores próximos são geograficamente também próximas, ou negativa, isto é, regiões com valores próximos tendem a estar distantes no espaço.

A fim de verificar a estrutura de correlação entre as regiões em termos do número per capita ($\times 10^5$) de homicídios dolosos no estado paulista no período analisado, analisa-se o gráfico de Moran ([MORAN, 1950](#)) de tal índice por ano. Esse gráficos são apresentados na [Figura 8](#). A estatística do teste de Moran é o coeficiente angular da regressão linear que tem como variável resposta o número per capita de vítimas normalizado, ou seja, obtido a partir da subtração da

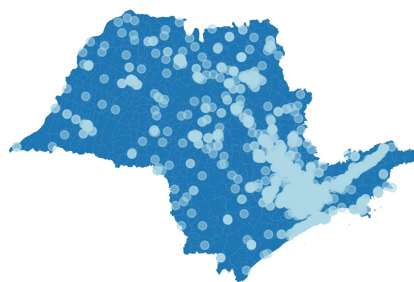
média e divisão pelo desvio-padrão, e como variável dependente o *lag* espacial de tais valores. Nesse sentido, o *lag* espacial de uma região considerada corresponde à média de seus vizinhos, ou seja, localidades que fazem fronteira.

Figura 5 – Localidades dos homicídios dolosos relatados em boletins de ocorrência no estado de São Paulo entre 2017 e 2022.

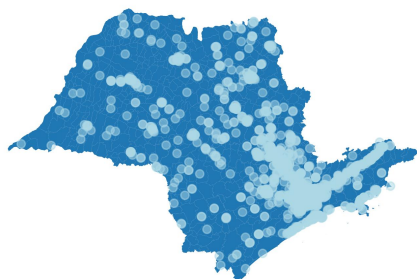
(a) Em 2017.



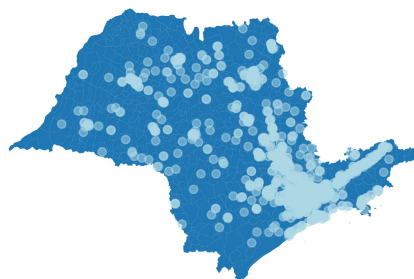
(b) Em 2018.



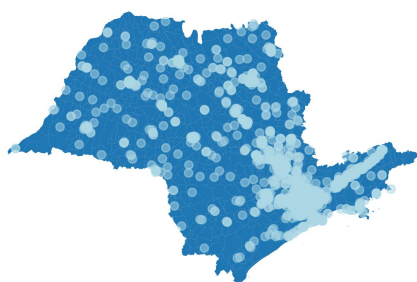
(c) Em 2019.



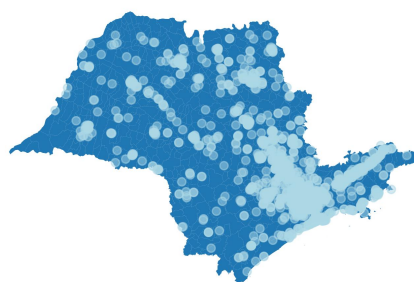
(d) Em 2020.



(e) Em 2021.

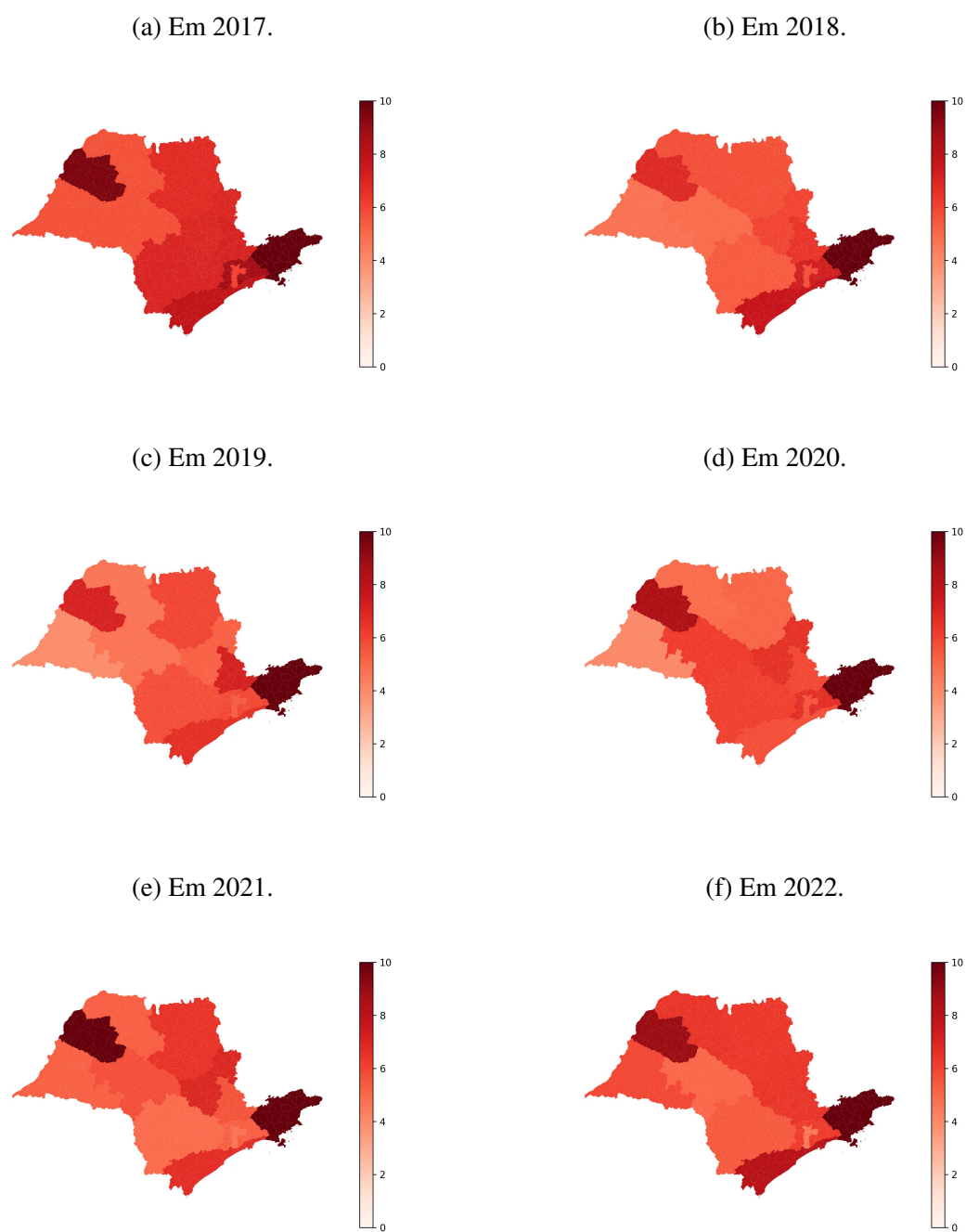


(f) Em 2022.



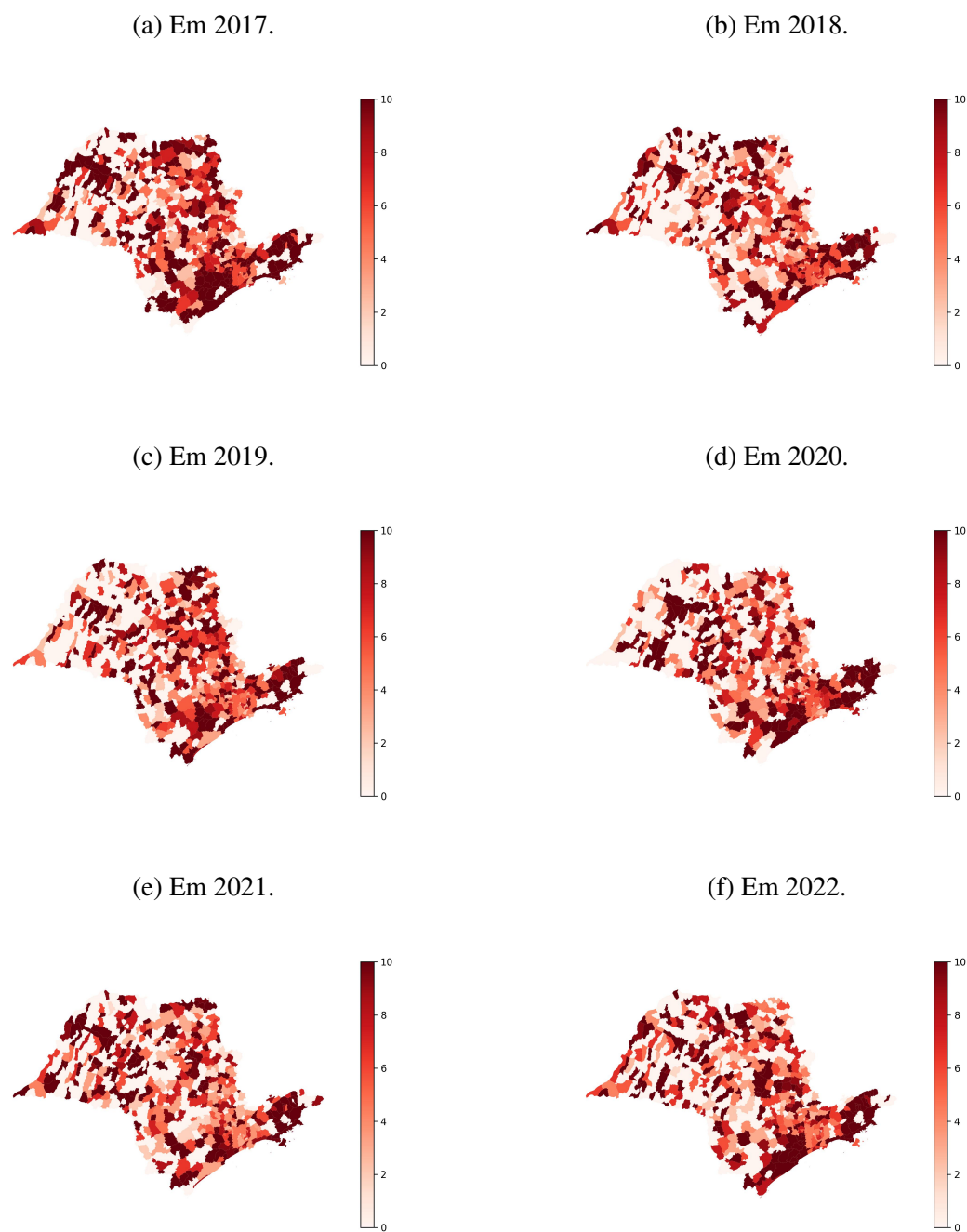
Fonte: Elaborada pela autora.

Figura 6 – Número per capita ($\times 10^5$) de vítimas de homicídios dolosos no estado de São Paulo entre 2017 e 2022 por região.



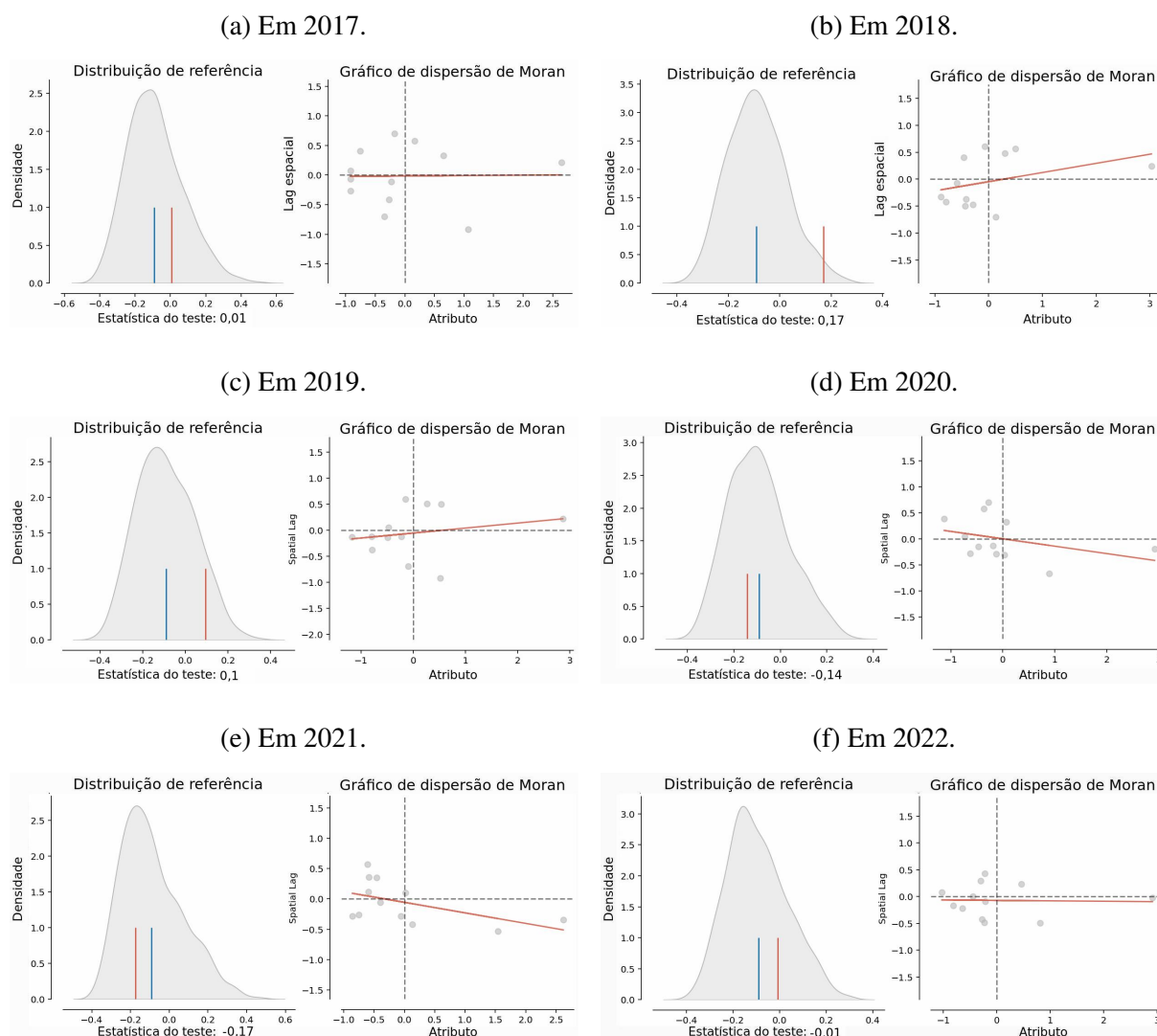
Fonte: Elaborada pela autora.

Figura 7 – Número per capita ($\times 10^5$) de vítimas de homicídios dolosos no estado de São Paulo entre 2017 e 2022 por município.



Fonte: Elaborada pela autora.

Figura 8 – Gráfico de Moran para o número per capita ($\times 10^5$) de vítimas de homicídios dolosos no estado de São Paulo entre 2017 e 2022 por região.



Fonte: Elaborada pela autora.

Assim, nos gráficos à esquerda da [Figura 8](#), tem-se a distribuição de referência dessa estatística de Moran, e quanto mais distante o traço vermelho (estatística observada) do traço azul (valor de referência), maior é a evidência de que a hipótese nula de independência espacial é falsa, ou seja, há dependência espacial entre as regiões. À direita, tem-se no eixo X os valores observados normalizados e, no eixo Y, o *lag* espacial de tais valores. A reta destacada em vermelho refere-se ao ajuste da regressão linear entre esses valores, conforme explicado anteriormente.

Conforme os valores-p dos testes de Moran para cada ano, apresentados na [Tabela 5](#), não se rejeita a hipótese nula em nenhum caso ao nível de significância de 5%, e considera-se que não há relação espacial entre os números per capita de vítimas de homicídios dolosos nas regiões do estado.

Tabela 5 – Valores-p dos testes de Moran para dependência espacial do número de homicídios per capita por região.

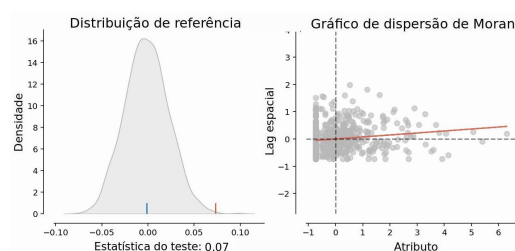
Ano	Valor-p
2017	0,58
2018	0,14
2019	0,29
2020	0,77
2021	0,64
2022	0,64

Fonte: Elaborada pela autora.

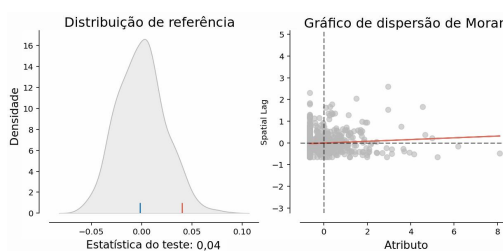
Também verifica-se o gráfico de Moran para o número per capita a cada 100 mil habitantes do número de vítimas de homicídios dolosos no estado por município para cada ano do período estudado, na [Figura 9](#).

Figura 9 – Gráfico de Moran para o número per capita ($\times 10^5$) de vítimas de homicídios dolosos no estado de São Paulo entre 2017 e 2022 por município.

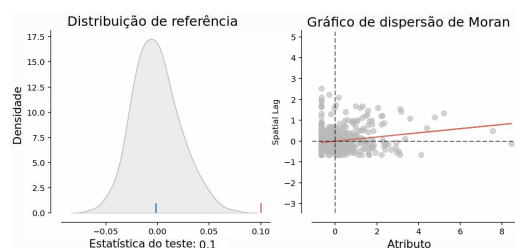
(a) Em 2017.



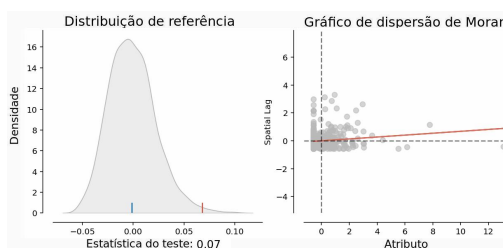
(b) Em 2018.



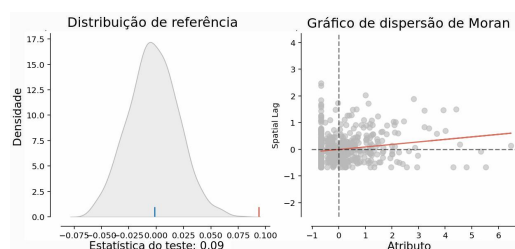
(c) Em 2019.



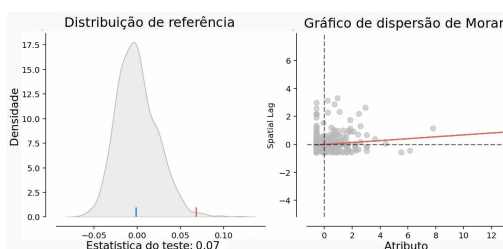
(d) Em 2020.



(e) Em 2021.



(f) Em 2022.



Fonte: Elaborada pela autora.

