Modelo Fatorial Espaço-Temporal para Dados de Contagem Multivariados: uma aplicação aos dados de criminalidade no Estado do Rio de Janeiro

Beatriz Rodrigues Pinna

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Métodos Estatísticos

2023

Modelo Fatorial Espaço-Temporal para Dados de Contagem Multivariados: uma aplicação aos dados de criminalidade no Estado do Rio de Janeiro

Beatriz Rodrigues Pinna

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Aprovada por:

Prof. João Batista de Morais Pereira

PhD - IM - UFRJ - Orientador.

Kelly Cristina Mota Gonçalves

PhD - IM - UFRJ.

Jony Arrais Pinto Junior

PhD - IM - UFRJ.

Rio de Janeiro, RJ - Brasil

2022

FICHA CATALOGRÁFICA

*** ***

"Life is not easy for any of us. But what of that?
We must have perseverance and above all confidence in ourselves.
We must believe that we are gifted for something, and that this thing,
at whatever cost, must be attained". (Marie Curie)

Agradecimentos

Agradeço a minha família, principalmente meus pais e minhas irmãs, pelo apoio e suporte em toda minha trajetória acadêmica e profissional.

Agradeço também aos meus amigos pelo amparo nos momentos difíceis e por entender algumas ausências devido aos estudos. Agradeço, em especial, ao meu orientador João Batista de Morais Pereira por toda ajuda e parceria no desenvolvimento desta pesquisa. Agradeço aos professores do programa de pós-graduação em Estatística da UFRJ por terem contribuído na minha formação.

Por fim, agradeço aos professores Kelly Cristina Mota Gonçalves e Jony Arrais Pinto Junior por aceitarem fazer parte da minha banca.

Resumo

Neste trabalho foi proposto um modelo denominado Modelo Fatorial Espaço-Temporal para Dados de Contagem Multivariados, com o objetivo de lidar com dados que possuem estruturas multivariada, espacial e temporal. Em particular, foram considerados dados de contagens agregados por área que seguem distribuição Poisson. Nesta abordagem, são considerados diversos tipos criminais pelas Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro no período de 2012 a 2020. O procedimento de inferência foi realizado sob enfoque bayesiano utilizando o método Monte Carlo Hamiltoniano para obtenção das amostras das distribuições *a posteriori* do modelo. Na modelagem proposta, assume-se que os fatores comuns incorporam a variação espacial e temporal existente entre as observações. Estudos com dados simulados e dados reais são discutidos a fim de avaliar a qualidade de ajuste do modelo, a suposição do número de fatores e compreender o fenômeno da criminalidade no estado.

Palavras-Chaves: modelos fatoriais, modelos espaço-temporais; dados de contagem; inferência bayesiana; CAR Intrínseco; Monte Carlo Hamiltoniano.

Abstract

In this work, a model called Space-Temporal Factorial Model for Multivariate Counting Data was proposed to deal with data that have multivariate, spatial, and temporal structures. In particular, area-aggregated counts data that follow the Poisson distribution were considered. In this approach, several criminal types are considered by the Integrated Areas of Public Security (AISP) of the state of Rio de Janeiro in the period from 2012 to 2020. The inference procedure was carried out under a Bayesian approach using the Monte Carlo Hamiltonian method to obtain samples from the posterior distributions of the model. For the proposed model, it is assumed that the common factors incorporate the spatial and temporal variation between observations. Studies with simulated data and real data are discussed to evaluate the model's goodness of fit, the assumption of the number of factors, and to understand the crime phenomenon in the state.

Keywords: factor models; spatio-temporal models; count data; Bayesian inference; Intrinsic CAR; Hamiltonian Monte Carlo.

Sumário

1 Introdução **1**

1.1 Organização da Dissertação 4

2 Revisão Bibliográfica **6**

3 Revisão Metodológica **10**

3.1 Inferência Bayesiana 10

3.2 Monte Carlo Hamiltoniano 12

3.3 Modelo Fatorial 13

3.4 Modelos Dinâmicos 15

3.4.1 Modelos Lineares Dinâmicos 16

3.4.2 Modelos Lineares Dinâmicos Generalizados 17

3.5 Modelos para dados de Área 18

3.5.1 Modelo CAR 19

3.6 Critérios de comparação de modelos 20

3.6.1 Deviance Information Criterion (DIC) 20

3.6.2 Watanabe-Akaike Information Criterion (WAIC) 21

4 Modelos Fatoriais Espaço-Temporais para Dados de Contagem Multivariados **22**

4.1 Modelo Fatorial Espacial 22

4.2 Modelo Fatorial Espaço-Temporal Multivariado 24

4.2.1 Procedimento de Inferência 25

4.3 Estudo Simulado 27
4.3.1 Modelo com 1 fator 28
4.3.2 Modelo com 2 fatores 29
4.3.3 Modelo com 3 fatores 33
4.3.4 Comparando diferentes números de fatores 36

5 Aplicação **45**

6 Conclusões **57**
6.1 Trabalhos Futuros 58

A Traços das Cadeias dos Parâmetros do Estudo Simulado **59**
A.1 Modelo com 1 fator 59
A.2 Modelo com 2 fatores 60
A.3 Modelo com 3 fatores 61

B Componentes dos Critérios de Comparação e Traços das Cadeias dos Parâmetros do Estudo Comparativo **62**
B.1 Modelo com 1 fator 64
B.2 Modelo com 2 fatores 65
B.3 Modelo com 3 fatores 67

C Aplicação **70**
C.1 Modelo com 3 fatores 70

D Algoritmo *Stan* **73**

Lista de Tabelas

4.1 Sumário da distribuição *a posteriori* para os parâmetros fixos do MFETM com 1 fator. 29

4.2 Sumário da distribuição *a posteriori* para os parâmetros fixos do MFETM com 2 fatores. 33

4.3 Sumário da distribuição *a posteriori* para os parâmetros fixos do MFETM com 3 fatores. 36

4.4 Comparação do MFETM com 1, 2 e 3 fatores para os dados gerados assumindo 2 fatores. Os menores valores do critério DIC, em itálico, indicam o melhor ajuste do modelo. 37

4.5 Comparação do MFETM com 1, 2 e 3 fatores para os dados gerados assumindo 2 fatores. Os menores valores do critério WAIC, em itálico, indicam o melhor ajuste do modelo. 37

5.1 Critérios de comparação DIC e WAIC do MFETM ajustado com 2 e 3 fatores da análise dos dados reais. 49

5.2 Sumário da distribuição *a posteriori* para os parâmetros do MFETM com 3 fatores da análise dos dados reais. 53

5.3 Média *a posteriori* da matriz de cargas fatoriais da análise dos dados reais. 54

B.1 Valores de $\overline{D}$ do critérios de comparação DIC para cada amostra de cada modelo. 62

B.2 Valores de p_D do critérios de comparação DIC para cada amostra de cada modelo. 63

B.3 Valores de *lppd* do critérios de comparação WAIC para cada amostra de cada modelo. 63
B.4 Valores de p_{WAIC} do critérios de comparação WAIC para cada amostra de cada modelo. 63

Lista de Figuras

1.1 Mapa das Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com os quintis da distribuição do número de vítimas de Lesão Corporal Dolosa em janeiro de 2021. 3

1.2 Série mensal do número de vítimas de Lesão Corporal Dolosa por algumas AISP. 4

4.1 Média *a posteriori* do nível do MFETM com 1 fator (linha cheia preta) com respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo. 28

4.2 Média *a posteriori* as cargas fatoriais do MFETM com 1 com seus respectivos intervalos de credibilidade de 95%. 29

4.3 Média *a posteriori* de cada nível do MFETM com 2 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo. 31

4.4 Média *a posteriori* das cargas fatoriais do MFETM com 2 fatores com seus respectivos intervalos de credibilidade de 95%. 32

4.5 Média *a posteriori* de cada nível do MFETM com 3 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo. 34

4.6 Média *a posteriori* das cargas fatoriais do MFETM com 3 fatores com seus respectivos intervalos de credibilidade de 95%. 35
4.7 Média *a posteriori* do nível do MFETM com 1 fator (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo para cada amostra. 38
4.8 Média *a posteriori* de cada nível do MFETM com 2 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo para cada amostra. 39
4.9 Média *a posteriori* de cada nível do MFETM com 3 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo para cada amostra. 40
4.10 Média *a posteriori* das cargas fatoriais do MFETM com 1 fator com seus respectivos intervalos de credibilidade de 95% para cada amostra. 41
4.11 Média *a posteriori* das cargas fatoriais do MFETM com 2 fatores com seus respectivos intervalos de credibilidade de 95% para cada amostra. 41
4.12 Média *a posteriori* das cargas fatoriais do MFETM com 3 fatores com seus respectivos intervalos de credibilidade de 95% para cada amostra. 42
4.13 Distribuição *a posteriori* marginal de $\lambda_{1,3}$, $\lambda_{2,10}$, $\phi_{1,35}$ e $\phi_{2,96}$ para algumas amostras do modelo com 2 fatores. 43
4.14 Distribuição *a posteriori* marginal de $\lambda_{1,3}$, $\lambda_{2,10}$, $\lambda_{3,9}$, $\phi_{1,35}$, $\phi_{2,96}$ e $\phi_{3,66}$ para algumas amostras do modelo com 3 fatores. 44

5.1 Divisão Territorial da base de segurança pública por Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro. 46
5.2 Focalização da capital do estado do Rio de Janeiro no mapa da Divisão Territorial da base de segurança pública por AISP. 46
5.3 Séries mensais de títulos criminais do estado Rio de Janeiro. 47

5.4 Mapa das Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com os quintis da distribuição do número de registros de alguns crimes em determinados períodos de tempo. 48
5.5 Distribuição *a posteriori* marginal de $\lambda_{1,2}$, $\lambda_{2,6}$, $\lambda_{3,8}$, $\phi_{1,28}$, $\phi_{2,105}$ e $\phi_{3,58}$ da análise dos dados reais usando o modelo com 3 fatores. 50
5.6 Média *a posteriori* de cada nível do MFETM com 3 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) ao longo do tempo da análise dos dados reais. 51
5.7 Média *a posteriori* das cargas fatoriais do MFETM com 3 fatores com seus respectivos intervalos de credibilidade de 95% da análise dos dados reais. 52
5.8 Mapa das Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com os quintis da distribuição dos valores de cada um dos 3 fatores em dezembro de 2020. 55
5.9 Séries mensais de cada um dos 3 fatores nas 3ª, 7ª, 25ª e 33ª AISP. . . . 56

A.1 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{\tau_1}$ e ρ_1 do modelo com 1 fator, a linha vermelha representa o valor verdadeiro. 59
A.2 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, ρ_1 e ρ_2 do modelo com 2 fatores, a linha vermelha representa o valor verdadeiro. 60
A.3 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{u_3}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, $\sqrt{\tau_3}$, ρ_1, ρ_2 e ρ_3 do modelo com 3 fatores, a linha vermelha representa o valor verdadeiro. 61

B.1 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{\tau_1}$ e ρ_1 do modelo com 1 fator para cada amostra, a linha vermelha representa o valor verdadeiro. 64
B.2 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, ρ_1 e ρ_2 do modelo com 2 fatores para cada amostra, a linha vermelha representa o valor verdadeiro. 65

B.3 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{u_3}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, $\sqrt{\tau_3}$, ρ_1, ρ_2 e ρ_3 do modelo com 3 fatores para cada amostra, a linha vermelha representa o valor verdadeiro. 67

C.1 Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{u_3}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, $\sqrt{\tau_3}$, ρ_1, ρ_2 e ρ_3 do modelo com 3 fatores da análise dos dados reais. . . . 70

C.2 Séries mensais de cada um dos 3 fatores por AISP do modelo com 3 fatores da análise dos dados reais. 72

Capítulo 1

Introdução

A obtenção de dados complexos e de alta dimensão tem se tornado cada vez mais frequente e exige o uso de técnicas avançadas que permitam capturar e analisar a sua informação. Os modelos fatoriais vem sendo explorados devido sua capacidade de modelar a estrutura de dependência entre as variáveis. Estes modelos visam reduzir a dimensionalidade de um grande conjunto de variáveis observáveis em um número menor de fatores não observáveis, que são utilizados para identificar as associações entre as variáveis.

Os modelos fatoriais sob enfoque bayesiano ganharam destaque por causa da facilidade das ferramentas computacionais atuais, como o método Monte Carlo via cadeias de Markov (MCMC) a partir dos anos 2000. Aguilar e West (2000) utilizaram o modelo fatorial dinâmico na aplicação de séries temporais financeiras multivariadas. Lopes e West (2004) consideraram a incerteza do número de fatores na estimação através do algoritmo MCMC com saltos reversíveis, RJMCMC, que trata o número de fatores como parâmetro. Sáfadi e Peña (2008) ajustaram um modelo fatorial dinâmico na aplicação de múltiplas séries temporais, cujos fatores seguem um modelo autorregressivo e os erros são independentes com distribuição normal.

Em muitos casos, os conjuntos de dados podem ser multivariados com estrutura temporal, espacial ou espaço-temporal, sendo necessário modelos que incorporem estas diferentes fontes de variação. Lopes et al. (2008) propuseram uma nova classe de modelos espaços-temporais, derivados de modelos fatoriais dinâmicos, para dados gaussianos sob enfoque bayesiano em que a matriz de cargas dos fatores é estática, ou seja, invariante no

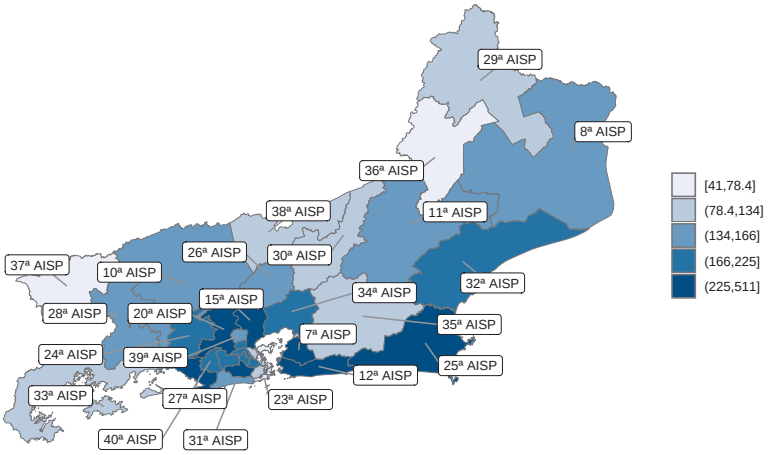
tempo e o número de fatores é tratado como um parâmetro desconhecido. A inferência do número de fatores é realizada via o algoritmo RJMCMC. Nesta nova classe de modelos, a dependência temporal é modelada pelos fatores comuns, enquanto a dependência espacial é modelada pelas cargas fatoriais.

Os estudos dos modelos fatoriais podem ser aplicados em diversas áreas das ciências, como, por exemplo, na área de segurança pública que engloba diferentes tipos de crimes cometidos em um determinado período e local. No caso, para os agentes públicos é de interesse identificar e entender algum padrão espacial relacionado a análise quantitativa dos crimes, assim como a dinâmica temporal deste fenômeno.

A violência urbana atinge diversas cidades brasileiras, em especial as grandes regiões metropolitanas que possuem áreas com muita variabilidade demográfica em seu espaço geográfico. O estado do Rio de Janeiro se destaca por seu cenário crítico em relação a criminalidade, no qual questões sociais e econômicas podem estar associadas a alta incidência da violência no estado (Clemente et al., 2022). Com a disponibilidade de dados criminais é possível realizar estudos quantitativos que possam identificar padrões e regularidades das ocorrências criminais registradas, já que podem variar bastante de acordo com o seu tipo, período e localidade. Conforme Miranda et al. (2006), reconhecer os determinantes que influenciam os crimes em certas áreas e horários é crucial para o planejamento de ações e estratégias de controle da violência.

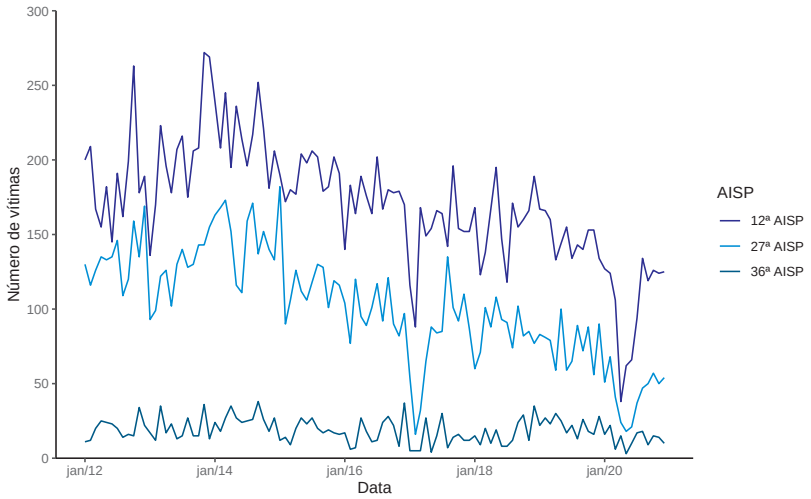
O Instituto de Segurança Pública (ISP, 2022) é o órgão responsável por centralizar, consolidar e disponibilizar as estatísticas oficiais de segurança pública do estado do Rio de Janeiro. O ISP divulga estes dados mensalmente, sendo um total de 52 títulos de ocorrências criminais e não criminais, abrangendo todo o território do estado. Por exemplo, a Figura 1.1 apresenta a distribuição espacial dos quintis do número de vítimas de Lesão Corporal Dolosa em janeiro de 2021 pelas Áreas Integradas de Segurança Pública (AISP), pode-se observar que as AISP referente a região metropolitana e a baixada fluminense do Rio de Janeiro possuem uma maior concentração de vítimas deste tipo de crime.

Figura 1.1: Mapa das Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com os quintis da distribuição do número de vítimas de Lesão Corporal Dolosa em janeiro de 2021.



Além do estudo espacial que descreve e visualiza distribuições espaciais a fim de identificar associação espacial, os dados de criminalidade variam no tempo. A análise temporal tem por objetivo verificar a existência de tendências e ciclos do fenômeno no tempo. Como exemplo, a Figura 1.2 apresenta a série temporal mensal do número de vítimas de Lesão Corporal Dolosa das 12ª, 27ª e 36ª AISP entre 2012 e 2020. Observa-se que a 12ª AISP possui os maiores valores no período analisado, essa região engloba principalmente o município de Niterói. Já a 27ª AISP compreende bairros da zona oste do município do Rio de Janeiro, e a 36ª AISP que apresenta os menores valores abarca principalmente municípios do Noroeste Fluminense do estado. Assim, é notória a diferença na evolução temporal do mesmo tipo de crime para diferentes regiões.

Figura 1.2: Série mensal do número de vítimas de Lesão Corporal Dolosa por algumas AISP.



Portanto, este trabalho propõe um modelo fatorial espaço-temporal que considera a a estrutura multivariada, espacial e temporal do conjunto de dados. Neste modelo, são utilizados dados de contagens que seguem distribuição Poisson sob abordagem completamente bayesiana, no qual a variação espacial e temporal existente entre as observações é incorporada através dos fatores comuns. O método Monte Carlo Hamiltoniano é utilizado para obter amostras das distribuições *a posteriori*. Um estudo com dados simulados é feito com a finalidade de avaliar o ajuste do modelo proposto. Também é realizada uma aplicação do modelo em dados reais de criminalidade, cujo o objetivo é reconhecer padrões dos crimes nas divisões territoriais de segurança pública do estado do Rio de Janeiro, e utilizar os fatores comuns como indicadores de criminalidade de múltiplos crimes.

1.1 Organização da Dissertação

Esta dissertação de Mestrado está organizado da seguinte forma:

No capítulo 2 é feita uma breve revisão da literatura acerca dos modelos fatorias dinâmicos espaciais sob enfoque bayesiano. O objetivo é ressaltar as diferentes especificações das estruturas temporal e espacial adotadas no modelo fatorial e a discussão dos problemas de identificabilidade desta classe de modelos.

O capítulo 3 introduz brevemente o conceito de Inferência Bayesiana. Em seguida,

é realizado um resumo sobre o procedimento de inferência Monte Carlo Hamiltoniano (HMC) na sua variante *Not U Turn* (NUTS), que é utilizado no *software* Stan. E é realizada uma revisão sobre Modelo Fatorial sob abordagem bayesiana, Modelos Dinâmicos e Modelos para dados de Área. Também é descrito os critérios DIC e WAIC de comparação de modelos utilizados neste trabalho.

O capítulo 4 apresenta o modelo proposto neste trabalho, denominado de Modelo Fatorial Espaço-Temporal para Dados de Contagem Multivariados. Além disso, apresenta o estudo simulado e os principais resultados obtidos em que vários cenários são avaliados e comparados.

No capítulo 5, é apresentado a aplicação do modelo proposto em um conjunto de dados reais para verificar o ajuste e avaliar o seu desempenho. O conjunto de dados utilizado se refere a diferentes títulos criminais do estado do Rio de Janeiro.

No capítulo 6 são apresentadas as principais conclusões e possíveis propostas de trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

Nesta seção, apresentaremos alguns trabalhos acerca dos modelos fatorias sob abordagem bayesiana ressaltando sua estrutura e as restrições utilizadas pelos autores a fim de evitar problemas de identificabilidade. De acordo com a literatura, muitos autores impõem aos modelos fatoriais bayesianos restrições para que seja definido um modelo único, sem problemas de identificação. A restrição mais geral é limitar os elementos da diagonal principal da matriz de cargas a assumirem somente valores positivos conforme Geweke e Zhou (1996)[1].

Aguilar e West (2000) apresentam o modelo fatorial dinâmico para séries temporais financeiras em que a parte dinâmica é inserida na variância dos fatores. Para evitar problemas de identificabilidade, a matriz de cargas fatoriais possui a restrição de ser triangular inferior de posto completo, que consiste em limitar os elementos da diagonal principal a assumirem valores iguais a 1 e em fixar em zero os elementos do triângulo superior. Os autores também utilizam a restrição para que o número de parâmetros livres não exceda $p(p+1)/2$ parâmetros para evitar sobreparametrização. E por último, asseguram a invariância do modelo sob transformações lineares inversíveis dos fatores comuns.

Em Wang e Wall (2003) é proposto um modelo fatorial espacial generalizado com um único fator comum. Ao contrário do modelo fatorial usual no qual os dados são

[1]Tal restrição advém de um teorema da teoria de matrizes que pode ser encontrado em Muirhead (1982), p. 592, Teorema A9.8.

normalmente distribuídos e os fatores são independentes entre as observações, os autores estendem o modelo para dados da família exponencial, particularmente para as distribuições Poisson e Binomial, além de assumirem que os fatores são espacialmente correlacionados para explicar as duas correlações: intra e inter localizações. A estrutura espacial é incorporada na matriz de covariância do fator comum, que possui distribuição normal multivariada. Além disso, para contornar os problemas de identificabilidade do modelo, os autores adotam que o fator comum espacial segue distribuição normal multivariada com média zero e uma estrutura de covariância espacial com um parâmetro de variância unitária, e adicionam a restrição de soma zero dos elementos do vetor do fator comum. De forma geral, o objetivo é encontrar um único fator comum espacial que as variáveis observadas compartilham. O estudo do modelo com mais de um fator contém problemas de identificabilidade que não foram trabalhados no artigo.

Já Sáfadi e Peña (2008) ajustaram um modelo fatorial dinâmico sob enfoque bayesiano na aplicação de múltiplas séries temporais, cujos fatores seguem um modelo autorregressivo e os erros são independentes com distribuição normal. O objetivo dos autores foi verificar a associação entre poluição do ar e mortalidade de doenças respiratórias e cardíacas na cidade de São Paulo. Além disso, para realizar a inferência acerca dos parâmetros foi utilizado o algoritmo amostrador de Gibbs. Neste trabalho, os autores também adotaram a restrição da matriz de cargas fatoriais possuir a restrição triangular inferior de posto completo com elementos da diagonal principal iguais a 1 para evitar problemas de identificação.

Lopes et al. (2008) propõem uma nova classe de modelos espaço-temporais a partir dos modelos fatoriais dinâmicos para dados gaussianos, onde a dependência temporal é modelada pelos fatores comuns através de um processo autorregressivo de ordem 1. A dependência espacial, por sua vez, é considerada ao permitir-se que as cargas fatoriais sigam um processo gaussiano. O objetivo deste modelo é identificar grupos de localizações no espaço considerando a estrutura temporal das observações. Além disso, os autores utilizam o algoritmo MCMC com saltos reversíveis (do inglês, *Reversible Jump Markov Chain Monte Carlo*, RJMCMC), que trata o número de fatores como parâmetro.

Em Tzala e Best (2008) é proposto três tipos de modelos fatoriais para dados de

contagens multivariados espaço-temporais sob abordagem bayesiana. O primeiro modelo é o mais simples e apresenta apenas um fator comum no preditor linear para diferentes tempos e locais, além disso nesse modelo não há um erro aleatório específico das variáveis aleatórias. O segundo modelo é uma extensão do primeiro onde são adicionados dois erros aleatórios, um espacial e outro temporal, para capturar os efeitos que não são explicados pelo fator comum. Nesses modelos, o fator comum segue distribuição normal padrão. Já no último modelo, o fator comum é separado em duas componentes independentes, uma espacial e outra temporal, e a estrutura espaço-temporal é modelada na escala dos fatores comuns. O objetivo do estudo é encontrar os padrões espaciais através de um fator comum espaço-temporal ou dos fatores espacial e temporal.

