

Universidade Federal do Rio de Janeiro
Instituto de Matemática

**MODELO FATORIAL DINÂMICO COM CARGAS
ESPACIALMENTE ESTRUTURADAS**

Esther Salazar Gonzales

2008

Livros Grátis

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.



UFRJ

MODELO FATORIAL DINÂMICO COM CARGAS ESPACIALMENTE ESTRUTURADAS

Esther Salazar Gonzales

Tese de Doutorado submetida ao programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do grau de Doutora em Estatística.

Orientadores: **Dani Gamerman e Hedibert Freitas Lopes**

Rio de Janeiro, Fevereiro de 2008

MODELO FATORIAL DINÂMICO COM CARGAS ESPACIALMENTE ESTRUTURADAS

Esther Salazar Gonzales

Orientadores: Dani Gamerman e Hedibert Freitas Lopes

Tese de Doutorado submetida ao programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do grau de Doutora em Estatística.

Aprovada por:

Presidente Prof. Dani Gamerman
IM-UFRJ

Prof. Hedibert Freitas Lopes
GSB-University of Chicago

Prof^a. Alexandra M. Schmidt
IM-UFRJ

Prof^a. Flávia M. F. P. Landim
IM-UFRJ

Prof. Carlos M. Carvalho
GSB-University of Chicago

Prof. Peter Müller
University of Texas

Rio de Janeiro, Fevereiro de 2008

Salazar, Esther

Modelo fatorial dinâmico com cargas espacialmente estruturadas/ Esther Salazar. – Rio de Janeiro: UFRJ/IM, 2008.

xvii, 133 f. : il. ; 31cm.

Tese (Doutorado) – UFRJ/IM. Programa de Pós-Graduação em Estatística, 2008.

Orientadores: Dani Gamerman, Hedibert F. Lopes

Referências bibliográficas: p.134–140.

1. Estatística Matemática - Tese. I. Gamerman, Dani. II. Universidade Federal do Rio de Janeiro. Instituto de Matemática. III. Título.

Dedicatória

*Ao Juan Carlos, meu esposo,
e aos meus pais e irmã.*

Amo vocês demais!

Agradecimentos

Aproveito estas linhas para agradecer a todas aquelas pessoas que colaboraram de alguma forma, direta ou indiretamente, na realização deste trabalho. Muitas delas, mesmo desde longe, contribuíram de forma significativa para que eu pudesse concluir esta etapa importantíssima da minha vida.

Agradeço especialmente aos meus orientadores Dani Gamerman e Hedibert Freitas Lopes pela ajuda nestes anos de pesquisa. O apoio que eles me brindaram foi essencial no meu crescimento profissional e pessoal e por isso sempre serei grata a eles. Ao Juan Carlos, meu esposo, pelo carinho e paciência, e por ter alegrado meus dias, sempre. Agradeço profundamente aos amigos que, mesmo desde longe, estiveram sempre presentes. Às amigadas que surgiram durante estes últimos anos, por termos compartilhado momentos de alegria que sempre guardarei no meu coração. Aos professores do DME por me incentivarem a continuar neste difícil, porém estimulante, mundo da pesquisa. Agradeço também ao programa de Pós-Graduação do IM pela oportunidade e a FAPERJ pelo apoio financeiro.

Dedico este trabalho aos meus pais, Teresa e Julio e a minha irmã Maria, por serem sempre meus cúmplices nas minhas aventuras, e por me fazerem sentir a pessoa mais importante das suas vidas. Por tudo isso, muito obrigada!

Resumo

MODELO FATORIAL DINÂMICO COM CARGAS ESPACIALMENTE ESTRUTURADAS

Esther Salazar Gonzales

Orientadores: Dani Gamerman e Hedibert Freitas Lopes

Resumo da tese de Doutorado submetida ao programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do grau de Doutora em Estatística.

Nesta tese, uma nova classe de modelos espaço-temporais derivados do modelo fatorial dinâmico é proposta. A dependência temporal é modelada através dos fatores latentes enquanto que a dependência espacial é modelada através das cargas dos fatores. A evolução temporal dos fatores é descrita de diversas formas as quais permitem a inclusão de tendência e sazonalidade. A dependência espacial é incorporada nas cargas dos fatores combinando elementos determinísticos e estocásticos dando maior flexibilidade aos modelos e generalizando abordagens previas. Inferência Bayesiana é feita para tratar o problema de estimação dos parâmetros, incluindo a estimação do número de fatores para a qual um algoritmo Monte Carlo via cadeias de Markov com saltos reversíveis é proposto.

O modelo proposto é utilizado para o tratamento de dados espaciais contínuos Gaussianos e não Gaussianos. Numa primeira fase, o modelo é proposto para tratar dados Gaussianos considerando cargas estáticas. Numa segunda fase, uma extensão é feita ao considerar cargas variando no tempo. Finalmente o modelo é estendido para dados pertencentes à família exponencial considerando cargas estáticas na sua estrutura.

Estudos simulados são apresentados para testar a aplicabilidade dos modelos e algoritmos propostos. Adicionalmente, três aplicações são mostradas. Na primeira o modelo para dados Gaussianos é utilizado na modelagem de concentrações de SO_2 na região leste dos EUA. Nas duas últimas, modelos com resposta Bernoulli são especificados para modelar as ocorrências de chuva numa região da Oceania e no estado de Minas Gerais.

Palavras-chave: Inferência Bayesiana, modelos fatoriais, modelos dinâmicos, métodos MCMC, processo Gaussiano.

Abstract

DYNAMIC FACTOR MODEL WITH SPATIALLY STRUCTURED LOADINGS

Esther Salazar Gonzales

Orientadores: Dani Gamerman e Hedibert Freitas Lopes

Resumo da tese de Doutorado submetida ao programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do grau de Doutora em Estatística.

In this work, a new class of spatio-temporal models derived from standard dynamic factor models is proposed. The temporal dependence is modeled by latent factors while the spatial dependence is modeled by the factor loadings. The temporal evolution of the factors is described in a number of forms which allows inclusion of trend and seasonality. The spatial dependence is incorporated into the factor loadings by a combination of deterministic and stochastic elements, giving more flexibility to the models and generalizing previous approaches. Bayesian inference is done to account with the problem of parameter estimation, including the estimation of the number of factors where a reversible jump Markov Chain Monte Carlo algorithm is proposed.

The proposed model is used to deal with Gaussian and non Gaussian spatial continuous data. In a first stage, the model is proposed to deal with Gaussian data considering static loadings. In the second stage, an extension is done considering time-varying loadings. Finally, the model is extended to deal with data belonging to the exponential family considering static loadings into its framework.

Simulation studies are showed to test the applicability of proposed models and algorithms. Additionally, three applications are developed. In the first, a model for Gaussian data is used to modeling SO_2 concentrations in eastern US. In the last two, models with Bernoulli response are specified to modeling the occurrence of rainfall in Oceania and Minas Gerais state.

Key-words: Bayesian Inference, factor models, dynamic models, MCMC methods, Gaussian process.

Sumário

Lista de Tabelas	xiv
Lista de Figuras	xvii
1 Introdução	1
1.1 Análise Fatorial	2
1.2 Modelos dinâmicos espaço-temporais	3
1.3 Modelos lineares dinâmicos generalizados	4
1.4 Objetivos	5
1.5 Organização da tese	6
2 Modelo fatorial espacial dinâmico	7
2.1 Introdução	7
2.2 Modelo proposto	10
2.2.1 Efeito de covariáveis no nível médio e nas cargas	13
2.2.2 Inclusão de fatores sazonais	14
2.2.3 Função de verossimilhança	14
2.2.4 Não separabilidade	15
2.3 Interpolação e previsão	16
2.3.1 Interpolação	16
2.3.2 Previsão	17
2.4 Procedimento de inferência	18
2.4.1 Distribuições a priori	18

2.4.2	Inferência a posteriori	19
2.4.3	Inferência para o número de fatores	19
2.5	Estudo simulado	21
2.5.1	Exemplo 1: Correlação exponencial	22
2.5.2	Exemplo 2: Correlação Matérn	26
2.6	Considerações finais	28
3	Modelo fatorial espacial com cargas dinâmicas	31
3.1	Introdução	31
3.2	Modelo proposto	32
3.2.1	Não separabilidade	34
3.2.2	Previsão	34
3.3	Procedimento de inferência	35
3.3.1	Inferência a posteriori	36
3.3.2	Inferência para o número de parâmetros	38
3.4	Estudo Simulado	38
3.5	Considerações finais	41
4	Aplicação: Concentrações de SO₂ na região leste dos EUA	43
4.1	Introdução	43
4.2	Modelagem com cargas estáticas	46
4.2.1	Modelos utilizados	46
4.2.2	Principais resultados	50
4.3	Modelagem com cargas dinâmicas	60
4.3.1	Modelos utilizados	60
4.3.2	Distribuições a priori	61
4.3.3	Principais resultados	61
4.4	Comparação entre as duas abordagens	64
4.5	Considerações finais	66
4.6	Apêndice	68

5	Modelo fatorial espacial dinâmico generalizado	70
5.1	Introdução	70
5.2	Modelo proposto	73
5.2.1	Função de verossimilhança	74
5.2.2	Interpolação	74
5.2.3	Previsão	75
5.3	Procedimento de inferência	76
5.3.1	Distribuições a priori	76
5.3.2	Inferência a posteriori	76
5.3.3	Esquema de amostragem proposto	77
5.3.4	Inferência para o número de fatores	84
5.4	Estudo simulado	85
5.4.1	Exemplo 1: Dados com distribuição Gama	85
5.4.2	Exemplo 2: Dados com distribuição de Bernoulli	92
5.5	Considerações finais	97
6	Aplicações	99
6.1	Eventos de chuva na região norte da Oceania	100
6.1.1	Análise descritiva dos dados	102
6.1.2	Modelos utilizados	104
6.1.3	Principais resultados	107
6.2	Eventos de chuva no estado de Minas Gerais	111
6.2.1	Análise descritiva dos dados	113
6.2.2	Modelos utilizados	116
6.2.3	Principiais resultados	117
6.3	Considerações finais	119
7	Considerações finais e extensões	124
7.1	Considerações finais	124
7.2	Extensões	125

7.2.1	MFEDG com cargas dinâmicas	126
7.2.2	MFED para dados de área	127
7.2.3	MFED com cargas dinâmicas para dados de área	129
7.2.4	MFEDG para dados de área	129
7.2.5	MFEDG com cargas dinâmicas para dados de área	129
A	Distribuições condicionais completas	131
A.1	Modelo com cargas estáticas	131
A.2	Modelo com cargas dinâmicas	133
	Referências Bibliográficas	140

Lista de Tabelas

2.1	<i>Exemplo 1:</i> Critérios de comparação de modelos.	23
2.2	<i>Exemplo 1:</i> Sumário a posteriori dos parâmetros estimados. \hat{R} : diagnóstico de Gelman & Rubin.	25
2.3	<i>Exemplo 2:</i> Critérios de comparação de modelos.	27
2.4	<i>Exemplo 2:</i> Sumário a posteriori dos parâmetros estimados. \hat{R} : diagnóstico de Gelman & Rubin.	27
3.1	<i>Exemplo:</i> Critérios de comparação de modelos.	39
3.2	<i>Exemplo:</i> Sumário a posteriori dos parâmetros do modelo. \hat{R} : diagnóstico de Gelman & Rubin.	40
4.1	<i>Dados CASTNet:</i> Lista das estações monitoradoras selecionadas.	45
4.2	<i>Dados CASTNet:</i> Critérios de comparação de modelos.	51
4.3	<i>Dados CASTNet:</i> Avaliação da previsão utilizando as medidas: LogS, CRPS, MA90.	52
4.4	<i>Dados CASTNet:</i> Sumário a posteriori dos parâmetros que caracterizam a dinâmica dos fatores comuns na especificação MFEDS(4,1)-cov-GP.	54
4.5	<i>Dados CASTNet:</i> Sumário a posteriori dos parâmetros que caracterizam as colunas da matriz de cargas dos fatores na especificação MFEDS(4,1)-cov-GP.	55
4.6	<i>Dados CASTNet:</i> Critérios de comparação de modelos.	62
4.7	<i>Dados CASTNet:</i> Sumário a posteriori dos parâmetros na especificação MFEDcd(3)-cov.	63
4.8	<i>Dados CASTNet:</i> Critérios de comparação entre modelos com cargas estáticas e dinâmicas.	65

4.9	<i>Dados CASTNet: Avaliação da previsão utilizando as medidas: LogS, CRPS e MA90.</i>	66
5.1	<i>Dados Gama: Critérios de comparação de modelos.</i>	89
5.2	<i>Dados Gama: Sumário a posteriori dos parâmetros do modelo escolhido.</i>	91
5.3	<i>Dados Bernoulli: Critérios de comparação de modelos.</i>	94
5.4	<i>Dados Bernoulli: Sumário a posteriori dos parâmetros do modelo escolhido.</i> . .	97
6.1	<i>Chuva na Oceania: Lista das estações monitoradoras de chuva selecionadas.</i> .	101
6.2	<i>Chuva na Oceania: Medidas resumo (máximo, média e desvio padrão) da quantidade de chuva (em mm) das 17 estações monitoradoras analisadas.</i> . .	103
6.3	<i>Chuva na Oceania: Critérios de comparação de modelos.</i>	108
6.4	<i>Chuva na Oceania: Sumário a posteriori dos parâmetros do modelo MFEDG(3) que caracterizam os fatores comuns e as cargas espacialmente estruturadas.</i> .	109
6.5	<i>Chuva em Minas Gerais: Lista das 17 estações monitoradoras de chuva e medidas resumo (máximo, média e desvio padrão) da quantidade de chuva diária (em mm).</i>	113
6.6	<i>Chuva em Minas Gerais: Critérios de comparação de modelos.</i>	118
6.7	<i>Chuva em Minas Gerais: Sumário a posteriori dos parâmetros do modelo MFEDG-CC(2) que caracterizam a componente comum, os fatores e as cargas espacialmente estruturadas.</i>	119

Lista de Figuras

2.1	Exemplo da redução da dimensão do vetor de observações y_t a um número menor de fatores latentes.	9
2.2	<i>Dados simulados</i> : y_t segue um modelo fatorial espacial dinâmico com 3 fatores.	12
2.3	<i>Exemplo 1</i> : Dados gerados do MFED. Cada gráfico representa a série temporal nos locais s_1, \dots, s_6	22
2.4	<i>Exemplo 1</i> : Interpolação espacial das componentes $\beta_{(j)}$ para $j = 1, \dots, 4$	24
2.5	<i>Exemplo 1</i> : Fatores simulados vs. fatores estimados	25
2.6	<i>Exemplo 2</i> : Interpolação espacial das componentes $\beta_{(j)}$ para $j = 1, \dots, 3$	28
2.7	<i>Exemplo 2</i> : Fatores simulados vs. fatores estimados	29
3.1	<i>Exemplo</i> : Valores simulados e estimados da componente temporal	40
3.2	<i>Exemplo</i> : Fatores simulados vs. fatores estimados	40
3.3	<i>Exemplo</i> : Interpolação espacial das componentes $\beta_{(j)t}$ e $\beta_{(2)t}$ para $t = 5, 40, 80$	41
4.1	<i>Dados CASTNet</i> : Localização das estações monitoradoras.	44
4.2	<i>Dados CASTNet</i> : Séries temporais semanais do $\log(SO_2)$	46
4.3	<i>Dados CASTNet</i> : Periodograma ajustado de algumas séries temporais.	48
4.4	<i>Dados CASTNet</i> : Diagrama da acurácia das previsões semanais do $\log(SO_2)$ para as especificações MFEDS(2,1)-Exp, MFEDS(2,1)-Matérn, MFED(4)-cov, MFED(4)-cov-GP, MFEDS(4,1)-cov-GP, MGETP e MGFP(4).	53
4.5	<i>Dados CASTNet</i> : Médias a posteriori dos fatores para a especificação MFEDS(4,1)-cov-GP.	56
4.6	<i>Dados CASTNet</i> : Interpolação Bayesiana das cargas dos fatores.	57

4.7	<i>Dados CASTNet</i> : Valores interpolados nas estações SPD e BWR deixadas de fora da análise utilizando a especificação MFEDS(4,1)-cov-GP.	58
4.8	<i>Dados CASTNet</i> : Valores previstos no período 2004:1–2004:30.	58
4.9	<i>Dados CASTNet</i> : Mapas das concentrações de SO ₂ utilizando a especificação MFEDS(4,1)-cov-GP.	59
4.10	<i>Dados CASTNet</i> : Médias a posteriori dos coeficientes do nível médio para a especificação MFEDcd(3)-cov.	63
4.11	<i>Dados CASTNet</i> : Médias a posteriori dos fatores para a especificação MFEDcd(3)-cov.	64
4.12	<i>Dados CASTNet</i> : Médias a posteriori da componente temporal de β_t para a especificação MFEDcd(3)-cov.	65
4.13	<i>Dados CASTNet</i> : Interpolação Bayesiana da componente espaço-temporal de β_t nas semanas 2002-1, 2003-1, 2003-26 e 2003-52.	67
4.14	<i>Dados CASTNet</i> : Trajetória das cadeias dos parâmetros γ_j e λ_j ($j = 1, \dots, 4$).	68
4.15	<i>Dados CASTNet</i> : Trajetória das cadeias dos parâmetros τ_j^2 e ϕ_j ($j = 1, \dots, 5$).	69
5.1	<i>Dados Gama</i> : Dados gerados do modelo Gama.	87
5.2	<i>Dados Gama</i> : Interpolação espacial das componentes β_1 e β_2	90
5.3	<i>Dados Gama</i> : Fatores simulados vs. fatores estimados	90
5.4	<i>Dados Gama</i> : Gráficos de autocorrelação de f_t , $t = 10, 35, 60, 85$	91
5.5	<i>Dados Bernoulli</i> : Dados gerados do modelo Bernoulli.	94
5.6	<i>Dados Bernoulli</i> : Interpolação espacial das componentes β_1 e β_2	95
5.7	<i>Dados Bernoulli</i> : Valores simulados e estimados da componente comum e dos fatores.	96
5.8	<i>Dados Bernoulli</i> : Gráficos de autocorrelação de f_t , $t = 10, 35, 60, 85$	96
6.1	<i>Chuva na Oceania</i> : Localização das estações monitoradoras.	101
6.2	<i>Chuva na Oceania</i> : Valores diários da quantidade de chuva e do evento no ano 2001.	102
6.3	<i>Chuva na Oceania</i> : <i>Box plots</i> da quantidade de chuva e do total de dias chuvosos por mês.	103

6.4	<i>Chuva na Oceania: Fatores estimados para a especificação MFEDG(3)</i>	110
6.5	<i>Chuva na Oceania: Interpolação Bayesiana nas cargas dos fatores.</i>	111
6.6	<i>Chuva na Oceania: Probabilidades ajustadas.</i>	111
6.7	<i>Chuva na Oceania: Interpolação espacial nas estações EFM14 e IM5.</i>	112
6.8	<i>Chuva em Minas Gerais: Localização das estações monitoradoras.</i>	113
6.9	<i>Chuva em Minas Gerais: Valores diários da quantidade de chuva e do evento no ano 2005.</i>	115
6.10	<i>Chuva em Minas Gerais: Gráfico do total de dias chuvosos por mês e por estação monitoradora.</i>	115
6.11	<i>Chuva em Minas Gerais: Componente comum e fatores estimados para a especificação MFEDG(2)-CC</i>	120
6.12	<i>Chuva em Minas Gerais: Interpolação Bayesiana nas cargas dos fatores.</i>	120
6.13	<i>Chuva em Minas Gerais: Probabilidades ajustadas.</i>	121
6.14	<i>Chuva em Minas Gerais: Interpolação espacial nas estações PI e CA.</i>	121
6.15	<i>Chuva em Minas Gerais: Mapas de probabilidade para alguns dias do ano 2005.</i>	122
7.1	Representação geral da classe de modelos proposta para diferentes tipos de dados espaço-temporais.	126

Capítulo 1

Introdução

Nesta tese é formulado um modelo fatorial espacial dinâmico para dados contínuos Gaussianos e não Gaussianos seguindo uma abordagem completamente Bayesiana. Neste tipo de modelo, para cada instante de tempo tem-se um vetor de observações continuamente distribuídas numa determinada área de estudo. Desta forma, para cada localização tem-se uma série temporal associada. A idéia é reduzir a dimensão do vetor de observações através de fatores latentes. Os fatores são ponderados por coeficientes chamados cargas do fator que medem a proporção com a qual o fator contribui à variância da observação. A variação espacial existente entre as observações é modelada através da matriz de cargas dos fatores. O comportamento dinâmico é modelado através dos fatores comuns que seguem processos auto-regressivos de ordem um. Devido a flexibilidade do modelo, diversas estruturas podem ser incorporadas nestas duas componentes tais como tendência e sazonalidade, tanto na média do processo quanto nas cargas espacialmente estruturadas.

Nesta tese, o modelo proposto será chamado de Modelo Fatorial Espacial Dinâmico (MFED). Inferência para este tipo de modelo será feita usando uma abordagem completamente Bayesiana. Amostras das distribuições a posteriori dos parâmetros serão obtidas utilizando métodos de Monte Carlo via cadeias de Markov (MCMC), em particular, o amostrador de Gibbs e o algoritmo de Metropolis-Hastings. Para o caso não Gaussiano, um algoritmo que permite amostrar os fatores latentes em bloco é proposto. Em geral, esta proposta é sugerida como uma alternativa para estimar o vetor de estados em modelos dinâmicos espaço-temporais.

Numa primeira abordagem, a matriz de cargas dos fatores é invariante no tempo (estática), neste caso as colunas da matriz de cargas são espacialmente estruturadas seguindo processos Gaussianos (PG). Aqui, inferência no número de fatores é introduzida via o algoritmo Monte Carlo via cadeias de Markov com saltos reversíveis (abreviação do inglês RJMCMC)

(ver [Green, 1995](#), para mais detalhes). Numa segunda abordagem, a matriz de cargas dos fatores tem um comportamento dinâmico. Para este modelo são propostos algoritmos para fazer inferência de todas as quantidades desconhecidas, considerando o número de fatores conhecido. Finalmente, o MFED é estendido para dados pertencentes à família exponencial, em que um esquema de amostragem para os fatores latentes é proposto como alternativa para amostrar os estados em modelos dinâmicos não lineares.

Em geral, três tipos de modelos são utilizados e generalizados ao longo da tese: *i*) a análise fatorial para dados espaciais, *ii*) os modelos dinâmicos espaço-temporais e *iii*) os modelos lineares dinâmicos generalizados. A seguir, estes três temas são apresentados de forma resumida.

1.1 Análise Fatorial

A análise fatorial é uma ferramenta estatística bastante utilizada na modelagem de dados multivariados, na qual o principal propósito é a redução de uma série de variáveis a um conjunto menor chamado de fatores latentes. Este tipo de modelagem, utilizado inicialmente para tratar dados relacionados ao comportamento humano, tem sido, nos últimos anos, amplamente desenvolvido e expandida a outras áreas da ciência. Isto, devido basicamente ao aparecimento de ferramentas computacionais adequadas e acessíveis. No contexto Bayesiano, a aparição de técnicas como o método MCMC, representou um avanço importante na implementação de algoritmos de inferência para modelos com muitos parâmetros, incluindo o modelo fatorial. Algumas abordagens nessa linha podem ser encontradas em [Polasek \(1997\)](#), [Arminger & Muthén \(1998\)](#) e [Press & Shigemasa \(1997\)](#). Adicionalmente, foram feitas extensões utilizando fatores dinâmicos na modelagem de séries temporais financeiras como podem ser vistas em [Aguilar \(1998\)](#), [Pitt & Shephard \(1999\)](#), [Aguilar & West \(2000\)](#) e [Lopes & West \(2004\)](#).

O modelo fatorial básico é definido da seguinte forma. Seja y_t um vetor de dimensão $N \times 1$, $t = 1, \dots, T$. Para algum inteiro positivo $m \leq N$, um modelo m -fatorial relaciona cada y_t a um vetor m -dimensional de variáveis aleatórias f_t , chamadas de fatores comuns, com a seguinte equação

$$y_t = \beta f_t + \epsilon_t$$

onde β é a matriz de cargas do fator de dimensão $N \times m$, os fatores f_t m -dimensionais são

independentes com $f_t \sim N(0, I_m)$, ϵ_t são vetores normais independentes tal que $\epsilon_t \sim N(0, \Sigma)$, $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$, e ϵ_t e $f_{t'}$ são independentes para todo t e t' .

Mais especificamente, cada y_{it} é expressado pela combinação linear dos fatores comuns, tal que $y_{it} = \beta_{i1}f_{1t} + \beta_{i2}f_{2t} + \dots + \beta_{im}f_{mt} + \epsilon_{it}$. Sob este modelo, a estrutura de covariância dos dados é dada por $V(y_t|\beta, \Sigma) = \beta\beta' + \Sigma$, com

$$\begin{aligned} \text{var}(y_{it}|\beta, \Sigma) &= \sum_{j=1}^m \beta_{ij}^2 + \sigma_i^2, \quad \forall i, \\ \text{cov}(y_{it}, y_{kt}|\beta, \Sigma) &= \sum_{j=1}^m \beta_{ij}\beta_{kj} \quad \forall i, k, i \neq k, \end{aligned}$$

ou seja, a variância de y_{it} é dividida em duas partes: *i*) variância explicada pelos fatores comuns (comunalidade) e *ii*) variância única ou específica.

Dados variando no espaço e no tempo, têm sido também incorporados na modelagem fatorial. Na literatura existem muitas aplicações referidas principalmente à análise de dados espaciais multivariados, nas quais o objetivo é encontrar estruturas latentes não observadas (fatores) que preservam a correlação espacial dos dados. Por exemplo, na análise de dados puramente espaciais, [Christensen & Amemiya \(2002, 2003\)](#) propuseram uma análise de variáveis latentes para modelar estruturas subjacentes (fatores) em dados multivariados georeferenciados. Em [Wang & Wall \(2001\)](#) a idéia é estimar um único fator que explique a correlação espacial dos dados. Já em [Wang & Wall \(2003\)](#) este modelo é estendido, permitindo que o fator comum seja estimado a partir da média dos dados com uma função de ligação apropriada. Este modelo é chamado de modelo fatorial espacial generalizado.

1.2 Modelos dinâmicos espaço-temporais

Os modelos lineares dinâmicos (MLD) ([West & Harrison, 1997](#)) se caracterizam por representar variáveis com evolução temporal. Nos modelos dinâmicos lineares uma estrutura de evolução temporal é facilmente incorporada na estrutura dos dados. Este modelo é caracterizado por um par de equações, chamadas de observação e de evolução, representadas pelo seguinte sistema:

$$y_t = F_t' \theta_t + \epsilon_t, \quad \epsilon_t \sim N(0, V_t) \quad (1.1)$$

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim N(0, W_t) \quad (1.2)$$

tal que y_t é uma seqüência de observações ao longo do tempo, $t = 1, \dots, T$ onde T é o último tempo observado da série y_t , F_t é uma matriz de valores conhecidos, θ_t é chamado de vetor de estados do modelo e G_t é uma matriz que descreve a evolução temporal dos parâmetros. As matrizes V_t e W_t representam as matrizes de covariância do erro de observação e de evolução respectivamente. Inferência nesta classe de modelos pode ser feita de maneira seqüencial seguindo o paradigma Bayesiano. Se $\{F, G, V, W\}_t$ são conhecidos pode ser utilizado o filtro de Kalman que fornece a distribuição condicional de θ_t dada a informação disponível até o tempo t .

Este tipo de estrutura pode ser generalizado ao considerar as matrizes $\{F, G, V, W\}_t$ desconhecidas e incorporando inúmeras estruturas nas equações de observação e de evolução para representar fenômenos mais complexos. A extensão para o caso espaço-temporal é facilitada ao considerar y_t como o vetor de observações que contém medidas em N locais (dados contínuos) ou N regiões (dados de área).

1.3 Modelos lineares dinâmicos generalizados

Os MLD podem ser estendidos para definir a classe dos modelos lineares dinâmicos generalizados (MLDG) introduzida por [West, Harrison & Migon \(1985\)](#), na qual é permitido que as observações sigam qualquer distribuição pertencente à família exponencial. A equação de observação (1.1) é substituída por

$$p(y_t|\eta_t) \propto \exp\{\psi[y_t'\eta_t - b(\eta_t)] + c(y_t, \psi)\} \quad (1.3)$$

$$g(\mu_t) = F_t'\theta_t \quad (1.4)$$

onde η_t é o parâmetro natural ou canônico, ψ é chamado de parâmetro de dispersão, a função $b(\eta_t)$ é convexa e duas vezes diferenciável e c é uma função conveniente que depende de y_t e ψ . A média e variância dependem de b da seguinte forma $E(y_t|\eta_t) = \mu_t = b'(\eta_t) = \tau(\eta_t)$ e $V(y_t|\eta_t) = \Sigma_t = b''(\eta_t)/\psi$. Desde que b é convexa, b' é monótona crescente. Isto implica que τ é injetiva com inversa τ^{-1} e, desta forma, temos que a variância pode ser expressa como uma função da média.

Esta nova classe é uma extensão natural dos modelos lineares generalizados (MLG) e permite uma dependência temporal nos parâmetros do preditor linear. Os MLDG têm sido amplamente estudados na literatura como por exemplo [West, Harrison & Migon \(1985\)](#),

Kitagawa (1987), Gamerman (1998) e Geweke & Tanizaki (2001) para o caso univariado e; Fahrmeir (1992), Fahrmeir & Tutz (1994) e Knorr-Held (1999) para o caso multivariado.

O MLDG é especificado da seguinte forma:

- (i) Componente aleatória: y_t .
- (ii) Componente sistemática: F_t e θ_t influenciam a distribuição de y_t através do preditor linear $\varphi_t = F_t'\theta_t$.
- (iii) Ligação entre as componentes aleatória e sistemática: a ligação entre a média μ_t e φ_t é especificada através da função resposta h tal que $\mu_t = h(\varphi_t)$. A inversa da função resposta é a função de ligação g onde $g(\mu_t) = \varphi_t$.

Aqui, a função que transforma o preditor linear no parâmetro natural é dada por $\nu(\varphi_t) = \eta_t$. Com $u = \nu^{-1}$ temos que $\varphi_t = u(\eta_t)$. Se u é a identidade podemos dizer que g é uma função de ligação canônica e, portanto, $\eta_t = \varphi_t$.

O modelo é completado com a equação de evolução em (1.2). O procedimento de inferência para o MLDG não é trivial. A principal dificuldade está na estimação dos estados, sendo necessárias técnicas como, por exemplo Linear Bayes, análise utilizando prioris conjugadas (West, Harrison & Migon, 1985), passos de Metropolis-Hastings com distribuições propostas eficientes (Gamerman, 1998), entre outras.

1.4 Objetivos

Os objetivos gerais desta tese são: *i)* propor uma nova classe de modelos espaço-temporais para dados Gaussianos e não Gaussianos baseada na análise fatorial, *ii)* propor um algoritmo de estimação eficiente para os modelos propostos, *iii)* aplicar esta nova classe a dados reais que apresentem estruturas e dinâmicas complexas.

Os objetivos específicos da tese são: *i)* estudar diversas especificações do modelo proposto tais como fatores sazonais, tendências, inclusão de covariáveis (estáticas e dinâmicas) tanto na média das observações quanto na média do processo Gaussiano; *ii)* investigar e implementar técnicas de amostragem recursiva para o modelo com cargas dinâmicas; e *iii)* propor um algoritmo de amostragem em blocos para estimar os estados latentes em modelos dinâmicos não lineares.

1.5 Organização da tese

Esta tese está organizada da seguinte forma. No Capítulo 2, o modelo fatorial espacial dinâmico para dados Gaussianos é proposto. Um esquema MCMC com saltos reversíveis é especificado para fazer inferência completa do modelo, incorporando o número de fatores como parâmetro. No Capítulo 3, uma extensão do modelo para dados Gaussianos é proposta considerando as cargas dos fatores variando no tempo. Esta generalização está baseada na decomposição da coluna dos fatores em duas componentes, uma puramente temporal e a outra espaço-temporal. Um esquema MCMC é desenvolvido para este caso considerando técnicas eficientes de amostragem para reduzir o tempo computacional. No Capítulo 4, uma aplicação com dados reais é apresentada, utilizando dados de concentrações de SO_2 na região leste dos EUA. No Capítulo 5, o modelo fatorial espacial dinâmico é estendido permitindo o tratamento de dados pertencentes à família exponencial. Neste capítulo, uma alternativa para amostrar os estados latentes dos modelos dinâmicos não normais é sugerida. No Capítulo 6, duas aplicações são apresentadas para modelar ocorrências de chuva em estações situadas na Oceania e no estado de Minas Gerais. Por último, no Capítulo 7 são apresentadas as conclusões e possíveis extensões.

Capítulo 2

Modelo fatorial espacial dinâmico

Neste capítulo, uma nova classe de modelos espaço-temporais para dados Gaussianos é proposta. Esta classe, derivada do modelo fatorial dinâmico, modela a dependência temporal através dos fatores latentes, enquanto que a dependência espacial é modelada através das cargas dos fatores. Esta especificação permite a inclusão de diversas estruturas temporais e espaciais, tais como tendência e sazonalidade na evolução dos fatores e elementos determinísticos e estocásticos nas cargas espacialmente estruturadas. Desta forma, o modelo proposto permite a incorporação de estruturas mais flexíveis e generaliza modelos já propostos na literatura. Inferência para esta classe de modelos é feita utilizando o paradigma Bayesiano considerando dois casos: (i) o número de fatores fixo e (ii) o número de fatores como um parâmetro do modelo. No segundo caso, inferência completamente Bayesiana é desenvolvida via o algoritmo RJMCMC (do inglês *Reversible Jump Markov Chain Monte Carlo*). Finalmente, exemplos com dados simulados são apresentados para mostrar a aplicabilidade do modelo e dos algoritmos propostos.

2.1 Introdução

A análise fatorial e a estatística espacial vêm recebendo especial atenção da comunidade científica nos últimos anos. Isto deve-se, em grande parte, ao aparecimento de esquemas computacionalmente eficientes que permitem a modelagem de estruturas mais complexas nestas áreas da estatística. No contexto Bayesiano, os métodos de simulação MCMC ([Gamerman & Lopes, 2006](#); [Robert & Casella, 2004](#)) tornaram acessível o tratamento completamente Bayesiano de modelos fatoriais e espaciais, como feito em [Lopes & West \(2004\)](#) e [Banerjee, Carlin & Gelfand \(2004\)](#), respectivamente.

A utilização destas duas ferramentas na modelagem de dados espaciais multivariados, já

foi explorada em abordagens anteriores, tanto para dados Gaussianos como não Gaussianos, em que mais de uma variável de interesse é medida em cada local ou região. Por exemplo, Wang & Wall (2001) introduziram um modelo fatorial espacial comum, o qual estende a análise fatorial tradicional permitindo estruturas espaciais nos fatores latentes. Em Wang & Wall (2003) foi ajustado um modelo fatorial espacial às taxas de mortalidade de três tipos de câncer, medidas em várias cidades de Minnesota, utilizando um modelo fatorial espacial comum generalizado. Christensen & Amemiya (2002, 2003) propuseram um método denominado *shift-factor analysis* para modelar dados espaciais multivariados, com variação temporal modelada através de componentes auto-regressivas. Por outro lado, Hogan & Tchernis (2004) ajustaram um modelo fatorial espacial considerando um único fator que admite diversas formas de dependência espacial. Em todas estas aplicações a análise fatorial é utilizada de forma usual e para dados espaciais multivariados, ou seja, os fatores comuns são responsáveis por reduzir potencialmente a dimensão total do vetor resposta em cada local.

Nesta tese, a análise fatorial é utilizada em outro contexto. Suponha que uma única variável de interesse é medida em N locais s_1, \dots, s_N em vários instantes de tempo $t = 1, \dots, T$. Assim, a variável resposta y_{it} ($i = 1, \dots, N$) medida no local i e no tempo t é univariada com $y_t = (y_{1t}, \dots, y_{Nt})'$ representando as observações medidas em todos os locais. Assim, a análise fatorial é utilizada para reduzir a dimensão de y_t a um conjunto menor de m fatores latentes $f_t = (f_{1t}, \dots, f_{mt})'$, com o propósito de identificar *clusters*/grupos de locais/regiões. Na Figura 2.1 esta idéia é representada de forma gráfica. O comportamento temporal é descrito através dos m **fatores latentes dinâmicos** f_t onde $m \ll N$. Um aspecto importante da modelagem proposta é que, informação a priori espacialmente estruturada com respeito aos grupos ou *clusters*, pode ser introduzida através das colunas da matriz de cargas dos fatores β_t (**cargas espacialmente estruturadas**). Pode ser mostrado que esta nova classe pertence à classe de modelos espaço-temporais não separáveis e não estacionários. Não separáveis porque em geral a estrutura de covariância não pode ser expressada em termos da soma ou produto de uma função de covariância puramente espacial e puramente temporal.

O modelo proposto se diferencia das abordagens anteriores nos seguintes dois aspectos: (i) dado um instante de tempo, as observações univariadas de todos os locais ($y_{it}, i = 1, \dots, N$) são agrupados no vetor y_t , representando o vetor de dados ou atributos da análise fatorial usual e (ii) dependência espacial é introduzida através das colunas da matriz de cargas dos fatores. Com esta especificação, os fatores dinâmicos podem ser interpretados como

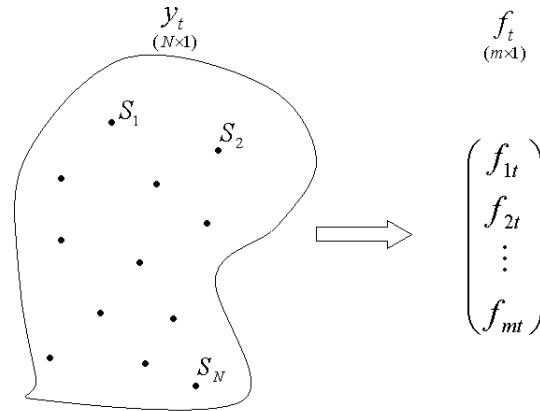


Figura 2.1: Exemplo da redução da dimensão do vetor de observações y_t a um número menor de fatores latentes.

medidas que descrevem similaridades entre as N séries temporais, tais como ciclos sazonais ou tendências (estacionárias ou não estacionárias). A influência de cada fator em descrever a observação medida em cada local é quantificada através das componentes da matriz de cargas do fator, as quais são modeladas através de processos Gaussianos (PG) ou campos aleatórios Gaussianos (CAG). Por outro lado, a capacidade do modelo de interpretar naturalmente a dinâmica dos fatores comuns e o comportamento das cargas é um aspecto muito importante a ser ressaltado. A interpretabilidade de ambas componentes tem papel fundamental na análise de dados espaço-temporais e em geral na análise de dados com estruturas complexas.

Outra característica relevante do modelo proposto é a capacidade de representar, de forma geral, diversos modelos existentes na literatura, os quais estão restritos a incluir, na maioria dos casos, ou fatores comuns ou matrizes de cargas não estocásticas. Mais especificamente, quando os fatores comuns são não estocásticos (por exemplo, modelos de regressão multivariada ou núcleos ponderados) e as cargas estocásticas têm prioris estruturadas hierárquicas, as seguintes duas classes caem na classe proposta nesta tese:

- (i) modelos com estrutura espacial nos coeficientes de regressão, como proposto em [Gaman, Moreira & Rue \(2003\)](#) e [Nobre, Schmidt & Lopes \(2005\)](#),
- (ii) modelos dinâmicos para dados espaço-temporais, como proposto em [Stroud, Müller & Sansó \(2001\)](#).

Adicionalmente, quando a matriz de cargas dos fatores é fixa, completamente determinística ou uma função que depende de um número pequeno de parâmetros, o modelo

proposto pode incorporar as seguintes estruturas:

- (i) β fixo utilizando funções ortogonais empíricas, como proposto em [Mardia, Goodall, Redfern & Alonso \(1998\)](#) (modelo chamado de *Kriged Kalman filter*),
- (ii) β construído a partir de bases ortonormais conhecidas, como em [Wikle & Cressie \(1999\)](#),
- (iii) β baseado em núcleos (*kernels*) suavizados como em [Sansó, Schmidt & Nobre \(2008\)](#) e [Calder \(2007\)](#).

Nesta tese, ambas formas, a determinística e a estocástica, são consideradas na especificação das cargas dos fatores, podendo incorporar facilmente informação como covariáveis através de funções de regressão. Por outro lado, nota-se que o número de fatores comuns tem um papel fundamental na redução da dimensionalidade, ou seja, as N séries temporais são parcimoniosamente representadas por um pequeno conjunto de m séries temporais e, portanto, um número “adequado” de fatores tem que ser considerado de acordo com o número de locais. Nesta tese, o número de fatores é tratado como um parâmetro adicional do modelo onde um algoritmo MCMC com saltos reversíveis é derivado. A proposta para estimar o número de fatores representa uma das principais contribuições da tese no contexto da modelagem espaço-temporal e da análise fatorial dinâmica.

Este capítulo está organizado da seguinte forma. Na Seção 2.2 é apresentado o modelo proposto. Na Seção 2.3, propostas para fazer interpolação espacial e previsão são apresentadas. Na Seção 2.4, um algoritmo MCMC é proposto para fazer inferência a posteriori de todos os parâmetros do modelo, incluindo o número de fatores. Na Seção 2.5, um estudo simulado é apresentado para testar os algoritmos de estimação propostos. Finalmente, na Seção 2.6, as considerações finais são apresentadas.

2.2 Modelo proposto

Seja N o número de locais numa determinada área S e $y_t = (y_{1t}, \dots, y_{Nt})$ o vetor de observações de dimensão N medido nos locais s_1, \dots, s_N ($s_i \in S, i = 1, \dots, N$) no tempo $t = 1, \dots, T$. O **modelo fatorial espacial dinâmico (MFED)** proposto é

$$y_t = \mu_t^{y^*} + \beta f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \quad (2.1a)$$

$$f_t = \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \quad (2.1b)$$

$$\beta_{(j)} \sim PG(\mu_j^{\beta^*}, \tau_j^2 \rho_{\phi_j}(\cdot)), \quad j = 1, \dots, m \quad (2.1c)$$

onde μ_t^{y*} representa o nível médio do processo, f_t é um vetor de dimensão $m \times 1$ chamado de fator comum, $m \leq N$ representa o número de fatores comuns tal que m é muito menor que N e $\beta = (\beta_{(1)}, \dots, \beta_{(m)})$ é a matriz de cargas dos fatores de dimensão $N \times m$. A matriz Γ caracteriza a evolução dinâmica dos fatores comuns, enquanto que Σ e Λ representam a variância de observação e de evolução, respectivamente. Por simplicidade é assumido que $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ e $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$. A evolução dinâmica dos fatores é caracterizada por $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_m)$, a qual pode ser facilmente estendida considerando valores fora da diagonal, como, por exemplo, na modelagem de fatores sazonais. Além disso, fatores estacionários ($|\gamma_i| < 1$) ou não estacionários ($|\gamma_i| = 1$) podem ser considerados.

Dependência temporal

Analogamente à análise fatorial, é assumido que os m fatores são condicionalmente independentes e capturam toda a estrutura de covariância temporal de y_t .

Dependência espacial

A dependência espacial é modelada através das colunas da matriz de cargas dos fatores β . Mais especificamente, cada coluna de β denotada por $\beta_{(j)} = (\beta_{(j)}(s_1), \dots, \beta_{(j)}(s_N))'$, para $j = 1, \dots, m$, é modelada através de um *processo Gaussiano condicionalmente independente*. O processo Gaussiano é especificado pelas funções $m(s_l)$ e $C(s_l, s_k)$ ($l, k \in \{1, \dots, N\}$) que correspondem, respectivamente, à média na localização s_l e a covariância do processo entre as localizações s_l e s_k (ver Billingsley, 1986, para mais detalhes) e, portanto, $\beta_{(j)}(\cdot) \sim PG(m(\cdot), C(\cdot, \cdot))$. Se $\mu_j^{\beta*} = (m(s_1), \dots, m(s_N))'$ é o vetor de médias de dimensão $N \times 1$ e $C(s_l, s_k) = \tau_j^2 \rho_{\phi_j}(|s_l - s_k|)$ é o elemento (l, k) da matriz R_{ϕ_j} , então

$$\beta_{(j)} \sim PG(\mu_j^{\beta*}, \tau_j^2 \rho_{\phi_j}(\cdot)) \equiv N(\mu_j^{\beta*}, \tau_j^2 R_{\phi_j})$$

onde R_{ϕ_j} representa a matriz de correlação espacial e $\rho_{\phi_j}(\cdot)$ define algum tipo de função de correlação espacial.

Os parâmetros ϕ_j 's são tipicamente escalares ou vetores de dimensão 2. Por exemplo, ϕ é univariado quando a função de correlação é exponencial $\rho_\phi(d) = \exp\{-d/\phi\}$ ou esférica $\rho_\phi(d) = (1 - 1.5(d/\phi) + 0.5(d/\phi)^3)1_{\{d/\phi \leq 1\}}$. Por outro lado, ϕ é bivariado ($\phi = (\phi_1, \phi_2)$) quando a função de correlação é exponencial $\rho_\phi(d) = \exp\{-(d/\phi_1)^{\phi_2}\}$ ou Matérn $\rho_\phi(d) = 2^{1-\phi_2} \Gamma(\phi_2)^{-1} (d/\phi_1)^{\phi_2} \mathcal{K}_{\phi_2}(d/\phi_1)$, onde $\mathcal{K}_{\phi_2}(\cdot)$ é a função de Bessel modificada de segundo tipo de ordem ϕ_2 . Em cada uma destas funções, d representa a distância euclidiana entre as localizações ($|s_l - s_k|$), o parâmetro de escala $\phi_1 > 0$ controla a taxa de decaimento da correlação quando a distância entre locais se incrementa, enquanto que o parâmetro de forma

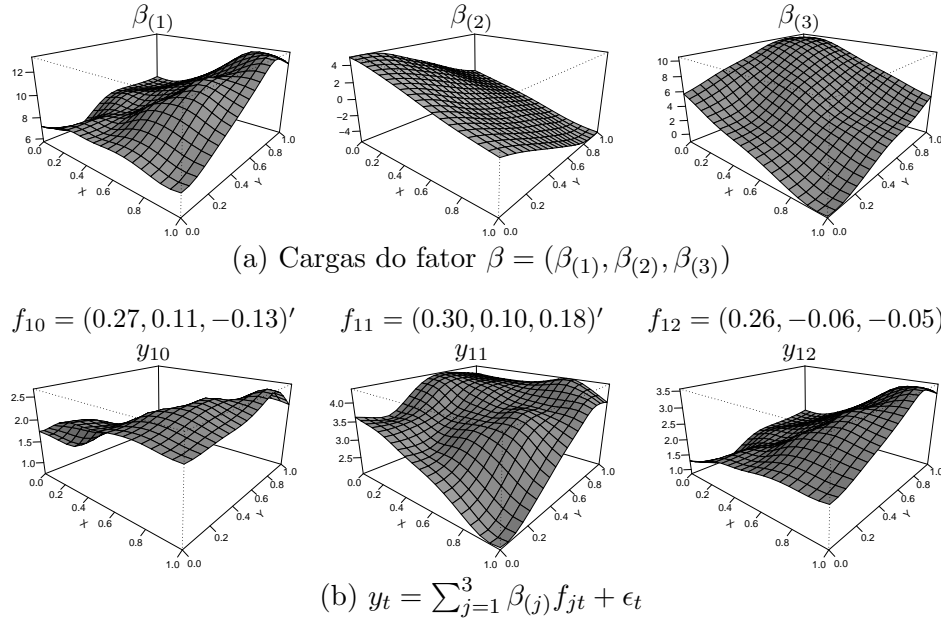


Figura 2.2: *Dados simulados:* y_t segue um modelo fatorial espacial dinâmico com 3 fatores. (a) Processos Gaussianos para as 3 colunas da matriz de cargas dos fatores. (b) Processo de y_t para $t = 10, 11, 12$.

(ou suavidade) ϕ_2 controla o grau de diferenciabilidade do processo latente (ver [Cressie, 1993](#); [Stein, 1999](#), para mais detalhes). Por outro lado, formulações não paramétricas poderiam ser incluídas no modelo, como por exemplo as introduzidas em [Gelfand, Kottas & MacEachern \(2005\)](#).

A Figura 2.2 ilustra a dinâmica do modelo através de dados simulados a partir de um MFED com três fatores. Estes dados foram gerados considerando como área de análise um quadrado $[0, 1] \times [0, 1]$. Nesta figura pode-se perceber o comportamento das superfícies de y_t ao longo de três instantes de tempo $t = 10, 11, 12$. Nota-se que a primeira coluna da matriz de cargas do fator, $\beta_{(1)}$, sugere que o primeiro fator é mais importante no nordeste que no sudoeste da região. Por outro lado, dada a superfície de $\beta_{(2)}$ e $\beta_{(3)}$, tem-se que o segundo e o terceiro fatores são mais importantes nas regiões sudoeste e noroeste, respectivamente. Nesta figura pode-se perceber claramente a interação das cargas e dos fatores ao longo do tempo, por exemplo nota-se que as superfícies de y_{12} e $\beta_{(1)}$ são bastante parecidas, isto devido à influência positiva do primeiro fator e a pouca ou quase nula influência do segundo e terceiro fatores.

Comparação com outros modelos

Modelos alternativos podem ser vistos como casos particulares do MFED. Entre eles podem ser mencionados os seguintes:

- [Mardia, Goodall, Redfern & Alonso \(1998\)](#) introduziram o método denominado *kriged Kalman filter*. Com esta abordagem as colunas de β são fixadas no início da análise baseadas em funções ortogonais empíricas.
- [Wikle & Cressie \(1999\)](#) considera a redução da dimensão do vetor de estados através de projeções de polinômios ortogonais empíricos. Neste caso, as colunas de β são determinísticas, baseadas em bases ortonormais conhecidas.
- [Calder \(2007\)](#) propôs um modelo de convolução fatorial dinâmico Bayesiano na qual os $\beta_{(j)}$'s são núcleos suavizados conhecidos.
- [Sansó, Schmidt & Nobre \(2008\)](#) consideraram uma classe de modelos baseado em processos de convolução independentes com núcleos discretos, na qual β representa a raiz quadrada da matriz de correlação.

2.2.1 Efeito de covariáveis no nível médio e nas cargas

Muitas especificações para o nível médio do processo espaço temporal e do processo Gaussiano podem ser incorporadas. As mais comuns são baseadas em covariáveis variando no tempo e covariáveis que dependem da localização (por exemplo a latitude e longitude do local). Para o nível médio do processo, μ_t^{y*} , (equação 2.1a) algumas alternativas são:

(i) nível médio constante: $\mu_t^{y*} = \mu^y, \forall t$ (em geral $\mu^y = 0$);

(ii) modelo de regressão: $\mu_t^{y*} = X_t^y \mu^y$, onde $X_t^y = (1_N, X_{1t}^y, \dots, X_{qt}^y)$ contém q covariáveis variando no tempo;

(iii) modelo com coeficientes dinâmicos: $\mu_t^{y*} = X_t^y \mu_t^y$, onde $\mu_t^y \sim N(\mu_{t-1}^y, W)$.

Analogamente, covariáveis podem ser incluídas na média das cargas do fator μ_j^β . Dado o comportamento estático de β , covariáveis variando no espaço podem ser consideradas unicamente. Nesta tese, são consideradas as seguintes especificações: (i) $\mu_j^{\beta*} = 0$; (ii) $\mu_j^{\beta*} = \mu_j^\beta 1_N$ e (iii) $\mu_j^{\beta*} = X_j^\beta \mu_j^\beta$, onde X_j^β é a matriz de covariáveis de dimensão $N \times p_j$. No último caso,

mais flexibilidade é incorporada na modelagem de β ao permitir diferentes covariáveis para cada processo Gaussiano. Adicionalmente, especificações determinísticas podem ser incorporadas como covariáveis tais como núcleos suavizados (Calder, 2007; Sansó et al., 2008) ou bases ortogonais (Stroud et al., 2001; Wikle & Cressie, 1999).

2.2.2 Inclusão de fatores sazonais

Padrões cíclicos ou periódicos estão presentes em muitas aplicações e podem ser naturalmente incluídos no MFED. Por exemplo, combinações lineares de funções trigonométricas (formas de Fourier) podem ser usadas para modelar sazonalidade (como em West & Harrison, 1997, Capítulo 8). Padrões sazonais podem ser incorporados de duas maneiras (i) através dos fatores dinâmicos ou (ii) através do nível médio do processo. Na primeira abordagem, os fatores sazonais comuns são ponderados pelas colunas da matriz de cargas associadas ao fator sazonal, sugerindo diferentes padrões sazonais em cada local. Na segunda abordagem, o mesmo padrão sazonal é assumido para todos os locais. Por exemplo, um fator sazonal de período p ($p = 12$ para dados mensais com ciclos anuais) pode ser incorporado considerando $\beta = (\beta_{(1)}, 0, \dots, \beta_{(1)}, 0)$ e $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_h)$ onde

$$\Gamma_l = \begin{pmatrix} \cos(2\pi l/p) & \sin(2\pi l/p) \\ -\sin(2\pi l/p) & \cos(2\pi l/p) \end{pmatrix}, \quad l = 1, \dots, h = p/2,$$

tal que $h = p/2$ é o número de harmônicos necessários para capturar o comportamento sazonal da série temporal (ver West & Harrison, 1997, Capítulo 8 para mais detalhes). Neste caso a matriz de covariância Λ não é mais diagonal dado que agora os fatores são correlacionados, ou seja, $\Lambda = \text{diag}(\Lambda_1, \dots, \Lambda_h)$, em que cada Λ_l ($l = 1, \dots, h$) é uma matriz de covariância de dimensão 2×2 . Na prática, poucos harmônicos são necessários para descrever adequadamente o padrão sazonal e, portanto, a dimensão da componente sazonal é pequena.

