

Trabalho de Conclusão de Curso

Análise Hierárquica para Dados de Bacias Hidrográficas

Verônica Stamatto Peres

Maio de 2022

Verônica Stamatto Peres

Análise Hierárquica para Dados de Bacias Hidrográficas

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dr. Sílvia Regina Costa Lopes

Porto Alegre
Maio de 2022

Agradecimentos

Agradeço à professora Sílvia Regina Costa Lopes pelo incentivo, apoio e orientação na realização deste trabalho.

Agradeço aos professores da banca Márcio Valk e Márcia Barbian.

Agradeço ao Instituto de Matemática e Estatística (IME), ao Laboratório Computacional em Pesquisa Matemática (LCPM) e a UFRGS.

Agradeço à minha família e amigos.

Resumo

Este trabalho tem como objetivo desenvolver e analisar modelos lineares generalizados para séries temporais, aplicando a metodologia Hierárquica Bayesiana. Utilizamos dois bancos de dados para comparar os modelos, os dados do Rio dos Sinos, que descrevem o índice da qualidade da água e os dados do Rio Colorado, que medem a vazão da água. Iniciamos as análises considerando apenas o efeito aleatório e no final acrescentamos o efeito correlacionado espacialmente para realizar a análise espaço-temporal. Para a comparação da eficiência dos modelos utilizamos diversos critérios de avaliação e gráficos das séries temporais com os estimadores. Com isso, foi possível observar quais modelos descrevem melhor os dados de acordo com suas características. Esse conhecimento pode ser utilizado para facilitar e melhorar a eficiência da modelagem de dados em outras aplicações futuras.

Palavras-Chave: Séries temporais, Inferência Bayesiana, Modelos lineares generalizados, Análise espaço-temporal, Análise Hierárquica.

Abstract

This work aims to develop and analyze generalized linear models for time series, applying the Bayesian Hierarchical methodology. We used two databases to compare the models, data from Rio dos Sinos, that describe the water quality index and data from Rio Colorado, that measure water flow. We started the analysis considering only the random effect and at the end we added the spatially correlated effect to perform the spatio-temporal analysis. To compare the efficiency of the models, we used several evaluation criteria and time series graphs with estimators. With this, it was possible to observe which models best describe the data according to their characteristics. This knowledge can be used to facilitate and improve the efficiency of data modeling in other future applications.

Keywords: Time series, Bayesian inference, Generalized linear models, Spatio-temporal analysis, Hierarchical analysis.

Conteúdo

1	Introdução	5
2	Fonte de dados	6
2.1	Dados do Rio dos Sinos	6
2.2	Índice da Qualidade da Água (IQA)	7
2.3	Dados do Rio Colorado	7
3	Metodologia	10
3.1	Descrição dos modelos	11
3.2	Modelo Espacial	15
3.3	Análises dos Modelos	16
3.4	Critérios de Seleção de Modelo	16
4	Dados do Rio dos Sinos	18
4.1	Modelos	18
4.2	Análise dos modelos	19
5	Dados do Rio Colorado nos Estados Unidos	21
5.1	Modelos para os dados do Rio Colorado	21
5.2	Análise dos modelos	22
5.3	Análise espaço-temporal	27
6	Conclusão	36
	Referências Bibliográficas	37
	Apêndice A	38
	Apêndice B	43
	Apêndice C	47

1 Introdução

O tema desse trabalho é analisar e desenvolver novos modelos lineares generalizados para séries temporais, comparando a eficiência dos mesmos com os já existentes na literatura, aplicando a metodologia hierárquica Bayesiana (Gelman et al., 2014 e Congdon, 2007). Dado as características de cada série temporal, como a tendência e a sazonalidade, determinar o modelo que melhor estima cada uma dessas características é de interesse prático e teórico. Existem muitos modelos lineares generalizados que são utilizados para analisar séries temporais na literatura, propostos por Waller et al. (1997). Porém, encontrar o modelo que melhor se adequa aos dados de acordo com suas características pode ser difícil. Portanto, esse conhecimento pode ser utilizado para facilitar e melhorar a eficiência da modelagem de dados em outras aplicações futuras.

O objetivo principal deste trabalho é desenvolver e analisar modelos lineares generalizados para séries temporais com diferentes características. Para o estudo utilizamos dois bancos de dados diferentes. Inicialmente aplicamos os modelos aos dados, que não apresentam sazonalidade, coletados em quatro pontos ao longo do Rio dos Sinos, que descrevem o Índice de Qualidade da Água (IQA) em quatro regiões. Já para analisar os modelos para as séries com sazonalidade utilizamos os dados do Rio Colorado em 29 regiões, que representam a vazão da água nos pontos de coleta.

A lista de modelos que visamos comparar compreende processos auto-regressivos aditivos de primeira ordem para a componente do efeito temporal, os modelos que consideram um processo auto-regressivos de primeira ordem através da média da distribuição *a priori*, proposto por Waller et al. (1997) e os que consideram um efeito de tendência linear aditiva ao efeito temporal. Também analisamos modelos auto-regressivos sazonais, permitindo a modelagem da sazonalidade que está presente nos dados.

A metodologia de interesse consiste da análise da estrutura espaço-temporal de Rampasso et al. (2016) nos modelos lineares generalizados, através da análise Bayesiana hierárquica (ver Gelman et al., 2014 e Congdon, 2007). A estrutura temporal é modelada através de processos auto-regressivos de primeira e segunda ordens com e sem sazonalidade. O efeito correlacionado espacialmente foi aplicado utilizando o modelo Intrínseco proposto por Besag et al. (1991).

Dentre os modelos analisados na primeira parte deste trabalho, acreditamos que aqueles que consideram um processo auto-regressivo aditivo de primeira ordem para a componente do efeito temporal apresentam melhores estimadores comparados com os modelos que consideram um processo auto-regressivo de primeira ordem através do valor da média da distribuição *a priori* e os modelos que consideram um efeito de tendência linear aditiva ao efeito temporal.

A partir desses resultados analisamos o melhor modelo, de acordo com os valores dos critérios de avaliação e dos gráficos dos dados com os estimadores, para séries temporais com sazonalidade e posteriormente realizamos uma análise multidimensional levando em consideração a informação espacial de cada região. Os critérios de avaliação utilizados para a comparação dos modelos foram: Logarithm of the Pseudo Marginal Likelihood - LPML (Congdon, 2007); Deviance Information Criterion - DIC (Spiegelhalter et al., 2002); Widely Applicable Information Criterion - WAIC (Gelman et al., 2014) e o Pearson Residuals - Q_p (Gelman et al., 2014).

2 Fonte de dados

2.1 Dados do Rio dos Sinos

Para o estudo utilizamos dois bancos de dados diferentes. Inicialmente aplicamos os modelos aos dados, que não apresentam sazonalidade, coletados em quatro pontos ao longo do Rio dos Sinos. Esses dados foram cedidos pela FEPAM (Fundação Estadual de Proteção Ambiental Henrique Luis Roessler), compreendendo quatro séries temporais que descrevem o Índice de Qualidade da Água (IQA) ao longo de 8 anos e meio (102 meses, durante o período de janeiro de 2000 a Junho de 2008) de cada um dos pontos de coleta.

Na Figura 1 abaixo, temos os pontos de coleta das 4 regiões do Rio dos Sinos.



Figura 1: Mapa dos quatro pontos de coleta do Rio dos Sinos.

Analisando a Figura 2 dos gráficos das séries temporais dos Índice de Qualidade da Água (IQA) observamos que no ponto 1 as medições são, em média, maiores em comparação aos outros pontos. É possível perceber também um decréscimo no IQA conforme o ponto de coleta se afasta da nascente. Observamos que o Ponto 3 apresenta a menor média do índice de qualidade da água.

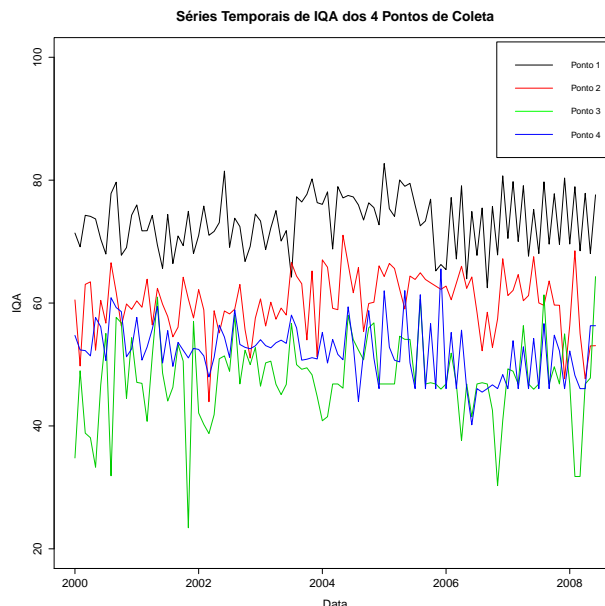


Figura 2: Gráfico das séries temporais dos IQA's nos quatro pontos de coleta.

2.2 Índice da Qualidade da Água (IQA)

O Índice da Qualidade da Água (IQA) leva em consideração nove parâmetros para ser calculado e varia entre 0 e 100, 0 sendo considerado o pior índice. Cada parâmetro de qualidade da água apresenta um peso diferente de acordo com a sua importância. São eles: Oxigênio dissolvido (0.17); coliformes termotolerantes (0.15); potencial hidrogeniônico - PH (0.12); demanda bioquímica de oxigênio - DBO5 (0.1); temperatura da água (0.1); nitrogênio total (0.1); fósforo total (0.1); turbidez (0.08) e resíduo total (0.08). O cálculo é feito com uma média ponderada, dado por

$$IQA = \prod_{i=1}^9 q_i^{w_i}, \quad (2.1)$$

onde q_i é a qualidade e w_i o peso do i -ésimo parâmetro.

2.3 Dados do Rio Colorado

Para analisar os modelos para as séries com sazonalidade utilizamos os dados do Rio Colorado em 29 regiões dos estados de Wyoming, Utah, Colorado, Novo México e Arizona, todos nos Estados Unidos. Esse banco é composto por 29 regiões que descrevem a vazão da água nos pontos de coleta durante 110 anos e 3 meses, ou seja, 1323 observações para cada região, durante o período de Outubro de 1905 a Dezembro de 2015.

Na Figura 3 abaixo, temos os pontos de coleta das 29 regiões de bacias hidrográficas dos estados de Wyoming, Utah, Colorado, Novo México e Arizona, todos nos Estados Unidos.



Figura 3: Mapa das 29 regiões.

Podemos perceber com a Figura 3 que algumas regiões estão bem próximas umas das outras, enquanto outras estão mais afastadas. Além disso, nem todas as regiões estão localizadas diretamente no Rio Colorado, mas em alguns rios menores conectados. Outro fator importante é a grande área em que os dados foram coletados, indicando que pode existir grande diferença geográfica entre as regiões. Por este motivo, analisamos o comportamento das séries temporais e separamos as 29 regiões em 3 grupos, de acordo com as respectivas médias.

Tabela 1: Valores das médias e respectivos desvios padrões, para cada uma das 29 regiões.

Região	Média	Desvio Padrão	Região	Média	Desvio Padrão
1	176667.237	213610.327	16	28410.228	47839.855
2	120842.667	137414.386	17	15154.767	21498.629
3	12670.546	14377.056	18	97409.760	116613.222
4	77936.921	92516.747	19	77396.859	76421.411
5	17396.062	23264.354	20	33300.459	91820.518
6	85596.859	94707.279	21	1747.346	2085.981
7	67075.252	90495.411	22	14385.825	29047.680
8	4953.349	28037.718	23	14473.337	28553.048
9	109167.369	125131.990	24	14204.726	15249.975
10	7334.005	11202.504	25	25627.855	55965.276
11	46743.565	56124.984	26	15340.041	28715.634
12	102614.670	145772.923	27	7927.070	23435.070
13	38634.252	50627.109	28	5989.737	30067.602
14	65289.295	74464.078	29	7241.291	37740.981
15	47429.373	41652.591			

De acordo com a Tabela 1, percebemos que algumas médias estão bem distantes umas das outras e os valores dos desvios padrões são muito altos, o que indica que as regiões apresentam uma grande dispersão ao longo do tempo.

Os grupos foram separados da seguinte forma:

Grupo 1: Regiões 1, 2, 9 e 12. As médias dessas regiões variam entre 102.614,70 até 176.698,10.

Grupo 2: Regiões 4, 6, 7, 11, 13, 14, 15, 16, 18, 20 e 25. As médias dessas regiões variam entre 25.627,85 até 97.409,76.

Grupo 3: Regiões 3, 5, 8, 10, 17, 21, 22, 23, 24, 26, 27, 28, 29. As médias dessas regiões variam entre 1.747,35 até 17.396,06.

Tabela 2: Divisão dos Dados em 3 Grupos, conforme as médias das regiões.

Grupo	Regiões	Médias
1	1, 2, 9 e 12	102,614.70 até 176,698.10
2	4, 6, 7, 11, 13, 14, 15, 16, 18, 20 e 25	25,627.85 até 97,409.76
3	3, 5, 8, 10, 17, 21, 22, 23, 24, 26, 27, 28, 29	1,747.35 até 17,396.06

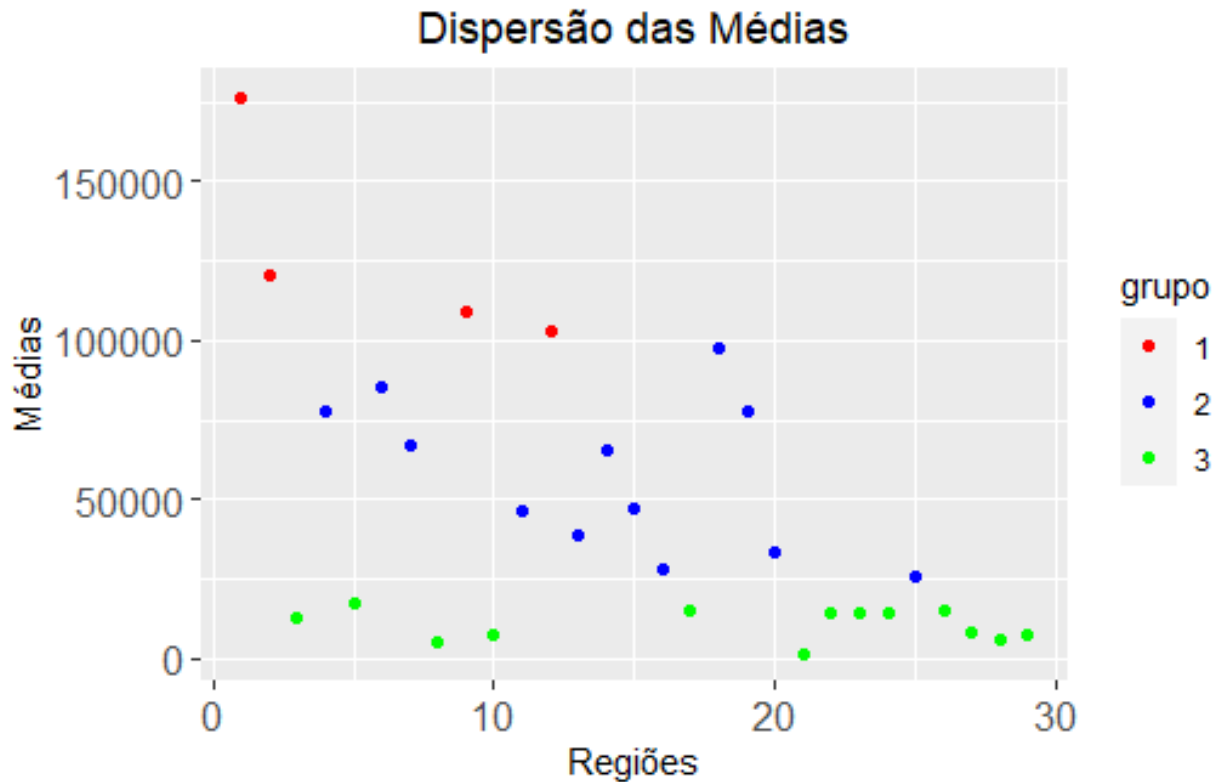


Figura 4: Gráfico de dispersão das médias das 29 regiões.

Podemos perceber pela Figura 4 que as médias das regiões do Grupo 1, representadas pelos pontos vermelhos, são as maiores. O Grupo 2, representado pelos pontos azuis, apresenta uma grande dispersão das médias das regiões. O Grupo 3, representado pelos pontos verdes, apresenta as regiões com as menores médias e a dispersão entre elas também é pequena.

3 Metodologia

Este trabalho caracteriza-se por desenvolver e analisar modelos que podem ser aplicados em outros trabalhos futuros envolvendo dados numéricos. A metodologia de interesse consiste da análise da estrutura espaço-temporal de Rampasso et al. (2016) nos modelos lineares generalizado, através da análise Bayesiana hierárquica (ver Gelman et al., 2014 e Congdon, 2007). A estrutura temporal é modelada através de processos auto-regressivos de primeira e segunda ordens, como proposto em Martínez-Beneito et al. (2008).

- Fase 1: Análise descritiva dos dados coletados em quatro pontos de coleta ao longo do Rio dos Sinos;
- Fase 2: Seleção dos modelos a serem utilizados;
- Fase 3: Análises para os modelos foram realizadas e os resultados comparados;

- Fase 4: Análise descritiva dos dados coletados em 29 pontos do Rio Colorado, nos Estados Unidos;
- Fase 5: Seleção e análise dos modelos para séries temporais com sazonalidade;
- Fase 6: Aplicação de uma matriz de vizinhos e análise utilizando a estrutura espaço-temporal.