Conforme os valores-p dos testes de Moran para cada ano apresentados na [Tabela 6](#), rejeita-se a hipótese de independência espacial para o número de homicídios per capita no estado

por município em todo o período analisado ao nível de significância de 5%. Conclui-se, portanto, que os índices per capita desse tipo de violência urbana não estão relacionados espacialmente quando se consideram as regiões do estado, porém estão quando se consideram seus municípios.

Tabela 6 – Valores-p dos testes de Moran para dependência espacial do número de homicídios per capita por município.

Ano	Valor-p
2017	0,00
2018	0,08
2019	$1,77 * 10^{-5}$
2020	$5,37 * 10^{-5}$
2021	0,00
2022	0,00

Fonte: Elaborada pela autora.

Também realizou-se a análise dos gráficos de autocorrelação e autocorrelação parcial da série temporal referente ao número de vítimas per capita de cada região por mês e dia, porém não foram encontradas dependências significativas. Nesse sentido, optou-se por não apresentá-los no presente texto.

2.2 Inquéritos

Em relação aos inquéritos, foram obtidos do site do Ministério Público do estado os dados referentes aos IP cujo tema era “Crimes contra a vida”. Tem-se 139.317 observações, contendo informações como o tipo de procedimento (inquérito policial e procedimento investigatório criminal), a unidade em que o procedimento foi realizado, a situação (arquivado, denunciado, em andamento etc), o assunto (como “DIREITO PENAL - Crimes contra a vida - Homicídio Simples” e “DIREITO PENAL - Crimes contra a vida - Homicídio Qualificado”), as partes envolvidas, a data da instauração entre outras informações.

Ressalta-se que não há a referência à qual boletim de ocorrência o inquérito está associado, ou seja, a relação entre ambos os procedimentos não está clara, sendo necessária a utilização de uma estratégia aproximada para calcular o índice de impunidade proposto neste trabalho, conforme será discutido no [Capítulo 5](#).

METODOLOGIA

Neste capítulo, as principais metodologias utilizadas para a análise dos dados são descritas. Divide-se na [Seção 3.1](#), que contém uma breve explicação a respeito da definição de dados espaciais e como as informações analisadas no presente estudo estão associadas a ela; na [Seção 3.2](#), em que aspectos teóricos são discutidos e os modelos CAR e ICAR são definidos e, por fim, na [Seção 3.3](#), que apresenta uma breve introdução aos modelos de sobrevivência - em especial os semi-paramétricos de riscos e chances proporcionais, com e sem fragilidade espacial.

Além disso, a fragilidade espacial também é definida na [Seção 3.3](#), bem como a implementação do método *Monte Carlo Markov Chain* (MCMC) utilizado para a análise dos dados.

3.1 Dados espaciais

De acordo com [Banerjee, Carlin e Gelfand \(2014\)](#), os dados espaciais são do tipo *areal data* (ou “dados de área”, em tradução livre) nos casos em que as regiões variam em D , um subconjunto fixo, de formato regular ou irregular, particionado em um número finito de unidades de áreas com limites bem definidos. Esses dados são chamados também de *lattice data* ou “dados de grade”, em tradução livre, e define-se o conceito de vizinhança baseando-se na disposição dos blocos no mapa.

Os dados aqui analisados, conforme a [Figura 2](#), podem ser representados por uma região, de formato irregular, particionada em doze áreas com limites bem definidos no mapa. Então, são considerados como *lattice data*. Em geral, dois modelos são mais utilizados para a análise desse tipo de dados: o *simultaneously* e o *conditional autoregressive* (SAR e CAR), desenvolvidos por [Whittle \(1954\)](#) e [Besag \(1974\)](#), respectivamente. Segundo [Banerjee, Carlin e Gelfand \(2014\)](#), o modelo SAR é mais computacionalmente conveniente para o uso com métodos de verossimilhança, enquanto o CAR para a amostragem de Gibbs usada na abordagem

bayesiana. O modelo usado nesse estudo é proposto por meio da ótica bayesiana e, portanto, optou-se pela modelagem dos dados pelo CAR.

3.2 Modelo CAR

3.2.1 Aspectos teóricos para a definição do modelo CAR

3.2.1.1 Definição de vizinhança

Dado um conjunto de m regiões, diz-se que i -ésima e j -ésima regiões são vizinhas se elas compartilham fronteira, e denota-se por $i \sim j$, para $i \neq j$, $i, j = 1, \dots, m$. Essa relação de vizinhança é simétrica, mas não reflexiva, ou seja, $i \sim j \iff j \sim i$, $i \neq j$, mas uma região não é vizinha de si mesma: $i \not\sim i, \forall i$ (MORRIS *et al.*, 2019). De forma equivalente, denota-se por $i \in \partial_j$ ou $j \in \partial_i$ se $i \sim j$, sendo ∂_i o conjunto de células vizinhas da i -ésima.

3.2.1.2 Lema de Brook

Dada uma distribuição conjunta $p(y_1, \dots, y_n)$, as distribuições condicionais completas $p(y_i | y_j, i \neq j), i = 1, \dots, n$ são unicamente determinadas. Entretanto, em geral, o contrário não é possível ser afirmado. O lema de Brook (BROOK, 1964) demonstra que é possível retornar à distribuição conjunta a partir de distribuições condicionais completas, desde que elas sejam compatíveis, ou seja, satisfaçam uma condição.

Lema 1 (*Lema de Brook*). Dado um ponto fixo $\mathbf{y}_0 = (y_{10}, y_{20}, \dots, y_{n0})$ e um ponto $\mathbf{y} = (y_1, y_2, \dots, y_n)$ qualquer, é posto que

$$\frac{p(\mathbf{y})}{p(\mathbf{y}_0)} = \prod_i^n \frac{p(y_i | y_1, \dots, y_{i-1}, y_{(i+1)0}, \dots, y_{n0})}{p(y_{i0} | y_1, \dots, y_{i-1}, y_{(i+1)0}, \dots, y_{n0})}.$$

Assim, é possível calcular a distribuição de probabilidade conjunta $p(\mathbf{y})$ como

$$p(\mathbf{y}) = \frac{\frac{p(\mathbf{y})}{p(\mathbf{y}_0)}}{\sum_{\mathbf{y}} \frac{p(\mathbf{y})}{p(\mathbf{y}_0)}},$$

e diz-se que a distribuição conjunta é válida se

$$\sum_{\mathbf{y}} \frac{p(\mathbf{y})}{p(\mathbf{y}_0)} = \sum_{\mathbf{y}} \prod_i^n \frac{p(y_i | y_1, \dots, y_{i-1}, y_{(i+1)0}, \dots, y_{n0})}{p(y_{i0} | y_1, \dots, y_{i-1}, y_{(i+1)0}, \dots, y_{n0})} < \infty.$$

Dessa forma, o lema auxilia na verificação e especificação dos casos em que as distribuições condicionais completas determinam uma distribuição conjunta válida.

3.2.1.3 Campo aleatório de Markov e distribuição de Gibbs

A especificação local para a determinação da distribuição conjunta de um conjunto de variáveis é conhecida como um *Markov Random Field* (MRF), ou campo aleatório de Markov (BANERJEE; CARLIN; GELFAND, 2014). Uma dessas especificações é a definição das condicionais completas a partir das vizinhanças, ou seja,

$$p(y_i | y_j, j \neq i) = p(y_i | y_j, j \in \partial_i), \forall i,$$

em que ∂_i corresponde ao conjunto de células vizinhas da i -ésima. Mais a esse respeito pode ser encontrado em ROZANOV (1982).

Entretanto, é necessário compreender se essa especificação determina uma única distribuição conjunta para $Y_1, \dots, Y_m$, o que pode ser feito através da condição oferecida pelo teorema de Brook (Subsubseção 3.2.1.2).

Além disso, o teorema de Hammersley-Clifford demonstra que, dado um MRF, então a distribuição conjunta é uma distribuição de Gibbs (BESAG, 1974; CLIFFORD, 1990). A saber, uma distribuição de Gibbs é aquela em que só depende de Y_i por meio de *potential on cliques*, sendo *cliques* um conjunto de células, ou índices, cujos elementos são vizinhos entre si e uma função *potential on cliques* de ordem k aquela função de k argumentos que é permutável neles (BANERJEE; CARLIN; GELFAND, 2014). Assim, uma distribuição de Gibbs toma a forma

$$p(y_1, \dots, y_n) \propto \exp \left\{ \gamma \sum_k \sum_{\alpha \in \mathcal{M}_k} \phi^{(k)}(y_{\alpha_1}, \dots, y_{\alpha_n}) \right\}, \quad (3.1)$$

em que $\phi^{(k)}$ é uma *potential function* de ordem k , $\mathcal{M}_k$ é uma coleção de todos os subconjuntos de tamanho k de $\{1, \dots, n\}$, $\alpha = (\alpha_1, \dots, \alpha_n)^\top$ corresponde aos índices dessa coleção e $\gamma > 0$ é um parâmetro de escala.

Enquanto isso, o teorema de Geman e Geman, apresentado em GEMAN e GEMAN (1984), afirma que, dada uma distribuição de Gibbs, de acordo com a Equação 3.1, tem-se especificado um MRF. Assim, com esses dois teoremas, afirma-se que uma distribuição de Gibbs é equivalente a um MRF (SPITZER, 1971).

3.2.2 Especificação dos modelos CAR e ICAR

A interação espacial entre pares de regiões pode ser descrita como um vetor aleatório m -dimensional $\mathbf{y} = (y_1, \dots, y_m)^\top$. Nesse sentido, define-se a distribuição condicional completa dessas variáveis por

$$y_i | y_j, j \neq i, \sim N \left(\sum_{j=1}^m w_{ij} y_j, \sigma^2 \right), \quad (3.2)$$

ou seja, é condicional na soma dos valores ponderados dos seus vizinhos e possui variância desconhecida σ^2 (BESAG, 1974). É necessário encontrar a distribuição do vetor aleatório $\mathbf{y}$.

Para isso, são usados os resultados teóricos apresentados anteriormente. Pelo lema de Brook, as distribuições condicionais completas da [Equação 3.2](#) formam uma distribuição única. Assim, pelo teorema de Hammersley-Clifford, a distribuição conjunta é uma distribuição de Gibbs, pois só depende de $y_i, i = 1, \dots, m$, por funções *potential on cliques*.

É obtido em [Besag \(1974\)](#) que

$$\mathbf{y} \sim N(0, Q^{-1}),$$

em que a matriz de precisão Q é simétrica e positiva definida, tomando a forma

$$Q = D(I - \alpha E),$$

sendo $0 < \alpha < 1$ o parâmetro referente à dependência espacial das regiões, D uma matriz diagonal e E a matriz de adjacência, que recebe valor $e_{ij} = 1$ se as regiões i e j são vizinhas ($i \sim j$) e $e_{ij} = 0$, caso contrário. Em geral, conforme explicitado em [Morris et al. \(2019\)](#), a matriz diagonal D recebe valor $d_{ii} = \sum_{j=1}^m e_{ij}$, correspondente ao número de vizinhos da i -ésima região.

No modelo *Intrinsic Conditional Autoregressive* (ICAR), o parâmetro α é estabelecido como 1, o que torna a distribuição *a priori* ICAR imprópria, mas a distribuição *a posteriori* é própria após a inclusão de dados ([MORRIS et al., 2019](#)).

3.3 Modelos de sobrevivência

Os modelos de sobrevivência são usados para modelagem do tempo até a ocorrência de um evento de interesse. Nesse sentido, é possível que essa informação esteja disponível apenas parcialmente para algum indivíduo no estudo, ou seja, o tempo exato até essa ocorrência seja desconhecido. Esses casos são conhecidos como censura ([EMMERT-STREIB; DEHMER, 2019](#)). A censura pode ser à esquerda, quando sabe-se somente que o evento precedeu certa data, à direita, quando sabe-se somente que o evento sucedeu tal tempo, ou intervalar, isto é, quando sabe-se somente o intervalo no qual o evento ocorreu.

Segundo [Moore et al. \(2016\)](#), estudos que utilizam a censura à direita podem ser categorizados em tipo I, II ou aleatório. Os do tipo I ocorrem quando são finalizados após um período fixo de tempo, enquanto os do tipo II, após uma parcela dos indivíduos apresentarem o evento de interesse. Por outro lado, os do tipo aleatório ocorrem quando uma observação é censurada por sair do estudo antes que o evento acontecesse, por motivos que não necessariamente estavam ligados à pesquisa, como por exemplo a perda de contato com um paciente ou a morte de um indivíduo por outras causas que não a doença em questão.

A existência ou não de censura é modelada por meio de uma variável δ tal que

$$\delta_i = \begin{cases} 0, & \text{se o } i\text{-ésimo tempo de sobrevivência é desconhecido (observação censurada);} \\ 1, & \text{se o } i\text{-ésimo tempo de sobrevivência é conhecido (evento de interesse observado).} \end{cases}$$

No contexto do estudo, o evento de interesse é a conversão de um boletim de ocorrência em inquérito, considerando os casos de homicídios dolosos no estado paulista. Nesse sentido, a censura é à direita, visto que acompanha-se o boletim de ocorrência desde a sua abertura até a possível conversão em inquérito, ou seja, pode haver cenários em que o tempo real de conversão seja maior do que aquele que foi possível observar. Assim, a censura é altamente relevante nesse contexto, pois trata-se de casos mais graves de impunidade. O estudo é também do tipo I, pois termina após um período fixo de tempo, a saber, 6 anos (de 2017 a 2022).

Um ponto de começo comum para a especificação de um modelo de regressão para tempos até a ocorrência de um evento é a definição de um função de sobrevivência basal (*baseline*) S_0 , que é modificada direta ou indiretamente pelas covariáveis $\mathbf{x}$ (MITRA; MÜLLER *et al.*, 2015). Assim, sendo T_0 um tempo aleatório de sobrevivência do grupo basal, ou seja, em que todas as covariáveis são zero, então tem-se que $S_0(t) = P(T_0 > t)$ é a função de sobrevivência basal.

A seguir, faz-se uma descrição de como as curvas de sobrevivência são representadas visualmente. Além disso, são apresentados esquemas que comparam como elas são interpretadas no contexto geral e no caso do presente estudo.

3.3.1 Análise do gráfico de sobrevivência

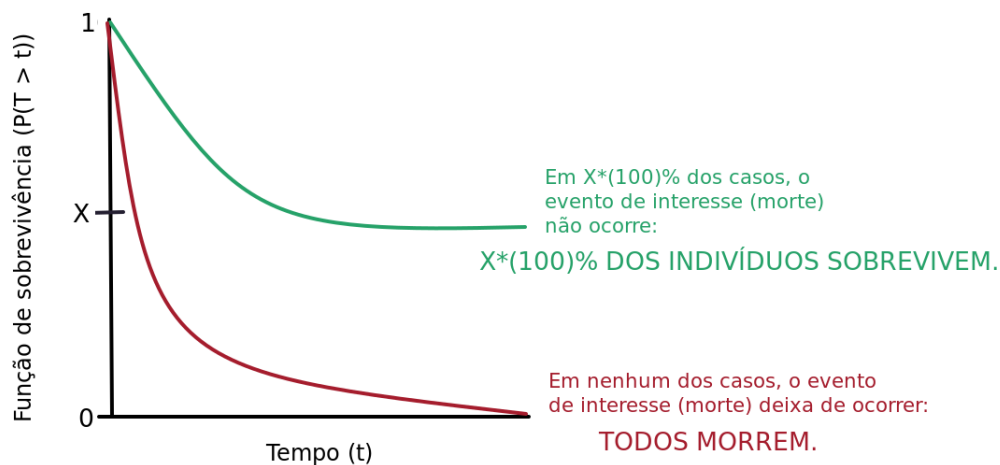
O gráfico de sobrevivência apresenta a probabilidade de uma observação demorar mais do que um tempo t para apresentar o evento de interesse. As curvas de sobrevivência sempre começam em 1, pois, no tempo $t = 0$, a probabilidade de que o evento de interesse demore mais do que t é igual a 1. Elas também terminam em 0 caso não haja censura, pois a estimação é a de que todos os indivíduos apresentem o evento de interesse até o último tempo observado. Por outro lado, elas não terminam em 0 quando há censura, pois estima-se que uma porcentagem dos dados apresentará o evento de interesse em um tempo maior do que o do final do estudo, podendo, inclusive, nunca haver essa ocorrência.

O esquema apresentado na Figura 10 representa duas curvas comuns em estudos que consideram como evento de interesse a morte de indivíduos ou a falha de equipamentos eletrônicos: em vermelho, o cenário mais grave e, em verde, o cenário mais favorável. No primeiro caso, a curva decai rapidamente, indicando que o evento de interesse (morte ou falha) foi observado para todos os indivíduos. Já no segundo caso, a curva decai mais lentamente, o que indica que a morte ou a falha não foi observada em uma parcela dos indivíduos em estudo.

Já no caso do presente estudo, a curva de sobrevivência indica a probabilidade de um boletim de ocorrência não se tornar inquérito até um tempo t , em determinada região. Sendo assim, caso a curva decaia rapidamente, entende-se que essa conversão é rápida. Inversamente, caso a curva demore a decair, compreende-se que, naquela região, a conversão de um boletim de ocorrência para inquérito policial é mais demorada. Sendo assim, os cenários representados na

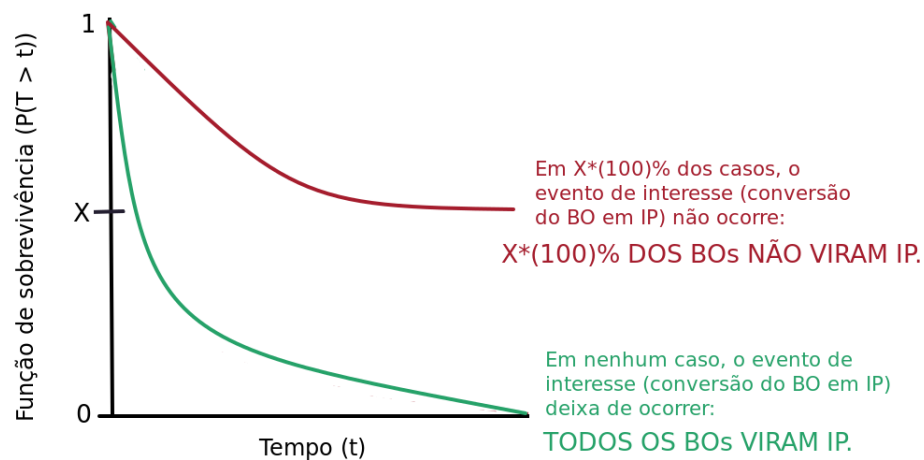
Figura 10 se invertem, como é esquematizado na Figura 11.

Figura 10 – Esquema de um gráfico de sobrevivência para um estudo cujo evento de interesse é a morte literal de um ser ou figurada (falha) de um aparelho eletrônico, por exemplo. Em vermelho, o caso mais grave. Em verde, o mais favorável.



Fonte: Elaborada pela autora.

Figura 11 – Esquema de um gráfico de sobrevivência para um estudo cujo evento de interesse é a conversão de um boletim de ocorrência (BO) em inquérito policial (IP). Em vermelho, o caso mais grave. Em verde, o mais favorável.



Fonte: Elaborada pela autora.