Quanto as restrições de identificabilidade, os autores assumiram que o fator comum dos dois primeiros modelos seguem a distribuição normal padrão. No último modelo o fator espacial e o fator temporal possuem variâncias fixas, e assim o problema que poderia surgir segundo os autores é a inversão do sinal na matriz de cargas. Para evitar tal problema os autores adotaram que a carga fatorial com maior predominância é estritamente positiva. Os autores testaram os dois últimos modelos com duas prioris diferentes para os resíduos espaciais e temporais. O modelo final escolhido dentre os testados foi o último modelo com estrutura espacial através do modelo CAR por ser um modelo simples e de fácil interpretação.

Thorson et al. (2015) apresentam um modelo fatorial espacial aplicado a dados de ecologia, em particular a dados de espécies de animais observados em diferentes localizações. Este modelo se difere dos demais discutidos anteriormente pelo fato da variável resposta seguir uma distribuição de Poisson com sobredisperção log-normal e por considerar dados espaciais de geoestatística. O vetor de valores log-esperados incorpora a parte fatorial do modelo, onde os fatores comuns seguem distribuição normal multivariada com média zero e a matriz de covariância inclui a estrutura espacial com a função de correlação Matérn. Os autores definem a matriz de cargas fatoriais como triangular inferior de posto completo devido aos problemas de identificabilidade dos modelos fatoriais. Além disso, os autores realizaram a rotação na matriz de cargas e nos fatores pelo método varimax para ajudar na interpretação.

A partir da revisão da literatura acerca do tema verificamos diferentes propostas de utilizar os modelos fatoriais sob abordagem bayesiana utilizando dados multivariados espaciais e/ou temporais. Desta forma, o principal objetivo deste trabalho é desenvolver um modelo espaço-temporal em que os fatores comuns incorporem a dependência espacial e temporal. A caracterização do modelo proposto acontecerá de forma similar ao modelo descrito em Wang e Wall (2003), em que os efeitos espaciais existente entre as observações sejam incluídos nos fatores comuns utilizando o modelo condicional autorregressivo - CAR. Já a dependência temporal será incorporada ao modelo através de um modelo dinâmico polinomial de primeira ordem. Tal escolha é fundamentada em artigos que utilizaram modelos fatoriais dinâmicos por meio dos fatores comuns que seguem processos auto-regressivos de ordem 1 descritos anteriormente.

Capítulo 3

Revisão Metodológica

Nesta seção são apresentados os métodos e modelos estatísticos usados para a modelagem dos dados.

3.1 Inferência Bayesiana

Na teoria bayesiana, a incerteza acerca do parâmetro θ é representada por modelos probabilísticos que traduzem a crença que se tem sobre ele. Desta forma, a ideia é fazer inferência sobre esta quantidade desconhecida utilizando o Teorema de Bayes que atualiza o conhecimento após receber novas informações.

Seja H a informação inicial sobre o parâmetro de interesse θ, assuma que a informação inicial pode ser expressa em termos de probabilidade que pode ser resumida por $p(\theta|H)$. Se a informação disponível em H for suficiente para realizar a inferência, então a descrição sobre a incerteza de θ está completa.

Caso a informação inicial não seja suficiente é necessário aumentá-la. Para que isto ocorra, assuma que um vetor de variáveis aleatórias Y relacionado com θ seja observado. Assim, a informação sobre θ passará a ser $H^* = H \cap \{Y = y\}$.

Portanto, podemos resumir a informação de θ por $p(\theta|y, H)$, e deseja-se encontrar $p(\theta|y, H)$ através de $p(\theta|H)$. O Teorema de Bayes é dado por:

$$p(\theta|y, H) = \frac{p(\theta, y|H)}{p(y|H)} = \frac{p(y|\theta, H)p(\theta|H)}{p(y|H)} \tag{3.1}$$

em que

$$p(y|H) = \int_{\Theta} p(y, \theta|H) d\theta \tag{3.2}$$

e Θ é espaço paramétrico de θ. A função do denominador $p(y|H)$, conhecida como distribuição preditiva de y, é apenas uma constante, ou seja, não depende de θ. Portanto, pode-se reescrever o Teorema de Bayes da seguinte forma:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{3.3}$$

Note que, a dependência de H, comum a todos os termos, não é considerada para simplificar a notação. A partir do teorema tem-se a regra para atualização de probabilidades sobre θ iniciando em $p(\theta)$ e chegando a $p(\theta|y)$, sendo denotadas, respectivamente, como distribuições *a priori* e *a posteriori*. Já $p(y|\theta)$ é conhecida como função de verossimilhança de θ, pois determina a função do parâmetro a partir das observações.

A distribuição *a priori* representa o conhecimento sobre θ, quantidade desconhecida, antes de observar o conjunto de dados. Existem formas alternativas de especificar a distribuição *a priori*, por exemplo, prioris conjugadas, prioris não informativas ou prioris hierárquicas. Para mais detalhes veja Migon et al. (2014).

A especificação de prioris conjugadas é uma forma simples de realizar a análise Bayesiana, pois as distribuições *a priori* e *a posteriori* pertencem a mesma classe de distribuições, e, assim a atualização do conhecimento que se tem de θ envolve apenas uma mudança nos hiperparâmetros.

A distribuição *a priori* pode ser considerada não informativa quando existe pouca ou nenhuma informação disponível a respeito dos parâmetros do modelo, ou quando se deseja omitir a opinião do pesquisador. A ideia é realizar a análise baseada no mínimo de informação subjetiva *a priori*.

O uso de prioris hierárquicas visa facilitar a especificação de distribuições *a priori* dividindo-as em estágios ou hierarquias. Com isso, a informação pode ser feita em duas etapas: a primeira estrutural, para a divisão dos estágios, e a segunda subjetiva, para a especificação de cada estágio.

Com a distribuição *a posteriori* e a distribuição preditiva é possível realizar inferências acerca do valor de θ e y, respectivamente, por meio de estimações pontuais ou intervalares. A estimação pontual é utilizada quando deseja-se sumarizar toda a informação presente através de um único número. Já a estimação intervalar é expressa de forma probabilística.

Na maioria das vezes a distribuição *a posteriori* não é conhecida e por isso é necessário recorrer a métodos numéricos. O método de Monte Carlo via cadeias de Markov (MCMC) tem sido amplamente utilizado quando estamos interessados em simular amostras de uma distribuição *a posteriori*, sem que se conheça sua função de densidade de probabilidade.

Os métodos mais utilizados para a construção de cadeias de Markov são: o amostrador de Gibbs, proposto por Geman e Geman (1984) e popularizado por Gelfand e Smith (1990) e o método de Metropolis-Hastings, introduzido por Metropolis et al. (1953) e extendido por Hastings (1970). Estes algoritmos satisfazem as condições da cadeia de Markov ser homogênea, irredutível e aperiódica. Em suma, o objetivo desses algoritmos é simular um passeio aleatório no espaço paramétrico que convirja para uma distribuição estacionária, a distribuição de interesse. Para maiores detalhes sobre os métodos MCMC veja Gamerman e Lopes (2006).

3.2 Monte Carlo Hamiltoniano

O algoritmo Monte Carlo Hamiltoniano (HMC) na sua variante *Not U Turn* (NUTS) foi utilizado para gerar amostras aleatórias da distribuição *a posteriori* através do *software* Stan (Stan Development Team, 2022). O HMC é um método MCMC que utiliza dinâmica hamiltoniana ao invés de uma distribuição de probabilidade para propor transições mais eficientes (Neal et al., 2011). Os métodos fundamentados em HMC são uma alternativa eficiente aos métodos tradicionais MCMC como o amostrador de Gibbs e Metropolis-Hastings na estimação dos parâmetros, pois conseguem gerar cadeias com maior velocidade de convergência e com baixa autocorrelação.

A proposta do algoritmo HMC com a variante NUTS foi introduzida por Hoffman et al. (2014), no qual a principal vantagem é não precisar definir os parâmetros do número de saltos e o tamanho de cada salto que são essenciais no funcionamento do HMC por

meio do algoritmo *Leapfrog*. Desta maneira, a forma adaptativa do HMC, o algoritmo NUTS, pode ser implementado sem a necessidade de ajuste manual.

A implementação de modelos bayesianos com o uso do algoritmo HMC na sua variante NUTS pode ser feita de forma direta pelo *software* Stan, que possui interação com o software *R* (R Core Team, 2022) através do pacote RStan (Stan Development Team, 2020). Com o Stan é necessário apenas especificar a função de verossimilhança do modelo, determinar os parâmetros e as distribuições a priori (de Almeida Inácio, 2017).

3.3 Modelo Fatorial

A análise fatorial é um método estatístico cujo o objetivo é reduzir a dimensionalidade dos dados, ou seja, descrever a estrutura de variabilidade de um conjunto de variáveis correlacionadas entre si através de um grupo menor de variáveis, denominados fatores comuns. Os fatores são ponderados por coeficientes chamados de cargas fatoriais. A parte da variabilidade que não é explicada pelos fatores comuns é associada ao erro aleatório (Johnson et al., 2002).

Atualmente os estudos sobre análise fatorial vêm sendo ampliado em outras áreas de pesquisa, mas tem ganhado espaço principalmente no contexto Bayesiano a partir dos anos 2000 devido ao surgimento de técnicas computacionais adequadas e acessíveis como o Método de Monte Carlo via cadeias de Markov (MCMC) (Geweke e Zhou, 1996; Aguilar e West, 2000; Lopes e West, 2004).

O modelo fatorial Bayesiano básico seguindo a metodologia apresentada em Lopes e West (2004) é definido a seguir. Considere o vetor $\boldsymbol{y}_k = (y_{1k}, y_{2k}, \ldots, y_{qk})'$ no qual para cada observação k $(k = 1, \ldots, K)$, haja q variáveis observáveis. Seja $\boldsymbol{f}_k = (f_{1k}, f_{2k}, \ldots, f_{mk})'$ um vetor aleatório de comprimento m $(m \leq q)$, em que m é o número de fatores que se relaciona a cada um dos elementos de $\boldsymbol{y}_k$. Portanto, o modelo fatorial descrito na forma vetorial é dado pela equação

$$\underset{q\times 1}{\boldsymbol{y_i}} = \underset{q\times m}{\boldsymbol{\beta}} \ \underset{m\times 1}{\boldsymbol{f_i}} + \underset{q\times 1}{\boldsymbol{\epsilon_i}} \tag{3.4}$$

em que:

- os fatores $\boldsymbol{f}_k$ são independentes $\forall k$ com $\boldsymbol{f}_k \sim N(\boldsymbol{0}, \boldsymbol{I}_m)$;
- os erros aleatórios $\boldsymbol{\epsilon}_k$ são independentes $\forall k$ com $\boldsymbol{\epsilon}_k \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, onde $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_q^2)$;
- $\boldsymbol{\epsilon}_k$ e $\boldsymbol{f}_s$ são independentes para todo k e s, $k, s = 1, \ldots, K$;
- $\boldsymbol{\beta}$ é a matriz de cargas fatoriais.

Considerando a forma vetorial do modelo fatorial, a estrutura de covariância dos dados é definida por $var(\boldsymbol{y}_k \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}$. É importante ter uma boa aproximação da matriz de covariâncias que mantenha ao máximo a estrutura dos dados para reduzir a dimensão do problema e possibilitar uma boa interpretação dos fatores.

Cada observação y_{ik}, $i = 1, \ldots, q$, pode ser descrita pela combinação linear dos fatores comuns, tal que $y_{ik} = \beta_{i1}f_{1k} + \beta_{i2}f_{2k} + \ldots + \beta_{im}f_{mk} + \epsilon_{ik}$. Os fatores comuns são capazes de descrever a estrutura de dependência entre as m variáveis, e os coeficientes β_{ij} definem o impacto que cada fator f_{jk} causa em y_{ik}, para $j = 1, \ldots, m$. O termo ϵ_{ik} impacta somente a variável y_{ik} e é descrito como a parte de y_{ik} não explicada pelos fatores comuns.

A estrutura de covariância dos dados sob o modelo é dada por:

- $cov(y_{ik}, y_{jk} \mid \boldsymbol{f}_k, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \begin{cases} \sigma_i^2, & i = j \\ 0, & i \neq j \end{cases}$
- $var(y_{ik} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{l=1}^{m} \beta_{il}^2 + \sigma_i^2 \quad \forall i;$
- $cov(y_{ik}, y_{jk} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{l=1}^{m} \beta_{il}\beta_{jl} \quad \forall i, j, \quad i \neq j;$
- $cov(y_{ik}, f_{jk}) = \beta_{ij} \quad i = 1, \ldots, q, \quad j = 1, \ldots, m.$

Dessa forma, a variância de y_{ik} é dividida em duas partes: a variância explicada pelos fatores comuns e a variância idiossincrática ou específica.

O modelo (3.4) pode ser reescrito na forma matricial como

$$\underset{K \times q}{\boldsymbol{y}} = \underset{K \times m}{\boldsymbol{F}} \; \underset{m \times q}{\boldsymbol{\beta}'} + \underset{K \times q}{\boldsymbol{\epsilon}} \tag{3.5}$$

onde $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)'$, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)'$ e $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_K)'$. As matrizes $\boldsymbol{\epsilon}$ e $\boldsymbol{F}$ seguem distribuição normal matriz-variada e são mutuamente independentes, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}_K, \boldsymbol{\Sigma})$. Então, a função de verossimilhança de $(\boldsymbol{F}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ é dada por

$$p(\boldsymbol{y} \mid \boldsymbol{F}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{K}{2}} \, etr \left\{ -\frac{1}{2} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}')'(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}') \right\}, \tag{3.6}$$

onde $etr(\boldsymbol{X}) = exp(\text{traço}(\boldsymbol{X}))$ para alguma matriz $\boldsymbol{X}$.

Um ponto que merece atenção são as restrições de identificabilidade do modelo. Em Aguilar e West (2000) a matriz de cargas fatoriais é definida sendo triangular inferior de posto completo com os elementos da diagonal principal iguais a 1, e em fixar em zero os elementos do triângulo superior. No entanto, Lopes e West (2004) utilizaram essa restrição alterando os elementos da diagonal principal de $\boldsymbol{\beta}$ a assumirem somente valores positivos.

Para a escolha do número de fatores analisam-se os dados e utilizam-se critérios para decidir qual seria o número razoável de fatores capazes de descrever as variáveis de interesse. Uma sugestão, é utilizar a comparação de modelos com diferentes números de fatores e escolher aquele que mais se adéqua às observações. Uma outra alternativa para a incerteza em relação ao número de fatores é utilizar o algoritmo MCMC com saltos reversíveis (RJMCMC) proposto por Lopes e West (2004) para inferir sobre a quantidade de fatores.

O procedimento de inferência utilizado para o modelo fatorial com o número de fatores conhecido é dado de forma direta utilizando o método MCMC amostrador de Gibbs, para simular amostras da distribuição *a posteriori* a partir das condicionais completas dos parâmetros $\boldsymbol{F}$, $\boldsymbol{\beta}$ e $\boldsymbol{\Sigma}$.

3.4 Modelos Dinâmicos

Os modelos dinâmicos são frequentemente utilizados em diferentes aplicações que variam ao longo do tempo, ou seja, dados de séries temporais. A principal subclasse desses modelos é a dos modelos lineares dinâmicos (MLD) no qual a variável resposta e a evolução dos parâmetros dos estados seguem distribuição normal (West e Harrison, 1997). Além disso, a combinação de modelos dinâmicos com modelos espaço-temporais tem sido cada vez mais abordada em diversas áreas da literatura. Por exemplo

Nesta seção faremos uma breve revisão sobre os modelos lineares dinâmicos (MLD) e modelos lineares dinâmicos generalizados (MLDG).

3.4.1 Modelos Lineares Dinâmicos

Os modelos lineares dinâmicos (MLD) têm como característica a evolução temporal dos parâmetros e a suposição de normalidade para a variável resposta. De acordo com West e Harrison (1997) o modelo na forma matricial pode ser definido através das seguintes equações:

$$
\begin{array}{lll}
\text{Equação das observações:} & \boldsymbol{y}_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t, & \boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{V}_t) \\
\text{Equação de evolução:} & \boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, & \boldsymbol{w}_t \sim N(0, \boldsymbol{W}_t) \\
\text{Distribuição inicial:} & \boldsymbol{\theta}_0 \mid D_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0) &
\end{array}
$$

em que $\boldsymbol{y}_t$ representa as observações de uma série temporal, onde $t = 1, \ldots, T$, $\boldsymbol{F}_t$ é a matriz de desenho conhecida que pode conter variáveis explicativas, componentes de nível, tendência, sazonalidade, entre outros. $\boldsymbol{G}_t$ é a matriz de evolução temporal dos parâmetros, $\boldsymbol{\theta}_t$ é o vetor de estados do modelo, $\boldsymbol{V}_t$ é a matriz de covariância observacional, $\boldsymbol{W}_t$ representa a matriz de covariância de evolução e D_t é o conjunto de informação disponível até o tempo t. Os hiperparâmetros $\boldsymbol{m}_0$ e $\boldsymbol{C}_0$ são conhecidos e representam, respectivamente, a média e a variância da distribuição *a priori* inicial normal multivariada em $t = 0$.

O procedimento de inferência sobre os estados nesta classe de modelos pode ser realizado de forma sequencial sob o ponto de vista bayesiano, como, por exemplo, pode ser utilizado o Filtro de Kalman quando as matrizes $\{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{V}, \boldsymbol{W}\}_t$ são conhecidas. O Filtro de Kalman é um algoritmo que fornece a distribuição condicional de $\boldsymbol{\theta}_t$ dada a informação disponível até o tempo t.

Quando as matrizes $\{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{V}, \boldsymbol{W}\}_t$ não são conhecidas, não é possível utilizar somente o Filtro de Kalman para realizar a inferência sobre os parâmetros. Neste caso, os métodos de Monte Carlo via Cadeias de Markov (MCMC) são utilizados para gerar

amostras da distribuição *a posteriori* do vetor paramétrico. Desta forma, para obter a distribuição condicional completa *a posteriori* do vetor de estados $\boldsymbol{\theta}_t$ utiliza-se o algoritmo FFBS (*Forward Filtering Backward Sampler*).

3.4.2 Modelos Lineares Dinâmicos Generalizados

A classe dos modelos lineares dinâmicos generalizados (MLDG), introduzida por West et al. (1985), é uma extensão dos MLD apresentados na Subseção 3.4.1 que considera dados pertencentes a família exponencial. Dessa forma, com este modelo é possível trabalhar com distribuições assimétricas, como, exponencial e gamma, ou mesmo distribuições discretas, como, binomial e Poisson, dado que é relaxada a suposição de normalidade na evolução dos parâmetros de estado e da variável resposta.

O MLDG pode ser especificado através das seguintes equações:

$$
\begin{aligned}
&\text{Equação das observações:} && p(\boldsymbol{y}_t \mid \eta_t, \phi_t) = exp\left\{\phi_t[\boldsymbol{y}_t\eta_t - b(\eta_t)]\right\} c(\boldsymbol{y}_t\phi_t) \\
&\text{Função de ligação:} && g(\eta_t) = \boldsymbol{F}_t'\boldsymbol{\theta}_t \\
&\text{Equação de evolução:} && \boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \qquad \boldsymbol{w}_t \sim (0, \boldsymbol{W}_t) \\
&\text{Distribuição inicial:} && \boldsymbol{\theta}_0 \mid D_0 \sim (\boldsymbol{m}_0, \boldsymbol{C}_0)
\end{aligned}
$$

em que η_t é o parâmetro natural, ϕ_t é a parâmetro de escala conhecido, $b(\cdot)$ e $c(\cdot,\cdot)$ são funções conhecidas, sendo $b(\eta_t)$ uma função convexa e duas vezes diferenciável. A média é definida por $E(\boldsymbol{y}_t \mid \eta_t, \phi_t) = \vartheta_t = b'(\eta_t)$ e a variância por $V(\boldsymbol{y}_t \mid \eta_t, \phi_t) = \Sigma_t = \frac{b''(\eta_t)}{\phi_t}$, no qual $b'(\cdot)$ e $b''(\cdot)$ são, respectivamente, a primeira e a segunda derivadas da função $b(\cdot)$.

Note que a equação da evolução dos estados é mesma especificada no MLD, diferindo apenas pela distribuição dos erros de evolução ($\boldsymbol{w}_t$), que agora não seguem necessariamente a distribuição normal.

O procedimento de inferência no MLDG sobre oa parâmetros de estados pode ser realizado somente de forma aproximada devido a complexidade. Pode-se utilizar técnicas, como, *Linear Bayes* (Hartigan, 1969) quando as demais quantidades envolvidas no modelo são conhecidas, exceto os parâmetros de estado. Outra alternativa são os métodos Monte Carlo via cadeias de Markov (MCMC) assim como nos MLD.

3.5 Modelos para dados de Área

A modelagem espacial é a área da estatística que envolve o desenvolvimento de modelos cuja estrutura espacial dos dados influencia nos resultados da análise. A incorporação das localizações espaciais na modelagem tem como objetivo descrever ou explicar o comportamento do fenômeno de interesse através dos padrões espaciais encontrados.

Segundo Cressie (1993), a estatística espacial pode ser dividida em três grandes áreas: padrões de pontos, dados de área e geoestatística.

Na análise de padrões de pontos a localização da ocorrência de um determinado processo é aleatória, por exemplo, a localização de crimes em uma cidade. Para dados de área, o espaço contínuo é dividido em áreas das quais os limites estão bem definidos e os dados estão agrupados de acordo com a unidade de análise, por exemplo, o número de óbitos por estado no Brasil. Já a geoestatística considera que o processo aleatório a ser estudado varia ao longo de uma superfície contínua, e que as localizações observadas do processo são conhecidas e podem ser identificadas. Por exemplo, as medições de um poluente em diversos pontos fixos de uma determinada cidade.

O enfoque deste trabalho é na modelagem de dados de área, no qual a região de interesse N é subdividida em um número finito n de sub-regiões. No caso, $N = \cup_{i=1}^{n} N_i$ com $N_i \cap N_j = \emptyset$ se $i \neq j$, tal que $i = 1, \ldots, n$. Portanto, mede-se a variável aleatória estudada em cada subdivisão da região de interesse. Além disso, é necessário especificar na modelagem se há dependência ou não das sub-regiões vizinhas.

De forma geral, quando existe estrutura espacial é de se esperar que as sub-regiões mais próximas sigam comportamento de modo semelhante. Enquanto que para sub-regiões mais distantes, espera-se que as observações sejam menos relacionadas. Para avaliar tal comportamento, utiliza-se a matriz de proximidade espacial ou matriz de vizinhança W, que é uma matriz de pesos que pode ser descrita através das localizações vizinhas. Para este estudo, a matriz W tem os elementos $w_{ij} = 1$ se as sub-regiões i e j são vizinhas, e $w_{ij} = 0$, caso contrário.

O modelo autoregressivo condicional (CAR) introduzido por Besag (1974) é um dos principais modelos para dados de área, sendo usado como estrutura de dependência

espacial em modelos hierárquicos de forma a capturar as possíveis correlações entre as observações. Mais especificamente, neste trabalho usaremos uma subclasse dos modelos CAR, o CAR Intrínseco (ICAR) proposto por Besag et al. (1991).

3.5.1 Modelo CAR

Dado um conjunto de observações em diferentes sub-regiões de uma região N, as interações espaciais entre um par de sub-regiões podem ser modeladas condicionalmente como uma variável aleatória espacial $\phi = (\phi_1, \ldots, \phi_n)'$.

A distribuição condicional completa do processo na área i dados os vizinhos para cada ϕ_i é especificada por:

$$\phi_i | \phi_j, j \neq i, \sim N\left(\sum_{j=1}^{n} w_{ij}\phi_j, \sigma^2\right) \tag{3.7}$$

A distribuição conjunta para os $\phi_{i'}$ é uma variável aleatória normal multivariada centrada em zero (Besag, 1974) conforme:

$$\phi \sim NM\left(\mathbf{0}, Q^{-1}\right) \tag{3.8}$$

em que Q corresponde a matriz de precisão. Para que o modelo seja válido, Q deve ser simétrica e positiva definida. Dessa forma, Q corresponde:

$$Q = [D_\tau (I - \alpha B)] \tag{3.9}$$

no qual W é a matriz de vizinhança $n \times n$ com $w_{ij} = 1$ se as sub-regiões i e j são vizinhas, e $w_{ij} = 0$, caso contrário. $D_\tau = \tau D$, D é a matriz diagonal $n \times n$ em que as entradas $\{i, i\}$ são o número de vizinhos da área i, e as entradas fora da diagonal assumem valor 0. I é a matriz identidade de ordem n. B é a matriz de adjacência escalonada $D^{-1}W$. α é o parâmetro que controla a dependência espacial, $\alpha = 0$ implica em independência espacial e $\alpha = 1$ em correlação espacial completa. Quando $\alpha \in (0, 1)$, a matriz de precisão Q é positiva definida, e, portanto a distribuição conjunta é própria.