3.1 Descrição dos modelos

Os modelos utilizados para as análises do banco de dados do Rio dos Sinos e para o banco de dados do Rio Colorado estão descritos abaixo.

Os modelos lineares são aqueles que apresentam uma relação linear nos parâmetros entre as variáveis, impondo que uma variação em um dos parâmetros não implica em uma variação nos demais. O modelo de regressão linear é dado a seguir

$$\mu_{it} \equiv E(Y_{it}) = X_{it}'\beta, \quad (3.1)$$

onde $Y_{it} \sim \mathcal{N}(\mu_{it}, \sigma^2)$, X_{it} o vetor de variáveis explicativas, β o vetor de parâmetros desconhecidos e $X_{it}'\beta$ o preditor linear.

Para obtermos o modelo linear generalizado proposto por Nelder et al. (1972) adaptamos a equação (3.1) de duas formas. A primeira é a inserção da função monótona e diferenciável $g(\cdot)$ e a segunda é que Y_{it} agora passa a ter distribuição da família exponencial uniparamétrica.

O modelo linear generalizado de interesse neste trabalho é dado por:

$$\mu_{it} \equiv E(Y_{it}),$$

$$g(\mu_{it}) = \alpha_0 + S_i(t) + R_i, \quad (3.2)$$

onde Y_{it} é a contagem de um fenômeno aleatório que segue uma distribuição que pertence à família exponencial uniparamétrica, μ_{it} é sua esperança matemática, α_0 é uma constante que representa a média total das regiões, definida como $\alpha_0 \sim \mathcal{N}(\mu, 0.001)$, onde μ representa a média do IQA, $S_i(t)$ é o efeito temporal associado a qualquer região i e $g(\cdot)$, uma função de ligação que relaciona as componentes espaço-temporal e aleatória, corresponde a qualquer função monótona e diferenciável.

A componente R_i é a soma de efeitos correlacionados espacialmente (b_i) e efeitos aleatórios não-correlacionados (h_i), onde $h_i \sim \mathcal{N}(0, \sigma_h^2)$ e b_i será modelado através de processos auto-regressivos condicionais Intrínsecos (*ICAR*), descritos na Seção 3.2. Primeiramente incluiremos apenas os efeitos não-correlacionados (h_i) nas análises.

Os modelos para explicar a componente de efeito temporal, $S_i(t)$ que consideramos neste trabalho são: modelos auto-regressivo aditivo de primeira ordem, ou seja, consideram uma simples tendência linear aditiva; auto-regressivo de primeira ordem através da média da distribuição *a priori*, podendo apresentar um parâmetro ρ extra na estrutura auto-regressiva; modelos com efeito de tendência linear aditiva, podendo apresentar o parâmetro ρ extra nesta estrutura auto-regressiva aditiva.

Os modelos estão descritos detalhadamente abaixo. Dentre eles, temos que os modelos $M4$, $M4^*$, $M5$ e $M5^*$ foram propostos neste trabalho.

Modelo 1: Este modelo descreve um efeito de tendência linear aditiva para todas as N regiões e é dado por

$$\mathbf{M1} : g(\mu_{it}) = \alpha_0 + \delta t + R_i, \quad (3.3)$$

onde δ é uma constante com o mesmo valor para todas as N regiões tal que $\delta | \sigma_\delta^2 \sim \mathcal{N}(0, \sigma_\delta^2)$, com σ_δ^2 o hiperparâmetro de dispersão.

Modelo 1*: Este modelo descreve um efeito de tendência linear aditiva para cada uma das i -ésimas regiões e é dado por

$$\mathbf{M1}^* : g(\mu_{it}) = \alpha_0 + \delta_i t + R_i, \quad (3.4)$$

onde δ_i é uma constante dependendo da i -ésima região tal que $\delta_i | \sigma_{\delta_i}^2 \sim \mathcal{N}(0, \sigma_{\delta_i}^2)$, com $\sigma_{\delta_i}^2$ o hiperparâmetro de dispersão, para todo $i \in \{1, \dots, N\}$.

Modelo 2: Este modelo considera um processo auto-regressivo de primeira ordem através do valor da média da distribuição *a priori* $A_1(t)$, com um hiperparâmetro extra ρ , para todas as N regiões, proposto por Waller et al. (1997), e é dado por

$$\mathbf{M2} : g(\mu_{it}) = \alpha_0 + A_1(t) + R_i, \quad (3.5)$$

onde a distribuição *a priori* $A_1(t) \sim \mathcal{N}(\phi_t + \rho(A_1(t-1) - \phi_{t-1}), \sigma_{\epsilon_t}^2)$, com $\phi_t = \phi_0 + \phi_1 t$, $\phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$ e $\sigma_{\epsilon_t}^2$ o hiperparâmetro de dispersão.

Modelo 2*: Este modelo considera um processo auto-regressivo de primeira ordem através do valor da média da distribuição *a priori* $A_1(t)$, com um hiperparâmetro extra ρ , para a i -ésima região, proposto por Waller et al. (1997), e é dado por

$$\mathbf{M2}^* : g(\mu_{it}) = \alpha_0 + A_1(i, t) + R_i, \quad (3.6)$$

onde a distribuição *a priori* $A_1(i, t) \sim \mathcal{N}(\phi_t + \rho(A_1(i, t-1) - \phi_{t-1}), \sigma_{\epsilon_{i,t}}^2)$, com $\phi_t = \phi_0 + \phi_1 t$, $\phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$ e $\sigma_{\epsilon_{i,t}}^2$ o hiperparâmetro de dispersão.

Modelo 3: Este modelo considera um processo auto-regressivo de primeira ordem através do valor da média da distribuição *a priori* $A_2(t)$, sem o hiperparâmetro extra ρ , para todas as N regiões, proposto por Waller et al. (1997), e é dado por

$$\mathbf{M3} : g(\mu_{it}) = \alpha_0 + A_2(t) + R_i, \quad (3.7)$$

onde a distribuição *a priori* $A_2(t) \sim \mathcal{N}(\phi_1 A_2(t-1), \sigma_{\epsilon_t}^2)$, com $\phi_1 \in \mathbb{R}$ e $\sigma_{\epsilon_t}^2$ o hiperparâmetro de dispersão.

Modelo 3*: Este modelo considera um processo auto-regressivo de primeira ordem através do valor da média da distribuição *a priori* $A_2(t)$, sem o hiperparâmetro extra ρ , para a i -ésima região, proposto por Waller et al. (1997), e é dado por

$$\mathbf{M3}^* : g(\mu_{it}) = \alpha_0 + A_2(i, t) + R_i, \quad (3.8)$$

onde a distribuição *a priori* $A_2(i, t) \sim \mathcal{N}(\phi_1 A_2(i, t-1), \sigma_{\epsilon_t}^2)$, com $\phi_1 \in \mathbb{R}$ e $\sigma_{\epsilon_t}^2$ o hiperparâmetro de dispersão.

Modelo 4: Este modelo considera um processo auto-regressivo aditivo de primeira ordem para a componente do efeito temporal $S_i(t)$, para todas as N regiões, proposto neste trabalho, e é dado por

$$\mathbf{M4} : g(\mu_{it}) = \alpha_0 + A_3(t) + R_i, \quad (3.9)$$

onde $A_3(t) = \phi_1 A_3(t-1) + \epsilon_t$, com $|\phi_1| < 1$ e $\epsilon_t \sim \mathcal{N}(0, \sigma_{\epsilon_t}^2)$ e $\sigma_{\epsilon_t}^2$ o hiperparâmetro de dispersão.

Modelo 4*: Este modelo considera um processo auto-regressivo aditivo de primeira ordem para a componente do efeito temporal $S_i(t)$, para a i -ésima região, proposto neste trabalho, e é dado por

$$\mathbf{M4}^* : g(\mu_{it}) = \alpha_0 + A_3(i, t) + R_i, \quad (3.10)$$

onde $A_3(i, t) = \phi_1 A_3(i, t-1) + \epsilon_{i,t}$, com $|\phi_1| < 1$ e $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_{\epsilon_{i,t}}^2)$ e $\sigma_{\epsilon_{i,t}}^2$ o hiperparâmetro de dispersão.

Modelo 5: Este modelo considera um processo auto-regressivo aditivo de primeira ordem para a componente do efeito temporal $S_i(t)$, com um hiperparâmetro extra ρ , para todas as N regiões, proposto neste trabalho, e é dado por

$$\mathbf{M5} : g(\mu_{it}) = \alpha_0 + A_4(t) + R_i, \quad (3.11)$$

onde $A_4(t) = \phi_t + \rho(A_4(t-1) - \phi_{t-1}) + \epsilon_t$, com $\phi_t = \phi_0 + \phi_1 t$, $\phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$ e $\epsilon_t \sim \mathcal{N}(0, \sigma_{\epsilon_t}^2)$ e $\sigma_{\epsilon_t}^2$ o hiperparâmetro de dispersão.

Modelo 5*: Este modelo considera um processo auto-regressivo aditivo de primeira ordem para a componente do efeito temporal $S_i(t)$, com o hiperparâmetro extra ρ , para a i -ésima região, proposto neste trabalho, e é dado por

$$\mathbf{M5}^* : g(\mu_{it}) = \alpha_0 + A_4(i, t) + R_i, \quad (3.12)$$

onde $A_4(i, t) = \phi_t + \rho(A_4(i, t-1) - \phi_{t-1}) + \epsilon_{i,t}$, com $\phi_t = \phi_0 + \phi_1 t$, $\phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$ e $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_{\epsilon_{i,t}}^2)$ e $\sigma_{\epsilon_{i,t}}^2$ o hiperparâmetro de dispersão.

Modelo 6: Este modelo considera um processo auto-regressivo para uma série temporal com sazonalidade para a componente do efeito temporal $S_i(t)$, com um hiperparâmetro extra ρ , para todas as N regiões, proposto neste trabalho, e é dado por

$$\mathbf{M6} : g(\mu_{it}) = \alpha_0 + A_5(t) + R_i, \quad (3.13)$$

onde $A_5(t) \sim \mathcal{N}(\phi_t + \rho(\phi A_5(t-1)) + \rho(A_5(t-12)) - \rho(\phi A_5(t-13)) - \rho(\phi_{t-1} - \phi_1)),$ com $\phi_t = \phi_0 + \phi_1 t$, $\phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$ e $\sigma_{\epsilon_t}^2$ o hiperparâmetro de dispersão.

Modelo 6*: Este modelo considera um processo auto-regressivo para uma série temporal com sazonalidade para a componente do efeito temporal $S_i(t)$, com o hiperparâmetro extra

ρ , para a i -ésima região, proposto neste trabalho, e é dado por

$$\mathbf{M6}^* : g(\mu_{it}) = \alpha_0 + A_5(i, t) + R_i, \quad (3.14)$$

onde $A_5(i, t) \sim \mathcal{N}(\phi_t + \rho(\phi A_5(i, t-1)) + \rho(A_5(i, t-12)) - \rho(\phi A_5(i, t-13)) - \rho(\phi_{t-1} - \phi_1))$, com $\phi_t = \phi_0 + \phi_1 t$, $\phi, \phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$ e $\sigma_{\varepsilon_{i,t}}^2$ o hiperparâmetro de dispersão.

Os modelos $M6$ e $M6^*$ envolvem os processos SARIMA(p, d, q) \times (P, D, Q) $_s$ que estão na Definição 3.1, a seguir.

Definição 3.1 *O modelo SARIMA(p, d, q) \times (P, D, Q) $_s$ geral é dado por*

$$\Phi_P(\mathcal{B}^s)\phi_p(\mathcal{B})\nabla_s^D\nabla^d X_t = \delta + \theta_q(\mathcal{B})\Theta_Q(\mathcal{B}^s)\varepsilon_t, \quad \text{para todo } t = 1, \dots, n, \quad (3.15)$$

onde os polinômios acima são definidos por

$$\begin{aligned} \phi_p(\mathcal{B}) &= 1 - \phi_1(\mathcal{B}) - \phi_2(\mathcal{B}^2) - \dots - \phi_p(\mathcal{B}^p), \quad |\phi_j| < 1, \quad 1 \leq j \leq p, \\ \Phi_P(\mathcal{B}^s) &= 1 - \Phi_1(\mathcal{B}^s) - \Phi_2(\mathcal{B}^{2s}) - \dots - \Phi_P(\mathcal{B}^{Ps}), \quad |\Phi_l| < 1, \quad 1 \leq l \leq P, \\ \theta_q(\mathcal{B}) &= 1 + \theta_1(\mathcal{B}) + \theta_2(\mathcal{B}^2) + \dots + \theta_q(\mathcal{B}^q), \quad |\theta_i| < 1, \quad 1 \leq i \leq q, \\ \Theta_Q(\mathcal{B}^s) &= 1 + \Theta_1(\mathcal{B}^s) + \Theta_2(\mathcal{B}^{2s}) + \dots + \Theta_Q(\mathcal{B}^{Qs}), \quad |\Theta_h| < 1, \quad 1 \leq h \leq Q, \end{aligned}$$

$\nabla_s^D \equiv (1 - \mathcal{B}^s)^D$, $\nabla^d \equiv (1 - \mathcal{B})^d$ com $d, D \in \mathbb{N} - \{0\}$ e onde $s \in \mathbb{N} - \{0\}$ é a sazonalidade.

Temos a seguir um exemplo para este modelo.

Exemplo 3.1: Especificamos o modelo SARIMA(1, 0, 0) \times (0, 1, 1) $_{12}$. De acordo com a equação (3.15), temos:

$$\phi_1(\mathcal{B})\nabla_{12}^1 X_t = \delta + \theta_1(\mathcal{B}^{12})\varepsilon_t,$$

ou seja,

$$(1 - \phi_1\mathcal{B})(1 - \mathcal{B}^{12})X_t \equiv \delta + (1 + \theta_1\mathcal{B}^{12})\varepsilon_t$$

$$(1 - \phi_1\mathcal{B} - \mathcal{B}^{12} + \phi_1\mathcal{B}^{13})X_t \equiv \delta + \varepsilon_t + \theta_1\varepsilon_{t-12}. \quad (3.16)$$

Utilizamos a notação $\phi_1 \equiv \phi$ e $\theta_1 \equiv \theta$. Então, (3.16) pode ser reescrita por:

$$X_t - \phi X_{t-1} - X_{t-12} + \phi X_{t-13} = \delta + \varepsilon_t + \theta\varepsilon_{t-12}. \quad (3.17)$$

Ou seja, o modelo SARIMA(1, 0, 0) \times (0, 1, 1) $_{12}$ é dado por

$$X_t = \phi X_{t-1} + X_{t-12} - \phi X_{t-13} + \delta + \varepsilon_t + \theta \varepsilon_{t-12}. \quad (3.18)$$

□

Para os modelos estamos considerando que δ é uma constante com o mesmo valor para todas as N regiões tal que $\delta | \sigma_\delta^2 \sim \mathcal{N}(0, \sigma_\delta^2)$, δ_i é uma constante dependendo da i -ésima região tal que $\delta_i | \sigma_{\delta_i}^2 \sim \mathcal{N}(0, \sigma_{\delta_i}^2)$, $\phi_t = \phi_0 + \phi_1 t$, $\phi_0, \phi_1 \in \mathbb{R}$, $|\rho| < 1$, $\varepsilon_t \sim \mathcal{N}(0, \sigma_{\varepsilon_t}^2)$, $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_{\varepsilon_{i,t}}^2)$ e $\sigma_{\varepsilon_t}^2$ o hiperparâmetro de dispersão. Temos que $\phi_0 \sim \mathcal{N}(0, 0.01)$; $\phi_1 \sim \mathcal{N}(0, 0.1)$; $\rho \sim \mathcal{N}(0, 1)$ e $\sigma_{\varepsilon_t}^2 = \frac{1}{\tau_t}$, onde $\tau_t \sim \mathcal{N}(0.1, 0.1)$. Os modelos A_k , para $k \in \{1, \dots, 5\}$ são especificados como distribuições *a priori* e foram descritos nos seus respectivos modelos.

O parâmetro ρ foi considerado por Waller et al. (1997) como livre de qualquer imposição. Vamos considerar $|\rho| < 1$ para garantir a convergência das cadeias de Markov quando usamos o algoritmo Monte Carlo Markov Chain (MCMC). Ao leitor, indicamos Geyer (1992) e Tierney (1994) para o algoritmo MCMC.

Por fim, assumindo independência *a priori* dos parâmetros e hiperparâmetros dos modelos e as distribuições *a prioris* definidas anteriormente, temos que a distribuição a posteriori conjunta é proporcional a

$$\begin{aligned} \Pi(\theta, y) \propto \mathcal{L}(\theta, y) \pi(S(t)) & \left(\frac{1}{\sigma_b^2} \right)^{\frac{N}{2} + \gamma_1 + 1} \exp \left(\frac{-1}{\sigma_b^2} \left(\gamma_2 + p \sum_{i=1}^N \frac{N_i (b_i - \bar{b}_i)^2}{2} \right) \right) \\ & \times \left(\frac{1}{\sigma_h^2} \right)^{\frac{N}{2} + \gamma_1 + 1} \exp \left(\frac{-1}{\sigma_h^2} \left(\gamma_2 + \sum_{i=1}^N \frac{h_i^2}{2} \right) \right) \exp \left(-\frac{p^2}{2} \right) I_p(0, 1), \end{aligned} \quad (3.19)$$

onde $\pi(\cdot)$ é uma função de distribuição *a priori*, $S_i(t)$ é definido a partir dos modelos descritos na Seção 3.1, $\mathcal{L}(\theta, y)$ é a função verossimilhança para o vetor de dados y e o vetor de parâmetros dados por $\theta = \{\theta_i\}$. Os componentes do vetor de parâmetros θ são p , h_i , b_i , α_0 , σ_b^2 , σ_h^2 , com $i \in \{1, \dots, N\}$, além dos parâmetros δ , δ_i , σ_δ , α . Para o efeito correlacionado espacialmente b , que está definido como um modelo Espacial Intrínseco na Seção 3.2, temos que $p = 1$.