3.3.2 Modelo semi-paramétrico de riscos proporcionais de Cox

No modelo de riscos proporcionais de Cox para dados contínuos, de acordo com [Lee e Wang \(2003\)](#), a função de sobrevivência para uma observação com covariáveis $\mathbf{x}$ é dada por

$$S_{\mathbf{x}}(t) = S_0(t)^{\exp(\mathbf{x}^\top \boldsymbol{\beta})}, \quad (3.3)$$

em que $\boldsymbol{\beta}$ são os parâmetros do modelo.

A função de risco $h(t)$ de um modelo de sobrevivência é definida pela probabilidade “instantânea” do evento de interesse ocorrer no tempo t , dado que o evento não ocorreu até o tempo t . Em termos mais formais, tem-se

$$\begin{aligned} h(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t)}{P(T > t) \Delta t} \\ &= \frac{1}{P(T > t)} \cdot \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t)}{\Delta t} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \quad (3.4)$$

em que $f(t)$ é a função densidade de probabilidade do tempo até o evento de interesse, ou seja, $f(t) = \frac{d}{dt}P(T < t) = \frac{d}{dt}(1 - S(t)) = -S'(t)$. Por fim, voltando na [Equação 3.4](#), tem-se que

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log(S(t)).$$

Novamente, voltando em [3.3](#), é obtido que

$$h_{\mathbf{x}}(t) = -\frac{d}{dt} \exp(\mathbf{x}^\top \boldsymbol{\beta}) \log(S_0(t)) = \exp(\mathbf{x}^\top \boldsymbol{\beta}) h_0(t).$$

Assim, para duas observações com covariáveis $\mathbf{x}_1$ e $\mathbf{x}_2$, tem-se que a razão entre os riscos nesse modelo é dada por

$$\frac{h_{\mathbf{x}_1}}{h_{\mathbf{x}_2}} = \frac{\exp(\mathbf{x}_1^\top \boldsymbol{\beta}) h_0(t)}{\exp(\mathbf{x}_2^\top \boldsymbol{\beta}) h_0(t)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)^\top \boldsymbol{\beta}),$$

que é uma constante. Assim, nesse modelo, a razão entre os riscos de duas observações com diferentes valores das variáveis explicativas não varia em função do tempo analisado t , mas somente das covariáveis $\mathbf{x}_1$ e $\mathbf{x}_2$.

3.3.3 Modelo semi-paramétrico de chances proporcionais

No modelo de chances proporcionais, tem-se que a relação entre a função de sobrevivência de um indivíduo, o vetor de covariáveis $\mathbf{x}$ associado a ele e a função de sobrevivência basal é

dada por

$$\frac{S_{\mathbf{x}(t)}}{1 - S_{\mathbf{x}(t)}} = \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right). \quad (3.5)$$

Assim, tomando observações com vetores de variáveis explicativas $\mathbf{x}_1$ e $\mathbf{x}_2$, tem-se que

$$\frac{\frac{S_{\mathbf{x}_1(t)}}{1 - S_{\mathbf{x}_1(t)}}}{\frac{S_{\mathbf{x}_2(t)}}{1 - S_{\mathbf{x}_2(t)}}} = \frac{\exp(-\mathbf{x}_1^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right)}{\exp(-\mathbf{x}_2^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right)} = \exp((\mathbf{x}_2 - \mathbf{x}_1)^\top \boldsymbol{\beta}).$$

Dessa forma, nesse modelo, a razão de chances de dois indivíduos é uma constante, e não depende do tempo t a ser analisado.

A seguir, encontra-se a expressão da função de sobrevivência em função dos parâmetros neste modelo. De 3.5, tem-se que

$$\begin{aligned} S_{\mathbf{x}}(t) &= \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right) (1 - S_{\mathbf{x}(t)}) \\ &= \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right) - S_{\mathbf{x}(t)} \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right) \implies \\ &\implies \left(1 + \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right) \right) S_{\mathbf{x}}(t) = \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right) \end{aligned}$$

E, por fim,

$$\begin{aligned} S_{\mathbf{x}}(t) &= \frac{\exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right)}{\left(1 + \exp(-\mathbf{x}^\top \boldsymbol{\beta}) \left(\frac{S_0(t)}{1 - S_0(t)} \right) \right)} \\ &= \frac{\left(\frac{\exp(-\mathbf{x}^\top \boldsymbol{\beta}) S_0(t)}{1 - S_0(t)} \right)}{\left(\frac{1 - S_0(t) + \exp(-\mathbf{x}^\top \boldsymbol{\beta}) S_0(t)}{1 - S_0(t)} \right)} \\ &= \frac{\exp(-\mathbf{x}^\top \boldsymbol{\beta}) S_0(t)}{1 - S_0(t) + \exp(-\mathbf{x}^\top \boldsymbol{\beta}) S_0(t)} \\ &= \frac{\exp(-\mathbf{x}^\top \boldsymbol{\beta}) S_0(t)}{1 + (\exp(-\mathbf{x}^\top \boldsymbol{\beta}) - 1) S_0(t)}. \end{aligned}$$

3.3.4 Fragilidade espacial

A fragilidade é uma quantidade aleatória não observada que afeta multiplicativamente o risco do evento de interesse nas observações (DUCROCQ; CASELLA, 1996). Sendo assim, é um efeito aleatório v_i que modela a heterogeneidade das observações após a inclusão das variáveis explicativas, de forma que o preditor linear é aumentado para $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + v_i$, $i = 1, \dots, n$ (MITRA; MÜLLER *et al.*, 2015). Em geral, é comum o uso de modelos de sobrevivência com efeitos aleatórios cujos termos de fragilidade são independentes e identicamente distribuídas quando o evento de ocorrência pode afetar o indivíduo múltiplas vezes (YAU, 2001).

Segundo [Mitra, Müller et al. \(2015\)](#), porém, uma vez que as observações de um modelo de sobrevivência são espacialmente correlacionadas, um dos interesses pode ser investigar a possível dependência espacial entre os tempos até as ocorrências do evento, depois de ajustar o modelo para os efeitos conhecidos de covariáveis específicas das unidades de observação. Nesse sentido, uma das abordagens possíveis é a utilização da fragilidade espacial v , considerando que $v \sim \text{CAR}(\tau)$ ou $v \sim \text{ICAR}(\tau)$, por exemplo.

3.3.5 Modelo semi-paramétrico com fragilidade espacial ICAR

3.3.5.1 Riscos proporcionais

O artigo de [Zhou, Hanson e Zhang \(2020\)](#) apresenta que, no modelo de riscos proporcionais com fragilidade espacial, a função de sobrevivência é dada por

$$S_{\mathbf{x}_{ij}}(t) = S_0(t)^{\exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + v_i)},$$

em que $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ é o vetor referente aos p coeficientes da regressão, v_i é a fragilidade (não observada) associada à i -ésima região $\mathbf{s}_i$ e $S_0(t)$ é a função de sobrevivência basal, isto é, quando $\mathbf{x}_{ij} = \mathbf{0}$ e $v_i = 0$.

3.3.5.2 Chances proporcionais

Em relação ao modelo com chances proporcionais, de acordo com [Zhou, Hanson e Zhang \(2020\)](#), tem-se que a função de sobrevivência é dada por

$$S_{\mathbf{x}_{ij}}(t) = \frac{\exp(-\mathbf{x}_{ij}^\top \boldsymbol{\beta} - v_i) S_0(t)}{1 + (\exp(-\mathbf{x}_{ij}^\top \boldsymbol{\beta} - v_i) - 1) S_0(t)}.$$

3.3.5.3 Distribuições a priori dos parâmetros

O pacote *spBayesSurv* utilizado ([ZHOU; HANSON; ZHANG, 2020](#)) implementa as seguintes distribuições *a priori*:

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, W_0),$$

$$S_0(\cdot) | \alpha, \boldsymbol{\theta} \sim TBP_L(\alpha, S_\theta(\cdot)), \alpha \sim \Gamma(a_0, b_0), \boldsymbol{\theta} \sim N_2(\boldsymbol{\theta}_0, V_0),$$

e

$$(\nu_1, \dots, \nu_m)^\top | \tau \sim \text{ICAR}(\tau^2), \tau^{-2} \sim \Gamma(a_\tau, b_\tau),$$

em que $N_p(\boldsymbol{\mu}, \Sigma)$ denota uma distribuição normal p -variada com média $\boldsymbol{\mu}$ e matriz de variâncias e covariâncias Σ , e $\Gamma(a, b)$ uma distribuição gama com média $\frac{a}{b}$. Além disso, a distribuição *a priori* $TBP_L(\alpha, S_\theta(\cdot))$ é discutida a seguir.

Uma das formas de modelar $S_0(\cdot)$ na abordagem bayesiana é usar distribuições *a priori* não paramétricas (CHEN; HANSON; ZHANG, 2014). Uma delas é a *Transformed Bernstein Polynomial* (TBP). Para um inteiro positivo fixo L , a distribuição *a priori* $TBP_L(\alpha, S_\theta(\cdot))$ é definida como

$$S_0(t) = \sum_{j=1}^L w_j I(S_\theta(t)|j, L-j+1), \quad \mathbf{w}_L \sim \text{Dirichlet}(\alpha, \dots, \alpha),$$

em que $\mathbf{w}_L = (w_1, \dots, w_L)^\top$ é um vetor de pesos positivos, $I(\cdot|a, b)$ é a distribuição acumulada da Beta com parâmetros a e b e $\{S_\theta(\cdot) : \theta \in \Theta\}$ é uma família paramétrica de funções de sobrevivência com suporte nos reais positivos $\mathbb{R}^+$. Três delas estão implementadas na função *survregbayes* do pacote *spBayesSurv*:

- Log-logística:

$$S_\theta(t) = \{1 + (e^{\theta_1} t)^{\exp(\theta_2)}\}^{-1};$$

- Log-normal:

$$S_\theta(t) = 1 - \Phi\{(\log(t) + \theta_1)\exp(\theta_2)\};$$

- Weibull:

$$S_\theta(t) = 1 - \exp\left\{-(e^{\theta_1} t)^{\exp(\theta_2)}\right\}.$$

Assim, $\theta = (\theta_1, \theta_2)^\top$. Além disso, a distribuição de $S_0(\cdot)$ está centrada em $S_\theta(\cdot)$, ou seja, $E(S_0(t)|\alpha, \theta) = S_\theta(t)$. O parâmetro α controla o quão perto os pesos $\mathbf{w}_L$ estão da média da distribuição $\text{Dirichlet}(\alpha, \dots, \alpha)$, $\frac{1}{L}$, pois, quanto maior o valor de α , menores as variâncias e covariâncias dessa distribuição. Isso está diretamente relacionado com o quão perto o formato da função de sobrevivência *baseline* $S_0(\cdot)$ está em relação à distribuição *a priori* $S_\theta(\cdot)$, pois $w_j = \frac{1}{L}$, para $j = 1, \dots, L$ implica que $S_0(t) = S_\theta(t)$ para $t \geq 0$ (ZHOU; HANSON; ZHANG, 2020). Conforme $\alpha \rightarrow \infty$, $S_0(\cdot) \rightarrow S_\theta(\cdot)$ com probabilidade 1.

3.3.5.4 Implementação do método Monte Carlo Markov Chain (MCMC)

A implementação do algoritmo de *Monte Carlo Markov Chain* (MCMC) do modelo semi-paramétrico com fragilidade ICAR utilizado neste trabalho é apresentada em Zhou e Hanson (2018) e descrita a seguir.

Sejam $s_1, \dots, s_m$ as m localizações espaciais distintas de onde foram coletadas as N observações, n_i o número de unidades experimentais na i -ésima região ($N = \sum_{i=1}^m n_i$), t_{ij} o tempo de sobrevivência da j -ésima unidade experimental na região s_i e $\mathbf{x}_{ij}$ o vetor p -dimensional referente às covariáveis utilizadas. Assumindo que os tempos de sobrevivência t_{ij} estão no intervalo (a_{ij}, b_{ij}) , $0 \leq a_{ij} \leq b_{ij} \leq \infty$, então os cenários de censura à direita são da forma (a_{ij}, ∞) , em que a_{ij} é o maior tempo avaliado em que é conhecido que o evento ainda não aconteceu para

a j -ésima observação da i -ésima localização s_i . Além disso, para os casos em que o tempo de sobrevivência é conhecido, tem-se que $a_{ij} = b_{ij}$ (ZHOU; HANSON, 2018).

Dessa forma, sendo $\mathcal{D} = \{(a_{ij}, b_{ij}, \mathbf{x}_{ij}, s_i; i = 1, \dots, m; j = 1, \dots, n_i)\}$ a representação dos dados observados e assumindo as distribuições *a priori* detalhadas na Subsubseção 3.3.5.3, tem-se que a função de verossimilhança para $(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu})$ é dada por

$$L(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu}) = \prod_{i=1}^m \prod_{j=1}^{n_i} [S_{x_{ij}}(a_{ij}) - S_{x_{ij}}(b_{ij})]^{I\{a_{ij} < b_{ij}\}} f_{x_{ij}}(a_{ij})^{I\{a_{ij} = b_{ij}\}}.$$

Assim, a densidade *a posteriori* é dada por

$$p(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu}, \alpha, \tau^2 | \mathcal{D}) \propto L(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu}) p(\mathbf{w}_L | \alpha) p(\alpha) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}) p(\boldsymbol{\nu} | \tau^2) p(\tau^2), \quad (3.6)$$

em que $p(\cdot)$ denota cada distribuição *a priori* definida como na Subsubseção 3.3.5.3.

Ressalta-se que $w_j = \frac{1}{L}$ implica que $S_0(t) = S_\theta(t)$, ou seja, o caso paramétrico, então o modelo paramétrico estabelece bons chutes iniciais para o modelo de sobrevivência considerando a distribuição *a priori Transformed Bernstein Polynomial* (TBP) (ZHOU; HANSON, 2018). No cenário do estudo, utiliza-se tanto o modelo paramétrico com fragilidade espacial quanto sem.

Considerando a densidade *a posteriori* dos parâmetros conforme visto em 3.6, tem-se que

$$\begin{aligned} p(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu}, \alpha, \tau^2 | \mathcal{D}) &\propto L(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\ &\times \frac{\Gamma(\alpha L)}{\Gamma(\alpha)^L} \prod_{j=1}^L (w_j)^{\alpha-1} \\ &\times \alpha^{a_\alpha-1} \exp(-b_\alpha \alpha) \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' V_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\} \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' W_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\ &\times \left(\tau^{-2}\right)^{\frac{\text{posto}(C)}{2}} \exp\left\{-\frac{1}{2\tau^2} \boldsymbol{\nu}' C \boldsymbol{\nu}\right\} \\ &\times (\tau^{-2})^{a_\tau-1} \exp(-b_\tau \tau^{-2}), \end{aligned}$$

em que $a_\alpha = b_\alpha = 1$, $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}$ é o estimador paramétrico de $\boldsymbol{\theta}$, $V_0 = 10\hat{V}$, $\hat{V}$ é a matriz de covariâncias estimada de $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $W_0 = 10^{10}I_p$, I_p a matriz identidade ordem p , p o número de parâmetros do modelo, $C = D - E$, sendo D é matriz diagonal $m \times m$ cujos elementos da diagonal correspondem à quantidade de vizinhos da i -ésima região, ou seja, $d_{ii} = e_{i+} = \sum_{j=1}^m e_{ij}$ e, por fim, $a_\tau = b_\tau = 0,001$.

Em seguida, seguem-se os seguintes passos:

Passo 1: Atualizar $\mathbf{w}_L$.

Seja $\mathbf{z}_{L-1} = (z_1, \dots, z_{L-1})$, considerando a transformação $z_j = \log(w_j) - \log(w_L)$. Então a distribuição condicional completa para $\mathbf{z}_{L-1}$ é dada por

$$p(\mathbf{z}_{L-1} | \Theta \setminus \mathbf{z}_{L-1}) \propto L(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \nu) \times \prod_{j=1}^L \left[\frac{e^{z_j}}{\sum_{k=1}^L e^{z_k}} \right]^\alpha,$$

em que $z_L = 0$ e $\Theta \setminus \mathbf{z}_{L-1}$ representa o conjunto formato por todos os parâmetros (Θ), exceto $\mathbf{z}_{L-1}$. Assim, o vetor $\mathbf{z}_{L-1}$ pode ser adaptado usando o método *Adaptive Metropolis* (Metropolis Adaptativo), utilizado quando a condição de reversibilidade das distribuições do método Metropolis-Hastings não é satisfeita (HAARIO; SAKSMAN; TAMMINEN, 2001). Nessa abordagem, a variância da distribuição proposta depende de todos os valores já amostrados.

Suponha que o algoritmo encontra-se na l -ésima iteração, ou seja, amostrou os estados $\mathbf{z}_{L-1}^{(1)}, \dots, \mathbf{z}_{L-1}^{(l-1)}$. Seleciona-se um índice l_0 para o tamanho de um período inicial e define-se

$$\Sigma_l = \begin{cases} \Sigma_0, & l \leq l_0 \\ \frac{(2,4)^2}{L-1} (C_l + 10^{-10} \mathbf{I}_{L-1}), & l > l_0, \end{cases}$$

em que C_l é a variância amostral de $\mathbf{z}_{L-1}^{(1)}, \dots, \mathbf{z}_{L-1}^{(l-1)}$ e Σ_0 é uma matriz de covariância diagonal inicial de $\mathbf{z}$, definida de tal forma que a variância de z_j é 0,16. Essa escolha é baseada em estudos extensivos de simulação, visto que outras escolhas, que não muito pequenas ou muito grandes, têm pequenos impactos na inferência da distribuição *a posteriori*, de acordo com o Apêndice A de Zhou e Hanson (2018). Dessa forma, gera-se $\mathbf{z}_{L-1}^* = (z_1^*, \dots, z_{L-1}^*)'$ de uma distribuição $N_{L-1}(\mathbf{z}_{L-1}^{(l-1)}, \Sigma_l)$, que é aceito com probabilidade

$$\min \left\{ 1, \frac{p(\mathbf{z}_{L-1}^* | \Theta \setminus \mathbf{z}_{L-1})}{p(\mathbf{z}_{L-1}^{(l-1)} | \Theta \setminus \mathbf{z}_{L-1})} \right\}.$$

Passo 2: Atualizar $\boldsymbol{\theta}$.

A distribuição condicional completa para $\boldsymbol{\theta}$ é

$$p(\boldsymbol{\theta} | \Theta \setminus \boldsymbol{\theta}) \propto L(\mathbf{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \nu) \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' V_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right\},$$

e é feita a amostragem pelo método Metropolis adaptativo, em que na l -ésima iteração, o valor candidato $\boldsymbol{\theta}^*$ é amostrado de uma distribuição $N_2(\boldsymbol{\theta}^{(l-1)}, \Sigma_l)$ e aceito com probabilidade

$$\min \left\{ 1, \frac{p(\boldsymbol{\theta}^* | \Theta \setminus \boldsymbol{\theta})}{p(\boldsymbol{\theta}^{(l-1)} | \Theta \setminus \boldsymbol{\theta})} \right\},$$

sendo Σ_l definida como no passo 1, porém com $\Sigma_0 = \hat{V}$.

Passo 3: Atualizar β .