Quando o parâmetro que controla a dependência espacial é definido em $\alpha = 1$, tem-se o modelo ICAR que é utilizado como distribuição *a priori* para os efeitos espaciais. Assim, a distribuição conjunta do modelo ICAR é expressa por

$$\phi \sim NM\left(\mathbf{0}, \left[D_\tau\left(I - B\right)\right]^{-1}\right) \tag{3.10}$$

no qual $D_\tau\left(I - B\right) = \tau\left(D - W\right)$, já que $D_\tau = \tau D$ e $B = D^{-1}W$. Note que (3.10) é derivada da distribuição conjunta do modelo CAR. A matriz de covariâncias resultante é singular, e, portanto, tem-se uma distribuição imprópria. Para tornar essa distribuição própria, adiciona-se a restrição $\sum_{i=1}^{n}\phi_i = 0$ ou a restrição para que $\tau \in \left(\lambda_{\min}^{-1}, \lambda_{\max}^{-1}\right)$, em que $\lambda_{\min}$ e $\lambda_{\max}$ são, respectivamente, os autovalores mínimo e máximo da matriz de vizinhança W (Schmidt e Nobre, 2014).

3.6 Critérios de comparação de modelos

Neste trabalho, serão utilizados dois critérios de comparação de modelos que tem por objetivo selecionar o modelo que melhor se ajusta aos dados. De acordo com a metodologia de inferência Bayesiana, cada um dos critérios é brevemente apresentado a seguir. Suponha um conjunto de n observações em que $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ é a variável resposta de um modelo com um conjunto de p parâmetros $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_p)^T$.

3.6.1 Deviance Information Criterion (DIC)

O critério *Deviance Information Criterion* (DIC) proposto por Spiegelhalter et al. (2002) é uma generalização do critério de informação de Akaike (AIC). Seja a deviance $D(\boldsymbol{\Theta}) = -2\log L\left(\boldsymbol{y} \mid \boldsymbol{\Theta}\right)$. O DIC é descrito por

$$DIC = \bar{D} + p_D = 2\bar{D} - D\left(\bar{\boldsymbol{\Theta}}\right), \tag{3.11}$$

em que $\bar{D}$ é a média *a posteriori* da deviance e $p_D = \bar{D} - D\left(\bar{\boldsymbol{\Theta}}\right)$, com $\bar{\boldsymbol{\Theta}}$ sendo as médias *a posteriori* dos elementos do vetor paramétrico $\boldsymbol{\Theta}$.

Observa-se que o DIC é calculado a partir da soma de duas parcelas, a primeira, $\bar{D}$, representa uma medida de qualidade do ajuste, já a segunda parcela, p_D, representa o número efetivo de parâmetros do modelo. Assim, quanto menor for o DIC, melhor será o ajuste do modelo.

3.6.2 Watanabe-Akaike Information Criterion (WAIC)

O critério *Watanabe-Akaike Information Criterion* (WAIC) proposto por Watanabe e Opper (2010) é um critério completamente bayesiano que também foi considerado para comparação entre os modelos. A contrário do DIC, esta medida calcula a média sobre a distribuição *a posteriori*, ao invés de condicionar a uma estimativa pontual. O WAIC conforme Gelman et al. (2014) é descrito por

$$WAIC = -2\left(lppd - p_{\text{WAIC}}\right), \tag{3.12}$$

em que *lppd* é o logaritmo da densidade preditiva pontual no qual indica o nível da qualidade do ajuste do modelo, e é calculado como $lppd = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}_i \mid \boldsymbol{\Theta}^{\boldsymbol{s}})\right)$, sendo S o número de iterações consideradas no MCMC para estimar $p(\boldsymbol{\Theta} \mid \boldsymbol{y})$. O segundo termo, p_{WAIC}, é uma penalização para o número efetivo de parâmetros, sendo calculado por $p_{\text{WAIC}} = \sum_{i=1}^{n} Var_{s=1}^{S}\left(\log(\boldsymbol{y}_i \mid \boldsymbol{\Theta}^{s})\right)$. A interpretação deste critério é semelhante ao DIC, no qual menores valores indicam um melhor ajuste do modelo.

Capítulo 4

Modelos Fatoriais Espaço-Temporais para Dados de Contagem Multivariados

Nesta seção será detalhada a proposta do modelo fatorial espacial dinâmico para dados de contagens multivariados sob a ótica bayesiana em que os fatores incorporam a dependência espacial e temporal existente entre as observações. Antes de descrever o modelo proposto, apresentamos o modelo fatorial espacial com distribuição Poisson elaborado por Wang e Wall (2003) que foi referência para o estudo.

4.1 Modelo Fatorial Espacial

Seja $\boldsymbol{Y}_i$ a variável aleatória multidimensional de dimensão $P \times 1$ observada em uma localização i, o modelo é descrito por

$$Y_{ij}|\theta_{ij} \sim Poisson(\theta_{ij}), \quad i = 1, \ldots, N \text{ e } j = 1, \ldots, P \tag{4.1a}$$

$$\log(\theta_{ij}) = \log(E_{ij}) + \lambda_j f_i \tag{4.1b}$$

em que Y_{ij}, com média θ_{ij}, é o j-ésimo elemento de $\boldsymbol{Y}_i$. A função logarítmica é utilizada como função de ligação no qual o termo E_{ij} é o número esperado, λ_j são as cargas fatoriais e f_i é o fator comum espacial.

O fator comum $\mathbf{f} = (f_1, \ldots, f_N)'$ possui distribuição normal multivariada, no qual a matriz de covariância incorpora a estrutura espacial através do modelo espacial autoregressivo condicional (CAR). Como discutido na subseção 3.5.1, no modelo CAR as correlações espaciais dependem da matriz de vizinhança dos dados. Portanto,

$$\mathbf{f} \sim MVN\left(\mu_f \mathbf{1}_n, \tau^2(\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}\right) \tag{4.2}$$

em que μ_f é a média de f_i, ρ é o parâmetro de correlação espacial do modelo CAR, τ^2 é a variância condicional de $f_i|\boldsymbol{f}_{-i}$, $\boldsymbol{I}_N$ é a matriz identidade de ordem N, e $\boldsymbol{W} = (w_{ij})$ é a matriz de vizinhança, onde $w_{ij} = 1$ se a área i compartilha um lado comum com a área j e $w_{ij} = 0$, caso contrário.

Conforme mencionado anteriormente, o modelo fatorial possui problemas de identificabilidade e por isso são impostas algumas restrições. Neste artigo (Wang e Wall, 2003), os autores fixam a variância $\tau^2 = 1$ devido à indeterminação da multiplicação dos coeficientes das cargas fatoriais com a variância do modelo, já que f_i é não observável. Além disso, é imposta também a restrição $\sum_{i=1}^{N} f_i = 0$.

Para realizar o procedimento de inferência no modelo fatorial espacial é necessário atribuir uma distribuição *a priori* para os parâmetros do modelo. Assumindo independência *a priori* entre os elementos da matriz de cargas e o parâmetro de correlação espacial, considerou-se

- $\lambda_j \sim N(0, \tau_\lambda), \quad j = 1, \ldots, P$. O parâmetro τ_λ é fixado em um valor grande de forma que a distribuição *a priori* seja pouco informativa.
- $\rho \sim Unif(\rho^L, \rho^U)$, em que $\rho^L = \min\left(\frac{1}{e_{min}}, 0\right) < \rho < \frac{1}{e_{max}} = \rho^U$, com e_{min} e e_{max} sendo o menor e o maior autovalores da matriz de vizinhança $\boldsymbol{W}$, respectivamente. Isso garante que a matriz seja positiva definida e o modelo CAR seja prórpio.

A obtenção de amostras das distribuições *a posteriori* e as estimativas dos parâmetros do modelo podem ser obtidas com o uso de métodos MCMC. No artigo, os autores

utilizaram o algoritmo de amostragem com rejeição adaptativa (ARS) para realizar a amostragem MCMC.

4.2 Modelo Fatorial Espaço-Temporal Multivariado

Seja $\boldsymbol{Y}_{it}$ a variável resposta multidimensional de dimensão $P \times 1$ observada em uma localização i e em um instante de tempo t. O modelo fatorial espaço-temporal para dados de contagens multivariados é representado por

$$Y_{ijt} \mid \theta_{ijt} \sim Poi(\theta_{ijt}) \tag{4.3a}$$

$$\log(\theta_{ijt}) = \log(E_{ijt}) + \lambda_j f_{it} \tag{4.3b}$$

$$\boldsymbol{f}_t = \boldsymbol{F}'\boldsymbol{\phi}_t + \boldsymbol{v}_t, \text{ com } \boldsymbol{v}_t \sim MVN(\boldsymbol{0}, \tau^2(\boldsymbol{D} - \rho\boldsymbol{W})^{-1}) \tag{4.3c}$$

$$\boldsymbol{\phi}_t = \boldsymbol{G}\boldsymbol{\phi}_{t-1} + \boldsymbol{u}_t, \text{ com } \boldsymbol{u}_t \sim N(\boldsymbol{0}, \boldsymbol{U}_t) \tag{4.3d}$$

$$\boldsymbol{\phi}_0 \mid D_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0) \tag{4.3e}$$

em que Y_{ijt}, com média θ_{ijt}, é o j-ésimo elemento de $\boldsymbol{Y}_{it}$ para $i =, 1 \ldots, N$, $j = 1, \ldots, P$ e $t = 1, \ldots, T$. O fator comum espacial e temporal é associado à média por meio da função de ligação logarítmica. O termo E_{ijt} é o número de contagens esperadas na região i no tempo t para a variável j, as cargas fatoriais λ_j são parâmetros fixos desconhecidos e f_{it} são os fatores latentes. $\boldsymbol{F}'$ é uma matriz de valores conhecidos, $\boldsymbol{\phi}_t$ é o vetor de estados do modelo e $\boldsymbol{G}$ é uma matriz que descreve a evolução temporal dos parâmetros. Os hiperparâmetros $\boldsymbol{m}_0$ e $\boldsymbol{C}_0$ são conhecidos e representam, respectivamente, a média e a variância da distribuição *a priori* inicial normal multivariada em $t = 0$. O fator comum possui distribuição normal multivariada, no qual a matriz de covariância incorpora a estrutura espacial através do modelo espacial CAR e $\boldsymbol{U}_t$ representa a matriz de covariância do erro de evolução. Nota-se que a distribuição do fator comum 4.3c difere da apresentada em Wang e Wall (2003) por 4.2.

Os dados criminais são constituídos de contagens de casos ou vítimas registrados em cada área. Entretanto, realizar análises somente com as contagens brutas não fará sentido com a realidade, pois dependem das populações de cada área. Portanto, é necessário

realizar o cálculo do valor esperado dos crimes por área de forma a permitir comparações entre as diferentes populações de cada área.

Suponha que temos inicialmente uma região de interesse dividida em N subregiões contíguas e que foram observados P títulos criminais em cada subregião em T períodos de tempo, assim temos que y_{ijt} representa o número de casos ou vítimas observados na subregião $i = 1, \ldots, N$ para cada um dos crimes $j = 1, \ldots, P$ em um instante de tempo $t = 1, \ldots, T$. Considerando que o número médio de casos ou vítimas em determinada subregião é proporcional ao número da população dessa subregião, e tendo observado o número total de casos ou vítimas na região de interesse para um determinado período de tempo t, pode-se calcular o número esperado de casos ou vítimas E_{ijt} em cada subregião i por cada crime j como

$$E_{ijt} = pop_{it} \frac{\sum_i y_{ijt}}{\sum_i pop_{it}}$$

em que pop_{it} é a população da subregião i no tempo t.

O procedimento de inferência para o modelo 4.3 será realizado sob enfoque bayesiano e discutido na próxima subseção, no qual serão apresentadas as distribuições *a priori* para os parâmetros do modelo, as distribuições *a priori* e os esquemas de amostragem utilizados.

4.2.1 Procedimento de Inferência

Considere o vetor paramétrico $\Theta = (\lambda_1, \ldots, \lambda_P, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_T, \boldsymbol{\phi}_1, \boldsymbol{\phi}_T, \boldsymbol{\phi}_0, \tau^2, \rho, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_T)$, que contém todas as quantidades desconhecidas a serem estimadas do modelo 4.3. O núcleo da distribuição *a posteriori* do vetor paramétrico Θ é dado por:

$$
\begin{aligned}
p\left(\Theta \mid \boldsymbol{y}\right) &\propto p\left(\boldsymbol{y} \mid \Theta\right) p\left(\Theta\right) \\
&\prod_{i=1}^{N}\prod_{j=1}^{P}\prod_{t=1}^{T} p\left(y_{ijt} \mid \lambda_j, f_{it}\right) \times \\
&\prod_{t=1}^{T}\left[p\left(\boldsymbol{f}_t \mid \boldsymbol{\phi}_t, \tau^2, \rho\right) p\left(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1}, \boldsymbol{U}_t\right)\right] \times \\
&\quad p\left(\boldsymbol{\phi}_0\right) p\left(\boldsymbol{U}_1\right) \ldots p\left(\boldsymbol{U}_T\right) p\left(\rho\right) p\left(\tau^2\right) p\left(\lambda_1\right) \ldots p\left(\lambda_P\right)
\end{aligned}
$$

A distribuição $p\left(\Theta \mid \boldsymbol{y}\right)$ não possui forma analítica fechada e, portanto, faz-se necessária a utilização de métodos se simulação estocástica para obter amostras desta distribuição *a posteriori*, como os métodos MCMC. Mais especificamente, será utilizado o método Monte Carlo Hamiltoniano apresentado na seção 3.2.

Como a inferência será realizada sob o enfoque bayesiano, é necessário atribuir uma distribuição *a priori* para o vetor paramétrico Θ. As distribuições *a priori* para $\boldsymbol{f}_t$ e $\boldsymbol{\phi}_t$, em que $t = 1, \ldots, T$, seguem diretamente da definição das estruturas dinâmicas em 4.3c, 4.3d e 4.3e, respectivamente. Já para os demais parâmetros de Θ, tem-se as seguintes distribuições *a priori*:

- $\lambda_j \sim N(0, \sigma_\lambda), \quad j = 1, \ldots, P$. σ_λ é fixado de forma que a distribuição *a priori* seja pouco informativa.

- $\rho \sim Unif(\rho^L, \rho^U)$, em que $\rho^L = \min\left(\frac{1}{e_{min}}, 0\right) < \rho < \frac{1}{e_{max}} = \rho^U$, com e_{min} e e_{max} sendo o menor e o maior autovalores da matriz de vizinhança $\boldsymbol{W}$, respectivamente. Isso garante que a matriz seja positiva definida e o modelo CAR seja próprio.

- A raiz quadrada de τ^2 e U_t seguem distribuição half-Cauchy padrão nos reais positivos com os parâmetros de escala sendo, respectivamente, σ_τ e σ_u. De acordo com Gelman (2006), é recomendável utilizar a distribuição *a priori* half-Cauchy em modelos hierárquicos quando há vários parâmetros de variância. Além disso, é uma priori para o desvio padrão pouco informativa e se iguala a distribuição uniforme quando seu parâmetro de escala tende a infinito.

4.3 Estudo Simulado

Nesta seção, serão apresentados dois estudos com dados artificialmente gerados a partir do modelo proposto. O primeiro estudo tem por objetivo de verificar a qualidade de ajuste do modelo com 1, 2 e 3 fatores. O segundo estudo tem por objetivo avaliar os impactos no ajuste do modelo quando a suposição do número de fatores não é adequada.

A rotina computacional do algoritmo do MCMC foi realizada pelos *softwares* R versão 4.1.2 (R Core Team, 2022) e Stan versão 2.21.1 (Stan Development Team, 2022). Além disso, o *software* R foi utilizado para construção dos gráficos e análise dos resultados. Para cada amostra foi rodada uma cadeia contendo 5.000 iterações, sendo descartas as primeiras 1.000 como período de aquecimento (*burn-in*), e espaçamento (*thin*) de 4 entre as iterações resultando em uma amostra final de tamanho 1.000, na qual foi realizada a inferência. Para avaliar a convergência das cadeias de cada parâmetro, utilizou-se o critério de inspeção visual que indicou que para a maioria das cadeias dos parâmetros a convergência foi alcançada.

Os dados foram gerados a partir do modelo fatorial espaço-temporal apresentado em 4.3 considerando $N = 39$ áreas correspondentes as AISP, $P = 11$ títulos criminais e $T = 108$ períodos de tempo. A matriz de vizinhança foi calculada de acordo com as bases cartográficas digitais por AISP divulgadas pelo ISP. Para o cálculo do número esperado de casos foram utilizadas as estimativas populacionais anuais e mensais divulgadas pelo ISP. Quanto à questão das restrições de identificabilidade do modelo, utilizamos a matriz de cargas sendo triangular inferior de posto completo com os elementos da diagonal principal iguais a 1, conforme Aguilar e West (2000).

Cabe ressaltar que também testamos outras restrições na matriz de cargas adotadas na literatura, porém não obtivemos bons resultados nos estudos simulados. Por exemplo, testamos a restrição sugerida por alguns autores do primeiro elemento ser positivo para resolver o problema de inversão de sinal das cargas e dos fatores, porém neste modelo essa restrição não foi suficiente e, portanto, adotamos a restrição mais forte do primeiro elemento ser 1. Além disso, as restrições impostas no artigo de Wang e Wall (2003) do fator comum espacial ter distribuição normal multivariada com média zero e estrutura de

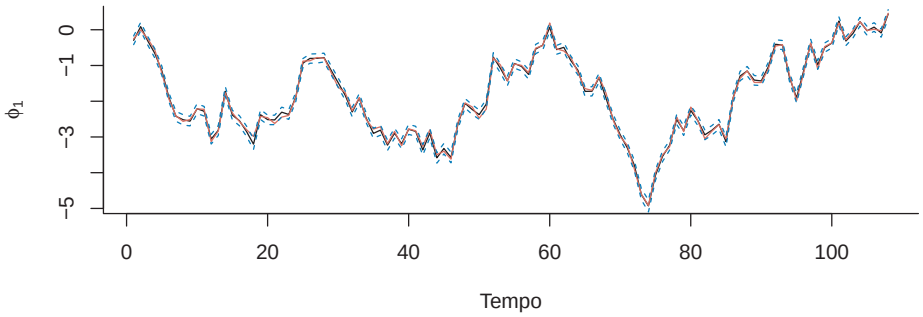
covariância espacial com um parâmetro de variância unitária, ou seja, fixar $\tau^2 = 1$, e da restrição de soma zero dos fatores, não fez diferença no problema de identificabilidade, uma vez que o modelo CAR é próprio

4.3.1 Modelo com 1 fator

Os parâmetros do modelo com 1 fator foram fixados em $\rho = 0,1$, $\tau^2 = 0,65$, $\phi_0 = 0,01$. A matriz $\boldsymbol{U}_t$ é uma matriz diagonal, sendo $U_t = diag(0,25)$. Assume-se que os elementos da matriz de cargas são independentes *a priori* de forma que $\lambda_j \sim N(0, 10^2)$, $j = 1, \ldots, 11$, e $\rho \sim Unif(-0,36; 0,17)$, que também é independente dos coeficientes. O desvio padrão dos parâmetros de variância τ^2 e U_t assumem distribuição half-Cauchy com parâmetro de escala igual a 10. A matriz de cargas foi especificada como $\boldsymbol{\lambda} = (1; 0,57; 0,32; 0,73; 0,56; 0,85; 0,57; 0,21; 0,39; 0,63; 0,19)'$, sendo $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{(1)}, \ldots, \boldsymbol{\lambda}_{(k)})'$ no qual $\boldsymbol{\lambda}_{(k)} = (\lambda_1, \ldots, \lambda_P)$, em que k é o número de fatores do modelo.

Observa-se na Figura 4.1 que a estimação da média *a posteriori* do nível do MFETM com 1 fator é realizada de forma satisfatória, pois os valores estimados se aproximam dos valores reais gerados e os os intervalos de credibilidade de 95% contém os valores verdadeiros em todo o período.

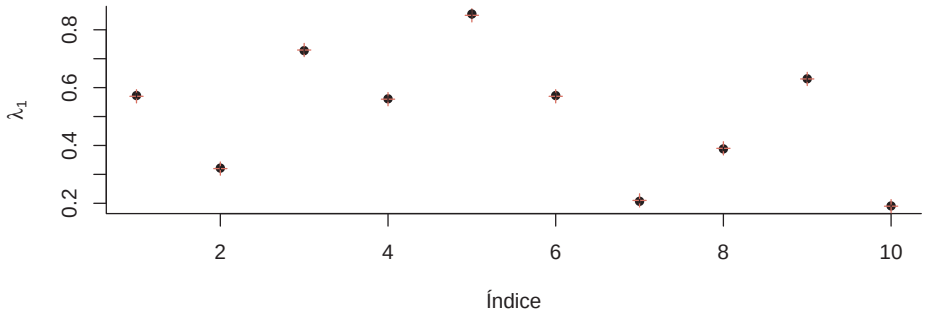
Figura 4.1: Média *a posteriori* do nível do MFETM com 1 fator (linha cheia preta) com respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo.



A Figura 4.2 apresenta o gráfico dos intervalos de credibilidade de 95% com a estimativa da média *a posteriori* e o seu valor verdadeiro das cargas fatoriais. Em todos os

índices, o intervalo de credibilidade contém o valor verdadeiro. Nota-se que o número de parâmetros a serem estimados é $(P-1)$ devido a restrição da matriz de cargas ser triangular inferior de posto completo, com os elementos da diagonal principal fixos iguais a 1.

Figura 4.2: Média *a posteriori* as cargas fatoriais do MFETM com 1 com seus respectivos intervalos de credibilidade de 95%.



A Tabela 4.1 apresenta os valores reais, as médias, medianas e os limites inferiores e superiores dos intervalos de 95% de credibilidade *a posteriori* dos parâmetros do modelo utilizados para gerar os dados artificiais. Observa-se que a maioria dos parâmetros fixos foram bem estimados, ressalta-se que o parâmetro de correlação espacial obteve uma diferença maior comparado ao valor verdadeiro. Entretanto, todos os parâmetros estão contidos no intervalo de credibilidade. Os traços das cadeias dos parâmetros do modelo com 1 fator podem ser vistos na seção A.1.

Tabela 4.1: Sumário da distribuição *a posteriori* para os parâmetros fixos do MFETM com 1 fator.

Parâmetro	Valor Real	Média	Mediana	Quantil 2,5%	Quantil 97,5%
$\sqrt{u_1}$	0,50	0,5038	0,5026	0,4380	0,5780
$\sqrt{\tau_1}$	0,8062	0,7963	0,7959	0,7752	0,8184
ρ_1	0,15	0,1001	0,1059	0,0048	0,1653

4.3.2 Modelo com 2 fatores

Os parâmetros do modelo com 2 fatores foram fixados em $\rho_1 = 0,15$, $\rho_2 = 0,05$, $\tau_1^2 = 0,65$, $\tau_2^2 = 0,25$, $\phi_{1_0} = 0,01$, $\phi_{2_0} = -0,01$, $U_t = diag(0,25; 0,18)$. Assume-

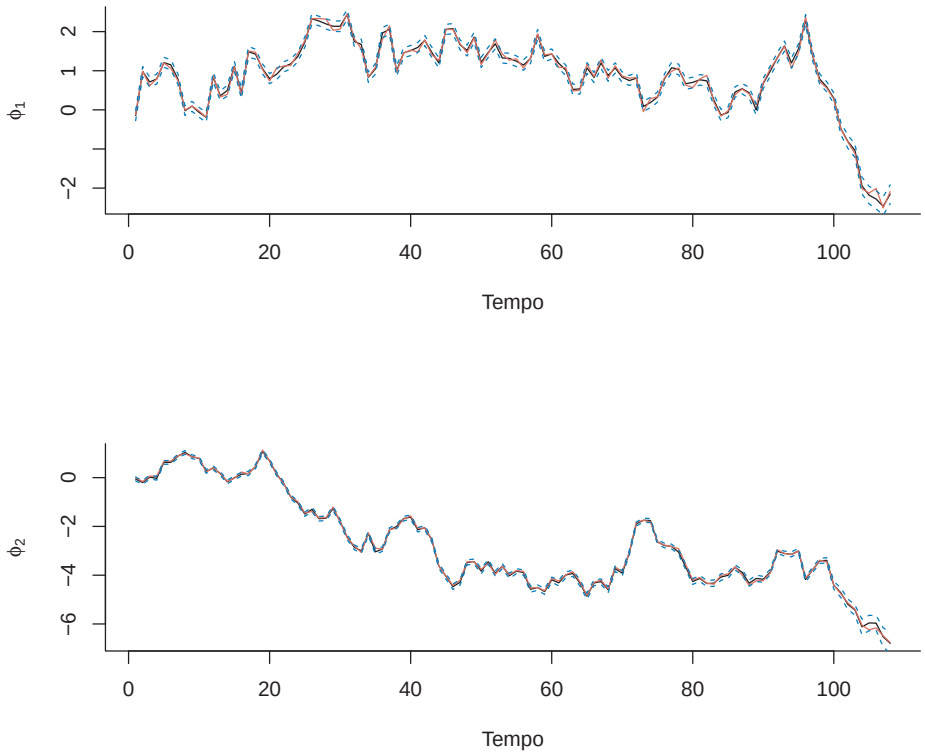
se que os elementos da matriz de cargas são independentes *a priori* de forma que $\lambda_j \sim N(0, 10^2)$, $j = 1, \ldots, 11$, e $\rho_1 \sim Unif(-0,36; 0,17)$, $\rho_2 \sim Unif(-0,36; 0,17)$, que também são independente dos coeficientes. O desvio padrão dos parâmetros de variância τ^2 e U_t assumem distribuição half-Cauchy com parâmetro de escala igual a 10. A matriz de cargas foi especificada conforme a seguir:

$$\boldsymbol{\lambda}' = \begin{bmatrix} 1 & 0,57 & 0,32 & 0,73 & 0,56 & 0,85 & 0,57 & 0,21 & 0,39 & 0,63 & 0,19 \\ 0 & 1 & 0,72 & 0,33 & 0,26 & 0,75 & 0,17 & 0,71 & 0,59 & 0,83 & 0,89 \end{bmatrix}$$

O valor verdadeiro dos vetores de estados do MFETM com 2 fatores, e os vetores de estado estimados são apresentados na Figura 4.3. Verifica-se que as estimativas *a posteriori* conseguem captar a dinâmica do vetor de estados dos dois fatores ao longo do tempo, com intervalos de credibilidade contendo os valores verdadeiros.