3.2 Modelo Espacial

A distribuição condicional completa de um modelo para análise espacial geral (CAR), proposto por Torabi e Rosychuk (2010), é definida por

$$b_i | b_{-i} \sim \mathcal{N} \left(\mu_i + p \sum_{j \sim i} C_{ij} (b_j - \mu_j), \sigma_b^2 m_{ii} \right), \quad (3.20)$$

onde p é a dependência espacial, $j \sim i$ a presença de vizinhança entre as áreas j e i ; C_{ij} a matriz de associação espacial que apresenta a estrutura de vizinhos; σ_b^2 a variância geral do modelo; m_{ii} uma matriz diagonal e b , proposto por Cressie (2015), é definido por (3.21)

$$b \sim \mathcal{NM}(\mu, \Sigma). \quad (3.21)$$

onde $\mu_i = (\mu_1, \dots, \mu_N)'$ é um vetor de médias, Σ é a matriz de covariâncias que considera a estrutura de vizinhos e $b_i = (b_1, \dots, b_N)'$ o vetor de efeitos correlacionados espacialmente.

O modelo Intrínseco (ICAR) proposto por Besag et al. (1991) considera que $C_{ij} = 1/n_i$ se as áreas i e j forem adjacentes e 0 caso contrário; $m_{ii} = 1/n_i$; $p = 1$; n_i é o número de vizinhos da área i , $\mu_i = 0$ e $b_{-i} = (b_1, \dots, b_{i-1}, b_{i+1}, \dots, b_N)'$ o vetor b sem a i -ésima componente. Assim, a distribuição condicional completa do modelo Intrínseco é dada por

$$b_i | b_{-i} \sim \mathcal{N}\left(\frac{1}{n_i} \sum_{j \sim i} b_j, \frac{\sigma_b^2}{n_i}\right), \quad (3.22)$$

3.3 Análises dos Modelos

Realizamos as análises dos modelos propostos na Seção 3.1, para avaliar os seus ajustes aos dados. As análises foram baseadas no método de Monte Carlo através de cadeias de Markov, o MCMC (Geyer, 1992; Tierney, 1994). Para isso, utilizamos a inferência Bayesiana que, diferente da estatística clássica, nos permite associar uma probabilidade a um número aleatório com qualquer grau de incerteza. Isso nos permite definir probabilidades *a priori* e *a posteriori*.

A probabilidade *a priori* é obtida a partir de informações que cada indivíduo traz do passado, enquanto a probabilidade *a posteriori* incorpora os novos dados obtidos enquanto o experimento é realizado. Assim, as probabilidades associadas com algum evento são pré-definidas e atualizadas ao introduzir estas novas informações à análise. Quando os dados são mais informativos do que as suposições iniciais, ou seja, quando o tamanho da amostra aumenta, as probabilidades convergem independente da distribuição inicial (*a priori*). É com esta abordagem que procedemos com a análise dos modelos propostos na Seção 3.1 utilizando o software *OpenBUGS*.

No início de cada análise definimos um Burn-in, que corresponde ao número da iteração em que os parâmetros começam a ser utilizados para calcular a distribuição *a posteriori*. O Thin, que corresponde ao intervalo l em que as observações são utilizados, o restante é descartado. E o número de interações, que corresponde ao tamanho da cadeia.

Com essas análises calculamos os critérios de seleção dos modelos descritos na Seção 3.4 a seguir, possibilitando a sua comparação.

3.4 Critérios de Seleção de Modelo

Os critérios de comparação dos modelos utilizados nas análises do banco de dados do Rio dos Sinos e para o banco de dados do Rio Colorado estão descritos abaixo.

Deviance Information Criterion (DIC):

De acordo com o *Deviance Information Criterion*, quanto menor o seu valor, melhor o modelo proposto. O DIC é descrito em Spiegelhalter et al. (2002) por

$$DIC = D(\mathbb{E}_{\theta,y}(\theta)) - 2p_D = \mathbb{E}_{\theta,y}(D) + p_D, \quad (3.23)$$

onde

$$p_D = \mathbb{E}_{\theta|y}(D) - D(\mathbb{E}_{\theta|y}(\theta)) \quad e \quad D = -2 \ln(L(y, \theta)).$$

Logarithm of the Pseudo Marginal Likelihood (LPML):

De acordo com o *Logarithm of the Pseudo Marginal Likelihood*, quanto maior o seu valor, melhor o modelo proposto. O LPML é descrito em Congdon (2007) por

$$\widehat{LPML}_{it} = \sum_{i=1}^N \sum_{t=1}^T \ln(\widehat{CPO}_{it}), \quad (3.24)$$

onde

$$\widehat{CPO}_{it} = \left(\frac{1}{n} \sum_{k=1}^n \frac{1}{f(y_{it} | \theta^k)} \right)^{-1},$$

onde θ^k é o θ k-ésimo valor gerado pela distribuição *a posteriori* conjunta.

Pearson Residuals (Q_p):

De acordo com o *Pearson Residuals*, quanto menor o seu valor, melhor o modelo proposto. O Q_p é descrito em Paulino et al. (2009) por

$$Q_p = \sum_{i=1}^N \sum_{t=1}^T \frac{(y_{it} - N_{it}\theta_{it})^2}{(1 - \theta_{it})N_{it}\theta_{it}}. \quad (3.25)$$

Widely Applicable Information Criterion (WAIC):

De acordo com o *Widely Applicable Information Criterion*, quanto menor o seu valor, melhor o modelo proposto. O WAIC é descrito em Gelman et al. (2014), Vehtari et al. (2017), Watanabe (2013a) e Watanabe (2013b) por

$$WAIC = LPPD - \hat{p}_{WAIC}, \quad (3.26)$$

onde \hat{p}_{WAIC} pode ser definido por (3.27) e (3.28) e LPPD é definido por (3.29), como seguem

$$\hat{p}_{WAIC1} = 2 \sum_{i=1}^N \sum_{t=1}^T \left[\ln \left(\frac{1}{n} \sum_{k=1}^n f(y_{it} | \theta^k) \right) - \frac{1}{n} \sum_{k=1}^n f(y_{it} | \theta^k) \right], \quad (3.27)$$

$$\hat{p}_{WAIC2} = \sum_{i=1}^N \sum_{t=1}^T \left[\frac{1}{n} \sum_{k=1}^n \left(\ln(f(y_{it} | \theta^k)) - \frac{1}{n} \sum_{k=1}^n \ln(f(y_{it} | \theta^k)) \right)^2 \right], \quad (3.28)$$

$$LPPD = \ln \left(\prod_{i=1}^N \prod_{t=1}^T f(y_{it} | \theta) \right) = \sum_{i=1}^N \sum_{t=1}^T \int f(y_{it} | \theta) f(\theta | y_{it}) d(\theta), \quad (3.29)$$

onde θ^k é o θ k-ésimo valor gerado pela distribuição *a posteriori* conjunta.

4 Dados do Rio dos Sinos

Vamos começar a análise dos modelos com o banco de dados coletados em quatro pontos ao longo do Rio dos Sinos, descritos na Seção 2.

4.1 Modelos

Começamos pelos modelos que não possuem o efeito correlacionado espacialmente (b_i), mas apenas o efeito aleatório (h_i).

Consideramos dez modelos para explicar a componente de efeito temporal, $S_i(t)$, quando utilizamos os dados do Rio dos Sinos. A Tabela 1 abaixo fornece esses modelos de interesse.

Tabela 3: Dez diferentes modelos que descrevem o efeito temporal de interesse.

Modelo	$S_i(t)$	Especificação do Modelo
$M1$	δt	tendência linear aditiva para todos os 4 pontos de coleta
$M1^*$	$\delta_i t$	tendência linear aditiva para a i -ésima região
$M2$	$A_1(t)$	$A_1(t) \sim \mathcal{N}(\phi_t + \rho(A_1(t-1) - \phi_{t-1}), \sigma_{\epsilon_t}^2)$
$M2^*$	$A_1(i, t)$	$A_1(i, t) \sim \mathcal{N}(\phi_t + \rho(A_1(i, t-1) - \phi_{t-1}), \sigma_{\epsilon_{i,t}}^2)$
$M3$	$A_2(t)$	$A_2(t) \sim \mathcal{N}(\phi_1 A_2(t-1), \sigma_{\epsilon_t}^2)$
$M3^*$	$A_2(i, t)$	$A_2(i, t) \sim \mathcal{N}(\phi_1 A_2(i, t-1), \sigma_{\epsilon_{i,t}}^2)$
$M4$	$A_3(t)$	$A_3(t) = \phi_1 A_3(t-1) + \epsilon_t$
$M4^*$	$A_3(i, t)$	$A_3(i, t) = \phi_1 A_3(i, t-1) + \epsilon_{i,t}$
$M5$	$A_4(t)$	$A_4(t) = \phi_t + \rho(A_4(t-1) - \phi_{t-1}) + \epsilon_t$
$M5^*$	$A_4(i, t)$	$A_4(i, t) = \phi_t + \rho(A_4(i, t-1) - \phi_{t-1}) + \epsilon_{i,t}$

M_j denota o modelo para todas as 4 regiões enquanto que M_j^* denota aqueles para a i -ésima região, onde $i \in \{1, \dots, 4\}$ e $j \in \{1, \dots, 5\}$.

Os modelos $M1$ e $M1^*$ são aqueles que consideram apenas uma simples tendência linear aditiva ao efeito temporal. Os modelos $M2$, $M2^*$, $M3$ e $M3^*$ possuem um processo auto-regressivo imposto na média da distribuição *a priori*, porém, os modelos $M2$ e $M2^*$ apresentam um parâmetro ρ extra na sua estrutura auto-regressiva. Já os modelos $M4$, $M4^*$, $M5$ e $M5^*$ consideram o processo auto-regressivo imposto aditivamente na componente $S_i(t)$, mas apenas os modelos $M4$ e $M4^*$ apresentam um parâmetro ρ extra nesta estrutura auto-regressiva aditiva.

A diferença entre os modelos M_j e M_j^* , onde $j \in \{1, 2, 3, 4, 5\}$, é que para aqueles sem asterisco os modelos são considerados para todos os 4 pontos, enquanto os modelos com asterisco são considerados diferentemente para cada i -ésima região, onde $i \in \{1, 2, 3, 4\}$.

4.2 Análise dos modelos

Para comparar a eficiência dos modelos realizamos diferentes análises de acordo com a metodologia descrita na Seção 3.3 e calculamos os critérios de avaliação descritos na Seção 3.4. Além disso, geramos os gráficos dos estimadores para todas as regiões dos dez modelos para esse banco de dados. Com isso, conseguimos avaliar o desempenho dos modelos propostos na Seção 3.1 e concluir qual descreve melhor os dados.

As análises foram realizadas de acordo com a descrição em 3.3. Primeiramente definimos valores para o Thin, Burn-in e Iterações diferentes para cada análise e um número de cadeias sempre igual a 2. O tamanho amostral é calculado a partir das informações dos parâmetros anteriores. Esses valores definem onde vamos iniciar as análises, de quanto em quanto vamos analisar as observações e qual o total que vamos observar.

O sistema operacional utilizado para as análises é o Windows 10, o software R versão 3.5.3 e o programa OpenBUGS versão 3.2.3. O tempo das análises para este banco de dados depende dos parâmetros definidos, quanto maiores os parâmetros selecionados, maior o tempo computacional.

A Tabela 4 descreve os resultados de algumas análises, indicando inicialmente os parâmetros utilizados. Além dos critérios de avaliação, a Tabela 4 apresenta os valores para as variâncias, muito importante para avaliar o desempenho dos modelos.

Não foi possível obter os valores para os critérios de avaliação para os modelos M_3 e M_3^* e ainda não sabemos o motivo para isso ter acontecido. Portanto, esses modelos foram avaliados somente pelos gráficos dos estimadores, apresentados no Apêndice A.

Podemos perceber com a Tabela 4 que as análises com um tamanho amostral maior apresentam resultados com variâncias menores. Valores para o Burn-in e Thin altos não apresentam melhoras tão significativas para os critérios de avaliação. Portanto, não é necessário definir um Burn-in e um Thin muito elevados, mas aumentar o tamanho amostral pode trazer resultados positivos. Contudo, ao aumentar o tamanho amostral ou o número de Iterações o tempo computacional de cada análise aumentará.

Tabela 4: Resultados das Análises para os dados do Rio dos Sinos

Análise 1: Thin = 1; Burn-in = 2000; Iterações = 4000; Amostra = 2000						
Modelo	Q_p	LPML	DIC	WAIC1	WAIC2	σ_h^2
M1	217.5351	-519.6783	2539	1039.3360	1039.3570	110.0175
M1*	213.2928	-517.8481	2533	1035.6280	1035.6960	110.3386
M2	196.3001	-510.8455	2516	1021.2810	1021.6470	281.8832
M2*	189.6201	-509.0046	2518	1017.3440	1017.9190	278.1900
M4	196.7200	-509.0100	2511	1017.3400	1017.9100	200.0100
M4*	187.1335	-507.6382	2513	1014.6400	1015.2290	1257.4284
M5	187.6688	-508.8739	2388	1016.4780	1017.7650	2586.9272
M5*	111.3786	-500.1784	2594	944.3579	961.4162	66.9694
Análise 2: Thin = 1; Burn-in = 2000; Iterações = 6000; Amostra = 4000						
Modelo	Q_p	LPML	DIC	WAIC1	WAIC2	σ_h^2
M1	217.5784	-519.6987	2539	1039.3760	1039.3970	115.5254
M1*	213.2840	-517.8284	2533	1035.5900	1035.6560	117.8288
M2	196.7975	-511.0326	2516	1021.6540	1022.0220	244.3963
M2*	189.3268	-508.6709	2516	1016.6060	1017.2630	241.7721
M4	195.8270	-506.9870	2513	1013.2110	1013.9080	228.5820
M4*	185.2142	-507.0529	2511	1013.3260	1014.0430	1030.5601
M5	188.1707	-508.9363	2469	1016.7330	1017.8340	3114.7207
M5*	93.4685	-480.7546	2574	920.7275	930.9065	77.9592
Análise 3: Thin = 15; Burn-in = 2000; Iterações = 4000; Amostra = 113						
Modelo	Q_p	LPML	DIC	WAIC1	WAIC2	σ_h^2
M1	217.5805	-519.7098	2539	1039.3970	1039.4200	120.1427
M1*	213.3124	-517.8366	2539	1035.6070	1035.6720	118.3244
M2	196.8719	-510.9740	2515	1021.5560	1021.9120	227.8500
M2*	191.7137	-509.4827	2515	1018.2230	1018.8600	293.5846
M4	194.6083	-509.4827	2512	1018.2230	1018.8600	210.6531
M4*	190.8636	-509.5754	2493	1019.0270	1018.9710	3257.8919
M5	196.4516	-510.5921	2511	1020.9840	1021.1660	6993.3167
M5*	13.9146	-414.7167	-13780	829.22620	829.4034	215.9935
Análise 4: Thin = 5; Burn-in = 500; Iterações = 4500; Amostra = 800						
Modelo	Q_p	LPML	DIC	WAIC1	WAIC2	σ_h^2
M1	217.5589	-519.6960	2539	1039.3710	1039.3920	125.7081
M1*	213.3131	-517.8501	2533	1035.6330	1035.7000	133.7193
M2	196.7041	-510.8716	2515	1021.3620	1021.7100	251.9210
M2*	175.4700	-503.7100	-7200	1016.4900	1005.8700	257.8383
M4	195.5126	-507.0529	2513	1013.3260	1014.0430	417.1062
M4*	185.6200	-507.2300	2511	1013.8600	1014.3700	417.1000
M5	195.0970	-510.7094	2504	1021.2090	1021.2470	3269.7027
M5*	17.1277	-416.1335	1767	832.1872	832.2592	2.4332
Análise 5: Thin = 2; Burn-in = 1000; Iterações = 5000; Amostra = 2000						
Modelo	Q_p	LPML	DIC	WAIC1	WAIC2	σ_h^2
M1	217.5700	-519.6900	2539	1039.3700	1039.3900	119.7500
M1*	213.3300	-517.8500	2533	1035.6400	1035.7100	126.7100
M2	196.9900	-510.9700	2515	1021.5800	1021.9100	275.7000
M2*	191.6000	-509.3200	2515	1017.9700	1018.5900	229.4400
M4	195.2800	-510.2200	2513	1020.1600	1020.4200	227.0100
M4*	191.9700	-510.1700	2486	1019.9100	1020.3000	1395.9600
M5	194.1800	-511.0570	2279	1021.4930	1022.0060	3430.4050
M5*	35.4680	-426.5400	1442	852.2860	852.9200	59.6670

De acordo com os resultados da Tabela 4, podemos perceber que as análises parecem apresentar um comportamento similar entre todos os modelos. Em quase todos os casos, os modelos com *, ou seja, que levam em consideração cada região separadamente, apresentam índices para os critérios de avaliação e variância melhores que os modelos que consideram todas as região juntas.