A distribuição condicional de β é dada por

$$p(\beta|\Theta \setminus \beta) \propto L(\mathbf{w}_L, \theta, \beta, \nu) \times \exp \left\{ -\frac{1}{2}(\beta - \beta_0)' W_0^{-1}(\beta - \beta_0) \right\},$$

e é feita a amostragem pelo método Metropolis adaptativo, em que na l -ésima iteração, o valor candidato β^* é amostrado de uma distribuição $N_p(\beta^{(l-1)}, \Sigma_l)$ e aceito com probabilidade

$$\min \left\{ 1, \frac{p(\beta^*|\Theta \setminus \beta)}{p(\beta^{(l-1)}|\Theta \setminus \beta)} \right\},$$

sendo Σ_l definida como no passo 1, porém com $\Sigma_0 = \hat{W}$.

Passo 4: Atualizar α .

A distribuição condicional de α é dada por

$$p(\alpha|\Theta \setminus \alpha) \propto \frac{\Gamma(\alpha L)}{\Gamma(\alpha)^L} \prod_{j=1}^L (w_j)^{\alpha-1} \times \alpha^{a_{\alpha}-1} \exp(-b_{\alpha}\alpha)$$

e é feita a amostragem pelo método Metropolis adaptativo, em que na l -ésima iteração, o valor candidato α^* é amostrado de uma distribuição $N_1(\alpha^{(l-1)}, \Sigma_l)$ e aceito com probabilidade

$$\min \left\{ 1, \frac{p(\alpha^*|\Theta \setminus \alpha)}{p(\alpha^{(l-1)}|\Theta \setminus \alpha)} \right\},$$

sendo Σ_l definida como no passo 1, porém com $\Sigma_0 = 0,16$.

Passo 5: Atualizar ν . Seja $L(\mathbf{w}_L, \theta, \beta, \nu) = \prod_{i=1}^m \prod_{j=1}^{n_i} L_{ij}(\mathbf{w}_L, \theta, \beta, \nu)$. Então a distribuição condicional completa para ν_i , $i = 1, \dots, m$ é dada por

$$p(\nu_i|\Theta \setminus \nu_i) \propto \prod_{j=1}^{n_i} L_{ij}(\mathbf{w}_L, \theta, \beta, \nu) \times \exp \left\{ -\frac{e_{i+}}{2\tau^2} \left(\nu_i - \sum_{j=1}^m \frac{e_{ij}\nu_j}{e_{i+}} \right)^2 \right\}.$$

Assim, ν_i é adaptado por meio dos passos de Metropolis-Hasting com distribuição proposta $\nu_i^* \sim N(\nu_i^{(l-1)}, \frac{\tau^2}{e_{i+}})$. O valor candidato ν_i^* é aceito com probabilidade

$$\min \left\{ 1, \frac{p(\nu_i^*|\Theta \setminus \nu_i)}{p(\nu_i^{(l-1)}|\Theta \setminus \nu_i)} \right\}.$$

Após a atualização de cada fragilidade individual, o vetor ν é atualizado para manter a restrição $\sum_{i=1}^m \nu_i = 0$ por meio da transformação $\nu \leftarrow \nu - \frac{1}{m} \mathbf{1}_m' \nu$. Reitera-se que a estratégia é criada por um propósito específico, podendo ser considerada como *ad hoc*, porém tem efeito insignificante na distribuição *a posteriori* e foi defendida por diversos autores, conforme apresentado no Apêndice A de [Zhou e Hanson \(2018\)](#).

Passo 6: Atualizar τ^{-2} .

A distribuição condicional de τ^{-2} é dada por

$$p(\tau^{-2} | \Theta \setminus \tau) \propto (\tau^{-2})^{a_\tau^*} \times \exp\{-b_\tau^* \tau^{-2}\},$$

em que $a_\tau^* = a_\tau + \frac{\text{posto}(C)}{2} - 1$ e $b_\tau^* = b_\tau + \frac{1}{2} \nu' C \nu$. Dessa forma, τ^{-2} é amostrado de uma distribuição $\Gamma(a_\tau^*, b_\tau^*)$.

Mais a respeito da abordagem utilizada para estimação dos valores das distribuições *a posteriori* por meio do método MCMC pode ser encontrado no Apêndice A de [Zhou e Hanson \(2018\)](#).

No presente estudo, foram utilizadas 10.000 iterações, sendo 2.000 utilizadas como *burn-in*. Considerando o tamanho do salto como 4, foram obtidos 2.000 valores simulados.

3.3.5.5 Métricas avaliativas do modelo

Como critérios de seleção de modelos, foram utilizados o Critério de Informação de Desvio (*Deviance Information Criterion* - DIC) ([SPIEGELHALTER et al., 2002](#)), o Critério de Informação de Watanabe-Akaike (*Watanabe-Akaike Information Criterion* - WAIC) ([WATANABE; OPPER, 2010](#)) - quanto menores, melhor indicativo - e o Log da Pseudo Verossimilhança Marginal (*Log Pseudo Marginal Likelihood* - LPML) ([GEISSER; EDDY, 1979](#)) - quanto maior, melhor indicativo. A fim de avaliar a qualidade de ajuste do modelo, utiliza-se os resíduos de Cox-Snell ([COX; SNELL, 1968](#)). Quanto mais próximos os valores estão da reta identidade, melhor é o ajuste do modelo. E, por fim, para a investigação da convergência do método MCMC aplicado, considera-se o critério de Gelman e Rubin ([GELMAN; RUBIN, 1992](#)), que, quanto mais próximo de 1, maior é o indício de que as cadeias convergiram.

SIMULAÇÃO

A fim de compreender se o modelo utilizado é capaz de acessar adequadamente os verdadeiros parâmetros, é feita uma simulação de dados. Inicialmente, ajusta-se um modelo de Cox com distribuição exponencial aos dados originais. Em seguida, utiliza-se os parâmetros obtidos para simular dados, de acordo também com os dados originais. A partir deles, ajusta-se um modelo de Cox com distribuição exponencial e avalia-se medidas como viés, erro quadrático médio, erro-padrão e probabilidade de cobertura.

Para investigar como o modelo se comporta com a variação do tamanho amostral, toma-se amostras de tamanho $n_1^* = 1.000$, $n_2^* = 5.000$, $n_3^* = 10.000$ e o tamanho original dos dados, ou seja, $n_4^* = 17.817$. Essas amostras são tomadas de forma estratificada, ou seja, respeitando a proporção dos casos (homicídios em cada município) observada na amostra original. Para cada tamanho amostral, simulam-se $m_1^* = 1.000$ e $m_2^* = 5.000$ conjuntos de dados e, para cada um deles, ajusta-se o modelo de Cox com distribuição exponencial. A escolha de m_1^* e m_2^* se deu a fim de verificar o comportamento das medidas consideradas conforme o número de simulações aumenta. A execução dos códigos foi realizada por meio dos recursos do *cluster* Aguiar4, disponibilizado pela Superintendência de Tecnologia da Informação da Universidade de São Paulo (USP).

4.0.1 Aspectos teóricos

A função de sobrevivência de um modelo de riscos proporcionais de Cox é dada por

$$S(t|x) = \exp(-H_0(t) \exp(\beta^T x)).$$

Dessa forma, a função distribuição acumulada é dada por

$$F(t|x) = 1 - \exp(-H_0(t) \exp(\beta^T x)).$$

Seja Y uma variável aleatória com função distribuição F , então $U = F(Y)$ segue uma distribuição uniforme entre 0 e 1, ou seja, $U \sim \text{Uniforme}(0, 1)$ (GENTLE, 2013). Além disso, se

$U \sim \text{Uniforme}(0, 1)$, então $(1 - U) \sim \text{Uniforme}(0, 1)$ (GENTLE, 2013). Sendo assim, sendo T o tempo de sobrevivência de um modelo de Cox, tem-se que

$$U = \exp(-H_0(t) \exp(\beta^\top x)) \sim \text{Uniforme}(0, 1).$$

Na sequência, segundo Bender, Augustin e Blettner (2005), se a função de risco basal $h_0(t) > 0, \forall t$, então a função de risco acumulado H_0 é inversível e o tempo T de sobrevivência do modelo de Cox pode ser expressa como

$$T = H_0^{-1}(-\log(U) \exp(-\beta^\top x)). \quad (4.1)$$

Para a distribuição exponencial com parâmetro $\lambda > 0$, tem-se que a sua função de risco basal é dada por

$$h_0(t) = \lambda$$

e a sua função de risco acumulado é equivalente a

$$H_0(t) = \lambda t.$$

Dessa forma, a inversa da função de risco acumulado é igual a

$$H_0^{-1}(t) = \frac{t}{\lambda}$$

e a fórmula para simular tempos de sobrevivência em um modelo de Cox usando a distribuição exponencial com parâmetro $\lambda > 0$, de acordo com Equação 4.1, é

$$T = -\frac{\log(U) \exp(-\beta^\top x)}{\lambda} = -\frac{\log(U)}{\lambda \exp(\beta^\top x)}. \quad (4.2)$$

Gera-se, então, um valor u da distribuição $U \sim \text{Uniforme}(0, 1)$ e, dados os vetores de covariáveis x e de parâmetros β , bem como o parâmetro λ da distribuição exponencial, é possível gerar, a partir da Equação 4.2, valores de tempos de sobrevivência de um modelo de Cox com distribuição exponencial.

O código utilizado para a modelagem e simulação de dados está disponibilizado em <https://github.com/ferbartolo/projeto-CIL>.

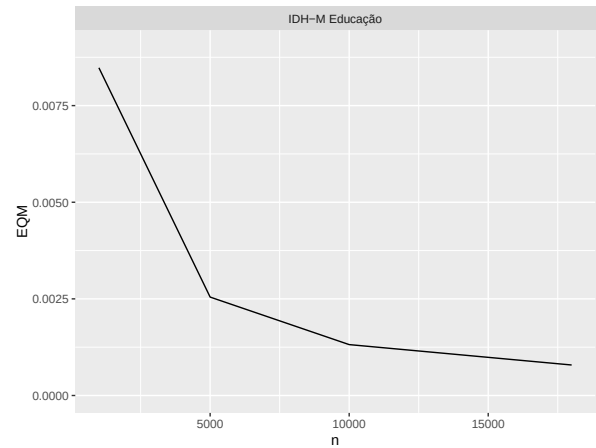
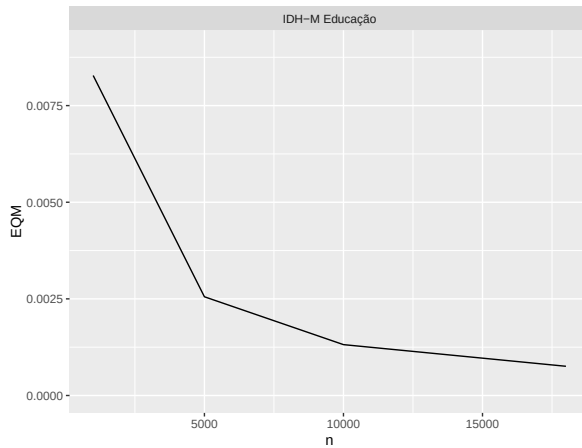
4.0.2 Resultados

A seguir, apresentam-se os resultados da simulação. Reitera-se que essa análise foi feita a fim de verificar o comportamento de medidas como erro quadrático médio (EQM), erro-padrão, viés e probabilidade de cobertura, conforme o número de simulações aumenta.

Nas Figuras numeradas de 12 a 16, tem-se o erro quadrático médio (EQM) para cada parâmetro estimado de acordo com o modelo selecionado, considerando $m_1^* = 1.000$ e $m_2^* = 5.000$ conjuntos de dados simulados.

Figura 12 – Erro quadrático médio (EQM) para IDH-M em relação à educação.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

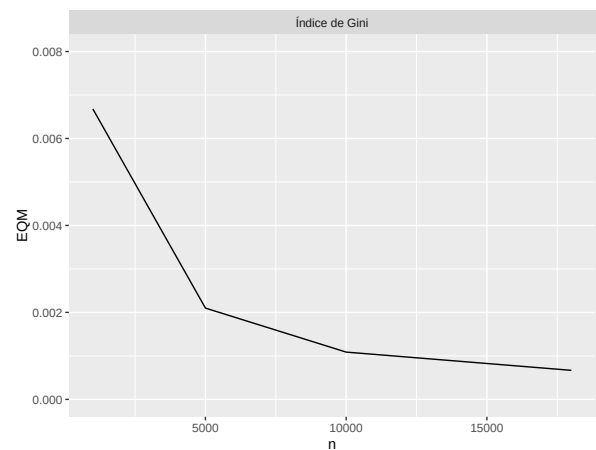
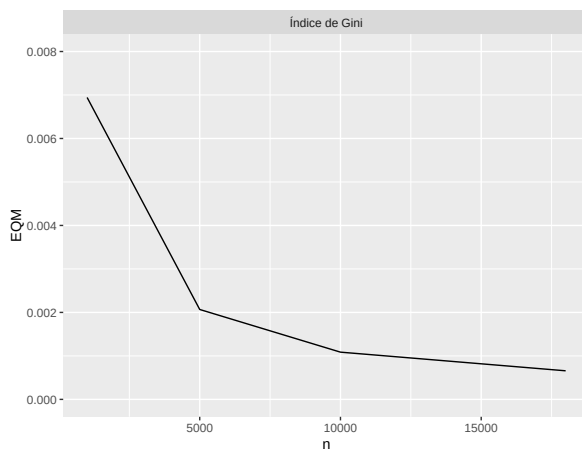


Fonte: Elaborada pela autora.

É possível perceber que o EQM de cada parâmetro manteve-se bem semelhante entre as duas simulações - conforme o tamanho da amostra aumenta, o erro quadrático médio do modelo tende a diminuir.

Figura 13 – Erro quadrático médio (EQM) para o índice de Gini.

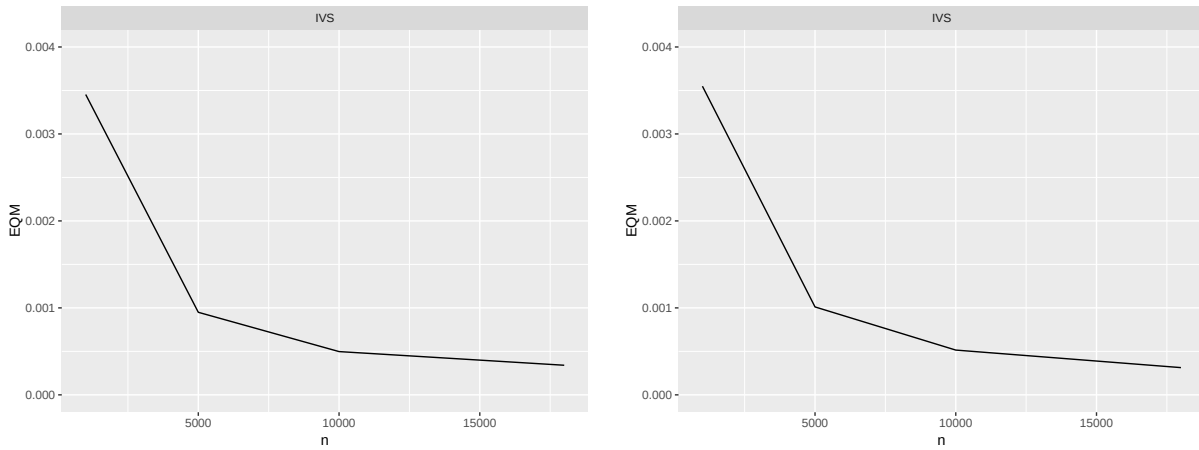
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 14 – Erro quadrático médio (EQM) para o índice de Vulnerabilidade Social (IVS).

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

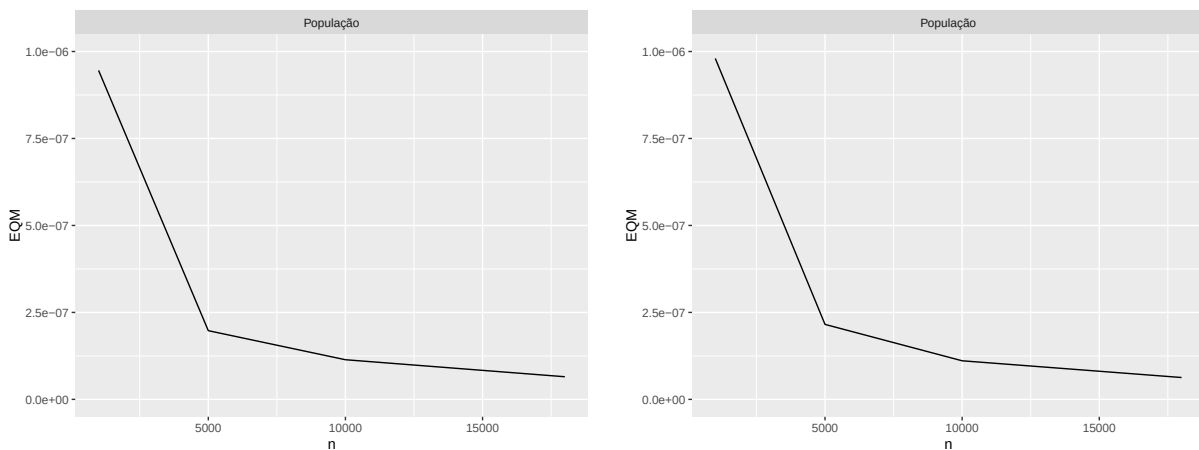


Fonte: Elaborada pela autora.

Esse comportamento está de acordo com o esperado: o erro quadrático médio (EQM) tende a zero conforme o tamanho amostral diminui, o que ocorre tanto para $m_1^* = 1.000$ como para $m_2^* = 5.000$ simulações realizadas. Assim, esse fato corrobora para a capacidade do modelo em acessar os verdadeiros parâmetros.

Figura 15 – Erro quadrático médio (EQM) para a população estimada.