2.2.3 Função de verossimilhança

Sem perda de generalidade, e para facilitar a notação é assumido que $\mu_t^{y*} = 0$ e $\mu_j^{\beta*} = X_j^\beta \mu_j^\beta$. Portanto, dado f_t , para $t = 1, \dots, T$, o modelo (2.1a) pode ser reescrito seguindo uma notação matricial como $y = F\beta' + \epsilon$, onde $y = (y_1, \dots, y_T)'$ e $F = (f_1, \dots, f_T)'$ são matrizes de dimensão $T \times N$ e $T \times m$, respectivamente. A matriz de erro, ϵ , tem dimensão $T \times N$ e segue uma distribuição normal matriz-variada denotada por $\epsilon \sim N(0, I_T, \Sigma)$ (Dawid, 1981; Brown,

Vannucci & Fearn, 1998), assim, a função de verossimilhança de (Θ, F, β, m) é dada por

$$p(y|\Theta, F, \beta, m) = (2\pi)^{-TN/2} |\Sigma|^{-T/2} \text{etr} \left\{ -\frac{1}{2} \Sigma^{-1} (y - F\beta)' (y - F\beta) \right\}, \quad (2.2)$$

onde $\Theta = (\sigma, \lambda, \gamma, \mu, \tau, \phi)$ (Θ vetor de hiperparâmetros), $\sigma = (\sigma_1^2, \dots, \sigma_N^2)'$, $\lambda = (\lambda_1, \dots, \lambda_m)'$, $\gamma = (\gamma_1, \dots, \gamma_m)'$, $\mu = (\mu_1^\beta, \dots, \mu_m^\beta)$, $\tau = (\tau_1^2, \dots, \tau_m^2)'$, $\phi = (\phi_1, \dots, \phi_m)'$ e $\text{etr}(X) = \exp(\text{traço}(X))$ para alguma matriz quadrada X . Nota-se que a dependência do número de parâmetros m é feita de maneira explícita e posteriormente considerada como outro parâmetro do modelo na Seção 2.4.3.

2.2.4 Não separabilidade

As funções de covariância separáveis de processos espaço-temporais podem ser escritas como o produto ou soma de uma função de covariância puramente espacial e uma função de covariância puramente temporal. Mais especificamente, seja $Z(s, t)$ um processo aleatório indexado no espaço e no tempo. O processo é separável se

$$\text{Cov}(Z(s_1, t_1), Z(s_2, t_2)) = \begin{cases} \text{Cov}_s(u|\theta) \text{Cov}_t(h|\theta) & \text{ou} \\ \text{Cov}_s(u|\theta) + \text{Cov}_t(h|\theta) \end{cases}$$

onde $\theta \in \Theta \subset \mathbb{R}^p$, $u = \|s_2 - s_1\|$ e $h = |t_2 - t_1|$.

No MFED, quando $m = 1$ e $\mu_t^{y*} = 0$, a covariância entre locais i, j nos instantes de tempo $t, t + h$ é dada por

$$\begin{aligned} \text{Cov}(y_{it}, y_{j,t+h}) &= \text{E}((\beta_{i1} f_{1t}) (\beta_{j1} f_{1,t+h})) = \text{E}_\beta(\beta_{i1} \beta_{j1} \text{E}_f(f_{1t} f_{1,t+h})) \\ &= (\lambda \gamma^h) (1 - \gamma^2)^{-1} \text{E}_\beta(\beta_{i1} \beta_{j1}) \\ &= (\lambda \gamma^h) (1 - \gamma^2)^{-1} (\tau^2 \rho(u, \phi) + \mu_i^\beta \mu_j^\beta). \end{aligned}$$

Neste caso, as funções de covariância espacial e temporal são separáveis e identificáveis.

Em Cressie & Huang (1999) é enfatizado que estruturas separáveis são usualmente escolhidas mais por conveniência que pela sua capacidade de ajustar fenômenos sob estudo. De fato, estruturas de covariância separáveis são muito limitadas pela incapacidade de modelar interações espaço-tempo. Para contornar este problema, alguns tipos de estruturas não separáveis já foram propostas na literatura. Cressie & Huang (1999), por exemplo, introduziram uma classe de funções de covariância estacionárias não separáveis que permitem iterações

espaço-tempo baseadas em inversões da transformada de Fourier. Já [Gneiting \(2002\)](#) estendeu a classe anterior através de construções de funções avaliadas diretamente no espaço e no tempo.

No MFED, quando $m = 2$ e $\mu_t^{y^*} = 0$, a covariância entre locais i, j nos instantes de tempo $t, t + h$ é dada por

$$\begin{aligned} \text{Cov}(y_{it}, y_{j,t+h}) &= \text{E}((\beta_{i1}f_{1t} + \beta_{i2}f_{2t})(\beta_{j1}f_{1,t+h} + \beta_{j2}f_{2,t+h})) \\ &= \text{E}(\beta_{i1}\beta_{j1}f_{1t}f_{1,t+h}) + \text{E}(\beta_{i2}\beta_{j2}f_{2t}f_{2,t+h}) \\ &= (\lambda_1\gamma_1^h)(1 - \gamma_1^2)^{-1}(\tau_1^2\rho(u, \phi_1) + \mu_{i1}^\beta\mu_{j1}^\beta) \\ &\quad + (\lambda_2\gamma_2^h)(1 - \gamma_2^2)^{-1}(\tau_2^2\rho(u, \phi_2) + \mu_{i2}^\beta\mu_{j2}^\beta). \end{aligned}$$

Em geral, para $m > 1$, o resultado anterior pode ser generalizado e, portanto, pode ser mostrado que

$$\text{Cov}(y_{it}, y_{j,t+h}) = \sum_{k=1}^m (\lambda_k\gamma_k^h)(1 - \gamma_k^2)^{-1}(\tau_k^2\rho(u, \phi_k) + \mu_{ik}^\beta\mu_{jk}^\beta)$$

indicando claramente uma estrutura não separável. Portanto, uma importante propriedade do MFED é que o modelo permite formas não separáveis na função de covariância quando o número de fatores comuns é maior que 1.

2.3 Interpolação e previsão

2.3.1 Interpolação

Neste ponto, o interesse está em interpolar a resposta para N_n localizações onde a variável resposta y não tenha sido observada. Mais especificamente, seja y^o o vetor de observações medido nos locais $S = \{s_1, \dots, s_N\}$ e y^n o vetor (latente) de medidas (não observadas) nos locais $S_n = \{s_{N+1}, \dots, s_{N+N_n}\}$. Analogamente, sejam $\beta_{(j)}^o$ e $\beta_{(j)}^n$ a j -ésima coluna da matriz de cargas do fator correspondentes aos valores observados e não observados respectivamente.

Interpolação, neste contexto, consiste em encontrar a distribuição a posteriori de β^n (krigagem Bayesiana),

$$p(\beta^n|y^o) = \int p(\beta^n|\beta^o, \Theta)p(\beta^o, \Theta|y^o)d\beta^o d\Theta,$$

onde $p(\beta^n|\beta^o, \Theta) = \prod_{j=1}^m p(\beta_{(j)}^n|\beta_{(j)}^o, \mu_j^\beta, \tau_j^2, \phi_j)$ e $\Theta = (\sigma, \lambda, \gamma, \mu, \tau, \phi)$. Para $j = 1, \dots, m$, as distribuições $p(\beta_{(j)}^n|\beta_{(j)}^o, \mu_j^\beta, \tau_j^2, \phi_j)$ podem ser facilmente obtidas utilizando as propriedades

da normal multivariada. Logo, dado Θ ,

$$\begin{pmatrix} \beta_{(j)}^o \\ \beta_{(j)}^n \end{pmatrix} \sim N \left[\begin{pmatrix} X_j^{\beta^o} \\ X_j^{\beta^n} \end{pmatrix} \mu_j^\beta, \tau_j^2 \begin{pmatrix} R_{\phi_j}^o & R_{\phi_j}^{o,n} \\ R_{\phi_j}^{n,o} & R_{\phi_j}^n \end{pmatrix} \right]$$

onde $R_{\phi_j}^n$ é a matriz de correlação, de dimensão $N_n \times N_n$, entre os locais onde as observações não foram medidas, $R_{\phi_j}^{o,n}$ é uma matriz de dimensão $N \times N^n$ onde cada elemento representa a correlação entre o local medido i e o local não medido j , para $i = 1, \dots, N$ e $j = N + 1, \dots, N + N_n$. Portanto,

$$\beta_{(j)}^n | \beta_{(j)}^o, \Theta \sim N \left(X_j^{\beta^n} \mu_j^\beta + R_{\phi_j}^{n,o} R_{\phi_j}^o{}^{-1} (\beta_{(j)}^o - X_j^{\beta^o} \mu_j^\beta); \tau_j^2 (R_{\phi_j}^n - R_{\phi_j}^{n,o} R_{\phi_j}^o{}^{-1} R_{\phi_j}^{o,n}) \right)$$

e a aproximação Monte Carlo de $p(\beta^n | y^o)$ é dada por

$$\hat{p}(\beta^n | y^o) \approx \frac{1}{L} \sum_{l=1}^L p(\beta^n | \beta^{o(l)}, \Theta^{(l)})$$

onde $\{(\beta^{o(1)}, \Theta^{(1)}), \dots, (\beta^{o(L)}, \Theta^{(L)})\}$ é uma amostra de $p(\beta^o, \Theta | y)$ (ver Seção 2.4 para mais detalhes). Se $\beta^{n(l)}$ é amostrado de $p(\beta^n | \beta^{o(l)}, \Theta^{(l)})$, para $l = 1, \dots, M$, então $\{\beta^{n(l)}, \dots, \beta^{n(L)}\}$ é uma amostra de $p(\beta^n | y^o)$. Assim, o valor esperado das medidas não observadas y^n pode ser aproximado por

$$\hat{E}(y^n | y^o) \approx \frac{1}{L} \sum_{l=1}^L \beta^{n(l)} f^{(l)}.$$

2.3.2 Previsão

Fazer previsão nesta classe de modelos é teoricamente direto. Frequentemente surge o interesse em aprender sobre densidade preditiva, h passos a frente $p(y_{T+h} | y)$, ou seja

$$p(y_{T+h} | y) = \int p(y_{T+h} | f_{T+h}, \beta, \Theta) p(f_{T+h} | f_T, \beta, \Theta) p(f_T, \beta, \Theta | y) df_{T+h} df_T d\beta d\Theta,$$

onde $(y_{T+h} | f_{T+h}, \beta, \Theta) \sim N(\beta f_{T+h}, \Sigma)$, $(f_{T+h} | f_T, \beta, \Theta) \sim N(\mu_h, V_h)$, $\mu_h = \Gamma^h f_T$ e $V_h = \sum_{k=1}^h \Gamma^{k-1} \Lambda (\Gamma^{k-1})'$, para $h > 0$. Portanto, se $\{(\beta^{(1)}, \Theta^{(1)}, f_T^{(1)}), \dots, (\beta^{(L)}, \Theta^{(L)}, f_T^{(L)})\}$ é uma amostra de $p(f_T, \beta, \Theta | y)$ (ver Seção 2.4 para mais detalhes), é fácil amostrar $f_{T+h}^{(j)}$ de $p(f_{T+h} | f_T^{(j)}, \beta^{(j)}, \Theta^{(j)})$, para todo $j = 1, \dots, L$, tal que

$$\hat{p}(y_{T+h} | y) = \frac{1}{L} \sum_{j=1}^L p(y_{T+h} | f_{T+h}^{(j)}, \beta^{(j)}, \Theta^{(j)})$$

é a aproximação via Monte Carlo de $p(y_{T+h} | y)$. Analogamente, se $y_{T+h}^{(j)}$ é amostrado de $p(y_{T+h} | f_{T+h}^{(j)}, \beta^{(j)}, \Theta^{(j)})$, para $j = 1, \dots, L$, então $\{y_{T+h}^{(1)}, \dots, y_{T+h}^{(L)}\}$ representa uma amostra de $p(y_{T+h} | y)$.

2.4 Procedimento de inferência

Nesta seção um procedimento de inferência baseado no paradigma Bayesiano é proposto. Inicialmente distribuições a priori para todos os hiperparâmetros são apresentadas. Em seguida, dado o número de fatores, o algoritmo MCMC é utilizado para obter amostras a posteriori de todos os parâmetros do modelo. Por último, o número de fatores é tratado como um parâmetro desconhecido e uma análise completamente Bayesiana é desenvolvida através de um algoritmo MCMC com saltos reversíveis.

2.4.1 Distribuições a priori

Por simplicidade, distribuições a priori conjugadas são utilizadas para todos os parâmetros que definem a dinâmica dos fatores, enquanto que dois diferentes tipos de priori são consideradas para os parâmetros que definem os processos espaciais. A distribuição a priori para os fatores comuns são especificadas na equação (2.1b) e completada pela informação inicial f_0 tal que $f_0 \sim N(m_0, C_0)$ onde m_0 e C_0 são hiperparâmetros conhecidos. As distribuições a priori para os hiperparâmetros σ e λ são as seguintes: *i*) $\sigma_i^2 \sim GI(n_\sigma/2, n_\sigma s_\sigma/2)$, $i = 1, \dots, N$; e *ii*) $\lambda_j \sim GI(n_\lambda/2, n_\lambda s_\lambda/2)$, $j = 1, \dots, m$, onde $n_\sigma, s_\sigma, n_\lambda$ e s_λ são hiperparâmetros conhecidos.

Para γ , várias especificações a priori podem ser consideradas. Por exemplo, *i*) $\gamma_j \sim Ntr_{(-1,1)}(m_\gamma, S_\gamma)$, em que $Ntr_{c,d}(a, b)$ refere-se a uma distribuição normal $N(a, b)$ truncada nos valores do intervalo $[c, d]$; e *ii*) $\gamma_j \sim \alpha Ntr_{(-1,1)}(m_\gamma, S_\gamma) + (1 - \alpha)\delta_1(\gamma_j)$, onde m_γ, S_γ e $\alpha \in (0, 1]$ são hiperparâmetros conhecidos, $\delta_1(\gamma_j) = 1$ se $\gamma_j = 1$ e $\delta_1(\gamma_j) = 0$ se $\gamma_j \neq 1$. No primeiro caso, é assumido que todos os fatores dinâmicos são estacionários. No segundo caso, possíveis fatores não estacionários podem ser incorporados no modelo. Nota-se que quando $\alpha = 1$, o caso *i*) é contemplado. Ver [Peña & Poncela \(2004, 2006\)](#) para mais detalhes de modelos fatoriais dinâmicos não estacionários.

Dois tipos de priori são especificadas para os parâmetros μ_j^β , ϕ_j e τ_j^2 , $j = 1, \dots, m$: *i*) prioris próprias mas vagas *ii*) prioris de referência. No primeiro caso é assumido que, $\mu_j^\beta \sim N(m_\mu, S_\mu)$, $\phi_j \sim GI(2, b)$ e $\tau_j^2 \sim GI(n_\tau/2, n_\tau s_\tau/2)$, $j = 1, \dots, m$, onde m_μ, S_μ, n_τ e s_τ são hiperparâmetros conhecidos, $b = \rho_0/(-2 \ln(0.05))$ e $\rho_0 = \max_{i,j=1,\dots,N} |s_i - s_j|$ (ver [Banerjee, Carlin & Gelfand \(2004\)](#) e [Schmidt & Gelfand \(2003\)](#), para mais detalhes). Ou seja, $\pi(\mu_j^\beta, \tau_j^2, \phi_j) = \pi_N(\mu_j^\beta)\pi_{GI}(\tau_j^2, \phi_j)$ com

$$\pi_{GI}(\tau_j^2, \phi_j) = \pi_{GI}(\tau_j^2)\pi_{GI}(\phi_j) \propto \tau_j^{-(n_\tau+2)} e^{-0.5n_\tau s_\tau/\tau_j^2} \phi_j^{-3} e^{-b/\phi_j}, \quad (2.3)$$

onde N e GI denotam as densidades normal e gama inversa respectivamente. No segundo caso, uma análise de referência proposta por [Berger, Oliveira & Sansó \(2001\)](#) é considerada. Mais especificamente, $\pi_R(\mu_j^\beta, \tau_j^2, \phi_j) = \pi_R(\mu_j^\beta | \tau_j^2, \phi_j) \pi_R(\tau_j^2, \phi_j)$, com $\pi_R(\mu_j^\beta | \tau_j^2, \phi_j) = 1$ e

$$\pi_R(\tau_j^2, \phi_j) = \pi_R(\tau_j^2) \pi_R(\phi_j) \propto \tau_j^{-2} \left\{ \text{tr}(W_{\phi_j}^2) - \frac{1}{N - p_j} [\text{tr}(W_{\phi_j})]^2 \right\}^{1/2}, \quad (2.4)$$

com $W_{\phi_j} = ((\partial/\partial\phi_j)R_{\phi_j})R_{\phi_j}^{-1}(I_N - X_j^\beta(X_j^{\beta'}R_{\phi_j}^{-1}X_j^\beta)^{-1}X_j^{\beta'}R_{\phi_j}^{-1})$. Vale a pena ressaltar que $\pi_{GI}(\tau_j^2)$ e $\pi_R(\tau_j^2)$ são similares quando n_τ tende para 0. [Berger et al. \(2001\)](#) propõem e recomendam o uso de prioris de referência para os parâmetros da função de correlação. A justificativa é que a priori de referência garante distribuições a posteriori próprias comparadas com outras prioris não informativas. Neste ponto, vale a pena enfatizar que esta especificação define uma priori de referência condicionada à matriz de cargas dos fatores.

2.4.2 Inferência a posteriori

Condicional ao número de fatores m , a distribuição a posteriori conjunta de (F, β, Θ) é dada por

$$\begin{aligned} p(F, \beta, \Theta | y) &\propto \prod_{t=1}^T p(y_t | f_t, \beta, \sigma) p(f_0 | m_0, C_0) \prod_{t=1}^T p(f_t | f_{t-1}, \lambda, \gamma) \\ &\times \prod_{j=1}^m p(\beta_{(j)} | \mu_j^\beta, \tau_j^2, \phi_j) p(\gamma_j) p(\lambda_j) p(\mu_j^\beta) p(\tau_j^2, \phi_j) \prod_{i=1}^N p(\sigma_i^2) \end{aligned} \quad (2.5)$$

a qual é analiticamente intratável e, portanto, inferência a posteriori exata é feita utilizando o esquema MCMC. Neste caso, os fatores comuns são amostrados conjuntamente utilizando o algoritmo FFBS (do inglês *forward filtering backward sampling*) de [Carter & Kohn \(1994\)](#) e [Frühwirth-Schnatter \(1994\)](#). Para o resto de parâmetros, Θ e β , todas as distribuições condicionais completas são: distribuições normais (univariadas ou multivariadas) ou distribuições Gama inversa, exceto para o parâmetro ϕ que caracteriza a correlação espacial. Neste último caso a amostragem do parâmetro ϕ é feita com passos de Metropolis-Hastings. Todas as condicionais completas e densidades propostas estão detalhadas no Apêndice [A.1](#).

2.4.3 Inferência para o número de fatores

Inferência para o número de fatores latentes é obtida através do cálculo da probabilidade a posteriori do modelo (PPM), conhecido por ter um papel importante na seleção Bayesiana de modelos.

Lopes & West (2004) introduziram um esquema completamente Bayesiano ao tratar o número de fatores como parâmetro em modelos fatoriais dinâmicos normais. Isto é feito através de um esquema MCMC com saltos reversíveis, denotado por RJMCMC. O algoritmo proposto por eles é construído a partir de um conjunto preliminar de resultados do algoritmo MCMC rodado para um conjunto de valores de m pré-especificado. As cadeias resultantes produzem um conjunto de amostras a posteriori para $\Psi_m = (F_m, \beta_m, \Theta_m)$ que aproxima a distribuição a posteriori de $p(\Psi_m|m, y)$. Os primeiros momentos destas amostras são utilizados para construir distribuições propostas que são usadas para gerar valores candidatos a cada passo do algoritmo RJMCMC.

Nesta seção é adaptada e generalizada esta abordagem para o MFED proposto. Para o MFED a distribuição proposta conjunta é

$$\begin{aligned} q_m(\Psi_m) &= \prod_{j=1}^m f_N(f_{(j)}; M_{f_{(j)}}, aV_{f_{(j)}}) f_N(\beta_{(j)}; M_{\beta_{(j)}}, bV_{\beta_{(j)}}) f_N(\gamma_j; M_{\gamma_j}, cV_{\gamma_j}) \\ &\times \prod_{j=1}^m f_{GI}(\lambda_j; d, dM_{\lambda_j}) f_N(\mu_j; M_{\mu_j}, eV_{\mu_j}) f_{GI}(\phi_j; f, fM_{\phi_j}) \\ &\times \prod_{j=1}^m f_{GI}(\tau_j^2; g, gM_{\tau_j}) f_{GI}(\sigma_j^2; h, hM_{\sigma_j}), \end{aligned} \quad (2.6)$$

onde a, b, c, d, e, f, g e h são chamadas de constantes sintonizadoras e M_x e V_x são as médias e variâncias amostrais baseadas nos resultados preliminares do algoritmo MCMC. A escolha da constante sintonizadora depende da forma da distribuição a posteriori. Por exemplo, são recomendados valores menores que 1 para a, b, c e e utilizados nas distribuições propostas normais, e valores maiores que 1,5 para d, f, g e h utilizados nas distribuições propostas Gama inversa.

Considerando $p(y, m, \Psi_m) = p(y|m, \Psi_m)p(\Psi_m|m)Pr(m)$ e dando valores iniciais para m e Ψ_m , o algoritmo RJMCMC se comporta de forma similar ao algoritmo de Metropolis-Hastings, ou seja, um valor candidato m' é amostrado da distribuição proposta $q(m, m')$ e então, dado $m', \Psi_{m'}$ é amostrado de $q_{m'}(\Psi_{m'})$. O par $(m', \Psi_{m'})$ é aceito com probabilidade

$$\alpha = \min \left\{ 1, \frac{p(y, m', \Psi_{m'})}{p(y, m, \Psi_m)} \frac{q_m(\Psi_m)q(m', m)}{q_{m'}(\Psi_{m'})q(m, m')} \right\}. \quad (2.7)$$

A distribuição proposta $q(m, m')$ pode tomar inúmeras formas, a mais simples é considerar que $q(m, m')$ tem distribuição uniforme discreta com probabilidade $1/M$, em que M é o número máximo de fatores. Mais especificamente e com fins ilustrativos, o algoritmo é

resumido nos seguintes passos:

Algoritmo 2.1. (*Algoritmo RJMCMC para o MFED*)

1. Escolher valores iniciais para m e Ψ_m . Uma escolha natural para Ψ_m são as médias a posteriori obtidas das análises preliminares utilizando o algoritmo MCMC (ver Subseção 2.4.2).

2. Entre modelos:

2.a Gerar um valor candidato m' a partir da distribuição proposta $q(m, m')$.

2.b Dado m' , amostrar $\Psi_{m'}$ da distribuição proposta $q_{m'}(\Psi_{m'})$.

2.c O valor proposto m' é aceito com probabilidade

$$\alpha = \min \left\{ 1, \frac{p(y, m', \Psi_{m'})}{p(y, m, \Psi_m)} \frac{q_m(\Psi_m)q(m', m)}{q_{m'}(\Psi_{m'})q(m, m')} \right\}.$$

3. Se o modelo com m' fatores for aceito, atualizar o vetor paramétrico Ψ_m com os valores candidatos $\Psi_{m'}$. Caso contrário manter o modelo com m fatores e os valores correntes de Ψ_m .

4. Repetir os passos 2 e 3 um número grande de vezes.

Para este algoritmo é assumido que $Pr(m) = 1/M$, em que M é o número máximo de fatores comuns a ser considerado. Vale a pena enfatizar que a escolha das distribuições propostas, $q_m(\Phi_m)$, não garante sempre aproximações acuradas da posteriori condicional $p(\Phi_m|m, y)$. Por outro lado, o algoritmo RJMCMC proposto pode ser considerado um caso particular do método *Metropolized Carlin & Chib* (Godsill, 2001 e Dellaportas, Forster & Ntzoufras, 2002) onde as distribuições propostas, que geram tanto uma nova dimensão do modelo como novos parâmetros, dependem do estado atual da cadeia somente através de m . A denominação mais correta do algoritmo proposto é métodos RJMCMC independentes, análoga à terminologia padrão para métodos de Metropolis-Hastings independentes (como indicado em Gamerman & Lopes, 2006, Capítulo 7).

2.5 Estudo simulado

Dois exemplos com dados simulados são apresentados nesta seção. O primeiro considerando que a função de correlação do processo Gaussiano é a exponencial e o segundo com função de

correlação Matérn. No primeiro caso, estruturas mais simples do processo Gaussiano foram consideradas, ou seja, sem covariáveis e considerando a média do processo fixa no valor zero. Além disso, dois tipos de prioris (vagas e de referência) foram utilizadas para fazer inferência a posteriori dos parâmetros do processo Gaussiano. No segundo exemplo, covariáveis estáticas são incorporadas na média do processo Gaussiano. Em ambos estudos o algoritmo RJMCMC é utilizado para estimar o número de fatores. Adicionalmente, outros critérios de seleção de modelos são utilizados.

2.5.1 Exemplo 1: Correlação exponencial

Foram simulados dados considerando $N = 50$ localizações geradas num quadrado $[0, 1] \times [0, 1]$ com $m = 4$ fatores para $T = 100$ instantes de tempo. Para cada instante de tempo foram geradas observações $y_t = (y_{1t}, \dots, y_{Nt})$ seguindo um MFED. A função de correlação do processo Gaussiano $\beta_{(j)}$ é a exponencial da forma $\rho(d; \phi_j) = \exp(-d/\phi_j)$. Os parâmetros considerados foram: $\Lambda = \text{diag}(0, 15, 0, 1, 0, 2, 0, 07)$, $\Gamma = \text{diag}(0, 6, 0, 5, 0, 2, 0, 3)$, $\tau = (1, 0, 6, 0, 8, 0, 5)$ e $\phi = (0, 2, 0, 5, 0, 3, 0, 1)$. Valores para $\sigma_i^2, i = 1, \dots, 50$, foram simulados da distribuição uniforme no intervalo $[0, 01; 0, 05]$. A Figura 2.3 mostra a evolução dos dados ao longo do tempo em 6 locais. A Figura 2.5 os fatores comuns (linha verde cheia) e a Figura 2.4 as cargas dos fatores interpoladas em todo o quadrado unitário. Nota-se nestas figuras a influência dos fatores no comportamento das observações geradas.

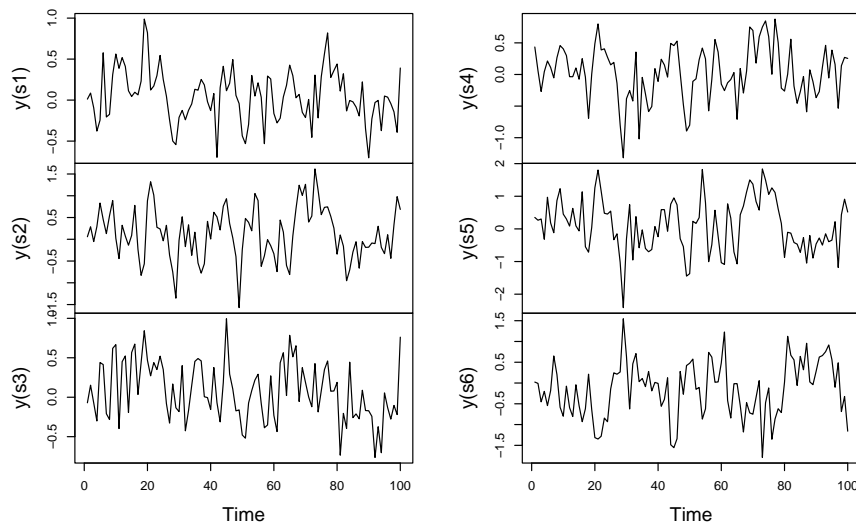


Figura 2.3: *Exemplo 1:* Dados gerados do MFED. Cada gráfico representa a série temporal nos locais s_1, \dots, s_6 .

As distribuições a priori utilizadas foram: σ , λ e τ têm distribuição gama inversa centradas no valor absoluto da média de y e desvio padrão igual a um, γ tem distribuição normal truncada com média igual ao valor verdadeiro e desvio padrão igual a um, e $\phi_j \sim GI(2, b)$, $j = 1, \dots, 4$ com $b = \max(dist)/(-2 \log 0.05)$. Adicionalmente foram consideradas as priors de referência para estimar os parâmetros do processo Gaussiano. Neste caso nenhum hiperparâmetro a priori tem que ser definido. O algoritmo MCMC proposto na Seção 2.4 foi utilizado para gerar amostras a posteriori de todos os parâmetros do modelo, incluindo o número de fatores, estimado via o algoritmo RJMCMC. Para cada parâmetro foram geradas duas cadeias paralelas (iniciando de valores diferentes) de tamanho 50.000, descartando-se as primeiras 10.000 iterações, consideradas período de aquecimento, guardando-se uma a cada 10 iterações. Portanto, o tamanho final das amostras é 4.000. Para verificar a convergência das cadeias foi utilizada estatística \hat{R} (ver Gelman & Rubin, 1992, para mais detalhes), em que valores próximos de 1 indicam convergência das cadeias. Os dados simulados foram ajustados considerando modelos com 3, 4, 5 e 6 fatores. Os resultados para estes 4 modelos são comparados utilizando critérios de informação como o AIC (Akaike, 1974) e o BIC (Schwarz, 1978), assim como medidas de erro como, por exemplo, o erro quadrático médio (EQM). O algoritmo RJMCMC fornece uma estimativa da probabilidade a posteriori do modelo (PPM), útil na seleção do número ótimo de fatores.

Tabela 2.1: *Exemplo 1:* Critérios de comparação de modelos. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Erro Quadrático Médio (EQM) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *itálico*.

<i>Priors vagas próprias</i>				
Critério	Número de fatores			
	$m = 3$	$m = 4$	$m = 5$	$m = 6$
AIC	7503,10	<i>6296,00</i>	6596,50	7510,90
BIC	10839,80	<i>10636,50</i>	11940,60	13858,70
EQM	305,40	<i>144,47</i>	144,49	154,36
PPM	0,000	<i>0,840</i>	0,158	0,002
<i>Priors de Referência</i>				
AIC	8957,30	<i>6263,20</i>	6829,40	6891,91
BIC	12294,10	<i>10603,70</i>	12173,50	13239,70
EQM	295,40	<i>140,60</i>	140,70	146,20
PPM	0,005	<i>0,755</i>	0,180	0,060

A Tabela 2.1 mostra os diferentes critérios de comparação para todos os modelos ajustados e para os dois tipos de priori utilizadas. Nota-se que o modelo com maior PPM é o modelo com

4 fatores, como esperado. A probabilidade a posteriori do modelo com 5 fatores é a segunda maior, sugerindo que este modelo pode ser considerado, por exemplo, no procedimento de mistura de modelos. Adicionalmente, os critérios de informação e a medida de erro também indicam o modelo com 4 fatores como o melhor. Nota-se também que o ajuste com 4 fatores e utilizando prioris de referência fornece os menores valores tanto para os critérios de informação quanto para a medida de erro. A comparação entre estas medidas é possível dado que o cálculo das duas primeiras está baseado na verossimilhança do modelo e a terceira em medidas de similaridade entre os valores ajustados e observados.

Para efeitos demonstrativos, os resultados apresentados a seguir estão baseados no ajuste do modelo com 4 fatores e com prioris de referência. A Figura 2.4 mostra as superfícies simuladas e estimadas de todas as colunas da matriz de cargas, cada par de figuras apresenta a mesma escala de variação e, portanto, os valores podem ser comparados. As cores mais claras indicam os valores mais altos e as mais escuras os valores mais baixos. Nota-se que as estimativas são bastantes parecidas indicando um bom ajuste destes valores.

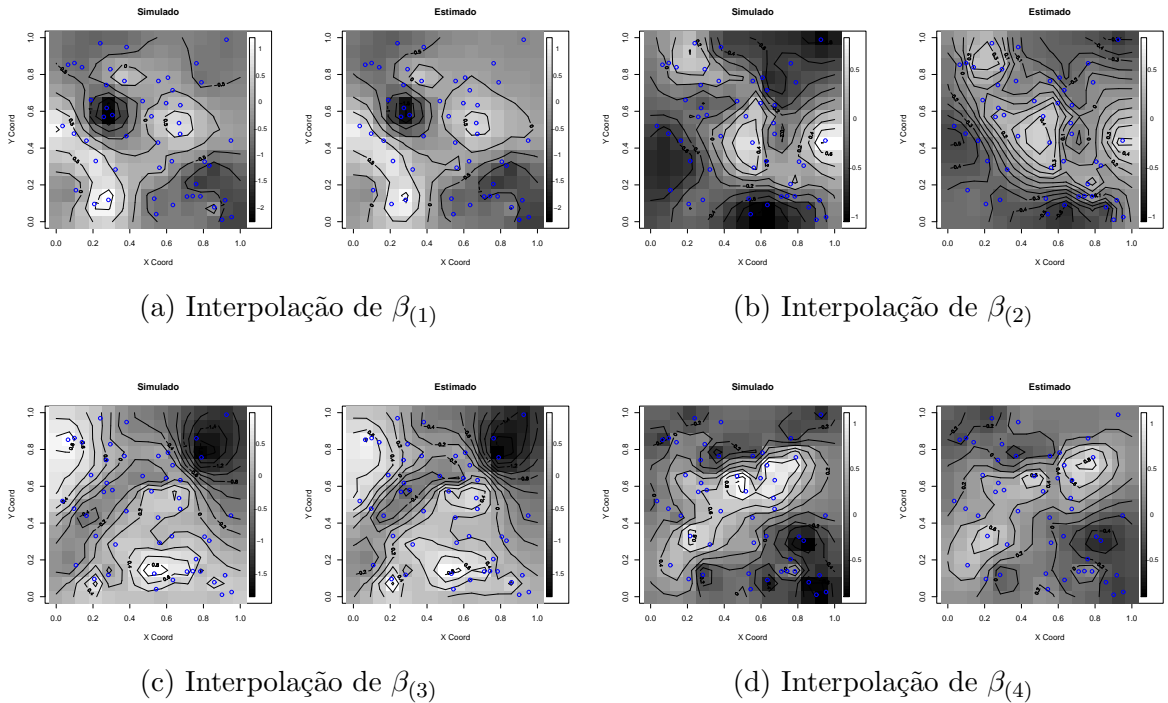


Figura 2.4: *Exemplo 1:* Interpolação espacial das componentes $\beta_{(j)}$ para $j = 1, \dots, 4$. Em cada gráfico, a figura da esquerda representa a superfície simulada e o da direita a superfície estimada via krigagem Bayesiana. Os pontos azuis indicam os locais onde as observações foram consideradas.

Na Figura 2.5 são apresentadas as médias a posteriori e os intervalos de credibilidade de

95% dos fatores comuns. Nota-se que as estimativas a posteriori captaram a dinâmica dos fatores ao longo do tempo, com intervalos de credibilidade contendo os valores verdadeiros em todo o período. A Tabela 2.2 mostra as estimativas a posteriori dos parâmetros do modelo. Em geral, todos os parâmetros foram bem estimados exceto λ_2 e λ_4 que apresentam médias a posteriori acima do valor verdadeiro.

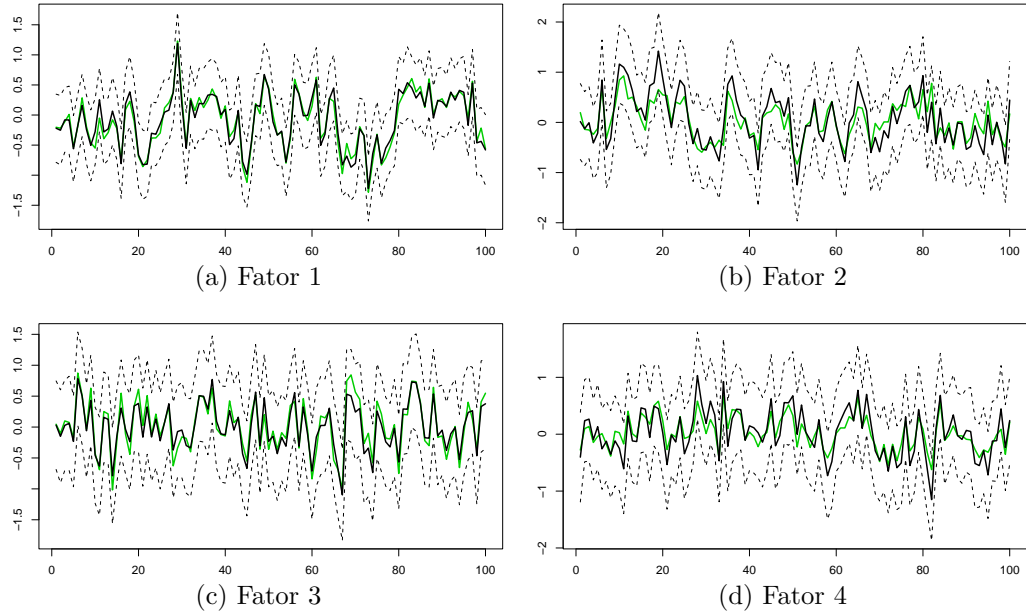


Figura 2.5: *Exemplo 1:* Fatores simulados vs. fatores estimados. A linha cheia verde representa os valores simulados, a linha cheia preta representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

Tabela 2.2: *Exemplo 1:* Sumário a posteriori dos parâmetros estimados. \hat{R} : diagnóstico de Gelman & Rubin.

θ	Valor	$E(\theta)$	$\sqrt{Var(\theta)}$	\hat{R}	θ	Valor	$E(\theta)$	$\sqrt{Var(\theta)}$	\hat{R}
λ_1	0,15	0,16	0,07	1,02	γ_1	0,60	0,54	0,09	1,01
λ_2	0,10	0,24	0,14	1,01	γ_2	0,50	0,50	0,09	1,00
λ_3	0,20	0,15	0,05	1,02	γ_3	0,20	0,18	0,09	1,01
λ_4	0,07	0,17	0,06	1,08	γ_4	0,30	0,29	0,10	1,01
τ_1^2	1,00	0,83	0,43	1,02	ϕ_1	0,2	0,19	0,08	1,00
τ_2^2	0,60	0,31	0,23	1,03	ϕ_2	0,5	0,44	0,48	1,01
τ_3^2	0,80	0,64	0,34	1,03	ϕ_3	0,3	0,22	0,16	1,02
τ_4^2	0,50	0,27	0,11	1,04	ϕ_4	0,1	0,17	0,08	1,03

2.5.2 Exemplo 2: Correlação Matérn

Para este segundo exemplo, $N = 25$ locais foram aleatoriamente selecionados no quadrado $[0, 1] \times [0, 1]$. Logo, para $t = 1, \dots, 100$, vetores y_t de dimensão 25×1 foram simulados a partir do MFED com 3 fatores comuns onde *i*) $\Gamma = \text{diag}(0, 6; 0, 4; 0, 3)$ e $\Lambda = \text{diag}(0, 02; 0, 03; 0, 01)$, *ii*) as colunas da matriz de cargas do fator seguem processos Gaussianos com função de correlação Matérn com $\phi = (0, 15; 0, 4; 0, 25)$, $\kappa = 1, 5$ e $\tau = (1, 00; 0, 75; 0, 56)$, e *iii*) $\mu_j^{\beta*} = X^\beta \mu_j^\beta$, $\mu_1^\beta = (5, 5, 4)'$, $\mu_2^\beta = (5, -6, -7)'$, $\mu_3^\beta = (5, -8, 6)'$, X^β contém na primeira coluna um termo constante igual a 1 e na segunda e terceira colunas os valores correspondentes a latitude e longitude do local e *iv*) σ_i^2 , $i = 1, \dots, 25$, foram geradas da distribuição uniforme no intervalo $(0, 01; 0, 05)$.

Neste exemplo, as últimas 10 observações foram deixadas de fora da análise para efeitos de comparação e previsão. Os dados gerados neste exemplo são os mesmos apresentados na Figura 2.2. Nota-se neste gráfico as superfícies dos processos Gaussianos influenciados diretamente pelos valores da latitude (eixo y) e longitude (eixo x) através dos valores de μ_j^β ($j = 1, 2, 3$). Na terceira coluna da matriz de cargas, nota-se a forte influência negativa da longitude (valores maiores na região oeste) e a influência positiva da latitude (valores maiores na região norte). Para $i = 1, \dots, 25$, a distribuição a priori para σ_i^2 e $GI(\epsilon, \epsilon)$ com $\epsilon = 0.01$. Para $j = 1, \dots, 3$, $\lambda_j \sim IG(\epsilon, \epsilon)$, $\gamma_j \sim N(0, 5; 1, 0)$, $\tau_j^2 \sim GI(2; 0, 75)$ e $\phi_j \sim GI(2; b)$ onde $b = \max(\text{dist})/(-2 \ln(0.05))$ e $\max(\text{dist})$ representa a máxima distância entre dois locais. Por último μ_j^β é normalmente distribuído com média igual ao valor verdadeiro e desvio padrão igual a 5. Neste exemplo, o parâmetro de forma da Matérn não foi estimado, permanecendo fixo no seu verdadeiro valor durante todo o processo de estimação.

Modelos com 2, 3, 4 e 5 fatores foram ajustados. A comparação entre os resultados destes modelos está baseada na probabilidade a posteriori do modelo (PPM) e nos critérios de informação utilizados no exemplo anterior. O algoritmo MCMC é utilizado para obter amostras a posteriori dos parâmetros de interesse seguindo as mesmas especificações do exemplo anterior. Na Tabela 2.3, estimativas para todos os critérios de comparação de modelos são apresentadas. Todos os critérios apontam o modelo com três fatores como o melhor, tanto no ajuste quanto na previsão (EQM_P). Entretanto, as probabilidades a posteriori dos modelos com 2 e 4 fatores não podem ser ignoradas, podendo ser utilizadas em técnicas como mistura Bayesiana de modelos para melhorar as estimativas de interpolação e previsão (ver [Raftery, Madigan & Hoeting, 1997](#) e [Clyde, 1999](#) para mais detalhes).

Tabela 2.3: *Exemplo 2:* Critérios de comparação de modelos. Akaike's information criterion (AIC); Schwartz's information criterion (BIC); Erro Quadrático Médio (EQM); Erro Absoluto Médio (EAM); EQM baseado nos últimos 10 valores previstos (EQM_P) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *itálico*.

m	AIC	BIC	EQM	EAM	EQM_P	PPM
2	1127.9	2643.3	0.11195	0.23822	2.3145	0.394
3	<i>-1005.5</i>	<i>1196.2</i>	<i>0.029108</i>	<i>0.13397</i>	2.3030	<i>0.437</i>
4	27589.7	30477.6	0.31112	0.41101	<i>2.2565</i>	0.101
5	42183.2	45757.3	0.46835	0.52360	2.2614	0.068

Tabela 2.4: *Exemplo 2:* Sumário a posteriori dos parâmetros que caracterizam a dinâmica dos fatores comuns e as colunas da matriz de cargas dos fatores. \hat{R} : diagnóstico de Gelman & Rubin.

θ	Valor	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			\hat{R}
				2.5%	50%	97.5%	
γ_1	0,60	0,504	0,091	0,325	0,504	0,687	1,03
γ_2	0,40	0,491	0,095	0,303	0,492	0,671	1,05
γ_3	0,30	0,416	0,100	0,209	0,417	0,623	1,03
λ_1	0,02	0,028	0,010	0,014	0,026	0,053	1,01
λ_2	0,03	0,019	0,004	0,011	0,018	0,028	1,03
λ_3	0,01	0,016	0,005	0,009	0,015	0,028	1,04
μ_{11}	5,00	4,44	0,89	2,78	4,43	6,21	1,01
μ_{21}	5,00	3,44	0,90	1,74	3,41	5,23	1,05
μ_{31}	4,00	4,10	0,85	2,44	4,09	5,76	1,01
τ_1^2	1,00	1,13	0,99	0,32	0,86	3,53	1,02
ϕ_1	0,15	0,20	0,08	0,10	0,19	0,40	1,01
μ_{12}	5,00	6,12	0,60	4,80	6,15	7,21	1,02
μ_{22}	-6,00	-6,00	0,80	-7,62	-5,99	-4,49	1,00
μ_{32}	-7,00	-7,51	0,86	-9,25	-7,45	-5,93	1,03
τ_2^2	0,75	0,51	0,91	0,11	0,30	2,29	1,10
ϕ_2	0,40	0,24	0,08	0,12	0,23	0,43	1,08
μ_{13}	5,00	4,53	0,58	3,39	4,52	5,67	1,05
μ_{23}	-8,00	-7,68	0,88	-9,43	-7,68	-6,01	1,05
μ_{33}	6,00	5,09	0,95	3,28	5,08	6,98	1,03
τ_3^2	0,56	0,44	0,40	0,15	0,34	1,36	1,04
ϕ_3	0,25	0,18	0,06	0,09	0,17	0,33	1,05

Os seguintes resultados estão baseados no modelo com 3 fatores. A Tabela 2.4 mostra as estimativas a posteriori dos parâmetros do modelo. Como uma indicação de que o MFED está capturando corretamente a estrutura dos dados, todos os parâmetros foram em geral bem estimados, pois os valores verdadeiros caíram dentro do intervalo de credibilidade de 95%. A estatística \hat{R} de Gelman & Rubin toma valores próximos de 1 para todos os parâmetros

sugerindo que as cadeias convergiram.

A Figura 2.6 mostra as superfícies simuladas e estimadas das colunas de β . Os três pares de gráficos sugerem que os valores foram em geral bem estimados, capturando a estrutura espacial presente nos dados. A Figura 2.7 mostra as estimativas a posteriori dos fatores comuns e os intervalos de credibilidade de 95%. Os três fatores foram bem ajustados sendo quase indistinguível a diferença entre os valores simulados e as médias a posteriori estimadas.

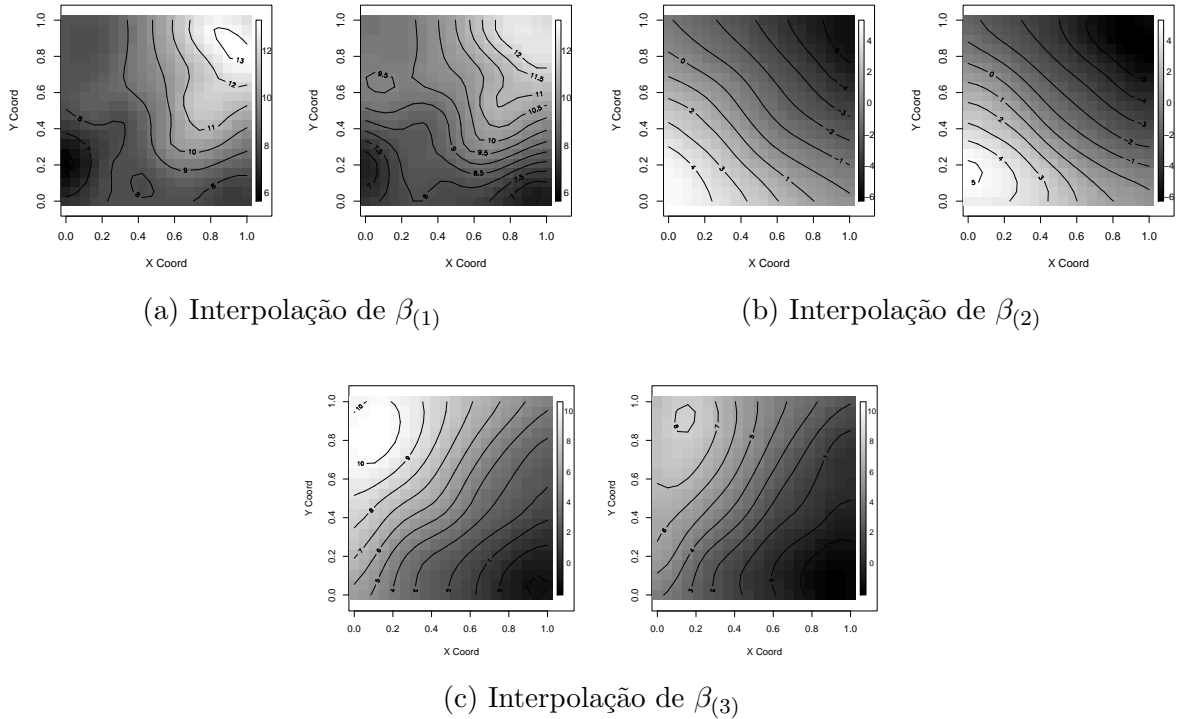


Figura 2.6: *Exemplo 2:* Interpolação espacial das componentes $\beta_{(j)}$ para $j = 1, \dots, 3$. Em cada gráfico, a figura da esquerda representa a superfície simulada e o da direita a superfície estimada via krigagem Bayesiana.

2.6 Considerações finais

Neste Capítulo foi introduzido o modelo fatorial espacial dinâmico para dados contínuos Gaussianos, o qual é apresentado como uma nova classe de modelos espaço-temporais não estacionários e com estrutura não separável que generaliza várias das alternativas existentes na literatura. O modelo faz uso da análise fatorial para explorar dependências espaciais e temporais. A variação espacial é modelada através das colunas da matriz de cargas dos fatores, enquanto que a dinâmica temporal é capturada através dos fatores comuns que seguem

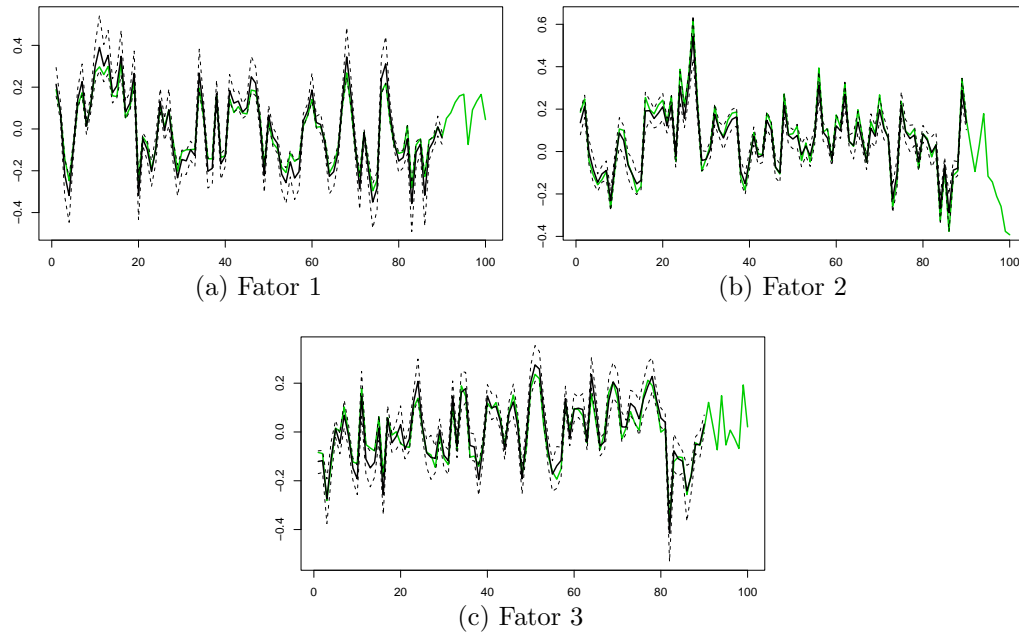


Figura 2.7: *Exemplo 2:* Fatores simulados vs. fatores estimados. A linha cheia verde representa os valores simulados, a linha cheia preta representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

processos auto-regressivos estacionários ou não estacionários. Assim, esta nova estrutura é capaz de separar naturalmente a variação espacial da variação temporal. O modelo toma vantagem de técnicas já sugeridas na literatura para processos espaciais e processos de séries temporais multivariadas. O MFED proposto se mostra flexível ao incluir diversas estruturas, tais como tendências (local e global), ciclos e sazonalidade, tanto na média do processo quanto na dinâmica dos fatores comuns.

Condicional ao número de fatores, inferência a posteriori é facilitada através do algoritmo MCMC que combina esquemas de amostragem, tais como o FFBS e o amostrador de Gibbs com passos de Metropolis-Hastings. Inferência para o número de fatores é desenvolvida através do algoritmo RJMCMC (MCMC com saltos reversíveis) que permite o cálculo da probabilidade a posteriori do modelo. Neste caso a maior probabilidade a posteriori indica o número ótimo de fatores.

Os estudos simulados sugerem que as técnicas de inferência propostas funcionam adequadamente e, portanto, podem ser utilizadas em situações práticas. O algoritmo de estimação requer métodos computacionalmente intensivos que podem ser implementados em alguma linguagem de programação (C, Ox ou Fortran, por exemplo). Devido a relação direta entre as cargas e os fatores, o sinal de ambas componentes pode ser trocado. Isto foi obser-

vado nos estudos simulados, nos quais os fatores e as cargas correspondentes eram estimados com o sinal trocado. Isto não representa um problema na modelagem dos dados já que o interesse é estimar o produto destas duas componentes para explicar a média do processo.