Os valores que apresentaram índices bons para as análises foram destacados na Tabela 4. Consideramos bons os modelos que apresentam valores baixos para os critérios de seleção DIC, WAIC1, WAIC2, e Q_p e valores altos para o LPML. Além disso, os modelos com as menores variâncias se destacam positivamente. Dessa forma, temos interesse em encontrar os modelos que apresentam bons índices para os critérios de avaliação e boas variâncias.

Na maioria das análises os modelos $M1$ e $M1^*$ apresentam os piores índices, encontrando altos valores de Q_p , WAIC1 e WAIC2. Porém, apresentam altos valores de LPML e as variâncias para esses modelos estão baixas. Os modelos $M2$, $M4$, $M4^*$ e $M5$ apresentam valores medianos para os critérios de avaliação dos modelos e para a variância quando comparados aos demais modelos. As variâncias de $M5$ e $M4^*$ são as mais elevadas.

Os modelos $M2^*$ e $M5^*$ apresentam os melhores índices para quase todos os critérios. De acordo com essas análises, $M5^*$ parece ser o melhor modelo para ajustar os dados, apresentando baixas variâncias e bons índices para os critérios de avaliação.

Os gráficos dos estimadores para todas as regiões dos dez modelos analisados estão apresentados no Apêndice A. Para cada gráfico temos que a linha preta representa a série temporal do IQA da região e a linha vermelha o seu estimador.

Podemos concluir que apenas o modelo $M2^*$ apresentou bons resultados, como pode ser visto na Figura A.4. Seus estimadores acompanham bem as séries temporais de todas as 4 regiões do Rio dos Sinos, indicando uma boa estimação do modelo. Os modelos $M4$, $M4^*$, $M5$ que queríamos testar, não apresentaram os melhores resultados, indo contra nossa hipótese inicial.

Com os resultados das análises e observando os gráficos dos estimadores para os modelos, podemos concluir que o modelo $M2^*$ apresentou os melhores resultados. O modelo $M5^*$, apesar de apresentar bons valores para os critérios de avaliação e para a variância, não conseguiu estimar bem os dados para as séries temporais.

Com o objetivo de melhorar os resultados das análises, foi desenvolvido o modelo $M6$ e $M6^*$, dados na Seção 3.1, incluindo os efeitos aleatórios espacialmente correlacionados (b_i) nos modelos estudados. Para realizar as análises outro banco de dados será utilizado, descritos na Seção 5.

5 Dados do Rio Colorado nos Estados Unidos

Vamos analisar os modelos com o banco de dados dos 29 pontos de coleta no Rio Colorado, Estados Unidos, descritos na Seção 2.3.

5.1 Modelos para os dados do Rio Colorado

Começamos pelos modelos que apresentam apenas o efeito aleatório (h_i). O efeito correlacionado espacialmente (b_i) foi incluído na Seção 5.3.

Para analisar esses dados consideramos quatro modelos para explicar a componente de efeito temporal, $S_i(t)$ quando utilizamos os dados do Rio Colorado. A Tabela 5 abaixo fornece esses modelos de interesse.

Tabela 5: Quatro diferentes modelos que descrevem o efeito temporal no processo de interesse.

Modelo	$S_i(t)$	Especificação do Modelo
$M2$	$A_1(t)$	$A_1(t) \sim \mathcal{N}(\phi_t + \rho(A_1(t-1) - \phi_{t-1}), \sigma_{\epsilon_t}^2)$
$M2^*$	$A_1(i, t)$	$A_1(i, t) \sim \mathcal{N}(\phi_t + \rho(A_1(i, t-1) - \phi_{t-1}), \sigma_{\epsilon_{i,t}}^2)$
$M6$	$A_5(t)$	$A_5(t) \sim \mathcal{N}(\phi_t + \rho(\phi A_5(t-1)) + \rho(A_5(t-12)) - \rho(\phi A_5(t-13)) - \rho(\phi_{t-1} - \phi_1))$
$M6^*$	$A_5(i, t)$	$A_5(i, t) \sim \mathcal{N}(\phi_t + \rho(\phi A_5(i, t-1)) + \rho(A_5(i, t-12)) - \rho(\phi A_5(i, t-13)) - \rho(\phi_{t-1} - \phi_1))$

M_j denota o modelo para todas as 29 regiões enquanto que M_j^* denota aqueles para a i -ésima região, onde $i \in \{1, \dots, 29\}$ e $j \in \{2, 6\}$.

Os modelos $M2$ e $M2^*$ possuem um processo auto-regressivo imposto na média da distribuição *a priori*. Já os modelos $M6$ e $M6^*$ consideram um processo auto-regressivo para uma série temporal com sazonalidade para a componente do efeito temporal $S_i(t)$. Ambos os modelos analisados para esse banco de dados apresentam um parâmetro ρ extra na sua estrutura auto-regressiva.

A diferença entre os modelos M_j e M_j^* , onde $j \in \{2, 6\}$, é que para aqueles sem asterisco os modelos são considerados para todas as 29 regiões, enquanto os modelos com asterisco são considerados diferentemente para cada i -ésima região, onde $i \in \{1, \dots, 29\}$.

Concluimos com os dados do Rio dos Sinos, analisados na Seção 4.2, que os modelos com asterisco apresentam melhores resultados em comparação com os modelos sem asterisco. Portanto, focamos as análises seguintes nos modelos que são considerados diferentemente para cada região.

5.2 Análise dos modelos

Primeiramente analisamos os dados para esse modelo considerando um banco de dados único para todas as regiões. Percebemos que em alguns pontos de coleta os estimadores obtiveram um ajuste melhor e que isso poderia estar acontecendo devido a influência das outras regiões no modelo. Por este motivo, comparamos os modelos para as 29 região do Rio Colorado divididas nos três grupos descritos na Seção 2.3. Para isso, geramos os gráficos dos dados com os estimadores para os modelos $M2^*$ e $M6^*$. A linha preta representa a série temporal da região e a linha vermelha o estimador. O Apêndice B contém as figuras desses gráficos para o modelo $M2^*$.

Podemos perceber, analisando os gráficos do Apêndice B, que o modelo $M2^*$ não estima bem os pontos altos e baixos das séries temporais. O Grupo 1 apresentou os melhores estimadores para esse modelo e os estimadores para o Grupo 3 foram os mais distantes dos dados reais.

Para as regiões do banco de dados podemos notar dois comportamentos diferentes: o primeiro com apenas picos altos e o segundo com picos positivos e negativos.

Seguem abaixo os gráficos dos estimadores para os três grupos para o modelo $M6^*$, que apresentou resultados melhores.

Grupo 1:

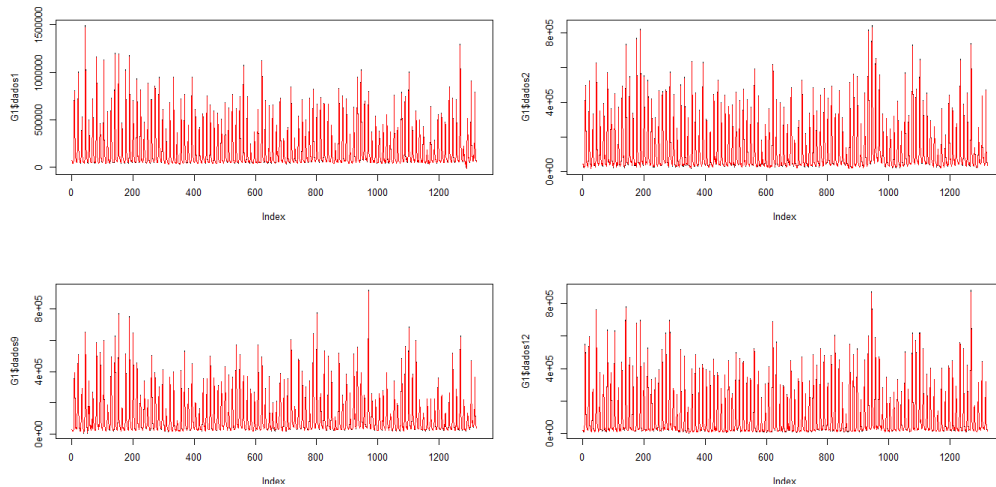


Figura 5: Gráfico dos dados com os estimadores para as regiões do Grupo 1 - modelo $M6^*$.

Podemos perceber pela Figura 5 que para as 4 regiões do Grupo 1 os estimadores acompanham bem os dados. Percebemos um comportamento parecido nessas 4 séries temporais.

Grupo 2:

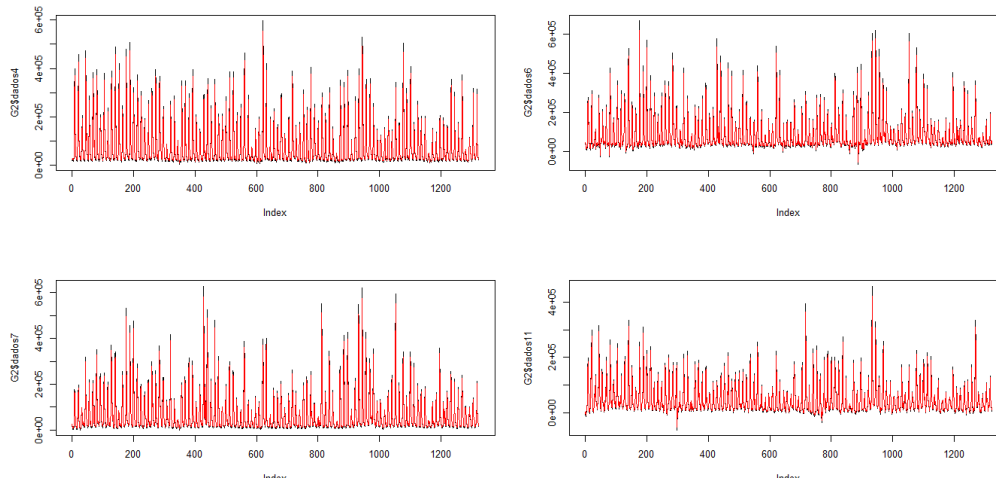


Figura 6: Gráfico dos dados com os estimadores para as regiões 4, 6, 7 e 11 do Grupo 2 - modelo $M6^*$.

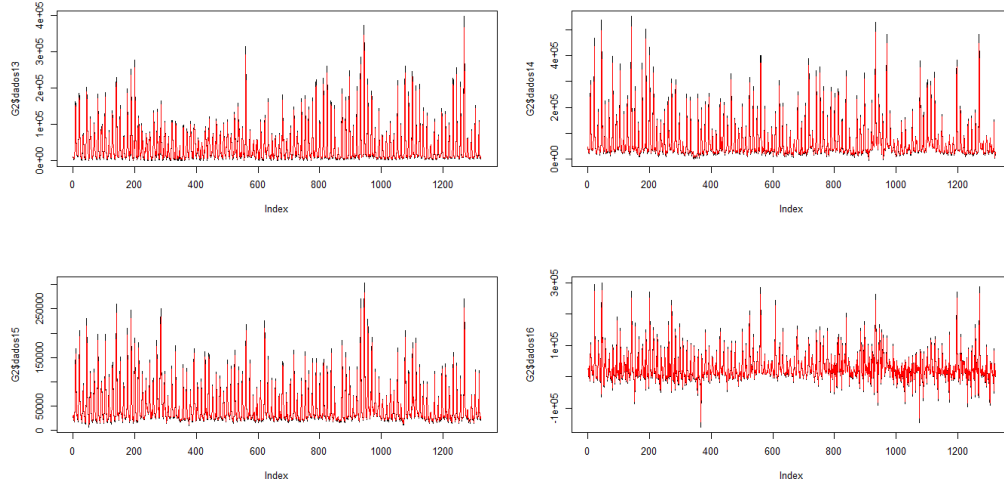


Figura 7: Gráfico dos dados com os estimadores para as regiões 13, 14, 15 e 16 do Grupo 2 - modelo M6*.

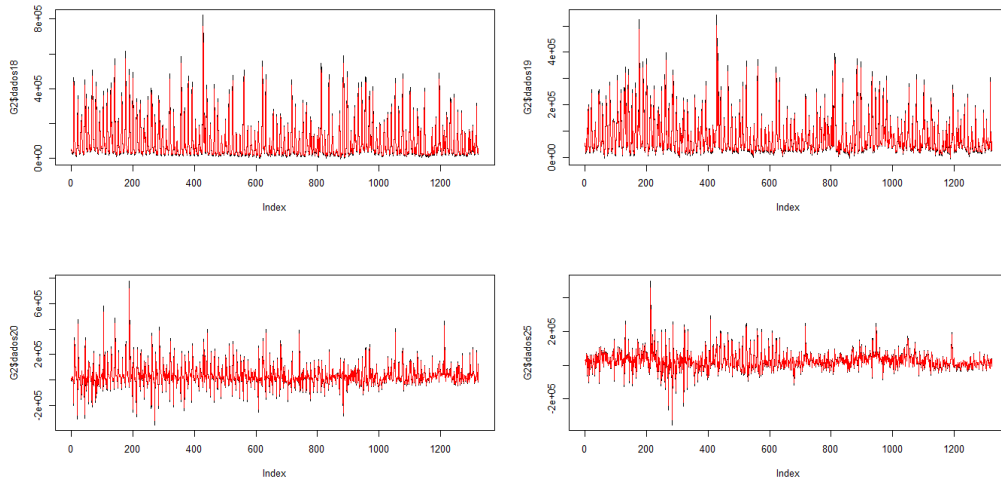


Figura 8: Gráfico dos dados com os estimadores para as regiões 18, 19, 20 e 25 do Grupo 2 - modelo M6*.

Para o Grupo 2 podemos notar que os estimadores acompanham bem os dados. Essas 12 séries temporais apresentam dois comportamentos diferentes, o primeiro muito semelhante as regiões do Grupo 1. O segundo tem um comportamento com picos altos e baixos, apresentando também valores negativos.

Grupo 3:

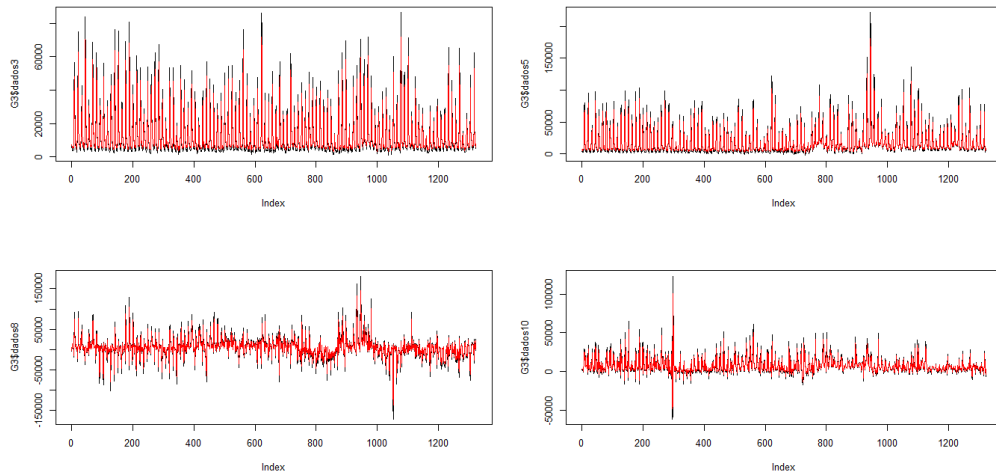


Figura 9: Gráfico dos dados com os estimadores para as regiões 3, 5, 8 e 10 do Grupo 3 - modelo M6*.

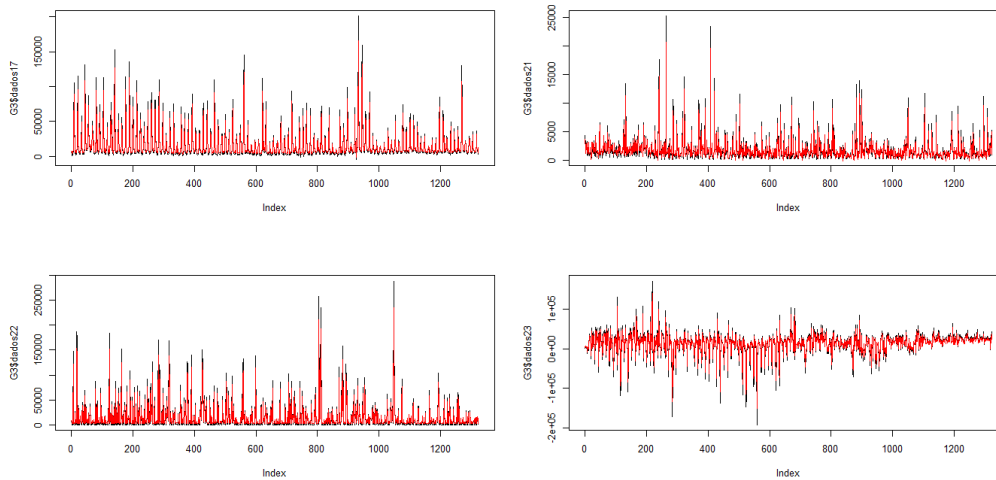


Figura 10: Gráfico dos dados com os estimadores para as regiões 17, 21, 22, 23 do Grupo 3 - modelo M6*.