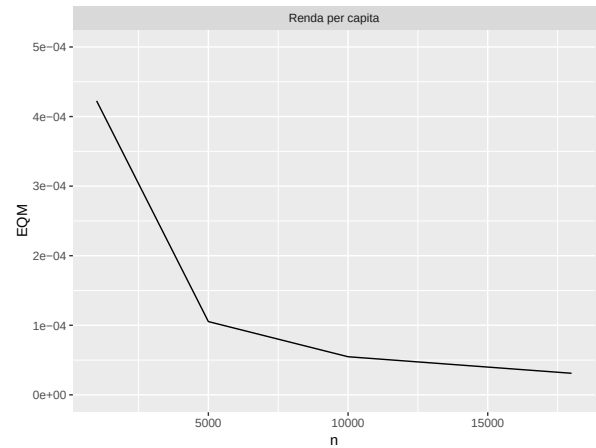
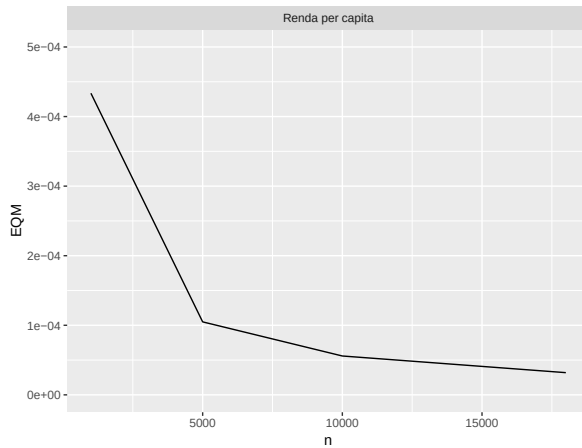
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 16 – Erro quadrático médio (EQM) para a renda per capita.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



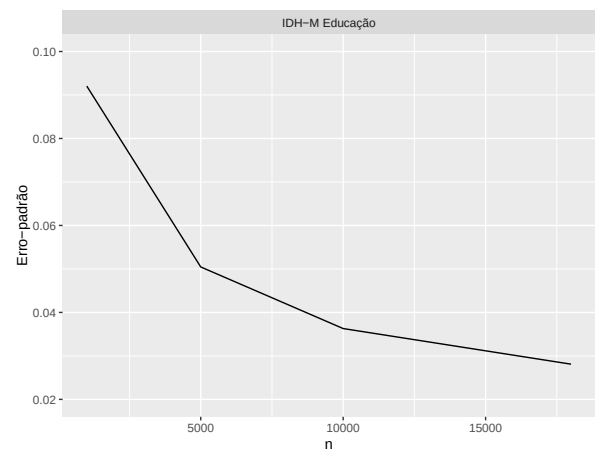
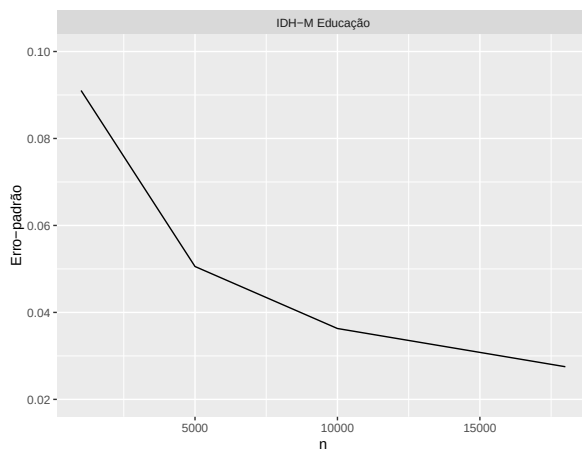
Fonte: Elaborada pela autora.

Nas Figuras numeradas de 17 a 21, apresenta-se o erro-padrão de cada parâmetro estimado, de acordo com os m_1^* e m_2^* conjunto de dados simulados. Novamente, é perceptível que o erro-padrão de cada parâmetro manteve-se semelhante entre as duas simulações.

Da mesma forma, o comportamento da curva do erro-padrão em função do tamanho amostral foi também de acordo com o esperado: conforme o tamanho amostral aumenta, o erro-padrão diminui. Conforme visto, não parece haver diferenças significativas entre os erros-padrão em ambos os cenários para todas as variáveis preditoras.

Figura 17 – Erro-padrão para IDH-M em relação à educação.

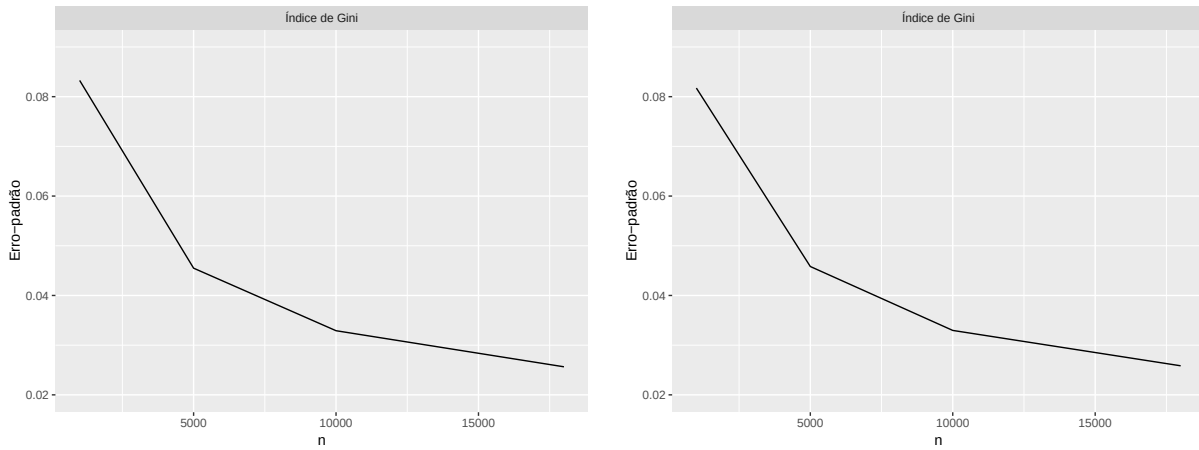
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 18 – Erro-padrão para o índice de Gini.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

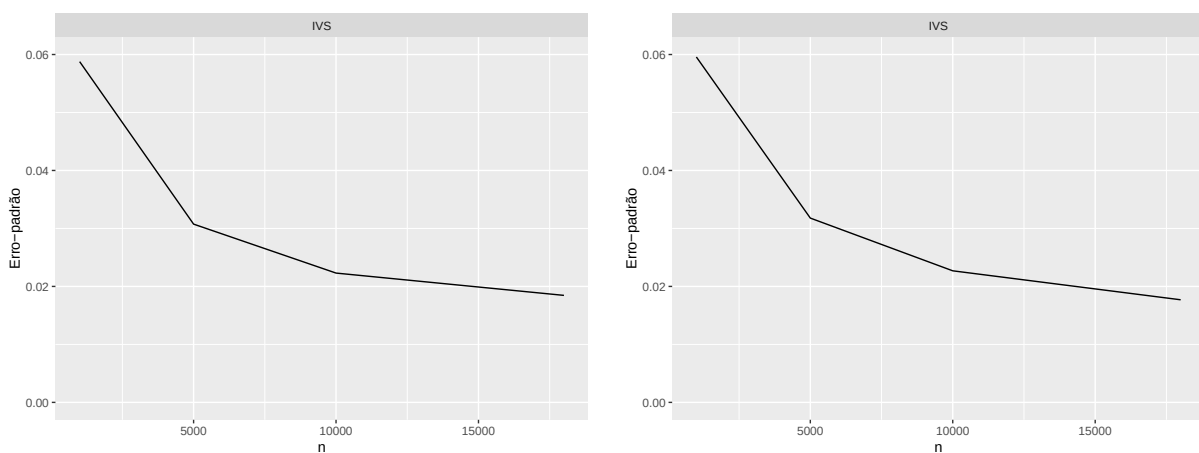


Fonte: Elaborada pela autora.

Novamente, o fato de o erro-padrão diminuir ao longo dos tamanhos amostrais para ambas as simulações realizadas, tendendo a zero, corrobora para a adequabilidade do modelo em acessar os verdadeiros parâmetros.

Figura 19 – Erro-padrão para o índice de Vulnerabilidade Social (IVS).

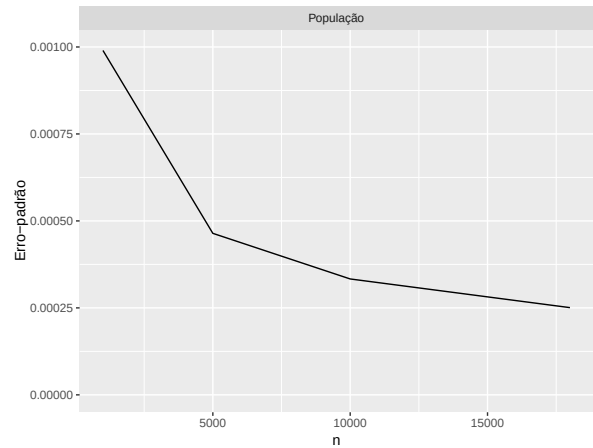
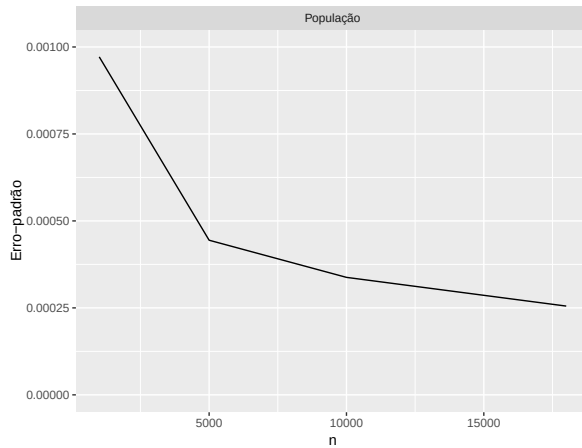
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 20 – Erro-padrão para a população estimada.

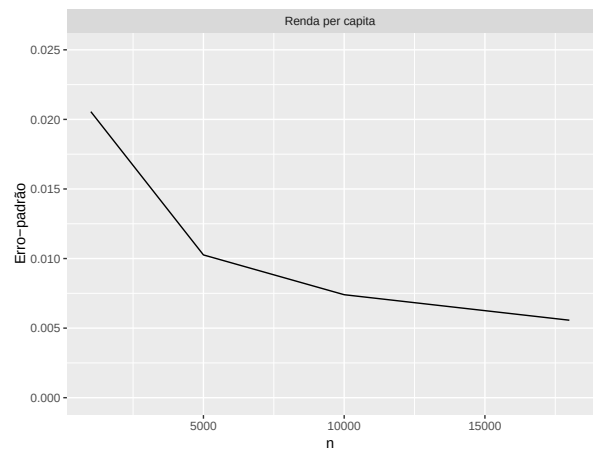
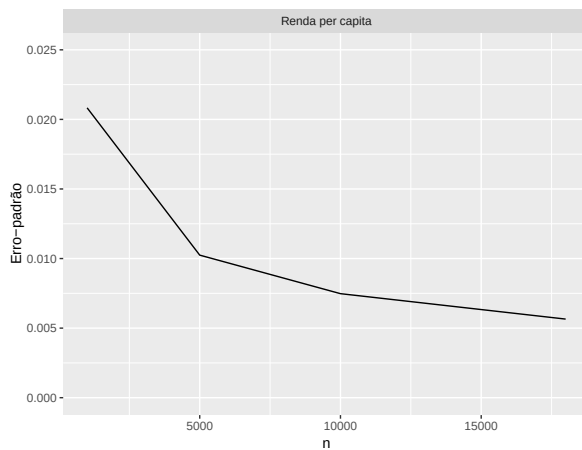
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 21 – Erro-padrão para a renda per capita.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



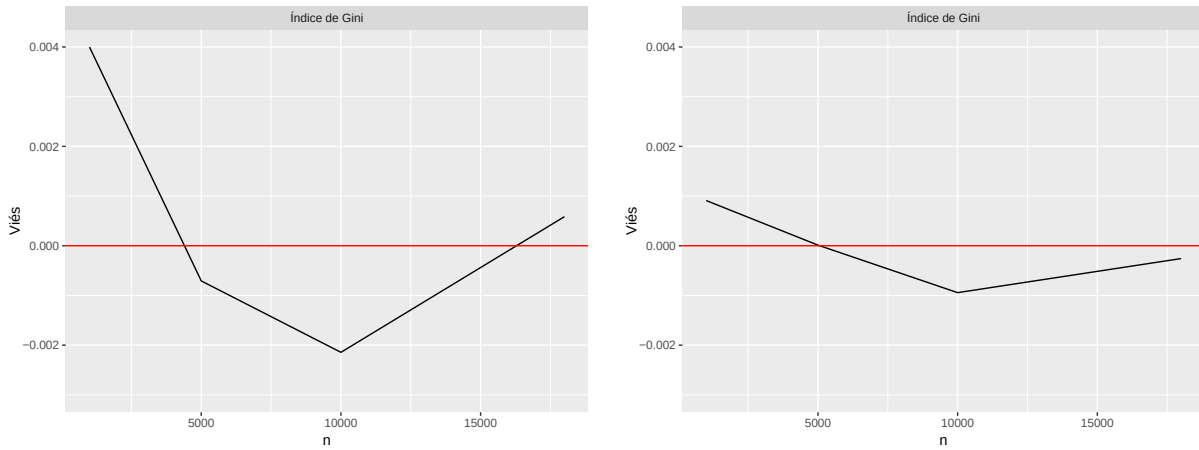
Fonte: Elaborada pela autora.

Agora, nas Figuras de 22 a 26, tem-se o viés para cada parâmetro considerando m_1^* e m_2^* conjuntos de dados simulados. É esperado que, conforme o tamanho amostral aumente, o viés se aproxime de zero.

Além disso, destaca-se em vermelho a reta $y = 0$, a fim de facilitar a visualização da oscilação do viés em torno do zero de acordo com a variação do tamanho amostral nos cenários de m_1^* e m_2^* conjuntos de dados simulados.

Figura 23 – Viés para o índice de Gini.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

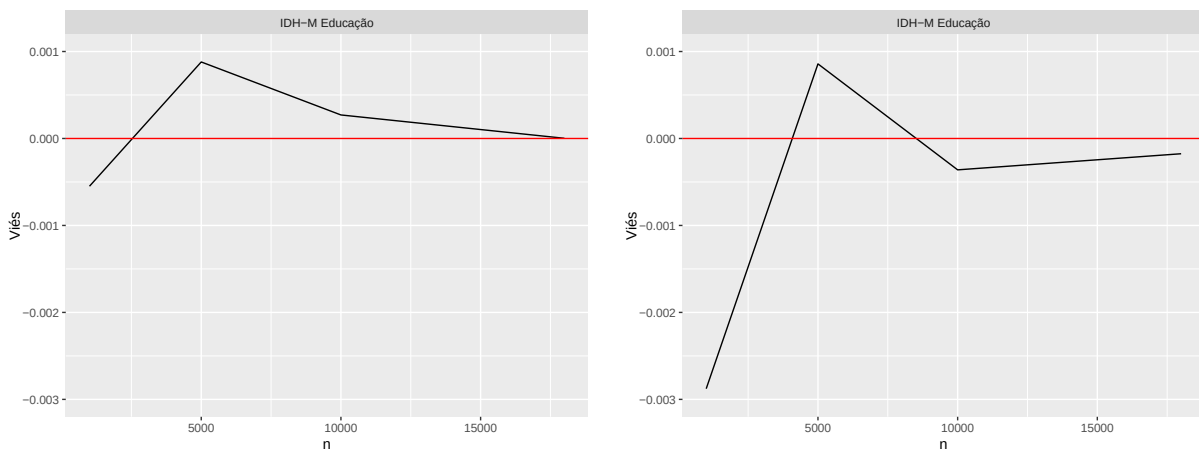


Fonte: Elaborada pela autora.

Observa-se que, para quase todos os parâmetros, com exceção do IDH-M em relação à educação e da renda per capita, ao passar de m_1^* para m_2^* conjuntos de dados, o viés varia menos em torno de zero ao longo dos tamanhos amostrais. Além disso, em geral, os vieses se aproximam mais de zero no cenário m_2^* em relação ao cenário imposto por m_1^* .

Figura 22 – Viés para IDH-M em relação à educação.

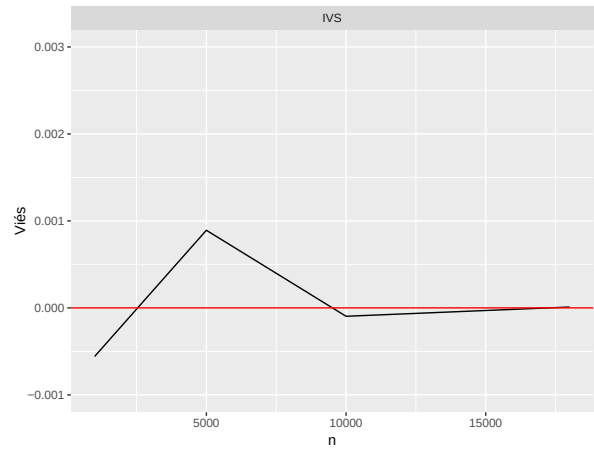
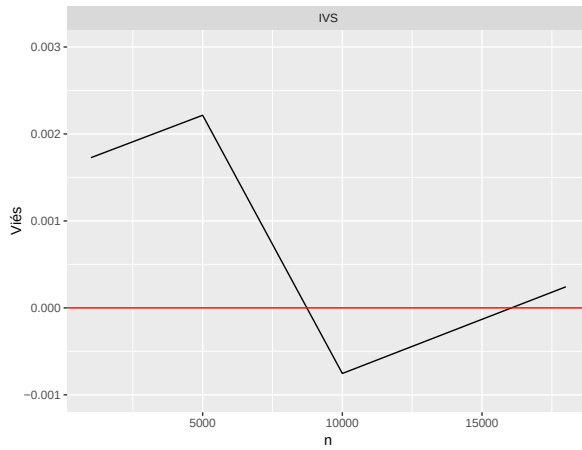
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 24 – Viés para o índice de Vulnerabilidade Social (IVS).

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

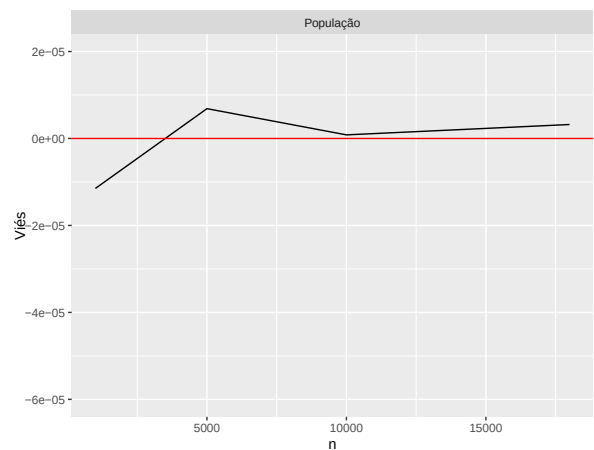
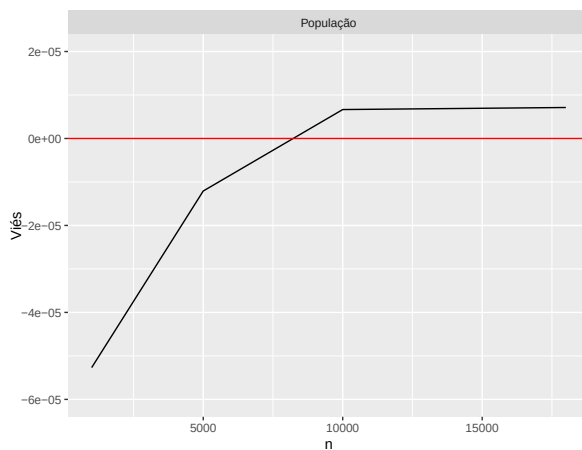


Fonte: Elaborada pela autora.

O fato de o viés variar em torno de zero, em geral, com ainda menos oscilações ao redor dele no cenário de $m_2^* = 5.000$ simulações realizadas em comparação a $m_1^* = 1.000$, também contribui para a adequabilidade do modelo.

Figura 25 – Viés para a população estimada.