Capítulo 3

Modelo fatorial espacial com cargas dinâmicas

Neste capítulo, uma extensão natural do MFED é proposta ao permitir que a matriz de cargas dos fatores varie ao longo do tempo. O comportamento dinâmico da componente espacial é modelado através de duas componentes, uma puramente temporal e a outra espaço-temporal. Inferência para esta classe de modelos é sugerida utilizando um esquema MCMC que incorpora técnicas eficientes para a amostragem das cargas dinâmicas. Finalmente um estudo simulado é apresentado para mostrar a performance do modelo proposto.

3.1 Introdução

Modelos fatoriais com cargas dinâmicas já foram propostos na literatura. [Lopes & Migon \(2002\)](#) e [Lopes & Carvalho \(2007\)](#) mostraram as vantagens de modelar a matriz de cargas dos fatores variando no tempo. Eles apontaram que a principal motivação para incorporar cargas dinâmicas no modelo fatorial é permitir mudanças na estrutura de covariância ao longo do tempo. Por exemplo, no modelo com cargas estáticas, variações temporais da estrutura de covariância devem-se exclusivamente a mudanças na variância do fator comum. Uma estrutura mais rica da estrutura de covariância pode ser de grande ajuda na modelagem de dados espaço-temporais. Assim, uma generalização da abordagem anterior é feita nessa linha.

A estrutura dinâmica das cargas do fator pode ser especificada de várias maneiras. Em particular, nesta tese foi utilizada a proposta de [Gelfand, Barnerjee & Gamerman \(2005\)](#). Eles propuseram a decomposição dos coeficientes de um modelo de regressão linear em duas componentes, uma puramente temporal e a outra espaço-temporal. Estas duas componentes foram utilizadas para estender o MFED considerando cargas dinâmicas na especificação do

modelo. As duas componentes seguem processos auto-regressivos de ordem um tal que a informação inicial da componente espaço-temporal é um processo Gaussiano, como especificado no modelo com cargas estáticas. Inferência para esta nova classe é proposta utilizando uma abordagem complementemente Bayesiana, incluindo a estimação do número de fatores com o algoritmo RJMCMC.

Este capítulo está dividido da seguinte forma. Na Seção 3.2 é apresentado o MFED com cargas dinâmicas e algumas características do modelo. Na Seção 3.3, o procedimento de inferência é descrito, inclusive o algoritmo RJMCMC para estimar o número de parâmetros. Na Seção 3.4, um estudo simulado é apresentado para testar os algoritmos propostos. Por último, na Seção 3.5, encontram-se as considerações finais.

3.2 Modelo proposto

O modelo fatorial espacial com cargas dinâmicas é representado pelas seguintes equações:

$$y_t = \mu_t^{y*} + \beta_t f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma), \quad (3.1a)$$

$$f_t = \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda), \quad (3.1b)$$

$$\beta_{(j)t} = \delta_{(j)t} + X_{j,t}^\beta \mu_{j,t}, \quad (3.1c)$$

$$\delta_{(j)t} = \delta_{(j)t-1} + \eta_{(j)t}, \quad \eta_{(j)t} \sim N(0, \tau_j^2 R_{\phi_j}), \quad (3.1d)$$

$$\mu_{j,t} = \mu_{j,t-1} + \eta_{j,t}^\mu, \quad \eta_{j,t}^\mu \sim N(0, \Sigma_{\mu_j}). \quad (3.1e)$$

Nota-se que estas equações se diferenciam de (2.1) nas equações (3.1a) e (3.1c). Neste caso, cada coluna $\beta_{(j)t}$ é decomposta em duas componentes: *i*) uma componente puramente temporal denotada por $\mu_{j,t}$ e *ii*) outra espaço-temporal denotada por $\delta_{(j)t}$. $X_{j,t}^\beta$ representa a matriz de p covariáveis de dimensão $N \times p$. Por outro lado, a evolução $\eta_{(j)t}$ segue um processo Gaussiano como em (2.1c). O modelo é completado com a especificação das seguintes priors:

$$\delta_{(j)0} \sim N(0, \tau_{j0}^2 R_{\phi_{j0}}),$$

$$\mu_{j,0} \sim N(m_0^\mu, C_0^\mu),$$

$$f_0 \sim N(m_0, C_0)$$

onde ϕ_{j0} , τ_{j0}^2 ($j = 1, \dots, m$), m_0^μ , C_0^μ , m_0 e C_0 são hiperparâmetros conhecidos. Mais especificamente, os parâmetros ϕ_{j0} , τ_{j0}^2 podem ser fixados nas médias a posteriori estimadas do modelo com cargas estáticas.

O modelo pode ser reescrito da seguinte forma. Sejam $\delta_t = (\delta'_{(1)t}, \dots, \delta'_{(m)t})'$, $\mu_t = (\mu'_{1t}, \dots, \mu'_{mt})'$ e $\delta_t^* = (\mu'_t, \delta'_t)'$ vetores de dimensão $Nm \times 1$, $mp \times 1$ e $m(N+p) \times 1$, respectivamente. Seja $X_{j,t}^\beta = X_t^\beta, \forall j$. Assim, podemos reescrever as equações (3.1a) e (3.1d) como:

$$y_t = f_t^* \delta_t^* + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \quad (3.2a)$$

$$\delta_t^* = \delta_{t-1}^* + \eta_t^*, \quad \eta_t^* \sim N(0, \Sigma_{\delta^*}) \quad (3.2b)$$

onde $f_t^* = (f_t' \otimes X_t^\beta \mid f_t' \otimes I_N)$ tem dimensão $N \times m(N+p)$ e Σ_{δ^*} é uma matriz bloco diagonal tal que $\Sigma_{\delta^*} = \text{diag}(W_\mu, \tau_1^2 R_{\phi_1}, \dots, \tau_m^2 R_{\phi_m})$ com $W_\mu = \text{diag}(\Sigma_{\mu_1}, \dots, \Sigma_{\mu_m})$.

Segundo esta representação, temos que a componente δ^* combina, em um único vetor, a informação das duas componentes da matriz β_t . Isto será muito útil para propor algoritmos de inferência eficientes a ser tratados na Seção seguinte.

Função de verossimilhança

Sem perda de generalidade, e para facilitar a notação é assumido que $\mu_t^{y^*} = 0$. Aqui, definimos $\Theta = (\sigma, \lambda, \gamma, \tau, \phi, \sigma_\mu)$ como o vetor que contém os parâmetros livres do modelo onde $\sigma = (\sigma_1^2, \dots, \sigma_N^2)'$, $\lambda = (\lambda_1, \dots, \lambda_m)'$, $\gamma = (\gamma_1, \dots, \gamma_m)'$, $\tau = (\tau_1^2, \dots, \tau_m^2)'$, $\phi = (\phi_1, \dots, \phi_m)'$ e $\sigma_\mu = (\Sigma_{\mu_1}, \dots, \Sigma_{\mu_m})$. Por outro lado, seja $F = (f_1, \dots, f_T)'$ e $\delta^* = (\delta_1^*, \dots, \delta_T^*)'$. Assim, da equação (3.2a) temos que a função de verossimilhança de (Θ, F, δ^*, m) é dada por

$$p(y|\Theta, F, \delta^*, m) = (2\pi)^{-TN/2} |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^T (y_t - f_t^* \delta_t^*)' \Sigma_{\delta^*}^{-1} (y_t - f_t^* \delta_t^*) \right\} \quad (3.3)$$

Inclusão de covariáveis no modelo

Analogamente ao modelo com cargas estáticas, a inclusão de covariáveis pode ser feita tanto no nível médio $\mu_t^{y^*}$ quanto na componente puramente temporal $\mu_{j,t}$ através da matriz $X_{j,t}^\beta$. Para o nível médio do processo algumas alternativas já foram apresentadas na Subseção 2.2.1 que sugere a inclusão de covariáveis variando no tempo e covariáveis que dependem da localização.

A influência de covariáveis na especificação do processo Gaussiano é dada através da matriz $X_{j,t}^\beta$. Uma característica importante a ser ressaltada é que, diferente do caso estático, covariáveis variando no tempo podem ser consideradas. Por exemplo, algumas especificações são: *i)* $X_{j,t}^\beta = 1_N$, *ii)* $X_{j,t}^\beta = X_j^\beta, \forall t$ onde X_j^β é uma matriz de dimensão $N \times p_j$ e *iii)*

$X_{j,t} = (1_N, X_{1,t}^j, \dots, X_{p_j,t}^j)$, tal que $X_{l,t}^j$, $l = 1, \dots, p_j$ são vetores de dimensão $N \times 1$. Nota-se que diferentes covariáveis podem ser incorporadas na especificação de β_t , dando maior flexibilidade ao modelo.

Fatores sazonais

A inclusão de padrões cíclicos podem ser naturalmente incorporados no MFED com cargas dinâmicas seguindo as especificações apresentadas na Subseção 2.2.2 em que os fatores sazonais são ponderados pelas colunas da matriz de cargas. Neste caso, a ponderação é feita pelas cargas dinâmicas sugerindo diferentes padrões sazonais em cada local.

3.2.1 Não separabilidade

No MFED com cargas dinâmicas, pode ser mostrado que a covariância entre locais i, j nos instantes de tempo $t, t + h$ é:

- para $m = 1$

$$\text{Cov}(y_{it}, y_{j,t+h}) = (\lambda\gamma^h)(1 - \gamma^2)^{-1} \mathbf{E}(\beta_{i,t}\beta_{j,t})$$

onde $\mathbf{E}(\cdot)$ denota a função esperança, $\beta_{i,t}$ é a i -ésima componente da matriz β_t . Neste caso as funções de covariância espacial e temporal são separáveis e identificáveis, e

- para $m > 1$

$$\text{Cov}(y_{it}, y_{j,t+h}) = \sum_{k=1}^m (\lambda_k \gamma_k^h) (1 - \gamma_k^2)^{-1} \mathbf{E}(\beta_{i,(k)t} \beta_{j,(k)t})$$

onde $\beta_{i,(k)t}$ é a i -ésima componente da coluna $\beta_{(k)t}$, $k = 1, \dots, m$, apresentando neste caso uma estrutura claramente não separável.

3.2.2 Previsão

Previsão é feita de forma análoga ao caso estático. Neste caso a densidade preditiva h passos ao frente $p(y_{t+h})$ é dada por:

$$p(y_{T+h}|y) = \int p(y_{T+h}|f_{T+h}, \delta_{T+h}^*, \Theta) p(f_{T+h}|f_T, \delta_{T+h}^*, \Theta) p(\delta_{T+h}^*|\delta_T^*, \Theta) p(f_T, \delta_T^*, \Theta|y) df_{T+h} df_T d\delta_{T+h}^* d\delta_T^* d\Theta$$

onde $(y_{T+h}|f_{T+h}, \delta_T^*, \Theta) \sim N(\beta_T f_{T+h}, \Sigma)$, $(f_{T+h}|f_T, \delta_{T+h}^*, \Theta) \sim N(\mu_h, V_h)$, $\mu_h = \Gamma^h f_T$, $V_h = \sum_{k=1}^h \Gamma^{k-1} \Lambda (\Gamma^{k-1})'$ e $(\delta_{T+h}^*|\delta_T^*, \Theta) \sim N(\delta_T^*, h\Sigma_{\delta^*})$ para $h > 0$.

Portanto, se $\{(f_T^{(1)}, \beta_T^{(1)}, \Theta^{(1)}), \dots, (f_T^{(L)}, \beta_T^{(L)}, \Theta^{(L)})\}$ é uma amostra de $p(f_T, \beta_T, \Theta|y)$ é fácil amostrar $f_{T+h}^{(j)}$ de $p(f_{T+h}|f_T^{(j)}, \beta_T^{(j)}, \Theta^{(j)})$ e $\beta_{T+h}^{(j)}$ de $p(\delta_{T+h}^*|\delta_T^{*(j)}, \Theta^{(j)})$, para todo $j = 1, \dots, L$. Desta forma temos que

$$\hat{p}(y_{T+h}|y) = \frac{1}{L} \sum_{j=1}^L p(y_{T+h}|f_{T+h}^{(j)}, \beta_T^{(j)}, \Theta^{(j)})$$

é a aproximação via Monte Carlo de $p(y_{T+h}|y)$. Analogamente, se $y_{T+h}^{(j)}$ é amostrado de $p(y_{T+h}|f_{T+h}^{(j)}, \beta_{T+h}^{(j)}, \Theta^{(j)})$, para $j = 1, \dots, L$, então $\{y_{T+h}^{(1)}, \dots, y_{T+h}^{(M)}\}$ representa uma amostra de $p(y_{T+h}|y)$.

3.3 Procedimento de inferência

Nesta seção é apresentado um esquema MCMC para fazer inferência a posteriori completa de todos os parâmetros do modelo proposto. O número de fatores é também considerado como um parâmetro do modelo e estimado através do algoritmo RJMCMC.

Distribuições a priori

Para este modelo podem ser utilizadas as mesmas distribuições a priori definidas na Subseção 2.4.1. Nota-se que a distribuição a priori para os fatores comuns é especificada na equação (3.1b) e completada por $f_0 \sim N(m_0, C_0)$. A distribuição a priori para a componente puramente temporal das cargas dinâmicas é especificada na equação (3.1e) com informação inicial dada por $\mu_{j,0} \sim N(m_0^\mu, C_0^\mu)$, enquanto que a priori para a componente espaço-temporal é dada pela equação (3.1d) e completada por $\delta_{(j)0} \sim N(0, \tau_{j0}^2, R_{\phi_{j0}})$, $j = 1, \dots, m$.

Os hiperparâmetros m_0 , C_0 , m_0^μ e C_0^μ são considerados conhecidos e fixados no início da análise. A informação inicial de $\delta_{(j)0}$ é um processo Gaussiano com parâmetros τ_{j0}^2 e ϕ_{j0} . Estes valores podem ser fixados ou estimados durante o procedimento de inferência. Para a segunda opção, testes prévios indicaram um incremento na incerteza associada a δ_t , principalmente nos primeiros valores da série. Para contornar este problema, se sugere fixar esses parâmetros nas médias a posteriori de resultados prévios do modelo com cargas estáticas.

3.3.1 Inferência a posteriori

Condicional ao número de fatores m a distribuição condicional completa conjunta de (Θ, F, δ^*) é dada por

$$\begin{aligned}
 p(\Theta, F, \delta^* | y) &\propto \prod_{t=1}^T p(y_t | f_t, \delta_t^*, \sigma) p(f_0 | m_0, C_0) \prod_{t=1}^T p(f_t | f_{t-1}, \lambda, \gamma) \times \\
 &\quad p(\delta_0^* | \tau_0, \phi_0) \prod_{t=1}^T p(\delta_t^* | \delta_{t-1}^*, \sigma_\mu, \tau, \phi) \prod_{j=1}^m p(\gamma_j) p(\lambda_j) \times \\
 &\quad \prod_{j=1}^m p(\tau_j) p(\phi_j) p(\sigma_{\mu_j}^2) \prod_{i=1}^N p(\sigma_i^2). \tag{3.4}
 \end{aligned}$$

Dado o número de fatores, um esquema MCMC é utilizado para gerar amostras a posteriori dos parâmetros de interesse. As distribuições condicionais completas de λ e γ são Gama inversa e normal respectivamente e amostras são geradas utilizando o amostrador de Gibbs (ver Apêndice A.1 para mais detalhes). Para os parâmetros σ , σ_u e τ as condicionais completas são Gama inversa. O parâmetro ϕ é amostrado com passos de Metropolis-Hastings com densidade proposta log-normal (mas detalhes no Apêndice A.2). Analogamente ao modelo com cargas estáticas, os fatores são amostrados conjuntamente utilizando o algoritmo FFBS detalhado no Apêndice A.1 com uma única variação: substituir β por β_t em todas as contas.

Amostragem de δ_t^*

Os valores da componente espaço-temporal, δ_t^* , são amostrados conjuntamente utilizando o algoritmo FIFBS (do inglês *Forward Information Filtering Backward Sampling*) proposto por Vivar (2007). Este algoritmo é similar ao FFBS (de Carter & Kohn, 1994 e Frühwirth-Schnatter, 1994). A diferença está na primeira etapa do algoritmo, na qual o filtro de Kalman é substituído pelo filtro de informação (Anderson & Moore (1979)). Na segunda etapa a amostragem é feita de forma retrospectiva como no FFBS. O filtro de informação é uma variante do filtro de Kalman que apresenta em sua formulação vantagens computacionais referidas principalmente ao tratamento das matrizes de precisão, evitando nesse caso a inversão de matrizes a cada passo da filtragem.

Utilizando a representação em (3.2), a priori para δ_0^* pode ser reescrita como

$$\delta_0^* \sim N(m_0^\delta, C_0^\delta)$$

onde $m_0^\delta = (m_0^{\mu*}, 0_{Nm})'$, $C_0^\delta = \text{diag}(C_0^{\mu*}, \tau_{10}^2 R_{\phi_{10}}, \dots, \tau_{m_0}^2 R_{\phi_{m_0}})$, $\mu_0^{\mu*} = 1_p \otimes \mu_0^\mu$ e $C_0^{\mu*} = C_0^\mu \otimes I_p$

O filtro de informação é descrito nos seguintes passos:

Algoritmo 3.1. (*Filtro de Informação para δ_t^**)

1. *Posteriori em $t - 1$:* $\delta_{t-1}^* | D_{t-1} \sim N(m_{t-1}^\delta, C_{t-1}^\delta)$.

2. *Priori em t :* $\delta_t^* | D_{t-1} \sim N(a_t^\delta, R_t^\delta)$ onde

$$a_t^\delta = m_{t-1}^\delta \quad e \quad R_t^{\delta^{-1}} = \Sigma_{\delta^*}^{-1} - \Sigma_{\delta^*}^{-1} (\Sigma_{\delta^*}^{-1} + C_{t-1}^{\delta^{-1}})^{-1} \Sigma_{\delta^*}^{-1}.$$

3. *Posteriori em t :* $\delta_t^* | y_t, D_{t-1} \sim N(m_t^\delta, C_t^\delta)$ onde

$$C_t^{\delta^{-1}} = R_t^{\delta^{-1}} + f_t^{*\prime} \Sigma^{-1} f_t^* \quad e \quad C_t^{\delta^{-1}} m_t^\delta = R_t^{\delta^{-1}} a_t^\delta + f_t^{*\prime} \Sigma^{-1} y_t^\beta.$$

Com estes resultados é fácil mostrar que $\delta_t^* | \delta_{t+1}^*, D_t \sim N(M_t^\delta, V_t^\delta)$

$$V_t^{\delta^{-1}} = C_t^{\delta^{-1}} + \Sigma_{\delta^*}^{-1} \quad e \quad V_t^{\delta^{-1}} M_t^\delta = C_t^{\delta^{-1}} m_t^\delta + \Sigma_{\delta^*}^{-1} \delta_{t+1}^*.$$

Em resumo, para obter amostras do bloco completo δ^* basta seguir os seguintes passos:

Algoritmo 3.2. (*FIFBS para δ^**)

1. Utilizar o filtro de informação para obter $C_t^{\delta^{-1}}$ e $C_t^{\delta^{-1}} m_t^\delta$.

2. Fazer $t = T$

3. Amostrar da distribuição $p(\delta_t^* | D_t)$.

4. Fazer $t \rightarrow t - 1$.

5. Amostrar da distribuição $p(\delta_t^* | \delta_{t+1}^*, D_t)$.

6. Voltar ao passo 4 até $k = 1$

3.3.2 Inferência para o número de parâmetros

Para fazer inferência para o número de parâmetros o algoritmo RJMCMC proposto na Seção 2.4.3 é estendido, considerando a seguinte distribuição proposta conjunta:

$$\begin{aligned}
q_m(\Psi_m) &= \prod_{j=1}^m f_N(f_{(j)}; M_{f_{(j)}}, aV_{f_{(j)}}) \prod_{t=1}^T f_N(\delta_t^*; M_{\delta_t^*}, bV_{\delta_t^*}) f_N(\gamma_j; M_{\gamma_j}, cV_{\gamma_j}) \\
&\times \prod_{j=1}^m f_{GI}(\lambda_j; d, dM_{\lambda_j}) f_N(\sigma_{\mu_j}; M_{\sigma_j}, eV_{\sigma_j}) f_{GI}(\phi_j; f, fM_{\phi_j}) \\
&\times \prod_{j=1}^m f_{GI}(\tau_j^2; g, gM_{\tau_j}) f_{GI}(\sigma_j^2; h, hM_{\sigma_j}).
\end{aligned} \tag{3.5}$$

A distribuição proposta da componente δ_t^* depende de $M_{\delta_t^*}$ e $V_{\delta_t^*}$ que representam a média e a variância a posteriori de um resultado preliminar do algoritmo MCMC com algum valor de m pré-especificado. Assim, de acordo com o algoritmo descrito em 2.4.3, depois de um valor candidato m' ser amostrado da distribuição proposta $q(m, m')$, é possível gerar um valor candidato $\Psi'_{m'}$ de $q'_{m'}(\Psi_{m'})$. Por último, o par $(m', \Psi_{m'})$ é aceito com probabilidade

$$\alpha = \min \left\{ 1, \frac{p(y, m', \Psi_{m'})}{p(y, m, \Psi_m)} \frac{q_m(\Psi_m)q(m', m)}{q_{m'}(\Psi_{m'})q(m, m')} \right\}. \tag{3.6}$$

3.4 Estudo Simulado

Dados do MFED com cargas dinâmicas foram simulados considerando $m = 2$ fatores, $N = 30$ locais (aleatoriamente escolhidos num quadrado $[0, 1] \times [0, 1]$) para $T = 80$ instantes de tempo. Os fatores evoluem de acordo com a especificação em (3.1b). Cada coluna da matriz de cargas $\beta_{(j)t}$, $j = 1, 2$, se comporta de acordo com (3.1c) onde cada componente $\mu_{j,t}$ e $\delta_{(j)t}$ evolui seguindo passeios aleatórios como em (3.1e) e (3.1d). O processo espaço-temporal foi inicializado em $\delta_{0t} \sim N(0, \tau_{j0}^2 R_{\phi_j})$ considerando uma função de correlação exponencial. Os parâmetros considerados foram $\Gamma = \text{diag}(0.95, 0.7)$, $\Lambda = \text{diag}(0, 49, 0, 81)$, $\sigma_i^2 \in [0, 02; 0, 93]$ ($i = 1, \dots, 30$), $\sigma_{\mu_1}^2 = 0, 64$ $\sigma_{\mu_2}^2 = 0, 25$), $\phi_{10} = 0, 15$, $\phi_{20} = 0, 4$, $\tau_{10}^2 = 0, 75$, $\tau_{20}^2 = 0, 9$, $\phi_1 = 0, 4$, $\phi_2 = 0, 1$, $\tau_1^2 = 1$ e $\tau_2^2 = 0, 49$.

As seguintes priors foram utilizadas: para $i = 1, \dots, 30$, $\sigma_i^2 \sim GI(0, 1; 0, 1)$, para $j = 1, 2$, $\gamma_j \sim N(0; 1)$ e $\lambda \sim GI(0, 1; 0, 1)$. As priors de referência propostas por Berger et al. (2001) foram utilizadas para os parâmetros do processo Gaussiano. Os hiperparâmetros ϕ_{j0} e τ_{j0}^2 foram assumidos conhecidos. O esquema MCMC proposto foi utilizado considerando 20.000

iterações, no qual as primeiras 10.000 foram descartadas. Amostras foram guardadas a cada 5 iterações e, portanto, os resultados estão baseados em cadeias de tamanho 2.000. Para efeitos de comparação foram rodados modelos com $m = 1, 2, 3, 4$ fatores. Todos estes modelos são comparados utilizando critérios de informação como o AIC (Akaike, 1974), BIC (Schwarz, 1978), o erro quadrático médio (EQM) e a probabilidade a posteriori do modelo calculada do algoritmo RJMCMC. A Tabela 3.1 mostra os resultados dos critérios de seleção de modelos. A maior probabilidade a posteriori é alcançada pelo modelo com 2 fatores como esperado. Os outros critérios também indicam este modelo como o melhor em termos de ajuste.

Tabela 3.1: *Exemplo:* Critérios de comparação de modelos. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Erro Quadrático Médio (EQM) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *italico*.

<i>Com prioris de Referência</i>				
Critério	Número de fatores			
	$m = 1$	$m = 2$	$m = 3$	$m = 4$
AIC	15848,09	<i>13928,00</i>	14565,40	16576,14
BIC	45294,47	<i>42844,10</i>	43934,50	46758,30
EQM	615,40	<i>603,52</i>	608,10	650,70
PPM	0,09	<i>0,56</i>	0,20	0,15

A Tabela 3.2 mostra os resultados a posteriori de todos os parâmetros do modelo. Nota-se que as estimativas foram bastantes acuradas com intervalos de credibilidade contendo os valores verdadeiros. As estatísticas \hat{R} de Gelman & Rubin ficaram próximas de 1 o que sugere que as cadeias convergiram.

As Figuras 3.1 e 3.2 mostram as estimativas da componente temporal de β_t e dos fatores respectivamente. Nota-se que as estimativas ficaram próximas dos valores simulados sendo mais evidente nas cargas. Os intervalos de credibilidade da componente μ_{jt} são bastante amplos porém, as médias a posteriori conseguem acompanhar a dinâmica dos valores simulados.

Finalmente, na Figura 3.3 são mostradas as superfícies estimadas das cargas dinâmicas β_t para três instantes de tempo $t = 5, 40, 80$. As duas primeiras colunas referem-se aos resultados de $\beta_{(1)t}$ e, as duas últimas, a $\beta_{(2)t}$. Para cada componente o primeiro gráfico representa a superfície simulada, o segunda, a superfície estimada considerando a mesma escala de cores para fazer possível a comparação. Nota-se que as estimativas a posteriori ficaram muito próximas dos valores simulados, capturando em todos os casos a variação temporal na região.

Tabela 3.2: *Exemplo:* Sumário a posteriori dos parâmetros do modelo. \hat{R} : diagnóstico de Gelman & Rubin.

θ	Valor	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			\hat{R}
				2.5%	50%	97.5%	
γ_1	0,95	0,927	0,037	0,849	0,930	0,992	1,01
γ_2	0,70	0,614	0,089	0,442	0,615	0,786	1,03
λ_1	0,49	0,459	0,085	0,327	0,450	0,647	1,03
λ_2	0,81	0,913	0,420	0,313	0,940	1,670	1,02
σ_{μ_1}	0,64	0,901	0,380	0,207	0,880	1,710	1,10
σ_{μ_2}	0,25	0,395	0,260	0,107	0,319	1,070	1,08
ϕ_1	0,40	0,563	0,120	0,378	0,541	0,873	1,04
ϕ_2	0,10	0,121	0,020	0,085	0,120	0,167	1,02
τ_1^2	1,00	1,410	0,260	0,977	1,380	1,980	1,02
τ_2^2	0,49	0,528	0,29	0,247	0,387	1,200	1,05

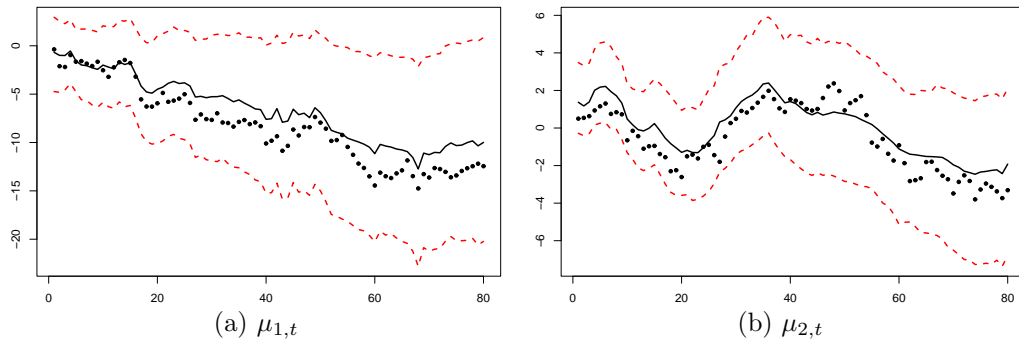


Figura 3.1: *Exemplo:* Médias a posteriori de $\mu_{j,t}$ para $j = 1, 2$ e $t = 5, 40, 80$. A linha cheia representa as médias a posteriori, as linhas pontilhadas os intervalos de credibilidade de 95% e os pontos os valores simulados.

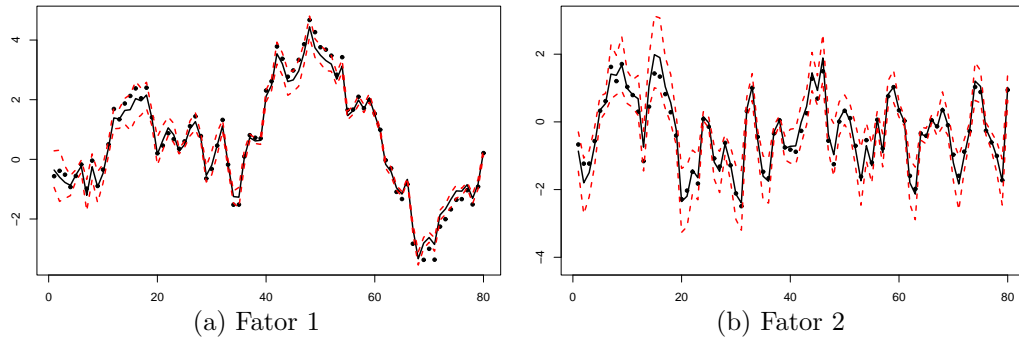


Figura 3.2: *Exemplo:* Fatores simulados vs. fatores estimados. A linha cheia representa as médias a posteriori, as linhas pontilhadas os intervalos de credibilidade de 95% e os pontos os valores simulados.

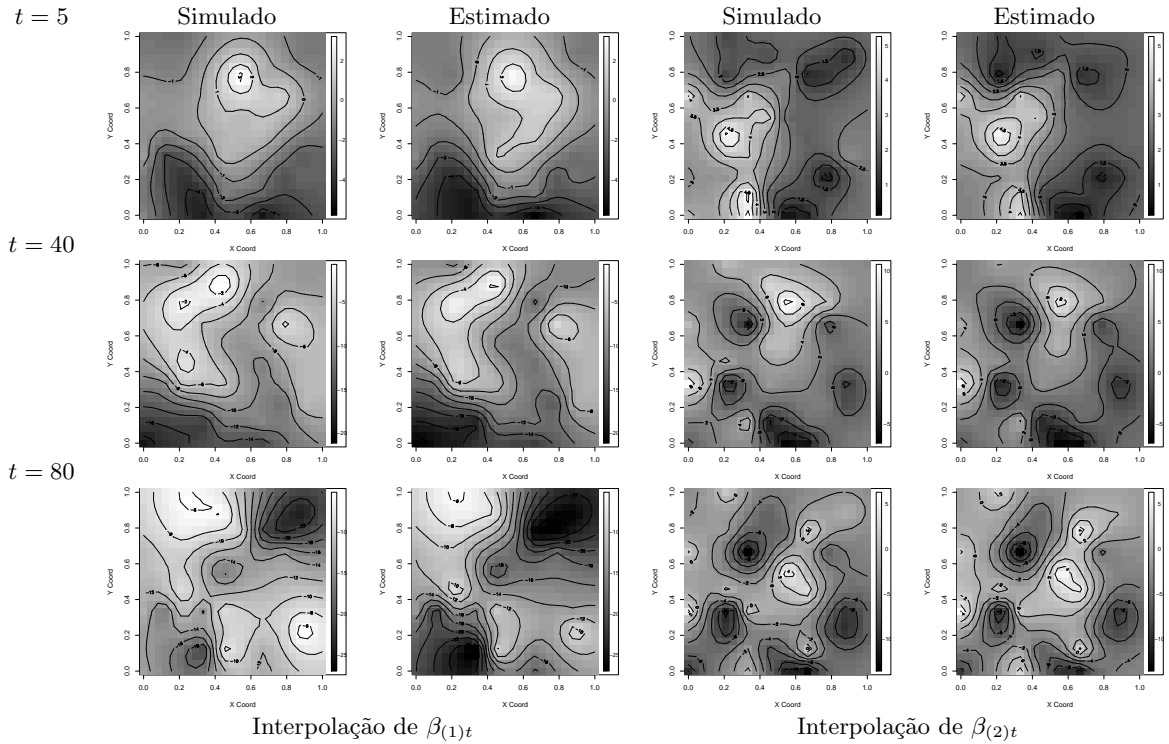


Figura 3.3: *Exemplo:* Interpolação espacial das componentes $\beta_{(j)t}$ e $\beta_{(2)t}$ para $t = 5, 40, 80$. Em cada gráfico, a figura da esquerda representa a superfície simulada e o da direita a superfície estimada via krigagem Bayesiana.

3.5 Considerações finais

Neste capítulo o MFED com cargas estáticas foi estendido ao permitir variações temporais na matriz de cargas dos fatores. Esta extensão parte do suposto de que cada componente de β_t pode ser explicada por duas componentes, uma puramente temporal e a outra espaço-temporal. Esta abordagem é uma generalização do modelo anterior e, portanto, todas as propriedades e características apresentadas no Capítulo 2 podem ser estendidas a este tipo de modelo.

Nota-se que este modelo é muito parametrizado, por esta razão algoritmos computacionalmente eficientes têm que ser utilizados para amostrar os parâmetros de interesse. Em particular, é proposto um esquema MCMC para fazer inferência do ponto de vista Bayesiano. Em cada passo do algoritmo MCMC, o algoritmo FIFBS é utilizado para amostrar conjuntamente a componente espaço-temporal, δ_t . Este algoritmo se mostra bastante útil dado que ajuda a reduzir o tempo computacional ao considerar as matrizes de precisão nas contas evitando a inversão de matrizes de grande dimensão. Por outro lado, o algoritmo RJMCMC

foi estendido permitindo a estimação do número de fatores nesta classe de modelos. O exemplo com dados simulados mostrou que os algoritmos de inferência funcionam adequadamente podendo ser aplicados a dados reais com estrutura espaço-temporal complexa.

Capítulo 4

Aplicação: Concentrações de SO₂ na região leste dos EUA

Nesta seção o MFED proposto é utilizado para ajustar os níveis de SO₂ medidos na região leste dos EUA. Numa primeira análise, o MFED com cargas estáticas é utilizado. Várias especificações são consideradas na modelagem baseadas principalmente na inclusão de covariáveis tanto na média do processo quanto na média do processo Gaussiano. Adicionalmente, dada a presença de ciclos anuais nos dados, fatores sazonais são considerados. Numa segunda análise, o MFED com cargas dinâmicas é ajustado. O esquema MCMC proposto é utilizado para fazer inferência a posteriori para todos os modelos especificados. A comparação entre as classes especificadas é feita utilizando diferentes critérios, entre eles a probabilidade a posteriori do modelo obtido através do algoritmo RJMCMC. Adicionalmente, e para efeitos de comparação, os resultados obtidos com o MFED são contrastados com resultados de modelos mais simples, comumente utilizados na modelagem espaço-temporal para dados normais.

4.1 Introdução

Os dados analisados neste capítulo foram disponibilizados pela *Clean Air Status and Trends Network* (CASTNet) que é uma rede de monitoramento pertencente à Agência de Proteção Ambiental (EPA) dos EUA. A rede conta com 80 estações monitoradoras espalhadas em toda a área dos EUA. Cada estação monitoradora da CASTNet mede concentrações atmosféricas semanais de sulfato, nitrato, amônia, dióxido sulfúrico e ácido nítrico. Mais detalhes da rede podem ser encontrados no sítio <http://www.epa.gov/castnet/>.

Dióxido sulfúrico é um dos maiores componentes da “chuva ácida”, associado com danos ao meio ambiente. Este gás é irritante, e a exposição a altas concentrações, por exemplo,

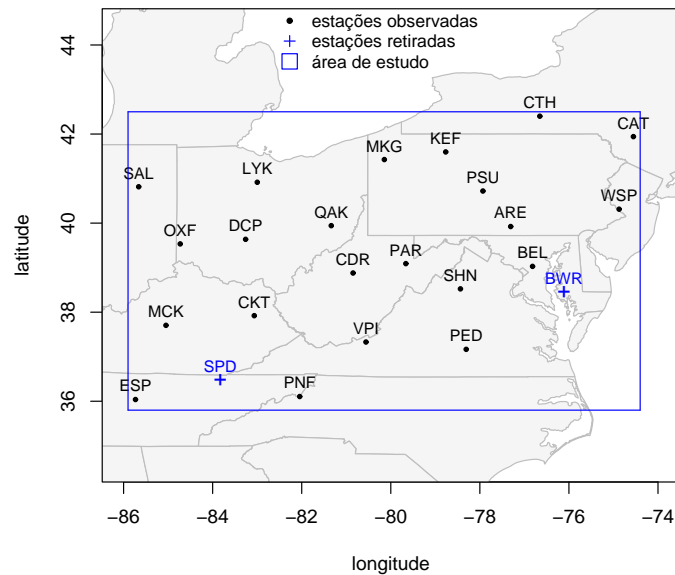


Figura 4.1: Dados CASTNet: Localização das estações monitoradoras. As estações SPD e BWR em azul foram retiradas da análise para efeitos de comparação.

durante dias de grande poluição, pode causar dificuldades de respiração. Altos níveis de dióxido sulfúrico também pioram os sintomas da asma. Nesta aplicação são analisadas as concentrações semanais de dióxido sulfúrico (SO₂) medidas em $\mu\text{g}/\text{m}^3$. A área de estudo corresponde à região leste dos EUA onde podem ser encontrados os registros mais antigos destas medidas. Mais especificamente, 24 estações monitoradoras foram selecionadas para fazer a análise espaço-temporal com medidas a partir da primeira semana de 1998 até a 30a semana de 2004, no total 342 observações em cada local. A Tabela 4.1 mostra a lista das estações. A Figura 4.1 mostra a área de estudo delimitada pelo quadrado e a localização das 24 estações. A área compreende os estados de Indiana, Ohio, Tennessee, Kentucky, Carolina do Norte, Virginia, Virginia Ocidental, Pennsylvania, Maryland, Nova Jersey e Nova York. Esta área se caracteriza por ter uma cadeia de montanhas paralela à costa do Atlântico, com áreas industrializadas principalmente na região norte da área delimitada. As medições nesta área indicam altos níveis de SO₂ principalmente em locais pertencentes ao estado de Ohio.

Para avaliar a performance de interpolação espacial do modelo as estações BWR e SPD foram retiradas da análise. Para avaliar a performance de previsão, as últimas 30 semanas (1a à 30a semana de 2004) foram retiradas e comparadas com os valores estimados. Em resumo, um total de $T = 312$ medidas ao longo do tempo e $N = 22$ estações são utilizadas na análise.

Tabela 4.1: *Dados CASTNet:* Lista das estações monitoradoras selecionadas.

Estação	Estado	Nome	Estação	Estado	Nome
SAL	IN	Salamonie Reservoir	BEL	MD	Beltsville
MCK	KY	Mackville	WSP	NJ	Wash. Crossing
CKT	KY	Crockett	CTH	NY	Connecticut Hill
PNF	NC	Cranberry	CAT	NY	Claryville
OXF	OH	Oxford	KEF	PA	Kane Exp. Forest
DCP	OH	Deer Creek	PSU	PA	Penn State
LYK	OH	Lykens	ARE	PA	Arendtsville
QAK	OH	Quaker City	SHN	VA	Shenandoah NP - Big Meadows
MKG	PA	M. K. Goddard	PED	VA	Prince Edward
ESP	TN	Edgar Evins	PAR	WV	Parsons
VPI	VA	Horton Station	SPD*	TN	Speed Well
CDR	WV	Cedar Creek	BWR*	MD	Blackwater NWR

* Locais de interpolação.

Pelo fato dos dados serem estritamente positivos, a transformação logaritmo foi utilizada para que os dados sejam aproximadamente normais. A Figura 4.2a mostra as séries temporais do logaritmo de SO₂ para quatro estações monitoradoras. A transformação logaritmo normaliza os dados muito razoavelmente como mostrada na Figura 4.1b. Visualmente, as séries temporais exibem um comportamento sazonal com um aparente ciclo anual com valores altos no inverno (início do ano). A sazonalidade pode ser explicada em termos gerais como resultado das altas taxas de oxidação de SO₂ a outros poluentes atmosféricos, durante o verão. Por outro lado parece haver um leve decréscimo na tendência das séries temporais ao longo dos anos, provavelmente devido a implementação do “Programa de Chuva Ácida” da EPA no Leste dos EUA em 1995 (Fase I) e 2000 (Fase II). Por outro lado, um procedimento de correção para o efeito de curvatura da terra, comumente presente em dados espacialmente distribuídos, é aplicado aos dados. Assim, os valores de latitude e longitude são convertidos em coordenadas UTM (*Universal Transverse Mercator*) tal que as coordenadas convertidas são medidas em quilômetros.

Nas seguintes duas seções, o MFED é utilizado no ajuste dos dados, a primeira abordagem considera que as cargas dos fatores são invariantes no tempo como proposto no Capítulo 2. Numa segunda abordagem, modelos com cargas variando no espaço e no tempo são consideradas como proposto no Capítulo 3. Estes modelos são comparados com modelos espaço-temporais mais simples baseados nos resultados da previsão h passos a frente.

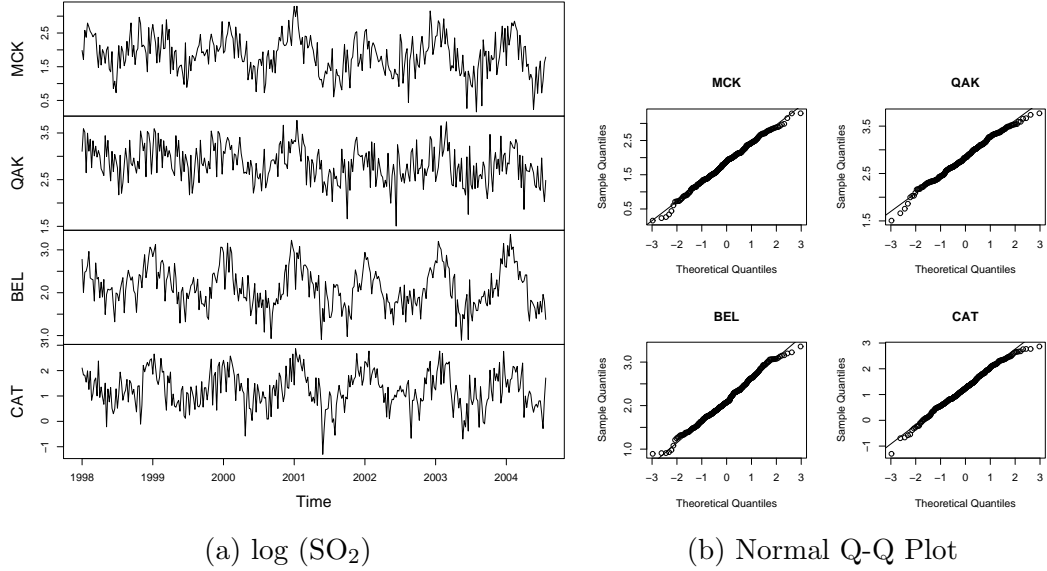


Figura 4.2: *Dados CASTNet:* (a) Séries temporais semanais do $\log(SO_2)$ nas estações MCK, QAK, BEL e CAT. (b) Q-Q plot normal das séries temporais mostradas em (a).

4.2 Modelagem com cargas estáticas

4.2.1 Modelos utilizados

Quatro classes de modelos fatoriais espaciais dinâmicos são consideradas:

- i)* MFEDS(m, h): modelo fatorial espacial dinâmico sazonal com m fatores comuns e h fatores sazonais. Se $h = 1$ temos que

$$\begin{aligned} y_t &= \beta f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \\ f_t &= \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \\ \beta_{(j)} &= \mu_j^\beta \mathbf{1}_N + \eta_j, \quad \eta_j \sim N(0, \tau_j^2 R_{\phi_j, \kappa}) \end{aligned}$$

onde $t = 1, \dots, T$, $j = 1, \dots, m$, $\beta = (\beta_{(1)}, \dots, \beta_{(m)}, \beta_{(m+1)}, 0_N)$, $f_t = (f_{1t}, \dots, f_{m+2,t})'$,

$$\Gamma = \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} \quad \Gamma_1 = \text{diag}(\gamma_1, \dots, \gamma_m) \quad \text{e} \quad \Gamma_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix}$$

ii) MFED(m)-cov: modelo fatorial espacial dinâmico com m fatores, covariáveis estáticas e sazonalidade comum na média do processo, tal que

$$\begin{aligned} y_t &= X^y \mu_t^y + \beta f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \\ \mu_t^y &= G \mu_{t-1}^y + \nu_t, \quad \nu_t \sim N(0, W) \\ f_t &= \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \\ \beta_{(j)} &= \mu_j^\beta \mathbf{1}_N + \eta_j, \quad \eta_j \sim N(0, \tau_j^2 R_{\phi_j, \kappa}) \end{aligned}$$

onde $X^y = (\mathbf{1}_N, lon, lat, lon^2, lon \times lat, lat^2, \mathbf{1}_N, 0_N)$,

$$\begin{aligned} G &= \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}, \quad G_1 = I_6, \quad G_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix}, \\ W &= \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}, \quad W_1 = \text{diag}(\omega_1, \dots, \omega_6) \text{ e } W_2 = \begin{pmatrix} \omega_7 & \omega_{7,8} \\ \omega_{8,7} & \omega_8 \end{pmatrix}. \end{aligned}$$

iii) MFED(m)-cov-GP: modelo fatorial espacial dinâmico com m fatores, sazonalidade comum na média do processo e covariáveis estáticas na média do PG, tal que

$$\begin{aligned} y_t &= X^y \mu_t^y + \beta f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \\ \mu_t^y &= G \mu_{t-1}^y + \nu_t, \quad \nu_t \sim N(0, W) \\ f_t &= \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \\ \beta_{(j)} &= X^\beta \mu_j^\beta + \eta_j, \quad \eta_j \sim N(0, \tau_j^2 R_{\phi_j, \kappa}) \end{aligned}$$

onde $X^y = (\mathbf{1}_N, \mathbf{1}_N, 0_N)$, $X^\beta = (\mathbf{1}_N, lon, lat, lon^2, lon \times lat, lat^2)$,

$$G = \begin{pmatrix} 1 & 0 \\ 0 & G_2 \end{pmatrix}, \quad G_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix} \text{ e } W = \begin{pmatrix} \varpi_1 & 0 & 0 \\ 0 & \varpi_2 & \varpi_{2,3} \\ 0 & \varpi_{3,2} & \varpi_3 \end{pmatrix}.$$

iv) MFEDS(m, h)-cov-GP: modelo fatorial espacial dinâmico sazonal com m fatores, h fatores sazonais e covariáveis estáticas na média do PG. Se $h = 1$ temos que

$$\begin{aligned} y_t &= \beta f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \\ f_t &= \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \\ \beta_{(j)} &= X^\beta \mu_j^\beta + \eta_j, \quad \eta_j \sim N(0, \tau_j^2 R_{\phi_j, \kappa}) \end{aligned}$$

onde $X^\beta = (1_N, lon, lat, lon^2, lon \times lat, lat^2)$, $f_t = (f_{1t}, \dots, f_{m+2,t})'$,
 $\beta = (\beta_{(1)}, \dots, \beta_{(m)}, \beta_{(m+1)}, 0_N)$,

$$\Gamma = \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix}, \Gamma_1 = \text{diag}(\gamma_1, \dots, \gamma_m) \text{ e } \Gamma_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix}$$

Cada modelo foi testado com no máximo 5 fatores e $h = 1$ componente harmônica com ciclos de 52 semanas como sugerido na análise do periodograma na qual aparecem os valores com maiores picos de sazonalidade (Figura 4.3). Modelos com mais fatores foram analisados, mas os resultados não são reportados, dado que não são estatisticamente relevantes ou significativos. A função de correlação do processo Gaussiano é Matérn para quase todas as especificações propostas, exceto para MFEDS(m, h) a qual também é ajustada com uma função de correlação exponencial.

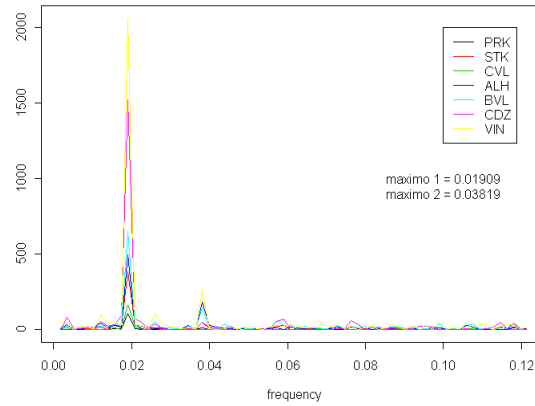


Figura 4.3: *Dados CASTNet:* Periodograma ajustado de algumas séries temporais onde $(\text{máximo}1)^{-1} = 52, 38$.

Adicionalmente, e para propósitos de comparação com os modelos propostos, duas especificações são consideradas. Estas representam modelos mais simples comumente utilizados no ajuste de modelos espaço-temporais (modelos *benchmark*).

i) MGETP: modelo geoestatístico espaço-temporal padrão,

$$y_t = X\mu_t^y + \nu_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2 I_N)$$

$$\mu_t^y = G\mu_{t-1}^y + \omega_t, \quad \omega_t \sim N(0, W)$$

$$\nu_t \sim PG(0, \tau^2 \rho_\phi(\cdot))$$

com $X = (\mathbf{1}_N, lon, lat, lon^2, lon \times lat, lat^2, \mathbf{1}_N, 0_N)$,

$$G = \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}, G_1 = I_6, G_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix},$$

$$W = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}, W_1 = \text{diag}(\omega_1, \dots, \omega_6), W_2 = \begin{pmatrix} \omega_7 & \omega_{7,8} \\ \omega_{8,7} & \omega_8 \end{pmatrix}.$$

ii) MGFP(m): modelo geoestatístico fatorial padrão com m fatores, $y_t = \beta f_t + X \mu_t^y + \nu_t + \epsilon_t$, $\beta_{j,j} = 1$, $\beta_{j,k} = 0$ ($k > j = 1, \dots, m$), μ_t^y , ν_t e ϵ_t como no MGETP.

Nota-se que no MGETP a variação temporal é explicada através de μ_t^{y*} e a variação espacial através de componentes geoestatísticas independentes ν_t ao invés do termo βf_t . MGFP é elaborado a partir de MGETP o qual incorpora fatores dinâmicos. A variação temporal é explicada através de μ_t^{y*} e a variação espacial através de βf_t e μ_t . A diferença com respeito ao modelo proposto é a ausência de dependência espacial nas cargas dos fatores β .

Distribuições a priori

Distribuições a priori relativamente vagas foram utilizadas na maioria dos parâmetros. Mais especificamente, σ 's são $GI(0, 01; 0, 01)$, λ 's são $GI(0, 01; 0, 01)$, γ 's são $Ntr_{(-1,1)}(0, 1)$, Λ 's são $WI(0, 01I_2; 2)$ (matriz de variância do fator sazonal). Adicionalmente, a priori mista para γ com $\alpha = 0, 5$ foi implementada para permitir possíveis fatores não estacionários no modelo. Priors de referência foram utilizadas para os parâmetros do processo Gaussiano com função de correlação exponencial. Para os modelos restantes, o parâmetro de ordem κ da função de correlação Matérn foi escolhido da seguinte forma. Primeiro, um modelo fatorial normal padrão com m fatores é ajustado para $m = 1, 2, 3$. As estimativas das colunas da matriz de cargas do fator são utilizadas para modelar processos Gaussianos com priors de referência. Cada modelo foi testado com valores $\kappa = 1, \dots, 10$ e utilizando o fator de Bayes, o modelo com $\kappa = 7$ foi selecionado na maioria dos casos. (ver Berger et al., 2001, para mais detalhes). Portanto, este valor é assumido em todos os modelos com distribuição de correlação Matérn. A especificação a priori para os parâmetros restantes do processo Gaussiano são: τ 's são $GI(2, 1)$, ϕ 's são $GI(2, b)$ e μ^β 's são $N(a, 1)$ onde $b = \max(dist)/(-2 \log(0, 05))$, $\max(dist)$ representa a maior distância entre os locais observados, e a é o valor absoluto da média das observações.

Esquema MCMC

Para obter amostras a posteriori de todos os parâmetros envolvidos no modelo, o algoritmo MCMC foi rodado. Foram realizadas 50.000 iterações considerando duas cadeias paralelas para cada parâmetro, as primeiras 25.000 foram excluídas como período de aquecimento guardando a cada 10 iterações para remover possíveis autocorrelações. Todas as rotinas computacionais foram escritas na linguagem de programação `0x` versão 3.40 (Doornik, 2002). A convergência das cadeias é verificada visualmente e utilizando a estatística \hat{R} (Gelman & Rubin, 1992), em que valores próximos de 1 indicam convergência.

4.2.2 Principais resultados

Todos os modelos propostos podem ser comparados utilizando diferentes critérios tais como a probabilidade a posteriori do modelo (PPM), a soma do erro quadrático (SEQ) e a soma do erro absoluto (SEA). Adicionalmente, os erros quadráticos médios (EQM), baseados nos valores interpolados e previstos, foram utilizados para fazer comparação de modelos. A Tabela 4.2 mostra os resultados dos critérios mencionados anteriormente para todas as especificações propostas. Nota-se que, entre todos os modelos com maior PPM de cada classe, o modelo MFEDS(4,1)-cov-GP tem a melhor performance em termos da previsão e interpolação. Adicionalmente, a tabela mostra os resultados para os modelos *benchmark*. Os resultados mostram que os modelos padrão têm uma melhor performance em termos de ajuste mas, se mostram pouco eficientes na previsão e interpolação. Estes resultados parecem indicar que a estrutura imposta pelos modelos propostos é, de fato, necessária para melhorar os resultados da previsão. Nota-se que resultados de interpolação não foram reportados para o MGFP, isto devido à ausência de estrutura espacial nas cargas dos fatores.