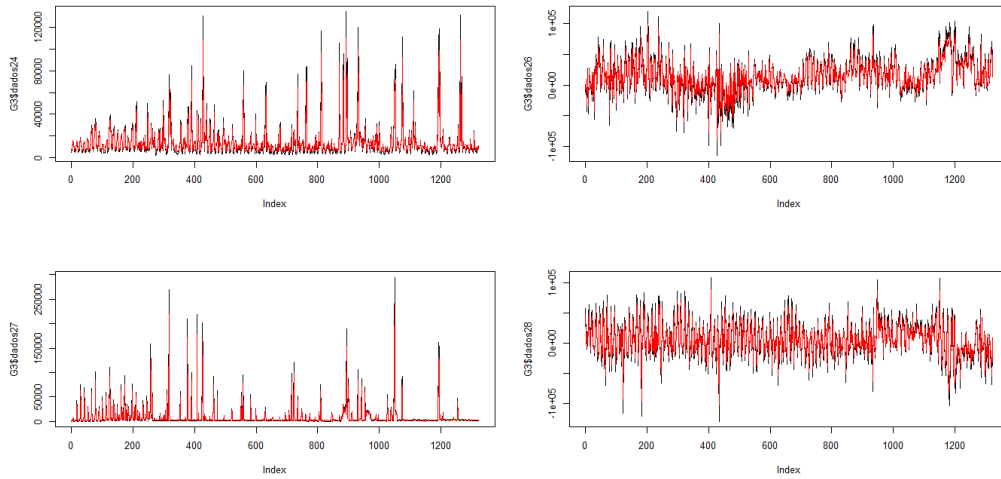


Figura 11: Gráfico dos dados com os estimadores para as regiões 24, 26, 27 e 28 do Grupo 3 - modelo $M6^*$.

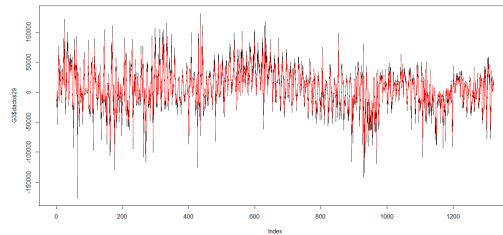


Figura 12: Gráfico dos dados com os estimadores para a região 29 do Grupo 3 - modelo $M6^*$.

Para o Grupo 3 podemos notar que os estimadores não acompanham os dados tão bem como no Grupo 1. As 13 séries temporais apresentam os mesmos comportamentos, o primeiro com apenas picos altos e o segundo com picos altos e baixos, apresentando também valores negativos. Podemos concluir com os gráficos dos 3 grupos que os estimadores acompanham melhor os dados no Grupo 1 e não estimam bem os pontos altos para o Grupo 3.

Comparando os gráficos dos dados com os estimadores para os dois modelos, podemos concluir que o $M6^*$ apresentou resultados melhores. Os estimadores acompanham bem as séries temporais, mas em algumas regiões do Grupo 3 ainda não detectam os picos altos e baixos.

Com o objetivo de melhorar os resultados das análises, incluímos os efeitos aleatórios espacialmente correlacionados (b_i) na Seção 5.3 para os modelos estudados.

5.3 Análise espaço-temporal

Vamos agora analisar esses mesmos dados levando em consideração a técnica espacial descrita na Seção 3.2. Consideramos um modelo que analise as mudanças de cada grupo em relação ao espaço e ao tempo. Para isso criamos uma matriz de vizinhos levando em consideração a proximidade dos pontos de coleta em relação ao curso do rio entre as regiões.

Para selecionar os vizinhos de cada região analisamos o curso do rio. Os pontos diretamente conectados foram considerados como vizinhos, mesmo apresentando distâncias diferentes. Dessa forma, conseguimos analisar a influência das regiões que o curso da água passou imediatamente antes ou imediatamente depois de cada ponto de coleta. Por exemplo, observando o mapa da Figura 3 da Seção 2.3, podemos perceber que a região 2 é a única conectada com a região 1. Contudo, observando o curso do rio antes e depois da região 28, temos que os pontos 29 e 27 estão conectados.

Tabela 6: Matriz de vizinhos e a quantidade de vizinhos para cada região.

Região	Vizinhos	Nº Vizinhos	Região	Vizinhos	Nº Vizinhos
1	2	1	16	11,12,13,14,15,17	6
2	1,6,8	3	17	7,8,16,19,20	5
3	4	1	18	4,5,19	3
4	3,5,18,19	4	19	4,5,7,8,17,18,20	7
5	4,6,18,19	4	20	7,8,17,19,21	5
6	2,5,8	3	21	20,22,23	3
7	8,17,19,20	4	22	21,23	2
8	2,6,7,17,19,20	6	23	21,22,24,25	4
9	10	1	24	23,25	2
10	9,11	2	25	23,24,26	3
11	10,12,13,14,15,16	6	26	25,27	2
12	11,13,14,15,16	5	27	26,28	2
13	11,12,14,15,16	5	28	27,29	2
14	11,12,13,15,16	5	29	28	1
15	11,12,13,14,16	5			

De acordo com a matriz de vizinhos percebemos que algumas regiões estão mais distantes geograficamente que outras. Considerando esses dados espaciais é possível analisar se existe algum padrão na vazão da água de acordo com as proximidades das regiões. Para isso, um modelo que utiliza a matriz de vizinhos e que considere tanto o comportamento temporal como o espacial foi utilizado.

O comportamento temporal foi definido através dos modelos da Tabela 5. Já o comportamento espacial foi definido através do modelo Intrínseco (ICAR) definido na Seção 3.2. Dessa forma, vamos realizar uma análise espaço-temporal das 29 regiões.

Para analisar e comparar a eficiência dos modelos que utilizam a matriz de vizinhos e que consideram o efeito aleatório (h_i) e o efeito correlacionado espacialmente (b_i) realizamos

diferentes análises de acordo com a metodologia descrita na Seção 3.3. O Deviance Information Criterion (DIC), descrito na Seção 3.4, foi o critério considerado para essa comparação. Além disso, geramos os gráficos dos estimadores para todas as regiões dos modelos $M2^*$ e $M6^*$ propostos na Tabela 5, ordenada de acordo com o tamanho amostral da análise.

As análises foram realizadas de acordo com a descrição em 3.3. Primeiramente definimos valores para o Thin, Burn-in e Iterações diferentes. O tamanho amostral é calculado a partir das informações dos parâmetros anteriores. Esses valores definem onde vamos iniciar as análises, de quanto em quanto vamos analisar as observações e qual o total que vamos observar. O tempo para realizar essas análises foi significativamente maior comparado com as análises da Seção 4.2. Por esse motivo, selecionamos parâmetros menores e conseqüentemente um tamanho amostral menor.

O sistema operacional utilizado para as análises é o Windows 10, o software R versão 4.1.3 e o programa OpenBUGS versão 3.2.3. O tempo das análises para este banco de dados depende dos parâmetros definidos, quanto maiores os parâmetros selecionados, maior o tempo computacional. Para valores altos, as análises podem demorar até 2 ou 3 dias.

A Tabela 7 descreve os resultados de algumas análises, indicando inicialmente os parâmetros utilizados. Além do critério de avaliação DIC, a Tabela 7 apresenta os valores para as variâncias. Podemos perceber que o tamanho amostral sozinho não parece ter grande influência nos índices calculados. Valores para o Burn-in e Thin alto podem apresentar melhoras leves para o critério de avaliação DIC.

Tabela 7: Resultados das Análises para os dados do Rio Colorado

Análise 1: Thin = 2; Burn-in = 20; Iterações = 220; Amostra = 100		
Modelo	DIC	σ_h^2
M2*	165480.00	1.12061e-08
M6*	163999.50	1.34225e-02
Análise 2: Thin = 4; Burn-in = 40; Iterações = 440; Amostra = 100		
Modelo	DIC	σ_h^2
M2*	165478.88	1.13264e-08
M6*	163991.50	1.35331e-02
Análise 3: Thin = 5; Burn-in = 50; Iterações = 800; Amostra = 150		
Modelo	DIC	σ_h^2
M2*	165477.00	1.10570e-08
M6*	163992.93	1.29244e-02
Análise 4: Thin = 5; Burn-in = 50; Iterações = 1050; Amostra = 200		
Modelo	DIC	σ_h^2
M2*	165477.90	1.10826e-08
M6*	163993.20	1.30355e-02
Análise 5: Thin = 5; Burn-in = 10; Iterações = 1510; Amostra = 300		
Modelo	DIC	σ_h^2
M2*	165478.30	1.11746e-08
M6*	163992.87	1.32728e-02
Análise 6: Thin = 2; Burn-in = 10; Iterações = 1010; Amostra = 500		
Modelo	DIC	σ_h^2
M2*	165480.80	1.11934e-08
M6*	163997.50	1.35114e-02
Análise 7: Thin = 2; Burn-in = 10; Iterações = 1510; Amostra = 750		
Modelo	DIC	σ_h^2
M2*	165479.20	1.11312e-08
M6*	163996.67	1.32495e-02
Análise 8: Thin = 1; Burn-in = 100; Iterações = 1100; Amostra = 1000		
Modelo	DIC	σ_h^2
M2*	165479.40	1.11985e-08
M6*	163992.10	1.31833e-02

Os valores que apresentaram índices bons para as análises foram destacados na Tabela 7. Consideramos bons os modelos que apresentam valores baixos para o critério de seleção DIC e para as variâncias.

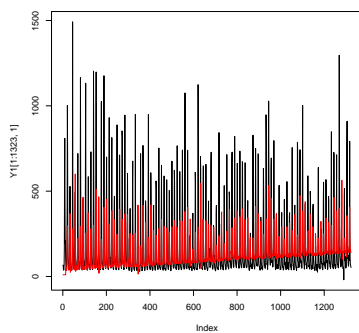
De acordo com os resultados da Tabela 7, podemos perceber que para todas as análises o modelo $M2^*$ apresentou valores para a variância bem inferiores ao modelo $M6^*$. Contudo, o modelo $M6^*$ apresentou índices do DIC inferiores ao modelo $M2^*$.

Os gráficos dos estimadores para todas as regiões para os dados do Rio Colorado do modelo $M2^*$ estão no Apêndice C. A linha preta em cada gráfico refere-se à série temporal e a linha vermelha ao estimador do modelo para a região.

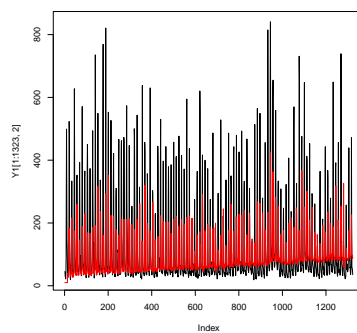
Podemos concluir que os gráficos dos estimadores não acompanham bem os dados do Rio Colorado para o modelo $M2^*$, não detectando pontos altos e baixos. Em alguns casos, como na região 3 na Figura C.2, os estimadores apresentam uma tendência inexistente na série

temporal.

Seguem abaixo os gráficos dos estimadores para todas as regiões para os dados do Rio Colorado do modelo $M6^*$. A linha preta em cada gráfico refere-se à série temporal e a linha vermelha ao estimador do modelo para a respectiva região.

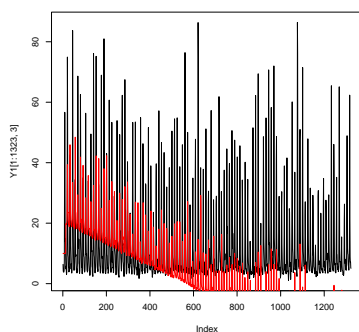


(a) Região 1.

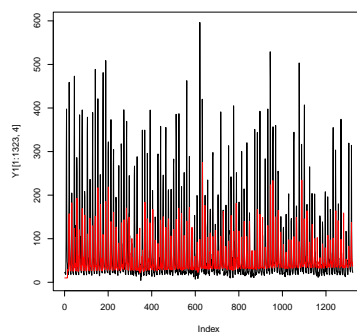


(b) Região 2.

Figura 13: Gráfico dos estimadores do modelo $M6^*$.

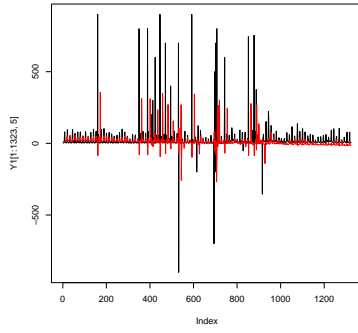


(a) Região 3.

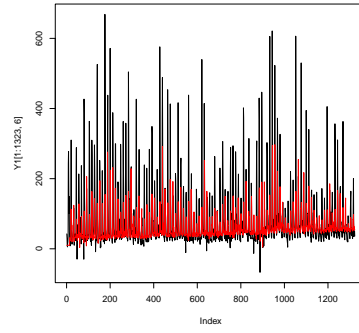


(b) Região 4.

Figura 14: Gráfico dos estimadores do modelo $M6^*$.

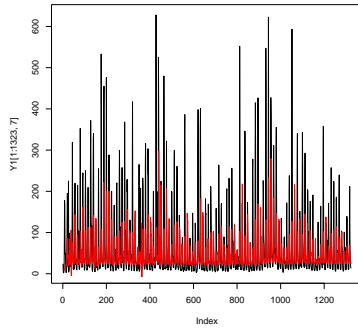


(a) Região 5.

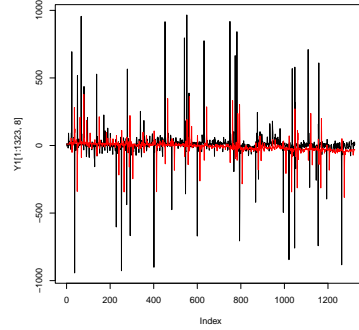


(b) Região 6.

Figura 15: Gráfico dos estimadores do modelo $M6^*$.

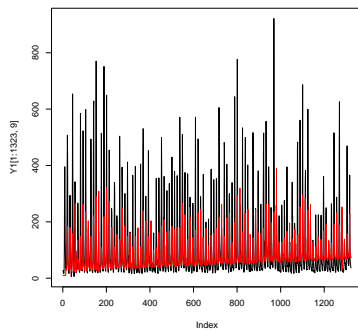


(a) Região 7.

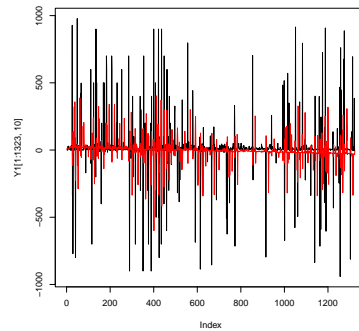


(b) Região 8.

Figura 16: Gráfico dos estimadores do modelo $M6^*$.

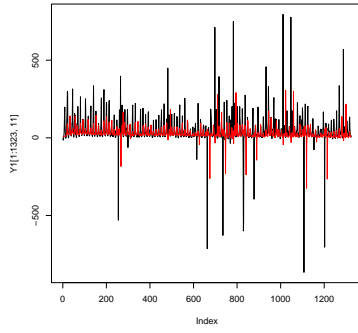


(a) Região 9.

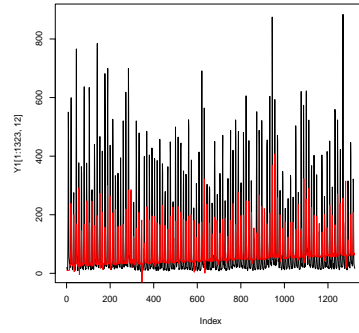


(b) Região 10.

Figura 17: Gráfico dos estimadores do modelo $M6^*$.

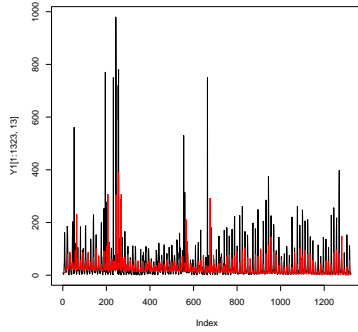


(a) Região 11.

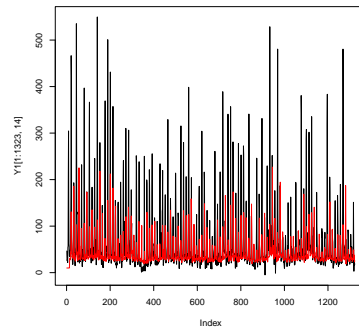


(b) Região 12.

Figura 18: Gráfico dos estimadores do modelo $M6^*$.

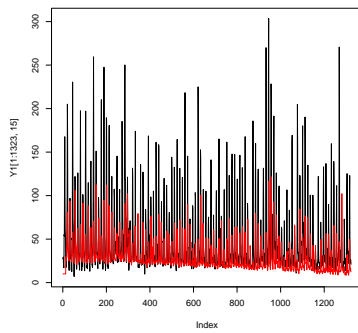


(a) Região 13.

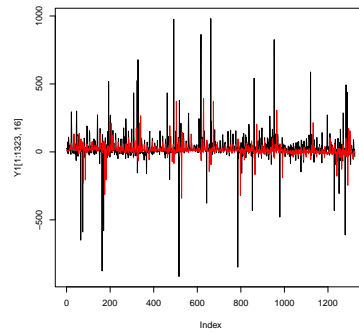


(b) Região 14.

Figura 19: Gráfico dos estimadores do modelo $M6^*$.