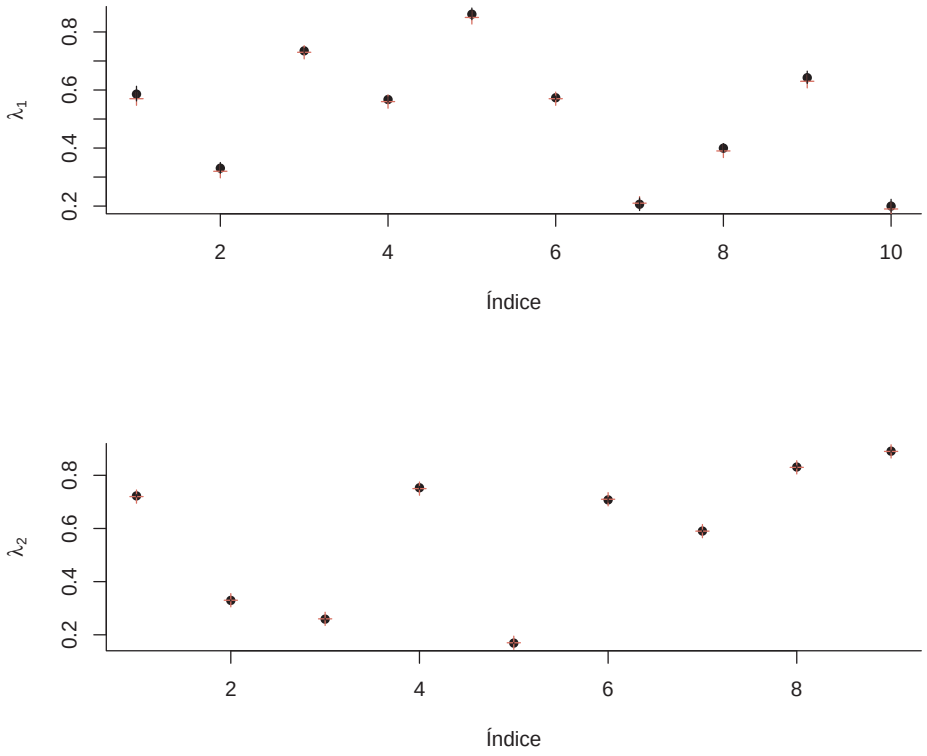
Pela Figura 4.3 observa-se os valores verdadeiros das observações referentes aos dois níveis do MFETM com 2 fatores estão contidos em seus respectivos intervalo de 95% de credibilidade. Nota-se também que as médias *a posteriori* de ϕ_1 e ϕ_2 são capazes de capturar a estrutura dos dados.

Figura 4.3: Média *a posteriori* de cada nível do MFETM com 2 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo.



A Figura 4.4 apresenta os gráficos dos intervalos de credibilidade de 95% com a estimativa média *a posteriori* e o seu valor verdadeiro das cargas fatoriais do MFETM com 2 fatores. Em todos os índices, o intervalo de credibilidade contém o valor verdadeiro.

Figura 4.4: Média *a posteriori* das cargas fatoriais do MFETM com 2 fatores com seus respectivos intervalos de credibilidade de 95%.



A Tabela 4.2 mostra os valores reais, as médias, medianas e os limites inferiores e superiores dos intervalos de 95% de credibilidade *a posteriori* dos parâmetros do modelo utilizados para gerar os dados artificiais. É possível notar que os parâmetros de correlação espacial foram subestimados, mas isto não impactou a estimação. De modo geral, o modelo estimou bem os valores dos parâmetros, com os intervalos de credibilidade contendo os valores verdadeiros. Os traços das cadeias dos parâmetros do modelo com 2 fatores podem ser vistos na seção A.2.

Tabela 4.2: Sumário da distribuição *a posteriori* para os parâmetros fixos do MFETM com 2 fatores.

Parâmetro	Valor Real	Média	Mediana	Quantil 2,5%	Quantil 97,5%
$\sqrt{u_1}$	0,5	0,4523	0,4509	0,3944	0,5195
$\sqrt{u_2}$	0,4243	0,4555	0,4532	0,4002	0,5252
$\sqrt{\tau_1}$	0,8062	0,8068	0,8070	0,7879	0,8262
$\sqrt{\tau_2}$	0,5	0,4846	0,4848	0,4697	0,5004
ρ_1	0,15	0,1392	0,1447	0,0814	0,1689
ρ_2	0,05	-0,0031	-0,0021	-0,1408	0,1282

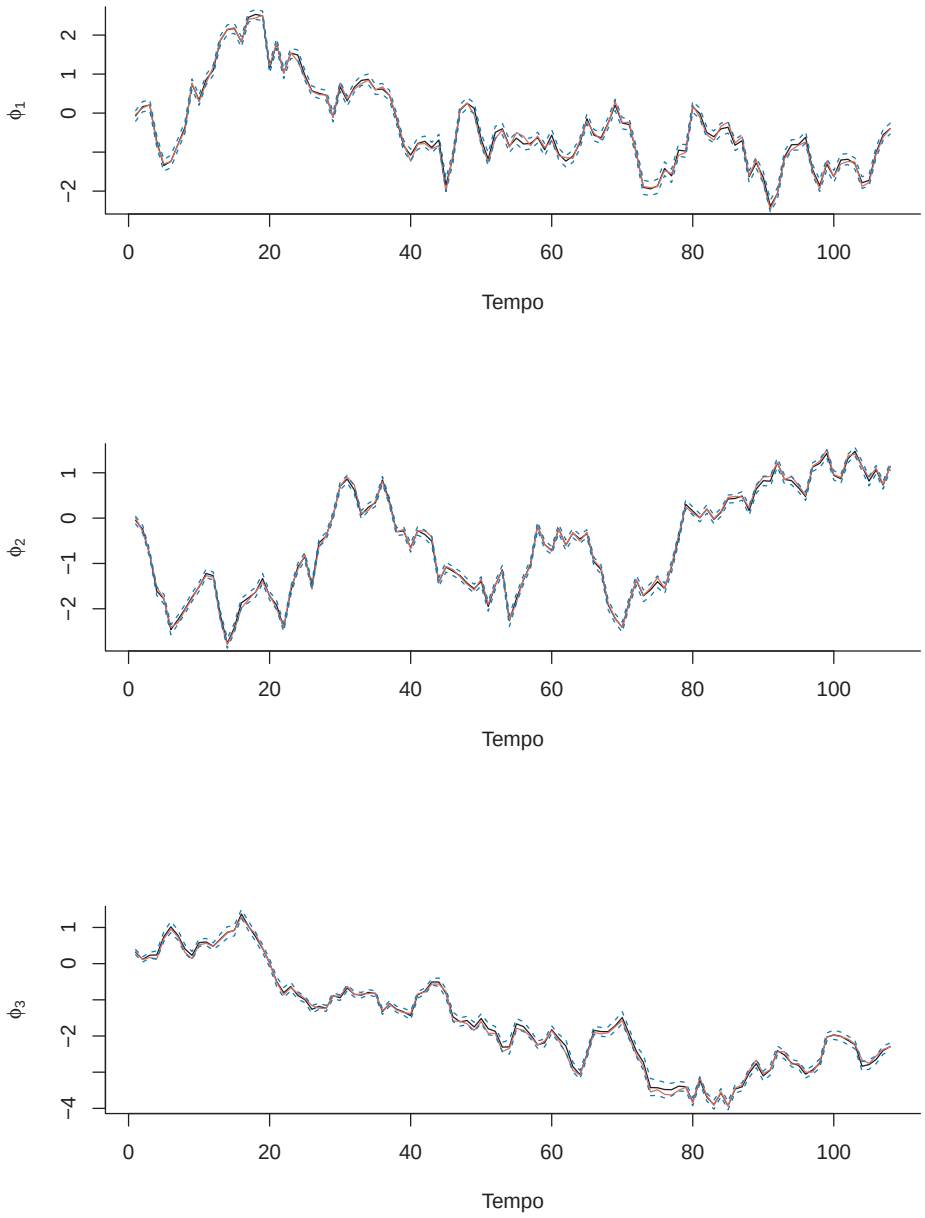
4.3.3 Modelo com 3 fatores

Os parâmetros do modelo com 2 fatores foram fixados em $\rho_1 = 0,15$, $\rho_2 = 0,05$, $\rho_3 = -0,1$, $\tau_1^2 = 0,65$, $\tau_2^2 = 0,25$, $\tau_3^2 = 0,22$, $\phi_{1_0} = 0,01$, $\phi_{2_0} = -0,01$, $\phi_{3_0} = -0,015$, $U_t = diag(0,25; 0,18; 0,12)$. Assume-se que os elementos da matriz de cargas são independentes *a priori* de forma que $\lambda_j \sim N(0, 10^2)$, $j = 1, \ldots, 11$, e $\rho_1 \sim Unif(-0,36; 0,17)$, $\rho_2 \sim Unif(-0,36; 0,17)$, $\rho_3 \sim Unif(-0,36; 0,17)$ que também são independente dos coeficientes. O desvio padrão dos parâmetros de variância τ^2 e U_t assumem distribuição half-Cauchy com parâmetro de escala igual a 10. A matriz de cargas foi especificada conforme a seguir:

$$\boldsymbol{\lambda}' = \begin{bmatrix} 1 & 0,57 & 0,32 & 0,73 & 0,56 & 0,85 & 0,57 & 0,21 & 0,39 & 0,63 & 0,19 \\ 0 & 1 & 0,72 & 0,33 & 0,26 & 0,75 & 0,17 & 0,71 & 0,59 & 0,83 & 0,89 \\ 0 & 0 & 1 & 0,54 & 0,86 & 0,35 & 0,27 & 0,61 & 0,33 & 0,52 & 0,21 \end{bmatrix}$$

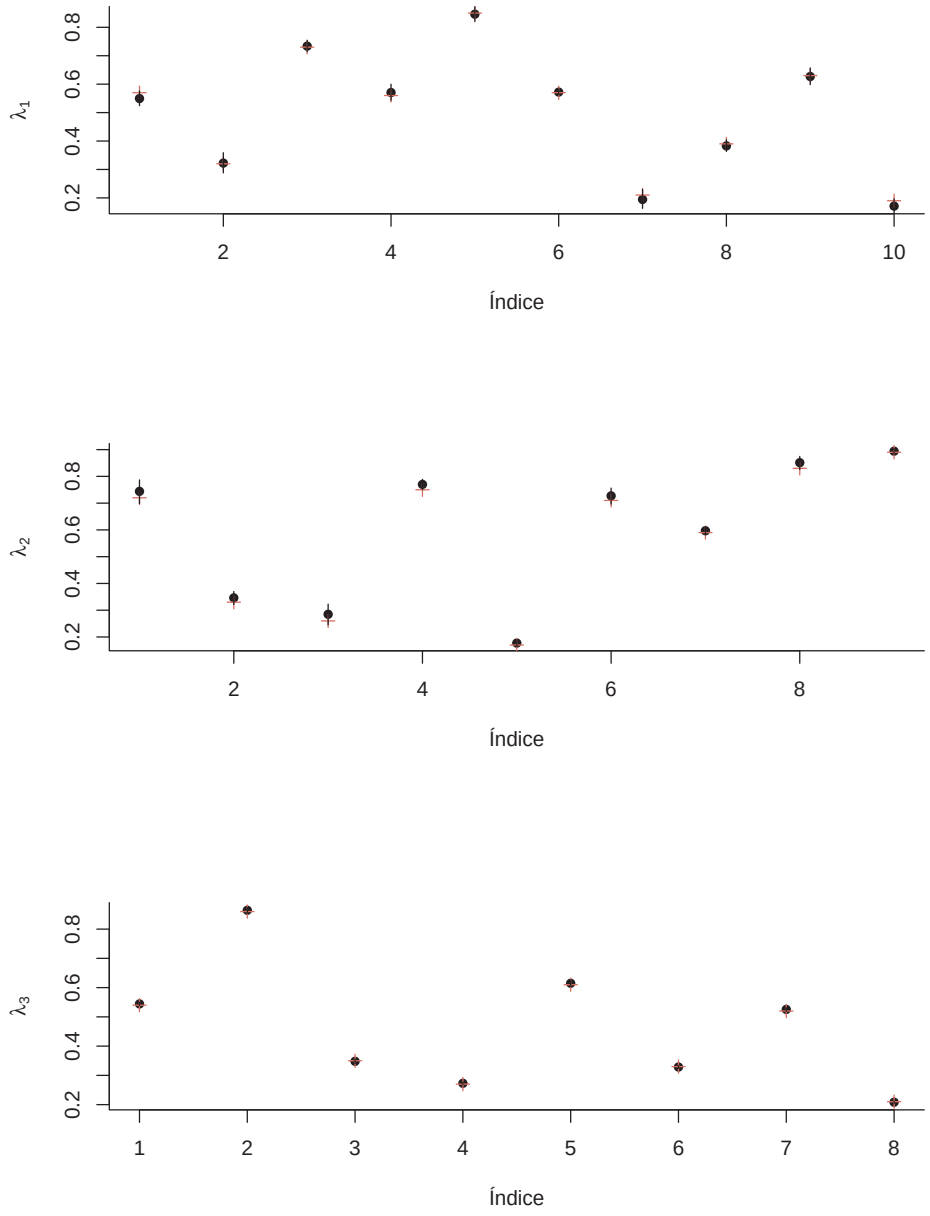
Na Figura 4.5 têm-se o valor verdadeiro para cada nível do MFETM com 3 fatores, com sua respectivas estimativas da média *a posteriori*. Observa-se que as estimativas *a posteriori* conseguem capturar a dinâmica dos dados gerados, com os intervalos de credibilidade contendo os valores verdadeiros e com uma amplitude pequena ao longo de todo o período.

Figura 4.5: Média *a posteriori* de cada nível do MFETM com 3 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo.



A Figura 4.6 apresenta os gráficos dos intervalos de credibilidade de 95% com a estimativa da média *a posteriori* e o seu valor verdadeiro das cargas fatoriais do MFETM com 3 fatores. Em todos os índices, o intervalo de credibilidade contém o valor verdadeiro.

Figura 4.6: Média *a posteriori* das cargas fatoriais do MFETM com 3 fatores com seus respectivos intervalos de credibilidade de 95%.



A Tabela 4.3 contém os valores reais, as médias, medianas e os limites inferiores e superiores dos intervalos de 95% de credibilidade *a posteriori* dos parâmetros do modelo utilizados para gerar os dados artificiais. O valor verdadeiro de cada parâmetro está contido no intervalo de 95% de credibilidade. Em geral, todos os parâmetros foram bem estimados, com exceção dos parâmetros de correlação espacial que apresentam as médias *a posteriori* abaixo do valor real. Os traços das cadeias dos parâmetros do modelo com

3 fatores podem ser vistos na seção A.3.

Tabela 4.3: Sumário da distribuição *a posteriori* para os parâmetros fixos do MFETM com 3 fatores.

Parâmetro	Valor Real	Média	Mediana	Quantil 2,5%	Quantil 97,5%
$\sqrt{u_1}$	0,5	0,5114	0,5114	0,4432	0,5900
$\sqrt{u_2}$	0,4243	0,4309	0,4297	0,3753	0,4955
$\sqrt{u_3}$	0,3464	0,3186	0,3176	0,2750	0,3704
$\sqrt{\tau_1}$	0,8062	0,8130	0,8124	0,7887	0,8358
$\sqrt{\tau_2}$	0,5	0,5050	0,5050	0,4915	0,5196
$\sqrt{\tau_3}$	0,4690	0,4662	0,4661	0,4469	0,4850
ρ_1	0,15	0,0738	0,0762	-0,0305	0,1589
ρ_2	0,05	0,1124	0,1197	0,0147	0,1674
ρ_3	-0,1	-0,1787	-0,1802	-0,3343	-0,0079

4.3.4 Comparando diferentes números de fatores

Este segundo estudo tem por objetivo analisar os resultados de fixar valores diferentes do número de fatores do MFETM com relação ao que foi usado na geração dos dados. Os dados foram gerados para cinco amostras considerando 2 fatores do MFETM. O modelo que gerou os dados possui os seguintes parâmetros: $\rho_1 = 0,15$, $\rho_2 = 0,05$, $\tau_1^2 = 0,65$, $\tau_2^2 = 0,25$, $\phi_{1_0} = 0,01$, $\phi_{2_0} = -0,01$, $U_t = diag(0,25; 0,18)$. A matriz de cargas foi especificada conforme a seguir:

$$\boldsymbol{\lambda}' = \begin{bmatrix} 1 & 0,57 & 0,32 & 0,73 & 0,56 & 0,85 & 0,57 & 0,21 & 0,39 & 0,63 & 0,19 \\ 0 & 1 & 0,72 & 0,33 & 0,26 & 0,75 & 0,17 & 0,71 & 0,59 & 0,83 & 0,89 \end{bmatrix}$$

Em seguida, os mesmos dados foram ajustados para o MFETM com 1 e 3 fatores. A comparação entre os modelos foi realizada usando os critérios DIC e WAIC. As Tabelas 4.4 e 4.5 apresentam os resultados dos critérios de comparação. Observa-se que a diferença dos valores do DIC e WAIC do modelo com 1 fator para os modelos com 2 e 3 fatores é significativa, indicando que o ajuste do MFETM com uma quantidade menor de fatores do que o ideal apresenta resultados piores. Já os resultados das medidas de comparação

para os modelos com 2 e 3 fatores são muito próximos, sendo em alguns casos o modelo com mais fatores como o de melhor ajuste do que o modelo gerador verdadeiro. Os demais componentes dos critérios de comparação DIC e WAIC dos modelos com 1, 2 e 3 fatores e os traços das cadeias de alguns parâmetros podem ser vistos no Apêndice B.

Tabela 4.4: Comparação do MFETM com 1, 2 e 3 fatores para os dados gerados assumindo 2 fatores. Os menores valores do critério DIC, em itálico, indicam o melhor ajuste do modelo.

Modelo	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1 fator	515.816,26	239.336,04	187.999,37	353.323,96	311.942,34
2 fatores	231.937,93	179.613,85	*161.504,45*	*185.836,84*	227.539,09
3 fatores	*231.466,19*	*179.611,58*	161.510,72	185.853,10	*227.533,90*

Tabela 4.5: Comparação do MFETM com 1, 2 e 3 fatores para os dados gerados assumindo 2 fatores. Os menores valores do critério WAIC, em itálico, indicam o melhor ajuste do modelo.

Modelo	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1 fator	555.091,15	245.658,06	191.901,88	375.584,70	323.349,68
2 fatores	232.070,14	179.737,37	161.665,28	*185.909,61*	227.691,86
3 fatores	*231.496,65*	*179.700,82*	*161.659,72*	185.914,86	*227.660,82*

De acordo com Lopes e West (2004) pode existir incerteza nos resultados apresentados pelos critérios de seleção de modelos, pois alguns deles tendem a preferir modelos com maiores números de fatores. Uma forma de verificar a superestimação do número de fatores é avaliar a multimodalidade das amostras *a posteriori* por meio de análises gráficas. Para isso, são apresentados a seguir os gráficos das estimativas *a posteriori* dos parâmetros do MFETM com diferentes números de fatores.

O comportamento do nível do MFETM para cada número de fatores ajustados são apresentados nas Figuras 4.7, 4.8 e 4.9. Observa-se que a média *a posteriori* do nível do modelo com 1 fator não é capaz de capturar bem a estrutura dos dados. O modelo gerado com 2 fatores realiza a estimação dos níveis do MFETM de forma muito precisa, com os intervalos de credibilidade de 95% *a posteriori* contendo os valores verdadeiros. Já no modelo com 3 fatores, Figura 4.9, as estimativas da média *a posteriori* para os dois

primeiros níveis do modelo estão próximas dos valores verdadeiros, indicando modelo conseguiu estimar de forma satisfatória ϕ_1 e ϕ_2 do MFETM com 3 fatores. Entretanto, as estimativas do nível do modelo associado ao terceiro fator, ϕ_3, parecem não ser significativas por estarem em torno de zero, o que é de se esperar pois os dados foram gerados considerando apenas 3 fatores.

Figura 4.7: Média *a posteriori* do nível do MFETM com 1 fator (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo para cada amostra.

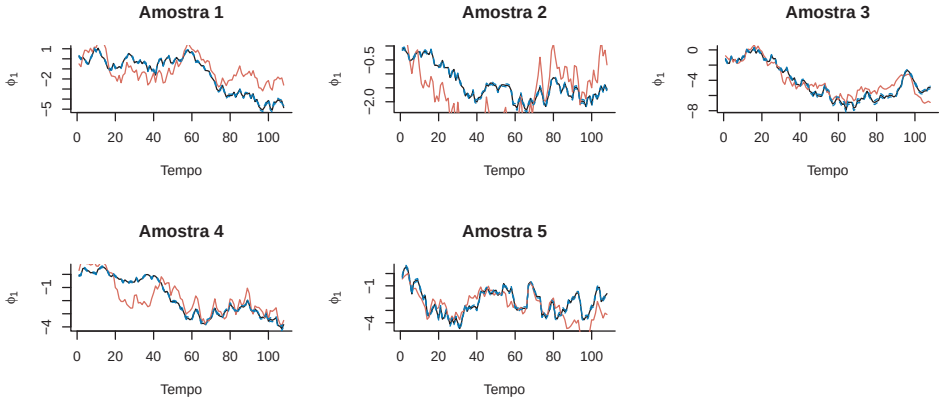


Figura 4.8: Média *a posteriori* de cada nível do MFETM com 2 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo para cada amostra.