Avaliações adicionais da performance da **distribuição preditiva** podem ser feitas utilizando os critérios propostos por Gneiting, Balabdaoui & Raftery (2007). Os principais critérios sugeridos estão baseados em medidas de acurácia (*sharpness*) e regras score. A acurácia é avaliada através da amplitude do intervalo de credibilidade da previsão, o menor valor indica o melhor modelo. A Tabela 4.3 e a Figura 4.4 mostram as diferentes medidas de acurácia. Ambos resultados apontam o modelo MFEDS(4,1)cov-GP como o melhor. Regras score foram também consideradas, com este critério é avaliado simultaneamente calibração

Tabela 4.2: *Dados CASTNet:* Criterios de comparação de modelos. Soma do Erro Quadrático - SEQ = $\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \hat{\mu}_{it})^2$; Soma do Erro Absoluto - SEA = $\sum_{i=1}^N \sum_{t=1}^T |y_{it} - \hat{\mu}_{it}|$; Erro quadrático médio da previsão (baseado nas últimas 30 semanas, 2004:01 - 2004:30) - EQM_P = $N^{-1}H^{-1} \sum_{i=1}^N \sum_{h=1}^H (y_{i,T+h} - E(y_{i,T+h}|y))^2$; Erro quadrático médio da interpolação (baseado nas 312 medidas para as estações BWR and SPD) - EQM_I = $N_n^{-1}T^{-1} \sum_{i=1}^{N_n} \sum_{t=1}^T (y_{N+i,t} - E(y_{i,t}|y))^2$, e Probabilidade a Posteriori do Modelo - PPM. Os melhores modelos para cada critério aparecem em *itálico*.

Modelo	m	SEQ	SEA	EQM _P	EQM _I	PPM
MFEDS($m, 1$)-Exp,	1	733,58	1651,8	0,76	0,15	0,43
	2	<i>594,30</i>	<i>1477,5</i>	<i>0,22</i>	0,16	<i>0,56</i>
	3	799,13	1738,2	0,23	<i>0,13</i>	0,01
MFEDS($m, 1$)-Matérn	1	733,66	1651,2	0,58	0,15	0,37
	2	<i>632,45</i>	<i>1539,6</i>	<i>0,25</i>	0,19	<i>0,54</i>
	3	802,20	1734,2	0,27	0,15	0,09
MFED(m)-cov	1	544,63	1418,0	0,368	0,167	0,13
	2	473,29	1306,0	<i>0,237</i>	0,178	0,18
	3	420,12	1217,6	0,249	0,171	0,20
	4	<i>375,10</i>	<i>1150,0</i>	0,271	0,160	<i>0,33</i>
	5	621,26	1486,1	0,245	<i>0,130</i>	0,09
	6	547,83	1412,7	0,251	0,147	0,07
MFED(m)-cov-GP	1	856,25	1811,8	0,384	<i>0,127</i>	0,04
	2	636,97	1549,0	0,638	0,167	0,13
	3	502,54	1348,5	0,308	0,148	0,26
	4	<i>462,57</i>	<i>1276,6</i>	<i>0,260</i>	0,192	<i>0,30</i>
	5	536,13	1425,6	0,498	0,213	0,16
	6	543,13	1415,9	0,304	0,216	0,11
MFEDS($m, 1$)-cov-GP	1	753,86	1673,3	0,651	0,153	0,00
	2	570,13	1450,7	0,288	0,161	0,23
	3	484,78	1320,2	0,255	<i>0,149</i>	0,31
	4	<i>450,95</i>	<i>1276,1</i>	0,229	0,158	<i>0,40</i>
	5	573,15	1446,4	<i>0,218</i>	0,165	0,06
MGETP	-	177,26	883,6	0,341	0,172	-
MGFP(m)	4	264,18	1069,7	0,322	-	-

e acurácia. O escore médio é definido por

$$S(F, y) = N^{-1}H^{-1} \sum_{i=1}^N \sum_{h=1}^H S(F_{i,T+h}, y_{i,T+h})$$

para alguma regra escore própria S . O modelo com menor $S(F, y)$ é considerado o melhor. Em particular, o escore logarítmico (LS) e o escore probabilístico de posto contínuo (CRPS) foram sugeridos por [Gneiting et al. \(2007\)](#), em que:

- LS: é o negativo do logaritmo da densidade preditiva.
- CRPS: para cada $y_{i,T+h}$, o EPPC é definido como

$$CRPS(F_{i,T+h}, y_{i,T+h}) = E_F |\hat{y}_{i,T+h} - y_{i,T+h}| - \frac{1}{2} E_F |\hat{y}_{i,T+h} - \tilde{y}_{i,T+h}|,$$

onde $\hat{y}_{i,T+h}$ e $\tilde{y}_{i,T+h}$ são valores de $p(y_{T+h}|y)$.

Note que a representação do CRPS é conveniente porque $p(y_{T+h}|y)$ é facilmente aproximada por uma amostra a posteriori do algoritmo MCMC (ver [Gschlöbl & Czado, 2005](#) para mais detalhes). Os valores de LS e CRPS para os 5 melhores modelos de cada classe são apresentados na Tabela 4.3, separando os resultados em dois grupos, um para os resultados do modelo proposto e outro para os modelos *benchmark*. Nota-se que a variação dentro de cada grupo é pequena comparada com a variação entre grupos. Este resultado sugere uma certa estabilidade na modelagem proposta ainda entre uma ampla gama de especificações derivadas do MFED. Nota-se que novamente a especificação SDFM(4,1)-cov-GP é apontada como a melhor com menores valores de LS e CRPS e, portanto, a análise dos resultados é baseada nesta especificação.

Tabela 4.3: *Dados CASTNet:* Avaliação da previsão utilizando as medidas: Escore logarítmico médio (LogS), escore probabilístico de posto contínuo (CRPS) e média da amplitude do intervalo de credibilidade de 90% (MA90).

Modelo	LogS	CRPS	MA90
MFEDS(2, 1)-Exp	11,286	0,762	3,517
MFEDS(2, 1)-Matérn	11,567	0,650	2,906
MFED(4)-cov	21,049	1,734	7,486
MFED(4)-cov-GP	11,463	0,813	3,619
MFEDS(4, 1)-cov-GP	10,045	0,644	2,855
MGETP	43,875	2,577	13,066
MGFP(4)	43,814	2,524	12,955

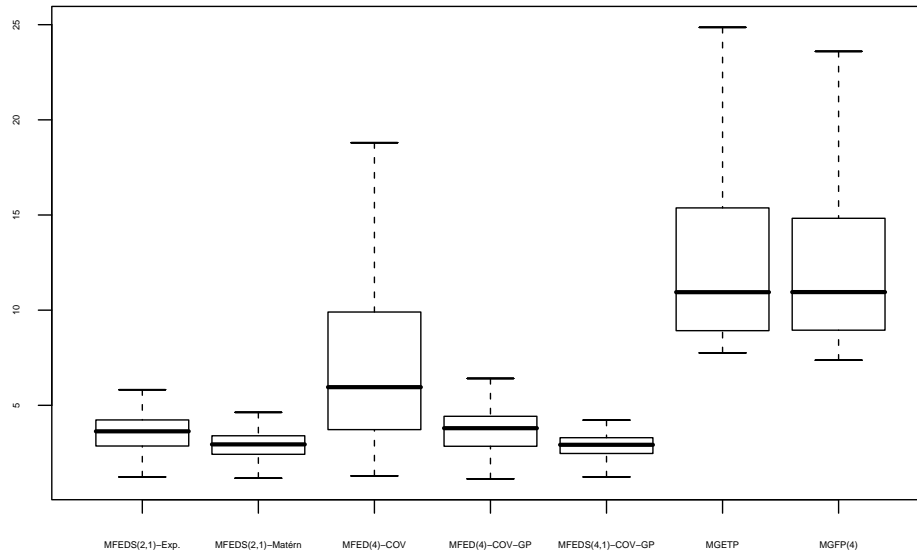


Figura 4.4: *Dados CASTNet*: diagrama de acurácia das previsões semanais do $\log(\text{SO}_2)$ para as especificações MFEDS(2,1)-Exp, MFEDS(2,1)-Matérn, MFED(4)-cov, MFED(4)-cov-GP, MFEDS(4,1)-cov-GP, MGETP e MGFP(4). Os *box plots* estão baseados nas amplitudes do intervalo de credibilidade de 90% dos valores previstos.

A Tabela 4.4 apresenta os resultados a posteriori dos parâmetros que conduzem a variação temporal dos fatores. A Figura 4.14 (no Apêndice deste capítulo) mostra as trajetórias das cadeias dos parâmetros γ_j e λ_j para $j = 1, \dots, 4$. Os resultados mostram uma ampla variedade de dependências auto-regressivas, variando de estruturas com pouca dependência ou ruído branco (fator 1) até estruturas não estacionárias (fator 4). O primeiro fator representa variações comuns em todos os locais, não devendo ser confundido com erros idiossincráticos, que são diferentes para cada local. Nota-se que os outros fatores exibem dependência temporal significativa.

Os fatores podem ser identificados de acordo com a sua contribuição em explicar a variabilidade das observações. Assim, pode-se verificar que em média, a maior proporção explicada pelos fatores está associada ao quarto fator e ao fator sazonal. Estes fatores representam 25% e 15% da variabilidade dos dados respectivamente. O terceiro fator vem depois com aproximadamente 11% seguido do primeiro fator com 4% e do segundo fator com 3%.

A Figura 4.5 mostra as estimativas dos fatores. Nota-se que o quarto fator representa a média sem o efeito da sazonalidade. Em análise fatorial é comum a presença de um fator com esta característica (ver Rencher, 2002 para mais detalhes). Este fator mede a variabilidade da tendência temporal global das séries. Os primeiros três fatores apresentam ruído, embora com variação limitada. O fator sazonal consegue captar os ciclos anuais das séries. O quarto

Tabela 4.4: *Dados CASTNet:* Sumário a posteriori dos parâmetros que caracterizam a dinâmica dos fatores comuns na especificação MFEDS(4,1)-cov-GP. \hat{R} : diagnóstico de Gelman & Rubin.

θ	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			\hat{R}
			2.5%	50%	97.5%	
γ_1	0,009	0,069	-0,122	0,010	0,147	1,01
γ_2	0,186	0,069	0,053	0,185	0,319	1,01
γ_3	0,354	0,091	0,175	0,356	0,522	1,02
γ_4	0,997	0,002	0,992	0,997	1,000	1,30
λ_1	0,005	0,003	0,002	0,004	0,011	1,04
λ_2	0,003	0,001	0,001	0,002	0,005	1,15
λ_3	0,002	0,002	0,001	0,002	0,007	1,26
λ_4	0,002	0,001	0,001	0,002	0,003	1,21
λ_5	0,004	0,003	0,001	0,002	0,012	1,38
λ_6	0,002	0,001	0,001	0,001	0,003	1,02

fator tem um comportamento quase não estacionário, enfatizado pela estimativa da densidade a posteriori de γ_4 concentrada ao redor de 1.

Com a finalidade de verificar a presença de fatores não estacionários, a priori mista para γ proposta na Subseção 2.4.1 foi implementada. Os resultados observados foram $\hat{p}(\gamma_1 = 1|y) = \hat{p}(\gamma_2 = 1|y) = \hat{p}(\gamma_3 = 1|y) = 0$ e $\hat{p}(\gamma_4 = 1|y) = 0.41$. Ou seja, os primeiros três fatores são estacionários enquanto que, o quarto fator é estacionário com 60% de probabilidade a posteriori. Adicionalmente, o modelo MFEDS(4,1)-cov-GP foi ajustado com $\gamma_4 = 1$ e comparado com o modelo com fatores estacionários. Os resultados mostraram valores maiores de SEQ, SAM e EQM baseados na previsão.

A Tabela 4.5 apresenta os resultados a posteriori dos parâmetros da regressão e da dependência temporal das cargas dos fatores. Gráficos das trajetórias de alguns destes parâmetros podem ser encontrados no Apêndice deste capítulo (Figura 4.15). Latitude e longitude são importantes para descrever o nível médio do PG, entretanto, o quadrado e o produto destes valores se mostraram não significativos. Estes resultados indicam que, para cada fator, as médias a posteriori da correlação espacial entre cargas de um par de locais distantes 100 quilômetros são 0,666, 0,703, 0,832, 0,671 e 0,696 respectivamente.

A Figura 4.6 apresenta a superfície média das colunas da matriz de cargas β obtidas via interpolação (krigagem Bayesiana) apresentada na Subseção 2.3.1. As cargas do quarto fator são maiores na parte central da área interpolada, principalmente ao redor da estação QAK, localizada em Ohio. A análise exploratória dos dados indicou que altos níveis de SO₂ foram

Tabela 4.5: *Dados CASTNet:* Sumário a posteriori dos parâmetros que caracterizam as colunas da matriz de cargas dos fatores na especificação MFEDS(4,1)-cov-GP. I.C. denota o intervalo de credibilidade de 95%. \hat{R} : diagnóstico de Gelman & Rubin.

	Coluna 1				Coluna 2			
	Média	Mediana	95% I.C.	\hat{R}	Média	Mediana	95% I.C.	\hat{R}
μ_{1j}^β	1,56	1,56	[0,61;2,49]	1,00	1,55	1,53	[0,62;2,52]	1,00
μ_{2j}^β	0,79	0,78	[0,00;1,71]	1,02	0,72	0,69	[-0,07;1,67]	1,05
μ_{3j}^β	0,92	0,91	[0,04;1,85]	1,01	1,06	1,05	[0,24;1,97]	1,07
μ_{4j}^β	-0,12	-0,12	[-0,27;0,01]	1,02	-0,17	-0,16	[-0,33;-0,04]	1,03
μ_{5j}^β	0,07	0,07	[-0,17;0,32]	1,00	0,03	0,03	[-0,18;0,24]	1,02
μ_{6j}^β	-0,31	-0,30	[-0,55;-0,08]	1,00	-0,14	-0,14	[-0,35;0,04]	1,07
τ_j^2	6,18	5,08	[1,52;18,20]	1,01	3,96	3,17	[0,96;11,00]	1,12
ϕ_j	32,70	32,32	[21,47;46,97]	1,00	35,51	35,06	[22,53;51,21]	1,03
	Coluna 3				Coluna 4			
	Média	Mediana	95% I.C.	\hat{R}	Média	Mediana	95% I.C.	\hat{R}
μ_{1j}^β	1,74	1,75	[0,77;2,68]	1,01	1,32	1,32	[0,62;2,00]	1,06
μ_{2j}^β	1,07	1,06	[0,14;2,08]	1,01	0,20	0,18	[-0,15;0,61]	1,02
μ_{3j}^β	1,04	1,08	[-0,30;2,17]	1,11	0,76	0,76	[0,31;1,27]	1,15
μ_{4j}^β	-0,16	-0,15	[-0,35;0,00]	1,01	-0,04	-0,04	[-0,09;0,00]	1,02
μ_{5j}^β	0,13	0,12	[-0,15;0,45]	1,00	0,03	0,03	[-0,03;0,11]	1,01
μ_{6j}^β	-0,16	-0,16	[-0,41;0,08]	1,00	-0,12	-0,11	[-0,22;-0,03]	1,05
τ_j^2	14,10	8,41	[0,91;48,80]	1,00	0,43	0,36	[0,16;1,07]	1,08
ϕ_j	52,79	54,38	[24,09;77,46]	1,19	33,04	32,87	[21,17;46,82]	1,05
	Coluna 5							
	Média	Mediana	95% I.C.	\hat{R}				
μ_{1j}^β	1,62	1,62	[0,85;2,36]	1,08				
μ_{2j}^β	0,09	0,06	[-0,25;0,59]	1,01				
μ_{3j}^β	0,19	0,17	[-0,26;0,75]	1,03				
μ_{4j}^β	0,00	0,00	[-0,06;0,03]	1,01				
μ_{5j}^β	0,00	0,00	[-0,07;0,07]	1,00				
μ_{6j}^β	-0,03	-0,02	[-0,12;0,05]	1,01				
τ_j^2	0,45	0,35	[0,14;1,29]	1,00				
ϕ_j	35,14	34,73	[21,22;51,95]	1,01				

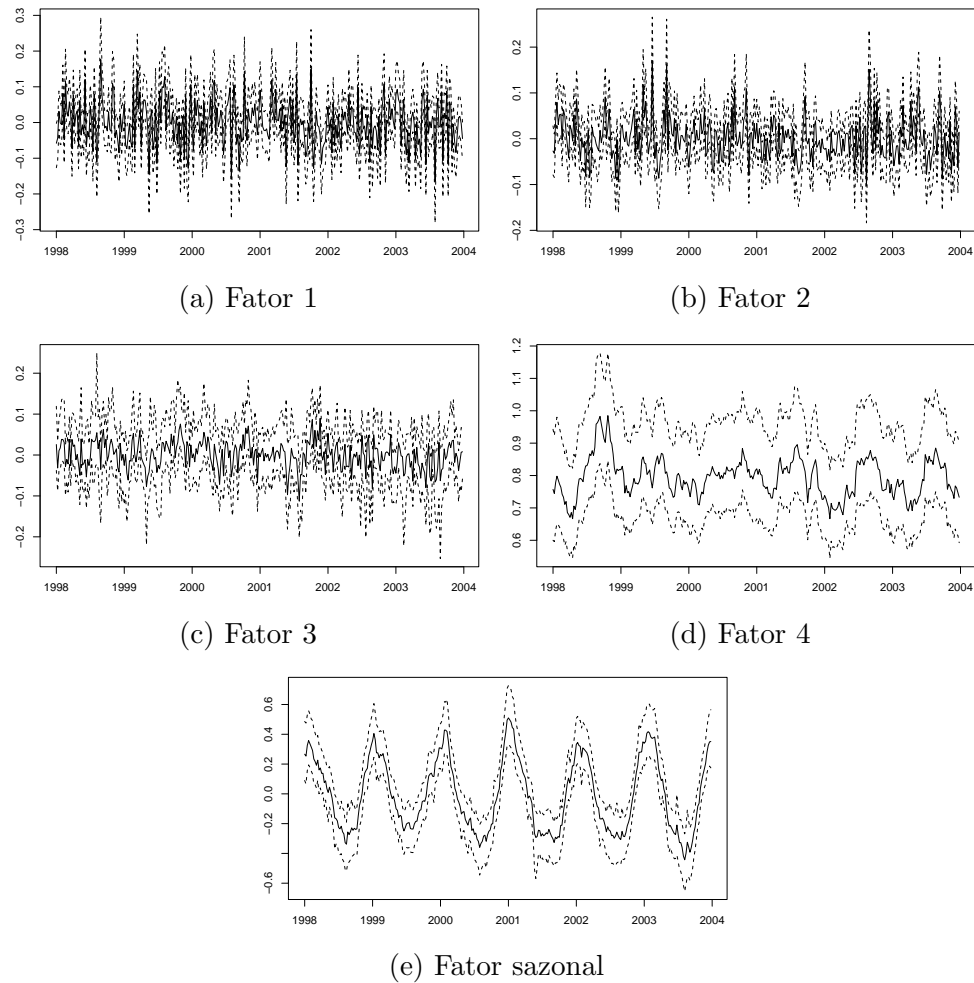


Figura 4.5: *Dados CASTNet:* Médias a posteriori dos fatores para a especificação MFEDS(4,1)-cov-GP. A linha cheia representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

medidos nesta estação. Este fato confirma a indicação do quarto fator como valor médio comum. Por outro lado, as cargas do fator sazonal são menores ao redor do industrializado estado de Ohio, sugerindo um comportamento cíclico menos evidente com altos níveis de SO_2 ao longo do ano (ver Figura 4.2a para mais detalhes). As cargas do primeiro e terceiro fatores parecem ser mais altas na porção sudoeste da área de estudo, enquanto que o segundo fator indica uma divisão entre leste e oeste com altos valores na parte oeste. A combinação de características temporais, representadas nos fatores comuns, e as características espaciais das colunas da matriz de cargas do fator, representa uma das principais características da modelagem proposta inerente da análise fatorial tradicional.

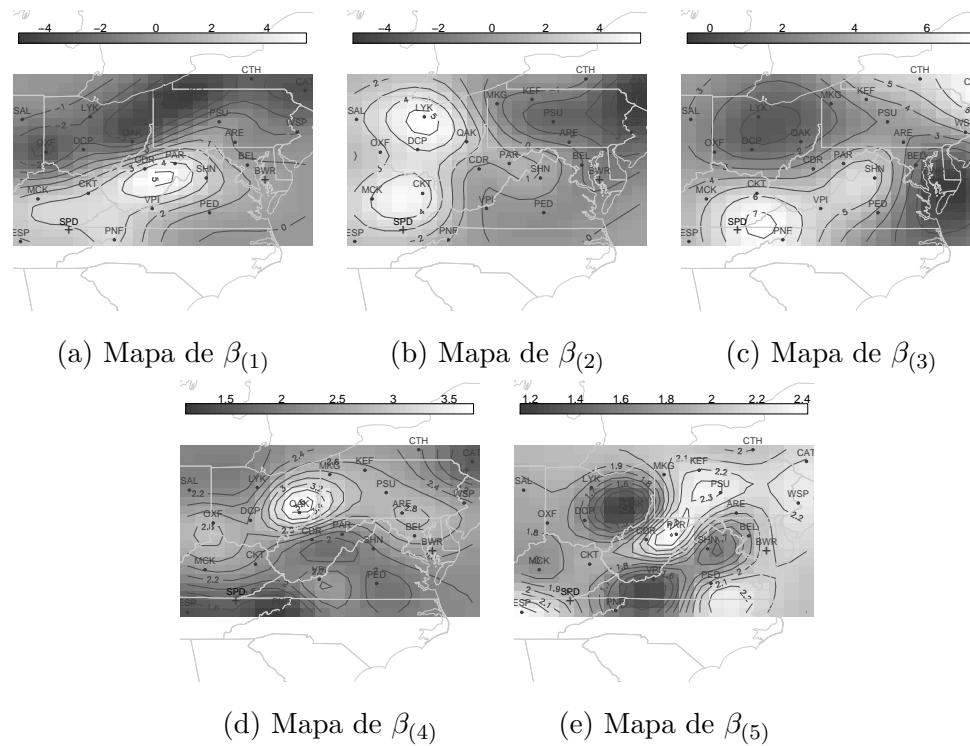


Figura 4.6: *Dados CASTNet:* Interpolação Bayesiana das cargas dos fatores. Os valores acima dos mapas representam a amplitude da variação das médias a posteriori.

Resultados de interpolação e previsão são apresentados nas Figuras 4.7 e 4.8, as quais exibem estimativas acuradas dos valores previstos e interpolados. Nota-se que os intervalos de credibilidade de 95%, tanto das previsões como das interpolações, não são simétricos. Este fato era esperado dado que ambas estimativas foram calculadas utilizando a transformação logaritmo dos dados. O procedimento de interpolação produz um bom ajuste em ambas estações, devido principalmente à presença de estrutura espacial entre as estações monitoradoras. Os valores previstos seguem a mesma tendência dos valores observados embora apresentem uma variação mais suave com intervalos de credibilidade maiores.

Finalmente, a Figura 4.9 apresenta a superfície média dos níveis de SO₂ para 9 semanas do ano 2003. Nota-se que algumas regiões do mapa são mais afetadas pelo fator sazonal com altos níveis em toda a região no início e no final do ano. A metade superior da área de estudo com variação temporal na direção leste-oeste define a região com altos níveis de SO₂ ao longo do tempo.

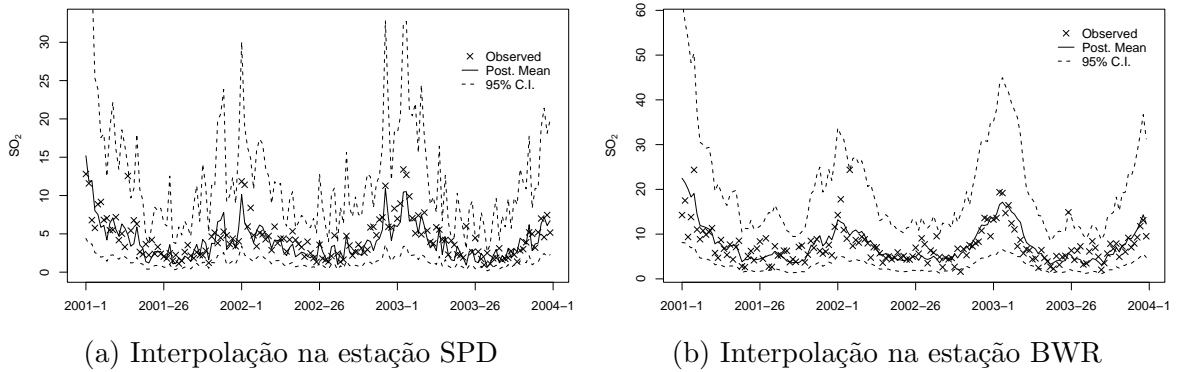


Figura 4.7: *Dados CASTNet:* Valores interpolados nas estações SPD e BWR deixadas de fora da análise. \times representam os valores observados. A linha cheia representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95% da especificação MFEDS(4,1)-cov-GP.

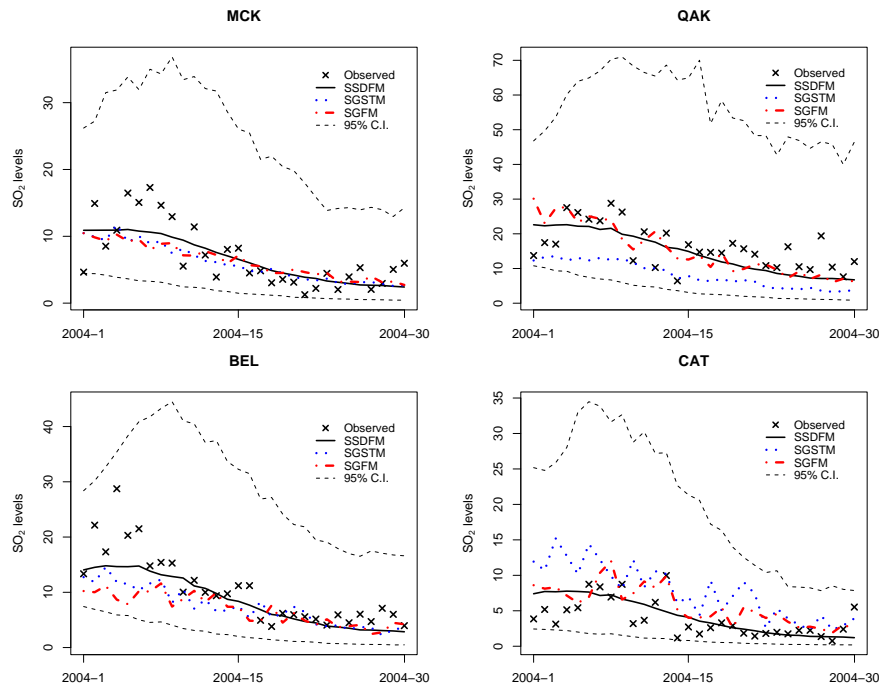


Figura 4.8: *Dados CASTNet:* Valores previstos no período 2004:1–2004:30. As linhas cheia, pontilhada (azul) e tracejada (vermelha) representam as médias a posteriori das especificações MFEDS(4,1)-cov-GP, MGETP e MGFP(4) respectivamente. As linhas tracejadas pretas representam os intervalos de credibilidade de 95% da especificação MFEDS(4,1)-cov-GP, \times os valores observados e as linhas azuis e vermelhas os valores previstos com os modelos *benchmark*.

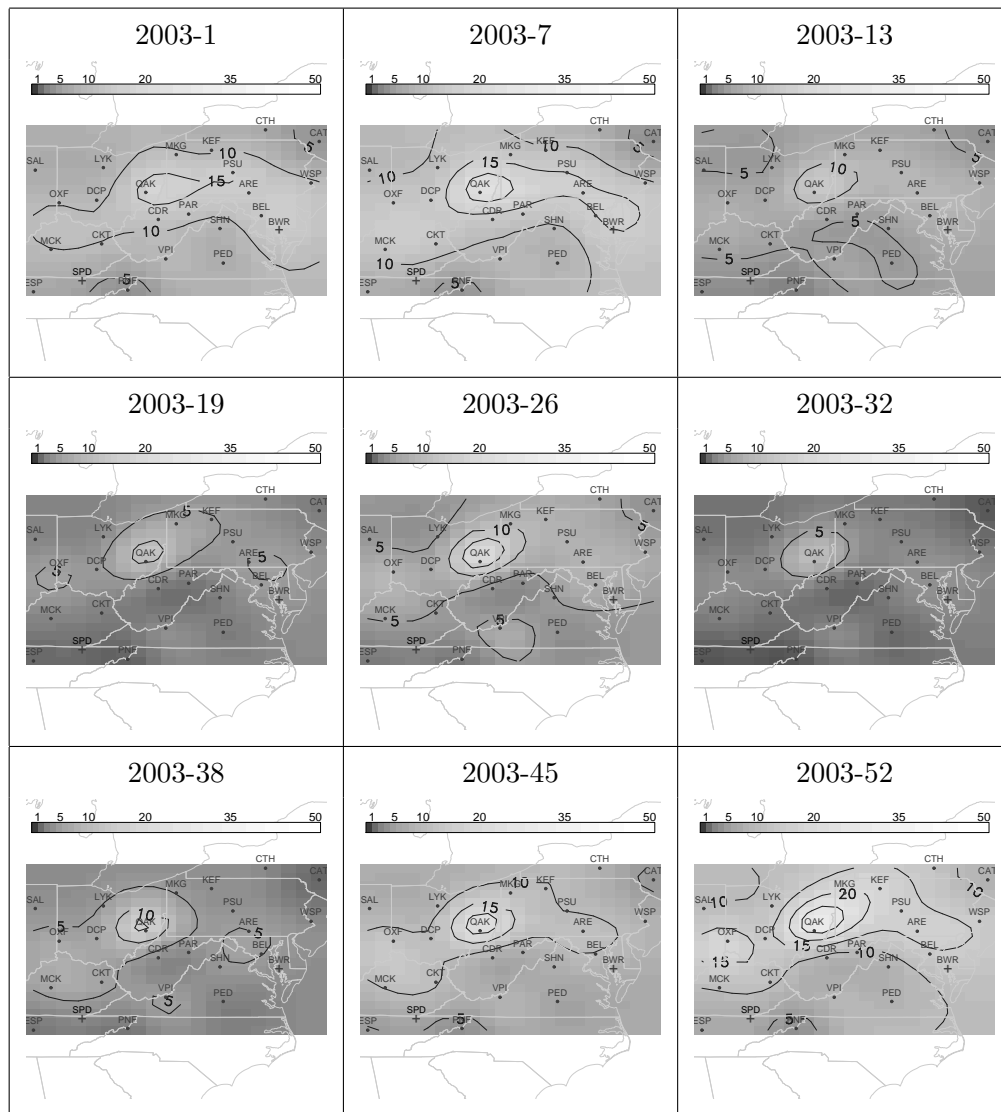


Figura 4.9: *Dados CASTNet:* Mapas das concentrações de SO₂ utilizando a especificação MFEDS(4,1)-cov-GP.

4.3 Modelagem com cargas dinâmicas

4.3.1 Modelos utilizados

Duas classes baseadas no MFED com cargas dinâmicas são consideradas:

- i)* MFEDcd(m)-cov: modelo fatorial espacial com cargas dinâmicas com m fatores e covariáveis estáticas e sazonalidade comum na média do processo, tal que:

$$\begin{aligned} y_t &= X\mu_t^y + \beta_t f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \\ \mu_t^y &= G\mu_{t-1}^y + \nu_t, \quad \nu_t \sim N(0, W) \\ f_t &= \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \\ \beta_{(j)t} &= \delta_{(j)t} + \mu_{j,t} \mathbf{1}_N \\ \delta_{(j)t} &= \delta_{(j)t-1} + \eta_{(j)t}, \quad \eta_{(j)t} \sim N(0, \tau_j^2 R_{\phi_j, \kappa}) \\ \mu_{j,t} &= \mu_{j,t-1} + \eta_{jt}^\mu, \quad \eta_{jt}^\mu \sim N(0, \sigma_{\mu_j}^2) \end{aligned}$$

onde $X = (\mathbf{1}_N, lon, lat, lon^2, lon \times lat, lat^2, \mathbf{1}_N, 0_N)$,

$$\begin{aligned} G &= \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}, \quad G_1 = I_6, \quad G_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix}, \\ W &= \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}, \quad W_1 = \text{diag}(\omega_1, \dots, \omega_6) \text{ e } W_2 = \begin{pmatrix} \omega_7 & \omega_{7,8} \\ \omega_{8,7} & \omega_8 \end{pmatrix}. \end{aligned}$$

- ii)* MFEDScd(m, h)-cov: modelo fatorial espacial com cargas dinâmicas com m fatores, h fatores sazonais e covariáveis estáticas na média do processo. Se $h = 1$ temos que: $X = (\mathbf{1}_N, lon, lat, lon^2, lon \times lat, lat^2)$, $G = I_6$, $W = \text{diag}(\omega_1, \dots, \omega_6)$, $f_t = (f_{1t}, \dots, f_{m+2,t})'$, $\beta_t = (\beta_{(1)t}, \dots, \beta_{(m+1)t}, 0)$,

$$\Gamma = \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix}, \quad \Gamma_1 = \text{diag}(\gamma_1, \dots, \gamma_m), \quad \Gamma_2 = \begin{pmatrix} \cos(2\pi/52) & \sin(2\pi/52) \\ -\sin(2\pi/52) & \cos(2\pi/52) \end{pmatrix}.$$

Analogamente à abordagem anterior, cada modelo foi testado com no máximo 5 fatores e $h = 1$ componente harmônica com ciclos de 52 semanas. A função de correlação para o processo Gaussiano é Matérn com parâmetro de ordem $\kappa = 7$ considerado no caso de cargas estáticas.

4.3.2 Distribuições a priori

O modelo é completado com a especificação das seguintes distribuições a priori:

$$\begin{aligned}\delta_{(j)0} &\sim N(0, \tau_{j0}^2 R_{\phi_{j0}, \kappa}) \\ \mu_{j,0} &\sim N(m_0^\mu, C_0^\mu) \\ f_0 &\sim N(m_0, C_0)\end{aligned}$$

onde $j = 1, \dots, m$, ϕ_{j0} e τ_{j0}^2 são fixados nas médias a posteriori estimadas no modelo com cargas estáticas, f_0 é $N(0, 1_m)$, $\mu_{j,0}$'s são $N(0, 1)$. Para o resto de parâmetros foram utilizadas prioris relativamente vagas. Mais especificamente, σ 's são $GI(0, 01; 0, 01)$, λ 's são $GI(0, 01; 0.01)$, Λ 's são $WI(0, 01I_2; 2)$. Para todas as classes especificadas, a priori mista para γ com $\alpha = 5$ foi utilizada, permitindo desta forma a inclusão de fatores não estacionários. Para o processo Gaussiano, prioris relativamente vagas foram utilizadas, mais especificamente, τ 's são $GI(2, 1)$ e ϕ 's são $GI(2, b)$ onde $b = \max(dist)/(-2 \log(0, 05))$.

Esquema MCMC

O algoritmo MCMC foi utilizado seguindo as mesmas especificações da abordagem anterior. Técnicas para reduzir o tempo computacional foram implementadas como, por exemplo, a decomposição espectral para inverter matrizes e o filtro de informação ([Anderson & Moore, 1979](#)) utilizado na amostragem das cargas dinâmicas.

4.3.3 Principais resultados

A comparação entre modelos é feita utilizando os mesmos critérios da análise anterior. A Tabela 4.6 mostra estas medidas para cada modelo proposto. Nota-se que entre os modelos com maior PPM a especificação MFEDcd(3)-cov apresenta os melhores resultados tanto no ajuste quanto na previsão e interpolação.

Os seguintes resultados estão baseados na especificação MFEDcd(3)-cov. A Figura 4.10 mostra o comportamento dos coeficientes das covariáveis ao longo do tempo. Note que a primeira componente representa a tendência comum das observações, indicando uma mudança no nível médio no final do ano 2001. Os coeficientes da latitude indicam a influência positiva desta variável na média. A latitude ao quadrado tem influência negativa ao longo do tempo. As outras medidas relacionadas a latitude e longitude são não significativas em quase todo o

Tabela 4.6: *Dados CASTNet:* Critérios de comparação de modelos: Soma do Erro Quadrático (SEQ), Soma do Erro Absoluto (SEA), Erro Quadrático Médio da previsão (EQM_P), EQM da interpolação (EQM_I) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *itálico*.

Modelo	m	SEQ	SEA	EQM _P	EQM _I	PPM
MFEDcd(m)-cov	1	358,90	1129,4	0,274	0,176	0,08
	2	322,16	1118,3	<i>0,243</i>	0,172	0,19
	3	<i>291,95</i>	<i>1046,5</i>	0,265	0,169	<i>0,38</i>
	4	299,36	1059,9	0,278	<i>0,164</i>	0,26
	5	302,80	1063,9	0,270	0,185	0,09
MFEDScd($m, 1$)-cov	1	475,90	1288,4	0,347	0,187	0,04
	2	388,12	1169,3	0,302	<i>0,169</i>	0,21
	3	<i>329,99</i>	<i>1104,5</i>	0,287	0,179	<i>0,41</i>
	4	345,09	1155,0	0,293	0,170	0,25
	5	385,10	1200,3	0,299	0,176	0,09

período. O último gráfico, representando a sazonalidade comum, exibe um padrão bastante regular com valores altos no início de cada ano.

As probabilidades a posteriori dos parâmetros auto-regressivos serem iguais a 1 foram $\hat{p}(\gamma_1 = 1|y) = 0$, $\hat{p}(\gamma_2 = 1|y) = 0.983$ e $\hat{p}(\gamma_3 = 1|y) = 0$, ou seja, o segundo fator é considerado não estacionário. A Tabela 4.7 mostra o sumário a posteriori dos parâmetros dos fatores e do processo Gaussiano. Nota-se que a dependência auto-regressiva do terceiro fator é bastante baixa enquanto que no primeiro fator a dependência é quase nula (não significativa). O segundo fator consegue captar uma certa estrutura sazonal e uma tendência decrescente até o final do ano 2001 (Figura 4.11). Este resultado sugere que o padrão sazonal comum estimado não consegue explicar totalmente o comportamento cíclico presente nos dados. Nota-se que este fator pode ser comparado com o quarto fator da abordagem anterior representando neste caso o fator de médias. O primeiro fator pode ser considerado como fator de ruído. O terceiro fator exibe um comportamento bastante variável nos anos 1998, 2000 e 2001, embora no resto do período exiba uma variação limitada.

A Figura 4.12 mostra as estimativas da componente puramente temporal da matriz de cargas dos fatores. A primeira série exibe um comportamento quase constante ao longo

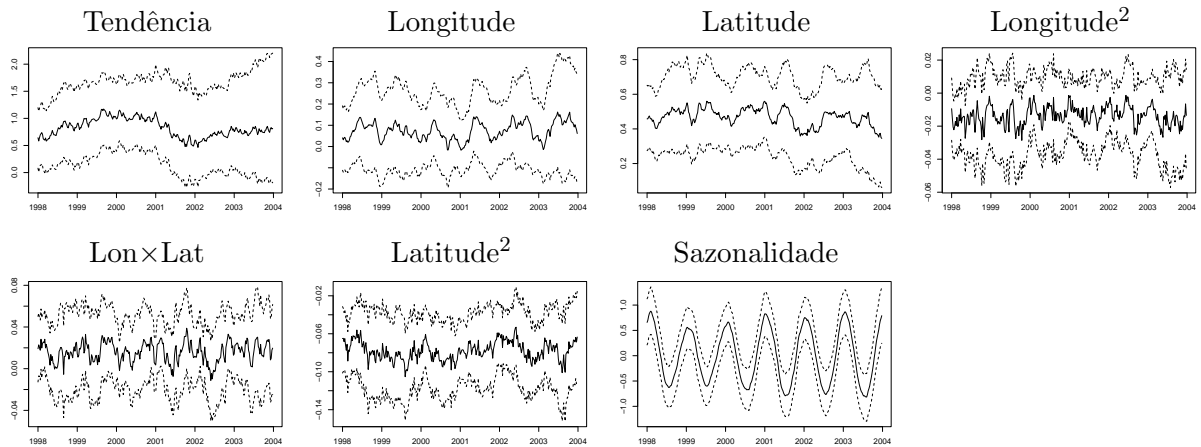


Figura 4.10: *Dados CASTNet:* Médias a posteriori dos coeficientes do nível médio para a especificação MFEDcd(3)-cov. A linha cheia representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

Tabela 4.7: *Dados CASTNet:* Sumário a posteriori dos parâmetros na especificação MFEDcd(3)-cov. \hat{R} : diagnóstico de Gelman & Rubin.

θ	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			\hat{R}
			2.5%	50%	97.5%	
γ_1	-0,040	0,088	-0,203	-0,041	0,142	1,03
γ_2^*	1,000	-	-	-	-	-
γ_3	0,100	0,094	0,038	0,097	0,279	1,04
λ_1	0,003	9e-4	0,001	0,003	0,005	1,04
λ_2	7e-4	2e-4	5e-4	7e-4	0,001	1,03
λ_3	0,004	0,002	0,002	0,004	0,008	1,08
$\sigma_{u_1}^2$	0,028	0,034	0,003	0,018	0,126	1,10
$\sigma_{u_1}^2$	0,054	0,051	0,004	0,037	0,182	1,05
$\sigma_{u_1}^2$	0,063	0,090	0,005	0,032	0,325	1,08
τ_1^2	0,105	0,030	0,057	0,100	0,175	1,03
τ_2^2	0,052	0,009	0,036	0,051	0,073	1,01
τ_3^2	0,279	0,120	0,132	0,249	0,610	1,01
ϕ_1	43,200	7,400	30,200	42,700	60,400	1,02
ϕ_2	43,100	4,300	34,900	43,200	51,200	1,01
ϕ_3	39,600	4,400	31,400	39,800	47,700	1,01

* Fator não estacionário com $\hat{p}(\gamma_2 = 1|y) = 0,98$.

dos anos, enquanto que a segunda e terceira séries variam ao longo do tempo. Estimativas da componente espaço-temporal de β_t foram calculadas para todas as semanas, a Figura 4.13 mostra as superfícies médias estimadas para alguma delas. Os mapas indicam pouca variação no padrão espacial ao longo do tempo. Nota-se que as superfícies de $\delta_{(1)t}$ e $\delta_{(2)t}$

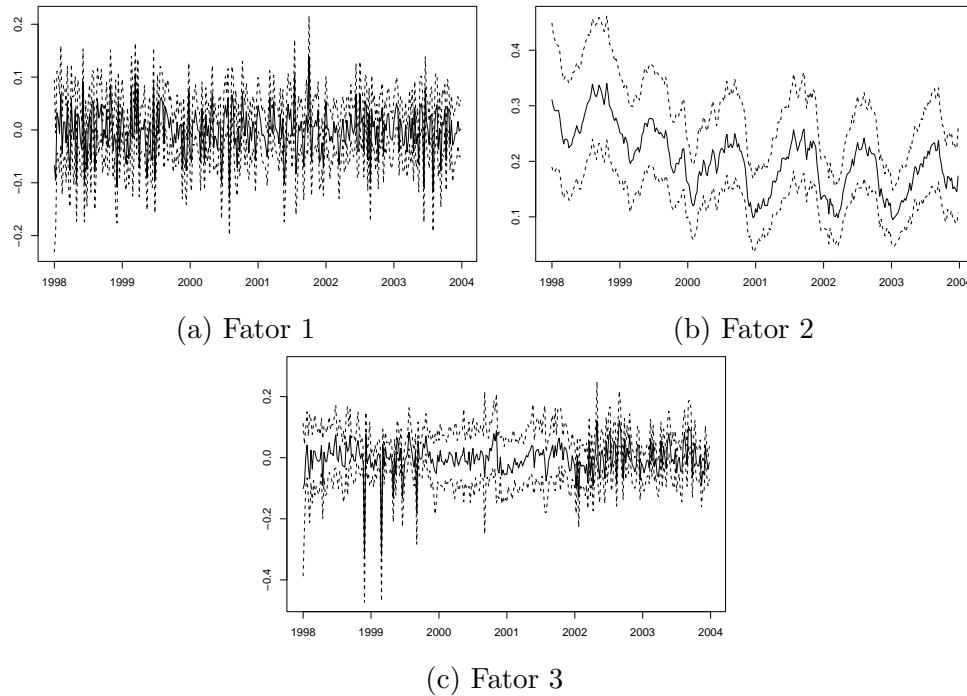


Figura 4.11: *Dados CASTNet:* Médias a posteriori dos fatores para a especificação MFEDcd(3)-cov. A linha cheia representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

são bastante parecidas, indicando um peso maior ao redor da estação QAK. As superfícies de $\delta_{(3)t}$ são bastante parecidas com a superfície de β_2 do modelo MFEDS(4,1)-cov-GP que separa a região em dois grandes setores, leste com pesos mais baixos e oeste com pesos mais altos.

4.4 Comparação entre as duas abordagens

A comparação é feita entre os melhores modelos selecionados de cada abordagem considerando critérios baseados no ajuste, previsão e interpolação. Da Tabela 4.8 concluímos que a especificação MFEDcd(3)-cov apresenta os melhores resultados em termos da previsão e interpolação, embora o ajuste não seja o melhor.

Comparações adicionais foram feitas ao considerar a performance preditiva de cada modelo utilizando os critérios propostos por Gneiting et al. (2007). Da Tabela 4.9 pode-se concluir que o modelo MFEDS(4,1)-cov-GP é o melhor. Com estes resultados concluímos que o modelo com cargas estáticas pode ser utilizado na modelagem dos dados de SO₂ tanto para

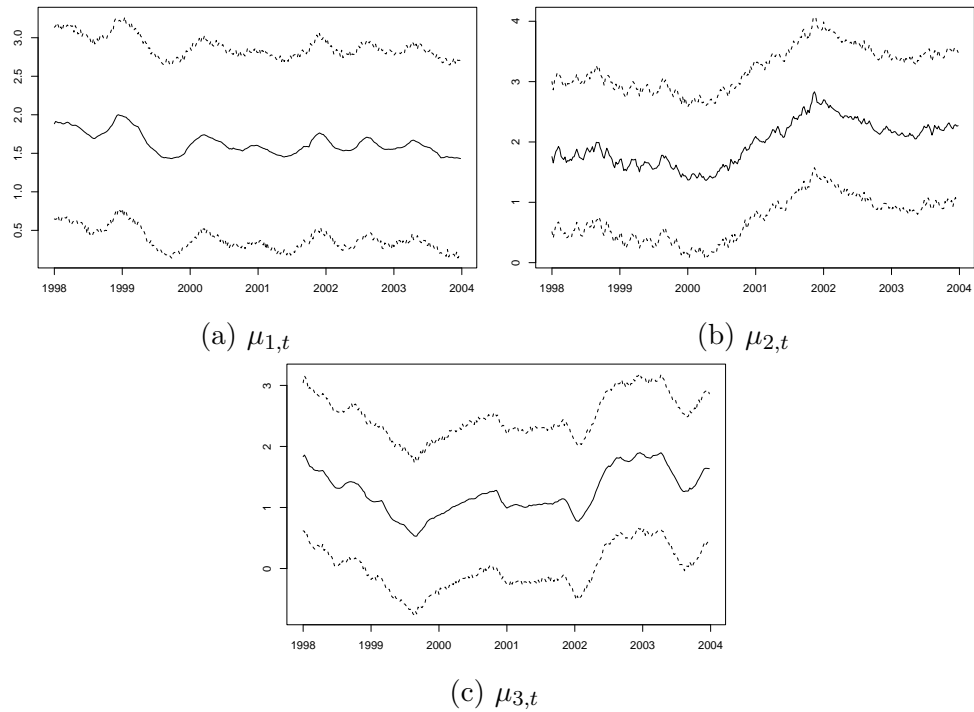


Figura 4.12: *Dados CASTNet:* Médias a posteriori da componente temporal de β_t para a especificação MFEDcd(3)-cov. A linha cheia representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%. $t = 1, \dots, 312$.

Tabela 4.8: *Dados CASTNet:* Critérios de comparação entre modelos com cargas estáticas e dinâmicas. Soma do Erro Quadrático (SEQ), Soma do Erro Absoluto (SEA), Erro Quadrático Médio da previsão (EQM_P) e EQM da interpolação (EQM_I).

Modelo	SEQ	SEA	EQM _P	EQM _I
MFEDS(4, 1)-cov-GP	450,95	1276,1	0,229	0,158
MFEDcd(3)-cov	291,95	1046,5	0,265	0,169

prever observações futuras, quanto para interpolar valores em locais não medidos. A inclusão de uma estrutura temporal nas cargas dos fatores não ajudou a melhorar os valores das previsões, mas ainda, incorporou variabilidade extra pelo fato de ser mais parametrizado.

Tabela 4.9: *Dados CASTNet:* Avaliação da previsão utilizando as medidas: Escore logarítmico médio (LogS), escore probabilístico de posto contínuo (CRPS) e média da amplitude do intervalo de credibilidade de 90% (MA90).

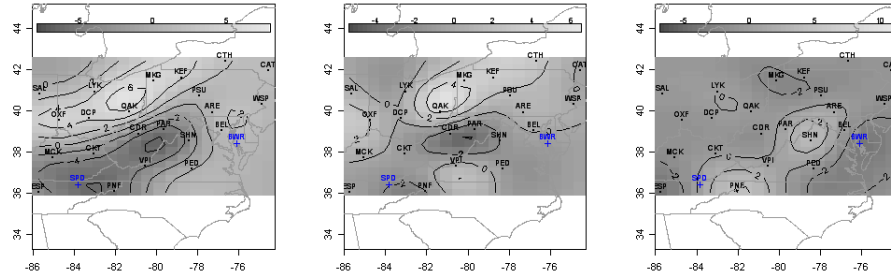
Modelo	LogS	CRPS	MA90
MFEDS(4, 1)-cov-GP	10,045	0,644	2,855
MFEDcd(3)-cov	12,568	0.919	4,166

4.5 Considerações finais

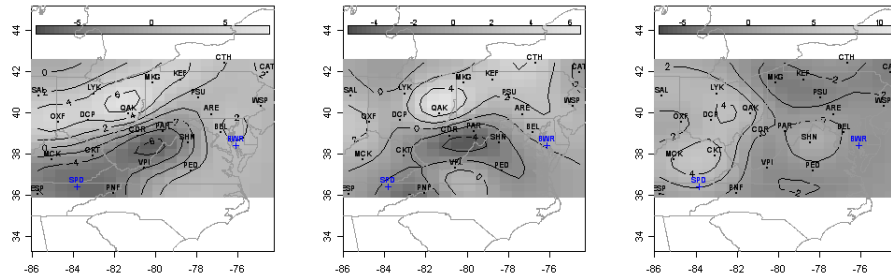
Neste capítulo duas abordagens para modelar os níveis de SO₂ na região leste dos EUA foram apresentadas. As duas abordagens foram baseadas nos modelos propostos nos Capítulos 2 e 3 com cargas estáticas e dinâmicas, respectivamente. Várias classes de modelos foram propostas em cada caso e comparadas com critérios baseados no ajuste, previsão e interpolação. Modelos com estruturas mais simples (modelos *benchmark*) também foram considerados, com o objetivo de avaliar se modelos menos estruturados eram suficientes para tratar a estrutura espaço-temporal presente nos dados. Os resultados mostraram que os modelos propostos se adaptaram bem à dinâmica dos dados, embora as medidas de ajuste resultaram melhores para os modelos mais simples. Comparações baseadas na previsão e interpolação indicaram que uma estrutura como a incorporada no modelo proposto é necessária, principalmente para fazer previsão.

Foram feitas comparações entres os melhores modelos de cada classe e entre modelos com cargas estáticas e dinâmicas. Para esta última comparação obtivemos que o modelo com cargas dinâmicas foi melhor em termos de ajuste, mas não em termos de interpolação e previsão. Comparações adicionais foram feitas utilizando critérios baseados nos valores previstos como o escore logarítmico médio e o critério CRPS proposto por [Gneiting et al. \(2007\)](#). Com estas medidas concluímos que o MFED com cargas estáticas é o mais indicado para representar a estrutura dos dados.