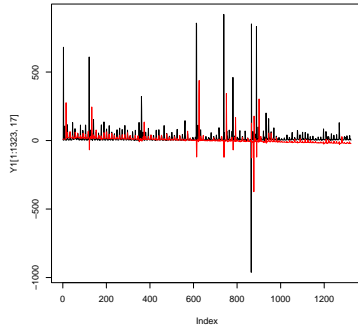


(a) Região 15.

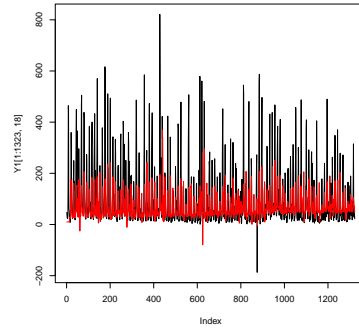


(b) Região 16.

Figura 20: Gráfico dos estimadores do modelo $M6^*$.

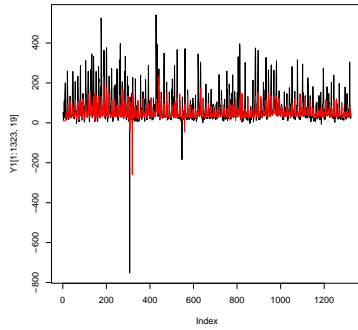


(a) Região 17.

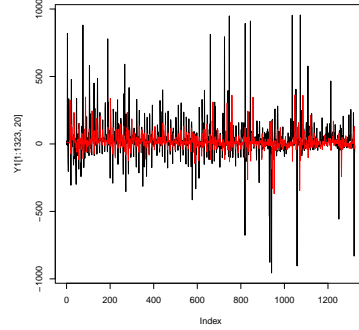


(b) Região 18.

Figura 21: Gráfico dos estimadores do modelo $M6^*$.

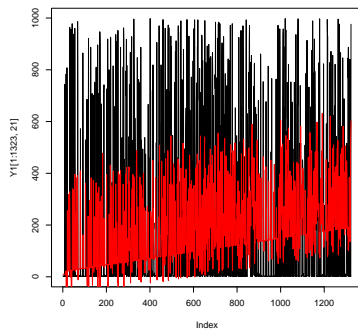


(a) Região 19.

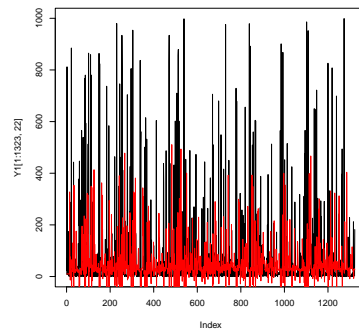


(b) Região 20.

Figura 22: Gráfico dos estimadores do modelo $M6^*$.

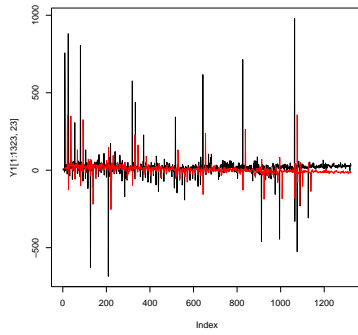


(a) Região 21.

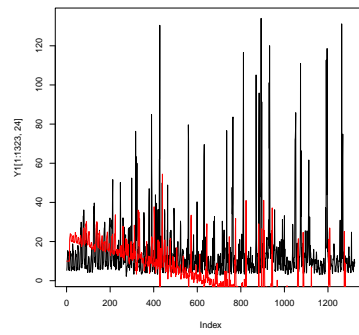


(b) Região 22.

Figura 23: Gráfico dos estimadores do modelo $M6^*$.

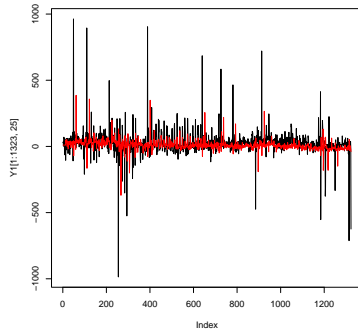


(a) Região 23.

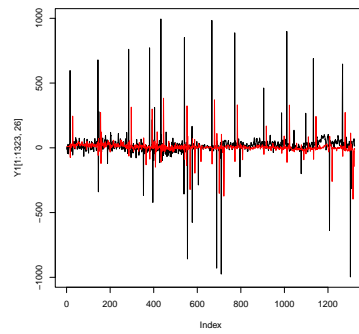


(b) Região 24.

Figura 24: Gráfico dos estimadores do modelo $M6^*$.

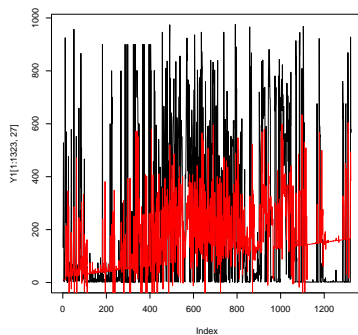


(a) Região 25.

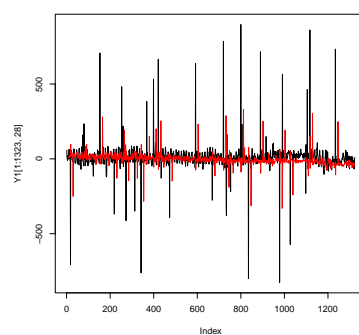


(b) Região 26.

Figura 25: Gráfico dos estimadores do modelo $M6^*$.



(a) Região 27.



(b) Região 28.

Figura 26: Gráfico dos estimadores do modelo $M6^*$.

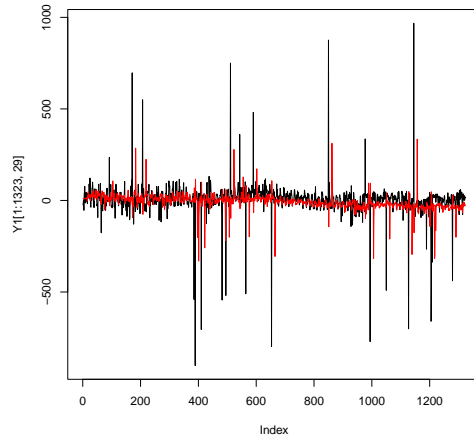


Figura 27: Região 29 - Gráfico dos dados com os estimadores do modelo $M6^*$.

Podemos concluir que os gráficos dos estimadores não acompanham bem os dados do Rio Colorado para o modelo $M6^*$ também, não detectando pontos altos e baixos. Em alguns casos, como na região 3 na Figura 14 e na região 24 na Figura 24, os estimadores apresentam uma tendência inexistente na série temporal. Comparando os gráficos dos dados das séries temporais com os estimadores dos dois modelos, conseguimos perceber uma leve melhora nas estimativas do modelo $M6^*$.

Analisando os gráficos dos dados com os estimadores dos modelos que levam em consideração apenas o efeito aleatório (h_i) percebemos que apresentaram detectaram melhor a série temporal comparado aos modelos que levam em consideração o efeito aleatório (h_i) e o efeito correlacionado espacialmente (b_i) juntos.

Para tentar melhorar os modelos que consideram o efeito aleatório (h_i) e o efeito correlacionado espacialmente (b_i) podemos analisar diferentes metodologias para o modelo espacial. O modelo espacial (CAR) aplicado neste trabalho foi o Intrínseco, considerado um modelo tradicional. Outros modelos espaciais que podem ser aplicados para o banco de dados são: modelo de Convolução; modelo de Leroux e modelo de LU.

6 Conclusão

O objetivo geral deste trabalho foi desenvolver e analisar modelos lineares generalizados para séries temporais. Utilizamos dois bancos de dados para comparar os modelos, sem e com sazonalidade. Iniciamos a análise com os dados do Rio dos Sinos e a partir das conclusões, passamos a comparar os modelos para os dados do Rio Colorado. No final, acrescentamos o efeito correlacionado espacialmente (b_i) para realizar uma análise espaço-temporal.

Ao analisar os dados do Rio dos Sinos para os modelos propostos, concluímos que o modelo $M2^*$, que considera um processo auto-regressivo de primeira ordem através do valor da média da distribuição *a priori* e o modelo $M5^*$, que considera um processo auto-regressivo aditivo de primeira ordem para a componente do efeito temporal $S_i(t)$ apresentaram os melhores resultados para as análises. Contudo, apenas o modelo $M2^*$ obteve gráficos dos dados com os estimadores para todas as regiões que acompanharam bem as séries temporais.

Sabendo disso, desenvolvemos o modelo $M6$ e $M6^*$ e utilizamos o segundo banco de dados, que apresenta sazonalidade, para a comparação dos modelos. Separamos os dados em grupos e analisamos os modelos separadamente, com o objetivo de melhorar os resultados. Foi possível concluir que o modelo $M6^*$, que considera um processo auto-regressivo para uma série temporal com sazonalidade para a componente do efeito temporal $S_i(t)$ apresentou os melhores resultados para os gráficos dos estimadores comparados com o modelo $M2^*$.

Com os resultados das duas comparações anteriores, incluímos o efeito correlacionado espacialmente nos modelos $M2^*$ e $M6^*$. Para isso, aplicamos a matriz de vizinhos e o modelo Espacial Intrínseco (ICAR). Ao acrescentar o efeito espacial, conseguimos notar uma piora nos gráficos dos seus respectivos estimadores.

Concluímos com esse trabalho que o modelo $M2^*$ apresentou os melhores resultados para os dados do Rio dos Sinos, ou seja, para os dados que não apresentam sazonalidade. Da mesma forma, o modelo $M6^*$ apresentou os melhores resultados para o banco do Rio Colorado, que apresenta sazonalidade. Além disso, podemos concluir que o modelo Espacial Intrínseco não mostrou melhoras para a estimação dos dados e seria interessante analisar outros modelos espaciais para comparar a sua eficiência, como por exemplo o modelo de Convolução; modelo de Leroux e modelo de LU.

Referências

- Besag, J., York, J. e Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, Vol. 43(1), 1-20.
- Congdon, P. (2007). *Bayesian statistical modelling*. Vol. 704. London: Wiley. John Wiley & Sons.
- Cressie, N. (2015). *Statistics for spatial data*. Vol. 2. Hoboken: Wiley. John Wiley & Sons.
- Gelman, A., Hwang, J. e Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, Vol. 24 (6), 997-1016.
- Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, Vol. 7, 473-511.
- Martínez-Beneito, M. A., López-Quilez, A. e Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, Vol. 27 (15), 2874-2889.
- Nelder, J. A. e Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Vol. 135 (3), 370-384.
- Paulino, C.D., Turkman, A.A. e Murteira, B. (2009). *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian.
- Rampaso, R. C., de Souza, A. D. P. e Flores, E. F. (2016). Bayesian analysis of spatial data using different variance and neighbourhood structures. *Journal of Statistical Computation and Simulation*, Vol. 86(3), 535-552.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. e Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 64(4), 583-639.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, Vol. 22, 1701-1762.
- Torabi, M. e Rosychuk, R. J. (2010). Spatio-temporal modelling of disease mapping of rates. *Canadian Journal of Statistics*, Vol. 38 (4), 698-715.
- Vehtari, A., Gelman, A. e Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, Vol. 27(5), 1413-1432.
- Waller, L. A., Carlin, B. P., Xia, H. e Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, Vol. 92 (438), 607-617.
- Watanabe, S. (2013a). Waic and wbic are information criteria for singular statistical model evaluation. In *Proceedings of the Workshop on Information Theoretic Methods in Science and Engineering*, 90-94.
- Watanabe, S. (2013b). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, Vol. 14(3), 867-897.

Apêndice A

As figuras neste Apêndice A representam os gráficos dos dados do Rio dos Sinos e seus estimadores para todos os modelos e todas as regiões. A linha preta em cada gráfico é a série temporal e a linha vermelha o estimador do modelo para a região.

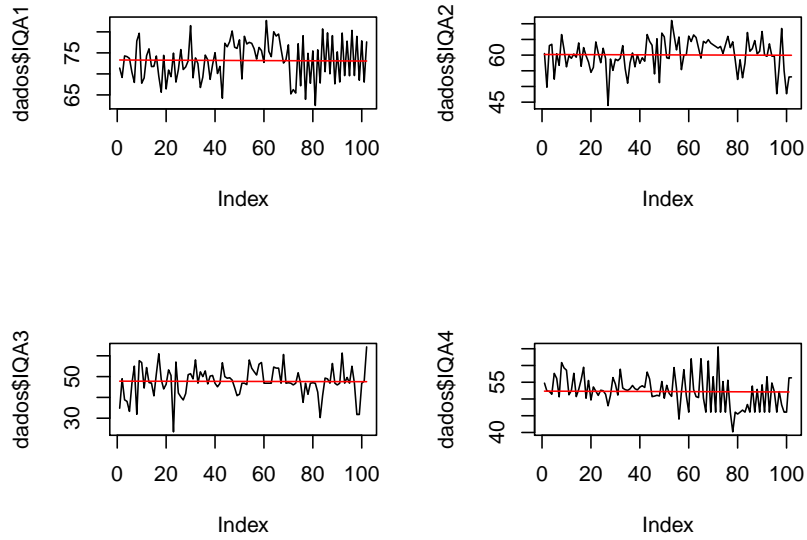


Figura A.1: Gráficos dos dados com os estimadores - $M1$.

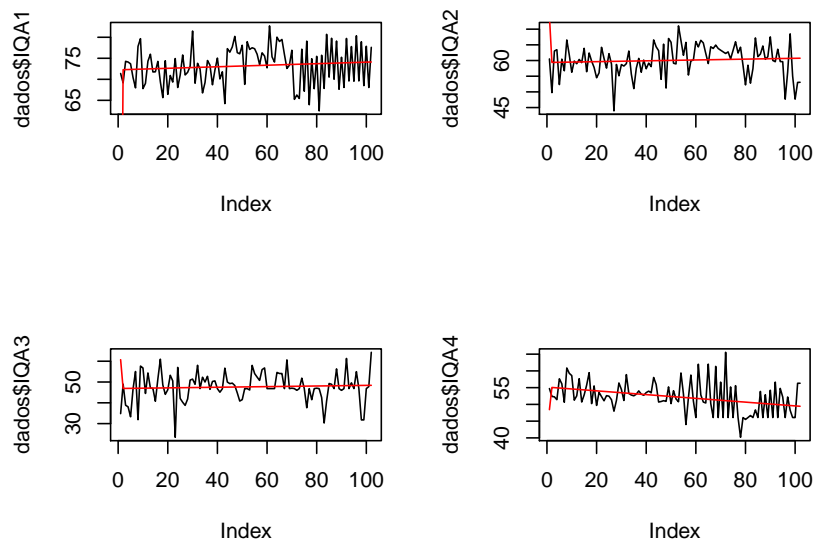


Figura A.2: Gráficos dos dados com os estimadores - $M1^*$.

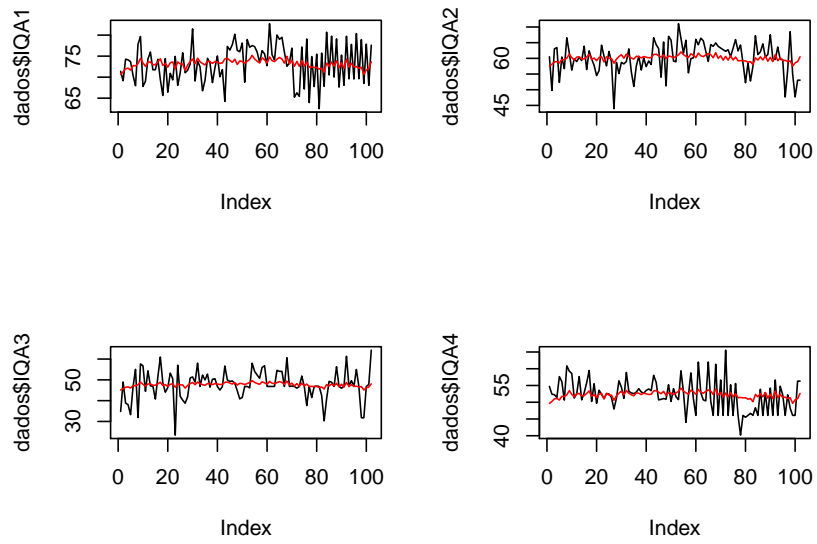


Figura A.3: Gráficos dos dados com os estimadores - $M2$.

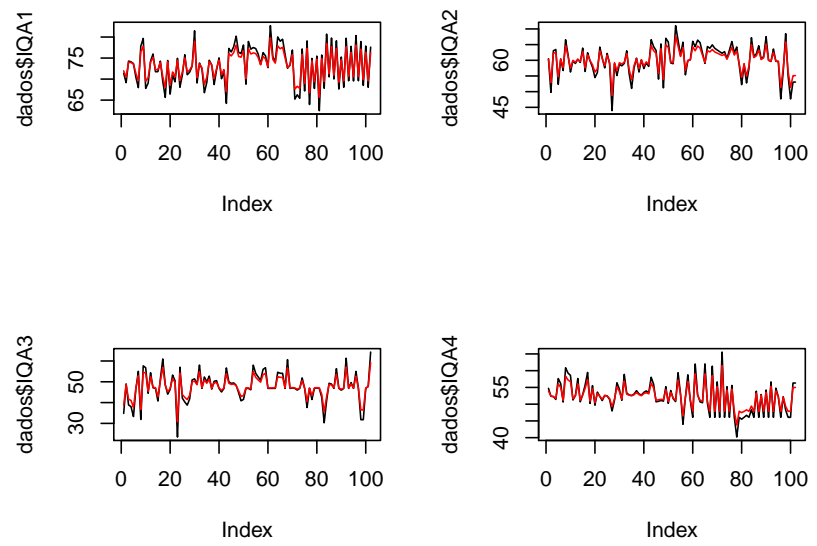


Figura A.4: Gráficos dos dados com os estimadores - $M2^*$.