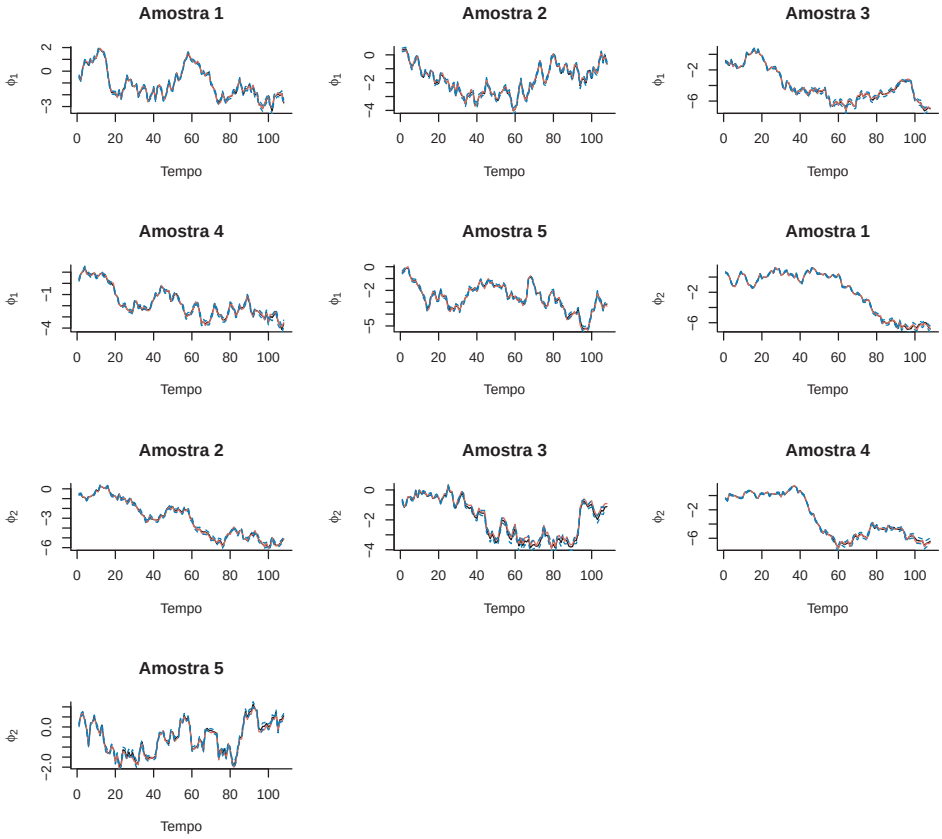
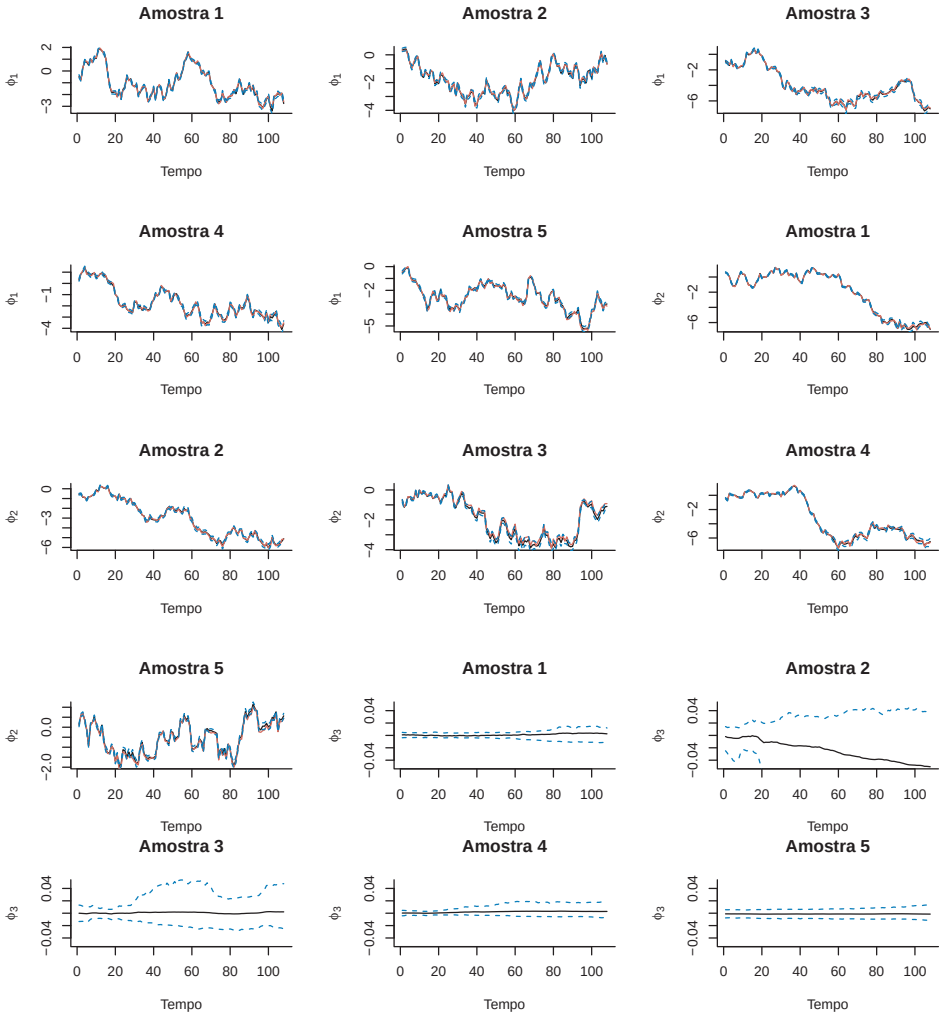


Figura 4.9: Média *a posteriori* de cada nível do MFETM com 3 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) e os valores verdadeiros (linha cheia vermelha) ao longo do tempo para cada amostra.



As Figuras 4.10, 4.11 e 4.12 ilustram a estimativa da média *a posteriori* das cargas fatorais para cada índice. O modelo ajustado com 1 fator não consegue estimar bem as cargas fatoriais. O modelo com 2 fatores, no qual os dados foram gerados, consegue realizar de forma precisa a estimação das cargas fatoriais. Já o modelo com 3 fatores obtém uma boa estimação das cargas fatoriais para λ_1 e λ_2, enquanto que para λ_3 há muita incerteza na estimação das cargas fatoriais pelos intervalos de 95% de credibilidade *a posteriori*, indicando um ajuste ruim.

Figura 4.10: Média *a posteriori* das cargas fatoriais do MFETM com 1 fator com seus respectivos intervalos de credibilidade de 95% para cada amostra.

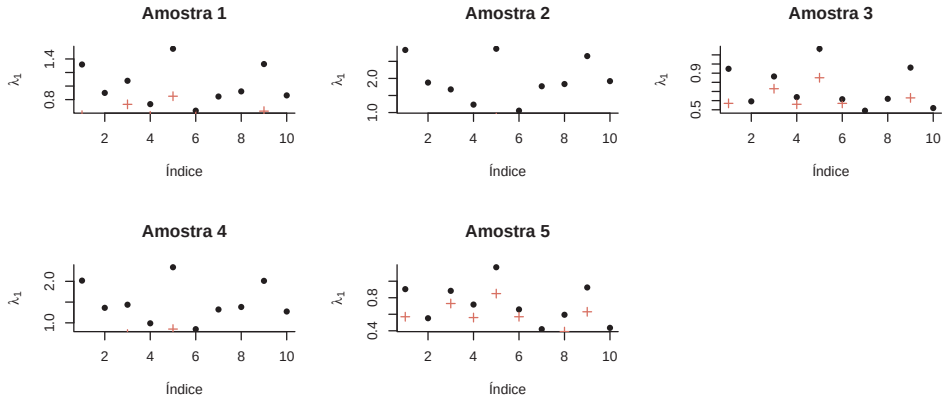


Figura 4.11: Média *a posteriori* das cargas fatoriais do MFETM com 2 fatores com seus respectivos intervalos de credibilidade de 95% para cada amostra.

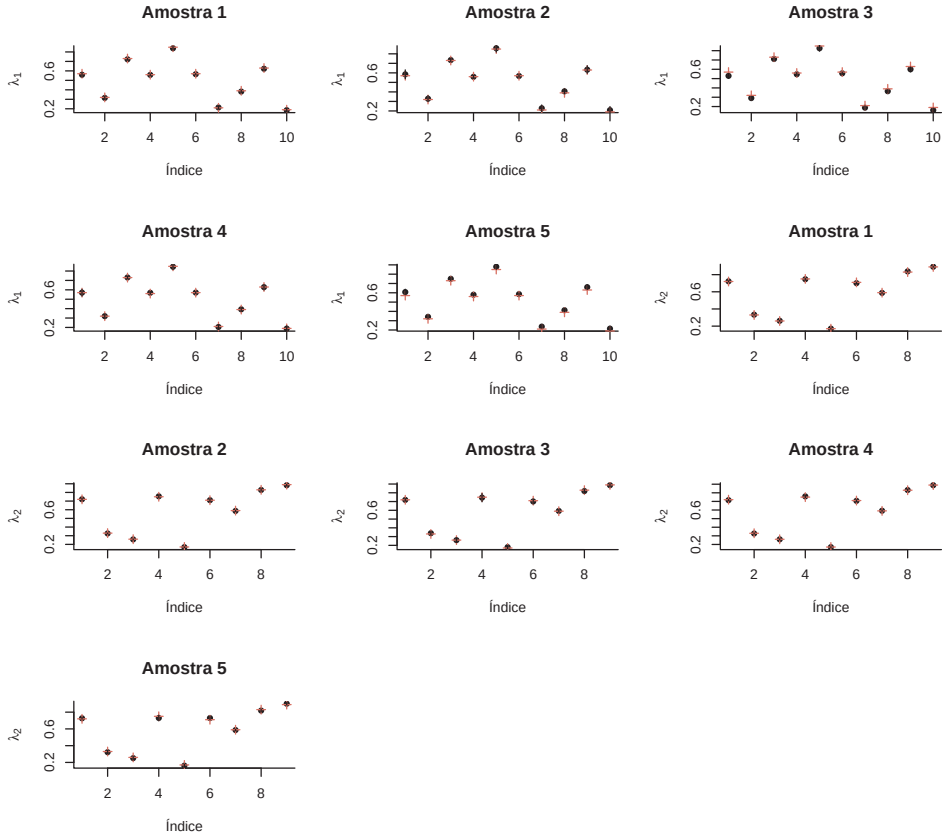
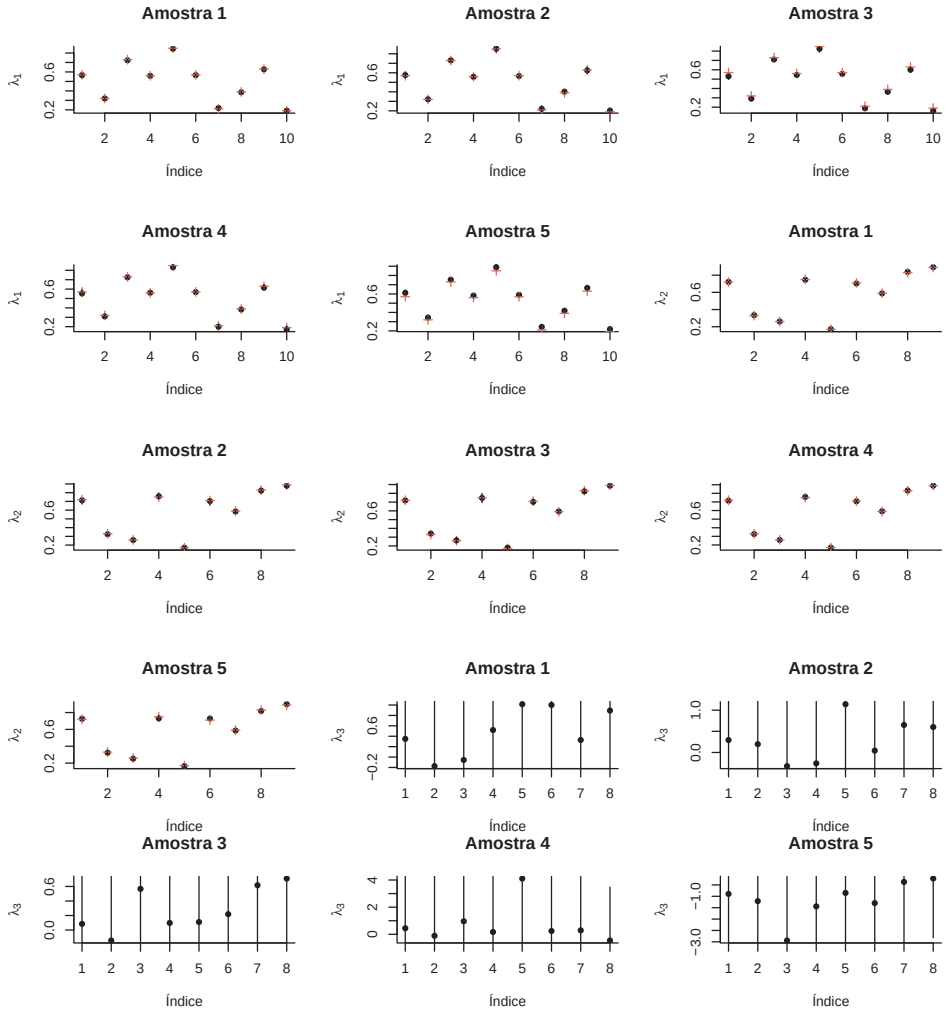


Figura 4.12: Média *a posteriori* das cargas fatoriais do MFETM com 3 fatores com seus respectivos intervalos de credibilidade de 95% para cada amostra.



Conforme analisado nos gráficos das médias *a posteriori* do nível e das cargas fatoriais do MFETM com diferentes números de fatores, o modelo ajustado com 3 fatores apresentou bons resultados para os parâmetros associados aos dois primeiros fatores, enquanto que os parâmetros relacionados ao terceiro fator não obtiveram boas estimativas. Apesar disso, os valores dos critérios de comparação, DIC e WAIC, indicam para algumas amostras que o modelo com 3 fatores possui melhor ajuste do que o modelo com 2 fatores, que é o modelo gerador dos dados. Vale ressaltar, que a diferença dos critérios de comparação entre os modelos com 2 e 3 fatores é muito próxima.

As Figuras 4.13 e 4.14 apresentam os histogramas da distribuição *a posteriori* de

alguns parâmetros da matriz de cargas fatoriais e dos níveis do modelo para algumas amostras do modelo com 2 e 3 fatores, respectivamente. Nota-se a existência de bimodalidade nas amostras dos parâmetros do modelo com 3 fatores. Portanto, embora o modelo com mais fatores indique uma melhor performance do que o modelo que gerou os dados é importante explorar o comportamento das amostras *a posteriori* verificando a existência de múltiplas modas para auxiliar na escolha correta do número de fatores.

Figura 4.13: Distribuição *a posteriori* marginal de $\lambda_{1,3}$, $\lambda_{2,10}$, $\phi_{1,35}$ e $\phi_{2,96}$ para algumas amostras do modelo com 2 fatores.

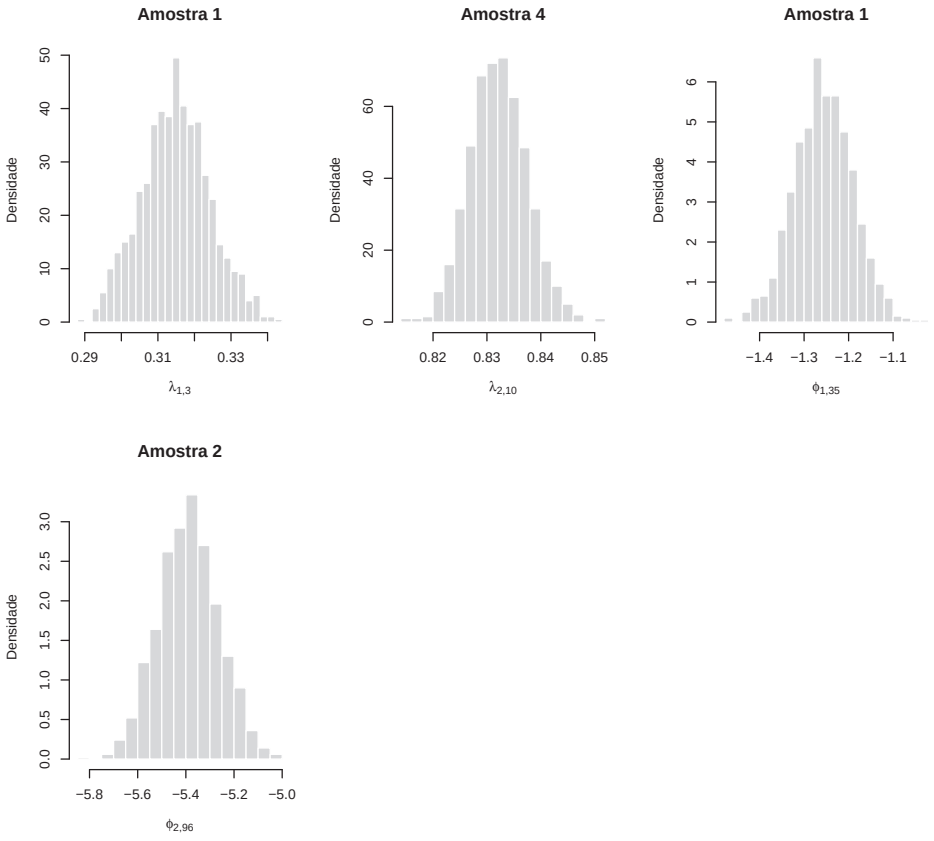
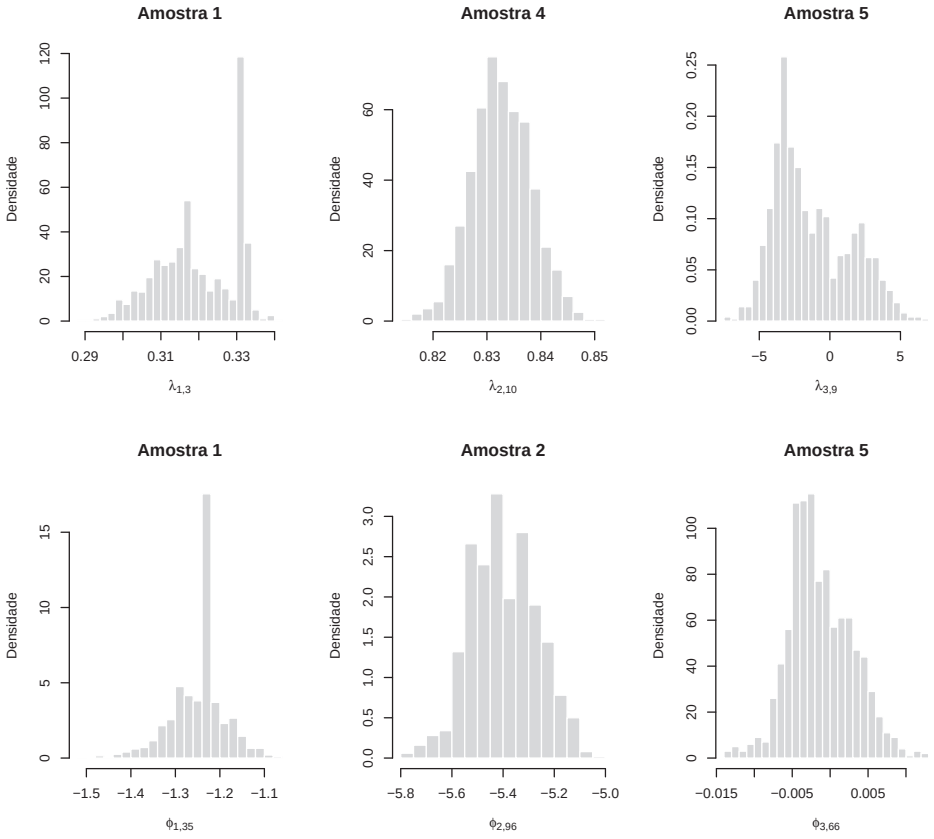


Figura 4.14: Distribuição *a posteriori* marginal de $\lambda_{1,3}$, $\lambda_{2,10}$, $\lambda_{3,9}$, $\phi_{1,35}$, $\phi_{2,96}$ e $\phi_{3,66}$ para algumas amostras do modelo com 3 fatores.



Capítulo 5

Aplicação

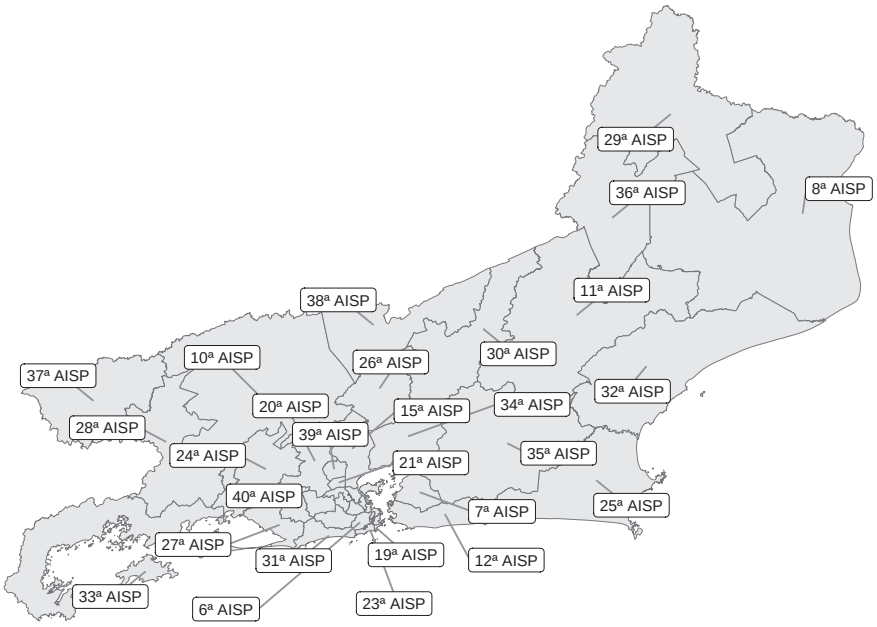
Neste capítulo, a modelagem proposta é aplicada ao conjunto de dados de contagens dos crimes do estado do Rio de Janeiro. A criminalidade nos estados brasileiros é um tema complexo que tem sido explorado por pesquisadores a fim de compreender seu fenômeno através dos dados para formulação de políticas públicas. A alta incidência de crimes nas grandes metrópoles se acentua pelo aumento populacional, desigualdades sociais e econômicas dos habitantes e a falta de infraestrutura das cidades (Pio et al., 2021). O estado do Rio de Janeiro se configura como um dos mais violentos do país, em 2018 o estado apresentou uma das maiores taxas de homicídio por 100 mil habitantes (Alves et al., 2021).

Os dados analisados encontram-se disponíveis no site do ISP[1] e correspondem as contagens mensais de 11 tipos de crimes no período de 2012 a 2020, ou seja, 108 observações mensais, nas 39 Áreas Integradas de Segurança Pública (AISP). Os 11 tipos criminais foram selecionados por serem os crimes mais frequentemente registrados. A Figura 5.1 apresenta a divisão das AISP no estado do Rio de Janeiro, essa divisão visa adequar os limites geográficos de atuação das unidades da Polícia Civil e Militar (ISP, 2022). As AISP correspondem às áreas de atuação de um batalhão da Polícia Militar e às circunscrições das delegacias contidas na área de cada batalhão. Tal área é escolhida devido ao fato das Circunscrições Integradas de Segurança Pública (CISP) poderem ser desmembradas ao

[1] https://www.ispdados.rj.gov.br:4432/

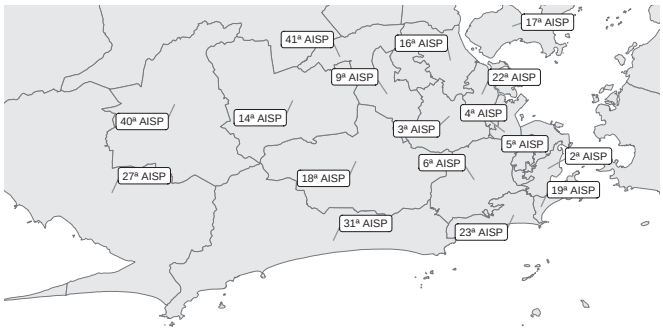
longo dos anos e impactar na análise espacial[2].

Figura 5.1: Divisão Territorial da base de segurança pública por Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro.



A Figura 5.2 mostra o mapa da Divisão Territorial da base de segurança pública por Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com foco na capital do estado a fim de observar as AISP correspondentes dessa região.

Figura 5.2: Focalização da capital do estado do Rio de Janeiro no mapa da Divisão Territorial da base de segurança pública por AISP.



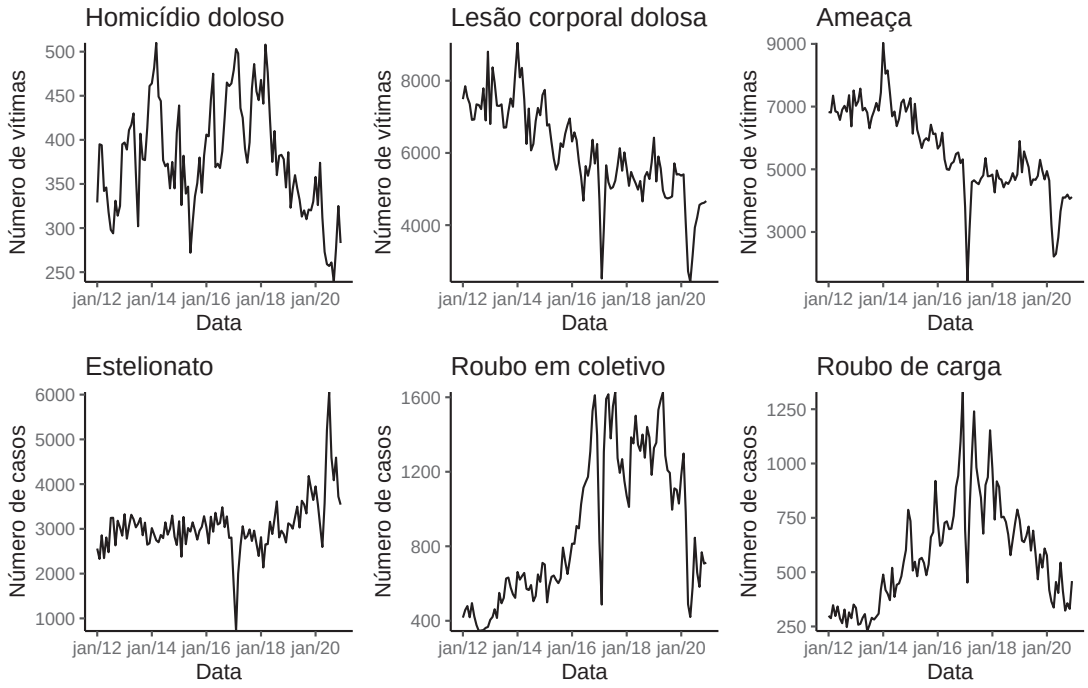
Os registros utilizados neste trabalho contam com o número absoluto de vítimas ou casos registrados em cada uma das 39 regiões, e também com a população mensal e

[2]A relação das AISP com regiões, bairros e municípios do estado do Rio de Janeiro pode ser encontrada em http://www.ispdados.rj.gov.br/Arquivos/Relacaodas%20RISP_AISP.pdf

anual de cada área divulgada pelo ISP. Os 11 crimes analisados possuem características distintas quanto ao tipo de ocorrência, são eles: homicídio doloso, lesão corporal dolosa, lesão corporal culposa (trânsito), roubo a transeunte, roubo de telefone celular, roubo em coletivo, roubo de veículo, roubo de carga, estelionato, apreensão de drogas e ameaça.