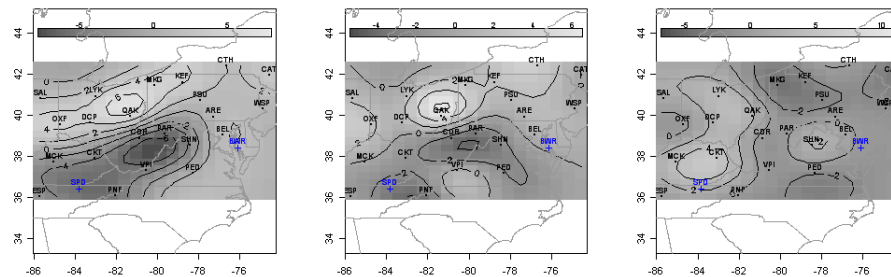
2002-1:



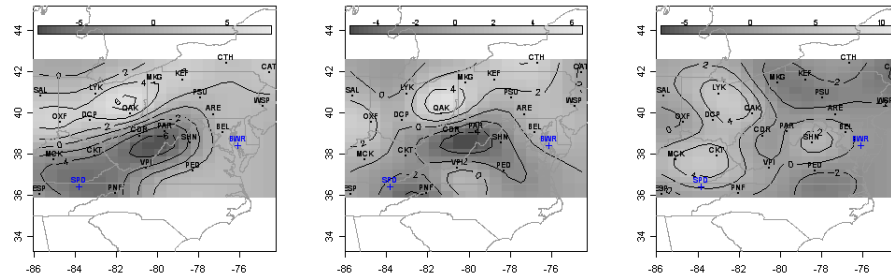
2003-1:



2003-26:



2003-52:



$\beta_{(1)t}$

$\beta_{(2)t}$

$\beta_{(3)t}$

Figura 4.13: *Dados CASTNet:* Interpolação Bayesiana da componente espaço-temporal de β_t nas semanas 2002-1, 2003-1, 2003-26 e 2003-52. Os valores acima dos mapas representam a amplitude da variação das médias a posteriori.

4.6 Apêndice

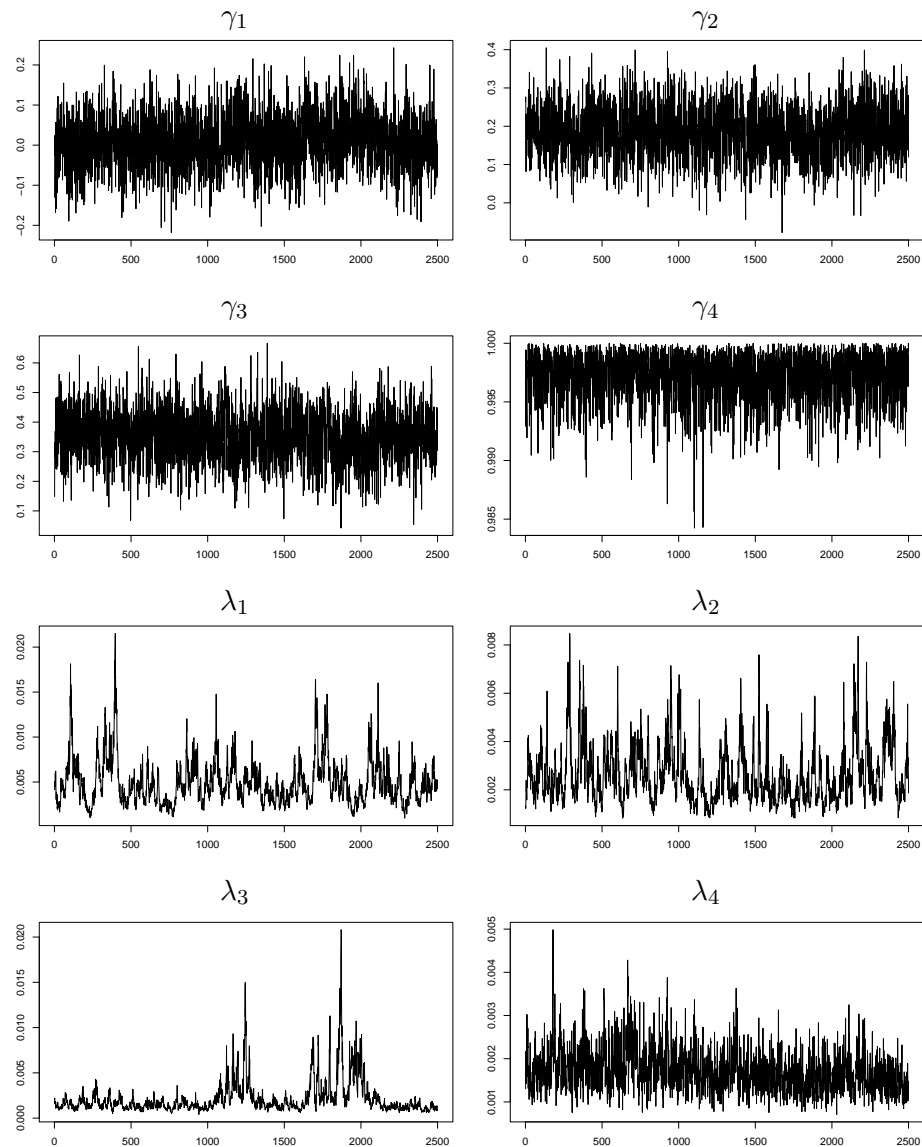


Figura 4.14: *Dados CASTNet:* Trajetória das cadeias dos parâmetros γ_j e λ_j ($j = 1, \dots, 4$). Tamanho das cadeias: 2.500

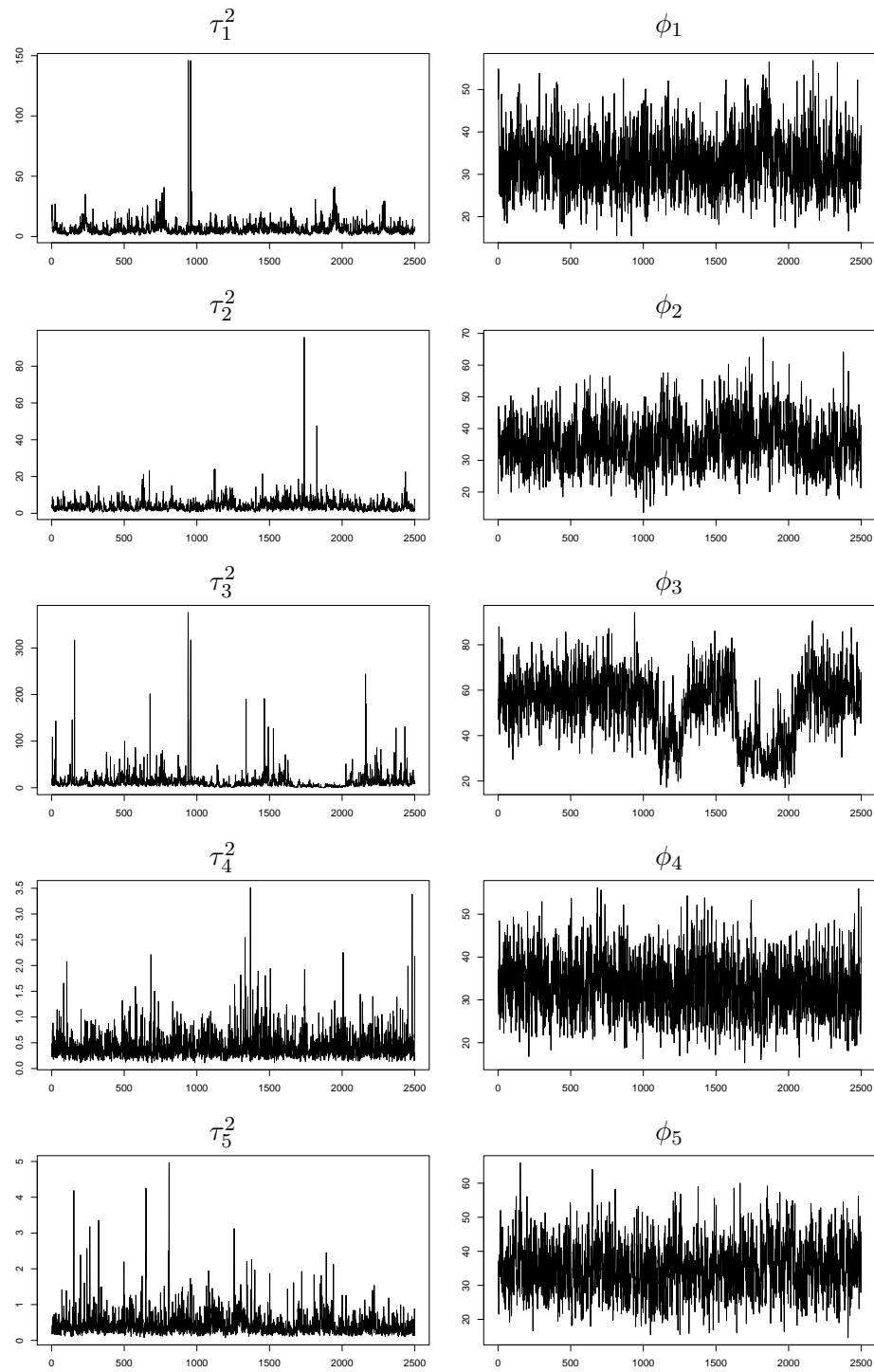


Figura 4.15: *Dados CASTNet:* Trajetória das cadeias dos parâmetros τ_j^2 e ϕ_j ($j = 1, \dots, 5$). Tamanho das cadeias: 2.500

Capítulo 5

Modelo fatorial espacial dinâmico generalizado

Neste capítulo o MFED para dados normais é estendido para dados pertencentes à família exponencial. Este tipo de modelo permite abordar situações em que os dados podem ser estritamente positivos (modelo Gama), binários (modelo Bernoulli), de contagem (modelo Poisson), entre outros. Neste caso, os fatores latentes e as cargas dos fatores são utilizados para modelar transformações da média através de uma função de ligação apropriada. A correlação temporal e espacial é modelada através dos fatores dinâmicos e cargas dos fatores respectivamente. Inferência para este tipo de modelo é proposta seguindo uma abordagem completamente Bayesiana, na qual técnicas como o Filtro de Kalman estendido e a amostragem em blocos dos parâmetros latentes do modelo são desenvolvidas para ser utilizadas dentro do algoritmo MCMC. O esquema proposto é mostrado como uma alternativa para amostrar os parâmetros latentes em modelos lineares dinâmicos generalizados multivariados, combinando as técnicas sugeridas em [Gamerman \(1998\)](#) e [Knorr-Held \(1999\)](#). A primeira delas refere-se à construção de densidades propostas para os passos de Metropolis-Hastings (M-H) e a segunda, a estratégias de amostragem em blocos para os estados do modelo dinâmico não linear. Finalmente, dois exemplos simulados com respostas Gama e Bernoulli são apresentados para mostrar a aplicabilidade dos modelos e algoritmos propostos.

5.1 Introdução

Modelos espaço-temporais não lineares e não Gaussianos vêm sendo amplamente utilizados em diversas aplicações, principalmente na modelagem de dados epidemiológicos, meteorológicos e de contaminação ambiental. Estes modelos permitem encarar problemas estruturalmente

mais complexos, nos quais as observações medidas no espaço pertencem à família exponencial, em particular dados Poisson, Gama e Binomial.

O MLDG tem sido utilizado para modelar dados variando no espaço e no tempo. Isto é possível impondo uma estrutura de correlação espaço-temporal nas componentes do preditor linear. Para dados espaciais univariados ou multivariados, nos quais uma ou mais variáveis são medidas em um local (dados contínuos) ou região (dados de área) em um período de tempo discreto, algumas abordagens recentes podem ser mencionadas. Por exemplo, [Hooten & Wikle \(2007\)](#) modelam a quantidade de uma espécie de pássaro registrada em locais espalhados em todo o território dos EUA através de um processo de Poisson. Já em [Fernandes, Schmidt & Migon \(2007\)](#) é utilizado um MLDG para modelar a quantidade de chuva medida em estações pluviométricas espalhadas na cidade do Rio de Janeiro, onde os processos espaço-temporais são inflacionados de zeros. Para o caso de dados de área, [Jin, Carlin & Banerjee \(2005\)](#) utilizam um modelo CAR multivariado generalizado (do inglês *GMCAR*) para ajustar as taxas de mortalidade por câncer de pulmão e esôfago e, mais recentemente, [Vivar \(2007\)](#) utiliza um MLDG com erros seguindo distribuições de campos aleatórios Markovianos Gaussianos próprios para modelar a presença ou ausência de uma espécie de pássaro nos EUA.

Nesta tese, o interesse é tratar esses tipos de estruturas espaço-temporais, utilizando a modelagem fatorial e a estrutura dos modelos dinâmicos (da mesma forma que no modelo normal) para modelar dados pertencentes à família exponencial. A modelagem proposta se diferencia das abordagens anteriores nos seguintes aspectos: *(i)* a redução da dimensionalidade através dos fatores comuns e *(ii)* a formação de grupos ou clusters na região de estudo. O modelo proposto sob este enfoque é chamado de **modelo fatorial espacial dinâmico generalizado** e denotado por **MFEDG**.

A análise fatorial já vem sendo utilizada na modelagem de dados espaciais. A maioria das abordagens propostas para dados na família exponencial considera o caso em que múltiplas variáveis são medidas em cada local (dados espaciais multivariados). Nestes casos, a análise fatorial é útil para modelar as variáveis observadas em termos de um número menor de variáveis latentes ou fatores assumindo que os fatores comuns são espacialmente correlacionados. [Knorr-Held & Best \(2001\)](#), por exemplo, propuseram um modelo de componentes comuns para o número de casos de duas doenças através de um processo de Poisson. Por outro lado, [Wang & Wall \(2003\)](#) ajustaram um modelo fatorial espacial às taxas de mor-

talidade de câncer de várias cidades de Minnesota. Em ambos os casos, a análise fatorial é utilizada em outro contexto e para um tipo diferente de dado. No MFEDG proposto, a análise fatorial é utilizada para reduzir a dimensionalidade de uma observação medida em vários locais. É essa a principal diferença com as abordagens anteriores.

Por outro lado, fazer inferência em MLDG espaço-temporais não é trivial e requer procedimentos computacionalmente intensivos. Entretanto, com o crescente desenvolvimento da computação e o uso cada vez mais crescente de métodos de amostragem eficientes como o algoritmo MCMC, este problema foi parcialmente resolvido, embora o procedimento de inferência ainda seja bastante demorado. Do ponto de vista Bayesiano, o principal problema está baseado na amostragem dos estados latentes cuja distribuição a posteriori não têm forma analítica fechada. Na literatura existem várias propostas para aproximar a distribuição a posteriori de interesse, a maioria delas para o caso de observações univariadas que podem ser estendidas ao caso multivariado com algumas modificações (ver [Migon, Gamerman, Lopes & Ferreira, 2005](#) para mais detalhes). Uma dessas propostas corresponde a [Ravines \(2007\)](#) na qual a amostragem do vetor de estados é feita em bloco utilizando o algoritmo denominado *CUBS* (do inglês *Conjugate Updating Backward Sampling*). Este algoritmo segue um esquema similar ao FFBS com a diferença de que o primeiro passo do FFBS é substituído pelo *Conjugate Updating* de [West et al. \(1985\)](#). Uma desvantagem deste método é a elicitação de prioris conjugadas multivariadas que faz com que o método seja inviável para o caso multivariado.

Em [Gamerman \(1998\)](#) a amostragem dos estados é feita com passos de Metropolis-Hastings utilizando uma distribuição proposta baseada em um modelo dinâmico normal aproximado. A amostragem pode ser feita de três formas: individualmente, em bloco ou amostrando individualmente dos erros. Por outro lado, [Geweke & Tanizaki \(2001\)](#) propuseram amostrar dos estados individualmente utilizando várias densidades propostas. Em [Knorr-Held \(1999\)](#) a matriz de estados é dividida em blocos permitindo que a amostragem de cada bloco seja feita com passos de Metropolis-Hastings. Nesta última abordagem a distribuição proposta não depende das observações e, portanto, a probabilidade de aceitação do algoritmo de M-H só depende da razão de verossimilhanças.

Alguns comentários podem ser extraídos destas propostas de amostragem. Primeiro, a amostragem individual pode apresentar convergência lenta das cadeias, sendo necessárias cadeias mais longas o que faz que o procedimento de inferência seja demorado. Segundo, a

amostragem conjunta dos estados é uma alternativa interessante. Entretanto, se a dimensão da matriz de estados for muito grande, a taxa de aceitação do algoritmo de M-H pode ser muito baixa ou quase nula. Por último, considerar a amostragem dos estados em blocos se mostra como uma alternativa eficiente, com taxas de aceitação mais razoáveis dependendo do número de blocos considerado. Portanto, nesta tese, uma proposta de amostragem que combina as técnicas sugeridas em [Gamerman \(1998\)](#) e [Knorr-Held \(1999\)](#) é sugerida.

Este capítulo está dividido da seguinte forma. Na Seção 5.2 é apresentado o MFEDG proposto. Na Seção 5.3 é descrito o procedimento de inferência e o esquema de amostragem proposto para os fatores latentes utilizando blocos. Dois estudos simulados para dados com distribuição Gama e Bernoulli são apresentados na Seção 5.4. Por último, considerações finais são apresentadas na Seção 5.5.

5.2 Modelo proposto

Seja N o número de localizações numa determinada área $S(\subset \mathbb{R}^2)$ e suponha que $y_t = (y_{1t}, \dots, y_{Nt})$ é um vetor N -dimensional que contém as observações medidas nos N locais $\{s_1, \dots, s_N\}$ ($s_i \in S, i = 1, \dots, N$) no tempo t . Logo, o modelo fatorial espacial dinâmico generalizado (MFEDG) é representado da seguinte forma:

$$p(y_t | \eta_t, \psi) \propto \exp\{\psi[y_t^T \eta_t - b(\eta_t)] + c(y_t, \psi)\} \quad (5.1a)$$

$$\eta_t = v(\theta_t) \quad (5.1b)$$

$$g(\mu_t) = \theta_t = \beta f_t + \mu_t^{y^*} \quad (5.1c)$$

$$f_t = \Gamma f_{t-1} + w_t, \quad w_t \sim N(0, \Lambda) \quad (5.1d)$$

$$\beta_{(j)} = \mu_j^\beta + \delta_j, \quad \delta_j \sim PG(0, \tau_j^2 \rho_{\phi_j}(\cdot)) \quad (5.1e)$$

$$f_0 \sim N(m_0, C_0)$$

onde η_t é o parâmetro natural e ψ é o parâmetro de dispersão, a média e variância de y_t dependem da função b da seguinte forma $E(y_t | \eta_t) = b'(\eta_t)$ e $V(y_t | \eta_t) = b''(\eta_t)/\psi$. v é uma função que transforma o preditor linear no parâmetro natural. A equação (5.1c) carrega parte da estrutura do MFED normal, $\mu_t^{y^*}$ é uma componente puramente temporal que pode tomar diferentes especificações de acordo ao problema (sazonalidade, tendências, etc) e inclusive considerar a inclusão de covariáveis estáticas e/ou dinâmicas. A matriz Γ descreve a evolução dos fatores latentes enquanto que λ representa a matriz de covariância das evoluções de f_t .

Da mesma forma que no modelo normal, a dependência espacial é modelada através das colunas da matriz de cargas dos fatores que seguem processos Gaussianos com média μ_j^β e função covariância $\tau_j^2 \rho_{\phi_j}(\cdot)$ ($j = 1, \dots, m$).

O modelo proposto permite a modelagem de uma ampla gama de dados provenientes de diversas distribuições pertencentes à família exponencial. Com a capacidade de incluir estruturas simples e/ou complexas, flexíveis para representar muitos fenômenos variando no espaço e no tempo. Dado que o MFEDG é uma generalização do MFED, todas as propriedades e características apresentadas no Capítulo 2 podem ser estendidas ao modelo não normal. Assim, pode-se afirmar que a função de covariância do processo espaço-temporal tem estrutura não separável. A inclusão de covariáveis é feita através das cargas e/ou a componente $\mu_t^{y^*}$ da mesma forma que na Subseção 2.2.1. Tendências mais específicas como padrões sazonais ao longo de tempo podem ser incorporadas de duas formas: (i) em $\mu_t^{y^*}$, em que a componente sazonal é comum em toda a região de estudo e (ii) através dos fatores dinâmicos. Neste último caso o fator sazonal é ponderado pelas cargas espacialmente estruturadas o que significa a incorporação de diferentes padrões sazonais para os diferentes locais no espaço (ver mais detalhes na Subseção 2.2.2).

5.2.1 Função de verossimilhança

Sem perda de generalidade é assumido que $\mu_t^{y^*} = 0$. Sejam $y = (y_1, \dots, y_T)'$, $\eta = (\eta_1, \dots, \eta_T)$, $\Theta^* = (\lambda, \gamma, \mu, \tau, \phi)$ e $\Psi = (F, \beta, \Theta^*)$. A função de verossimilhança de (Ψ, η, ψ, m) é

$$\begin{aligned} p(y|\Psi, \eta, \psi, m) &= \prod_{t=1}^T p(y_t|\Psi, \eta_t, \psi, m) \\ &= \prod_{t=1}^T p(y_t|\eta_t, \psi) \\ &= \exp \left\{ \psi \left[\sum_{t=1}^T y_t^T \eta_t - \sum_{t=1}^T b(\eta_t) \right] + \sum_{t=1}^T c(y_t, \psi) \right\} \end{aligned} \quad (5.2)$$

onde η_t é o parâmetro natural ou canônico e $\eta_t = v(\theta_t) = v(\beta f_t)$.

5.2.2 Interpolação

O procedimento de interpolação é feito seguindo as mesmas idéias apresentadas no modelo normal (Seção 2.3). O interesse é interpolar valores para N_n localizações onde a variável resposta y_t não foi observada. Seja y^o o vetor de observações nas localizações em S e y^n o vetor de observações não medidas em $S_n = \{s_{N+1}, \dots, s_{N+N_n}\}$. Analogamente, sejam β^o e β^n as matrizes de cargas dos fatores correspondentes as observações y^o e y^n respectivamente.

Do ponto de vista Bayesiano, o interesse é encontrar a distribuição a posteriori de y^n dada por

$$p(y^n|y^o) = \int p(y^n|y^o, \beta^n, f, \Theta) p(\beta^n|\beta^o, \Theta) p(\beta^o, f, \Theta|y^o) d\beta^n d\beta^o df d\Theta \quad (5.3)$$

onde $p(\beta^n|\beta^o, \Theta) = \prod_{j=1}^m p(\beta_{(j)}|\beta_{(j)}^o, \mu_j^\beta, \tau_j^2, \phi_j)$ e

$$p(\beta_{(j)}|\beta_{(j)}^o, \mu_j^\beta, \tau_j^2, \phi_j) = N(\mu_j^\beta + R_{\phi_j}^{n,o} R_{\phi_j}^o{}^{-1} (\beta_{(j)}^o - \mu_j^\beta); \tau_j^2 (R_{\phi_j}^n - R_{\phi_j}^{n,o} R_{\phi_j}^o{}^{-1} R_{\phi_j}^{o,n}))$$

com $R_{\phi_j}^o$ e $R_{\phi_j}^{n,o}$ definidos na Seção 2.3. A densidade $p(y^n|y^o)$ é aproximada utilizando os métodos de Monte Carlo de acordo com a seguinte formula

$$\hat{p}(y^n|y^o) = \frac{1}{L} \sum_{l=1}^L p(y^n|y^o, \beta^{n(l)}, f^{(l)}, \Theta^{(l)})$$

onde $\{(\beta^{n(1)}, f^{(1)}, \Theta^{(1)}), \dots, (\beta^{n(L)}, f^{(L)}, \Theta^{(L)})\}$ é uma amostra obtida do algoritmo MCMC.

Mais especificamente, para cada tempo t e para cada l -ésima iteração temos que

$$p(y_t^n|y_t^o, \beta^{n(l)}, f_t^{(l)}, \Theta^{(l)}) = p(y_t^n|\eta_t^{n(l)})$$

onde $\eta_t^{n(l)}$ é o parâmetro natural da distribuição correspondente às N_n localizações em S_n , $\eta_t^{n(l)} = v(\theta_t^{n(l)}) = v(\beta^{n(l)} f_t^{(l)} + \mu_t^{y^*(l)})$ e $\theta_t^{n(l)}$ é o preditor linear associado a $\eta_t^{n(l)}$.

5.2.3 Previsão

Analogamente ao caso normal, previsão h -passos a frente é feita seguindo o mesmo procedimento apresentado na Seção 2.3. Neste caso, uma aproximação da densidade preditiva h -passos a frente $p(y_{T+h}|y)$ é dada pela seguinte fórmula

$$\hat{p}(y_{T+h}|y) = \frac{1}{L} \sum_{l=1}^L p(y_{T+h}|f_{T+h}^{(l)}, \beta^{(l)}, \Theta^{(l)}).$$

Mais especificamente temos que para cada l -ésima iteração do algoritmo MCMC

$$p(y_{T+h}|f_{T+h}^{(l)}, \beta^{(l)}, \Theta^{(l)}) = p(y_{T+h}|\eta_{T+h}^{(l)})$$

onde $\eta_{T+h}^{(l)} = v(\theta_{T+h}^{(l)}) = v(\beta^{(l)} f_{T+h}^{(l)} + \mu_{T+h}^{y^*(l)})$ e, portanto, uma amostra $\{y_{T+h}^{(1)}, \dots, y_{T+h}^{(L)}\}$ de $p(y_{T+h}|y)$ é obtida amostrando $y_{T+h}^{(l)}$ de $p(y_{T+h}|\eta_{T+h}^{(l)})$ para $l = 1, \dots, L$.

5.3 Procedimento de inferência

Nesta seção é apresentado o procedimento de inferência proposto seguindo uma abordagem completamente Bayesiana. Para isso, distribuições a priori para todos os parâmetros latentes e hiperparâmetros são especificadas e o algoritmo MCMC novamente é utilizado para obter amostras a posteriori dos parâmetros de interesse combinando técnicas como o amostrador de Gibbs e o algoritmo de Metropolis-Hastings. Em seguida, um esquema de amostragem em blocos para os fatores latentes é proposto utilizando numa primeira etapa o filtro de Kalman estendido a fim de obter uma aproximação linear para a equação de observação e numa segunda etapa o FFBS (do inglês *Forward Filtering Backward Sampling*) de [Frühwirth-Schnatter \(1994\)](#) e [Carter & Kohn \(1994\)](#). Esta proposta é apresentada como uma alternativa na amostragem dos estados latentes em MLDG dentro do algoritmo MCMC.

5.3.1 Distribuições a priori

O modelo em (5.1) é completado com a especificação das distribuições a priori de todos os parâmetros. A priori para os fatores comuns é dada pela equação (5.1d) e completada com a priori para o estado inicial $f_0 \sim N(m_0, C_0)$ com m_0 e C_0 conhecidos. As distribuições a priori para os hiperparâmetros λ_j e γ_j ($j = 1, \dots, m$) são as mesmas apresentadas na Seção 2.4.1 com distribuição Gama Inversa para λ_j e uma priori mista para γ_j em que fatores não estacionários podem ser considerados.

Para os parâmetros do processo Gaussiano μ_j^β , ϕ_j e τ_j^2 ($j = 1, \dots, m$) prioris vagas foram especificadas tal que $\mu_j^\beta \sim N(m_\mu, S_\mu)$, $\tau_j^2 \sim GI(n_\tau/2, n_\tau s_\tau/2)$ e $\phi_j \sim GI(2, b)$. m_μ , S_μ , n_τ e s_τ são hiperparâmetros conhecidos e $b = \max(dist)/(-2 \log(0.05))$ onde $\max(dist)$ representa a máxima distância entre dois locais. Neste caso, as prioris de referência propostas por [Berger et al. \(2001\)](#) consideradas para o caso normal não foram utilizadas. Testes preliminares com dados simulados indicaram problemas de convergência nas cadeias. Notou-se também que a função de verossimilhança não fornecia suficiente informação para estimar os parâmetros e, portanto, prioris mais informativas tiveram que ser consideradas.

5.3.2 Inferência a posteriori

Um esquema de amostragem baseado no algoritmo MCMC é utilizado para estimar todos os parâmetros do modelo. Isto é feito combinando a função de verossimilhança com a informação

a priori considerando o número de fatores m conhecido. Sem perda de generalidade, considere $\mu_t^{y*} = 0 \forall t$. Assim, dado m , a distribuição a posteriori conjunta de (F, β, Θ) é

$$p(F, \beta, \Theta|y) \propto \prod_{t=1}^T p(y_t|\eta_t, F, \beta)p(f_0|m_0, C_0) \prod_{t=1}^T p(f_t|f_{t-1}, \lambda, \gamma) \\ \times \prod_{j=1}^m p(\beta_{(j)}|\mu_j^\beta, \tau_j^2, \phi_j)p(\gamma_j)p(\lambda_j)p(\mu_j^\beta)p(\tau_j^2)p(\phi_j). \quad (5.4)$$

A distribuição acima não tem forma fechada conhecida e é analiticamente intratável, portanto, o algoritmo MCMC será utilizado para amostrar os parâmetros de interesse. Neste ponto o interesse é propor um algoritmo que permita amostrar os fatores latentes, a matriz de cargas dos fatores e os hiperparâmetros Θ a partir das suas respectivas condicionais completas. Para os hiperparâmetros em Θ as distribuições condicionais completas são: normal para γ_j e μ_j^β , gama inversa para λ_j e τ_j^2 ($j = 1, \dots, m$). O parâmetro ϕ_j é amostrado com passos de Metropolis-Hastings com densidade proposta seguindo uma distribuição log-normal com parâmetro de locação $\log \phi_j$ e de escala Δ_ϕ , ou seja $q_j(\phi_j, \cdot) = f_{LN}(\cdot; \log \phi_j, \Delta_\phi)$ onde Δ_ϕ é uma constante sintonizadora utilizada para calibrar a densidade proposta. Dado F e β , as distribuições condicionais completas a posteriori de todos os hiperparâmetros em Θ são as mesmas que no MFED. Mais detalhes podem ser encontrados no Apêndice A.1.

A amostragem da matriz de cargas β é feita conjuntamente utilizando passos de Metropolis-Hastings. Para isso, reescrevemos as equações (5.1c) e (5.1e) da seguinte forma $\theta_t = f_t^* \beta^*$ e $\beta^* \sim N(\mu_{\beta^*}, \Sigma_{\beta^*})$, com $f_t^* = f_t' \otimes I_N$, $\beta^* = (\beta'_{(1)}, \dots, \beta'_{(m)})'$, $\mu_{\beta^*} = \mu \otimes 1_N$, $\Sigma_{\beta^*} = \Sigma_\beta \otimes R_\phi$ e $\Sigma_\beta = \text{diag}(\tau_1^2, \dots, \tau_m^2)$. A distribuição proposta é uma normal multivariada com média igual ao valor de β^* da iteração corrente e variância $\Delta_\beta I_{Nm}$ (I_{Nm} matriz identidade de dimensão Nm) onde Δ_β é uma constante sintonizadora conveniente para garantir taxas de aceitação razoáveis. Assim, um valor candidato $\tilde{\beta}^*$ é gerado da distribuição $N(\beta^*, \Delta_\beta I_{Nm})$ e aceito com probabilidade

$$\alpha(\beta^*, \tilde{\beta}^*) = \min \left\{ 1, \frac{\prod_{t=1}^T p(y_t|f_t^*, \tilde{\beta}^*) f_N(\tilde{\beta}^*; \mu_{\beta^*}, \Sigma_{\beta^*})}{\prod_{t=1}^T p(y_t|f_t^*, \beta^*) f_N(\beta^*; \mu_{\beta^*}, \Sigma_{\beta^*})} \right\}.$$

5.3.3 Esquema de amostragem proposto

A condicional completa dos fatores comuns não é analiticamente tratável e requer procedimentos computacionais mais elaborados. Assim, um esquema de amostragem em blocos com passos de Metropolis-Hastings é apresentado. O algoritmo proposto combina as idéias apresentadas em [Gamerman \(1998\)](#) e [Knorr-Held \(1999\)](#) referidas à construção de distribuições

propostas e à amostragem em blocos respectivamente. A idéia de sugerir um esquema em blocos surge da necessidade de obter taxas de aceitação razoáveis para o algoritmo de M-H. A amostragem individual mostrou-se pouco eficiente com alta correlação entre os f_t 's e apresentou convergência lenta das cadeias. Na literatura, a amostragem conjunta dos estados no MLDG se mostra como uma alternativa eficiente com melhoras na convergência das cadeias comparada ao caso de movimentos individuais. Entretanto, esta abordagem não garante taxas de aceitação razoáveis reportando-se valores muito baixos quando a dimensão do bloco F é muito alta. Assim, um ponto intermediário às amostragens individual e conjunta é a amostragem de F em blocos proposta nesta tese.

Especificação dos blocos

Seja $F = (f_1, \dots, f_T)'$ a matriz de dimensão $T \times m$ representando o bloco completo dos fatores comuns ao longo do tempo. A idéia geral do esquema de amostragem proposto é dividir a matriz F em B blocos tal que

$$F = \begin{pmatrix} F_1 \\ \vdots \\ F_B \end{pmatrix}$$

onde cada bloco F_b ($b = 1, \dots, B$) tem dimensão $T_b \times m$ e $T = B T_b$. Sejam L_1^b e L_2^b o limite inferior e superior do bloco b tal que $L_1^1 = 1$, $L_2^B = T$ e $T_b = L_2^b - L_1^b + 1$. A representação do b -ésimo bloco F_b é dada por

$$F_b = (f_{L_1^b}, \dots, f_{L_2^b})' = f_{L_1^b, L_2^b}.$$

Assim, a **condicional completa de F_b é dada pela seguinte fórmula**

$$p(f_{L_1^b, L_2^b} | y, \Theta, f_{1, L_1^b-1}, f_{L_2^b+1, T}) \propto \prod_{t=L_1^b}^{L_2^b} p(y_t | f_t, \Theta) p(f_{L_1^b, L_2^b} | \Theta, f_{1, L_1^b-1}, f_{L_2^b+1, T}). \quad (5.5)$$

A fim de aproximar a moda a posteriori de F_b , a função resposta h é linearizada e a equação de observação é aproximada por um modelo Gaussiano, este método é conhecido como Filtro de Kalman estendido (KFE) (ver [Anderson & Moore \(1979\)](#) para mais detalhes). Em seguida uma amostra de F_b é obtida a partir do algoritmo FFBS que consiste em amostrar os fatores de forma retrospectiva de $t = L_2^b$ até $t = L_1^b$. A densidade proposta para o algoritmo de Metropolis-Hastings é construída a partir desta aproximação como proposto em [Gamerman \(1998\)](#).

Distribuição condicional em priors auto-regressivas

A distribuição conjunta da priori auto-regressiva $p(F)$ é especificada por

$$p(F) \propto \exp \left\{ \frac{1}{2} \sum_{t=2}^T (f_t - \Gamma f_{t-1})' \Lambda^{-1} (f_t - \Gamma f_{t-1}) \right\}.$$

Seja \tilde{F} o vetor empilhado de F tal que $\tilde{F} = (f'_1, \dots, f'_T)'$ é de dimensão $mT \times 1$. Com esta representação a distribuição conjunta de F pode ser escrita como

$$p(F) \propto \exp \left\{ \frac{1}{2} \tilde{F}' K \tilde{F} \right\}$$

onde K é chamada de matriz de penalidades (Green, 1987) ou matriz de precisão. Nota-se que K é bem definida embora a matriz K^{-1} não exista devido a K não ter posto completo.

A matriz de precisão K tem papel fundamental na derivação de uma distribuição condicional para um sub-bloco de F . Para a definição dessa distribuição condicional definimos a matriz G de dimensão $m(T-1) \times mT$ como

$$G = \begin{pmatrix} -\Gamma & I_m & 0 & \dots & 0 \\ 0 & -\Gamma & I_m & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -\Gamma & I_m \end{pmatrix}$$

e uma matriz bloco-diagonal Q de dimensão $m(T-1) \times m(T-1)$ com elementos Λ tal que $Q = \text{diag}(\Lambda, \dots, \Lambda)$. Com estas matrizes é possível demonstrar que

$$\tilde{F}' K \tilde{F} = \tilde{F}' G' Q^{-1} G \tilde{F}$$

onde $K = G' Q^{-1} G$ é simétrica e de dimensão $mT \times mT$. Utilizando estes resultados, a distribuição condicional $p(f_{L_1^b, L_2^b} | \Theta, f_{1, L_1^b-1}, f_{L_2^b+1, T})$, que faz parte da equação (5.5), pode ser derivada.

Seja $K_{L_1^b, L_2^b}$ uma sub-matriz de K localizada entre as linhas e colunas $m(L_1^b - 1) + 1$ e mL_2^b associada aos elementos do bloco F_b . Além disso, sejam K_{1, L_1^b} e $K_{L_2^b, T}$ as matrizes localizadas à esquerda e à direita de $K_{L_1^b, L_2^b}$ tal que

$$K = \begin{pmatrix} \dots & K'_{1, L_1^b} & \dots \\ K_{1, L_1^b} & K_{L_1^b, L_2^b} & K_{L_2^b, T} \\ \dots & K'_{L_1^b+1, T} & \dots \end{pmatrix}.$$

Logo a distribuição condicional de $f_{L_1^b, L_2^b}$ dados f_{1, L_1^b-1} , $f_{L_2^b+1, T}$ e Θ é

$$p(f_{L_1^b, L_2^b} | \Theta, f_{1, L_1^b-1}, f_{L_2^b+1, T}) = N(\mu_{L_1^b, L_2^b}, \Sigma_{L_1^b, L_2^b}) \quad (5.6)$$

onde

$$\mu_{L_1^b, L_2^b} = \begin{cases} -K_{L_1^b, L_2^b}^{-1} K_{L_2^b+1, T} f_{L_2^b+1, T} & b = 1 \\ -K_{L_1^b, L_2^b}^{-1} K_{1, L_2^b-1} f_{1, L_1^b-1} & b = T \\ -K_{L_1^b, L_2^b}^{-1} (K_{1, L_2^b-1} f_{1, L_1^b-1} + K_{L_2^b+1, T} f_{L_1^b+1, T}) & \text{c.c.} \end{cases},$$

e $\Sigma_{L_1^b, L_2^b} = K_{L_1^b, L_2^b}^{-1}$. Mais detalhes e a derivação do resultado anterior podem ser encontradas em [Knorr-Held \(1999\)](#) assim como a generalização para prioris auto-regressivas de ordem k .

Filtro de Kalman estendido

O objetivo é aproximar a moda a posteriori m_t e a variância C_t da posteriori $p(F|y)$. A moda a posteriori maximiza

$$\log p(F, y) = \sum_{t=1}^T \log p(y_t | f_t) + \sum_{t=1}^T \log p(f_t | f_{t-1}) + \log p(f_0).$$

A derivada da log-verossimilhança da observação dados os estados é

$$\frac{\partial \log p(y_t | f_t)}{\partial f_t} = \{y_t - b'(\eta_t)\} \frac{1}{b''(\eta_t)} \frac{\partial h}{\partial \theta_t} \frac{\partial \theta_t}{\partial f_t} \\ \beta h'(\theta_t) \Sigma_t^{-1}(\theta_t) \{y_t - h(\theta)\}.$$

Se a função resposta h fosse linear poderia ser aplicado o Filtro de Kalman diretamente para calcular m_t e C_t mas não é o caso. Logo, a equação de observação é aproximada com um modelo Gaussiano. Isto é feito linearizando a função resposta h ao redor de um ponto $\hat{\theta}_t$. A média é aproximada por $\mu_t = h(\theta_t) \approx h(\hat{\theta}_t) + h'(\hat{\theta}_t)(\theta_t - \hat{\theta}_t)$ e variância por $\hat{\Sigma}_t = \Sigma_t(\hat{\theta}_t)$.

Usando esta aproximação tem-se que

$$\frac{\partial \log p(y_t | f_t)}{\partial f_t} \approx \beta h'(\hat{\theta}_t) \hat{\Sigma}_t^{-1} [y_t - h(\hat{\theta}_t) + h'(\hat{\theta}_t) \hat{\theta}_t - h'(\hat{\theta}_t) \theta_t] \\ = \beta \underbrace{h'(\hat{\theta}_t) \hat{\Sigma}_t^{-1} h'(\hat{\theta}_t)}_{\hat{V}_t^{-1}} \underbrace{[h'(\hat{\theta}_t)^{-1} (y_t - h(\hat{\theta}_t)) + \hat{\theta}_t - \theta_t]}_{\hat{y}_t}$$

Logo, a aproximação da equação de observação é feita da seguinte forma:

$$\hat{y}_t = \beta f_t + \nu_t, \quad \nu_t \sim N(0, \hat{V}_t) \quad (5.7)$$

onde $\hat{y}_t = h'(\hat{\theta}_t)^{-1}(y_t - h(\hat{\theta}_t)) + \hat{\theta}_t$ e $\hat{V}_t^{-1} = h'(\hat{\theta}_t)\hat{\Sigma}_t^{-1}h'(\hat{\theta}_t)$. Com estes resultados é possível aplicar o Filtro de Kalman diretamente da mesma forma que no caso normal. Logo, a filtragem para cada f_t , $t = L_1^b, \dots, L_2^b$, do bloco F_b é resumida nos seguintes passos:

Algoritmo 5.1. (FKE)

1. **Posteriori em $t - 1$:** $f_{t-1}|D_{t-1} \sim N(m_{t-1}, C_{t-1})$.

2. **Priori em t :** $f_t|D_{t-1} \sim N(a_t, R_t)$ onde $a_t = \Gamma m_{t-1}$ e $R_t = \Gamma C_{t-1} \Gamma' + \Lambda$.

3. **Ponto para a expansão de Taylor:** $\hat{\theta}_t = \beta f_t^{(l-1)}$.

4. **Aproximação da equação de observação:** Determine $\hat{y}_t = h'(\hat{\theta}_t)^{-1}[y_t - h(\hat{\theta}_t)] + \hat{\theta}_t$ e $\hat{V}_t^{-1} = h'(\hat{\theta}_t)\hat{\Sigma}_t^{-1}(\hat{\theta}_t)h'(\hat{\theta}_t)$.

5. **Previsão um passo ao frente:** $\hat{y}_t|D_{t-1} \sim N(f_t^*, Q_t)$ onde

$$f_t^* = \beta a_t \quad e \quad Q_t = \beta R_t \beta' + \hat{V}_t.$$

6. **Posteriori em t :** $f_t|D_t \sim N(m_t, C_t)$ onde

$$m_t = a_t + R_t \beta' Q_t^{-1} (\hat{y}_t - f_t^*) \quad e \quad C_t = R_t - R_t \beta' Q_t^{-1} \beta R_t.$$

$D_t = \{y_1, \dots, y_t\}$ denota o conjunto de observações até o tempo t . Dentro do algoritmo MCMC, se l denota a iteração corrente, o ponto para a expansão de Taylor $\hat{\theta}_t$ depende do valor de $f_t^{(l-1)}$ na iteração anterior. Logo, \hat{y}_t pode ser expressa em função de $f_t^{(l-1)}$ e denotada por $\hat{y}_t = \hat{y}_t(f_t^{(l-1)})$.

Amostragem em blocos

A amostragem do bloco F_b ($b = 1, \dots, B$) é feita utilizando passos de Metropolis-Hastings dentro do algoritmo MCMC. Assim, na l -ésima iteração, um valor candidato \tilde{F}_b é amostrado da densidade proposta $q(F_b^{(l-1)} \rightarrow \tilde{F}_b)$ dada pela condicional completa conjunta aproximada $\hat{p}(\tilde{F}_b) = p(\tilde{F}_b | \hat{y}, \Theta, f_{1, L_1^b-1}, f_{L_2^b+1, T})$. Expressões para $\hat{p}(\tilde{F}_b)$ são obtidas decompondo a condicional completa aproximada da seguinte forma:

- Para $b = 1$

$$\hat{p}(\tilde{F}_1) = p(\tilde{f}_{L_2^1} | f_{L_2^1+1}, \hat{D}_{L_2^1}, \Theta) \prod_{t=1}^{L_2^1-1} p(\tilde{f}_t | \tilde{f}_{t+1}, \hat{D}_t, \Theta).$$

- Para $1 < b < B$

$$\hat{p}(\tilde{F}_b) = p(\tilde{f}_{L_2^b} | \tilde{f}_{L_2^{b+1}}, \hat{D}_{L_2^b}, \Theta) \prod_{t=L_1^b+1}^{L_2^b-1} p(\tilde{f}_t | \tilde{f}_{t+1}, \hat{D}_t, \Theta) p(\tilde{f}_{L_1^b} | \tilde{f}_{L_1^{b+1}}, f_{L_1^{b-1}}, \hat{D}_{L_1^b}, \Theta).$$

- Para $b = B$

$$\hat{p}(\tilde{F}_B) = p(\tilde{f}_T | \hat{D}_T, \Theta) \prod_{t=L_1^B+1}^{L_2^B-1} p(\tilde{f}_t | \tilde{f}_{t+1}, \hat{D}_t, \Theta) p(\tilde{f}_{L_1^B} | \tilde{f}_{L_1^{B+1}}, f_{L_1^{B-1}}, \hat{D}_{L_1^B}, \Theta).$$

Neste caso, $\hat{D}_t = \{\hat{y}_1, \dots, \hat{y}_t\}$ denota o conjunto de observações artificiais até o tempo t definidas em (5.7) com $\hat{y}_t = \hat{y}_t(f_t^{(l-1)})$. Por outro lado, as médias e variâncias das distribuições retrospectivas são obtidas facilmente utilizando as propriedades da normal multivariada. Em geral, pode-ser mostrado que:

- $p(f_t | f_{t+1}, \hat{D}_t, \Theta) = N(m_t^s(f_{t+1}), C_t^s)$, onde

$$\begin{aligned} m_t^s(f_{t+1}) &= m_t + C_t \Gamma' R_{t+1}^{-1} (f_{t+1} - \Gamma m_t), \\ C_t^s &= C_t - C_t \Gamma' R_{t+1}^{-1} \Gamma C_t. \end{aligned}$$

- $p(f_t | f_{t+1}, f_{t-1}, \hat{D}_t, \Theta) = N(m_t^{s*}(f_{t+1}, f_{t-1}), C_t^{s*})$, onde

$$\begin{aligned} m_t^{s*}(f_{t+1}, f_{t-1}) &= m_t^*(f_{t-1}) + C_t^* \Gamma' (\Gamma C_t^* \Gamma' + \Lambda)^{-1} (f_{t+1} - \Gamma m_t^*(f_{t-1})), \\ C_t^{s*} &= C_t^* - C_t^* \Gamma' (\Gamma C_t^* \Gamma' + \Lambda)^{-1} \Gamma C_t^* \end{aligned}$$

tal que

$$\begin{aligned} C_t^{*-1} &= R_t^{-1} \Gamma C_{t-1} C_{t-1}^s C_{t-1} \Gamma' R_t^{-1} + C_t^{-1}, \\ m_t^*(f_{t-1}) &= C_t^{*-1} (R_t^{-1} \Gamma C_{t-1} C_{t-1}^s (f_{t-1} - E_t) + C_t^{-1} m_t), \end{aligned}$$

onde $E_t = (I_m - C_{t-1} \Gamma' R_t^{-1} \Gamma) m_{t-1}$.

Nas expressões acima, m_t e C_t representam as médias e variâncias *on-line* obtidas do FKE, $m_t^s(\cdot)$, $m_t^{s*}(\cdot)$ e $C_t^s(\cdot)$, $C_t^{s*}(\cdot)$ representam as médias e variâncias retrospectivas dado f_{t+1} e $\{f_{t+1}, f_{t-1}\}$, respectivamente.

O esquema de amostragem pode ser descrito da seguinte maneira. Para cada $b = 1, \dots, B$, um valor candidato \tilde{F}_b é amostrado da densidade proposta $q(F_b^{(l-1)}, \tilde{F}_b)$ utilizando as aproximações obtidas no FKE e as médias e variâncias retrospectivas do FFBS. Cada f_t é amostrado

seqüencialmente de $t = L_2^b$ até $t = L_1^b$ de suas distribuições retrospectivas. Portanto, uma amostra $F_b^{(l)}$ na l -ésima iteração do algoritmo MCMC é obtida de acordo com o seguinte algoritmo:

Algoritmo 5.2. (*Amostragem do bloco F_b*)

1. Amostragem $\tilde{f}_{L_2^b}$ de

$$\begin{aligned} N(m_{L_2^b}^s(f_{L_2^b+1}^{(l-1)}), C_{L_2^b}^s) & \quad \text{se } b = 1, \dots, B-1 \quad \text{ou de} \\ N(m_{L_2^b}, C_{L_2^b}) & \quad \text{se } b = B. \end{aligned}$$

2. Partindo de $t = L_2^b - 1$ até $t = L_1^b + 1$, amostrar \tilde{f}_t de $N(m_t^s(\tilde{f}_{t+1}), C_t^s)$.

3. Amostragem $\tilde{f}_{L_1^b}$ de

$$\begin{aligned} N(m_{L_1^b}^s(\tilde{f}_{L_1^b+1}), C_{L_1^b}^s) & \quad \text{se } b = 1 \quad \text{ou de} \\ N(m_{L_1^b}^{s*}(\tilde{f}_{L_1^b+1}, f_{L_1^b-1}^{(l-1)}), C_{L_1^b}^{s*}) & \quad \text{se } b = 2, \dots, B. \end{aligned}$$

4. Faça $F_b^{(l)} = \tilde{F}_b$ com probabilidade α_b e $F_b^{(l)} = F_b^{(l-1)}$ com probabilidade $1 - \alpha_b$, onde

$$\alpha_b = \min \left\{ 1, \frac{p(\tilde{F}_b | y, \Theta, f_{1, L_1^b-1}^{(l-1)}, f_{L_2^b+1, T}^{(l-1)}) q(\tilde{F}_b \rightarrow F_b^{(l-1)})}{p(F_b^{(l-1)} | y, \Theta, f_{1, L_1^b-1}^{(l-1)}, f_{L_2^b+1, T}^{(l-1)}) q(F_b^{(l-1)} \rightarrow \tilde{F}_b)} \right\}.$$

No algoritmo anterior, a probabilidade de aceitação α_b depende da distribuição condicional completa $p(F_b | y, \Theta, f_{1, L_1^b-1}, f_{L_2^b+1, T})$ definida na equação (5.5). Esta distribuição é facilmente calculada combinando os resultados das equações (5.2) e (5.6) correspondentes a função de verossimilhança e à priori condicional auto-regressiva respectivamente. A densidade proposta $q(F_b^r \rightarrow F_b^s)$ (r, s igual a $l-1, l$ ou $l, l-1$) é avaliada como o produto das densidades retrospectivas apresentadas anteriormente. Nota-se que na atualização de F_b^r a F_b^s , as médias e variâncias das distribuições propostas, dependem de todos os elementos do bloco F_b^r utilizados na construção das observações artificiais e nas médias e variâncias *on-line* do KFE.

Por último, o esquema de amostragem baseado no algoritmo MCMC, para obter amostras a posteriori de todos os parâmetros do modelo (F, β, Θ) , é resumido no seguinte algoritmo:

Algoritmo 5.3. (*Esquema MCMC para o MFEDG*)

a. Inicialização: dar valores iniciais $(F^{(0)}, \beta^{(0)}, \Theta^{(0)})$ e inicialize o contador de iterações em $l = 1$.

b. Amostre F : $F^{(l)}$ é amostrado em blocos como segue:

1. Faça $b = 1$
2. Amostre \tilde{F}_b da densidade proposta $q(F_b \rightarrow \tilde{F}_b)$ de forma seqüencial de $t = L_2^b$ até L_1^b (FKE + FFBS).
3. Faça $F_b^{(l)} = \tilde{F}_b$ com probabilidade α_b e $F_b^{(l)} = F_b^{(l-1)}$ com probabilidade $1 - \alpha_b$, onde

$$\alpha_b = \min \left\{ 1, \frac{p(\tilde{F}_b | y, \Theta, f_{1, L_1^b-1}^{(l-1)}, f_{L_2^b+1, T}^{(l-1)}) q(\tilde{F}_b \rightarrow F_b^{(l-1)})}{p(F_b^{(l-1)} | y, \Theta, f_{1, L_1^b-1}^{(l-1)}, f_{L_2^b+1, T}^{(l-1)}) q(F_b^{(l-1)} \rightarrow \tilde{F}_b)} \right\}.$$

4. Faça $b \rightarrow b + 1$ e volte para **b.2.** se $b < B$.

c. Amostre β^* : $\beta^{*(l)}$ é amostrado como segue:

1. Amostre $\tilde{\beta}^*$ da distribuição $N(\beta^{*(l-1)}, \Delta_\beta I_{Nm})$.
2. Faça $\beta^{*(l)} = \tilde{\beta}^*$ com probabilidade α e $\beta^{*(l)} = \beta^{*(l-1)}$ com probabilidade $1 - \alpha$, onde

$$\alpha = \min \left\{ 1, \frac{\prod_{t=1}^T p(y_t | f_t^*, \tilde{\beta}^*) f_N(\tilde{\beta}^*; \mu_{\beta^*}, \Sigma_{\beta^*})}{\prod_{t=1}^T p(y_t | f_t^*, \beta^*) f_N(\beta^*; \mu_{\beta^*}, \Sigma_{\beta^*})} \right\}.$$

d. Amostrando Θ : Os elementos de Θ são gerados separadamente utilizando o amostrador de Gibbs e passos de Metropolis-Hastings.

e. Atualização: Faça $l \rightarrow l + 1$ e volte para **b** até a convergência.

No Algoritmo 5.3 a atualização do vetor de hiperparâmetros Θ é feita seguindo os procedimentos descritos na Subseção 5.3.2. Na Seção 5.4 são apresentados dois exemplos com dados simulados para mostrar a aplicabilidade do algoritmo de inferência proposto nesta seção.