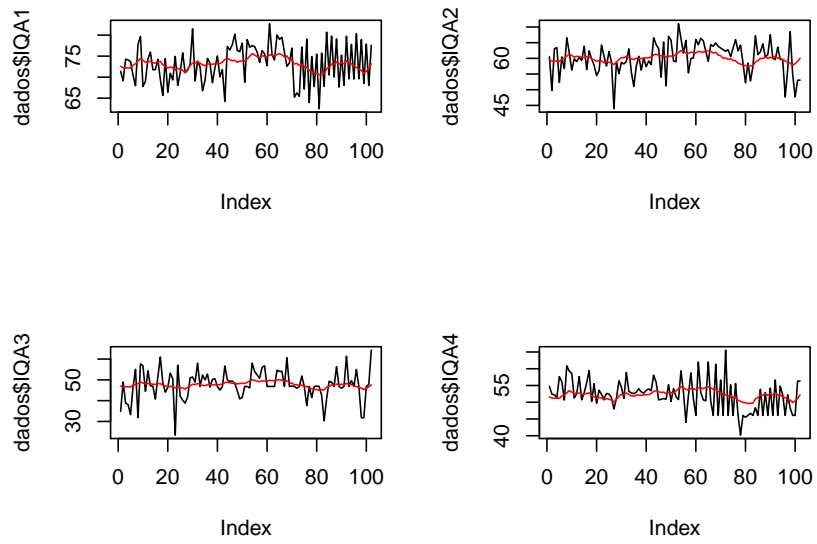


Figura A.5: Gráficos dos dados com os estimadores - $M3$.

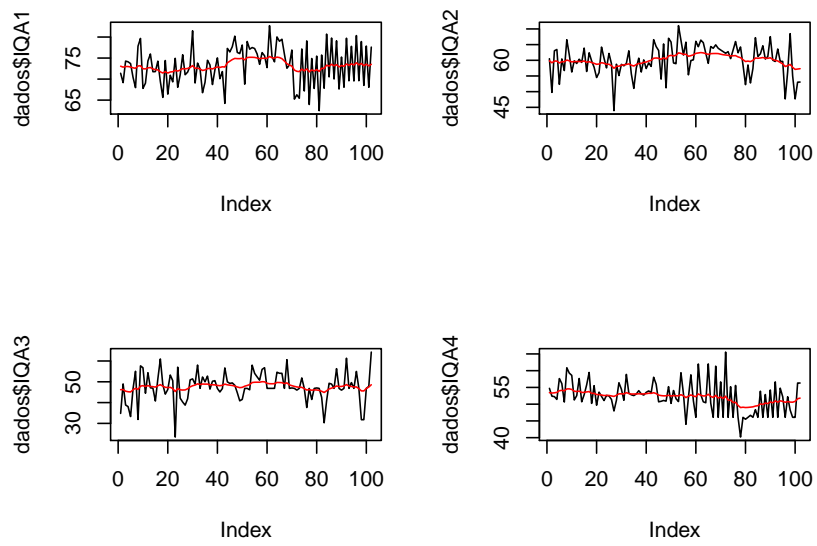


Figura A.6: Gráficos dos dados com os estimadores - $M3^*$.

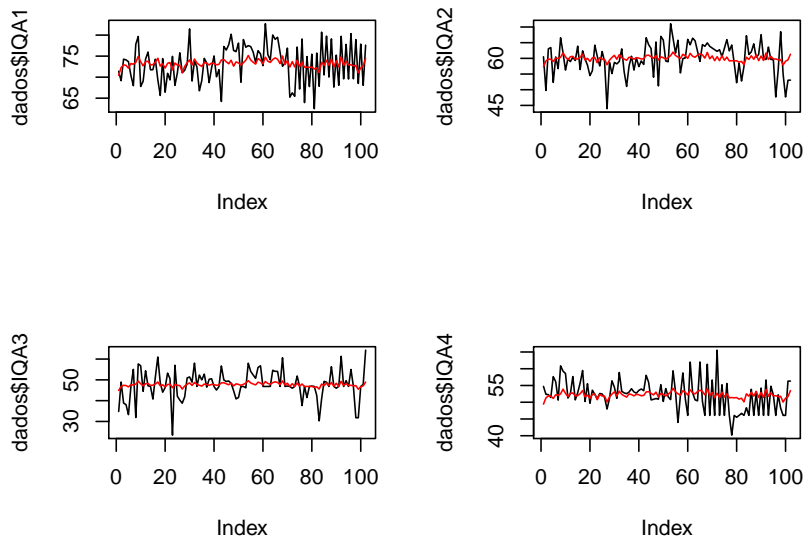


Figura A.7: Gráficos dos dados com os estimadores - $M4$.

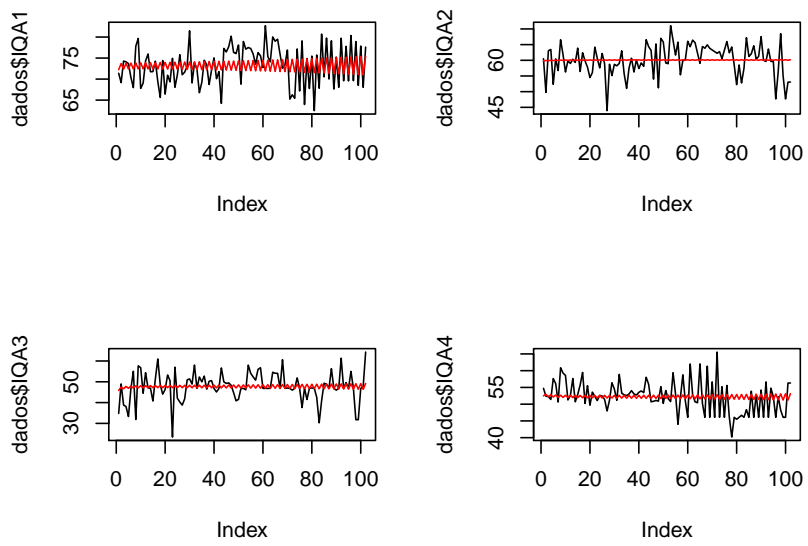


Figura A.8: Gráficos dos dados com os estimadores - $M4^*$.

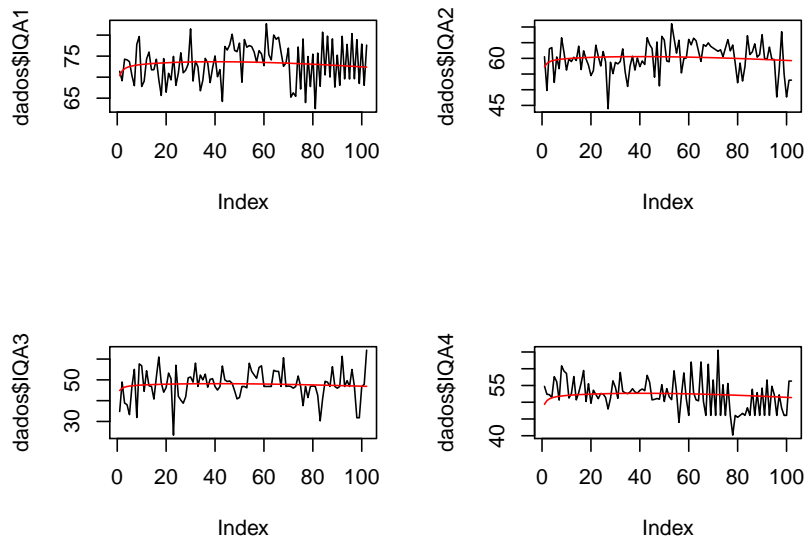


Figura A.9: Gráficos dos dados com os estimadores - $M5$.

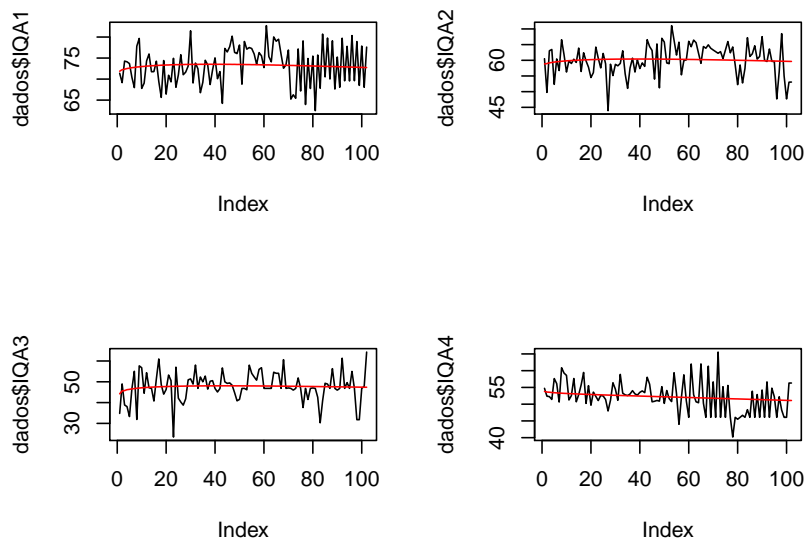


Figura A.10: Gráficos dos dados com os estimadores - $M5^*$.

Apêndice B

As figuras neste Apêndice B representam os gráficos dos dados do Rio Colorado e seus estimadores para o modelo $M2^*$ em todas as regiões divididas nos três grupos. A linha preta em cada gráfico refere-se à série temporal e a linha vermelha ao estimador do modelo para a região.

Seguem abaixo os gráficos dos estimadores para os três grupos para o modelo $M2^*$

Grupo 1:

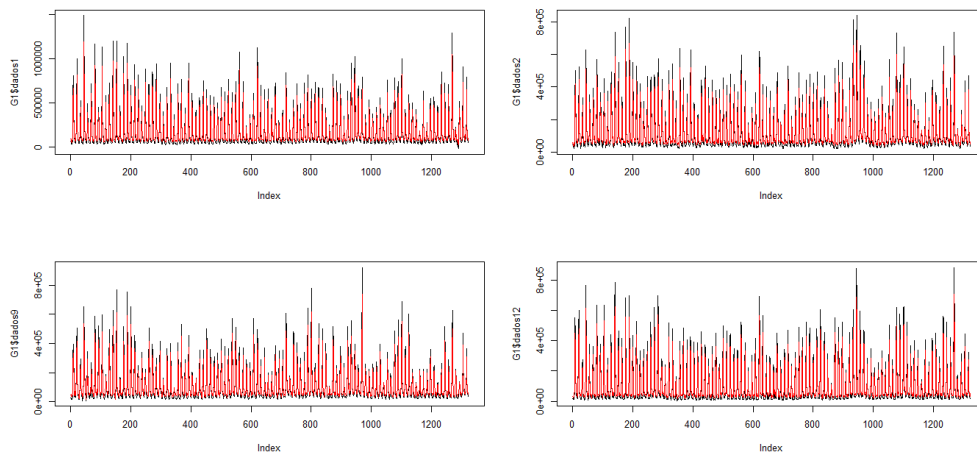


Figura B.1: Gráfico dos dados com os estimadores para as regiões do Grupo 1 - modelo $M2^*$.

Grupo 2:

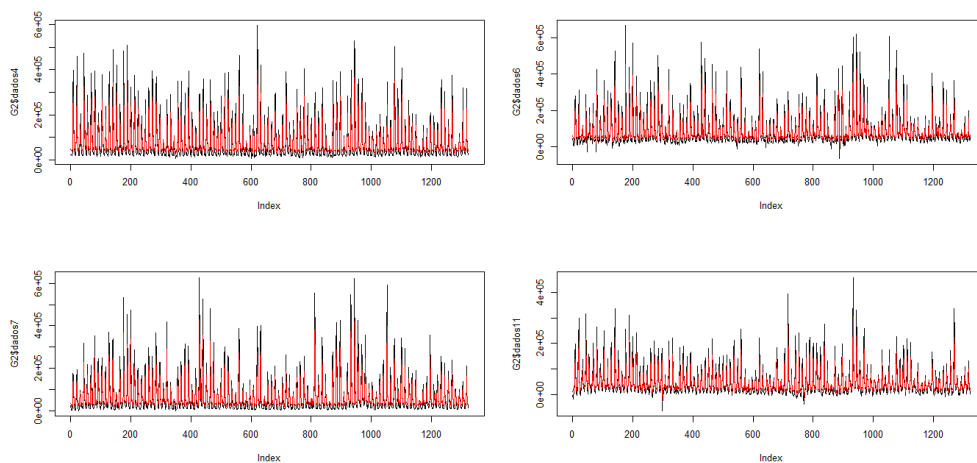


Figura B.2: Gráfico dos dados com os estimadores para as regiões 4, 6, 7 e 11 do Grupo 2 - modelo $M2^*$.

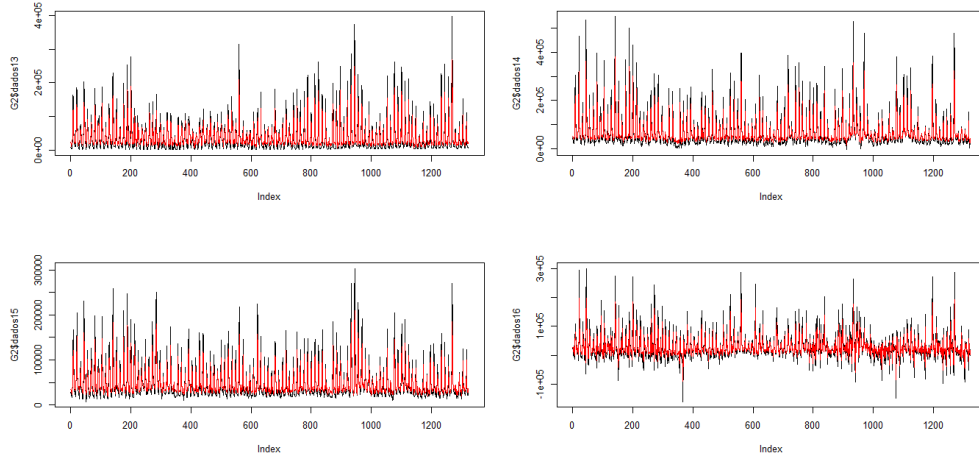


Figura B.3: Gráfico dos dados com os estimadores para as regiões 13, 14, 15 e 16 do Grupo 2 - modelo M2*.

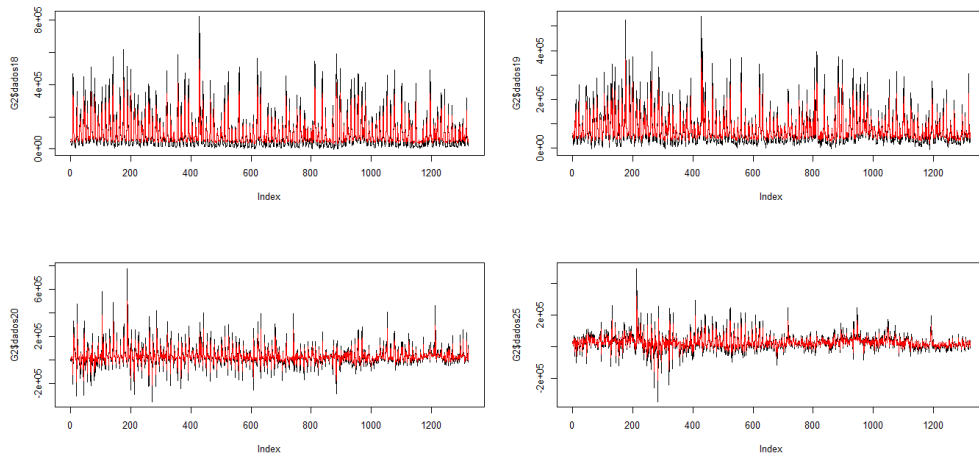


Figura B.4: Gráfico dos dados com os estimadores para as regiões 18, 19, 20 e 25 do Grupo 2 - modelo M2*.

Grupo 3:

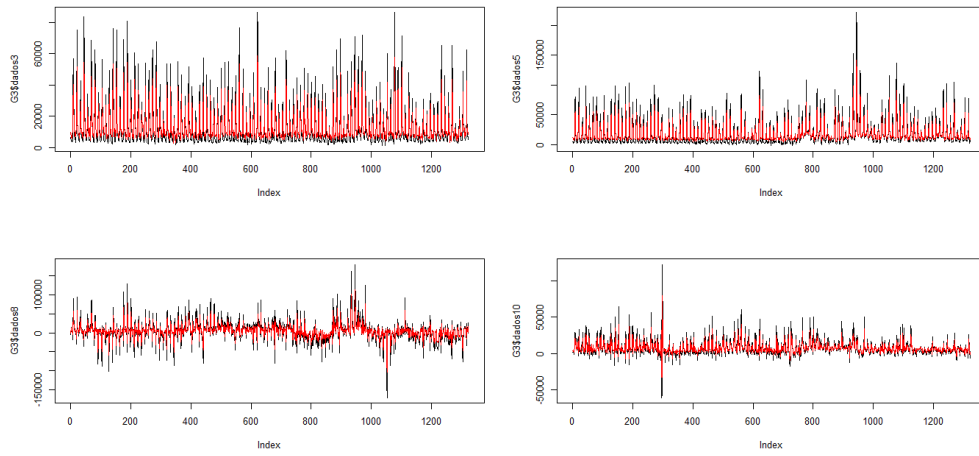


Figura B.5: Gráfico dos dados com os estimadores para as regiões 3, 5, 8 e 10 do Grupo 3 - modelo M2*.

