

# Análise Hierárquica Bayesiana com Aplicação à Qualidade da Água do Rio dos Sinos

Rafael Bernardoni Chaves

Orientadora: Dra. Sílvia Regina Costa Lopes

IME - UFRGS - Brasil



## Introdução

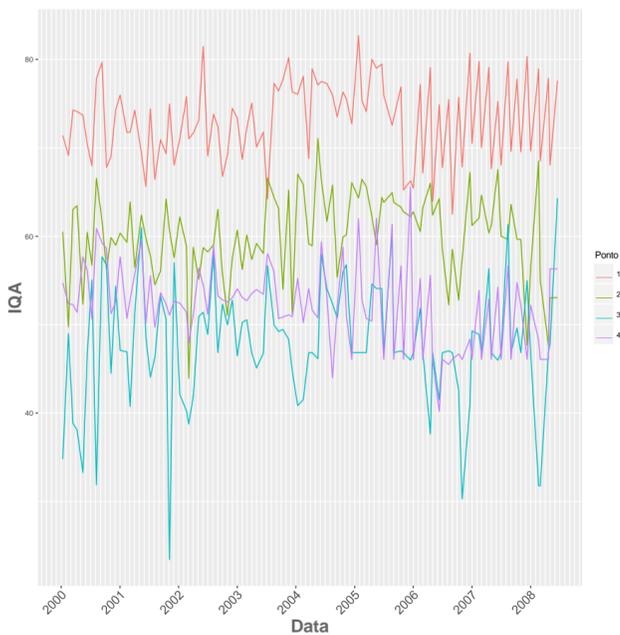
O objetivo do projeto de pesquisa é analisar os dados referentes ao Índice de Qualidade da Água (IQA), coletados em quatro pontos distintos distribuídos ao longo do curso do Rio dos Sinos. O período desse estudo é de janeiro de 2000 a junho de 2008, e os dados foram cedidos pela FEPAM (Fundação Estadual de Proteção Ambiental Henrique Luiz).

Normalmente, a qualidade da água é medida através da quantidade de coliformes (em mg/litro), pela quantidade de sólidos totais, pela turbidez, pela quantidade de sulfetos ou ainda pelo *iqg* (índice de qualidade da água). Neste trabalho utilizamos o *iqg* como forma de mensurar a qualidade da água.

Utilizando o ambiente R com o software *OpenBUGS*, foram realizadas simulações baseadas no método de Monte Carlo através de Cadeias de Markov (MCMC).

## Análise Descritiva das Séries Temporais

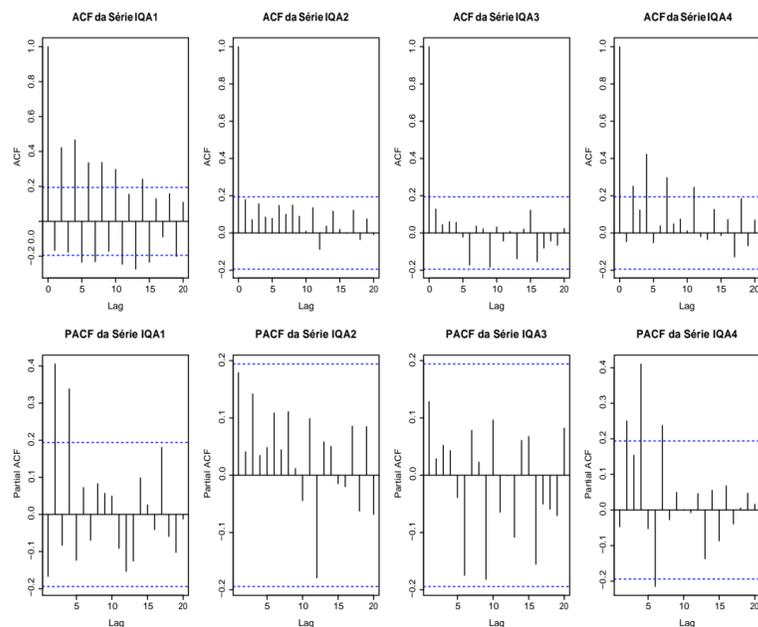
O trabalho iniciou-se com uma análise descritiva das séries temporais referentes aos quatro pontos de coleta ao longo do Rio dos Sinos:



Séries Temporais dos IQA's nos 4 Pontos de Coleta.

Nota-se maior qualidade média da água no ponto de coleta 1 em relação aos demais, seguido pelo ponto de coleta 2. Qual a causa desta degradação da qualidade da água no decorrer do curso do Rio dos Sinos?