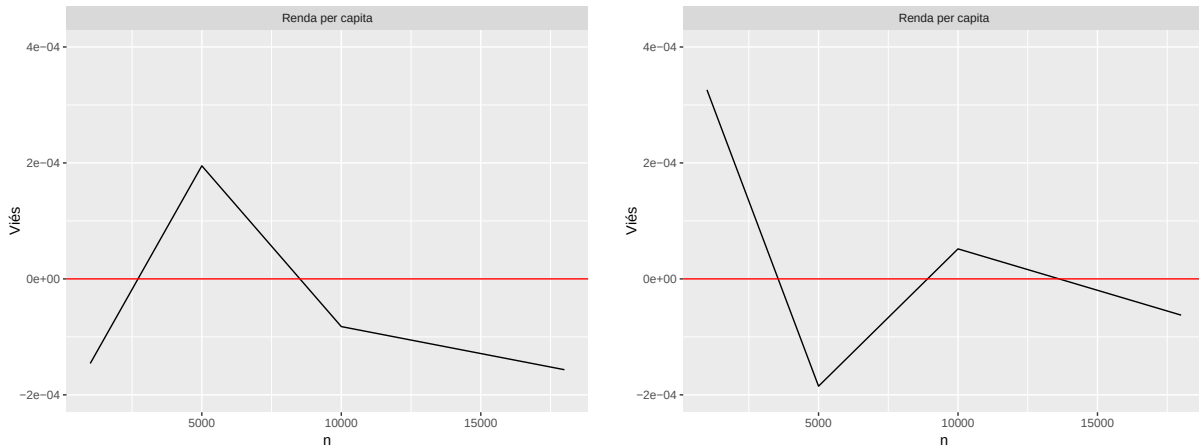
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 26 – Viés para a renda per capita.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

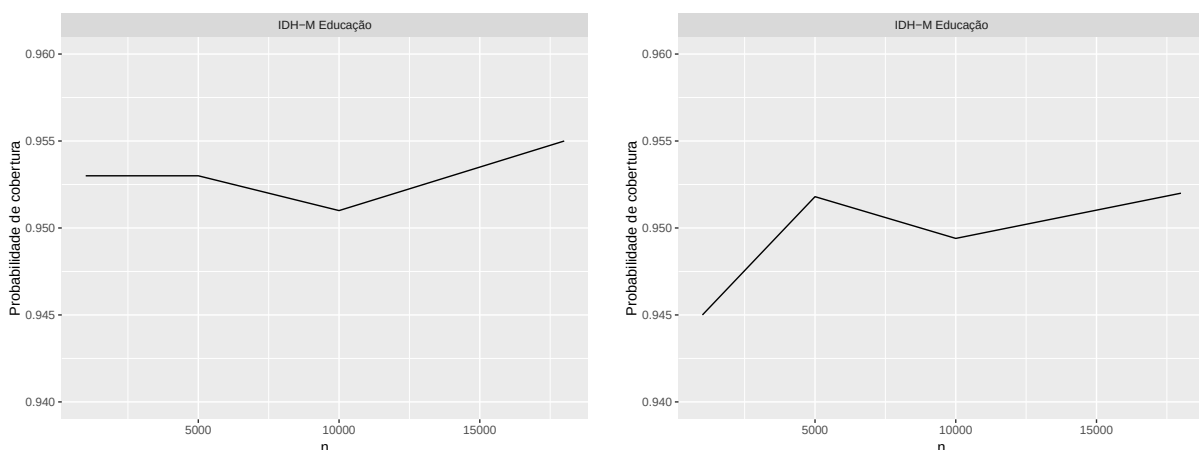


Fonte: Elaborada pela autora.

Por fim, apresenta-se, nas Figuras de 27 a 31, a probabilidade de cobertura para cada parâmetro - IDH-M em relação à educação, índice de Gini, índice de vulnerabilidade social (IVS), população estimada e renda per capita - para os dois cenários considerados ($m_1^* = 1.000$ e $m_2^* = 5.000$ conjuntos de dados simulados).

Figura 27 – Probabilidade de cobertura para IDH-M em relação à educação.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



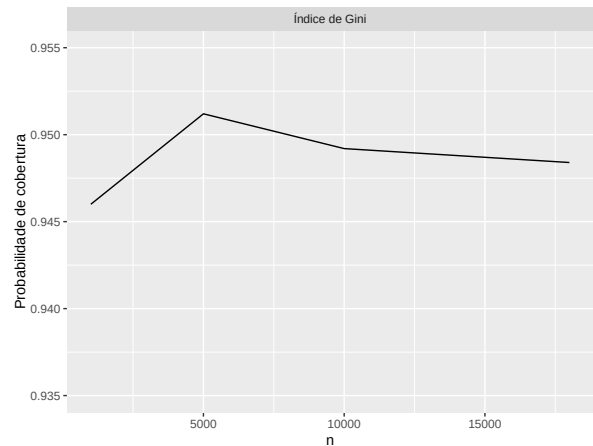
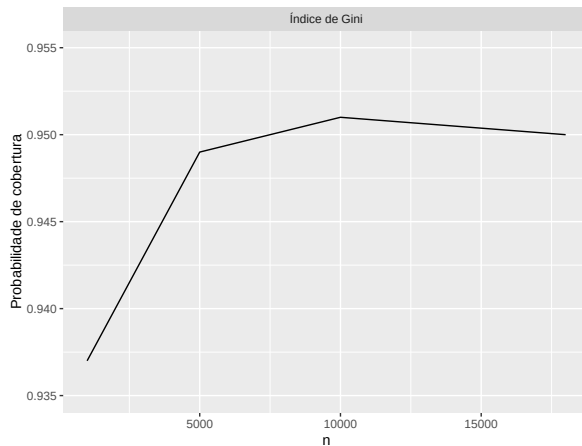
Fonte: Elaborada pela autora.

É possível perceber que, ainda que a probabilidade de cobertura não aumente em todos os cenários, entre os gráficos de $m_1^* = 1.000$ conjuntos de dados simulados e $m_2^* = 5.000$ conjuntos

de dados simulados, nota-se que a probabilidade de cobertura torna-se, em geral, mais estável conforme o tamanho da amostra aumenta, quando considera-se o cenário de m_2^* em relação ao de m_1^* .

Figura 28 – Probabilidade de cobertura para o índice de Gini.

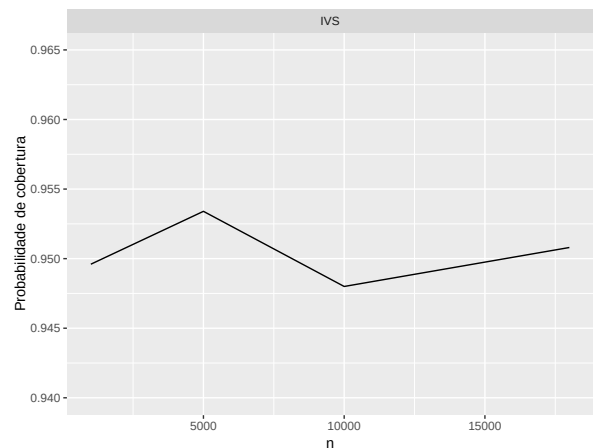
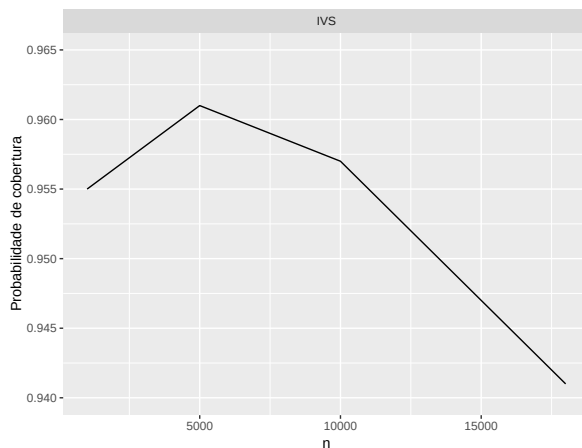
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 29 – Probabilidade de cobertura para o índice de Vulnerabilidade Social (IVS).

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.

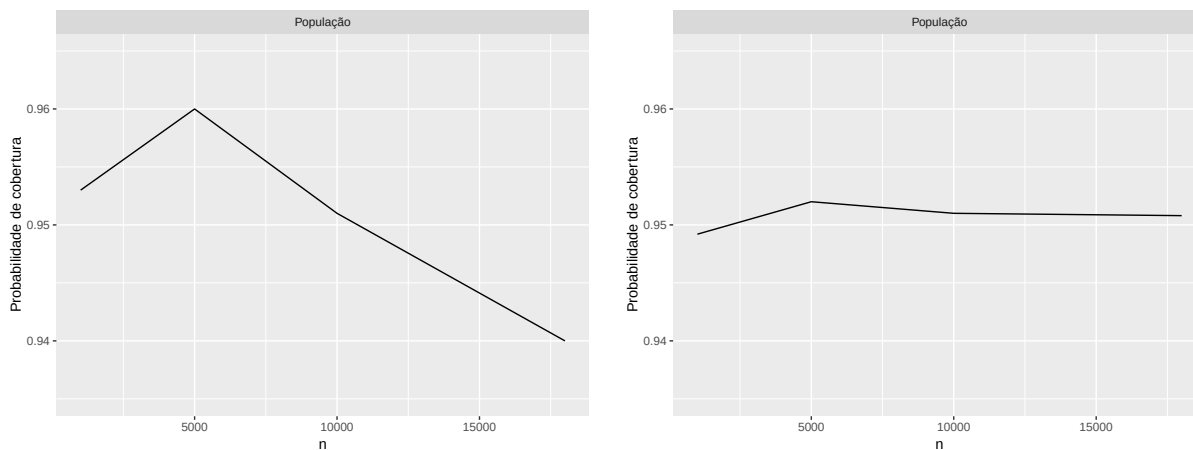


Fonte: Elaborada pela autora.

Dessa forma, de acordo com os gráficos observados a respeito de medidas como o EQM, o erro-padrão, a probabilidade de cobertura e o viés das estimativas dos parâmetros, conclui-se que os resultados estão de acordo com o esperado e comprovam a capacidade do modelo em recuperar os parâmetros adequadamente.

Figura 30 – Probabilidade de cobertura para a população estimada.

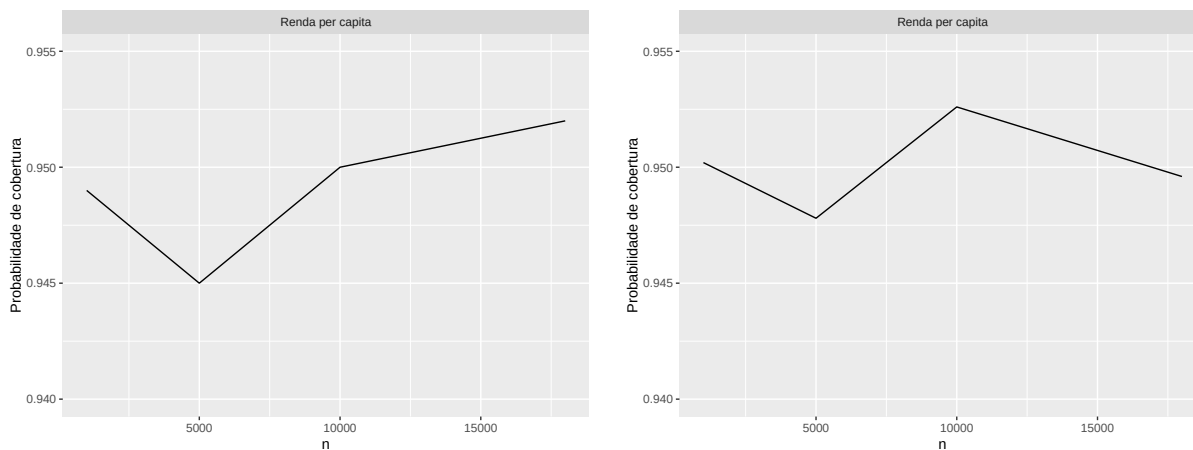
- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

Figura 31 – Probabilidade de cobertura para a renda per capita.

- (a) Considerando $m_1^* = 1.000$ conjuntos de dados simulados. (b) Considerando $m_2^* = 5.000$ conjuntos de dados simulados.



Fonte: Elaborada pela autora.

APLICAÇÃO

A seguir, na [Seção 5.1](#), a forma como os dados foram tratados é apresentada e as devidas limitações da estratégia abordada são discutidas. Na [Seção 5.2](#), a análise visual das curvas de sobrevivência é apresentada, bem como a visualização no mapa do índice de impunidade proposto. Além disso, os resultados, as medidas avaliativas, os resíduos e as estimativas dos modelos ajustados são apresentados e discutidos.

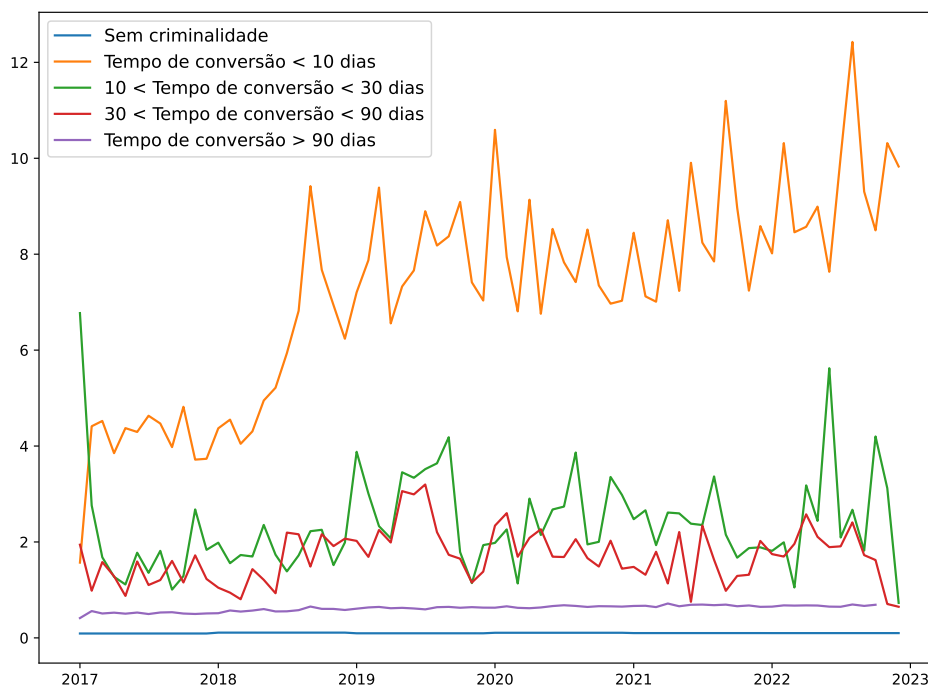
5.1 Tratamento dos dados

A fim de considerar o tempo até a conversão do boletim de ocorrência (BO) em inquérito policial (IP) dos casos de homicídios dolosos ocorridos no estado de São Paulo entre 2017 e 2022, foi feita uma associação entre a base dos BOs e dos IPs da seguinte forma: associou-se a cada observação de notificação de crime o primeiro inquérito instaurado na mesma cidade após a abertura do BO. Ressalta-se que a abordagem foi utilizada devido à falta de correspondência entre ambas as bases de dados, ou seja, não há uma coluna identificadora nas informações referentes aos inquéritos que mencione à qual boletim de ocorrência o crime está associado.

De acordo com o tratamento de dados considerado, obtém-se o tempo de conversão de BO em inquérito de cada notificação de crime em cada município do estado no período de estudo. Com isso, toma-se o tempo mediano de conversão para cada cidade ao mês. Assim, formam-se “grupos” ao longo do tempo, que possuem índices semelhantes de impunidade, a saber “Sem criminalidade no período”, “Tempo de conversão menor do que 10 dias”, “Tempo de conversão entre 10 e 30 dias”, “Tempo de conversão entre 30 e 90 dias” e “Tempo de conversão maior do que 90 dias”, sendo que esse último cenário inclui os casos em que não houve conversão, ou seja, são casos de censura conforme retratado no [Capítulo 3](#). Dessa forma, na [Figura 32](#), apresenta-se o tamanho médio populacional (em 100 mil habitantes) de tais grupos analisado por mês.

Percebe-se que o grupo com maior tamanho populacional médio é o que possui menor

Figura 32 – Tamanho populacional médio ($\times 10^5$) do grupo referente ao nível do índice de impunidade considerado por mês.



Fonte: Elaborada pela autora.

nível de impunidade (“Tempo de conversão menor do que 10 dias”), enquanto o segundo grupo com menor tamanho populacional médio está associado à maior impunidade (“Tempo de conversão maior do que 90 dias” - incluindo falta de conversão), sendo que o grupo com menor tamanho populacional médio está associado à falta de criminalidade no período analisado.

Isso se dá, pois a maior limitação da estratégia tomada é a possível falta de correspondência entre os procedimentos, o que pode ocasionar a inflação do índice de impunidade para municípios pequenos, cujos inquéritos são abertos, geralmente, em cidade vizinhas maiores; além da subnotificação da impunidade em municípios grandes, principalmente a capital paulista, em que pode-se associar um inquérito a um boletim de ocorrência referente a um caso posterior.

Além disso, considera-se a capital em conjunto com a região metropolitana na região “Grande São Paulo”, e toma-se as seguintes variáveis explicativas: população estimada (por cem mil habitantes), Índice de Desenvolvimento Humano do município (IDH-M) em relação à educação, renda per capita, índice de vulnerabilidade social (IVS) e índice de Gini do município.

A saber, o IVS é complementar ao IDH-M e refere-se a indicadores estruturados em três dimensões - infraestrutura humana, capital humano e renda e trabalho - cuja posse ou privação determina as condições de bem-estar das populações nas sociedades contemporâneas (COSTA; MARGUTI, 2015).

Já o índice de Gini é uma medida de desigualdade de renda que varia entre 0 e 1, sendo que quanto maior o valor, mais desigual é a localidade. Equivale ao dobro da área entre a reta

identidade e a curva de Lorenz, conforme descrito em [Gastwirth \(1972\)](#). Essa curva representa o gráfico da porcentagem da renda total em função da porção da população que recebe tal parcela da renda absoluta, quando a população está ordenada de forma crescente de acordo com sua renda ([GASTWIRTH, 1971](#)), ou seja, baseando-se no rol - a lista ordenada de valores.