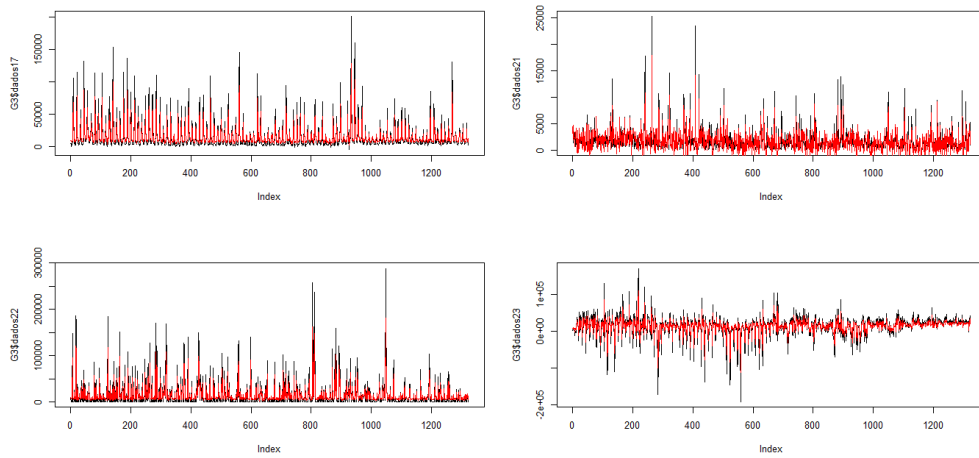


Figura B.6: Gráfico dos dados com os estimadores para as regiões 17, 21, 22, 23 do Grupo 3 - modelo M2*.

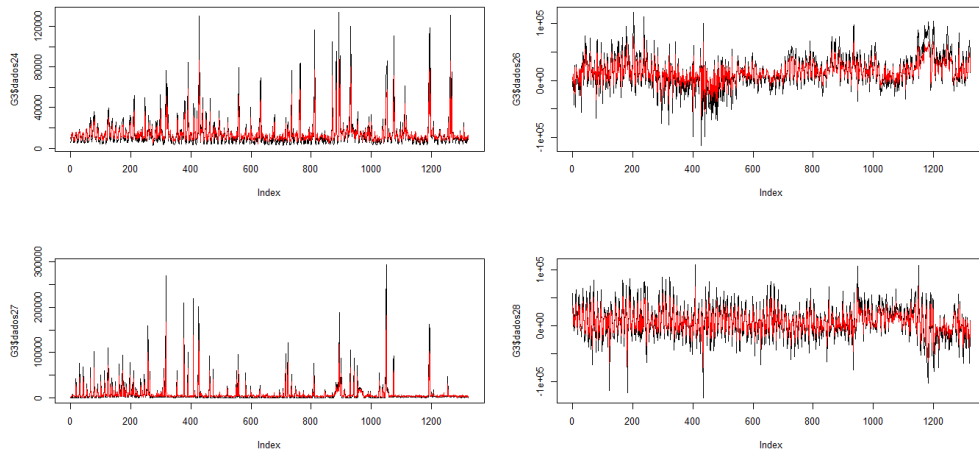


Figura B.7: Gráfico dos dados com os estimadores para as regiões 24, 26, 27 e 28 do Grupo 3 - modelo M2*.

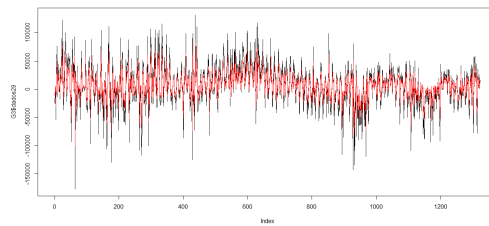
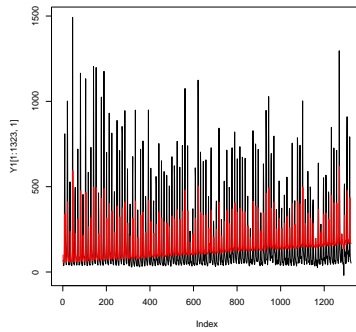


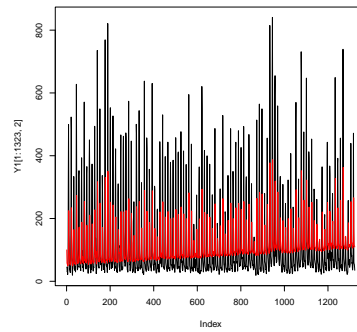
Figura B.8: Gráfico dos dados com os estimadores para a região 29 do Grupo 3 - modelo M2*.

Apêndice C

As figuras neste Apêndice C representam os gráficos dos dados do Rio Colorado e seus estimadores para o modelo $M2^*$ em todas as regiões. A linha preta em cada gráfico refere-se à série temporal e a linha vermelha ao estimador do modelo para a região.

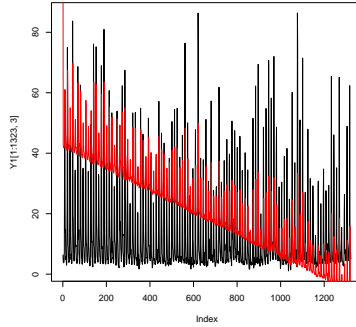


(a) Região 1.

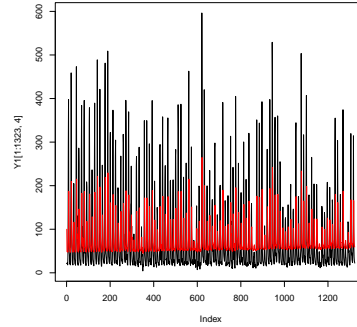


(b) Região 2.

Figura C.1: Gráfico dos estimadores do modelo $M2^*$.

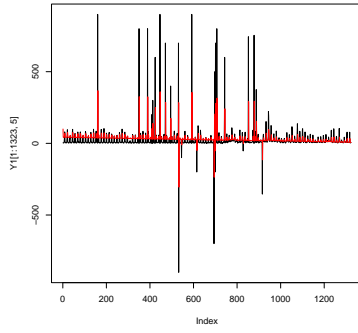


(a) Região 3.

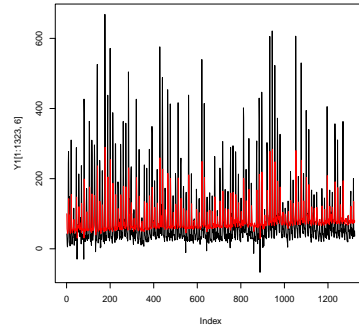


(b) Região 4.

Figura C.2: Gráfico dos estimadores do modelo $M2^*$.

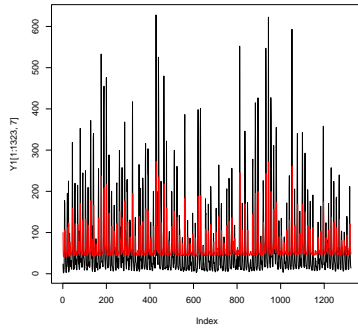


(a) Região 5.

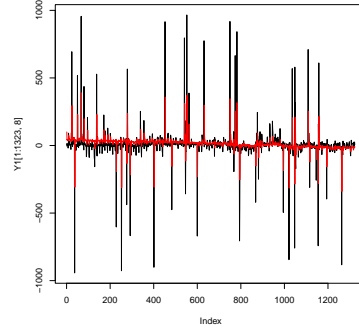


(b) Região 6.

Figura C.3: Gráfico dos estimadores do modelo $M2^*$.

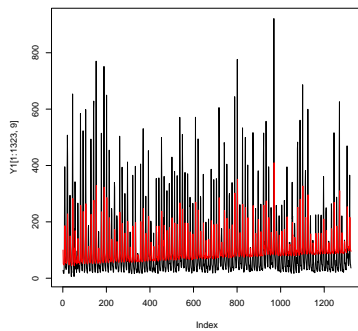


(a) Região 7.

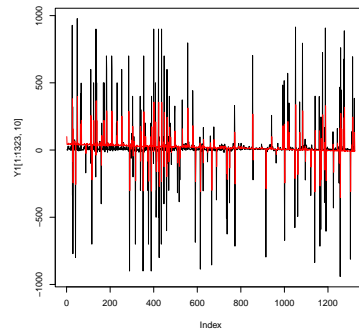


(b) Região 8.

Figura C.4: Gráfico dos estimadores do modelo $M2^*$.

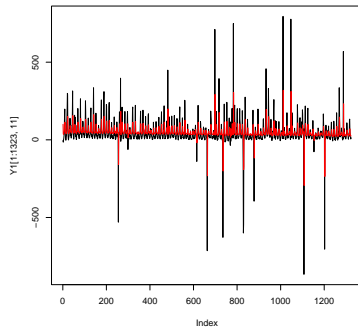


(a) Região 9.

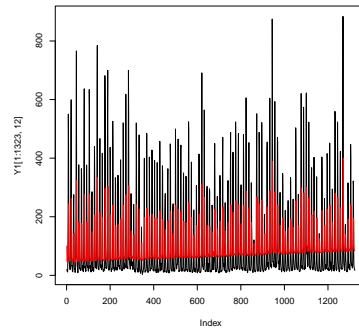


(b) Região 10.

Figura C.5: Gráfico dos estimadores do modelo $M2^*$.

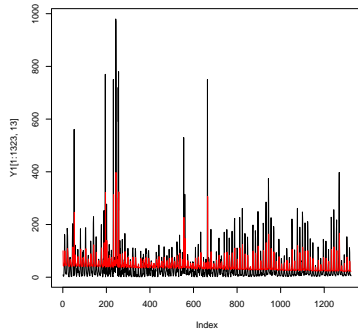


(a) Região 11.

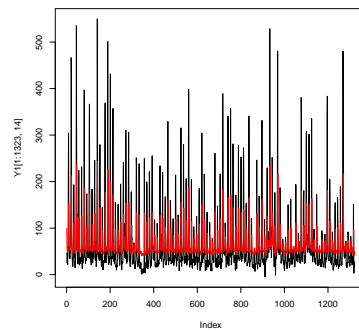


(b) Região 12.

Figura C.6: Gráfico dos estimadores do modelo $M2^*$.

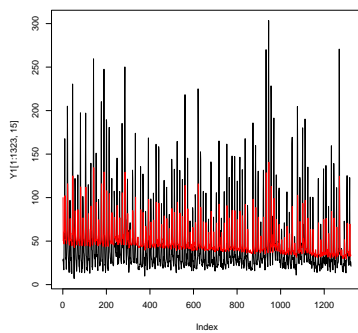


(a) Região 13.

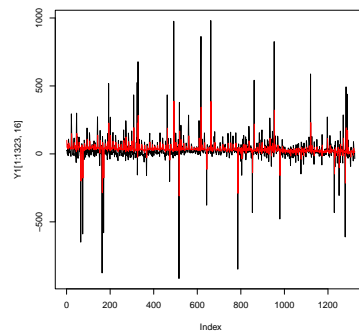


(b) Região 14.

Figura C.7: Gráfico dos estimadores do modelo $M2^*$.

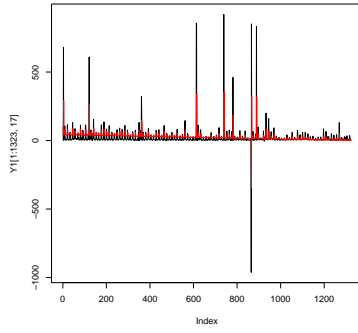


(a) Região 15.

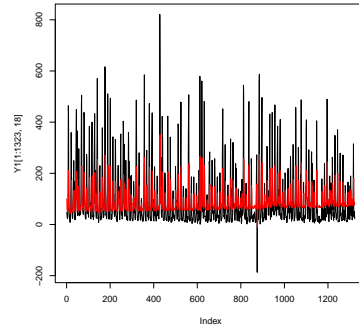


(b) Região 16.

Figura C.8: Gráfico dos estimadores do modelo $M2^*$.

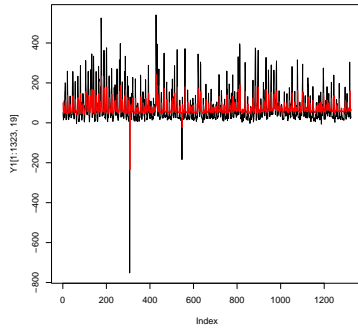


(a) Região 17.

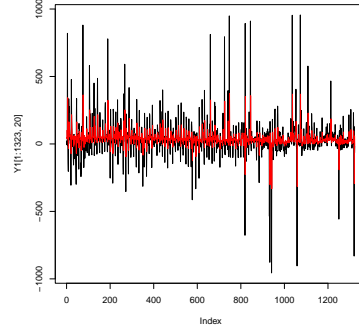


(b) Região 18.

Figura C.9: Gráfico dos estimadores do modelo $M2^*$.

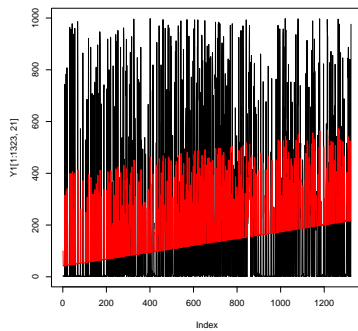


(a) Região 19.

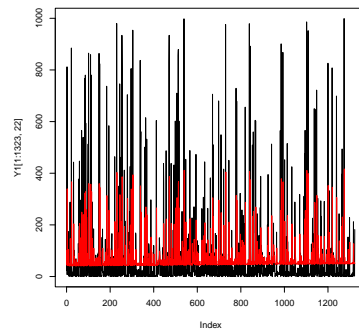


(b) Região 20.

Figura C.10: Gráfico dos estimadores do modelo $M2^*$.

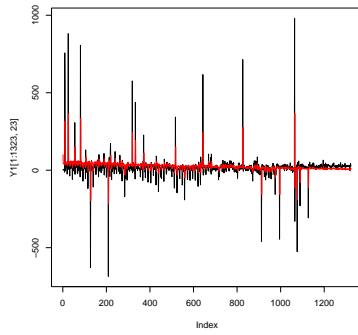


(a) Região 21.

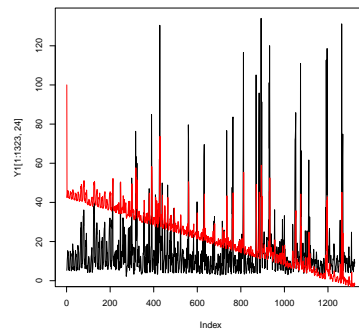


(b) Região 22.

Figura C.11: Gráfico dos estimadores do modelo $M2^*$.

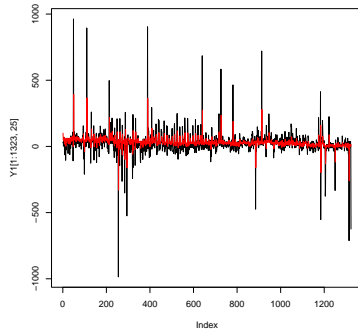


(a) Região 23.

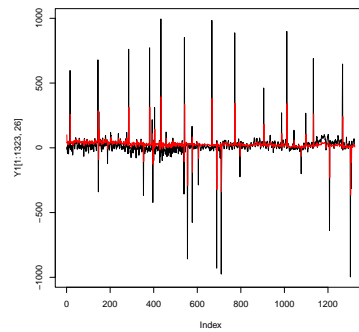


(b) Região 24.

Figura C.12: Gráfico dos estimadores do modelo $M2^*$.

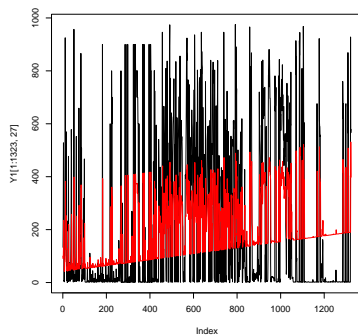


(a) Região 25.

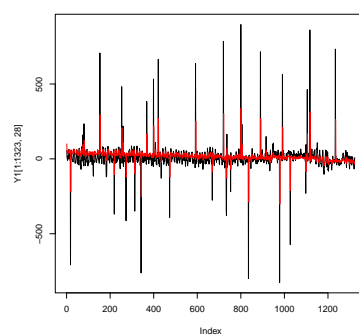


(b) Região 26.

Figura C.13: Gráfico dos estimadores do modelo $M2^*$.



(a) Região 27.



(b) Região 28.

Figura C.14: Gráfico dos estimadores do modelo $M2^*$.

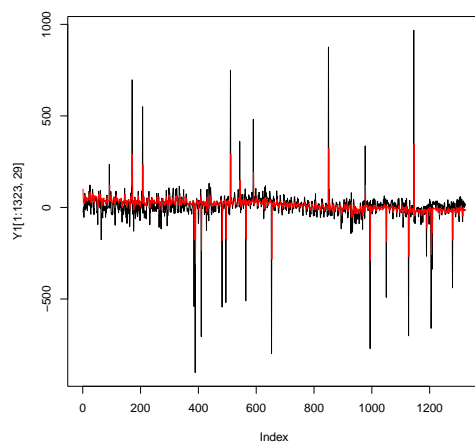


Figura C.15: Região 29 - Gráfico dos dados com os estimadores do modelo M2*.