A Figura 5.3 apresenta algumas das 11 séries temporais mensais de todo o estado do Rio de Janeiro que serão analisadas no período de janeiro de 2012 a dezembro de 2020. É possível observar uma redução do número de registros no início de 2017 e 2020. Nos três primeiros meses de 2017 houve uma greve dos policiais civis que pode estar relacionada a queda dos registros nesse período. Já em 2020, a pandemia de covid-19 impactou o número de crimes após a decretação da situação de emergência no estado e a aplicação de medidas de isolamento social.

Figura 5.3: Séries mensais de títulos criminais do estado Rio de Janeiro.

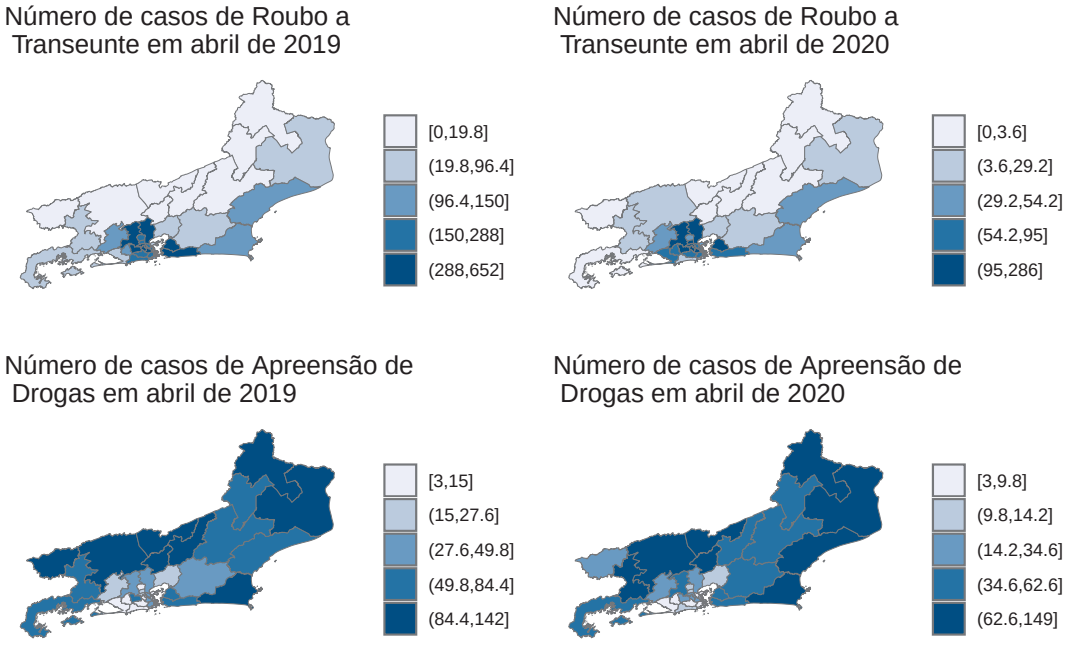


A Figura 5.4 apresenta os quintis da distribuição do número de registros de Roubo a Transeunte e Apreensão de Drogas por AISP em abril de 2019 e abril de 2020. A seleção desses períodos visa observar os efeitos da pandemia no estado do Rio de Janeiro. Dessa forma, nota-se que a distribuição espacial do número de casos de Roubo a Transeunte teve uma pequena variação nas AISP entre esses períodos, e nota-se também uma redução

acentuada do total de casos registrados em abril de 2020. De acordo com o estudo elaborado pelo ISP[3], o distanciamento social devido a pandemia afetou o número de casos de Roubo a Transeunte no estado, já que é um crime de oportunidade realizado nas ruas. Ao passo que o crime de Apreensão de Drogas apresenta uma distribuição espacial mais semelhante nos dois períodos analisados. Além do mais, o estudo do ISP indicou que para esse tipo crime, que está relacionado a proatividade policial, não houve impacto significativo nos registros por causa dos efeitos da pandemia.

Dessa maneira, é possível verificar que a dinâmica espacial varia de acordo com o tipo de crime e as regiões. E os dados considerados para análise são dados de séries multivariadas que são medidas em cada uma das áreas da divisão territorial. Portanto, o objetivo deste estudo é compreender a temporalidade e a espacialização de diferentes tipos criminais nas AISP do estado do Rio de Janeiro utilizando o Modelo Fatorial Espaço-Temporal Multivariado que lida com a complexidade desse tipo de conjunto de dados.

Figura 5.4: Mapa das Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com os quintis da distribuição do número de registros de alguns crimes em determinados períodos de tempo.



[3] http://arquivo.proderj.rj.gov.br/isp_imagens/uploads/impacto-covidNosCrimes2021.html

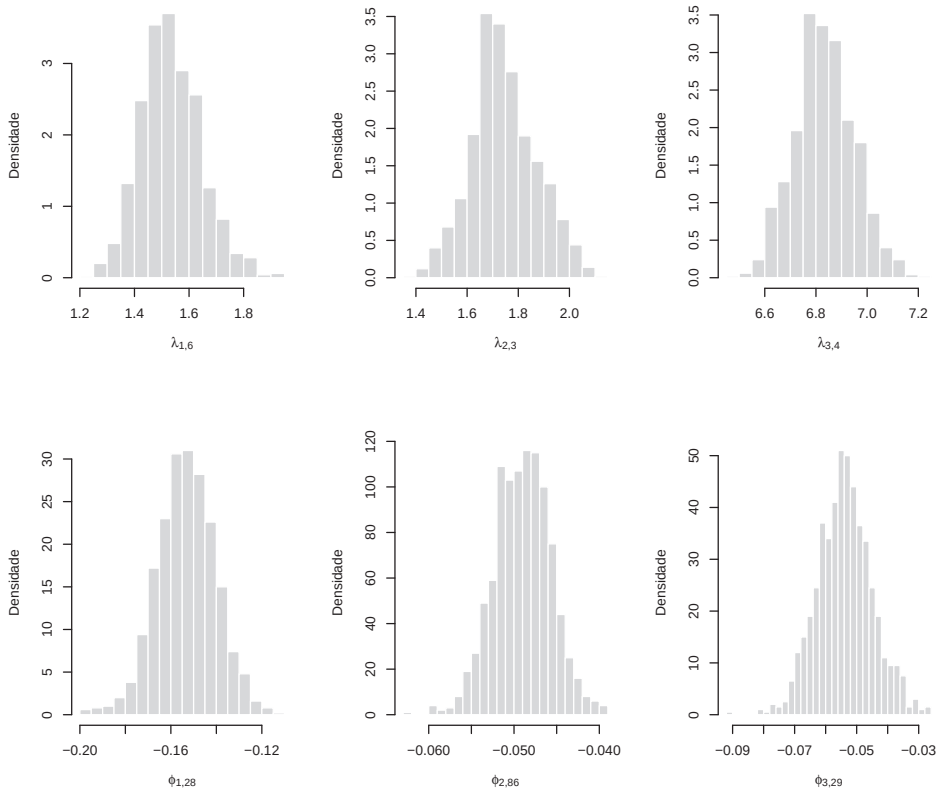
O modelo proposto em 4.3 foi ajustado com os dados reais considerando 2 e 3 fatores e as mesmas distribuições *a priori* para os parâmetros especificados nos exemplos simulados na seção 4.3. Novamente, foram utilizados os *softwares* R versão 4.1.2 e Stan versão 2.21.1 para o ajuste do modelo e análise dos resultados. Para cada modelo foi rodada uma cadeia contendo 10.000 iterações, sendo descartas as primeiras 3.000 como período de aquecimento (*burn-in*), e espaçamento (*thin*) de 7 entre as iterações resultando em uma amostra final de tamanho 1.000, na qual foi realizada a inferência. A convergência das cadeias foi verificada visualmente.

O número de fatores escolhido para análise tem como base o estudo realizado por dos Santos et al. (2021) em que é aplicada a técnica da análise fatorial clássica com 3 fatores para os dados de criminalidade do Rio de Janeiro nos anos de 2006, 2010 e 2016. Dessa forma, é razoável ajustar o MFETM com 2 e 3 fatores. A Tabela 5.1 apresenta os critérios de comparação DIC e WAIC para cada um dos modelos junto com seus componentes, no qual $\overline{D}$ e *lppd* medem a qualidade do ajuste, enquanto p_D e p_{WAIC} são as quantidades que penalizam o número de parâmetros. Observa-se que o modelo com 3 fatores apresenta menores valores, e, portanto, é definido como o melhor. Para confirmar a escolha do modelo são analisadas o comportamento da moda das distribuições *a posteriori* de alguns parâmetros do modelo pela Figura 5.5, e verifica-se que todas são unimodais, não caracterizando o problema de sobrestimação do número de fatores.

Tabela 5.1: Critérios de comparação DIC e WAIC do MFETM ajustado com 2 e 3 fatores da análise dos dados reais.

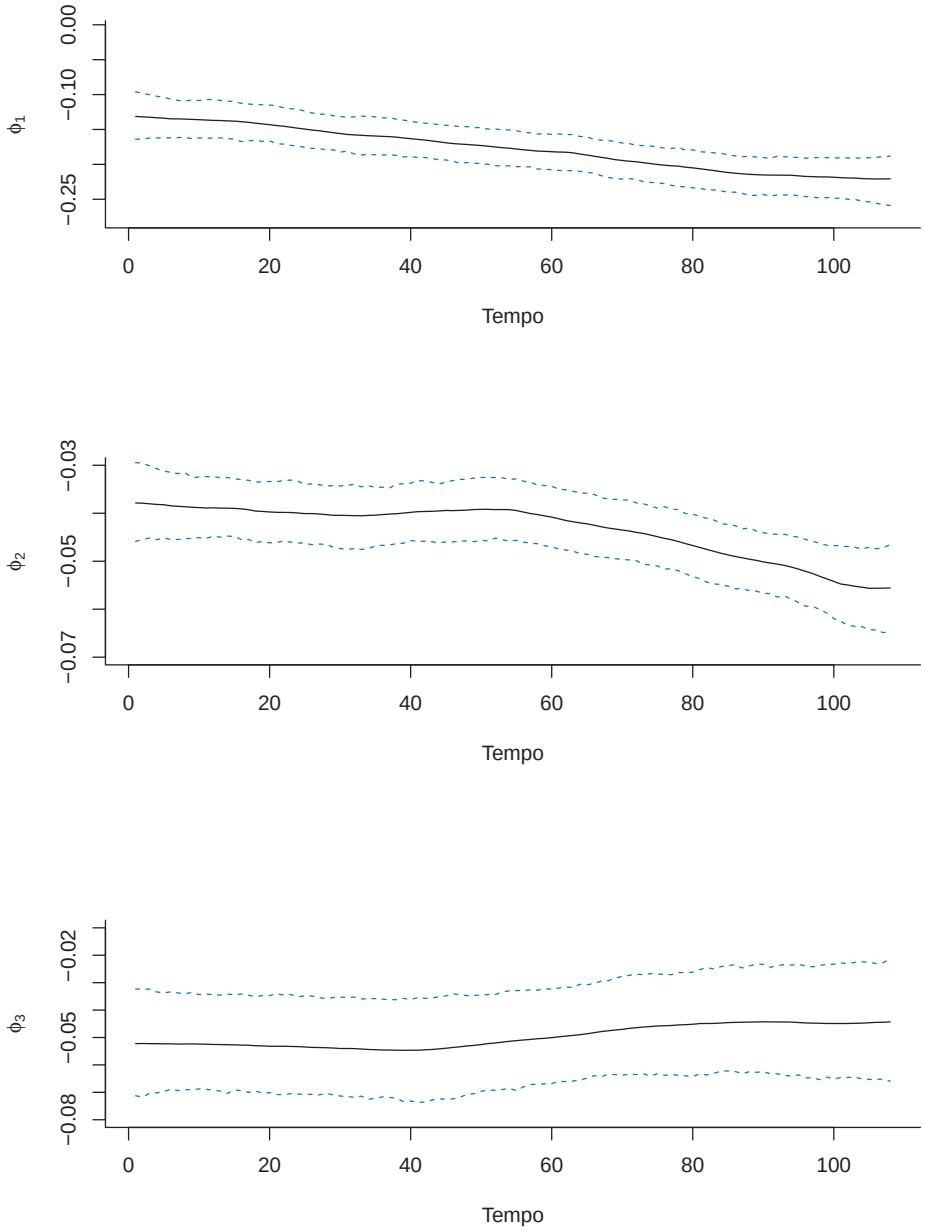
Modelo	$\overline{D}$	p_D	DIC	*lppd*	p_{WAIC}	WAIC
2 fatores	519.100,675	8.263,056	527.363,731	-240.907,250	48.906,557	579.627,614
3 fatores	438.914,818	12.114,170	451.028,987	-204.814,189	39.155,364	487.939,105

Figura 5.5: Distribuição *a posteriori* marginal de $\lambda_{1,2}$, $\lambda_{2,6}$, $\lambda_{3,8}$, $\phi_{1,28}$, $\phi_{2,105}$ e $\phi_{3,58}$ da análise dos dados reais usando o modelo com 3 fatores.



Sendo assim, serão analisados os resultados obtidos com ajuste do modelo com 3 fatores. A Figura 5.6 representa a média *a posteriori* de cada nível do modelo com 3 fatores com seus respectivos intervalos de 95% de credibilidade. Nota-se um decaimento do primeiro e segundo nível, ϕ_1 e ϕ_2, ao longo do período analisado, enquanto o terceiro nível, ϕ_3, apresenta um leve aumento a partir da metade do período de tempo considerado. No Apêndice D pode ser visto o algoritmo realizado no *software* Stan para o modelo com 3 fatores.

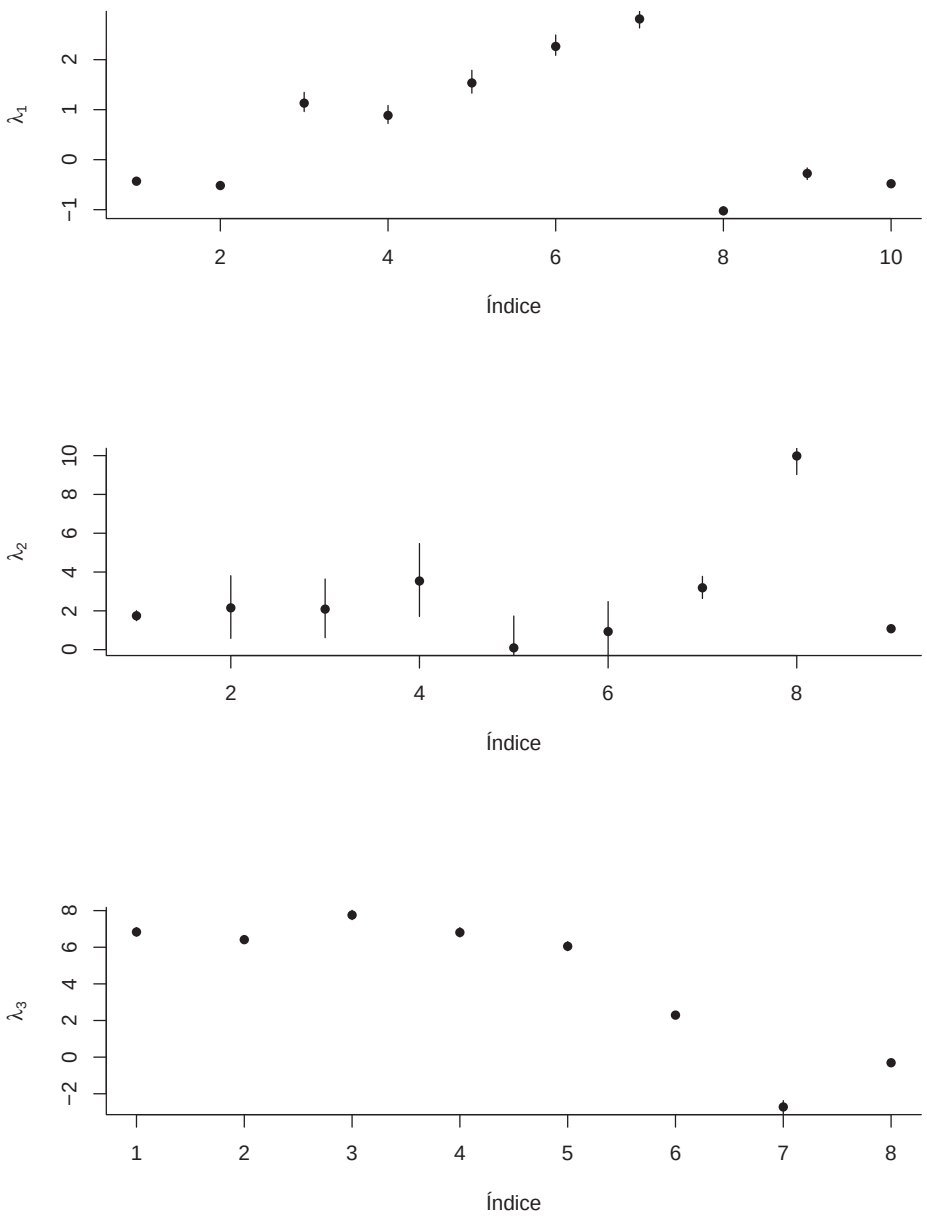
Figura 5.6: Média *a posteriori* de cada nível do MFETM com 3 fatores (linha cheia preta) com seus respectivos intervalos de 95% de credibilidade *a posteriori* (linhas pontilhadas) ao longo do tempo da análise dos dados reais.



A partir da Figura 5.7 podem ser observadas as estimativas da média *a posteriori* das cargas fatoriais do modelo com 3 fatores com seus respectivos intervalos de 95% de credibilidade. Observa-se que os valores referente a primeira e a terceira colunas da matriz de cargas fatoriais, λ_1 e λ_3, apresentam valores maiores para os índices intermediários

que estão associados aos crimes de roubos. Ao passo que o índice 8 em λ_2 se destaca, tal índice corresponde ao crime de apreensão de drogas.

Figura 5.7: Média *a posteriori* das cargas fatoriais do MFETM com 3 fatores com seus respectivos intervalos de credibilidade de 95% da análise dos dados reais.



A Tabela 5.2 contém as médias, medianas e os limites inferiores e superiores dos intervalos de 95% de credibilidade *a posteriori* dos parâmetros do modelo com 3 fatores. Os traços das cadeias dos parâmetros do modelo com 3 fatores apresentados na tabela

abaixo podem ser vistos no Apêndice C.

Tabela 5.2: Sumário da distribuição *a posteriori* para os parâmetros do MFETM com 3 fatores da análise dos dados reais.

Parâmetro	Média	Mediana	Quantil 2,5%	Quantil 97,5%
$\sqrt{u_1}$	0,0055	0,0051	0,0028	0,0095
$\sqrt{u_2}$	0,0013	0,0012	0,0006	0,0023
$\sqrt{u_3}$	0,0017	0,0015	0,0003	0,0040
$\sqrt{\tau_1}$	0,6587	0,6582	0,6299	0,6880
$\sqrt{\tau_2}$	0,1457	0,1457	0,1350	0,1573
$\sqrt{\tau_3}$	0,3874	0,3876	0,3697	0,4044
ρ_1	0,1684	0,1690	0,1641	0,1699
ρ_2	0,1675	0,1682	0,1607	0,1699
ρ_3	0,1693	0,1695	0,1675	0,1700

A média *a posteriori* das cargas fatoriais do MFETM com 3 fatores nos 11 tipos de crimes são apresentadas na Tabela 5.3. Observa-se que no primeiro fator o crime de Homicídio Doloso é predominantemente independente, enquanto o segundo fator está associado aos crimes de Lesão Corporal Dolosa, Lesão Corporal Culposa, Estelionato, Apreensão de drogas e Ameaça. Já no terceiro fator os crimes de roubos estão relacionados, são eles, Roubo a Transeunte, Roubo a Celular, Roubo de Veículo e Roubo de Carga.

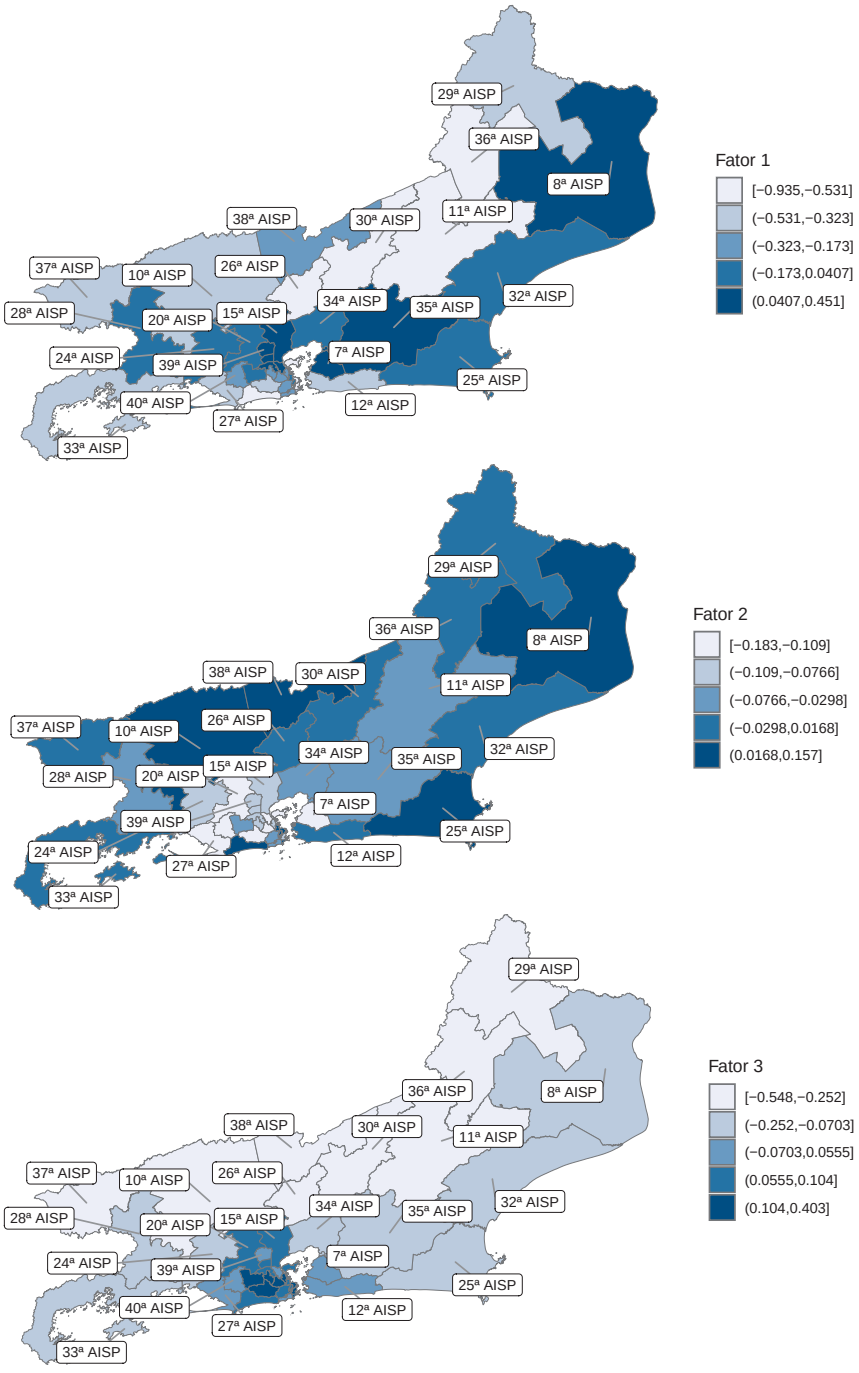
O delito de Homicídio Doloso comumente conhecido como assassinato é um tipo de crime hediondo diferentemente dos demais crimes analisados, assim esse primeiro fator estaria relacionado a crime violento contra a vida. O segundo fator engloba crimes relacionados a lesões contra a vítima e crimes contra a incolumidade pública, assim, esse fator pode ser representado como crimes contra a pessoa e sociedade. O terceiro e último fator agrupa somente variáveis que caracterizam roubo em geral, podendo ser caracterizado como crimes contra o patrimônio. Tais resultados obtidos vão de encontro a estudos realizados com análise fatorial clássica em dados de criminalidade do estado do Rio de Janeiro (Ventorim, 2020; dos Santos et al., 2021; Clemente et al., 2022).

Tabela 5.3: Média *a posteriori* da matriz de cargas fatoriais da análise dos dados reais.