Antes de testarmos como as séries temporais se ajustam aos modelos propostos, observamos os gráficos das funções de Autocorrelação (*acf*) e de Autocorrelação Parcial (*pacf*) para procurarmos algum padrão nas séries.



Gráficos das Funções de *acf* e *pacf* das Séries Temporais de IQA dos 4 Pontos de Coleta.

Identificamos as séries temporais de IQA dos pontos de coleta 2 e 3 como ruídos brancos, ou seja, não apresentam uma autocorrelação significativa. Os gráficos indicam, também, que as séries temporais dos pontos de coleta 1 e 4 aparentam ser processos Auto Regressivos.

## Metodologia

A metodologia de interesse deste projeto consiste na análise da estrutura espaço-temporal através dos modelos lineares generalizados, utilizando a metodologia Bayesiana. A estrutura temporal será modelada através de processos auto-regressivos de primeira ordem.

O modelo de interesse neste trabalho é dado por

$$g(\mu_{it}) \equiv E(Y_{it}) = a_0 + S_i(t) + R_i$$

onde  $Y_{it}$  é a contagem de um fenômeno aleatório,  $\mu_{it}$  é sua esperança matemática,  $a_0$  é uma constante,  $S_i(t)$  é o efeito temporal associado a qualquer região  $i$  e  $R_i$  é o efeito aleatório associado a  $i$ -ésima região, para  $i \in \{1, \dots, 4\}$  e  $t \in \{1, \dots, 102\}$  com  $g(\cdot)$  uma função que relaciona as componentes espaço-temporal e aleatória.

A componente  $R_i$  é a soma de efeitos aleatórios espacialmente correlacionados ( $b_i$ ) e efeitos aleatórios não-correlacionados ( $h_i$ ), onde  $b_i$  será modelado através de processos auto-regressivos condicionais intrínsecos (CAR). Primeiramente incluiremos apenas os efeitos não-correlacionados ( $h_i$ ) nas simulações.

Os dez modelos para explicar a componente de efeito temporal,  $S_i(t)$  que consideraremos neste trabalho são:

Modelo	$S_i(t)$	Especificação do Modelo
M1	$\delta t$	tendência linear aditiva para todas as 4 regiões
M1*	$\delta_{i,t}$	tendência linear aditiva para a $i$ -ésima região
M2	$A_1(t)$	$A_1(t) \sim N(\varphi_1 + \rho(A_1(t-1) - \varphi_{1-1}), \sigma_{S_1}^2)$
M2*	$A_1(i, t)$	$A_1(i, t) \sim N(\varphi_1 + \rho(A_1(i, t-1) - \varphi_{1-1}), \sigma_{S_{1,t}}^2)$
M3	$A_2(t)$	$A_2(t) \sim N(\varphi_1 A_2(t-1), \sigma_{S_1}^2)$
M3*	$A_2(i, t)$	$A_2(i, t) \sim N(\varphi_1 A_2(i, t-1), \sigma_{S_{1,t}}^2)$
M4	$A_3(t)$	$A_3(t) = \varphi_1 A_3(t-1) + S_t$
M4*	$A_3(i, t)$	$A_3(i, t) = \varphi_1 A_3(i, t-1) + S_{i,t}$
M5	$A_4(t)$	$A_4(t) = \varphi_1 + \rho(A_4(t-1) - \varphi_{1-1}) + S_t$
M5*	$A_4(i, t)$	$A_4(i, t) = \varphi_1 + \rho(A_4(i, t-1) - \varphi_{1-1}) + S_{i,t}$

Os modelos M1 e M1\* 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  $\rho$  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  $\rho$  extra nesta estrutura auto-regressiva aditiva.

O parâmetro  $\rho$  foi considerado por Waller et al. (1977b) como livre de qualquer imposição. Neste trabalho, consideramos  $\rho < 1$  para garantir a convergência das cadeias de Markov quando do uso do algoritmo Monte Carlo Markov Chain.

A diferença entre os modelos  $M_j$  e  $M_j^*$ , onde  $j \in \{1, 2, 3, 4, 5\}$ , é que para os modelos sem asterisco consideramos para todas as 4 regiões, enquanto os modelos com asterisco são considerados diferentemente para cada uma das regiões, onde  $j \in \{1, 2, 4\}$ .

## Resultados e Conclusão

Será apresentado um resultado entre as diversas simulações realizadas até o presente momento, as quais incluem os modelos  $M_{j,h}$ , onde  $j \in \{1, 2, 4\}$  sem efeitos aleatórios espacialmente correlacionados ( $b_i$ ), incluindo na componente  $R_i$  apenas os efeitos não-correlacionados ( $h_i$ ). Os efeitos aleatórios ( $b_i$ ) estão para ser adicionados em uma próxima etapa da pesquisa.