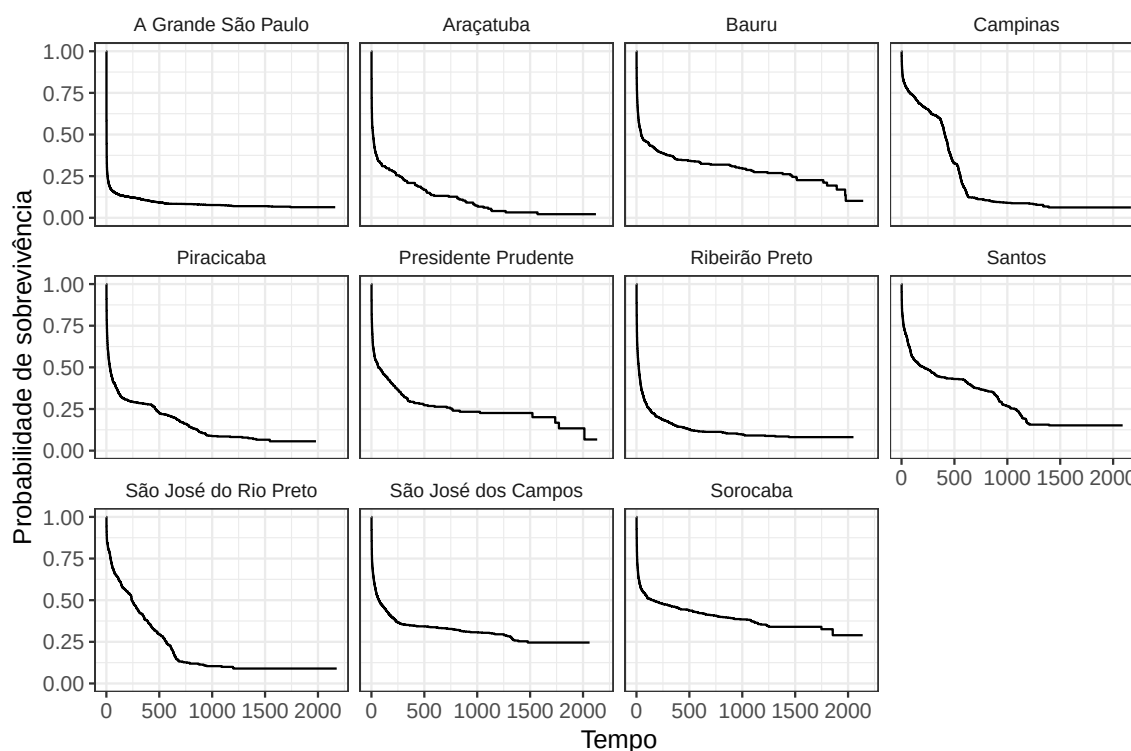
Em relação à extração das informações, a população estimada foi obtida por meio do *site* do Instituto Brasileiro de Geografia e Estatística (IBGE), o IDH-M por meio do portal Atlas Brasil e a renda per capita, índice de vulnerabilidade social e índice de Gini dos municípios da organização Base dos Dados, com informações provenientes do Instituto de Pesquisa Econômica Aplicada (IPEA).

5.2 Resultados

5.2.1 Análise visual das curvas de sobrevivência

A seguir, na [Figura 33](#), apresentam-se as curvas de sobrevivência das regiões ao longo do período estudado, de 2017 a 2022.

Figura 33 – Gráfico das curvas de sobrevivência por região ao longo do tempo (em dias).



Fonte: Elaborada pela autora.

É possível observar que, na região da Grande São Paulo, a curva decai rapidamente, ou seja, a probabilidade de que um boletim de ocorrência demore mais tempo para se tornar um

inquérito cai rapidamente. Isso significa, então, que essa conversão é, em geral, mais rápida do que para as outras regiões do estado.

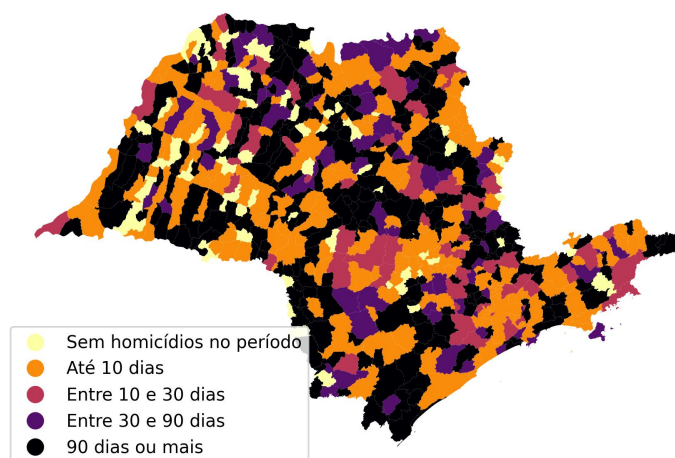
Por outro lado, em Sorocaba, cuja curva de sobrevivência está mais acima das demais, conclui-se que a conversão de um boletim de ocorrência em inquérito é, em geral, mais demorada do que nas demais, sendo que a tendência sugere que cerca de 30% dos boletins não viram inquérito. Isso pode ser percebido, pois a probabilidade de que a conversão demore mais do que um número alto de dias (o período de estudo, de 6 anos) é, aproximadamente, essa.

Outra região que se destaca pelo comportamento da curva é a de Campinas, pois, aproximadamente até 375 dias, a probabilidade de que a conversão demore mais tempo do que isso é muito alta e, em seguida, ela decai rapidamente. Isso indica que, em geral, a tendência é a de que os boletins de ocorrência demorem até 375 dias para serem convertidos em inquéritos, porém, quando passa esse período, eles são, de fato, convertidos.

5.2.2 *Análise visual do índice de impunidade considerado*

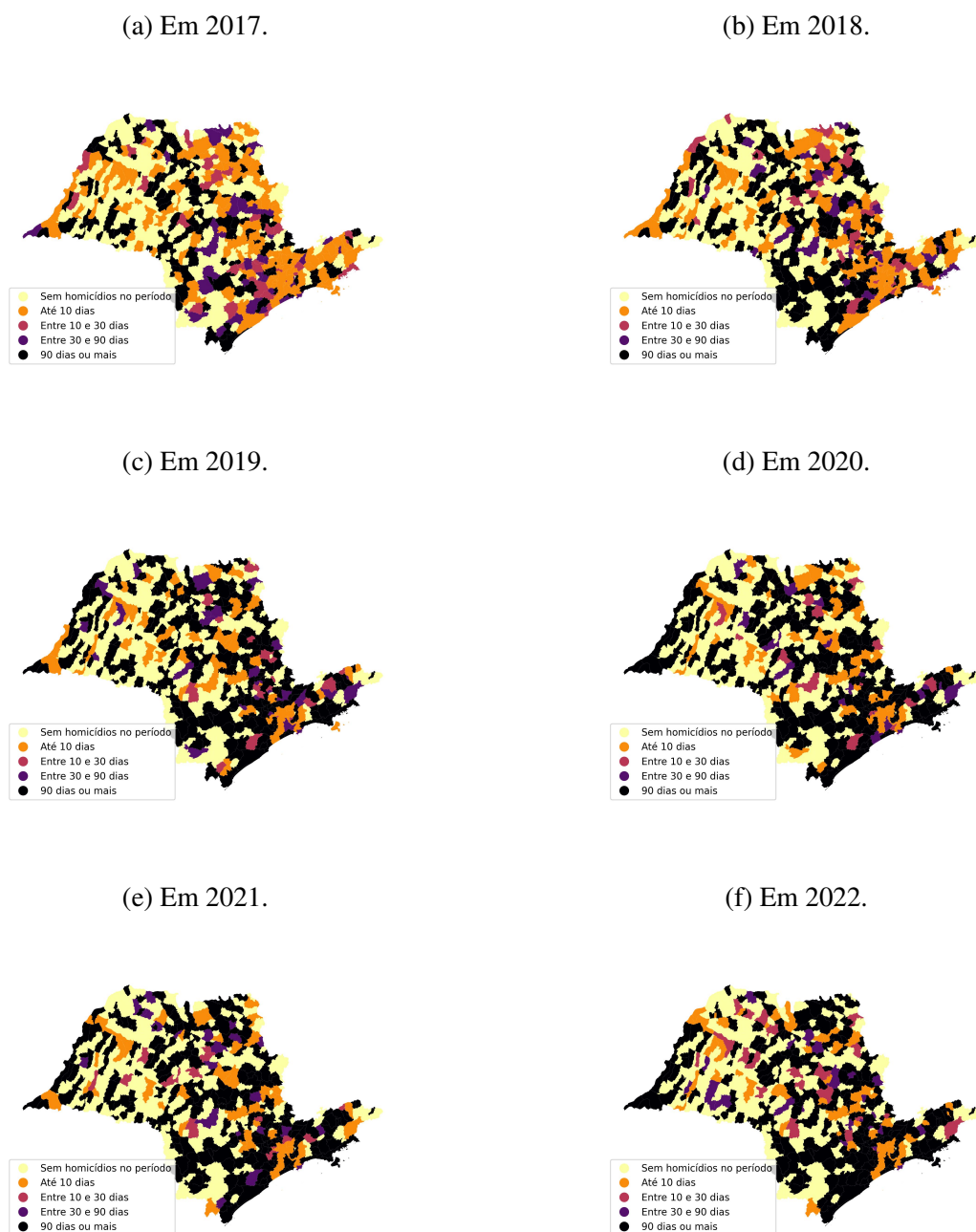
A seguir, nas [Figura 34](#) e [Figura 35](#), apresenta-se o nível de impunidade de cada município, de acordo com o índice construído, ou seja, o tempo mediano de conversão do BO em inquérito considerando o período em questão, respectivamente, em todo o período de estudo e por ano.

Figura 34 – Nível de impunidade de acordo com o tempo mediano de conversão do BO em inquérito em todo o período de estudo (2017 a 2022) para cada município do estado de SP.



Fonte: Elaborada pela autora.

Figura 35 – Nível de impunidade de acordo com o tempo mediano de conversão de BO em inquérito por ano entre 2017 e 2022 por município.



Fonte: Elaborada pela autora.

5.2.3 Medidas avaliativas dos modelos ajustados

A partir do índice de impunidade construído, considera-se um modelo bayesiano de sobrevivência com fragilidade espacial ICAR, conforme descrito no [Capítulo 3](#). A seguir, na [Tabela 7](#), apresenta-se as medidas avaliativas dos modelos ajustados, ou seja, com função de sobrevivência basal dos tipos log-logística, log-normal e Weibull para riscos (PH - *proportional*

hazards) e chances proporcionais (PO - *proportional odds*) para as cadeias 1 e 2. Elas foram obtidas, respectivamente, utilizando como valores iniciais o ajuste do modelo paramétrico sem fragilidade e o ajuste do modelo paramétrico com fragilidade.

Tabela 7 – Medidas avaliativas dos modelos ajustados com a fragilidade espacial ICAR.

Modelo	Baseline	Cadeia	WAIC	DIC	Log-verossimilhança
PH	Log-logística	1	123.517,20	123.516,00	-61.758,62
PH	Log-logística	2	123.443,60	123.440,70	-61.721,78
PH	Log-normal	1	123.108,40	123.106,30	-61.554,18
PH	Log-normal	2	123.105,90	123.102,40	-61.552,97
PH	Weibull	1	125.929,10	125.923,30	-62.964,53
PH	Weibull	2	125.448,80	125.443,40	-62.724,40
PO	Log-logística	1	124.140,80	124.136,80	-62.070,38
PO	Log-logística	2	124.362,80	124.358,00	-62.181,40
PO	Log-normal	1	123.406,30	123.402,90	-61.703,15
PO	Log-normal	2	123.781,80	123.778,00	-61.890,89
PO	Weibull	1	127.642,70	127.638,20	-63.821,35
PO	Weibull	2	127.770,10	127.763,70	-63.885,10

Fonte: Elaborada pela autora.

Assim, é possível perceber que o modelo com chances proporcionais (PH) e função baseline log-normal - destacado em negrito - possui menores valores de WAIC e DIC, bem como maior valor da log-verossimilhança entre aqueles ajustados com a fragilidade espacial ICAR. Na sequência, na [Tabela 8](#), verifica-se o ajuste do modelo de sobrevivência bayesiano sem a fragilidade espacial ICAR para a abordagem escolhida (com função baseline log-normal).

Tabela 8 – Medidas avaliativas dos modelos ajustados sem a fragilidade espacial ICAR.

Modelo	Baseline	Cadeia	WAIC	DIC	Log-verossimilhança
PH	Log-normal	1	123.159,50	123.159,20	-61.579,74
PH	Log-normal	2	123.025,50	123.024,90	-61.512,74
PO	Log-normal	1	123.707,30	123.705,50	-61.853,62
PO	Log-normal	2	123.659,40	123.657,00	-61.829,66

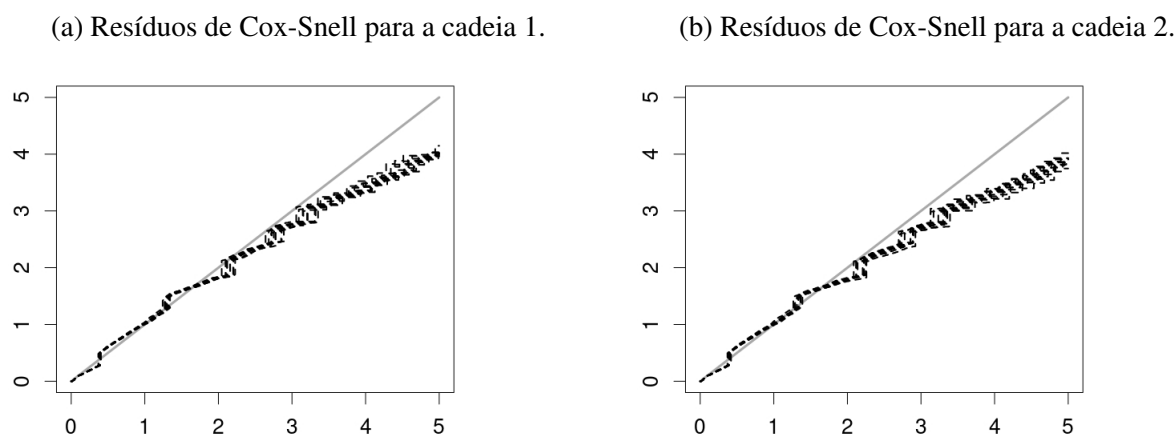
Fonte: Elaborada pela autora.

Os resultados das medidas avaliativas estão próximos entre as abordagens de riscos proporcionais e função baseline log-normal para os modelos com e sem a fragilidade espacial. Desse forma, opta-se por manter a fragilidade como componente da modelagem, visto que a sua interpretação é de interesse no estudo. Além disso, o custo computacional não foi significativamente maior entre os modelos com e sem a fragilidade. Dessa forma, segue-se com o modelo de sobrevivência com fragilidade espacial ICAR, função baseline log-normal e riscos proporcionais.

5.2.4 Resíduos do modelo escolhido

Na [Figura 36](#), apresentam-se os gráficos de resíduos de Cox-Snell do modelo escolhido. Apesar das curvas apresentarem um desvio da reta identidade, reitera-se que esse foi o modelo mais adequado entre os ajustados nesse estudo.

Figura 36 – Resíduos de Cox-Snell para as cadeias 1 e 2.



Fonte: Elaborada pela autora.

5.2.5 Convergência do método MCMC

Em seguida, na [Tabela 9](#), analisa-se, por meio do critério de Gelman e Rubin ([GELMAN; RUBIN, 1992](#)), a convergência das duas cadeias simuladas em cada um dos parâmetros em questão para o modelo escolhido.

Tabela 9 – Critério de Gelman e Rubin para a convergência das duas cadeias simuladas (β_1 : população estimada do município, em 100 mil habitantes; β_2 : IDH-M da educação; β_3 : índice de vulnerabilidade social; β_4 : renda per capita; β_5 : índice de Gini.)

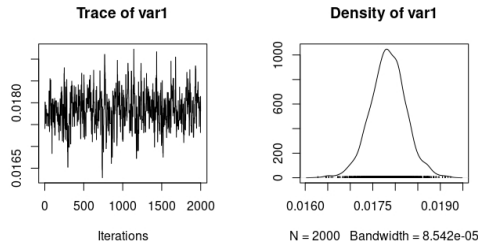
Parâmetro	Estimativa pontual	Limite superior do IC
β_1	1,05	1,21
β_2	1,00	1,00
β_3	1,06	1,25
β_4	1,02	1,08
β_5	1,01	1,06

Fonte: Elaborada pela autora.

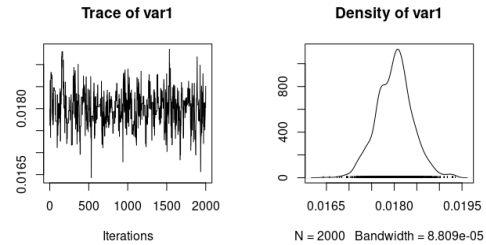
Também observa-se, nas [Figuras 37, 38, 39, 40 e 41](#), a trajetória dos parâmetros estimados para a avaliação da convergência do método MCMC do modelo escolhido.

Figura 37 – Trajetórias e densidades atenuadas dos valores estimados do parâmetro β_1 do modelo escolhido para cada uma das cadeias.

(a) Trajetória e densidade atenuada dos valores estimados de β_1 para a cadeia 1.



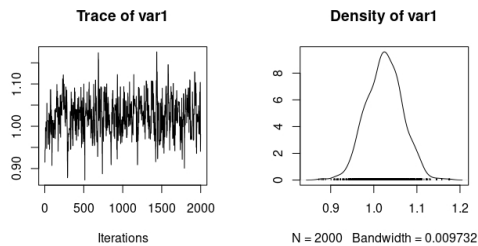
(b) Trajetória e densidade atenuada dos valores estimados de β_1 para a cadeia 2.



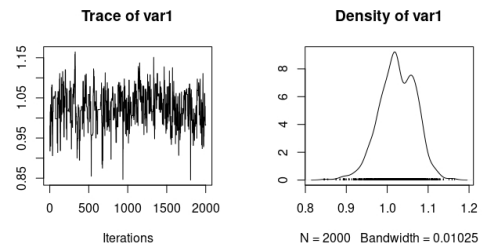
Fonte: Elaborada pela autora.

Figura 38 – Trajetórias e densidades atenuadas dos valores estimados do parâmetro β_2 do modelo escolhido para cada uma das cadeias.

(a) Trajetória e densidade atenuada dos valores estimados de β_2 para a cadeia 1.



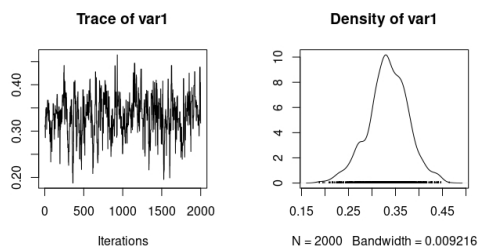
(b) Trajetória e densidade atenuada dos valores estimados de β_2 para a cadeia 2.



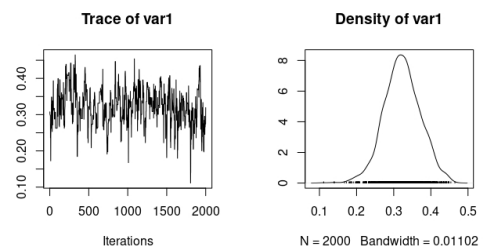
Fonte: Elaborada pela autora.

Figura 39 – Trajetórias e densidades atenuadas dos valores estimados do parâmetro β_3 do modelo escolhido para cada uma das cadeias.

(a) Trajetória e densidade atenuada dos valores estimados de β_3 para a cadeia 1.



(b) Trajetória e densidade atenuada dos valores estimados de β_3 para a cadeia 2.



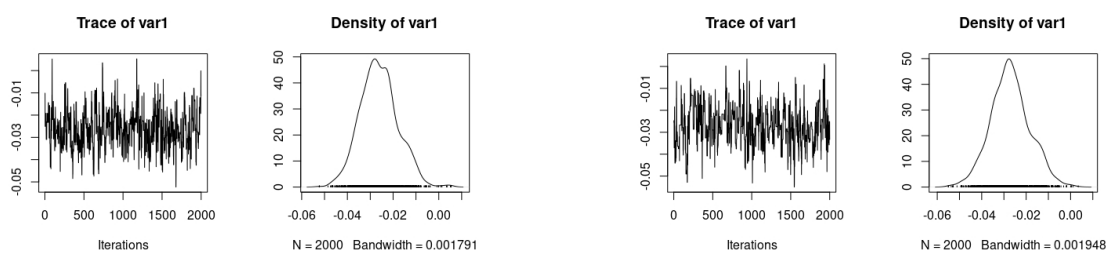
Fonte: Elaborada pela autora.