Crime	**Fator 1**	**Fator 2**	**Fator 3**
Homicídio Doloso	1	0	0
Lesão Corporal Dolosa	-0,4305	1	0
Lesão Corporal Culposa	-0,5172	*1,7442*	1
Roubo a Transeunte	1,1310	2,1532	*6,8337*
Roubo a Celular	0,8853	2,0881	*6,4142*
Roubo em Coletivo	1,5345	3,5374	*7,7573*
Roubo de Veículo	2,2646	0,0919	*6,8072*
Roubo de Carga	2,8139	0,9317	*6,0522*
Estelionato	-1,0239	*3,1883*	2,2937
Apreensão de drogas	-0,2755	*9,9825*	-2,7202
Ameaça	-0,4812	*1,0790*	-0,3060

Como os fatores incorporam a estrutura temporal e espacial do modelo proposto, podemos analisar o comportamento dos 3 fatores ajustados aos dados reais. A Figura 5.8 apresenta a distribuição espacial dos quintis de cada um dos fatores por AISP em dezembro de 2020. Observa-se que no Fator 1 as áreas das regiões Metropolitana e das Baixadas Litorâneas do estado do Rio de Janeiro apresentam valores mais altos, enquanto que as áreas situadas na região Serrana do estado apresentaram valores mais baixos. O Fator 2, que está relacionado a crimes contra a pessoa e sociedade, em especial tráfico de drogas, apresenta uma distribuição espacial diferente do primeiro fator. Nota-se que as regiões do Noroeste e Centro-Sul Fluminense, que englobam as AISP 8, 10 e 38, se destacam por apresentarem valores mais elevados. Essas regiões, são regiões de fronteira com outros estados do Sudeste, que estão relacionadas a entrada de drogas no estado (Novellino e Oliveira, 2019). O terceiro fator que está associado a crimes contra o patrimônio se difere dos demais fatores por estar concentrado principalmente na região Metropolitana do estado.

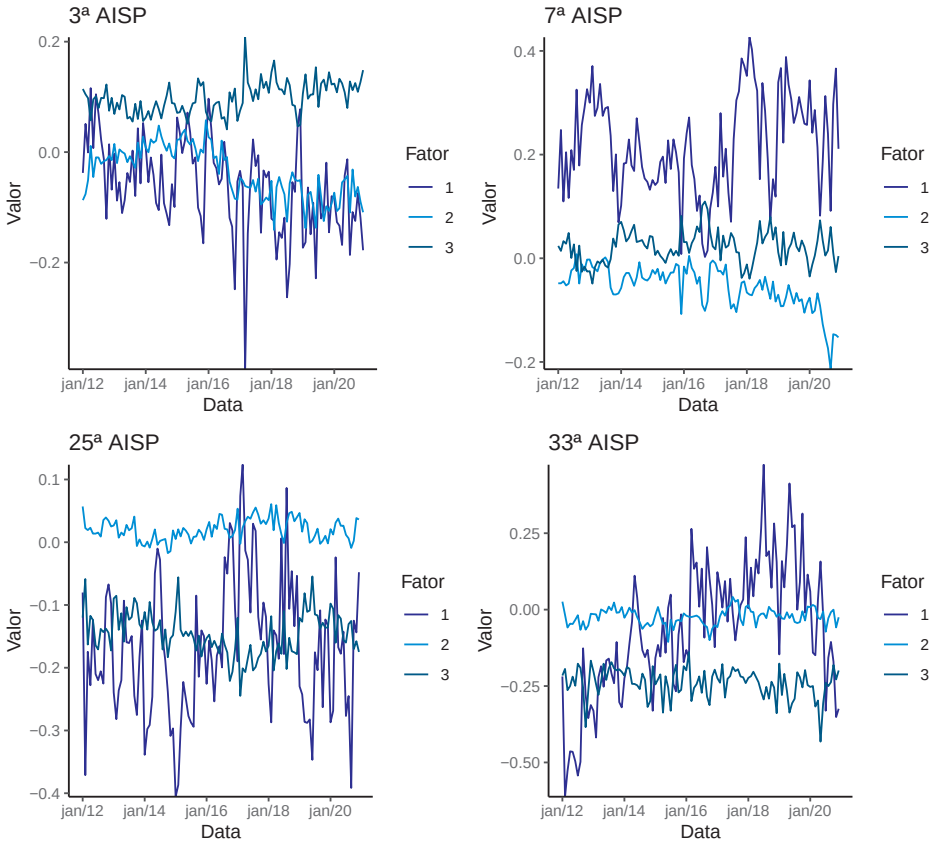
Figura 5.8: Mapa das Áreas Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro com os quintis da distribuição dos valores de cada um dos 3 fatores em dezembro de 2020.



A Figura 5.9 mostra a variação temporal mensal dos fatores no período entre 2012 e 2020 em algumas das AISP consideradas no estudo. Observa-se na 3ª AISP que contém

bairros da zona norte da cidade do Rio de Janeiro, como, Méier, Jacarezinho, Piedade, etc, que o terceiro fator se destaca dos demais. Em contrapartida, na 7ª AISP, associada ao município de São Gonçalo, o primeiro fator possui maiores valores ao longo do tempo. Já em relação a 25ª, o segundo fator se destaca dos demais, sendo a região das Baixadas Litorâneas integrante dessa área. Por último, a 33ª AISP que compreende a região da Costa Verde apresenta uma tendência de aumento do primeiro fator até 2018, enquanto no final do período analisado essa posição se inverte, em que o segundo e terceiro fator apresentam maiores valores. As demais outras históricas por AISP podem ser vistas no Apêndice C.

Figura 5.9: Séries mensais de cada um dos 3 fatores nas 3ª, 7ª, 25ª e 33ª AISP.



Capítulo 6

Conclusões

Este trabalho teve como objetivo propor um modelo fatorial bayesiano espaço-temporal para dados de contagens multivariados agregados por áreas. Essa metodologia se destaca por considerar a dependência espacial e temporal em conjuntos de dados multivariados. No modelo proposto, os fatores incorporam essas duas dependências através de modelos lineares dinâmicos generalizados com distribuição Poisson e do modelo autoregressivo condicional. O interesse deste estudo e estimar o processo latente subjacente existente em problemas multivariados espaço-temporais.

O procedimento de inferência foi realizado sob o enfoque bayesiano com o uso do algoritmo Monte Carlo Hamiltoniano como método de amostragem, sendo uma alternativa eficiente aos métodos tradicionais MCMC. Verificou-se pela metodologia proposta que os resultados obtidos nos estudos simulados resultaram em uma boa performance na estimação dos parâmetros, evidenciando que o modelo e o método de estimação propostos são adequados para modelar conjuntos de dados com esse tipo de estrutura.

Na análise com os dados artificiais realizou-se um estudo comparativo para verificar os efeitos resultantes de fixar um valor diferente do número de fatores daquele usado na geração dos dados. O estudo indicou que quando o número de fatores é sobrestimado ocorre a presença de multimodalidade nas amostras *a posteriori*. Além disso, alguns critérios de comparação de modelos foram utilizados para auxiliar na escolha do número de fatores.

Os resultados obtidos a partir da aplicação de diferentes tipos de crimes por Áreas

Integradas de Segurança Pública (AISP) do estado do Rio de Janeiro no período de 2012 a 2020 fornecem informações relevantes sobre a segurança pública do estado, que estão diretamente relacionadas a questões sociais. Apesar dos dados serem complexos por possuírem estrutura multivariada, temporal e espacial, é possível verificar a correlação espacial que certos tipos de crimes possuem e como variam ao longo tempo. Dessa forma, o ajuste do modelo com 3 fatores foi eficaz ao separar os diferentes crimes conforme a literatura desta temática, sendo possível ainda analisar os fatores pelas áreas de segurança e com os períodos de tempo considerados no estudo.

6.1 Trabalhos Futuros

Devido a flexibilidade do modelo, diversas estruturas podem ser incorporadas como a inclusão de sazonalidade, covariáveis e também a estimação do número de fatores. Em geral, padrões sazonais estão presentes em muitas séries temporais e precisam ser considerados na modelagem. No modelo proposto tentamos incluir a sazonalidade através dos fatores comuns, entretanto a estimação dos parâmetros apresentou desafios que precisam ser melhores estudados. A estimação do número de fatores pode ser considerada pelo algoritmo MCMC com saltos reversíveis, denotado por RJMCMC Lopes e West (2004).

Um outro interesse seria na otimização computacional do modelo realizado no *software* Stan, pois devido a complexidade do modelo e quantidade de parâmetros o custo computacional pode ser alto e, portanto, precisa ser aprimorado.

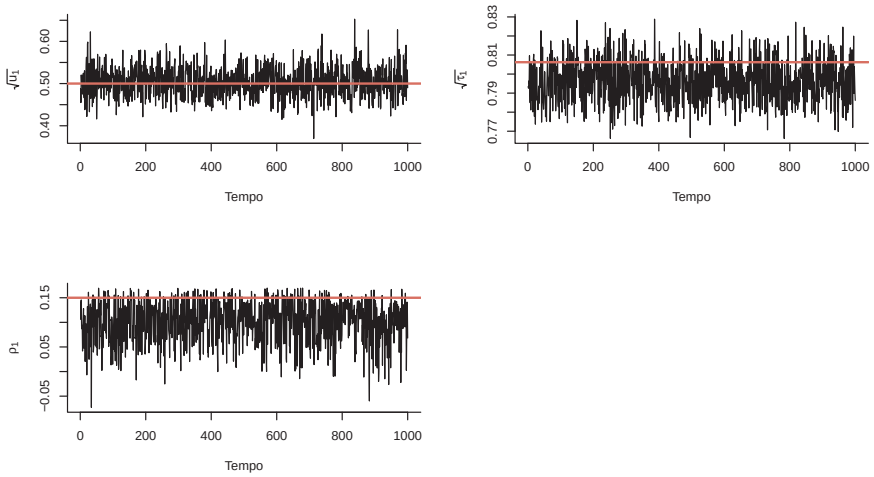
Apêndice A

Traços das Cadeias dos Parâmetros do Estudo Simulado

Neste apêndice, apresentamos os traços das cadeias de alguns parâmetros para cada um dos modelos ajustados nos Capítulos 4 e 5.

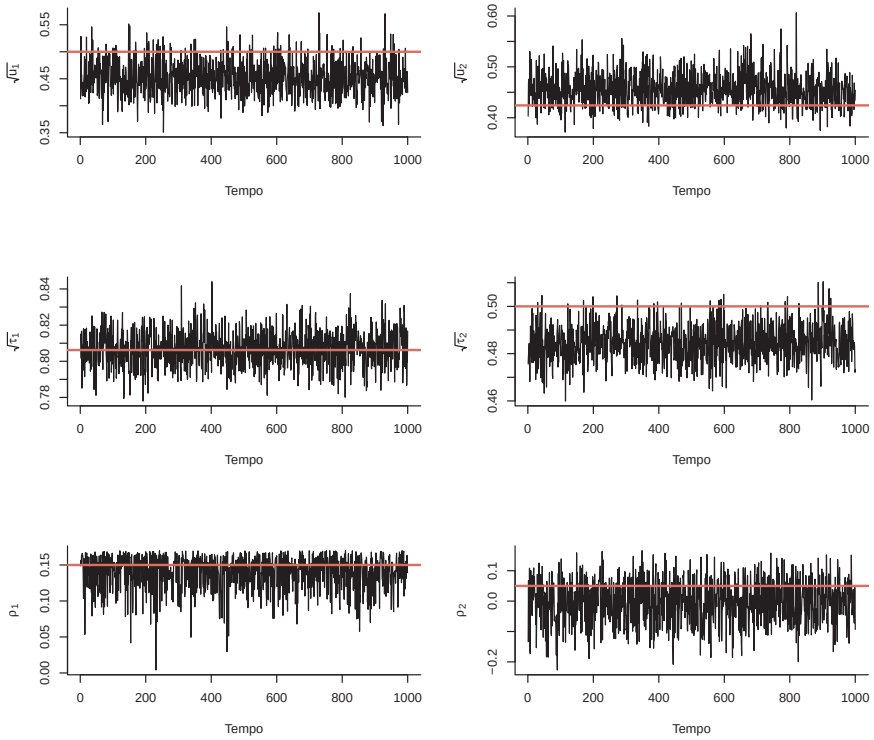
A.1 Modelo com 1 fator

Figura A.1: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{\tau_1}$ e ρ_1 do modelo com 1 fator, a linha vermelha representa o valor verdadeiro.



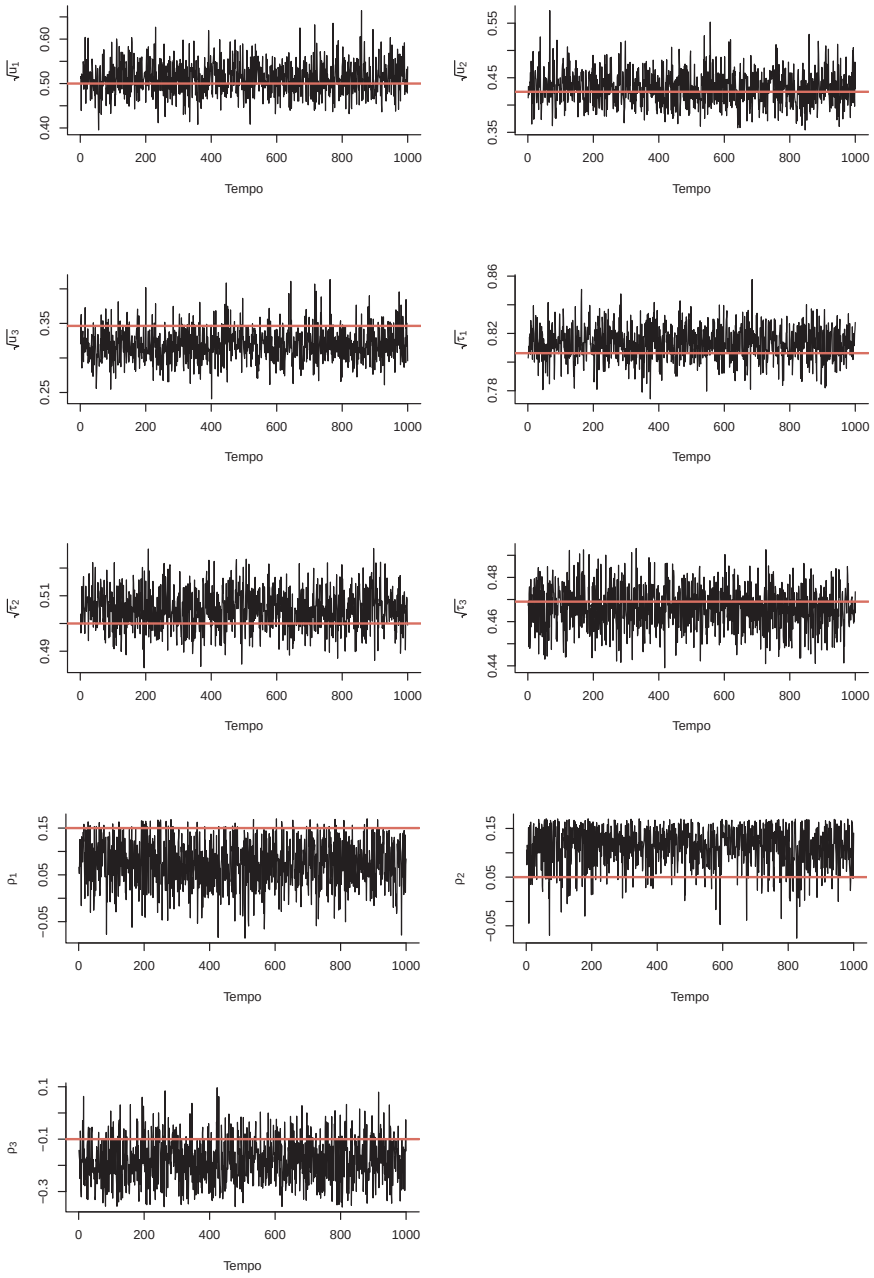
A.2 Modelo com 2 fatores

Figura A.2: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, ρ_1 e ρ_2 do modelo com 2 fatores, a linha vermelha representa o valor verdadeiro.



A.3 Modelo com 3 fatores

Figura A.3: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{u_3}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, $\sqrt{\tau_3}$, ρ_1, ρ_2 e ρ_3 do modelo com 3 fatores, a linha vermelha representa o valor verdadeiro.



Apêndice B

Componentes dos Critérios de Comparação e Traços das Cadeias dos Parâmetros do Estudo Comparativo

Neste apêndice, apresentamos os seguintes componentes dos critérios de comparação DIC e WAIC: $\overline{D}$, p_D, $lppd$ e p_{WAIC} referente ao estudo de comparação de diferentes números de fatores. Também apresentamos os traços das cadeias de alguns parâmetros para cada um dos modelos ajustados deste estudo de comparação.

Tabela B.1: Valores de $\overline{D}$ do critérios de comparação DIC para cada amostra de cada modelo.

Modelo	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1 fator	512.450,81	236.311,24	185.458,85	350.557,96	308.298,12
2 fatores	226.145,57	175.031,77	157.522,82	181.212,34	221.622,03
3 fatores	226.077,31	174.937,96	157.424,33	181.152,58	221.420,67

Tabela B.2: Valores de p_D do critérios de comparação DIC para cada amostra de cada modelo.

Modelo	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1 fator	3.365,44	3.024,80	2.540,53	2.766,00	3.644,22
2 fatores	5.792,36	4.582,09	3.981,62	4.624,50	5.917,06
3 fatores	5.388,88	4.673,63	4.086,39	4.700,52	6.113,22

Tabela B.3: Valores de *lppd* do critérios de comparação WAIC para cada amostra de cada modelo.

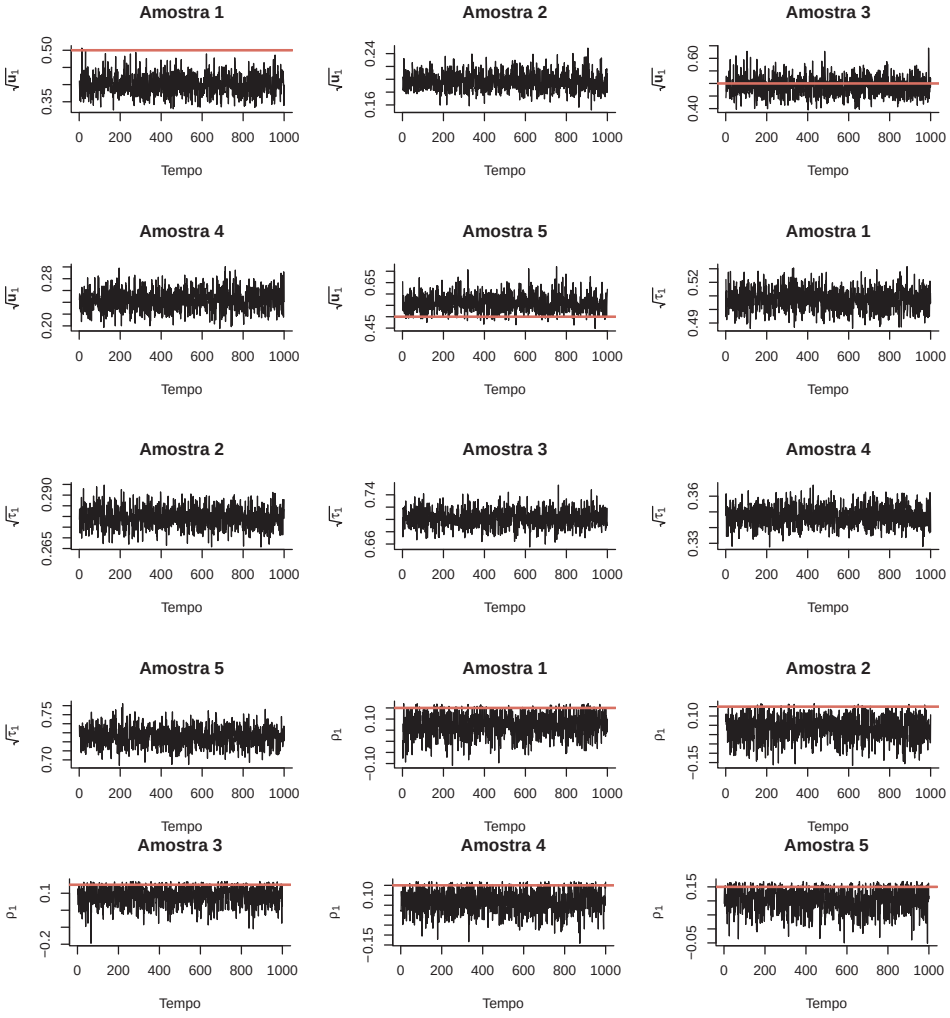
Modelo	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1 fator	-241.540,42	-114.521,14	-90.234,05	-166.623,83	-148.402,12
2 fatores	-111.038,79	-85.795,79	-77.251,34	-88.956,93	-108.669,07
3 fatores	-111.109,30	-85.725,11	-77.166,43	-88.903,35	-108.507,78

Tabela B.4: Valores de p_{WAIC} do critérios de comparação WAIC para cada amostra de cada modelo.

Modelo	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1 fator	36.005,16	8.307,90	5.716,89	21.168,52	13.272,72
2 fatores	4.996,28	4.072,89	3.581,30	3.997,87	5.176,86
3 fatores	4.639,03	4.125,29	3.663,43	4.054,08	5.322,63

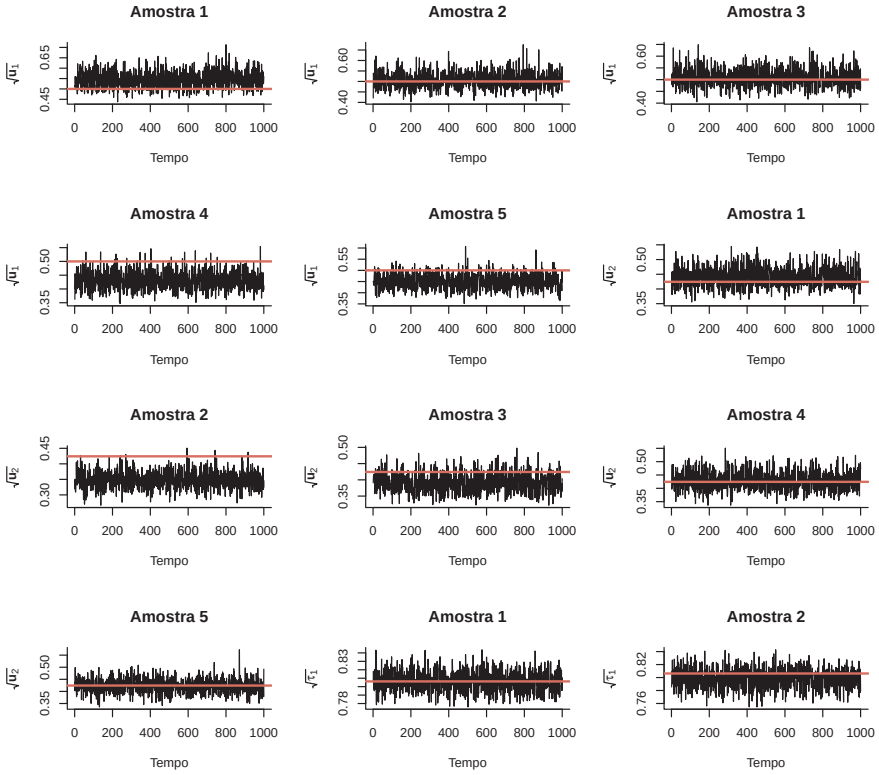
B.1 Modelo com 1 fator

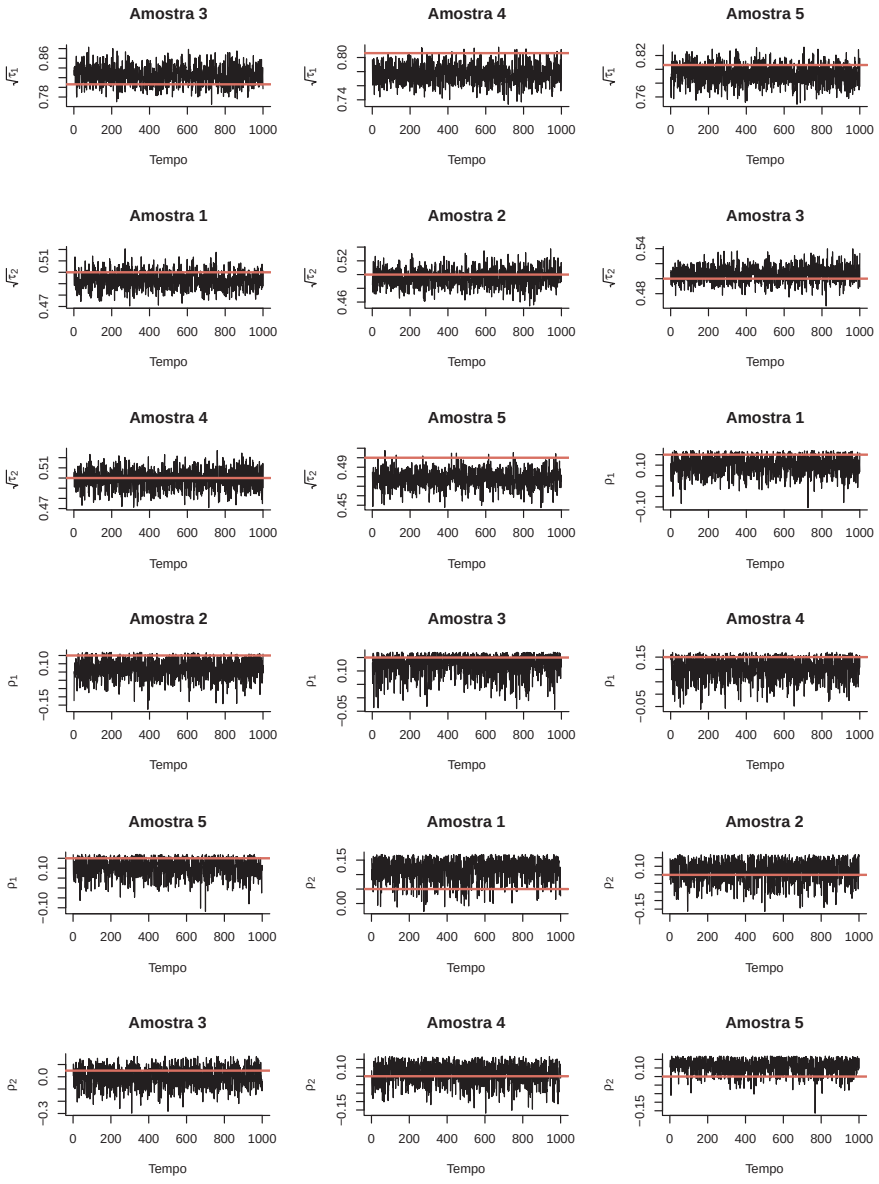
Figura B.1: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{\tau_1}$ e ρ_1 do modelo com 1 fator para cada amostra, a linha vermelha representa o valor verdadeiro.