5.3.4 Inferência para o número de fatores

Analogamente ao caso normal, o algoritmo RJMCMC é utilizado para estimar o número de fatores m . Este algoritmo (baseado na proposta de Lopes & West (2004)) fornece o cálculo da probabilidade a posteriori do modelo (PPM) utilizado na seleção de modelos. O algoritmo RJMCMC para o MFEDG é o mesmo apresentado na Subseção 2.4.3 com duas variantes.

Primeiro, a distribuição proposta conjunta de todos os parâmetros, dado m , denotado por $q_m(\Psi_m)$ é dada por:

$$\begin{aligned} q_m(\Psi_m) &= \prod_{j=1}^m f_N(f_{(j)}; M_{f_{(j)}}, aV_{f_{(j)}}) f_N(\beta_{(j)}; M_{\beta_{(j)}}, bV_{\beta_{(j)}}) f_N(\gamma_j; M_{\gamma_j}, cV_{\gamma_j}) \\ &\times \prod_{j=1}^m f_{IG}(\lambda_j; d, dM_{\lambda_j}) f_N(\mu_j; M_{\mu_j}, eV_{\mu_j}) f_{IG}(\phi_j; f, fM_{\phi_j}) \\ &\times \prod_{j=1}^m f_{IG}(\tau_j^2; g, gM_{\tau_j}), \end{aligned}$$

onde $\Psi_m = (F_m, \beta_m, \Theta)$, a , b , c , d , e , f e g são constantes sintonizadoras e M_x e V_x representam as médias e variâncias a posteriori baseadas em rodadas preliminares do algoritmo MCMC para diferentes valores de m .

Segundo, a distribuição a posteriori conjunta de Ψ_m e m denotada por $p(\Psi_m, m|y)$ é dada por $p(\Psi_m, m|y) \propto p(y|m, \Psi_m)p(\Psi_m)Pr(m)$ onde $p(y|m, \Psi_m)p(\Psi)$ é obtida de (5.4) e $Pr(m) = 1/M$ onde M é o número máximo de fatores testados.

5.4 Estudo simulado

Nesta seção dois exemplos com dados simulados são apresentados. O primeiro refere-se a dados com distribuição Gama e, o segundo, a dados com distribuição de Bernoulli, gerados seguindo a estrutura do MFEDG. O objetivo é testar a aplicabilidade dos algoritmos propostos considerando diferentes especificações do modelo geral. Em ambos os casos, N pontos foram gerados num quadrado unitário representando os locais, onde todas as observações foram simuladas. Além disso, diferentes especificações do MFEDG são ajustadas aos dados simulados para testar se a especificação utilizada na geração dos dados se ajustou melhor as observações. Isto é feito mediante critérios de seleção de modelos e utilizando o algoritmo RJMCMC proposto.

5.4.1 Exemplo 1: Dados com distribuição Gama

Em geral, se y segue uma distribuição Gama com parâmetros α (de forma) e β (de escala) a sua densidade é dada por:

$$p(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad y \geq 0$$

com média e variância dadas por $E(y) = \mu = \alpha/\beta$ e $Var(y) = \alpha/\beta^2$. Uma representação alternativa é obtida re-parametrizando β em função da média tal que $\beta = \alpha/\mu$ e, portanto, tem-se que

$$p(y|\mu, \alpha) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu} y\right)^\alpha \exp\left\{-\frac{\alpha}{\mu} y\right\} \frac{1}{y}.$$

A densidade acima é utilizada neste exemplo e é denotada por $Gama(\mu, \alpha)$. Para o caso de dados variando no espaço e no tempo, seja $y_t = (y_{1t}, \dots, y_{Nt})$ ($t = 1, \dots, T$) o vetor de observações nos N locais tal que $y_t \sim Gama(\mu_t, \alpha)$ com densidade

$$p(y_t|\mu_t, \alpha) = \prod_{i=1}^N \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_{it}} y_{it}\right)^\alpha \exp\left\{-\frac{\alpha}{\mu_{it}} y_{it}\right\} \frac{1}{y_{it}}$$

onde $\mu_t = (\mu_{1t}, \dots, \mu_{Nt})'$ e α é o parâmetro de forma escalar. O logaritmo da densidade de y_t é dada por

$$\begin{aligned} \log p(y_t|\mu_t, \alpha) &= \sum_{i=1}^N [-(\alpha/\mu_{it})y_{it} + \alpha \log(\alpha/\mu_{it}) + \alpha \log y_{it} - \log \Gamma(\alpha) - \log y_{it}] \\ &= \alpha [y_t' \eta_t - b(\eta_t)] + c(y_t, \alpha), \end{aligned}$$

onde $\eta_t = (\eta_{1t}, \dots, \eta_{Nt})'$, $\eta_{it} = -\frac{1}{\mu_{it}}$, $b(\eta_t) = -\sum_{i=1}^N \log(-\eta_{it})$ e $c(y_t, \alpha) = \sum_{i=1}^N (\alpha \log y_{it} - \log \Gamma(\alpha) - \log y_{it})$. As funções média e variância são dadas por

$$E(y_t|\eta_t) = b'(\eta_t) = \begin{pmatrix} -\frac{1}{\eta_{1t}} \\ \vdots \\ -\frac{1}{\eta_{Nt}} \end{pmatrix} \quad \text{e} \quad V(y_t|\eta_t) = b''(\eta_t)/\alpha = \text{diag} \begin{pmatrix} \frac{1}{\alpha \eta_{1t}^2} \\ \vdots \\ \frac{1}{\alpha \eta_{Nt}^2} \end{pmatrix}.$$

Modelo utilizado

O MFEDG com dados Gama, utilizado neste exemplo, é especificado da seguinte forma:

$$y_t|\mu_t, \alpha \sim Gama(\mu_t, \alpha) \tag{5.8a}$$

$$\eta_t = -1/\mu_t = -\exp(-\theta_t) \tag{5.8b}$$

$$\log \mu_t = \theta_t = \beta f_t \tag{5.8c}$$

$$f_t = \Gamma f_{t-1} + w_t, \quad w_t \sim N(0, \Lambda) \tag{5.8d}$$

$$\beta_{(j)} = \mu_j^\beta \mathbf{1}_N + \delta_j, \quad \delta_j \sim PG(0, \tau_j^2 \rho_{\phi_j}(\cdot)) \tag{5.8e}$$

$$f_0 \sim N(m_0, C_0)$$

Nota-se que a função de ligação é o logaritmo da média, $g(\mu_t) = \log(\mu_t) = \theta_t$ e, portanto, a função resposta é a função logaritmo. O parâmetro natural η_t depende do preditor linear θ_t através de $\eta_t = -\exp(-\theta_t)$. De acordo com o número de fatores considerado o modelo acima é denotado por MFEDG(m).

Dados simulados

Foram gerados dados a partir do modelo descrito em (5.8) considerando $N = 25$ locais (gerados aleatoriamente num quadrado unitário $[0, 1] \times [0, 1]$), 100 instantes de tempo e $m = 2$ fatores. Cada coluna da matriz de cargas do fator segue um processo Gaussiano com função de correlação Matérn com parâmetro de forma $\kappa = 1$. Os parâmetros utilizados foram: $\alpha = 1, 5$; $\Gamma = \text{diag}(0, 8; 0, 65)$; $\Lambda = \text{diag}(0, 05; 0, 1)$; $\mu^\beta = (0; 0)$; $\phi = (0, 15; 0, 1)$ e $\tau = (0, 25; 1)$.

Neste exemplo, os parâmetros κ e μ^β do processo Gaussiano não foram estimados, permanecendo fixos nos seus valores verdadeiros durante o procedimento de inferência. Por outro lado, as últimas 10 observações de cada série temporal foram deixadas de fora da análise para efeitos de previsão e comparação e, portanto, o tamanho final da amostra é $T = 90$. A Figura 5.1 mostra os dados gerados ao longo do tempo em 12 dos 25 locais, e a Figura 5.2 (primeira coluna) as superfícies simuladas de $\beta_{(1)}$ e $\beta_{(2)}$ junto com os 25 locais considerados (pontos azuis).

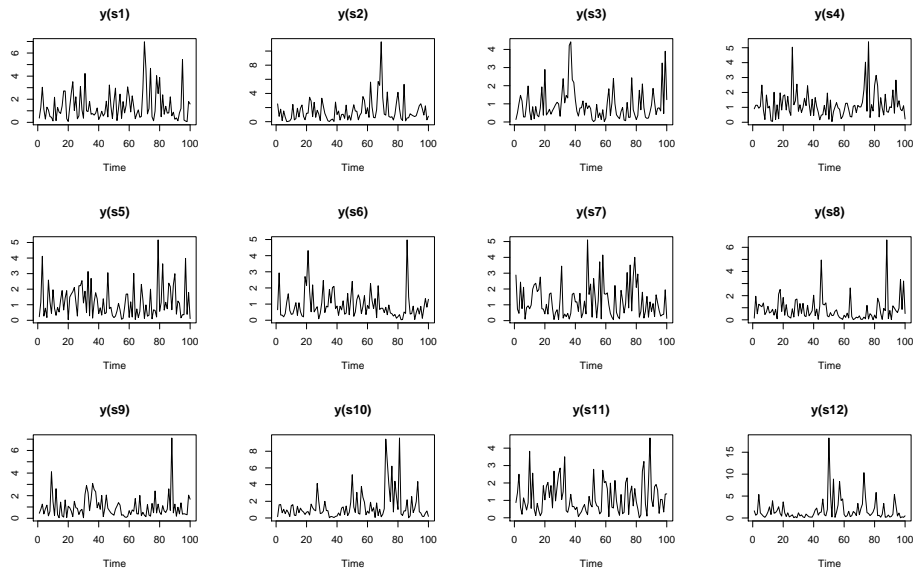


Figura 5.1: *Dados Gama:* Dados gerados do modelo Gama. Cada gráfico representa a série temporal simulada no local s_i , $i = 1, \dots, 12$.

As distribuições a priori utilizadas no procedimento de inferência foram as seguintes: para $j = 1, \dots, m$, $\gamma_j \sim N(0, 1)$, $\lambda_j \sim GI(0, 01; 0, 01)$, $\tau_j^2 \sim IG(2; 0, 5)$ e $\phi_j \sim GI(2; b)$ onde $b = \max(dist)/(-2 \log(0, 05))$ tal que $\max(dist)$ é a distância máxima entre locais. A priori para f_0 é $N(0_m, I_m)$ e para o parâmetro de forma α é $GI(0, 01; 0, 01)$. α é amostrado dentro do algoritmo MCMC com passos de Metropolis-Hastings. A proposta de transição é uma log-normal com parâmetro de forma $\log \alpha$ e parâmetro de escala Δ_α , ou seja $q(\alpha, \cdot) = f_{LN}(\cdot; \log \alpha, \Delta_\alpha)$. A amostragem dos fatores foi feita em blocos como descrita na Subseção 5.3.3. Foram testados 5, 10 e 15 blocos para garantir taxas de aceitação razoáveis. Com 5 blocos as taxas foram baixas ($< 0, 10$) e com 10 e 15 blocos, aumentaram. Dado que o tempo computacional é menor ao considerar menos blocos, $B = 10$ blocos foram considerados neste exemplo.

Amostras a posteriori para todos os parâmetros do modelo foram obtidas utilizando o esquema MCMC descrito no Algoritmo 5.3. Para cada parâmetro foram geradas duas cadeias paralelas de tamanho 20.000 onde as primeiras 5.000 iterações foram deixadas de fora como período de aquecimento, guardou-se amostras a cada 5 iterações. Desta forma o tamanho final das amostras a posteriori é 3.000. Todos os esquemas propostos foram implementados na linguagem de programação `0x` versão 3.40 (Doornik, 2002). A estatística \hat{R} de Gelman & Rubin (1992) foi utilizada para testar a convergência das cadeias, onde valores próximos de 1 sugerem que a convergência das cadeias foi atingida.

O conjunto de dados simulados foi ajustado considerando a especificação MFEDG(m) com $m = 1, 2, 3, 4$ e 5 fatores. Adicionalmente um modelo mais simples com estrutura espacial, denominado modelo espacial dinâmico generalizado com componente comum (MEDG-CC), é ajustado. Este modelo se diferencia da especificação em 5.8 na equação do preditor linear tal que $\theta_t = \mu_t^y 1_N + \delta_t$, $\delta_t \sim PG(0, \tau^2 \rho_\phi(\cdot))$ e $\mu_t^y \sim N(\mu_{t-1}^y, W)$.

A comparação entre modelos foi feita utilizando várias medidas, entre elas o Erro Quadrático Médio (EQM), o Erro Absoluto Médio (EAM) e o probabilidade a posteriori do modelo (PPM) estimada do algoritmo RJMCMC (ver Subseção 5.3.4 para mais detalhes). As duas primeiras medidas estão baseadas no erro entre as observações simuladas e ajustadas e no erro de previsão. Os valores previstos foram estimados utilizando o procedimento apresentado na Subseção 5.2.3. A Tabela 5.1 mostra estes valores para todos os modelos ajustados. Nota-se que para a especificação MFEDG(m), a maior probabilidade a posteriori (0,29) é para o modelo com 2 fatores, como desejado. As medidas de erro também indicam o modelo com

dois fatores como o melhor, tanto no ajuste quanto na previsão. A especificação MEDG-CC apresenta os piores resultados considerando todos os critérios utilizados, indicando que um modelo mais estruturado é necessário para o ajuste dos dados.

Tabela 5.1: *Dados Gama:* Critérios de comparação de modelos. Soma do Erro Quadrático (SEQ), Soma do Erro Absoluto (SEA), Erro Quadrático Médio da previsão (EQM_P), Erro Absoluto Médio da previsão (EAM_P) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *itálico*.

Modelo	m	SEQ	SEA	EQM_P	EAM_P	PPM
MFEDG(m)	1	2581,00	1604,10	1,47	0,92	0,22
	2	<i>2363,38</i>	<i>1557,50</i>	<i>1,35</i>	<i>0,89</i>	<i>0,29</i>
	3	2507,27	1587,27	1,43	0,91	0,20
	4	2675,68	1639,14	1,53	0,94	0,10
	5	2522,59	1606,57	1,44	0,92	0,19
MEDG-CC	-	5464,65	2613,67	3,12	1,49	-

Os seguintes resultados estão baseados nas amostras a posteriori da especificação MFEDG(2) (modelo utilizado na geração dos dados). Para a matriz de cargas dos fatores, a taxa de aceitação do algoritmo de Metropolis-Hastings foi 0,288. A Figura 5.2 mostra as superfícies simuladas e estimadas das componentes $\beta_{(1)}$ e $\beta_{(2)}$. A estimação de cada superfície foi feita via krigagem Bayesiana seguindo o procedimento descrito na Subseção 5.2.2. A coluna do lado direito de cada gráfico indica o intervalo de variação das estimativas na mesma magnitude para cada $\beta_{(j)}$, $j = 1, 2$, onde as cores mais claras indicam os valores mais altos. Nota-se que para $\beta_{(1)}$ ambas superfícies estimada e simulada são muito parecidas indicando valores próximos dos valores simulados. Nota-se também, que os valores estimados de $\beta_{(2)}$ foram subestimados embora eles apresentem a mesma estrutura espacial.

A amostragem dos fatores foi feita considerando 10 blocos, cada um deles com taxas de aceitação de 0,245, 0,359, 0,412, 0,412, 0,178, 0,304, 0,371, 0,394, 0,123 e 0,264 respectivamente. A Figura 5.3 mostra os valores simulados dos fatores assim como as médias a posteriori estimadas e os intervalos de credibilidade de 95%. O primeiro gráfico indica que todos valores estimados do fator 1 ficaram próximos dos simulados. Já o segundo gráfico indica que alguns valores foram superestimados, principalmente para t variando de 40 até 60. Este fato está diretamente relacionado com os resultados de $\beta_{(2)}$. Ou seja, dado que quase todos os valores de $\beta_{(2)}$ foram subestimados, era esperado valores superestimados no segundo fator. Adicionalmente, na Figura 5.4 são apresentados os gráficos de autocorrelação das amostras a posteriori de f_t , $t = 10, 35, 60, 85$. O interesse aqui é perceber se as amostras apresentam

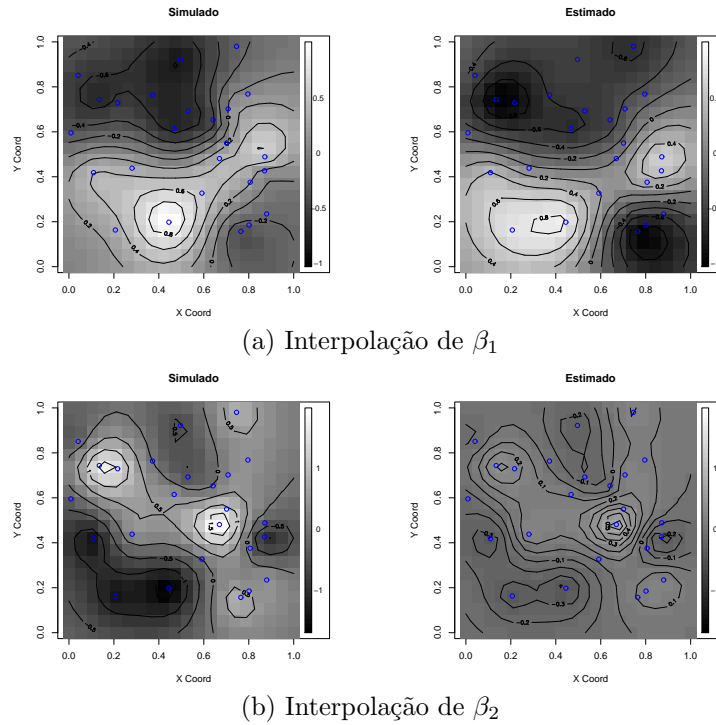


Figura 5.2: *Dados Gama:* Interpolação espacial das componentes β_1 e β_2 . Em (a) e (b) o gráfico da esquerda representa a superfície simulada e o da direita a superfície estimada via krigagem Bayesiana. Os pontos azuis indicam os locais onde as observações foram simuladas.

alta autocorrelação. Nota-se que a maioria dos gráficos apresenta baixa correlação sugerindo independência nos valores gerados.

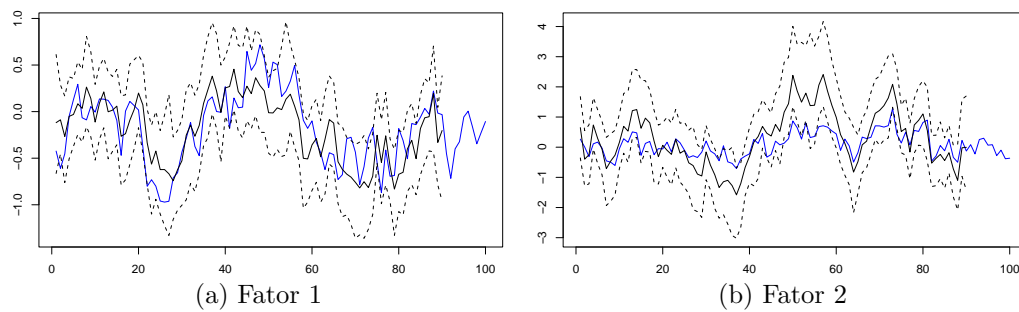


Figura 5.3: *Dados Gama:* Fatores simulados vs. fatores estimados. A linha cheia azul representa os valores simulados, a linha cheia preta representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

Finalmente, a Tabela 5.2 mostra algumas estatísticas a posteriori dos hiperparâmetros do modelo 5.8. Nota-se que quase todos os parâmetros (menos λ_2 e τ_2^2) foram bem estimados com intervalos de credibilidade contendo o valores verdadeiros. O parâmetro λ_2 que repre-

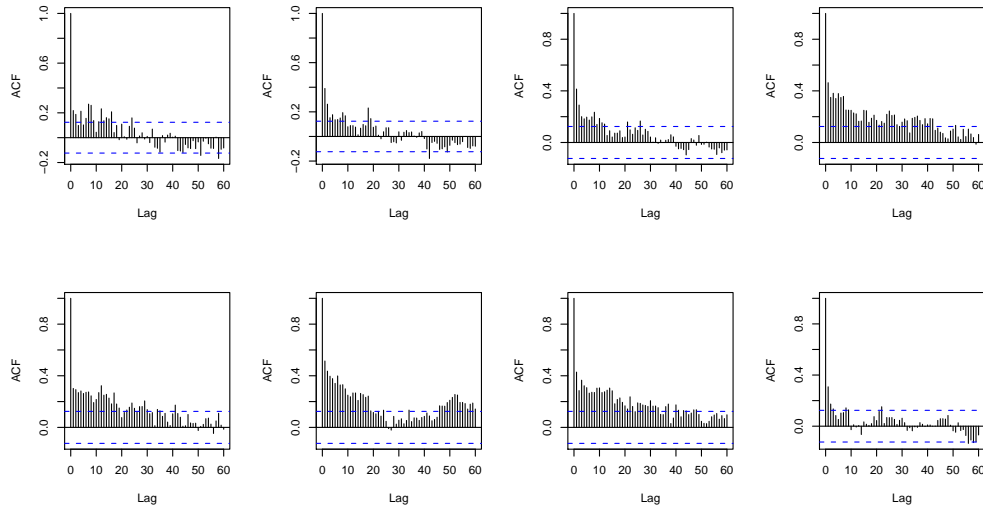


Figura 5.4: *Dados Gama:* Gráficos de autocorrelação de f_t , $t = 10, 35, 60, 85$ do primeiro (acima) e segunda fator (embaixo) respectivamente.

Tabela 5.2: *Dados Gama:* Sumário a posteriori dos parâmetros do modelo em (5.8). T.A.: Taxa de aceitação. \hat{R} : diagnóstico de Gelman & Rubin.

θ	Valor	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			T.A.	\hat{R}
				2,5%	50 %	97,5%		
α	1,50	1,460	0,042	1,380	1,460	1,540	0,513	1,00
γ_1	0,80	0,764	0,120	0,509	0,778	0,952	-	1,01
γ_2	0,65	0,765	0,110	0,527	0,779	0,946	-	1,01
λ_1	0,05	0,083	0,042	0,024	0,076	0,181	-	1,02
λ_2	0,10	0,460	0,210	0,189	0,425	0,936	-	1,03
τ_1^2	0,25	0,407	0,200	0,158	0,362	0,899	-	1,05
τ_2^2	1,00	0,139	0,064	0,057	0,127	0,287	-	1,03
ϕ_1	0,15	0,194	0,073	0,086	0,185	0,364	0,612	1,04
ϕ_2	0,10	0,095	0,047	0,032	0,086	0,212	0,719	1,05

senta a variância do segundo fator foi superestimado. Este resultado guarda relação com a superestimação de alguns valores de f_{2t} . Por outro lado, o parâmetro τ_2^2 foi subestimado indicando menor variabilidade no processo de β_2 como verificado nos resultados apresentados na Figura 5.2b. Ainda na Tabela 5.2, a última coluna mostra os resultados da estatística \hat{R} com valores muito próximos de 1, sugerindo convergência das cadeias.

5.4.2 Exemplo 2: Dados com distribuição de Bernoulli

Neste exemplo, suponha que a observação y_{it} (com valores 0 ou 1) medida no local s_i ($i = 1, \dots, N$) e no tempo t ($t = 1, \dots, T$) segue uma distribuição de Bernoulli com probabilidade p_{it} . Ou seja $y_{it} \sim \text{Bernoulli}(p_{it})$ com média p_{it} e variância $p_{it}(1 - p_{it})$. Se y_t representa o vetor que contém as observações nos N locais, a densidade de y_t é dada por

$$\begin{aligned} p(y_t|p_t) &= \prod_{i=1}^N p_{it}^{y_{it}} (1 - p_{it})^{1-y_{it}} \\ &= \exp \left\{ \sum_{i=1}^N y_{it} \log(p_{it}/(1 - p_{it})) + \sum_{i=1}^N \log(1 - p_{it}) \right\} \\ &= \exp \left\{ y_t' \eta_t - \sum_{i=1}^N \log(1 + \exp \eta_{it}) \right\} \end{aligned}$$

onde $\eta_{it} = \log(p_{it}/(1 - p_{it}))$ e $\eta_t = (\eta_{1t}, \dots, \eta_{Nt})'$. Dado que $b(\eta_t) = \sum_{i=1}^N \log(1 + \exp \eta_{it})$ temos que as funções média e variância são dadas por

$$E(y_t|\eta_t) = b'(\eta_t) = \begin{pmatrix} \frac{\exp \eta_{1t}}{1 + \exp \eta_{1t}} \\ \vdots \\ \frac{\exp \eta_{Nt}}{1 + \exp \eta_{Nt}} \end{pmatrix} \quad \text{e} \quad V(y_t|\eta_t) = b''(\eta_t) = \text{diag} \begin{pmatrix} \frac{\exp \eta_{1t}}{(1 + \exp \eta_{1t})^2} \\ \vdots \\ \frac{\exp \eta_{Nt}}{(1 + \exp \eta_{Nt})^2} \end{pmatrix}.$$

Modelo utilizado

O modelo utilizado na geração dos dados é o seguinte:

$$y_{it}|p_{it} \sim \text{Bernoulli}(p_{it}), \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (5.9a)$$

$$\log(p_{it}/(1 - p_{it})) = \theta_{it}, \quad (5.9b)$$

$$\theta_t = \mu_t^{y*} + \beta f_t, \quad (5.9c)$$

$$f_t = \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \quad (5.9d)$$

$$\beta_{(j)} = \mu_j^{\beta*} + \delta_j, \quad \delta_j \sim PG(0, \tau_j^2 \rho_{\phi_j}(\cdot)) \quad (5.9e)$$

$$f_0 \sim N(m_0, C_0),$$

onde $\theta_t = (\theta_{1t}, \dots, \theta_{Nt})'$ é chamado de preditor linear. A função de ligação g é o *logito* tal que $g(p_{it}) = \log(p_{it}/(1 - p_{it}))$. A componente μ_t^{y*} no preditor linear é igual a $\mu_t^{y*} = \mu_t^y \mathbf{1}_N$ com $\mu_t^y \sim N(\mu_{t-1}^y, W)$ e $\mu_0^y \sim N(m_0^y, C_0^y)$. Ou seja, μ_t^y representa a componente de tendência comum para todos os locais. O modelo em 5.9 é chamado de modelo fatorial espacial dinâmico

generalizado com componente comum e é denotado por MFEDG-CC(m) onde m indica o número de fatores considerado.

Dados simulados

Inicialmente foram gerados $N = 30$ pontos num quadrado unitário representando os locais, onde as observações são medidas. Em seguida, para $t = 1, \dots, 100$, observações y_t foram geradas a partir da especificação MFEDG-CC(2). Analogamente ao caso anterior, as colunas de β seguem processos Gaussianos com função de correlação Matérn e parâmetro de forma $\kappa = 1$. Os parâmetros utilizados na geração dos dados foram: $W = 0, 10$; $\Gamma = \text{diag}(0, 75; 0, 90)$; $\Lambda = \text{diag}(0, 10; 0, 05)$; $\mu^\beta = (-1; 1)$; $\phi = (0, 15; 0, 2)$ e $\tau = (0, 81; 1)$. Neste estudo, o parâmetro κ não é estimado permanecendo fixo no seu valor verdadeiro durante o procedimento de inferência. Novamente as últimas 10 observações de cada série foram retiradas da análise para efeitos de comparação.

A Figura 5.5 mostra os valores simulados em 12 locais. Em cada gráfico os pontos indicam as observações geradas ao longo do tempo e a linha preta o valor da média denotada por p_t . Nota-se que em todos os locais o valor da média têm o mesmo comportamento, ou seja, tendência decrescente a partir de $t = 10$ e tendência crescente a partir de $t = 50$. Este comportamento é explicado pela presença da componente comum μ_t^y que exibe uma trajetória similar.

As distribuições a priori utilizadas foram as mesmas especificadas no exemplo com dados Gama. Adicionalmente, distribuições a priori para μ^β e W foram: $\mu^\beta \sim N((-1, 1)', 5I_2)$ e $W \sim GI(2, 0.1)$. A amostragem dos fatores foi feita considerando $B = 9$ blocos. O algoritmo MCMC foi utilizado para gerar amostras a posteriori dos parâmetros seguindo o mesmo esquema do exemplo anterior.

No total, 5 modelos foram ajustados, os 4 primeiros seguindo a especificação MFEDG-CC(m) com $m = 1, 2, 3$ e 4 fatores e o último seguindo a especificação MEDG-CC definido no exemplo anterior (modelo mais simples). A comparação entre estes modelos é feita utilizando os critérios EQM e EAM baseados no ajuste e na previsão e a probabilidade a posteriori do modelo (PPM) calculada do algoritmo RJMCMC. A Tabela 5.3 mostra estes resultados. A especificação MFEDG-CC(2) apresenta a maior PPM como esperado e os menores valores de EQM e EAM tanto no ajuste quanto na previsão. A especificação MEDG-CC tem os maiores valores de EQM e EAM indicando ser o modelo com a pior performance.

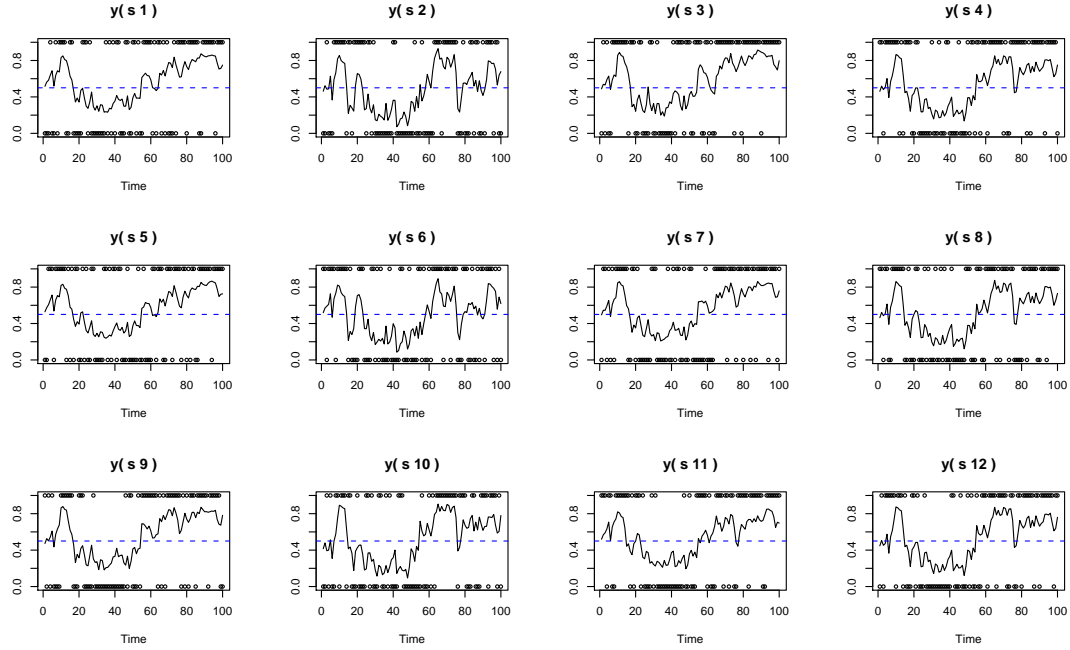


Figura 5.5: *Dados Bernoulli:* Dados gerados do modelo Bernoulli. Cada gráfico representa a série temporal simulada no local s_i , $i = 1, \dots, 12$.

Tabela 5.3: *Dados Bernoulli:* Critérios de comparação de modelos. Erro Quadrático Médio (EQM), Erro Absoluto Médio (EAM) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *italico*.

Modelo	m	EQM ¹	EAM ¹	EQM ²	EAM ²	PPM
	1	0,2001	0,4373	0,2380	0,4471	0,20
MFEDG-CC(m)	2	<i>0,1949</i>	<i>0,4000</i>	<i>0,2330</i>	<i>0,4190</i>	<i>0,38</i>
	3	0,2134	0,4212	0,2412	0,4360	0,27
	4	0,2221	0,4410	0,2490	0,4522	0,15
MEDG-CC	-	0,3520	0,6450	0,3998	0,6993	-

¹ Com valores ajustados dentro da amostra

² Com valores da previsão.

A taxa de aceitação de Metropolis-Hastings para β foi 0,636. A Figura 5.6 mostra as superfícies estimada e simulada de $\beta_{(1)}$ e $\beta_{(2)}$ obtidas via krigagem Bayesiana. O gráfico indica que as superfícies estimadas são bastante semelhantes às simuladas com algumas discrepâncias, principalmente nos pontos, em que os $\beta_{(j)}$'s apresentam os menores valores.

As taxas de aceitação dos 9 blocos de F foram 0,287, 0,213, 0,313, 0,248, 0,186, 0,223,

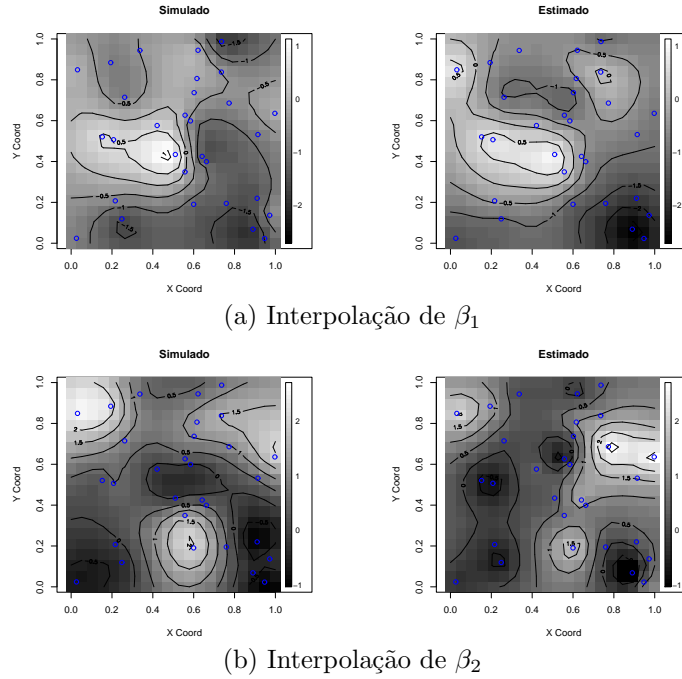


Figura 5.6: *Dados Bernoulli:* Interpolação espacial das componentes β_1 e β_2 . Em (a) e (b) o gráfico da esquerda representa a superfície simulada e o da direita a superfície estimada via krigagem Bayesiana. Os pontos azuis indicam os locais onde as observações foram simuladas.

0,321, 0,253 e 0,179. A Figura 5.7a apresenta a trajetória simulada e estimada da componente comum e dos fatores. O primeiro gráfico, correspondente a componente comum, indica que as médias a posteriori são bastante próximas dos valores simulados. O segundo gráfico, correspondente ao fator 1, indica que nem todos os valores foram bem estimados, mais especificamente no início e quase no final da série. O terceiro gráfico, correspondente ao terceiro fator, mostra valores mais próximos dos simulados embora no final da série o intervalo de credibilidade não contenha alguns valores.

A Figura 5.8 apresenta os gráficos de autocorrelação para alguns valores de f_t , $t = 10, 35, 65$ e 85 . Em geral os gráficos mostram correlações baixas, exceto nas componentes f_{1t} e f_{2t} (dois primeiros gráficos).

Finalmente, a Tabela 5.4 mostra o sumário a posteriori para todos os hiperparâmetros do modelo 5.9 obtidos a partir do algoritmo MCMC. Nota-se que quase todos os intervalos de credibilidade de 95% contêm os valores verdadeiros dos parâmetros, exceto γ_2 que foi subestimado.

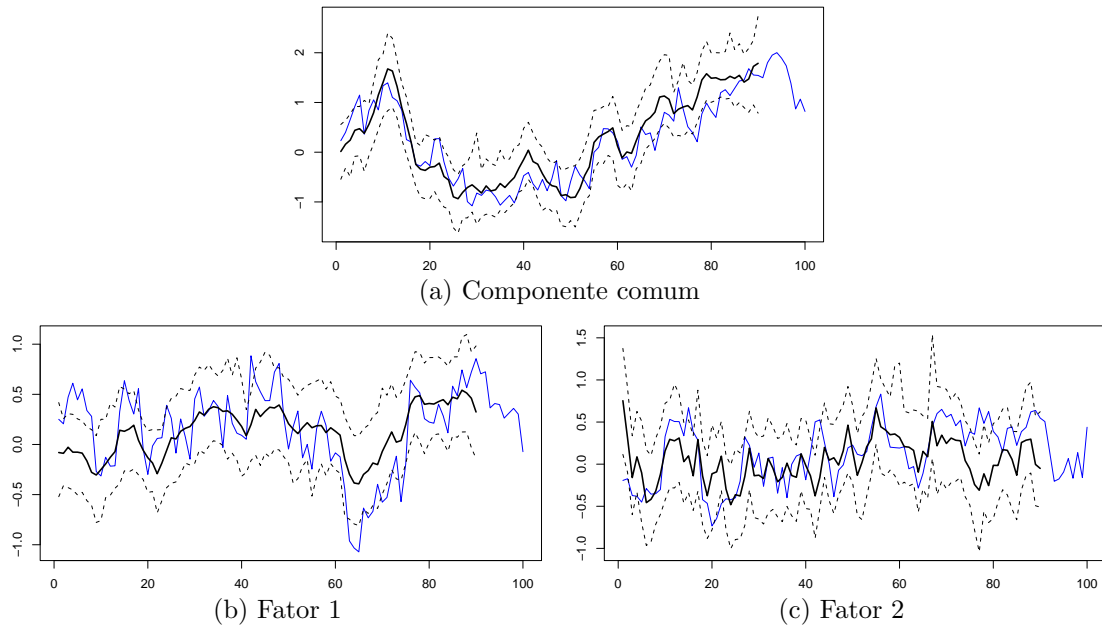


Figura 5.7: *Dados Bernoulli:* Valores simulados e estimados da componente comum e dos fatores. A linha cheia azul representa os valores simulados, a linha cheia preta representa as médias a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

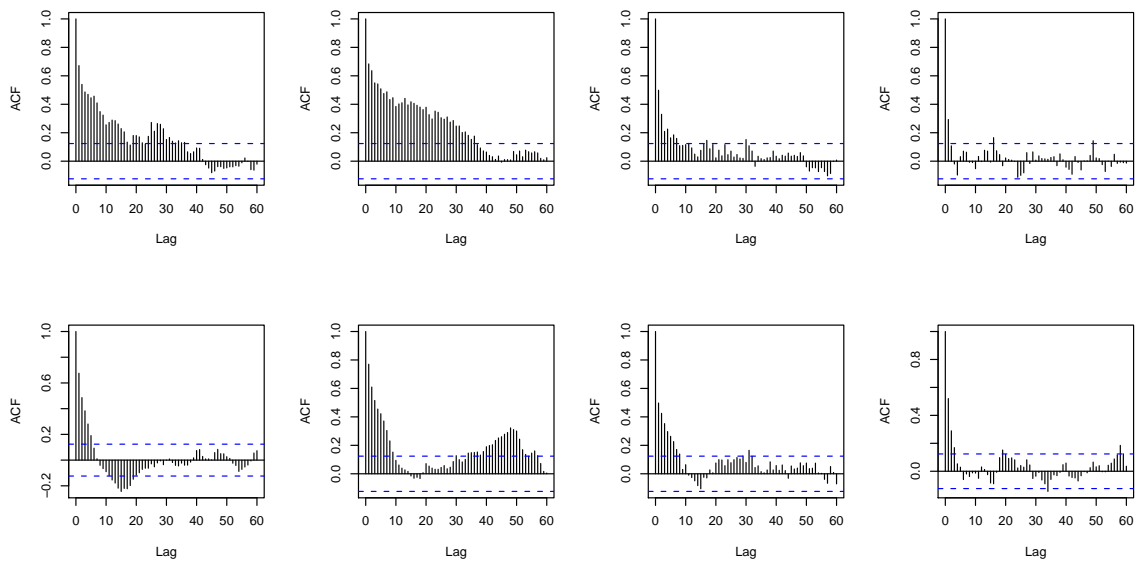


Figura 5.8: *Dados Bernoulli:* Gráficos de autocorrelação de f_t , $t = 10, 35, 60, 85$ do primeiro (acima) e segunda fator (embaixo) respectivamente.

Tabela 5.4: *Dados Bernoulli:* Sumário a posteriori dos parâmetros do modelo em (5.9). T.A.: Taxa de aceitação. \hat{R} : diagnóstico de Gelman & Rubin.

θ	Valor	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			T.A.	\hat{R}
				2,5%	50 %	97,5%		
W	0,10	0,105	0,043	0,044	0,100	0,217	-	1,01
γ_1	0,75	0,840	0,100	0,608	0,856	0,986	-	1,01
γ_2	0,90	0,663	0,190	0,466	0,688	0,862	-	1,03
λ_1	0,10	0,097	0,031	0,047	0,092	0,176	-	1,01
λ_2	0,05	0,034	0,017	0,011	0,030	0,075	-	1,01
μ_1^β	-1,00	-0,844	0,390	-1,680	-0,824	-0,115	-	1,00
μ_2^β	1,00	0,670	0,380	-0,023	0,644	1,520	-	1,01
τ_1^2	0,81	1,060	0,580	0,413	0,926	2,600	-	1,04
τ_2^2	1,00	1,400	0,500	0,735	1,310	2,570	-	1,03
ϕ_1	0,15	0,156	0,075	0,063	0,138	0,353	0,567	1,05
ϕ_2	0,20	0,104	0,053	0,038	0,086	0,249	0,683	1,06

5.5 Considerações finais

Neste capítulo foi apresentada uma extensão do MFED para dados na família exponencial. Isto é possível substituindo a equação de observação por uma representação na família exponencial em que o MFED é usado para modelar transformações da média através de uma função de ligação apropriada. O modelo, denominado MFEDG, se mostra bastante flexível na modelagem de dados espaço-temporais não Gaussianos estendendo diversas abordagens existentes na literatura. Esta classe de modelos preserva as propriedades e características descritas para o caso dos MFED.

Condicional ao número fatores, a inferência a posteriori é facilitada através do algoritmo MCMC. Em geral, fazer inferência para os estados latentes de um modelo dinâmico não linear e não normal não é trivial. Diversas abordagens têm sido propostas na literatura para contornar este problema, muitas delas referidas principalmente ao caso de observações univariadas. Entre as técnicas existentes para o caso multivariado podem ser mencionadas as encontradas em [Gamerman \(1998\)](#) e [Knorr-Held \(1999\)](#). Ambas foram testadas neste trabalho além de tentar a amostragem individual com passos de Metropolis-Hastings. As duas primeiras foram implementadas considerando a amostragem dos fatores conjuntamente. O principal problema observado foi a baixa taxa de aceitação do algoritmo de M-H, fato que era esperado por se tratar de um esquema em bloco. A amostragem individual não teve sucesso devido à alta correlação das cadeias resultantes. Para contornar este problema, um

esquema que permite a amostragem dos fatores em B blocos foi sugerida neste capítulo. O algoritmo proposto se mostra bastante eficiente com taxas de aceitação razoáveis além de obter correlações baixas das cadeias. Este esquema de amostragem foi incorporado dentro do algoritmo MCMC proposto. Adicionalmente o algoritmo RJMCMC proposto na Seção [2.4.3](#) foi estendido para esta classe de modelos. Os resultados com dados simulados mostram que estas técnicas funcionam podendo ser aplicáveis a situações reais com estrutura complexa.

Capítulo 6

Aplicações

Neste capítulo duas aplicações para modelar ocorrências de chuva são apresentadas. A primeira aplicação trata dados de chuva diários medidos em 19 estações monitoradoras situadas na região norte da Oceania. A segunda aplicação refere-se ao estado de Minas Gerais, onde conta-se com informação de 17 estações pluviométricas espalhadas em todo o estado. Com estes dados, modelos espaço-temporais com resposta Bernoulli, derivados da classe dos MFEDG, são utilizados no ajuste dos dados. Com esta metodologia espera-se encontrar padrões espaciais que indiquem, por exemplo, regiões mais ou menos chuvosas ou grupos de estações com características semelhantes.

A modelagem de ocorrência de chuva é de grande importância no estudo de fenômenos climatológicos dado que pode ajudar a entender a estrutura de probabilidade da precipitação. Estes modelos podem ser utilizados para simular processos em que a probabilidade de ocorrência é tratada como variável de entrada (*input*) para modelar enchentes e vazão ou de forma geral modelos hidrológicos. Os dados de chuva, tanto do ponto de vista da ocorrência como da quantidade, já foram analisados e estudados na literatura. Uma proposta é que os dados sejam analisados mediante a obtenção das frequências observadas (dias chuvosos) como feito em [Stern & Coe \(1984\)](#). Uma das abordagens pioneiras para descrever a ocorrência de dias com ou sem chuva através de um modelo matemático é a proposta por [Gabriel & Neumann \(1962\)](#) em que dados de chuva em Israel foram analisados. Eles encontraram que existia uma forte persistência entre dias chuvosos consecutivos e obtiveram um bom ajuste ao utilizar modelos baseados na cadeia de Markov de primeira ordem. Posteriormente estas idéias foram estendidas e na década passada [Hughes, Guttorp & Charles \(1999\)](#) propõem o denominado *non-homogeneous hidden Markov model* (NHMM) que relaciona medidas atmosféricas à ocorrência de chuva através de estados latentes que atuam como ligação entre as duas medidas. Uma abordagem mais recente na modelagem de chuva pode ser encontrada em

Fernandes et al. (2007). Eles propuseram o ajuste da quantidade de chuva utilizando uma mistura entre uma distribuição Bernoulli, com probabilidade de sucesso θ , e uma função de densidade contínua. Neste caso, θ representa de forma natural a probabilidade de ocorrência de chuva.

Neste capítulo, o interesse é modelar as ocorrências de chuva diárias para entender a estrutura espaço-temporal da probabilidade de chuva utilizando a estrutura do MFEDG. A modelagem proposta permite capturar correlações espaciais entre estações e a dinâmica temporal das probabilidades. O estudo de duas regiões (norte da Oceania e o estado de Minas Gerais) totalmente diferentes, tanto em características climatológicas como em tamanho geográfico, permite avaliar a capacidade do modelo de capturar padrões espaço-temporais em diferentes contextos.

6.1 Eventos de chuva na região norte da Oceania

Nesta seção são ajustados modelos espaço-temporais aos eventos de chuva diários ocorridos em 19 locais situados na região norte da Oceania. Os dados utilizados na análise pertencem ao projeto *PACRAIN* (do inglês *Pacific Rainfall Data Base*). O *PACRAIN* coleta medições diárias e mensais de precipitação em várias estações monitoradoras situadas principalmente em atóis e ilhas do Pacífico. Estes dados, assim como maiores detalhes do projeto e fontes de informação disponíveis podem ser encontrados no sitio web <http://pacrain.evac.ou.edu/>.

No total foram selecionadas 19 estações considerando aquelas com menos observações faltantes. Mais especificamente, 14 estações pertencem aos Estados Federados da Micronésia (EFM) e as outras 5 às Ilhas Marshall (IM). Esta região se caracteriza por ter um clima cálido e úmido, com fortes chuvas durante todo o ano, especialmente nas ilhas orientais, e com risco de tufões no período de junho a dezembro.

Para cada estação tem-se dados de precipitação diária medidas em milímetros (*mm*) para todos os dias do ano 2001, ou seja, 19 séries de tamanho 365. Na Tabela 6.1 encontra-se a lista completa das estações, o grupo/região à qual pertencem e o nome da estação. A Figura 6.1 mostra a localização das estações monitoradoras no mapa e a área de estudo delimitada pelo polígono. As estações EFM14 e IM5 foram retiradas da análise para avaliar a capacidade de interpolação espacial dos modelos propostos e, portanto, as restantes 17 estações são consideradas na modelagem.

Tabela 6.1: *Chuva na Oceania:* Lista das estações monitoradoras de chuva selecionadas.

Estação	Grupo	Nome da Estação	Estação	Grupo	Nome da Estação
EFM1	EFM / Yap	Gilman Elem. School	EFM11	EFM / Pohnpei	Metelanim (Ponape)
EFM2	EFM / Yap	Maap	EFM12	EFM / Kosrae	Utwa
EFM3	EFM / Yap	Ulithi	EFM13	EFM / Kosrae	Lele (Kosrae)
EFM4	EFM / Yap	Woleai	EFM14*	EFM / Pohnpei	Pingelap
EFM5	EFM / Yap	Puluwat	IM1	Ilhas Marshall	Kwajalein
EFM6	EFM / Chuuk	Chuuk WSO AP	IM2	Ilhas Marshall	Wotje
EFM7	EFM / Chuuk	Lukunor	IM3	Ilhas Marshall	Laura (Majuro)
EFM8	EFM / Pohnpei	Kapingamarangi	IM4	Ilhas Marshall	Majuro
EFM9	EFM / Pohnpei	Nukuoro	IM5*	Ilhas Marshall	Ailinglapalap
EFM10	EFM / Pohnpei	Ponape			

* Locais de interpolação.

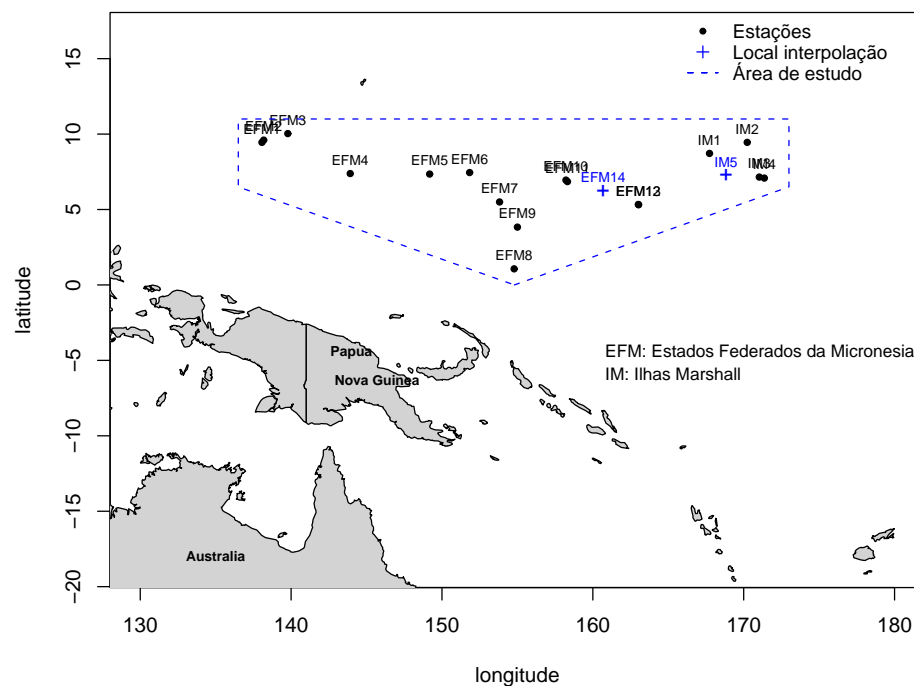


Figura 6.1: *Chuva na Oceania:* Localização das estações monitoradoras. O polígono delimita a região de estudo. As estações sinalizadas com + foram tiradas da análise para fazer interpolação espacial.

Nesta análise, o sucesso do evento é definido como a ocorrência de chuva tal que a precipitação total diária é maior ou igual a 1mm . De acordo com esta definição, foram construídas séries temporais com valores 0 (ausência de chuva) e 1 (ocorrência de chuva) para cada estação para serem utilizadas na modelagem espaço-temporal. Além disso, as coordenadas geográficas de latitude e longitude, que indicam a localização de cada estação, foram convertidas a coordenadas UTM (do inglês *universal transverse Mercator*). As coordenadas convertidas são medidas em quilômetros onde pode-se verificar que a máxima distância entre

estações é de 3.727,21 quilômetros (distância entre as estações EFM1 e IM4).

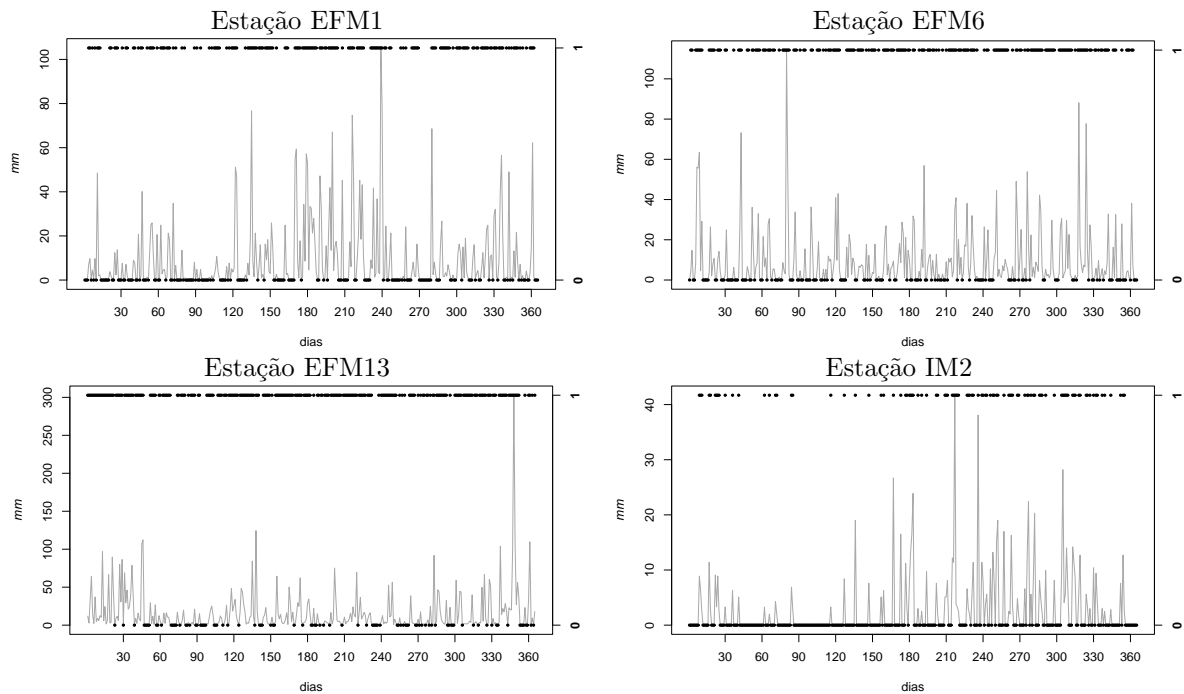


Figura 6.2: *Chuva na Oceania:* Valores diários da quantidade de chuva e do evento no ano 2001 de quatro estações monitoradoras. A linha cheia representa a precipitação diária medida em milímetros (mm). Os pontos representam a ocorrência dos eventos com valores 0–1.