A taxa de aceitação do método MCMC utilizado foi de 12.26% para a cadeia 1 e 9.33% para a cadeia 2. Apesar de não serem valores altos, destaca-se que o algoritmo é considerável-

mente complexo, conforme apresentado no [Capítulo 3](#), o que justifica a taxa de aceitação mais baixa.

Figura 40 – Trajetórias e densidades atenuadas dos valores estimados do parâmetro β_4 do modelo escolhido para cada uma das cadeias.

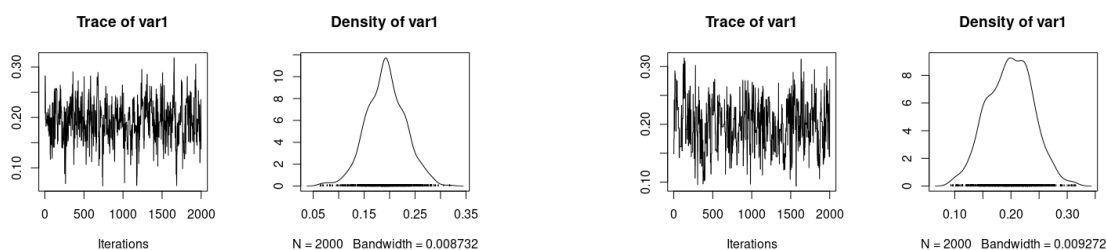
- (a) Trajetória e densidade atenuada dos valores estimados de β_4 para a cadeia 1. (b) Trajetória e densidade atenuada dos valores estimados de β_4 para a cadeia 2.



Fonte: Elaborada pela autora.

Figura 41 – Trajetórias e densidades atenuadas dos valores estimados do parâmetro β_5 do modelo escolhido para cada uma das cadeias.

- (a) Trajetória e densidade atenuada dos valores estimados de β_5 para a cadeia 1. (b) Trajetória e densidade atenuada dos valores estimados de β_5 para a cadeia 2.



Fonte: Elaborada pela autora.

5.2.6 Estimativas do modelo escolhido

Por fim, apresentam-se as estimativas dos parâmetros do modelo escolhido, bem como os limites do intervalo de credibilidade para cada parâmetro e seu respectivo erro-padrão, isto é, o desvio-padrão das estimativas, nas cadeias 1 e 2, respectivamente, nas [Tabela 10](#) e [Tabela 11](#). Nota-se que os valores são positivos para todas as variáveis, exceto a renda per capita. Isso indica que, quanto mais populoso é o município, quanto maior o índice de educação (IDH-M Educação), maior a vulnerabilidade social (IVS) e maior a desigualdade social (índice de Gini), maior é o risco do evento de interesse ocorrer, isto é, menos morosidade há na conversão de boletins de ocorrência em inquérito. Por outro lado, a estimativa do parâmetro associado à renda per capita é negativo, ou seja, ela impacta negativamente no risco de conversão: quanto maior a renda per capita do município, menor é o risco de conversão, o que representa maior morosidade do processo em questão.

Tabela 10 – Estimativa pontual, limites inferior (LI) e superior (LS) do intervalo de credibilidade de 95% e erro-padrão das estimativas dos parâmetros: População (estimada, em 100 mil habitantes), IDH-M da educação, índice de vulnerabilidade social (IVS), renda per capita e índice de Gini para a cadeia 1.

Estimativa	LI	Pontual (mediana)	LS	Erro-padrão
População	0,02	0,02	0,02	0,00
IDH-M Educação	0,95	1,02	1,11	0,04
IVS	0,24	0,33	0,42	0,04
Renda per capita	-0,04	-0,03	-0,01	0,01
Índice de Gini	0,12	0,19	0,27	0,04

Fonte: Elaborada pela autora.

Tabela 11 – Estimativa pontual, limites inferior (LI) e superior (LS) do intervalo de credibilidade de 95% e erro-padrão das estimativas dos parâmetros: População (estimada, em 100 mil habitantes), IDH-M da educação, índice de vulnerabilidade social (IVS), renda per capita e índice de Gini para a cadeia 2.

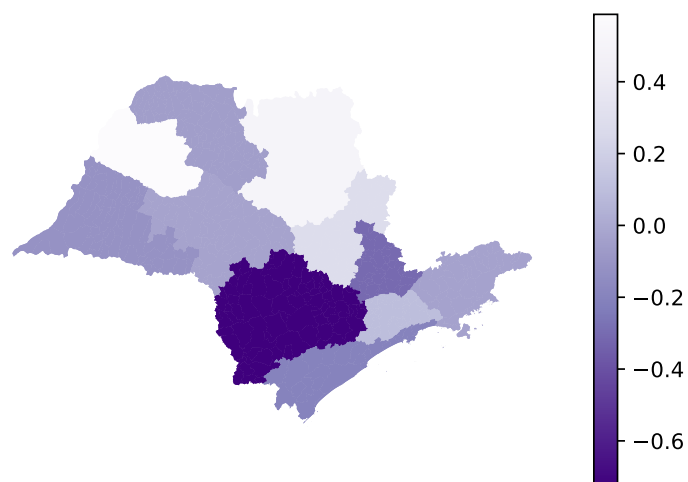
Estimativa	LI	Pontual (mediana)	LS	Erro-padrão
População	0,02	0,02	0,02	0,00
IDH-M Educação	0,93	1,03	1,11	0,04
IVS	0,22	0,32	0,42	0,05
Renda per capita	-0,04	-0,03	-0,01	0,01
Índice de Gini	0,12	0,20	0,28	0,04

Fonte: Elaborada pela autora.

Na [Figura 42](#), apresenta-se também uma análise visual das fragilidades simuladas. É notável que a região de Sorocaba possui o menor valor de fragilidade, ou seja, menor risco de que o evento de interesse ocorra. Isso significa que a morosidade do processo de conversão do boletim de ocorrência em inquérito policial é alto na localidade. Em seguida, apresentando níveis moderados de morosidade, tem-se Santos, Bauru, Presidente Prudente, São José do Rio Preto e Campinas. Por fim, tendo melhores índices (maior fragilidade e, portanto, maior risco de conversão do boletim em inquérito), as localidades de São José dos Campos, Grande São Paulo (capital e região metropolitana), Piracicaba, Ribeirão Preto e Araçatuba.

Ressalta-se que, como analisado no [Seção 5.1](#), não há um identificador que associe com precisão as bases dos boletins de ocorrência e dos inquéritos policiais. Conforme discutido, a aproximação tomada a fim de medir a morosidade do processo de conversão dos registros de ocorrência em processos administrativos no sistema judiciário pode acarretar na subnotificação da impunidade nos municípios grandes, como a capital paulista, e a inflação do índice para municípios pequenos. Esse mesmo processo pode influenciar diretamente as estimativas dos parâmetros observados nas [Tabela 10](#) e [Tabela 11](#), bem com a interpretação do gráfico das fragilidades simuladas, apresentado na [Figura 42](#).

Figura 42 – Mapa das fragilidades simuladas.



Fonte: Elaborada pela autora.

Por isso, é importante destacar que a interpretação aqui discutida está de acordo com os dados obtidos a partir da aproximação realizada, que está sujeita a erros. A fim de analisar e interpretar os dados de forma mais fidedigna, seria necessário a existência de uma relação direta entre as bases, o que tornaria possível a análise estatística com base em dados mais condizentes com a realidade. Reitera-se, por fim, que a análise do índice utilizado é nova na literatura para investigação da morosidade do sistema judiciário brasileiro na conversão de registros de ocorrências em inquéritos policiais.

DISCUSSÃO E TRABALHOS FUTUROS

Ressalta-se a importância do estudo enquanto enriquecimento da literatura em termos da quantificação da impunidade no país por meio da proposta do tempo de conversão de boletins de ocorrência em inquéritos policiais como seu indicativo. Esse tempo refere-se à morosidade do processo e representa o fluxo entre delegacias e o sistema judiciário brasileiro. Nesse sentido, uma limitação do estudo é a falta de correspondência entre as bases, provenientes de setores distintos do governo. Apesar disso, foi proposta uma aproximação entre as bases, cujas limitações foram discutidas no [Capítulo 5](#). Além disso, o estudo é relevante enquanto investiga a eficiência do sistema judiciário do país, utilizando como recorte o cenário no estado paulista.

Destaca-se que o estudo foi apresentado na Conferência *Spatial Statistics 2025: At the Dawn of IA* como apresentação de pôster enquanto aplicação do tema “*Crime, poverty, liveability mapping*”, ou seja, “Crime, pobreza e mapeamento da qualidade de vida”. O contexto, a metodologia e os resultados parciais foram discutidos e seus desdobramentos estão incorporados no presente documento.

Entende-se como trabalhos futuros o enfoque em aprendizado de máquina para a avaliação do poder preditivo do modelo, por meio da divisão dos dados em treino e teste, utilizando como métricas medidas como o *C-index*, índice de Brier e *Predictive Error Curve* (PEC). Também considera-se relevante o estudo de outros modelos de sobrevivência com fragilidade espacial CAR, como o *Accelerated Failure Time* (AFT), cuja relação da função de sobrevivência com as covariáveis depende também do tempo considerado, ou seja, não é constante como nos modelos de riscos e chances proporcionais.

Além disso, entende-se como de interesse a análise mais profunda da região de São José dos Campos, localizada no Vale do Paraíba. Segundo notícias mais recentes, de 2025, o Vale segue sendo cenário de alta criminalidade: apresentou mais de 200 mortes violentas, o que inclui homicídios dolosos e roubos seguidos de morte (latrocínios), de janeiro a setembro do ano ([G1, 2025a](#)); além de uma média de seis carros furtados por dia - um carro a cada 4 horas - no mesmo

ano (G1, 2025b); São José dos Campos também registrou o maior número de estupros da série histórica em 2025 (G1, 2025c). Sendo assim, verifica-se a necessidade de maior investigação acerca da região.

Destaca-se como possível próximo passo a inclusão de uma variável indicadora referente aos anos de ocorrência da pandemia de *COVID-19* no contexto da análise de sobrevivência. Esse estudo pode ser de interesse, pois considera um fator marcante tanto no cenário mundial quanto brasileiro, que pode ter impactado a morosidade dos processos retratados nesse estudo.

REFERÊNCIAS

ADORNO, S.; PASINATO, W. A justiça no tempo, o tempo da justiça. **Tempo social**, SciELO Brasil, v. 19, p. 131–155, 2007. Citado na página 18.

_____. Violência e impunidade penal: da criminalidade detectada à criminalidade investigada. **Dilemas-Revista de Estudos de Conflito e Controle Social**, Universidade Federal do Rio de Janeiro, v. 3, n. 7, p. 51–84, 2010. Citado na página 17.

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC, 2014. Citado nas páginas 33 e 35.

BENDER, R.; AUGUSTIN, T.; BLETTER, M. Generating survival times to simulate cox proportional hazards models. **Statistics in Medicine**, v. 24, n. 11, p. 1713–1723, 2005. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2059>>. Citado na página 48.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 192–225, 1974. Citado nas páginas 19, 33, 35 e 36.

BRICEÑO-LEÓN, R.; VILLAVECES, A.; CONCHA-EASTMAN, A. Understanding the uneven distribution of the incidence of homicide in latin america. **International journal of epidemiology**, Oxford University Press, v. 37, n. 4, p. 751–757, 2008. Citado na página 17.

BROOK, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. **Biometrika**, JSTOR, v. 51, n. 3/4, p. 481–483, 1964. Citado na página 34.

CHEN, Y.; HANSON, T.; ZHANG, J. Accelerated hazards model based on parametric families generalized with bernstein polynomials. **Biometrics**, Oxford University Press, v. 70, n. 1, p. 192–201, 2014. Citado na página 42.

CLERCQ, J. A. L.; LARA, G. R. S. Global impunity index 2020 (gii 2020). impunity levels in the world. **Impunity Levels in the World (April 30, 2021)**, 2021. Citado na página 18.

CLIFFORD, P. Markov random fields in statistics. **Disorder in physical systems: A volume in honour of John M. Hammersley**, p. 19–32, 1990. Citado na página 35.

COSTA, M. A.; MARGUTI, B. O. E. Atlas da vulnerabilidade social nos municípios brasileiros. Instituto de Pesquisa Econômica Aplicada (Ipea), 2015. Citado na página 60.

COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968. Citado na página 46.

CRIBARI-NETO, F.; SANTOS, M. A. An empirical analysis of worldwide impunity. **Humanities and Social Sciences Communications**, Palgrave, v. 11, n. 1, p. 1–13, 2024. Citado na página 19.

- DUCROCQ, V.; CASELLA, G. A bayesian analysis of mixed survival models. **Genetics Selection Evolution**, EDP Sciences, v. 28, n. 6, p. 505–529, 1996. Citado na página 40.
- EMMERT-STREIB, F.; DEHMER, M. Introduction to survival analysis in practice. **Machine Learning and Knowledge Extraction**, MDPI, v. 1, n. 3, p. 1013–1038, 2019. Citado na página 36.
- FAJNZYLBER, P.; LEDERMAN, D.; LOAYZA, N. Inequality and violent crime. **The journal of Law and Economics**, The University of Chicago Press, v. 45, n. 1, p. 1–39, 2002. Citado na página 17.
- G1. **Com 27 mortes em setembro, Vale do Paraíba ultrapassa 200 assassinatos em 2025**. 2025. Disponível em: <https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2025/10/31/com-27-mortes-em-setembro-vale-do-paraiba-ultrapassa-200-assassinatos-em-2025.ghtml>. Acesso em 10 fev. 2026. Citado na página 71.
- _____. **Com média de um furto a cada 4h, seis veículos são furtados por dia no Vale do Paraíba, aponta SSP**. 2025. Disponível em: <https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2025/11/02/com-media-de-um-furto-a-cada-4h-seis-veiculos-sao-furtados-por-dia-no-vale-do-paraiba-aponta-ssp.ghtml>. Acesso em 10 fev. 2026. Citado na página 72.
- _____. **Em sete meses, São José dos Campos registra o maior número de estupros da série histórica da SSP**. 2025. Disponível em: <https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2025/08/29/em-sete-meses-sao-jose-dos-campos-registra-o-maior-numero-de-estupros-da-serie-historica-da-ssp.ghtml>. Acesso em 10 fev. 2026. Citado na página 72.
- GASTWIRTH, J. L. A general definition of the lorenz curve. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 1037–1039, 1971. Citado na página 61.
- _____. The estimation of the lorenz curve and gini index. **The review of economics and statistics**, JSTOR, p. 306–316, 1972. Citado nas páginas 17 e 61.
- GEISSER, S.; EDDY, W. F. A predictive approach to model selection. **Journal of the American Statistical Association**, Taylor & Francis, v. 74, n. 365, p. 153–160, 1979. Citado na página 46.
- GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical science**, Institute of Mathematical Statistics, v. 7, n. 4, p. 457–472, 1992. Citado nas páginas 46 e 65.
- GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 6, p. 721–741, 1984. Citado na página 35.
- GENTLE, J. E. Theory of statistics. **George Mason University**, 2013. Citado nas páginas 47 e 48.
- HAARIO, H.; SAKSMAN, E.; TAMMINEN, J. An adaptive metropolis algorithm. 2001. Citado na página 44.
- HARRENDORF, S.; HEISKANEN, M.; MALBY, S. **International statistics on crime and justice**. [S.l.]: European Institute for Crime Prevention and Control, affiliated with the . . . , 2010. Citado na página 18.

HOELSCHER, K. Politics and social violence in developing democracies: Theory and evidence from brazil. **Political Geography**, Elsevier, v. 44, p. 29–39, 2015. Citado na página 17.

LEE, E. T.; WANG, J. **Statistical methods for survival data analysis**. [S.l.]: John Wiley & Sons, 2003. v. 476. Citado na página 39.

LOPES, J. R. d. L. Direitos humanos e tratamento igualitário: questões de impunidade, dignidade e liberdade. **Revista brasileira de Ciências sociais**, SciELO Brasil, v. 15, p. 77–100, 2000. Citado na página 18.

MITRA, R.; MÜLLER, P. *et al.* **Nonparametric Bayesian inference in biostatistics**. [S.l.]: Springer, 2015. Citado nas páginas 19, 37, 40 e 41.

MOORE, D. F. *et al.* **Applied survival analysis using R**. [S.l.]: Springer, 2016. v. 473. Citado na página 36.

MORAN, P. A. Notes on continuous stochastic phenomena. **Biometrika**, JSTOR, v. 37, n. 1/2, p. 17–23, 1950. Citado nas páginas 19 e 25.

MORRIS, M.; WHEELER-MARTIN, K.; SIMPSON, D.; MOONEY, S. J.; GELMAN, A.; DIMAGGIO, C. Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. **Spatial and Spatio-temporal Epidemiology**, v. 31, p. 100301, 2019. ISSN 1877-5845. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1877584518301175>>. Citado nas páginas 34 e 36.

NERY, F. S.; NADANOVSKY, P. Homicide impunity in brazil between 2006 and 2016. **Revista de saude publica**, SciELO Brasil, v. 54, p. 144, 2020. Citado nas páginas 18 e 19.

ROZANOV, Y. A. Markov random fields. In: **Markov random fields**. [S.l.]: Springer, 1982. p. 55–102. Citado na página 35.

SANTOS, M. R.; TESTA, A. Global trends in homicide. In: **Homicide and violent crime**. [S.l.]: Emerald Publishing Limited, 2018. v. 23, p. 199–222. Citado na página 17.

SILVA, A. C.; DEMARQUI, F. N.; SILVA, B. F.; PRATES, M. O. The analysis of criminal recidivism: a hierarchical model-based approach for the analysis of zero-inflated, spatially correlated recurring event data. **Journal of the Royal Statistical Society Series A: Statistics in Society**, Oxford University Press UK, p. qnaf061, 2025. Citado na página 19.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the royal statistical society: Series b (statistical methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 46.

SPITZER, F. Markov random fields and gibbs ensembles. **The American Mathematical Monthly**, Taylor & Francis, v. 78, n. 2, p. 142–154, 1971. Citado na página 35.

WATANABE, S.; OPPER, M. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of machine learning research**, v. 11, n. 12, 2010. Citado na página 46.

WHITTLE, P. On stationary processes in the plane. **Biometrika**, JSTOR, p. 434–449, 1954. Citado na página 33.

YAU, K. K. Multilevel models for survival analysis with random effects. **Biometrics**, Oxford University Press, v. 57, n. 1, p. 96–102, 2001. Citado na página 40.

ZHOU, H.; HANSON, T. A unified framework for fitting bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 522, p. 571–581, 2018. Citado nas páginas 42, 43, 44, 45 e 46.

ZHOU, H.; HANSON, T.; ZHANG, J. spbayessurv: Fitting bayesian spatial survival models using r. **Journal of Statistical Software**, v. 92, p. 1–33, 2020. Citado nas páginas 19, 41 e 42.