B.2 Modelo com 2 fatores

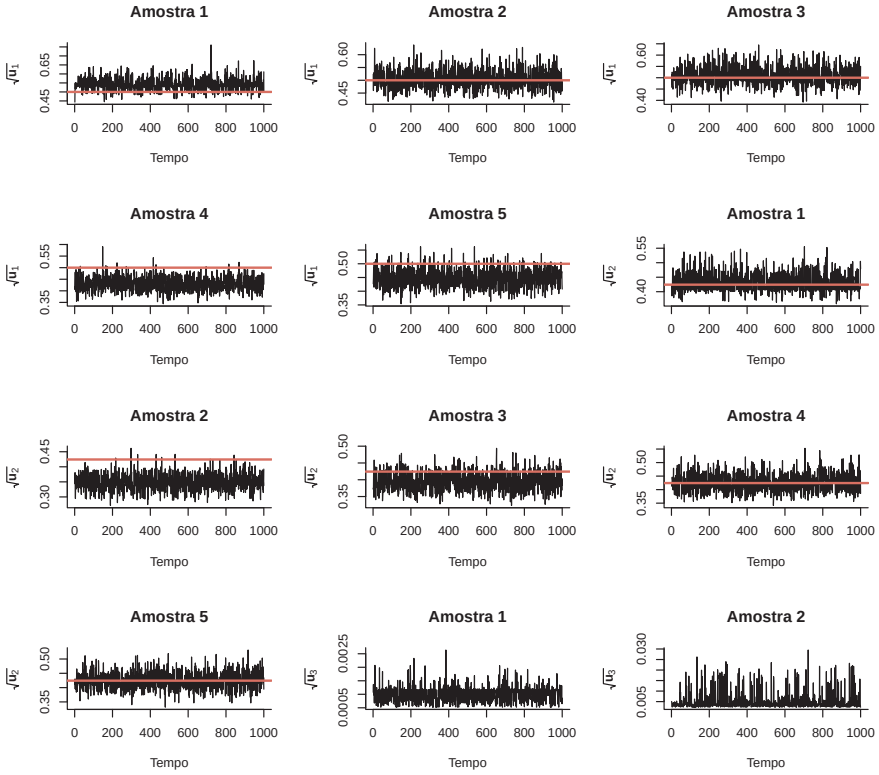
Figura B.2: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, ρ_1 e ρ_2 do modelo com 2 fatores para cada amostra, a linha vermelha representa o valor verdadeiro.

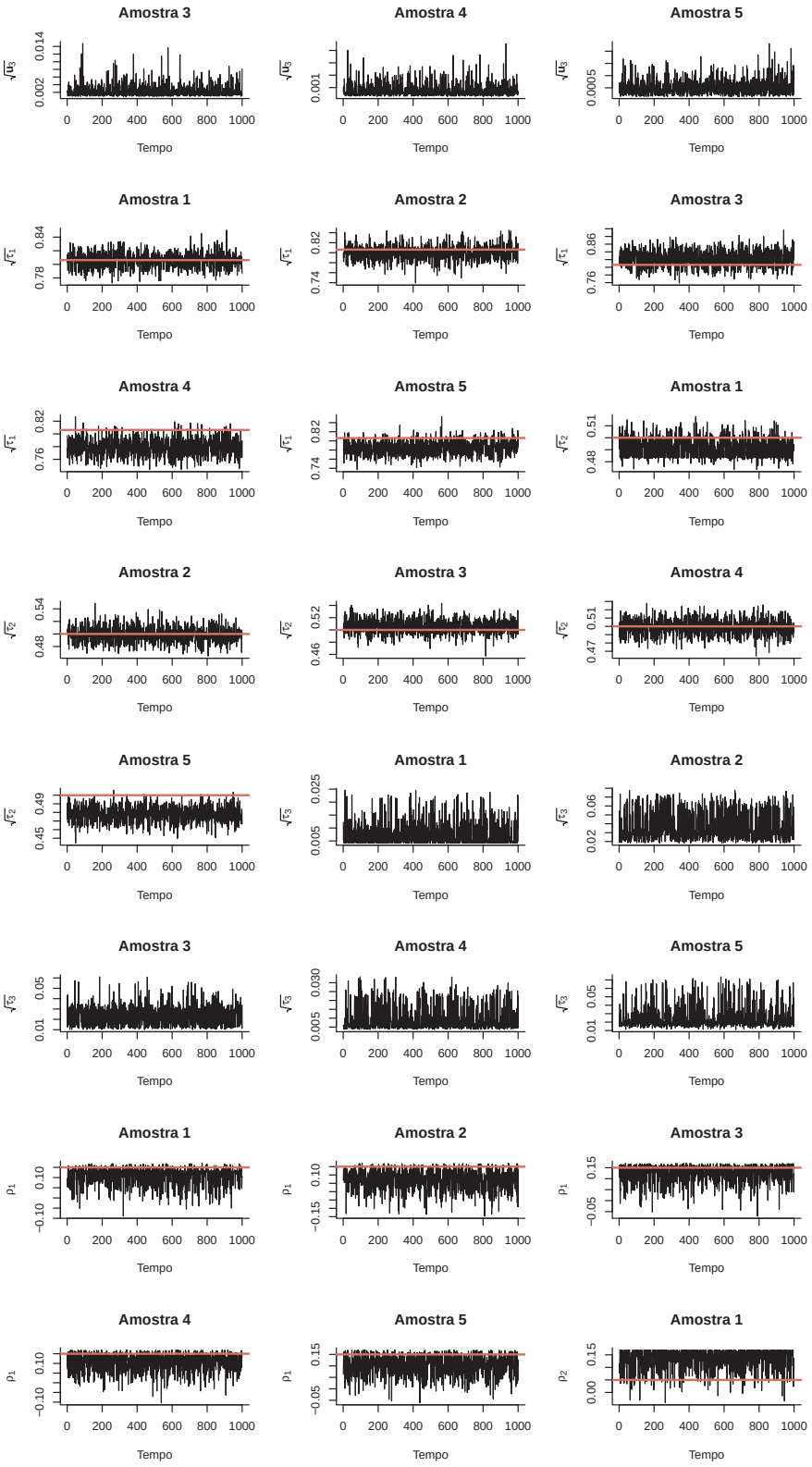


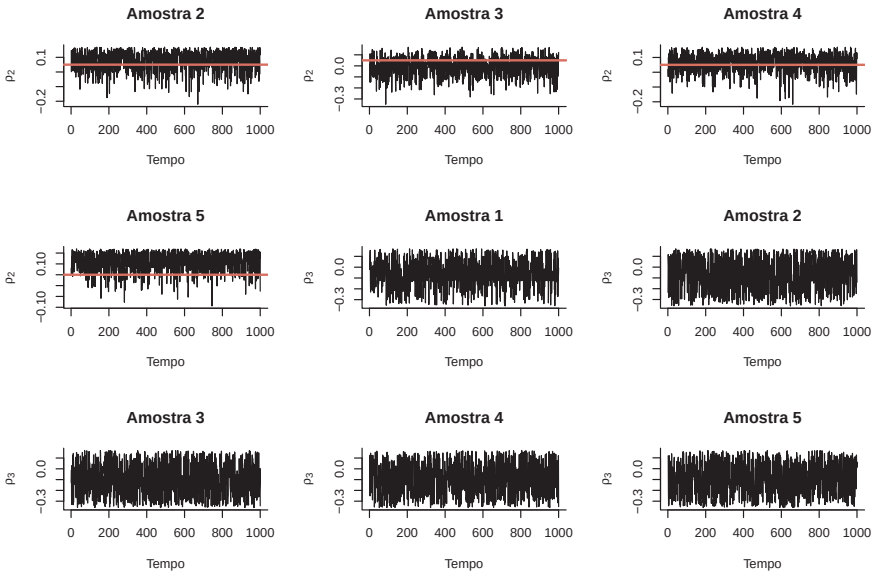

Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Tempo

B.3 Modelo com 3 fatores

Figura B.3: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{u_3}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, $\sqrt{\tau_3}$, ρ_1, ρ_2 e ρ_3 do modelo com 3 fatores para cada amostra, a linha vermelha representa o valor verdadeiro.




Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Tempo

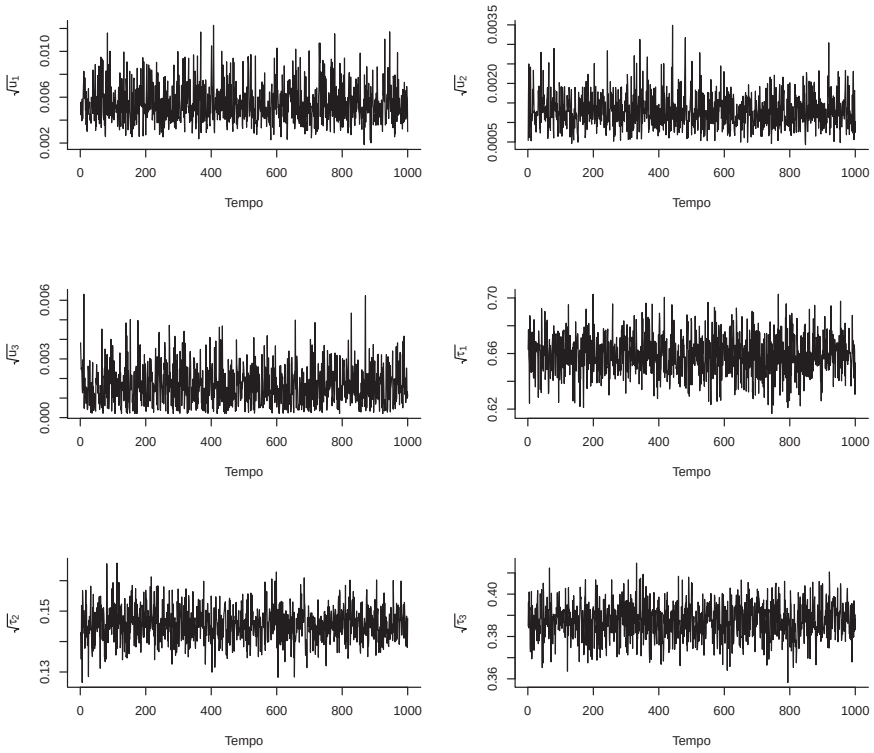

Amostra 2
Amostra 3
Amostra 4
Amostra 5
Amostra 1
Amostra 2
Amostra 3
Amostra 4
Amostra 5
ρ2
ρ3
Tempo
0
200
400
600
800
1000

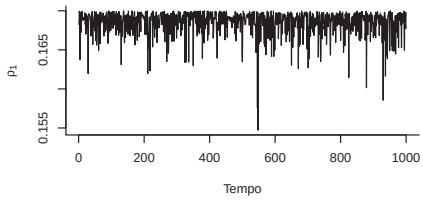
Apêndice C

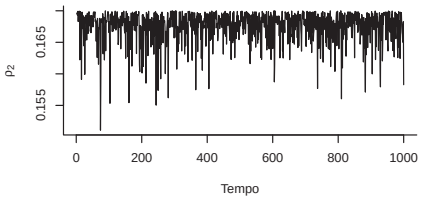
Aplicação

C.1 Modelo com 3 fatores

Figura C.1: Traço das cadeias *a posteriori* dos parâmetros $\sqrt{u_1}$, $\sqrt{u_2}$, $\sqrt{u_3}$, $\sqrt{\tau_1}$, $\sqrt{\tau_2}$, $\sqrt{\tau_3}$, ρ_1, ρ_2 e ρ_3 do modelo com 3 fatores da análise dos dados reais.




ρ1
0.155
0.165
0
200
400
600
800
1000
Tempo


ρ2
0.155
0.165
0
200
400
600
800
1000
Tempo

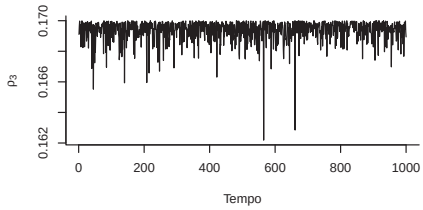
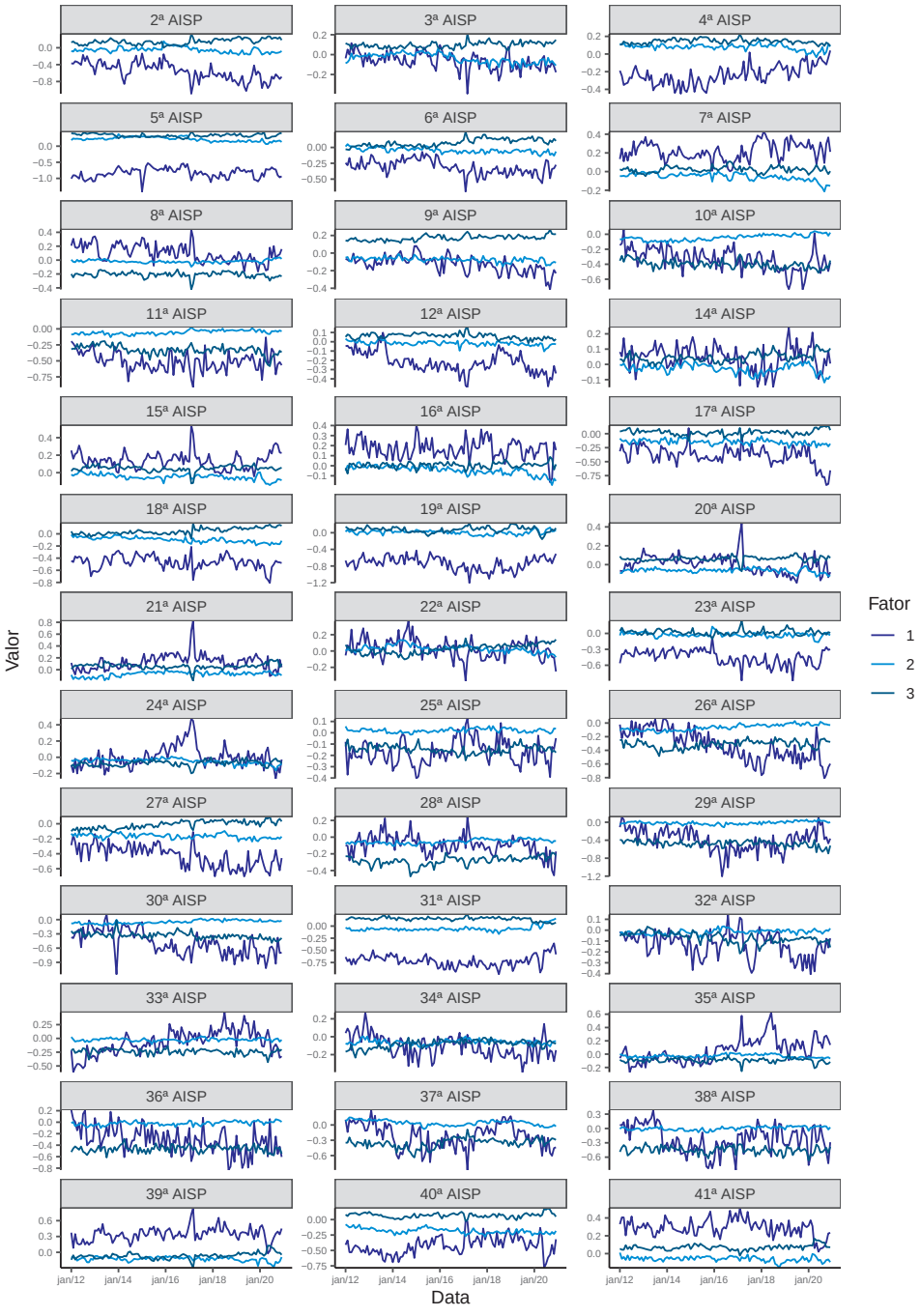

ρ3
0.162
0.166
0.170
0
200
400
600
800
1000
Tempo

Figura C.2: Séries mensais de cada um dos 3 fatores por AISP do modelo com 3 fatores da análise dos dados reais.



Apêndice D

Algoritmo *Stan*

A seguir é apresentado o código utilizado no *software* Stan para obter aproximações da distribuição *a posteriori* dos parâmetros para o modelo proposto considerando 3 fatores.

```
data{
  int<lower=0> N;
  int<lower=0> P;
  int<lower=0> L;
  int<lower=0> y[L,N,P];
  real<lower=0> log_E[L,N,P];
  matrix<lower = 0, upper = 1>[N, N] W;
  matrix<lower = 0, upper = 39>[N, N] D;
}

transformed data{
  vector[N] uns;
  uns = rep_vector(1,N);
}

parameters{
```

```
  vector[P-1] lambda1;
  vector[P-2] lambda2;
  vector[P-3] lambda3;
  matrix[L,N] f1;
  matrix[L,N] f2;
  matrix[L,N] f3;
  vector[L] phi1;
  vector[L] phi2;
  vector[L] phi3;
  real phi10;
  real phi20;
  real phi30;
  real<lower=0> ru1;
  real<lower=0> rtau1;
  real<lower=0> ru2;
  real<lower=0> rtau2;
  real<lower=0> ru3;
  real<lower=0> rtau3;
  real<lower=-0.36, upper=0.17> rho1;
  real<lower=-0.36, upper=0.17> rho2;
  real<lower=-0.36, upper=0.17> rho3;
}

transformed parameters{
  real theta[L,N,P];
  real<lower=0> tau1;
  real<lower=0> tau2;
  real<lower=0> tau3;
```

```
  for(t in 1:L){
    for(i in 1:N){
      theta[t,i,1] = log_E[t,i,1] + f1[t,i];
      theta[t,i,2] = log_E[t,i,2] + lambda1[1]*f1[t,i] + f2[t,i];
      theta[t,i,3] = log_E[t,i,3] + lambda1[2]*f1[t,i] + lambda2[1]*f2[t,
          i] + f3[t,i];
      for(j in 4:P){
        theta[t,i,j] = log_E[t,i,j] + lambda1[j-1]*f1[t,i] + lambda2[j
            -2]*f2[t,i] + lambda3[j-3]*f3[t,i];
      }
    }
  }

  tau1 = rtau1^2;
  tau2 = rtau2^2;
  tau3 = rtau3^2;
}

model{
  for(t in 1:L){
    for(i in 1:N){
      for(j in 1:P){
        y[t,i,j] ~ poisson_log(theta[t,i,j]);
      }
    }
  }

  for(t in 1:L){
    f1[t,] ~ multi_normal_prec(uns*phi1[t],1/tau1*(D - rho1 * W));
```

```
    f2[t,] ~ multi_normal_prec(uns*phi2[t],1/tau2*(D - rho2 * W));
    f3[t,] ~ multi_normal_prec(uns*phi3[t],1/tau3*(D - rho3 * W));
  }

  phi10 ~ normal(0,10);
  phi20 ~ normal(0,10);
  phi30 ~ normal(0,10);
  phi1[1] ~ normal(phi10,ru1);
  phi2[1] ~ normal(phi20,ru2);
  phi3[1] ~ normal(phi30,ru3);

  for(t in 2:L){
    phi1[t] ~ normal(phi1[t-1],ru1);
    phi2[t] ~ normal(phi2[t-1],ru2);
    phi3[t] ~ normal(phi3[t-1],ru3);
  }

  lambda1 ~ normal(0,10);
  lambda2 ~ normal(0,10);
  lambda3 ~ normal(0,10);
  ru1 ~ cauchy(0,10);
  ru2 ~ cauchy(0,10);
  ru3 ~ cauchy(0,10);
  rtau1 ~ cauchy(0,10);
  rtau2 ~ cauchy(0,10);
  rtau3 ~ cauchy(0,10);
  rho1 ~ uniform(-0.36, 0.17);
  rho2 ~ uniform(-0.36, 0.17);
  rho3 ~ uniform(-0.36, 0.17);
}
```

Referências Bibliográficas

Aguilar, O. e West, M. (2000) Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, **18**, 338–357.

de Almeida Inácio, M. H. (2017) Introdução ao stan como ferramenta de inferência bayesiana. URLhttps://marcoinacio.com/stan. Versão 1.4.3.

Alves, P. P., Lima, R. S. d., Marques, D., Silva, F. A. B. d., Lunelli, I. C., Rodrigues, R. I., Lins, G. d. O. A., Armstrong, K. C., Lira, P., Coelho, D. et al. (2021) Atlas da violência 2021.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **36**, 192–225.

Besag, J., York, J. e Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, **43**, 1–20.

Clemente, A., Clemente, L. T. e Clemente, A. K. (2022) Criminalidade nos municípios do estado do rio de janeiro: em busca da sua essência e da sua dinâmica. *Revista de Administração Pública*, **55**, 1392–1421.

Cressie, N. (1993) *Statistics for spatial data*. John Wiley e Sons.

Gamerman, D. e Lopes, H. F. (2006) *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

Gelfand, A. E. e Smith, A. F. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85**, 398–409.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. e Rubin, D. B. (2014) *Bayesian Data Analysis*. Chapman Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 3 edn.

Geman, S. e Geman, D. (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 721–741.

Geweke, J. e Zhou, G. (1996) Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, **9**, 557–587.

Hartigan, J. (1969) Linear bayesian methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, **31**, 446–454.

Hastings, W. K. (1970) *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press.

Hoffman, M. D., Gelman, A. et al. (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, **15**, 1593–1623.

ISP (2022) Instituto de segurança pública. URLhttps://www.ispvisualizacao.rj.gov.br:4434.

Johnson, R. A., Wichern, D. W. et al. (2002) *Applied multivariate statistical analysis*, vol. 5. Prentice hall Upper Saddle River, NJ.

Lopes, H. F., Salazar, E., Gamerman, D. et al. (2008) Spatial dynamic factor analysis. *Bayesian Analysis*, **3**, 759–792.

Lopes, H. F. e West, M. (2004) Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**, 41–68.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. e Teller, E. (1953) Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**, 1087–1092.

Migon, H. S., Gamerman, D. e Louzada, F. (2014) *Statistical inference: an integrated approach*. Chapman and Hall/CRC.

Miranda, A. P. M. d., Guedes, S. L., Borges, D., Beato, C., Souza, E. e Teixeira, P. (2006) A análise criminal e o planejamento operacional. *Rio de Janeiro: Instituto de Segurança Pública*, 15.

Muirhead, R. J. (1982) *Aspects of multivariate statistical theory*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Neal, R. M. et al. (2011) Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, **2**, 2.

Novellino, M. e Oliveira, L. (2019) Territórios-rede do crime organizado no rio de janeiro. *Anais XVIII ENANPUR*.

Pio, J. G., Brito, A. C. S. e Gomes, A. L. (2021) Criminalidade na cidade do rio de janeiro (rj) as influências das políticas públicas e as relações a curto e longo prazos. *Revista Brasileira de Ciências Sociais*, **36**.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/.

Sáfadi, T. e Peña, D. (2008) Bayesian analysis of dynamic factor models: an application to air pollution and mortality in são paulo, brazil. *Environmetrics: The official journal of the International Environmetrics Society*, **19**, 582–601.

dos Santos, P. S., Bezerra, É. C. D., de Freitas, C. A. e Becker, K. L. (2021) Criminalidade nos municípios do rio de janeiro: uma análise multivariada e espacial. *Revista de Economia*, **42**, 447–479.

Schmidt, A. M. e Nobre, W. S. (2014) Conditional autoregressive (car) model. *Wiley StatsRef: Statistics Reference Online*, 1–11.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. e Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, **64**, 583–639.

Stan Development Team (2020) RStan: the R interface to Stan. URLhttp://mc-stan.org/. R package version 2.21.2.

— (2022) *Stan Modeling Language Users Guide and Reference Manual, Version 2.31.* URLhttps://mc-stan.org.

Thorson, J., Scheuerell, M., Shelton, A., See, K., Skaug, H. e Kristensen, K. (2015) Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, **6**.

Tzala, E. e Best, N. (2008) Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research*, **17**, 97–118. PMID: 17855747.

Ventorim, F. C. (2020) Criminalidade e espaço urbano: uma análise das redes de relação entre tipos de crime, vítima e localização espacial no rio de janeiro.

Wang, F. e Wall, M. (2003) Generalized common spatial factor analysis. *Biostatistics (Oxford, England)*, **4**, 569–82.

Watanabe, S. e Opper, M. (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, **11**.

West, M. e Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer, 2nd edn.

West, M., Harrison, P. J. e Migon, H. S. (1985) Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, **80**, 73–83.