6.1.1 Análise descritiva dos dados

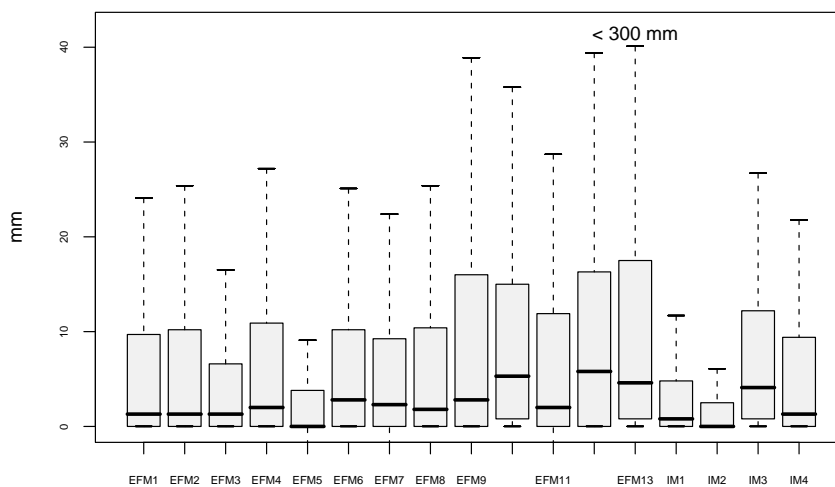
A análise exploratória dos dados constatou que três das séries contêm observações faltantes. Este fato pode ser resolvido durante o procedimento de inferência em que as observações faltantes podem ser tratadas como parâmetros do modelo. A Figura 6.2 mostra as séries temporais da precipitação diária e os eventos correspondentes a quatro estações monitoradoras denotadas por EFM1, EFM6, EFM13 e IM2. Nota-se que as séries de precipitação diária não apresentam ciclos ou padrões sazonais ao longo do tempo.

Numa primeira análise foram calculadas algumas medidas resumo como o máximo, média e desvio padrão da quantidade de chuva (em mm) e o número de dias chuvosos ao longo do ano. Vale a pena ressaltar que as estações EFM7 e EFM11 tem 1 e 34 observações faltantes respectivamente e, portanto, as medidas resumo não podem ser diretamente comparadas com as estações restantes. A Tabela 6.2 mostra estas medidas para cada estação, a Figura 6.3(a) os *box plots* da precipitação de chuva diária por estação e as Figuras 6.3(b)-(c) o total de dias

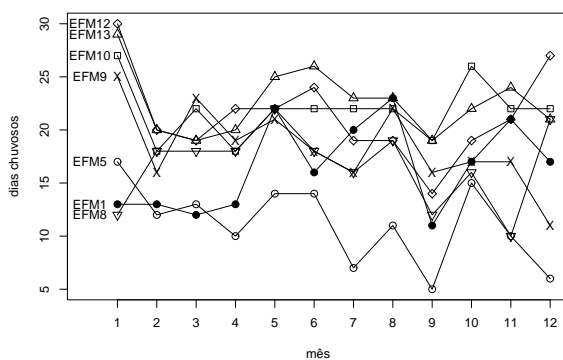
Tabela 6.2: *Chuva na Oceania:* Medidas resumo (máximo, média e desvio padrão) da quantidade de chuva (em mm) das 17 estações monitoradoras analisadas. As colunas “Nº dias” indicam o total de dias com chuva no ano de estudo.

Estação	Máximo	Média	D.P.	Nº dias	Estação	Máximo	Média	D.P.	Nº dias
EFM1	105,20	8,66	15,52	198	EFM10	130,60	12,13	18,94	262
EFM2	127,00	8,42	15,39	207	EFM11*	177,30	8,60	15,30	196
EFM3	128,80	8,04	16,97	194	EFM12	132,10	13,62	21,00	256
EFM4	135,90	9,05	16,32	203	EFM13	302,80	15,36	27,32	271
EFM5	110,70	5,00	11,87	134	IM1	96,00	5,27	10,44	174
EFM6	114,30	8,92	14,69	221	IM2	41,70	2,43	5,45	102
EFM7*	113,30	8,04	14,71	232	IM3	78,70	9,20	13,07	266
EFM8	109,50	9,99	18,40	200	IM4	137,40	8,66	17,91	199
EFM9	115,80	12,02	19,75	221					

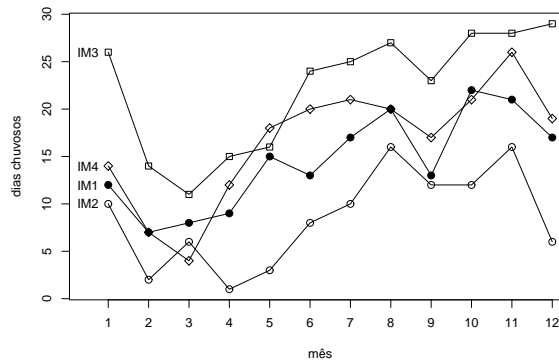
* Locais com observações faltantes.



(a) Gráficos *box plot*



(b) Estações EFM



(c) Estações IM

Figura 6.3: *Chuva na Oceania:* (a) *Box plots* da quantidade de chuva em milímetros (*mm*) para cada estação monitoradora. (b)-(c) Gráficos do total de dias chuvosos por mês e por estação monitoradora.

chuvosos por mês para 7 estações do grupo EFM e para todas as estações do grupo IM. Com estes resultados, algumas características relevantes podem ser mencionadas:

- Em quase todas as estações o total de dias com chuva (em todo o período) representa mais da terça parte do ano, exceto na estação IM2.
- As estações com maior precipitação média diária e mais dias chuvosos são: EFM10, EFM12 e EFM13 (no primeiro grupo) e IM3 e IM4 (no segundo grupo). Aqui pode-se afirmar que dentro de cada grupo as estações que ficam na região leste são as mais chuvosas.
- As estações EFM5 e IM2 apresentam em média os menores valores de precipitação diária ($5,00mm$ e $2,43mm$ respectivamente) e são também as estações com menos dias chuvosos (134 e 102 dias respectivamente).
- No grupo de estações EFM, a partir do mês de fevereiro, as séries de dias chuvosos por mês não apresentam uma tendência clara. Embora todas as estações apresentem um decréscimo no mês de setembro.
- No grupo de estações IM, as séries de dias chuvosos por mês apresentam um decréscimo nos meses de fevereiro e março indicando o período menos chuvoso do ano nessa região. A partir do mês de abril os valores se incrementam até o final do ano com uma pequena queda no mês de setembro.

Os pontos acima mencionados indicam características tanto espaciais como temporais nos dados, que sugerem a utilização de modelos espaço-temporais. Na subseção seguinte são apresentadas duas classes de modelos derivadas do MFEDG para os eventos de chuva assim como modelos mais simples comumente utilizados na literatura (modelos benchmark). Estes modelos, que não consideram fatores nem cargas espacialmente estruturadas na função média das observações, são comparados com os modelos propostos nesta tese.

6.1.2 Modelos utilizados

Diversas abordagens para modelar eventos de chuva têm sido propostas na literatura. Por exemplo [Kitagawa \(1987\)](#) e [Gamerman \(1998\)](#) modelaram o número de ocorrências de chuva diárias (precipitação maior de $1mm$) na cidade de Tóquio no período 1983–1984. Para este

conjunto de dados o objetivo foi estimar a probabilidade de ocorrência de chuva para cada dia do ano utilizando, para este propósito, modelos lineares dinâmicos generalizados para dados com resposta binomial. Nesta análise, ao invés de fazer uma modelagem puramente temporal, a idéia é utilizar o MFEDG com resposta Bernoulli para estimar a probabilidade de ocorrência de chuva levando em consideração a correlação espacial existente entre as estações monitoradoras. As vantagens são: (i) identificar através dos fatores latentes grupos de estações com características semelhantes e (ii) a capacidade do modelo para fazer interpolação espacial.

O modelo geral considerado tem resposta Bernoulli com função de ligação *logito* na média, isto é:

$$y_{it}|p_{it} \sim \text{Bernoulli}(p_{it}), \quad i = 1, \dots, 17, \quad t = 1, \dots, 365, \quad (6.1a)$$

$$\log(p_{it}/(1 - p_{it})) = \theta_{it}, \quad (6.1b)$$

$$\theta_t = \mu_t^{y*} + \beta f_t, \quad (6.1c)$$

$$f_t = \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \quad (6.1d)$$

$$\beta_{(j)} = \mu_j^{\beta*} + \delta_j, \quad \delta_j \sim N(0, \tau_j^2 R_{\phi_j}) \quad (6.1e)$$

$$f_0 \sim N(m_0, C_0)$$

onde $\theta_t = (\theta_{1t}, \dots, \theta_{17,t})'$ é o preditor linear, $f_t \in \mathbb{R}^m$ (m -número de fatores) e $\beta_{(j)} \in \mathbb{R}^{17}$ representa a j -ésima coluna da matriz de cargas dos fatores ($j = 1, \dots, m$). $R_{\phi_j} \in \mathbb{R}^{17 \times 17}$ é a matriz de correlação espacial com elementos $r_{lk} = \rho_{\phi_j}(|s_l - s_k|)$, ($l, k = 1, \dots, 17$) tal que $\rho_{\phi_j}(\cdot)$ é a função Matérn com parâmetro de ordem igual a 1.

Os vários modelos ajustados variam de acordo ao número de fatores considerado e a especificação das componentes μ_t^{y*} e $\mu_j^{\beta*}$. Desta forma, duas classes de modelos foram consideradas:

1. MFEDG(m): modelo fatorial espacial dinâmico generalizado com m fatores com $\mu_t^{y*} = 0$ e média do PG $\mu_j^{\beta*} = \mu_j \mathbf{1}_{17}$ tal que:

$$\theta_t = \beta f_t \quad \text{e} \quad \beta_{(j)} = \mu_j \mathbf{1}_{17} + \delta_j.$$

2. MFEDG(m)-CC: modelo fatorial espacial dinâmico generalizado com m fatores e componente comum no preditor linear onde $\mu_t^{y*} = \mu_t^y \mathbf{1}_{17}$, $\mu_t^y \sim N(\mu_{t-1}^y, W)$, $\mu_0^y \sim N(m_0^\mu, C_0^\mu)$

e média do PG $\mu_j^{\beta^*} = \mu_j \mathbf{1}_{17}$ tal que:

$$\begin{aligned}\theta_t &= \mu_t^y \mathbf{1}_{17} + \beta f_t, \\ \mu_t^y &= \mu_{t-1}^y + \nu_t, \quad \nu_t \sim N(0, W), \\ \beta_{(j)} &= \mu_j \mathbf{1}_{17} + \delta_j.\end{aligned}$$

Modelos benchmark

Para efeitos de comparação foram ajustados dois modelos com estruturas mais simples, comumente utilizados na modelagem de dados na família exponencial com variação temporal e espaço-temporal. O primeiro refere-se a um modelo temporal sem estrutura espacial com componente comum no preditor linear. O segundo refere-se a um modelo espaço-temporal sem considerar fatores comuns nem cargas espacialmente estruturadas no preditor linear. Estes modelos diferem do modelo geral definido em (6.1) nas equações (6.1c)-(6.1e) de acordo com as seguintes especificações:

1. MDG-CC: modelo dinâmico generalizado com componente comum no preditor linear e sem estrutura espacial tal que:

$$\begin{aligned}\theta_t &= \mu_t^y \mathbf{1}_{17}, \\ \mu_t^y &= \mu_{t-1}^y + \nu_t, \quad \nu_t \sim N(0, W), \\ \mu_0^y &\sim N(m_0^\mu, C_0^\mu).\end{aligned}$$

2. MEDG-CC: modelo espacial dinâmico generalizado com componente comum no preditor linear tal que:

$$\begin{aligned}\theta_t &= \mu_t^y \mathbf{1}_{17} + \delta_t, \\ \mu_t^y &= \mu_{t-1}^y + \nu_t, \quad \nu_t \sim N(0, W), \\ \delta_t &\sim PG(0, \tau^2 \rho_\phi(\cdot)), \\ \mu_0^y &\sim N(m_0^\mu, C_0^\mu).\end{aligned}$$

Vale a pena ressaltar que o modelo MEDG-CC é um caso particular do modelo MFEDG(1)-CC ao considerar o fator comum com valor constante igual a 1.

Para as classes MFEDG(m) e MFEDG(m)-CC foram testados modelos com 1, 2, 3 e 4 fatores. Entretanto, só serão reportados resultados considerando no máximo 3 fatores. Isto devido a que o quarto fator em ambas as classes resultou não significativo,

assim como os parâmetros relacionados a este. As distribuições a priori utilizadas em todos os casos foram relativamente vagas. Assim, para $j = 1, \dots, m$, considerou-se $\lambda_j \sim GI(0, 01; 0, 01)$, $\mu_j \sim N(0; 1)$, $\tau_j^2 \sim IG(2; 0, 5)$, $W \sim GI(0, 01; 0, 01)$ e $\phi_j \sim GI(2; b)$ onde $b = \max(dist)/(-2 \log(0, 05))$ tal que $\max(dist)$ é a distância máxima entre locais. Para o parâmetro γ_j utilizou-se a priori mista definida na Subseção 2.4.1 com $\alpha = 0, 5$, $m_\gamma = 0$ e $S_\gamma = 1$ que permite a inclusão de fatores não estacionários no modelo. As priors para f_0 e μ_0 foram $N(0_m, I_m)$ e $N(0, 1)$ respectivamente. A amostragem dos fatores foi feita em blocos como descrita na Subseção 5.3.3 com 10 e 15 blocos para a primeira e segunda classe de modelos, respectivamente. Finalmente, para os modelos MDG-CC e MEDG-CC foram utilizadas as mesmas distribuições a priori da classe MFEDG(m)-CC nos parâmetros com a mesma definição e estrutura. Neste caso, a amostragem dos estados ($\mu_t, t = 1, \dots, T$) também foi feita em blocos, mais especificamente 10 blocos para cada modelo com $m_0^\mu = 0$ e $C_0^\mu = 1$.

O algoritmo MCMC foi utilizado para obter amostras a posteriori de todos os parâmetros do modelo. Para cada parâmetro foram geradas duas cadeias (partindo de diferentes valores iniciais) de tamanho 30.000 descartando-se as primeiras 10.000 iterações como período de aquecimento e guardando os valores gerados a cada 5 iterações para quebrar possíveis autocorrelações nas amostras resultantes. Portanto, os resultados a posteriori dos parâmetros estão baseados em amostras de tamanho 4.000. Para verificar a convergência das cadeias foi utilizada a estatística \hat{R} de Gelman & Rubin (1992) onde valores próximos de 1 sugerem convergência das cadeias.

6.1.3 Principais resultados

A Tabela 6.3 mostra os critérios de comparação de modelos calculados para todos os modelos propostos. Neste caso utilizou-se o erro quadrático médio (EQM), erro absoluto médio (EAM) o logaritmo da verossimilhança (LV) e a probabilidade a posteriori do modelo (PPM) obtida utilizando o algoritmo RJMCMC. Os valores EQM e EAM foram calculados para os valores ajustados dentro da amostra e para os valores interpolados nas estações EFM14 e IM5. Dos resultados obtidos podemos inferir que para as duas primeiras especificações, o modelo com três fatores é o melhor dentro de cada classe, com probabilidades a posteriori 0,46 e 0,40 respectivamente. Observa-se também que, entre os modelos com maior PPM, a especificação MFEDG(3) apresenta o melhor resultado segundo os critérios LV, EQM e EAM, embora os critérios EQM e EAM, baseados nos valores interpolados, indiquem como melhor o

Tabela 6.3: *Chuva na Oceania:* Critérios de comparação de modelos para todas as classes de modelos especificadas. Erro Quadrático Médio (EQM), Erro Absoluto Médio (EAM), Logaritmo da Verossimilhança (LV) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *itálico*.

Modelo	m	EQM ¹	EAM ¹	EQM ²	EAM ²	LV	PPM
MFEDG(m)	1	0,2162	0,4391	0,2119	0,4158	-3844,4	0,28
	2	0,2037	0,4173	<i>0,1987</i>	<i>0,3922</i>	-3709,7	0,26
	3	<i>0,1901</i>	<i>0,3990</i>	0,2069	0,4018	<i>-3583,9</i>	<i>0,46</i>
MFEDG(m)-CC	1	0,2068	0,4227	0,1987	0,4058	-3742,6	0,23
	2	0,2000	0,4118	0,1986	0,3968	-3679,3	0,35
	3	<i>0,1979</i>	<i>0,4045</i>	<i>0,1982</i>	<i>0,3950</i>	<i>-3646,5</i>	<i>0,42</i>
MDG-CC	-	0,2233	0,4525	-	-	-3951,7	-
MEDG-CC	-	0,2098	0,4258	0,2015	0,4108	-3768,0	-

¹ Com valores ajustados dentro da amostra

² Com valores interpolados nas estações EFM14 e IM5.

modelo MFEDG(3)-CC. Entre os modelos benchmark, a especificação MEDG-CC apresenta os melhores resultados sugerindo que a inclusão de estrutura espacial nos dados tem uma melhor performance no ajuste. Contudo, segundo todos os critérios de comparação, exceto o EQM dos valores ajustados, a especificação MFEDG(3) apresenta melhores resultados. Isto sugere que um modelo espaço-temporal mais estruturado, como o proposto nesta tese, é necessário na modelagem deste conjunto de dados.

Baseado nos resultados anteriores, o modelo final considerado é MFEDG(3), ou seja, um modelo com 3 fatores no preditor linear e média comum nas cargas espacialmente estruturadas. A Tabela 6.4 mostra as estatísticas a posteriori dos parâmetros que caracterizam a variação temporal dos fatores comuns (primeiro grupo) e dos parâmetros do processo Gaussiano (segundo grupo) que caracterizam a dependência espacial através das cargas dos fatores. O primeiro grupo de parâmetros indica a presença de dois fatores estacionários (fatores 1 e 3) e um fator não estacionário (fator 2) com as seguintes probabilidades a posteriori: $\hat{p}(\gamma_1 = 1|y) = 0,006$, $\hat{p}(\gamma_2 = 1|y) = 0,899$ e $\hat{p}(\gamma_3 = 1|y) = 0,025$. As taxas de aceitação dos 10 blocos considerados na amostragem dos fatores foram 0,495, 0,118, 0,154, 0,180, 0,153, 0,152, 0,145, 0,140, 0,130 e 0,139 respectivamente. Os resultados para o segundo grupo de parâmetros indicam que a média a posteriori de cada processo Gaussiano é não significativa,

Tabela 6.4: *Chuva na Oceania:* Sumário a posteriori dos parâmetros do modelo MFEDG(3) que caracterizam os fatores comuns e as cargas espacialmente estruturadas. T.A.: Taxa de aceitação. \hat{R} : diagnóstico de Gelman & Rubin.

θ	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			T.A.	\hat{R}
			2,5%	50%	97,5%		
γ_1	0,629	0,120	0,417	0,616	0,844	-	1,06
γ_2	0,991	0,030	0,887	1,000	1,000	-	1,04
γ_3	0,864	0,041	0,785	0,846	0,965	-	1,02
λ_1	0,204	0,060	0,101	0,207	0,330	-	1,02
λ_2	0,146	0,038	0,084	0,140	0,238	-	1,10
λ_3	0,108	0,021	0,075	0,105	0,157	-	1,03
μ_1	-0,256	0,330	-0,943	-0,246	0,380	-	1,01
μ_2	0,460	0,240	-0,013	0,461	0,918	-	1,00
μ_3	0,379	0,310	-0,242	0,381	0,976	-	1,01
τ_1^2	0,758	0,360	0,313	0,673	1,700	-	1,05
τ_2^2	0,618	0,320	0,243	0,533	1,370	-	1,02
τ_3^2	0,667	0,320	0,276	0,588	1,490	-	1,02
ϕ_1	248,440	175,990	59,589	193,530	690,850	0,630	1,11
ϕ_2	102,780	44,477	50,035	93,740	222,630	0,605	1,05
ϕ_3	232,350	116,490	89,354	208,300	543,930	0,635	1,05

com valores próximos de zero. Os parâmetros ϕ_i , ($i = 1, 2, 3$) indicam que a média a posteriori da correlação espacial das cargas entre locais situados a 500 quilômetros de distância são 0,172, 0,015 e 0,201 respectivamente. Este resultado sugere que os fatores 1 e 3 têm maior influência em locais mais próximos. A taxa de aceitação do algoritmo de Metropolis-Hastings para β foi 0,31. Adicionalmente a Tabela 6.4 mostra a estatística \hat{R} (Gelman & Rubin, 1992) com valores próximos de 1 o que sugere que a convergência das cadeias dos parâmetros foi atingida.

A Figura 6.4 mostra os fatores estimados ao longo do tempo e os intervalos de credibilidade a posteriori de 95%. Na Figura 6.4(a) o primeiro fator exibe um padrão bastante irregular ao longo do tempo com uma tendência levemente decrescente nos primeiros 120 dias do ano. Na Figura 6.4(b) o segundo fator exibe o comportamento das estações mais chuvosas com valores altos nos primeiros dias do ano, uma tendência decrescente até meados de fevereiro e uma queda significativa no mês de setembro. Característica que foi percebida na análise exploratória de dados. Nota-se também uma variação intermensal com valores altos no início de cada mês. Na Figura 6.4(c) o terceiro fator mostra o comportamento comum do grupo de estações IM com uma tendência decrescente até o mês de fevereiro e depois crescente até o final do ano como mostrado na Figura 6.3(c).

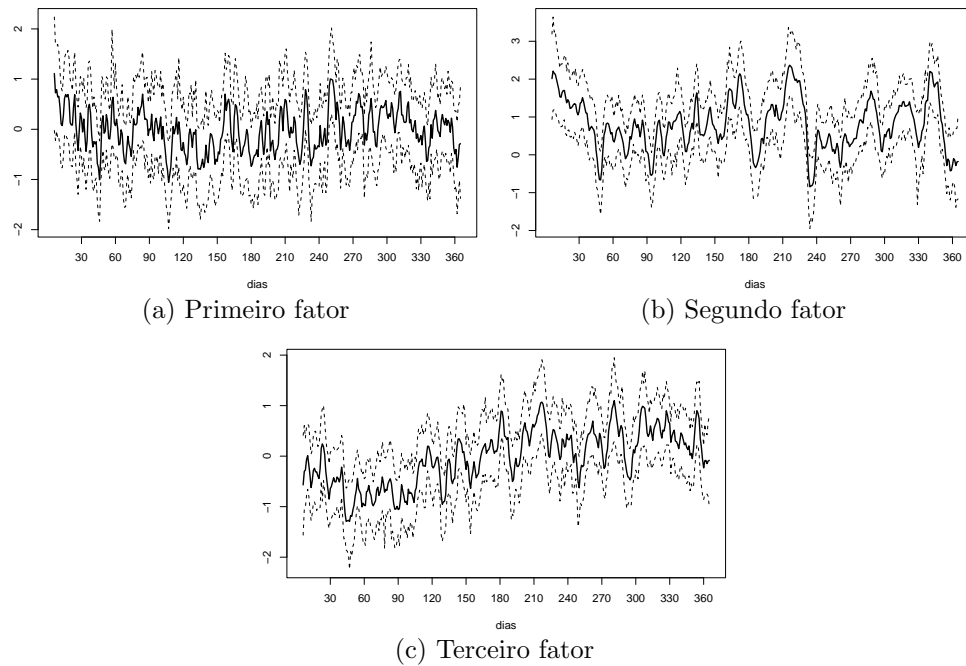


Figura 6.4: Chuva na Oceania: Fatores estimados para a especificação MFEDG(3). A linha cheia representa a média a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

A Figura 6.5 mostra a superfície média de cada coluna da matriz de cargas, estimada via interpolação (*krigagem*) Bayesiana. As cargas do primeiro fator indicam uma variação suave na região, com valores altos ao redor das estações EFM5 e IM2 (estações com menos dias chuvosos) e valores baixos na região nordeste. A superfície das cargas do segundo fator não é suave e indica basicamente a influência das estações mais chuvosas a regiões mais próximas, aqui as estações EFM5 e IM2 apresentam pesos menores tirando a influência das estações menos chuvosas do segundo fator. O mapa das cargas do terceiro fator mostra uma clara diferença entre grupos de estações, neste caso pode-se afirmar que o fator representa o comportamento das estações IM tirando a influência das outras estações na região.

A Figura 6.6 mostra as probabilidades a posteriori e os intervalos de credibilidade de 95% de quatro estações monitoradoras. Em cada gráfico observa-se que as probabilidades estimadas são muito variáveis ao longo do tempo, indicando probabilidades altas nos dias de ocorrência de chuva. Por exemplo, na estação EFM13 as probabilidades são altas nos primeiros dias do ano indicando o período mais chuvoso do ano. Já na estação IM2 as probabilidades estimadas são baixas nos primeiros três meses com uma tendência crescente nos últimos meses do ano. Por último, a Figura 6.7 mostra as probabilidades a posteriori

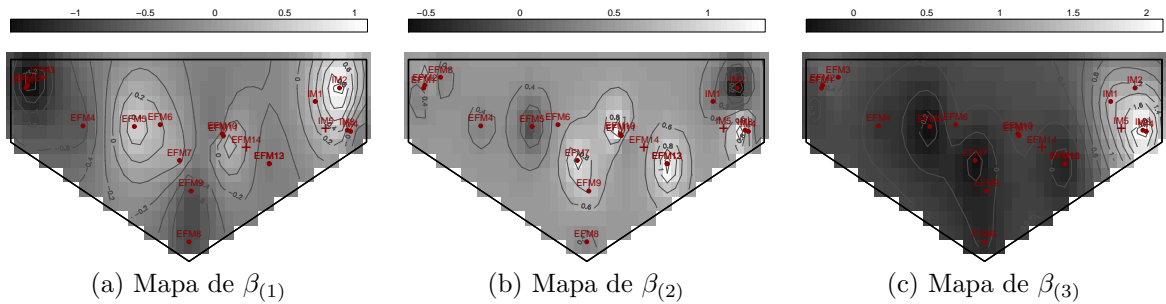


Figura 6.5: *Chuva na Oceania:* Interpolação Bayesiana nas cargas dos fatores. Os valores acima de cada gráfico indicam o intervalo de variação das médias a posteriori.

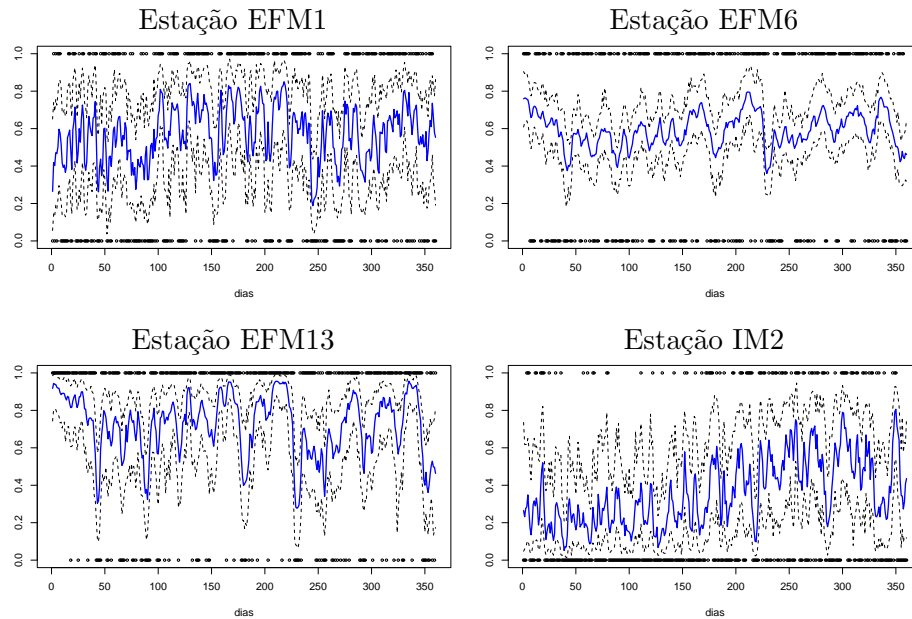


Figura 6.6: *Chuva na Oceania:* Probabilidades ajustadas de ocorrência de chuva em quatro estações monitoradoras. A linha cheia representa a média a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

estimadas para os dois locais que foram retirados da análise. Estes resultados foram obtidos utilizando os resultados da interpolação espacial da matriz de cargas dos fatores. Para a estação EFM14 observa-se a influência do fator 2 e na estação IM5 a influência do fator 3.

6.2 Eventos de chuva no estado de Minas Gerais

Nesta seção novamente são ajustadas duas classes de modelos derivadas do MFEDG para as ocorrências de chuva diárias medidas em 17 estações meteorológicas automáticas situadas

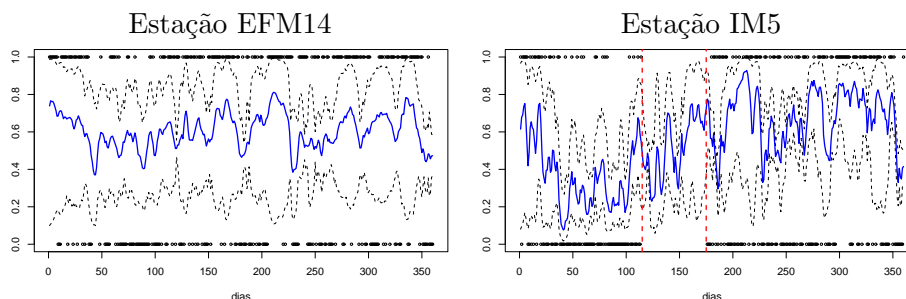


Figura 6.7: *Chuva na Oceania:* Interpolação espacial nas estações EFM14 e IM5. A linha azul representa a probabilidade de ocorrência de chuva e as linhas pontilhadas o intervalo de credibilidade de 95%. Em IM5 as linhas verticais indicam o período com dados faltantes.

no estado de Minas Gerais. Estes dados foram disponibilizados pela rede meteorológica e hidrológica telemétrica do SIMGE (Sistema de Meteorologia e Recursos Hídricos de Minas Gerais) que realiza transmissões de dados a cada 3 horas, através de dois satélites brasileiros SCD-1 e SCD-2. Maiores detalhes da rede, assim como os dados utilizados nesta análise, podem ser encontrados no sitio web <http://www.simge.mg.gov.br>.

Os dados analisados correspondem ao ano 2005, em que para cada dia do ano tem-se a precipitação total em milímetros (mm). Com esta informação foram construídas 17 séries temporais que determinam a ocorrência do evento chuva com valores 1 (se a precipitação total for maior a $1mm$) e 0 em caso contrário. A Tabela 6.5 mostra a lista das 17 estações monitoradoras e o município ao qual pertencem e a Figura 6.8 a localização geográfica das estações no mapa. A área de estudo compreende todo o estado de Minas Gerais. As estações Pirapora (PI) e Caratinga (CA) foram retiradas da análise para avaliar a capacidade de interpolação espacial do modelo e, portanto, somente as primeiras 15 estações são consideradas na modelagem espaço-temporal. Da mesma forma que a aplicação anterior, as coordenadas de latitude e longitude foram transformadas a coordenadas UTM nas quais pode-se verificar que a distância máxima entre estações é de 1.036,38 quilômetros, mais especificamente entre as estações Honorópolis (HO) e Nanuque (NA).

Características do estado de Minas Gerais

O estado de Minas Gerais tem uma extensão territorial de 586.528 km^2 e localiza-se no Sudeste do Brasil. Os climas predominantes são o tropical (temperaturas elevadas todos os meses do ano) e tropical de altitude (com ocasionais geadas e muito raramente precipitações de neve). As regiões mais altas e o sul do estado apresentam as temperaturas mais baixas,

chegando a atingir marcações próximas de 0°C . Nas regiões sul, sudeste, leste e central do estado são registrados os maiores índices pluviométricos. Por outro lado, nas regiões norte e nordeste, as chuvas escassas e as altas temperaturas tornam essas regiões muito susceptíveis à seca.

Tabela 6.5: *Chuva em Minas Gerais:* Lista das 17 estações monitoradoras de chuva e medidas resumo (máximo, média e desvio padrão) da quantidade de chuva diária (em *mm*).

Est.	Nome	Máx.	Média	D.P.	Est.	Nome	Máx.	Média	D.P.
SF	Santa Fé	66,00	3,56	10,14	LP	Leopoldina	81,75	3,64	10,03
MC	Montes Claros	68,75	3,00	8,89	HO	Honorópolis	70,75	5,49	12,10
AC	Araçuaí	74,00	2,09	7,39	AR	Araxá	72,75	4,81	11,69
NA	Nanuque	64,25	2,89	8,44	LA	Lavras	48,75	3,74	8,16
PR	Paracatu	101,00	4,54	14,75	MA	Machado	100,50	3,22	9,71
AN	Andrequice	103,75	4,57	13,06	IT	Itajubá	68,00	3,64	8,81
IP	Ipatinga	114,75	3,84	11,42	PI*	Pirapora	127,25	3,43	12,56
BH	Belo Horizonte	74,75	3,51	9,65	CA*	Caratinga	75,50	3,96	10,77
VC	Viçosa	96,00	3,93	10,89					

* Locais de interpolação.

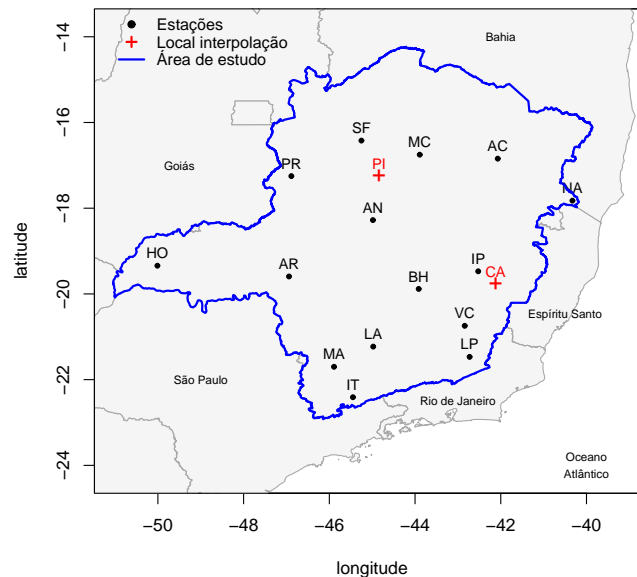


Figura 6.8: *Chuva em Minas Gerais:* Localização das 17 estações monitoradoras. A linha azul delimita a região de estudo (todo o estado de MG). As estações sinalizadas com + foram tiradas da análise para fazer interpolação espacial.

6.2.1 Análise descritiva dos dados

Com a análise descritiva dos dados pode-se verificar que todas as séries temporais têm observações faltantes principalmente nos meses de fevereiro, março e agosto. As estações com

mais observações faltantes são HO, SF, MA, NA, IT e PR com 91, 37, 37, 29, 26 e 23 dias respectivamente. A imputação destas observações é feita durante o procedimento de inferência no qual estes valores são considerados como parâmetros do modelo. A Figura 6.9 mostra as séries temporais da precipitação diária e do evento chuva das estações AC, AN, BH e MA. Nestes gráficos, pode-se perceber que as ocorrências de chuva são menos frequentes no período de abril a outubro, enquanto que os períodos mais chuvosos acontecem no início e no final do ano. A Tabela 6.5 mostra algumas medidas resumo (máximo, média e desvio padrão) da precipitação total diária para cada estação monitoradora, daqui pode-se verificar que: (i) as estações com menor precipitação média são MC, AC e NA as quais ficam no nordeste do estado e (ii) as estações com maior precipitação média são HO, AR (na região sudoeste), PR e AN (na região central).

A Figura 6.10 mostra o total de dias chuvosos por mês para todas as estações monitoradoras separadas em três grupos de acordo a sua localização geográfica. Nota-se que alguns valores não aparecem nos gráficos, isto devido a falta de informação em alguns meses. Algumas características relevantes são apresentadas a seguir:

- Em geral, todas as estações apresentam a mesma tendência, com mais dias chuvosos no início e no final do ano, mais especificamente nos períodos de janeiro a março e de novembro a dezembro, respectivamente.
- As estações do norte e nordeste (exceto a estação NA) apresentam menos dias chuvosos no mês de janeiro e no período de abril a outubro que as outras estações consideradas.
- As estações do centro e sudeste apresentam, em geral, mais dias chuvosos por mês que as estações do primeiro grupo.
- Por último, as estações do sul e sudoeste apresentam mais dias chuvosos nos meses de janeiro, setembro e outubro que as outras estações da rede.

Esta análise preliminar indica que as regiões sul e sudeste são as mais chuvosas e com os maiores índices pluviométricos, sugerindo uma tendência espacial na direção sul e sudeste. Fato que caracteriza o comportamento climático do estado de Minas Gerais.

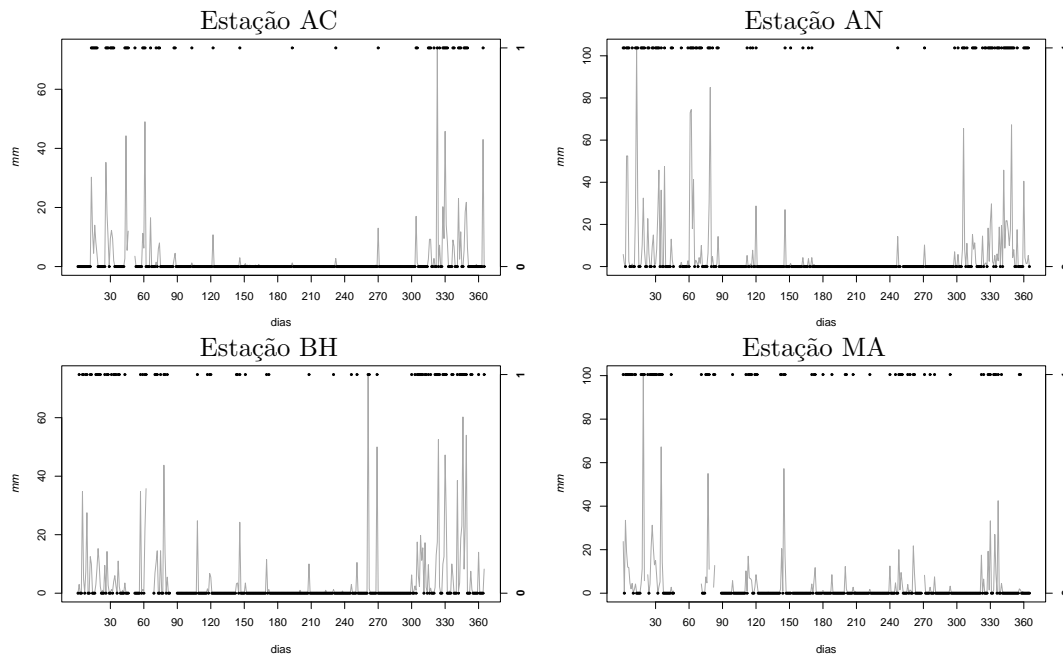


Figura 6.9: *Chuva em Minas Gerais:* Valores diários da quantidade de chuva e do evento no ano 2005 de quatro estações monitoradoras. A linha cheia representa a precipitação diária medida em milímetros (*mm*). Os pontos representam a ocorrência dos eventos com valores 0–1.

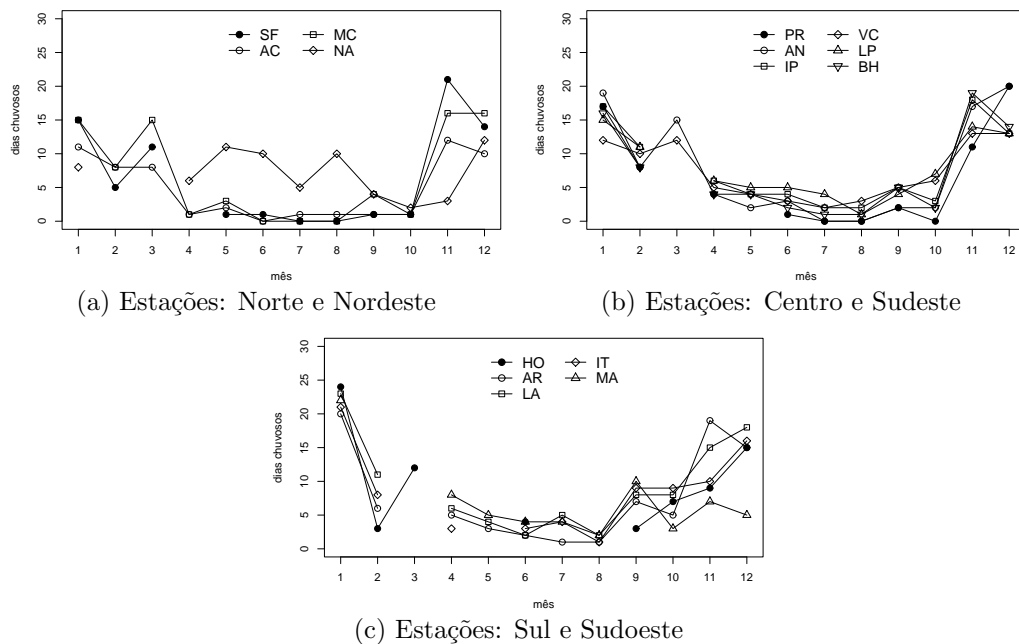


Figura 6.10: *Chuva em Minas Gerais:* Total de dias chuvosos por mês e por estação monitoradora. Cada gráfico representa grupo de estações localizadas em diferentes regiões do estado.

6.2.2 Modelos utilizados

Para este conjunto de dados novamente será utilizado o MFEDG com resposta Bernoulli para estimar a probabilidade de ocorrência de chuva diária do ano 2005 em todo o estado de Minas Gerais. O modelo espaço-temporal considerado segue as mesmas especificações das equações (6.1a)-(6.1e) onde

$$y_{it}|p_{it} \sim \text{Bernoulli}(p_{it}), \quad i = 1, \dots, 15, t = 1, \dots, 365.$$

A função de correlação espacial das cargas dos fatores é a Matérn com parâmetro de ordem igual a 1. De forma análoga à aplicação anterior as duas classes de modelos derivadas do MFEDG são:

1. MFEDG(m): modelo fatorial espacial dinâmico generalizado com m fatores com $\mu_t^{y*} = 0$ e média do PG $\mu_j^{\beta*} = \mu_j \mathbf{1}_{15}$.
2. MFEDG(m)-CC: modelo fatorial espacial dinâmico generalizado com m fatores, componente comum no preditor linear, $\mu_t^{y*} = \mu_t^y \mathbf{1}_{15}$ onde $\mu_t^y \sim N(\mu_{t-1}^y, W)$ e média do PG $\mu_j^{\beta*} = \mu_j \mathbf{1}_{15}$.

Adicionalmente, e para efeitos de comparação, foram ajustados dois modelos com estruturas mais simples (modelos benchmark) definidos na Subseção 6.1.2 denotados por:

1. MDG-CC: modelo dinâmico generalizado com componente comum no preditor linear e sem estrutura espacial.
2. MEDG-CC: modelo espacial dinâmico generalizado com componente comum no preditor linear.

As distribuições a priori consideradas foram as mesmas utilizadas para a aplicação de chuva em Oceania (ver Subseção 6.1.2). Para as classes MFEDG(m) e MFEDG(m)-CC foram testados modelos com 1, 2, 3 e 4 fatores. Entretanto, só serão apresentados resultados com até três fatores. Neste ponto, vale a pena ressaltar que modelos com mais de 3 fatores não foram bem estimados. Na amostragem em blocos dos fatores pode-se perceber que as taxas de aceitação para o algoritmo de Metropolis-Hastings foram diminuindo à medida que o número de fatores aumentava, sendo para alguns blocos praticamente nula. Para solucionar este problema foram considerados 15 e 20 blocos como forma de garantir taxas de aceitação

mais razoáveis. Os resultados com 20 blocos foram melhores em modelos com até 3 fatores em ambas as classes.

O algoritmo MCMC foi rodado com um total de 30.000 iterações, descartando-se as primeiras 10.000 e guardando os valores gerados a cada 5 iterações. As amostras resultantes de tamanho 4.000 foram utilizadas para calcular as principais estatísticas a posteriori. Para cada um dos parâmetros, duas cadeias foram geradas partindo de diferentes pontos do espaço paramétrico. Finalmente a convergência das cadeias foi testada utilizando o critério de [Gelman & Rubin \(1992\)](#).

6.2.3 Principiais resultados

A Tabela 6.6 mostra os diferentes critérios de comparação de modelos baseados nas estatísticas EQM, EAM e LV assim como a probabilidade a posteriori calculada do algoritmo RJMCMC. Para a primeira e segunda classe de modelos as especificações MFEDG(3) e MFEDG(2)-CC apresentam os melhores resultados em termos de ajuste além de apresentar as probabilidades a posteriori mais altas, 0,46 e 0,57 respectivamente. Entre estes dois modelos, todos os critérios de comparação sugerem a escolha da especificação MFEDG(2)-CC em termos de ajuste e interpolação. Ainda na Tabela 6.6, são apresentados os resultados dos modelos benchmark. Nota-se que a especificação MEDG-CC tem uma melhor performance que o modelo MDG-CC. Entretanto, a especificação MFEDG(2)-CC continua sendo a de melhor performance em termos de ajuste e interpolação espacial.

Na Tabela 6.7 são apresentadas as medidas resumo a posteriori dos hiperparâmetros do modelo final e a estatística \hat{R} com valores próximos de 1 sugerindo a convergência de todas as cadeias. A variância da componente comum μ_t é menor que as variâncias dos fatores. Segundo os resultados, ambos os fatores são estacionários com probabilidades a posteriori $\hat{p}(\gamma_1 = 1|y) = \hat{p}(\gamma_2 = 1|y) = 0$. As taxas de aceitação para os 20 blocos considerados na amostragem dos fatores variam de 0,08 a 0,49, enquanto que a taxa de aceitação das cargas é de 0,27. As médias a posteriori dos processos Gaussianos são não significativas. Os parâmetros ϕ_1 e ϕ_2 indicam que a correlação espacial das cargas dos fatores distantes em 200 quilômetros são 0,643 e 0,796 respectivamente e, portanto, sugere a presença de uma forte correlação espacial.

A Figura 6.11 apresenta as trajetórias estimadas da componente comum e dos fatores, assim como os intervalos de credibilidade a posteriori. No primeiro gráfico nota-se que a

Tabela 6.6: *Chuva em Minas Gerais:* Critérios de comparação de modelos para todas as classes de modelos especificadas. Erro Quadrático Médio (EQM), Erro Absoluto Médio (EAM), Logaritmo da Verossimilhança (LV) e Probabilidade a Posteriori do Modelo (PPM). Os melhores modelos para cada critério aparecem em *itálico*.

Modelo	m	EQM ¹	EAM ¹	EQM ²	EAM ²	LV	PPM
MFEDG(m)	1	0,1243	0,2530	0,1045	0,2285	-2118,0	0,15
	2	0,1057	0,2219	<i>0,0973</i>	<i>0,2164</i>	-1908,9	0,39
	3	<i>0,1039</i>	<i>0,2175</i>	0,1024	0,2239	<i>-1881,1</i>	<i>0,46</i>
MFEDG(m)-CC	1	0,1045	0,2128	0,1046	0,2178	-1849,5	0,09
	2	<i>0,0923</i>	<i>0,1986</i>	<i>0,0958</i>	0,2103	<i>-1740,0</i>	<i>0,57</i>
	3	0,0996	0,2074	0,0992	<i>0,2101</i>	-1766,8	0,34
MDG-CC	-	0,1250	0,2570	-	-	-2198,9	-
MEDG-CC	-	0,1244	0,2545	0,1080	0,2425	-2171,0	-

¹ Com valores ajustados dentro da amostra

² Com valores interpolados nas estações PI e CA.

componente comum capta a tendência observada em todas as estações monitoradoras, ou seja, uma diminuição na ocorrência de chuva no período de abril a setembro indicando uma quebra estrutural nesse período. Nota-se também uma variação irregular com valores baixos no início e no final de cada mês, o que sugere um aparente ciclo intermensal. O primeiro fator exibe pouca variação temporal com coeficiente auto-regressivo estimado em 0,353. O segundo fator apresenta um padrão bastante irregular, principalmente no período de abril a setembro, com aparentes variações intermensais ao longo do ano. Nota-se também que no final de abril e inícios de maio a série apresenta os valores mais baixos indicando uma variação importante neste período. Na Figura 6.10 observou-se que nos meses de abril e maio começa o período menos chuvoso do ano e, portanto, esta característica está relacionada diretamente ao comportamento do segundo fator.

A Figura 6.12 mostra os mapas das cargas dos fatores para todo o estado estimados via *krigagem* Bayesiana. Ambos os gráficos indicam uma suave variação espacial na direção sudeste para $\beta_{(1)}$ e centro-sul para $\beta_{(2)}$. Segundo as características climatológicas e da análise descritiva dos dados as regiões sul, sudeste e central apresentam os maiores índices pluviométricos e mais dias chuvosos ao longo do ano. Em conclusão, podemos afirmar que: (i) a interação do primeiro fator com $\beta_{(1)}$ indica que a ocorrência de chuva na região sudeste

Tabela 6.7: *Chuva em Minas Gerais:* Sumário a posteriori dos parâmetros do modelo MFEDG-CC(2) que caracterizam a componente comum, os fatores e as cargas espacialmente estruturadas. T.A.: Taxa de aceitação. \hat{R} : diagnóstico de Gelman & Rubin.

θ	$E(\theta)$	$\sqrt{Var(\theta)}$	Percentis			T.A.	\hat{R}
			2,5%	50 %	97,5%		
W	0,591	0,120	0,410	0,583	0,802	-	1,01
γ_1	0,353	0,068	0,220	0,354	0,484	-	1,02
γ_2	0,568	0,058	0,458	0,569	0,679	-	1,00
λ_1	0,751	0,110	0,562	0,741	0,996	-	1,01
λ_2	0,742	0,084	0,597	0,734	0,919	-	1,01
μ_1	-0,370	0,490	-1,230	-0,406	0,742	-	1,07
μ_2	0,251	0,690	-1,160	0,247	1,590	-	1,05
τ_1^2	0,908	0,560	0,327	0,748	2,490	-	1,01
τ_2^2	2,190	1,300	0,840	1,840	5,800	-	1,03
ϕ_1	250,720	135,460	72,147	221,050	603,370	0,574	1,08
ϕ_2	379,960	174,410	156,600	351,260	840,380	0,529	1,05

é pouco variável ao longo do tempo, fato que pode ser verificado com a Figura 6.10(b) na qual o total de dias chuvosos por mês varia suavemente ao longo do ano; e por último, (ii) a interação do segundo fator com $\beta_{(2)}$ indica que a ocorrência de chuva é mais variável e com maior probabilidade nas regiões centro e sul do estado.

A Figura 6.13 mostra as probabilidades ajustadas da ocorrência de chuva para 4 estações monitoradoras, AC, AN, BH e MA, respectivamente. Ao longo do tempo, pode-se verificar que as probabilidades de ocorrência de chuva são próximas de 1 naqueles dias em que o evento realmente aconteceu o que indica um bom ajuste do modelo no período amostral. A Figura 6.14 mostra os resultados da interpolação espacial. Cada gráfico indica as probabilidades ajustadas nas estações PI e CA retiradas da análise. Nota-se que os valores interpolados parecem acompanhar a tendência dos dados com valores próximos de 1 nos dias de ocorrência de chuva. Finalmente, a Figura 6.15 mostra os mapas de probabilidade para três dias consecutivos dos meses de janeiro, março, setembro e dezembro. Neste gráfico pode-se observar uma suave variação tanto espacial como temporal.