Para comparar os modelos através de simulações, utilizaremos alguns critérios de seleção de modelo propostos em Figueira et al (2016). Os critérios selecionados foram: *Deviance Information Criterion* (DIC) [ver Spiegelhalter et al (2002)], *Logarithm of the Pseudo Marginal Likelihood* (LPML) [ver Souza (2015) e Congdon (2006)], *Pearson Residuals* (Qp) [ver Paulino et al (2009)] e *Widely Applicable Information Criterion* (WAIC) [ver Vehtari e Gelman (2015), Gelman et al (2014) e Watanabe (2013a, b)].

Temos, então, os seguintes parâmetros de uma das nossas simulações:

- Número de Cadeias = 2;
- Pulo = 3;
- Burn-in = 1000;
- Número de Iterações = Burn-in + 4000;
- Amostra Efetiva = 1333.

### Resultado de uma simulação

Modelo	$Q_p$	LPML	DIC	WAIC1	WAIC2	$\sigma_{\eta}^2$
M1h	217.54	-519.69	2539.00	1039.37	1039.39	113.64
M1h*	213.28	-517.80	2533.00	1035.53	1035.60	121.52
M2h	196.60	-510.77	2515.00	1021.17	1021.53	277.24
M2h*	191.53	-509.15	2514.00	1017.68	1018.32	304.48
M4h	195.71	-509.15	2512.00	1017.68	1018.23	250.11
M4h*	179.47	-505.16	2509.00	1009.99	1009.86	607.11

Comparação do ajuste dos modelos  $M_{j,h}$ , onde  $j \in \{1, 2, 4\}$ .

Com as simulações realizadas até o momento, esse foi o melhor resultado. Os valores a serem escolhidos para os parâmetros devem ser baseados nas medidas descritivas, e a mudança de um modelo para o outro deve ser feita alterando apenas um dos parâmetros, para uma melhor observação da influência de cada um. Ainda precisamos realizar mais testes no futuro. Algumas das considerações mais importantes, seriam buscar um valor mais eficiente para estimar  $a_0$ , descobrir e resolver os problemas envolvendo os modelos M3h, M3h\*, M5h e M5h\*. Nossa proposta é analisar as simulações para determinar qual o modelo que melhor descreve os dados apresentados, dentre os modelos aqui estudados. Para isso, incluiremos em estudos futuros, os efeitos aleatórios espacialmente correlacionados ( $b_i$ ) nos modelos estudados.

## Bibliografia

- Congdon, P. (2006). *Bayesian statistical modelling*. London: Wiley.
- Figueira, C.V., S.R.C. Lopes e G.L. Silva (2017). *Spatiotemporal Analysis by Autoregressive Models*. Submetido.
- Gelman, A., J. Hwang e A. Vehtari (2014). *Understanding predictive information criteria for Bayesian models*. *Statistics and Computing*, vol. 24(6), 997-1016.
- Paulino, C.D., A.A. Turkman e B. Murteira (2009). *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian.
- Rampaso, R.C., A.D.P. Souza e E.F. Flores (2015). *Bayesian analysis of spatial data using different variance and neighbourhood structures*. *Journal of Statistical Computation and Simulation*, DOI:10.1080/00949655.2015.1022549.
- Spiegelhalter, D.J., N.G. Bestm, B.P. Carlin, e A. Van der Linde (2002). *Bayesian Measures of Model Complexity and Fit (with Discussion)*. *Journal of the Royal Statistical Society, Series B*, Vol. 64(4), 583-616.
- Vehtari, A. e A. Gelman (2015). *WAIC and cross-validation in Stan*. Disponível em [http://www.stat.columbia.edu/~gelman/research/unpublished/waic\\_stan.pdf](http://www.stat.columbia.edu/~gelman/research/unpublished/waic_stan.pdf).
- Watanabe, S. (2013a). *WAIC and WBIC are information criteria for singular statistical model evaluation*. Disponível em [http://www.me.inf.kyushu-u.ac.jp/witmse2013/abst\\_files/watanabe2013.pdf](http://www.me.inf.kyushu-u.ac.jp/witmse2013/abst_files/watanabe2013.pdf).
- Watanabe, S. (2013b). *A widely applicable Bayesian information criterion*. *Journal of Machine Learning Research*, vol. 14(3), 867-897.

VOZES DIVERSAS  
DIFERENTES SABERES



SALÃO DE INICIAÇÃO CIENTÍFICA  
XXX SIC

15 A 19  
OUTUBRO  
CAMPUS DO VALE



rafaelbechaves@gmail.com e  
silvia.lopes@ufrgs.br