6.3 Considerações finais

Neste capítulo várias especificações do MFEDG foram utilizadas para modelar as ocorrências de chuva em duas regiões com características climatológicas totalmente distintas. Nas duas

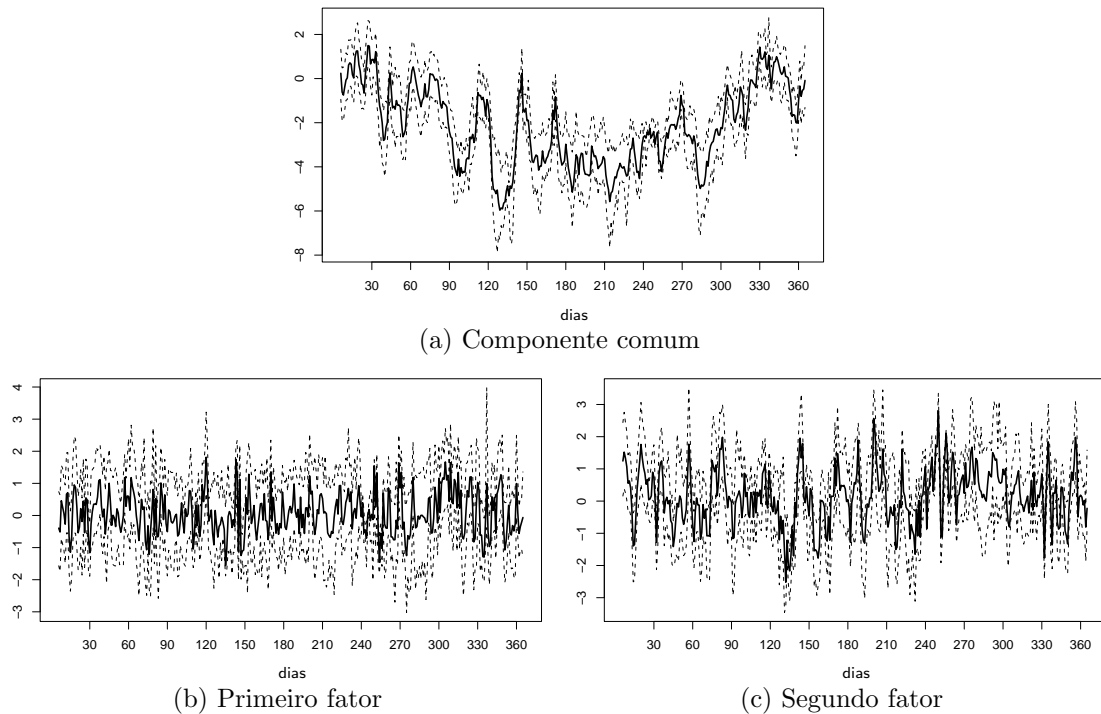


Figura 6.11: *Chuva em Minas Gerais:* Componente comum e fatores estimados para a especificação MFEDG(2)-CC. A linha cheia representa a média a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

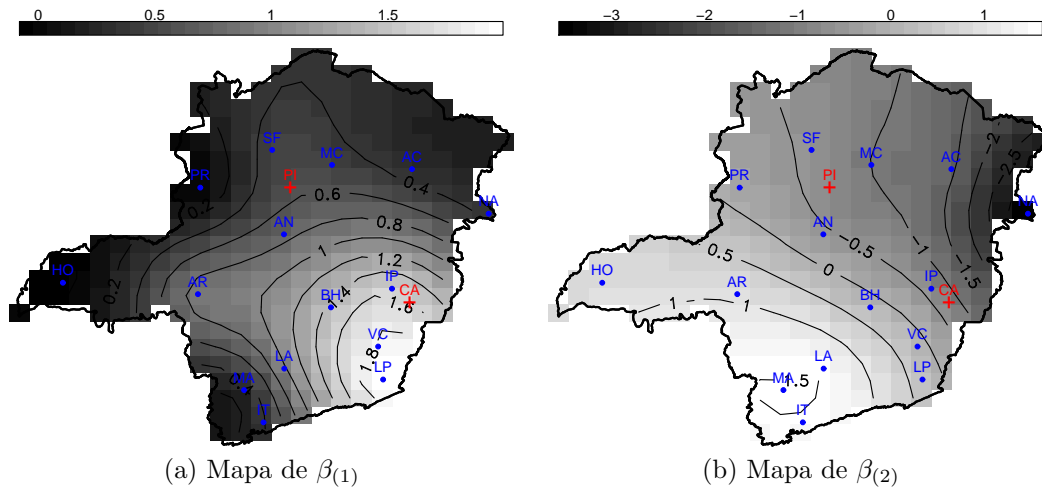


Figura 6.12: *Chuva em Minas Gerais:* Interpolação Bayesiana nas cargas dos fatores. Os valores acima de cada gráfico indicam o intervalo de variação das médias a posteriori.

aplicações apresentadas, o objetivo é estimar a probabilidade de ocorrência de chuva para entender e explicar através do modelos a dinâmica espacial e temporal da probabilidade.

Na primeira aplicação o modelo foi capaz de identificar dois tipos de padrões na dinâmica

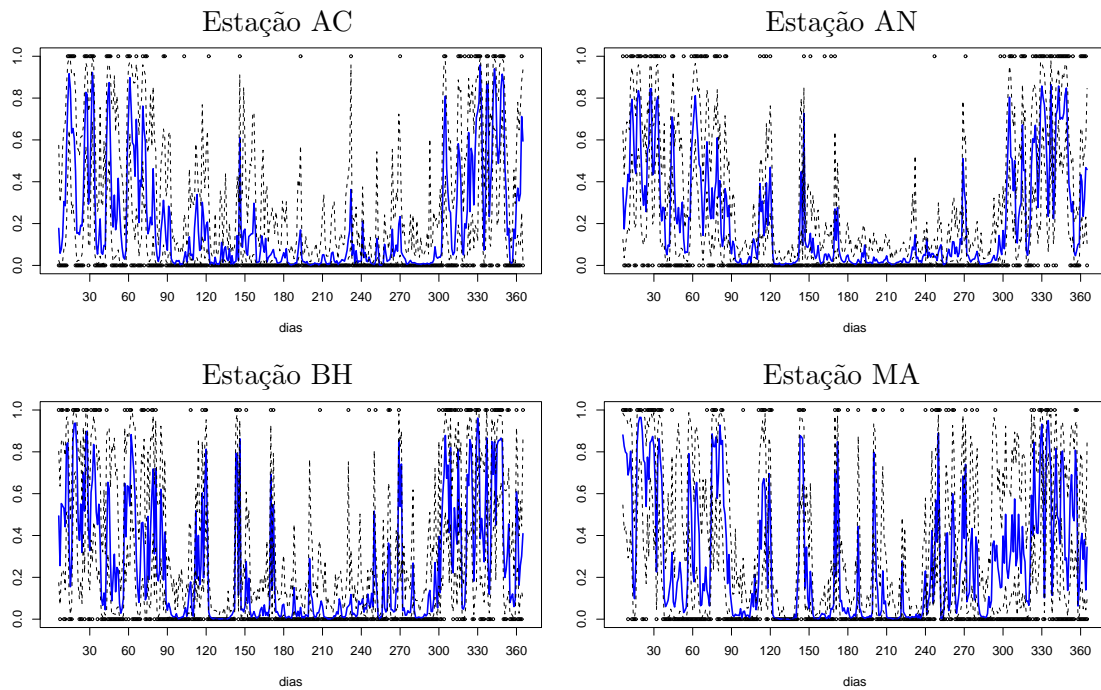


Figura 6.13: *Chuva em Minas Gerais:* Probabilidades ajustadas de ocorrência de chuva em quatro estações monitoradoras. A linha cheia representa a média a posteriori e as linhas pontilhadas os intervalos de credibilidade de 95%.

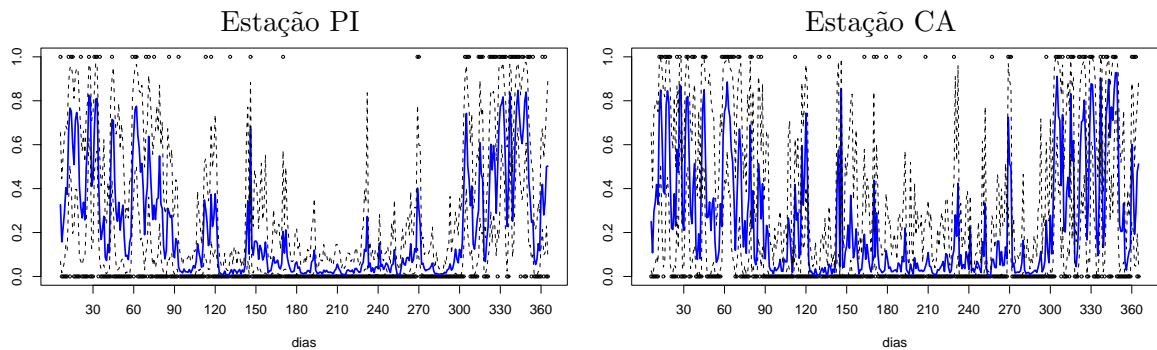


Figura 6.14: *Chuva em Minas Gerais:* Interpolação espacial nas estações PI e CA. A linha azul representa a probabilidade de ocorrência de chuva e as linhas pontilhadas o intervalo de credibilidade de 95%.

dos dados, uma delas na porção nordeste da área de estudo (locais pertencentes às Ilhas Marshall) e a outra na região oeste. Na segunda aplicação o modelo foi capaz de retratar a estrutura espacial do comportamento da chuva, estabelecendo diferenças claras entre as regiões norte, sul e leste do estado, refletindo as características climatológicas que predominam nestas regiões (zonas chuvosas).

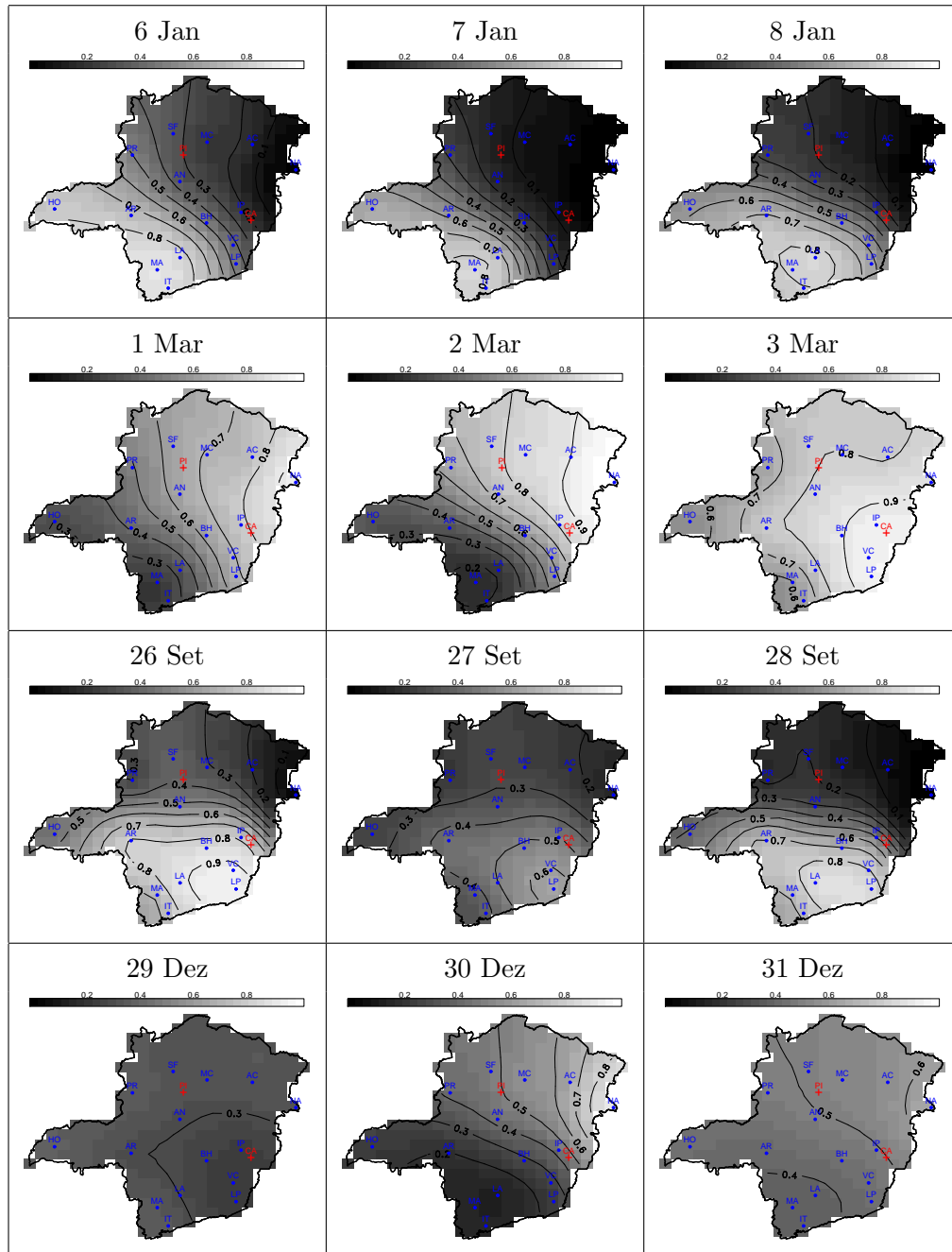


Figura 6.15: *Chuva em Minas Gerais:* Mapa de probabilidade para alguns dias do ano 2005.

Nas duas aplicações, interpolação via *krigagem* Bayesiana foi feita para estimar a dinâmica temporal da probabilidade de chuva em locais onde não se têm informação. Em ambos os casos a interpolação foi feita em locais deixados de fora da análise propositalmente para efeitos de comparação, os resultados mostraram que probabilidades relativamente altas foram

estimadas nos casos de ocorrência do evento chuva e desta maneira interpolação pode ser feita para estimar valores da probabilidade em qualquer ponto da área de análise.

O ajuste de todas as classes propostas foi feito utilizando uma abordagem completamente Bayesiana em que o número de fatores foi estimado utilizando o algoritmo RJMCMC. Devido à alta dimensionalidade dos parâmetros latentes, modelos com mais de 3 fatores não foram bem estimados, com taxas de aceitação para o algoritmo de Metropolis-Hastings nulas ou quase nulas. Este problema foi contornado ao considerar a amostragem em blocos dos fatores. O número máximo de blocos considerado foi 20. A utilização de um número maior de blocos tem a desvantagem de tornar mais lento o procedimento de inferência e de fornecer amostras a posteriori altamente correlacionadas.

Devido a flexibilidade do modelo de incorporar diferentes estruturas tanto na dinâmica temporal como na espacial, variáveis atmosféricas e climatológicas podem ser incorporadas naturalmente na modelagem, seguindo as idéias apresentadas em [Hughes et al. \(1999\)](#).

Capítulo 7

Considerações finais e extensões

7.1 Considerações finais

Nesta tese foi desenvolvida uma nova classe de modelos espaço-temporais para dados contínuos Gaussianos e não Gaussianos. A modelagem proposta se mostra flexível e incorpora diversas estruturas existentes da modelagem espaço-temporal e da análise fatorial. Inicialmente um modelo fatorial espacial dinâmico foi proposto para a modelagem de dados Gaussianos. Uma extensão natural desta abordagem foi apresentada no Capítulo 3, no qual é incorporada uma estrutura dinâmica às cargas dos fatores, modelada através de uma componente puramente temporal e outra espaço-temporal. Em seguida, a modelagem para dados Gaussianos foi estendida permitindo a modelagem de dados pertencentes à família exponencial em que toda a estrutura espaço-temporal descrita para o caso Gaussiano é facilmente adaptada para modelar uma função da média através de uma função de ligação apropriada.

Inferência para esta classe de modelos foi desenvolvida através de um esquema MCMC para obter uma aproximação da posteriori do modelo. Tanto para o caso Gaussiano quanto para o caso não Gaussiano, um tratamento completamente Bayesiano foi proposto para estimar o número de fatores. Os estudos simulados apresentados mostraram que os algoritmos de inferência funcionam com resultados satisfatórios em todos os casos. O algoritmo RJMCMC mostra-se como uma alternativa eficiente para a seleção de modelos espaço-temporais representando uma das contribuições mais importantes da tese. Por outro lado, para o MFEDG uma proposta para amostrar dos fatores latentes é desenvolvida. Este algoritmo se mostra como uma alternativa eficiente na amostragem de fatores (ou estados) latentes multivariados em modelos dinâmicos espaço-temporais generalizados.

Os principais produtos desta tese são três artigos. O primeiro, já submetido para publicação, intitulado “*Spatial Dynamic Factor Analysis*” tem como principal objetivo apresen-

tar a nova classe de modelos para o caso Gaussiano considerando as cargas invariantes no tempo, além de apresentar um tratamento completamente Bayesiano para estimar o número de fatores do modelo. Adicionalmente, uma aplicação para ajustar níveis de SO_2 na região leste dos EUA é apresentada. O segundo artigo, em fase de preparação, refere-se à extensão do MFED ao permitir que as cargas dos fatores tenham evolução temporal. O terceiro artigo, também em fase de preparação, refere-se à extensão do modelo Gaussiano para modelar dados pertencentes à família exponencial. Neste artigo, algumas aplicações para dados Gama e Bernoulli serão consideradas, assim como uma proposta de amostragem em blocos para os fatores latentes dentro de um esquema MCMC.

7.2 Extensões

Extensões do modelo proposto podem ser feitas em diferentes linhas. Uma delas é considerar a classe proposta na modelagem de dados de área. Os dados de área representam outro tipo de observações medidas no espaço, onde as observações são obtidas a partir de um número finito de subregiões ou áreas que compreendem toda a região sob estudo. Alguns exemplos típicos relacionados a este tipo de dado são: o número de casos de uma determinada doença nos bairros de uma cidade, a presença de uma espécie animal numa grade especificada, etc. Para tratar com este tipo de dado, todos os modelos propostos nesta tese podem ser estendidos ao considerar uma estrutura espacial que dependa da estrutura de vizinhança dos dados.

Na Figura 7.1 tem-se uma representação geral das possíveis variações e extensões da classe proposta. Cada vértice da base do paralelepípedo representa um tipo de dado espaço-temporal modelado através da classe geral proposta. Mais especificamente: *i*) DCN: dados contínuos normais, *ii*) DCNN: dados contínuos não normais, *iii*) DAN: dados de área normais e *iv*) DANN: dados de área não normais.

Segundo esta representação, no total, 8 classes de modelos podem ser consideradas. Para os 4 tipos de dados espaço-temporais, uma primeira abordagem é considerar as cargas dos fatores invariantes ao longo do tempo. Uma extensão natural vem depois ao considerar um comportamento dinâmico nas cargas dos fatores. Entre todas estas classes possíveis, somente três foram propostas nesta tese. As duas primeiras para dados contínuos normais com cargas estáticas e dinâmicas, respectivamente, e a terceira para dados contínuos não normais com cargas estáticas (representados em 3 vértices do paralelepípedo). Nota-se que o lado do

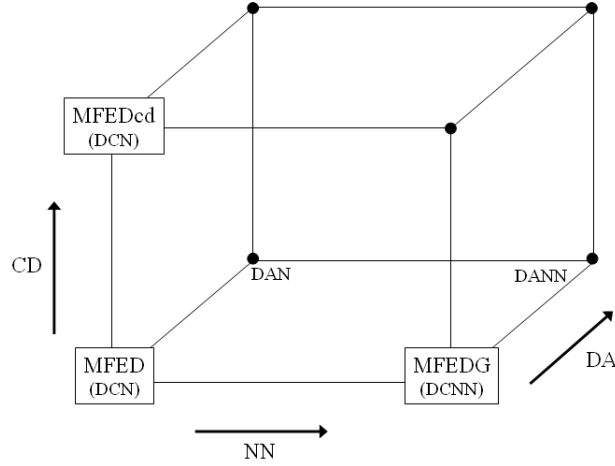


Figura 7.1: Representação geral da classe de modelos proposta para diferentes tipos de dados espaço-temporais.

paralelepípedo que representa a extensão para o modelo não normal (NN) é maior. Isto indica um grau de dificuldade maior ao considerar dados pertencentes à família exponencial, devido, principalmente, às dificuldades inerentes ao procedimento de inferência. Portanto, as extensões deste trabalho estão baseadas nas restantes 5 classes de modelos descritas a seguir.

7.2.1 MFEDG com cargas dinâmicas

Uma extensão natural do modelo proposto na Seção 5 é feita ao considerar as cargas dos fatores variando no tempo. Este modelo, chamado de modelo fatorial espacial dinâmico generalizado com cargas dinâmicas (MFEDGcd), é representado da seguinte forma:

$$p(y_t | \eta_t, \psi) \propto \exp\{\psi[y_t^T \eta_t - b(\eta_t)] + c(y_t, \psi)\} \quad (7.1a)$$

$$\eta_t = v(\theta_t) \quad (7.1b)$$

$$g(\mu_t) = \theta_t = \beta_t f_t + \mu_t^{y*} \quad (7.1c)$$

$$f_t = \Gamma f_{t-1} + w_t, \quad w_t \sim N(0, \Lambda) \quad (7.1d)$$

$$\beta_{(j)t} = \delta_{(j)t} + X_t^\beta \mu_{j,t}, \quad (7.1e)$$

$$\delta_{(j)t} = \delta_{(j)t-1} + \varepsilon_{(j)t}, \quad \varepsilon_{(j)t} \sim N(0, \tau_j^2 R_{\phi_j}), \quad (7.1f)$$

$$\mu_{j,t} = \mu_{j,t-1} + \varepsilon_{jt}^\mu, \quad \varepsilon_{jt}^\mu \sim N(0, \Sigma_{\mu_j}). \quad (7.1g)$$

O modelo é completado com a especificação das seguintes priors: $\delta_{(j)0} \sim N(0, \tau_{j0}^2 R_{\phi_{j0}})$, $\mu_{j,0} \sim N(m_0^\mu, C_0^\mu)$ e $f_0 \sim N(m_0, C_0)$. Analogamente ao modelo descrito em (3.1), as colunas

de β_t são decompostas em duas componentes: uma puramente temporal, denotada por $\mu_{j,t}$, e a outra espaço-temporal, denotada por $\delta_{(j)t}$. Assim, a dependência espacial é modelada através das cargas dinâmicas que seguem processos auto-regressivos de ordem um. Inferência Bayesiana completa pode ser feita utilizando os procedimentos descritos nas Seções 3.3 e 5.3. Mais especificamente, para os fatores comuns, o esquema de amostragem em blocos descrito na Subseção 5.3.3 pode ser utilizado, substituindo β por β_t . Para as cargas dinâmicas, o esquema em blocos também pode ser estendido re-escrevendo as equações (7.1c) e (7.1e) da seguinte forma:

$$\begin{aligned}\theta_t &= f_t^* \delta_t^* + \mu_t^{y*} \\ \delta_t^* &= \delta_{t-1}^* + \varepsilon_t^*, \quad \varepsilon_t^* \sim N(0, \Sigma_{\delta^*})\end{aligned}$$

onde $f_t^* = (f_t' \otimes X_t^\beta \mid f_t' \otimes I_N)$ tem dimensão $N \times m(p + N)$, $\delta_t^* = (\mu_t', \delta_t')'$ tem dimensão $m(p + N) \times 1$, $\mu_t = (\mu_{1t}', \dots, \mu_{mt}')'$, $\delta_t = (\delta_{(1)t}', \dots, \delta_{(m)t}')'$ e $\Sigma_{\delta^*} = \text{diag}(W_\mu, \tau_1^2 R_{\phi_j}, \dots, \tau_m^2 R_{\phi_m})$ com $W_\mu = \text{diag}(\Sigma_{\mu_1}, \dots, \Sigma_{\mu_m})$. Desta forma, o esquema em blocos é feita para o vetor δ_t^* como descrita na Subseção 5.3.3 substituindo β por f_t^* e f_t por δ_t^* .

7.2.2 MFED para dados de área

Seja S a região de estudo (com forma regular ou irregular) dividida em um número finito N de subregiões ou áreas regulares (grade) ou irregulares (municípios, cidades, etc) com limites bem definidos, cada uma delas denotada por s_i , $i = 1, \dots, N$. Seja $y_t = (y_{1t}, \dots, y_{Nt})$ o vetor de observações que contém as medidas em todas as áreas s_i no tempo t . Neste caso, a idéia é utilizar estruturas que especifiquem que a resposta do processo em uma determinada área está influenciado, de alguma forma, pelas observações de áreas vizinhas.

Assim, o MFED descrito em (2.1) pode ser utilizado na modelagem de dados de área ao considerar uma estrutura espacial distinta nas colunas da matriz de cargas dos fatores. Esta nova especificação considera uma estrutura de vizinhança na matriz de correlação espacial de cada $\beta_{(j)}$. O MFED para dados de área é especificado da seguinte forma:

$$y_t = \mu_t^{y*} + \beta f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \quad (7.2a)$$

$$f_t = \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda) \quad (7.2b)$$

$$\beta_{(j)} = \mu_j^\beta + \delta_{(j)}, \quad \delta_{(j)} \sim \text{CAR}(\tau_j^2) \quad (7.2c)$$

$$f_0 \sim N(m_0, C_0) \quad (7.2d)$$

onde a distribuição de $\delta_{(j)}$ segue uma priori auto-regressiva condicional (CAR) intrínseca (Besag, York & Mollié, 1991). Neste caso, a distribuição condicional de $\delta_{l(j)}$ na área l dados os vizinhos é dada por

$$\delta_{l(j)} | \delta_{k(j)}, l \neq k \sim N \left(\frac{\sum_{k \in \vartheta_l} \delta_{k(j)}}{\sum_{k \in \vartheta_l} \omega_{lk}}, \frac{\tau_j^2}{\sum_{k \in \vartheta_l} \omega_{lk}} \right)$$

onde ϑ_l representa o conjunto de áreas subjacentes a l e $\omega_{lk} = 1$ se l e k são vizinhos e 0 em caso contrário. Com esta especificação, a distribuição a priori conjunta de $\delta_{(j)}$ é

$$p(\delta_{(j)} | \tau_j^2) \propto \exp\{-0,5\tau_j^2 \delta'_{(j)}(D - W)\delta_{(j)}\}$$

onde $D = \text{diag}(\omega_{1+}, \dots, \omega_{N+})$, $\omega_{l+} = \sum_{k \in \vartheta_l} \omega_{lk}$ e o elemento (l, k) da matriz W é ω_{lk} .

Outra possível especificação para $\delta_{(j)}$ é considerar uma priori CAR própria tal que:

$$\delta_{(j)} \sim N(0, \tau_j^2 P_j)$$

onde $P_j^{-1} = I_N + \phi_j(D - W)$ e $\phi_j \geq 0$. Neste caso, o parâmetro ϕ_j controla o grau de associação entre as componentes de $\delta_{(j)}$ e P_j^{-1} é dominada pela diagonal e, por causa disso, é definida positiva (Harville, 1997). Assim, a distribuição condicional de $\delta_{l(j)}$ na área l dados os vizinhos é dada por

$$\delta_{l(j)} | \delta_{k(j)}, l \neq k \sim N \left(\frac{\phi_j \sum_{k \in \vartheta_l} \delta_{k(j)}}{1 + \phi_j \sum_{k \in \vartheta_l} \omega_{lk}}, \frac{\tau_j^2}{1 + \phi_j \sum_{k \in \vartheta_l} \omega_{lk}} \right).$$

Inferência para este tipo de modelo é feita utilizando os procedimentos descritos na Seção 2.4, exceto para os parâmetros de $\delta_{(j)}$. Se a priori CAR intrínseca for considerada, uma priori gama inversa para o parâmetro τ_j^2 pode ser utilizada, porem outras distribuições a priori podem ser especificadas, lembrando que a inferência pode ser sensível a escolha desta priori. Para o caso da priori CAR própria, a priori de referência conjunta para o par (τ_j^2, ϕ_j) proposta por Ferreira & De Oliveira (2007) é sugerida.

7.2.3 MFED com cargas dinâmicas para dados de área

Analogamente ao modelo apresentado no Capítulo 3, esta nova classe pode ser representada da seguinte forma:

$$y_t = \mu_t^{y*} + \beta_t f_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma), \quad (7.3a)$$

$$f_t = \Gamma f_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Lambda), \quad (7.3b)$$

$$\beta_{(j)t} = \delta_{(j)t} + X_{j,t}^\beta \mu_{j,t}, \quad (7.3c)$$

$$\delta_{(j)t} = \delta_{(j)t-1} + \eta_{(j)t}, \quad \eta_{(j)t} \sim \text{CAR}(\tau_j^2) \quad (7.3d)$$

$$\mu_{j,t} = \mu_{j,t-1} + \eta_{j,t}^\mu, \quad \eta_{j,t}^\mu \sim N(0, \Sigma_{\mu_j}). \quad (7.3e)$$

O modelo é completado com a especificação das seguintes priors: $\delta_{(j)0} \sim \text{CAR}(\tau_{j0}^2)$, $\mu_{j,0} \sim N(m_0^\mu, C_0^\mu)$ e $f_0 \sim N(m_0, C_0)$. Inferência a posteriori para esta classe é feita utilizando o algoritmo MCMC proposto na Seção 3.3. O hiperparâmetro τ_{j0}^2 é conhecido e pode ser fixado na média a posteriori obtida do modelo com cargas estáticas.

7.2.4 MFEDG para dados de área

Da mesma forma que no Capítulo 5, esta nova classe é especificada como:

$$p(y_t | \eta_t, \psi) \propto \exp\{\psi [y_t^T \eta_t - b(\eta_t)] + c(y_t, \psi)\} \quad (7.4a)$$

$$\eta_t = v(\theta_t) \quad (7.4b)$$

$$g(\mu_t) = \theta_t = \beta f_t + \mu_t^{y*} \quad (7.4c)$$

$$f_t = \Gamma f_{t-1} + w_t, \quad w_t \sim N(0, \Lambda) \quad (7.4d)$$

$$\beta_{(j)} = \mu_j^\beta + \delta_{(j)}, \quad \delta_{(j)} \sim \text{CAR}(\tau_j^2) \quad (7.4e)$$

$$f_0 \sim N(m_0, C_0).$$

Este modelo preserva as mesmas características e propriedades do modelo em (5.2). Inferência a posteriori é feita sob a abordagem Bayesiana da mesma forma que na Seção 5.3.

7.2.5 MFEDG com cargas dinâmicas para dados de área

A última classe especificada, é uma variante do modelo especificado em (7.1) ao considerar uma priori CAR intrínseca na evolução da componente espaço-temporal das cargas dinâmicas.

Este modelo é representado da seguinte forma:

$$p(y_t|\eta_t, \psi) \propto \exp\{\psi[y_t^T \eta_t - b(\eta_t)] + c(y_t, \psi)\} \quad (7.5a)$$

$$\eta_t = v(\theta_t) \quad (7.5b)$$

$$g(\mu_t) = \theta_t = \beta_t f_t + \mu_t^{y^*} \quad (7.5c)$$

$$f_t = \Gamma f_{t-1} + w_t, \quad w_t \sim N(0, \Lambda) \quad (7.5d)$$

$$\beta_{(j)t} = \delta_{(j)t} + X_t^\beta \mu_{j,t}, \quad (7.5e)$$

$$\delta_{(j)t} = \delta_{(j)t-1} + \varepsilon_{(j)t}, \quad \varepsilon_{(j)t} \sim \text{CAR}(\tau_j^2), \quad (7.5f)$$

$$\mu_{j,t} = \mu_{j,t-1} + \varepsilon_{jt}^\mu, \quad \varepsilon_{jt}^\mu \sim N(0, \Sigma_{\mu_j}), \quad (7.5g)$$

e completado com as seguintes distribuições a priori: $\delta_{(j)0} \sim \text{CAR}(\tau_{j0}^2)$, $\mu_{j,0} \sim N(m_0^\mu, C_0^\mu)$ e $f_0 \sim N(m_0, C_0)$. O procedimento de inferência, descrito em linhas gerais na Subseção 7.2.1, pode ser utilizado para este modelo.

Apêndice A

Distribuições condicionais completas

A.1 Modelo com cargas estáticas

As distribuições condicionais completas para todos os parâmetros no modelo (2.5) estão listadas a seguir. Aqui são consideradas as seguintes denominações: Variâncias idiossincráticas: σ , dinâmicas dos fatores comum: γ , variâncias do fator comum: λ , média das cargas: μ , hiperparâmetros espaciais: τ_j^2 e ϕ_j , matriz de cargas do fator: β , e fatores comuns: f_t , para $t = 1, \dots, T$. Neste apêndice, $[\theta]$ denota a densidade condicional completa de θ . Além disso, se A e B são matrizes de dimensão $m \times n$ e $s \times t$, o produto Kronecker $A \otimes B$ é uma matriz de dimensão de dimensão $ns \times nt$

Variâncias idiossincráticas: Da verossimilhança apresentada na Subseção 2.2.3 pode ser mostrado que $y_i | F, \sigma_i^2, \beta_i \sim N(F\beta_i, \sigma_i^2 I)$, $i = 1, \dots, N$, onde y_i é a i -ésima coluna de y , β_i é a i -ésima coluna de β . Portanto, $[\sigma_i^2] \sim IG((T + n_\sigma)/2, ((y_i - F\beta_i)'(y_i - F\beta_i) + n_\sigma s_\sigma)/2)$.

Variâncias dos fator comum: $[\lambda_j] \sim IG((T - 1 + n_\lambda)/2, (\sum_{t=2}^T (f_{jt} - \gamma_j f_{j,t-1})^2 + n_\lambda s_\lambda)/2)$.

Médias das cargas: $[\mu_j] \sim N(m_{\mu_j}^*, S_{\mu_j}^*)$, $m_{\mu_j}^* = S_{\mu_j}^* \left[\tau_j^{-2} \beta'_{(j)} R_{\phi_j}^{-1} \mathbf{1}_N + m_\mu S_\mu^{-1} \right]$ and $S_{\mu_j}^{*-1} = \tau_j^{-2} \mathbf{1}'_N R_{\phi_j}^{-1} \mathbf{1}_N + S_\mu^{-1}$.

Cargas do fator: A matriz de cargas do fator é amostrada conjuntamente. Assim, a equação (2.1a) é reescrita como $y_t = f_t^* \beta^* + \epsilon_t$, onde $f_t^* = f_t' \otimes I_N$ e $\beta^* = (\beta'_{(1)}, \dots, \beta'_{(m)})'$ são matrizes de dimensão $N \times Nm$ e $Nm \times 1$, onde $A \otimes B$ denota o produto Kronecker de matrizes A e B . Similarmente, a distribuição a priori de β^* é $\beta^* \sim N(\mu_{\beta^*}, \Sigma_{\beta^*})$, com $\mu_{\beta^*} = \mu \otimes \mathbf{1}_N$, $\Sigma_{\beta^*} = \Sigma_\beta \otimes R_\phi$ e $\Sigma_\beta = \text{diag}(\tau_1^2, \dots, \tau_m^2)$. Da regressão normal multivariada (Box & Tiao, 1973), pode ser mostrado que $(\beta^* | \dots) \sim N(\tilde{\mu}_{\beta^*}, \tilde{\Sigma}_{\beta^*})$, com $\tilde{\Sigma}_{\beta^*}^{-1} = \sum_{t=1}^T f_t^{*'} \Sigma^{-1} f_t^* + \Sigma_{\beta^*}^{-1}$ e

$$\tilde{\mu}_{\beta^*} = \tilde{\Sigma}_{\beta^*} \left(\sum_{t=1}^T f_t^{*\prime} \Sigma^{-1} y_t + \Sigma_{\beta^*}^{-1} \mu_{\beta^*} \right).$$

Dinâmicas do fator comum: Da equação (2.1b) segue que $f_{jt} \sim N(\gamma_j f_{j,t-1}, \lambda_j)$, $j = 1, \dots, m$ e $t = 2, \dots, T$. Portanto, $p(\gamma_i | \dots) \propto \prod_{t=2}^T p(f_{jt} | f_{j,t-1}, \gamma_i, \lambda_i) p(\gamma_i | m_\gamma, S_\gamma, \alpha)$, logo, i) se $\alpha = 1$, $(\gamma_j | \dots) \sim Ntr_{(-1,1)}(m_{\gamma_j}^*, S_{\gamma_j}^*)$ com $S_{\gamma_j}^{*-1} = \lambda_j^{-1} \sum_{t=2}^T f_{j,t-1}^2 + S_\gamma^{-1}$ e $m_{\gamma_j}^* = S_{\gamma_j}^* \left[\lambda_j^{-1} \sum_{t=2}^T f_{jt} f_{j,t-1} + m_\gamma S_\gamma^{-1} \right]$, e ii) se $\alpha \in (0, 1)$ amostrar γ_j com probabilidade α^* utilizando a distribuição normal $Ntr_{(-1,1)}(m_{\gamma_j}^*, S_{\gamma_j}^*)$ ou se $\gamma_i = 1$ com probabilidade $1 - \alpha^*$ em que $\alpha^* = A/(A+B)$, $A = \alpha C S_\gamma^{-1/2} S_{\gamma_j}^{*1/2} \exp\{-0.5[\lambda^{-1} \sum_{t=2}^T f_{jt}^2 + S_\gamma^{-1} m_\gamma - S_{\gamma_j}^{*-1} m_{\gamma_j}^{*2}]\}$, $C = [\Phi((1 - m_{\gamma_j}^*)/S_{\gamma_j}^{*1/2}) - \Phi((-1 - m_{\gamma_j}^*)/S_{\gamma_j}^{*1/2})][\Phi((1 - m_\gamma)/S_\gamma^{1/2}) - \Phi((-1 - m_\gamma)/S_\gamma^{1/2})]^{-1}$, $B = (1 - \alpha) \exp\{-0.5\lambda_j^{-1} \sum_{t=2}^T (f_{jt} - f_{j,t-1})^2\}$ e Φ é a probabilidade acumulada da normal padrão.

Fatores comuns: Os fatores comuns f_1, \dots, f_T , são amostrados conjuntamente utilizando o algoritmo FFBS (do inglês *forward filtering backward sampling*) proposto por Carter & Kohn (1994) e Frühwirth-Schnatter (1994), condicional sobre β e Θ a distribuição condicional completa conjunta segue a decomposição retrospectiva $p(F|y) = \prod_{t=0}^{T-1} p(f_t | f_{t+1}, D_t) p(f_T | D_T)$, onde $D_t = \{y_1, \dots, y_t\}$, $t = 1, \dots, T$ e D_0 representam a informação inicial. Iniciando com $f_0 \sim N(m_0, C_0)$, pode ser mostrado que $f_t | D_t \sim N(m_t, C_t)$, com $m_t = a_t + A_t(y_t - \tilde{y}_t)$, $C_t = R_t - A_t Q_t A_t'$, $a_t = \Gamma m_{t-1}$, $R_t = \Gamma C_{t-1} \Gamma' + \Lambda$, $\tilde{y}_t = \beta a_t$, $Q_t = \beta R_t \beta' + \Sigma$ e $A_t = R_t \beta' Q_t^{-1}$, para $t = 1, \dots, T$. f_T é amostrado de $p(f_T | D_T)$ (passo denominado *forward filtering*). Para $t = T - 1, \dots, 2, 1, 0$, f_t é amostrado de $p(f_t | f_{t+1}, D_t) = N(\tilde{a}_t, \tilde{C}_t)$, com $\tilde{a}_t = m_t + B_t(f_{t+1} - a_{t+1})$, $\tilde{C}_t = C_t - B_t R_{t+1} B_t'$ e $B_t = C_t \Gamma' R_{t+1}^{-1}$ (passo denominado *backward sampling*).

Hiperparâmetros espaciais: Combinando a densidade a priori Gama Inversa (2.3) ou a densidade da priori de referência (2.4) com a função de verossimilhança tem-se que $(\tau_j^2 | \dots) \sim IG(n_{\tau_j}^*/2, n_{\tau_j}^* s_{\tau_j}^*/2)$, com $n_{\tau_j}^* = N + n_\tau$ e $n_{\tau_j}^* s_{\tau_j}^* = (\beta_{(j)} - \mu_j 1_N)' R_{\phi_j}^{-1} (\beta_{(j)} - \mu_j 1_N) + n_\tau s_\tau$ quando a distribuição a priori Gama Inversa é utilizada, e $n_{\tau_j}^* = N$ e $n_{\tau_j}^* s_{\tau_j}^* = (\beta_{(j)} - \mu_j 1_N)' R_{\phi_j}^{-1} (\beta_{(j)} - \mu_j 1_N)$ quando a distribuição da priori de referência é utilizada. A condicional completa de ϕ_j não tem forma conhecida e, portanto, passos de Metropolis-Hastings são implementados. Um valor candidato $\tilde{\phi}_j$ é gerado da distribuição log-normal com parâmetro de locação $\log \phi_j$ e de escala Δ_ϕ , ou seja, $q_j(\phi_j, \cdot) = f_{LN}(\cdot; \log \phi_j, \Delta_\phi)$. Δ_ϕ é a constante sintonizadora frequentemente utilizada para calibrar a densidade proposta. O valor candidato é aceito com

probabilidade

$$\alpha(\phi_j, \tilde{\phi}_j) = \min \left\{ 1, \frac{f_N(\beta_{(j)}; \mu_j \mathbf{1}_N, \tau_j^2 R_{\tilde{\phi}_j}) \pi_P(\tilde{\phi}_j) \tilde{\phi}_j}{f_N(\beta_{(j)}; \mu_j \mathbf{1}_N, \tau_j^2 R_{\phi_j}) \pi_P(\phi_j) \phi_j} \right\},$$

onde π_P pode ser a priori gama inversa, ou seja, π_{IG} ou a priori de referência, ou seja, π_R .

A.2 Modelo com cargas dinâmicas

As condicionais completas para os parâmetros σ (variâncias idiossincráticas do processo), σ_μ (variâncias idiossincráticas de componente temporal) e, τ e ϕ (hiperparâmetros espaciais) são descritas a seguir.

Variâncias idiossincráticas do processo: Considere os dados da forma y_{it} onde $i = 1, \dots, N$ e $t = 1, \dots, T$. Assim, podemos reescrever a equação (3.1a) como: $y_{it} = \sum_{k=1}^m \beta_{i(k)t} f_{kt} + \epsilon_{it}$, com $\epsilon_{it} \sim N(0, \sigma_i^2)$. Portanto, $[\sigma_i^2] \sim GI(n_{\sigma_i}^*/2, n_{\sigma_i}^* s_{\sigma_i}^*/2)$ com $n_{\sigma_i}^* = T + n_\sigma$ e $n_{\sigma_i}^* s_{\sigma_i}^* = \sum_{t=1}^T (y_{it} - \sum_{k=1}^m \beta_{i(k)t} f_{kt}) + n_\sigma s_\sigma$.

Variâncias idiossincráticas da componente temporal: Da equação (3.1e) temos que $(\mu_{jt} | \mu_{j,t-1}, \sigma_{\mu_j}^2) \sim N(\mu_{j,t-1}, \sigma_{\mu_j}^2)$, $j = 1, \dots, m$, $t = 2, \dots, T$. Portanto, $[\sigma_{\mu_j}^2] \sim GI(n_{\mu_j}^*/2, n_{\mu_j}^* s_{\mu_j}^*/2)$ tal que $n_{\mu_j}^* = T - 1 + n_\mu$ e $n_{\mu_j}^* s_{\mu_j}^* = \sum_{t=2}^T (\mu_{jt} - \mu_{j,t-1})^2 + n_\mu s_\mu$.

Hiperparâmetros espaciais: Combinando a densidade a priori Gama Inversa (2.3) ou a densidade da priori de referência (2.4) com a função de verossimilhança tem-se que $[\tau_j^2] \sim IG(n_{\tau_j}^*/2, n_{\tau_j}^* s_{\tau_j}^*/2)$, com $n_{\tau_j}^* = N(T - 1) + n_\tau$ e $n_{\tau_j}^* s_{\tau_j}^* = \sum_{t=2}^T (\delta_{(j),t} - \delta_{(j),t-1})' R_{\phi_j}^{-1} (\delta_{(j),t} - \delta_{(j),t-1}) + n_\tau s_\tau$ quando a distribuição a priori Gama Inversa é utilizada, e $n_{\tau_j}^* = N(T - 1)$ e $n_{\tau_j}^* s_{\tau_j}^* = \sum_{t=2}^T (\delta_{(j),t} - \delta_{(j),t-1})' R_{\phi_j}^{-1} (\delta_{(j),t} - \delta_{(j),t-1})'$ quando a distribuição da priori de referência é utilizada. A condicional completa de ϕ_j não tem forma conhecida e, portanto, passos de Metropolis-Hastings são implementados. Um valor candidato $\tilde{\phi}_j$ é gerado da distribuição log-normal com parâmetro de locação $\log \phi_j$ e de escala Δ_ϕ , ou seja, $q_j(\phi_j, \cdot) = f_{LN}(\cdot; \log \phi_j, \Delta_\phi)$. Δ_ϕ é a constante sintonizadora freqüentemente utilizada para calibrar a densidade proposta. O valor candidato é aceito com probabilidade

$$\alpha(\phi_j, \tilde{\phi}_j) = \min \left\{ 1, \frac{\prod_{t=2}^T f_N(\delta_{(j)t}; \delta_{(j),t-1}, \tau_j^2 R_{\tilde{\phi}_j}) \pi_P(\tilde{\phi}_j) \tilde{\phi}_j}{\prod_{t=2}^T f_N(\delta_{(j)t}; \delta_{(j),t-1}, \tau_j^2 R_{\phi_j}) \pi_P(\phi_j) \phi_j} \right\},$$

onde π_P pode ser a priori gama inversa, ou seja, π_{IG} ou a priori de referência, ou seja, π_R .

Referências Bibliográficas

- Aguilar, O. (1998). *Latent structure in Bayesian multivariate time series models*. Tese de Doutorado, Duke University, Durham, NC 27708.
- Aguilar, O. & West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, 18, 338–357.
- Akaike, H. (1974). New look at the statistical model identification. *IEEE Transactions in Automatic Control*, AC-19, 716–723.
- Anderson, B. D. & Moore, J. B. (1979). *Optimal Filtering*. New Jersey: Prentice Hall.
- Arminger, G. & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63, 271–300.
- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press/Chapman & Hall.
- Berger, J. O., Oliveira, V., & Sansó, B. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96, 1361–1374.
- Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1–59.
- Billingsley, P. (1986). *Probability and Measure* (2nd. ed.). New York, USA: John Wiley & Sons.
- Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley.

- Brown, P. J., Vannucci, M., & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, 60, 627–41.
- Calder, C. A. (2007). Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. *Environmental and Ecological Statistics*, 14, 229–247.
- Carter, C. K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81, 541–553.
- Christensen, W. & Amemiya, Y. (2002). Latent variable analysis of multivariate spatial data. *Journal of the American Statistical Association*, 97(457), 302–317.
- Christensen, W. & Amemiya, Y. (2003). Modeling and prediction for multivariate spatial factor analysis. *Journal of Statistical Planning and Inference*, 115, 543–564.
- Clyde, M. (1999). Bayesian model averaging and model search strategies. In Bernardo, J. M., Berger, J. O., Dawid, A. P., & Smith, A. F. M. (Eds.), *Bayesian Statistics 6*, (pp. 157–85). London: Oxford University Press.
- Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.
- Cressie, N. & Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448), 1330–1340.
- Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68, 265–274.
- Dellaportas, P., Forster, J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12, 27–36.
- Doornik, J. A. (2002). *Object-Oriented Matrix Programming Using Ox* (3rd ed.). London: Timberlake Consultants Press and Oxford. www.doornik.com.
- Fahrmeir, L. (1992). Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87, 501–509.
- Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.

- Fernandes, M. V., Schmidt, A. M., & Migon, H. S. (2007). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*. to appear.
- Ferreira, M. A. R. & De Oliveira, V. (2007). Bayesian analysis for a class of gaussian markov random fields. *Journal of Multivariate Analysis*, 98, 789–812.
- Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2), 183–202.
- Gabriel, K. R. & Neumann, J. (1962). A markov chain model for daily rainfall occurrences at Tel Aviv. *Quarterly Journal Royal Meteorological Society*, 88, 90–95.
- Gamerman, D. (1998). Markov chain monte carlo for dynamic generalised linear models. *Biometrika*, 85, 215–227.
- Gamerman, D. & Lopes, H. F. (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference* (2nd ed.). Chapman&Hall/CRC.
- Gamerman, D., Moreira, A. R. B., & Rue, H. (2003). Space-varying regression models: specifications and simulation. *Computational Statistics and Data Analysis*, 42, 513–33.
- Gelfand, A. E., Barnerjee, S., & Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, 16, 465–479.
- Gelfand, A. E., Kottas, A., & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100, 1021–35.
- Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.
- Geweke, J. & Tanizaki, H. (2001). Bayesian estimation of state space models using metropolis-hastings algorithm within gibbs sampling. *Computacional Statistics & Data Analysis*, 37, 151–170.
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458), 590–600.
- Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B*, 69(2), 243–268.

- Godsill, S. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, *10*, 1–19.
- Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, *55*, 245–259.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.
- Gschlößl, S. & Czado, C. (2005). Spatial modelling of claim frequency and claim size in insurance. Discussion Paper 461, Ludwig-Maximilians-Universität München, Munich.
- Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag.
- Hogan, J. W. & Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, *99*, 314–324.
- Hooten, M. & Wikle, C. (2007). A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove. *Environmental and Ecological Statistics*. to appear.
- Hughes, J. P., Guttorp, P., & Charles, S. P. (1999). A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(1), 15–30.
- Jin, X., Carlin, B., & Banerjee, S. (2005). Generalized hierarchical multivariate car models for areal data. *Biometrics*, *61*, 950–961.
- Kitagawa, G. (1987). Non-gaussian state-space modeling of non-stationary time series. *Journal of the American Statistical Association*, *82*, 1032–1041.
- Knorr-Held, L. (1999). Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, *26*(1), 129–144.
- Knorr-Held, L. & Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A*, *164*, 73–85.

- Lopes, H. F. & Carvalho, C. M. (2007). Factor stochastic volatility with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference*, 137, 3082–3091.
- Lopes, H. F. & Migon, H. S. (2002). Comovements and contagion in emergent markets: Stock indexes volatilities. In Gatsonis, C., Kass, R., Carriquiry, A., Gelman, A., Verdinelli, Pauler, D., & Higdon, D. (Eds.), *Case Studies in Bayesian Statistics, Volume 6*, (pp. 287–302).
- Lopes, H. F. & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14, 41–67.
- Mardia, K. V., Goodall, C. R., Redfern, E., & Alonso, F. J. (1998). The kriged Kalman filter (with discussion). *Test*, 7, 217–85.
- Migon, H. S., Gamerman, D., Lopes, H. F., & Ferreira, M. A. R. (2005). Dynamic models. In Dey, D. K. & Rao, C. (Eds.), *Handbook of Statistics: Bayesian Thinking, Modeling and Computation*, volume 25, (pp. 553–588)., North Holland.
- Nobre, A. A., Schmidt, A. M., & Lopes, H. F. (2005). Spatio-temporal models for mapping the incidence of malaria in Pará. *Environmetrics*, 16, 291–304.
- Peña, D. & Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics*, 119, 291–321.
- Peña, D. & Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, 136, 1237–1257.
- Pitt, M. & Shephard, N. (1999). Time varying covariances: A factor stochastic volatility approach (with discussion). In Bernardo, J., Berger, J., Dawid, A., & Smith, A. (Eds.), *Bayesian statistics 6*, (pp. 547–570). London: Oxford University Press.
- Polasek, W. (1997). Factor analysis and outliers: A Bayesian approach. Discussion Paper, University of Basel.
- Press, S. J. & Shigemasu, K. (1997). Bayesian inference in factor analysis (Revised). Technical Report 243, Department of Statistics, University of California at Riverside.

- Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–91.
- Ravines, R. (2007). *Um esquema eficiente de amostragem em modelos dinâmicos generalizados com aplicações em funções de transferência*. Tese de Doutorado, Instituto de Matemática, DME-UFRJ, Rio de Janeiro, Brasil.
- Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.). New York: Wiley.
- Robert, C. R. & Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer.
- Sansó, B., Schmidt, A. M., & Nobre, A. A. (2008). Bayesian spatio-temporal models based on discrete convolutions. *Canadian Journal of Statistics*. To appear.
- Schmidt, A. M. & Gelfand, A. (2003). A Bayesian coregionalization model for multivariate pollutant data. *Journal of Geophysics Research*, *108* (D24). <http://www.agu.org/pubs/crossref/2003/2002JD002905.shtml>.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
- Stein, M. L. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.
- Stern, R. D. & Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society, Series A*, *147*, 1–34.
- Stroud, R., Müller, P., & Sansó, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B*, *63*(4), 673–689.
- Vivar, J. C. (2007). *Modelos espaço-temporais para dados de área na família exponencial*. Tese de Doutorado, Instituto de Matemática, DME-UFRJ, Rio de Janeiro, Brasil.
- Wang, F. & Wall, M. M. (2001). Modelling multivariate data with a common spatial factor. Technical Report 2001-008, Division of Biostatistics, University of Minnesota.
- Wang, F. & Wall, M. M. (2003). Generalized common spatial factor model. *Biostatistics*, *4*, 569–582.
- West, M. & Harrison, P. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

- West, M., Harrison, P. J., & Migon, H. S. (1985). Dynamic generalized linear models and bayesian forecasting (with discussion). *Journal of the American Statistical Association*, 80, 73–96.
- Wikle, C. K. & Cressie, N. (1999). A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86, 815–829.

Livros Grátis

(<http://www.livrosgratis.com.br>)

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)
[Baixar livros de Literatura de Cordel](#)
[Baixar livros de Literatura Infantil](#)
[Baixar livros de Matemática](#)
[Baixar livros de Medicina](#)
[Baixar livros de Medicina Veterinária](#)
[Baixar livros de Meio Ambiente](#)
[Baixar livros de Meteorologia](#)
[Baixar Monografias e TCC](#)
[Baixar livros Multidisciplinar](#)
[Baixar livros de Música](#)
[Baixar livros de Psicologia](#)
[Baixar livros de Química](#)
[Baixar livros de Saúde Coletiva](#)
[Baixar livros de Serviço Social](#)
[Baixar livros de Sociologia](#)
[Baixar livros de Teologia](#)
[Baixar livros de Trabalho](#)
[Baixar livros de Turismo](#)