Resumo - Iniciação Científica
Escalabilidade de métodos semiparamétricos para inferência em processos estocásticos espaço-temporais

Pedro Nasevicius Ramos
Orientador: Guilherme Vieira Nunes Ludwig

22 de outubro de 2020

Resumo

Modelos semiparamétricos são frequentemente utilizados em estatística espacial, mas o aumento de resolução e a alta dimensionalidade dos dados de experimentos científicos modernos gera problemas de escalabilidade computacional dos métodos semiparamétricos existentes. A proposta deste trabalho é considerar técnicas de fusão de dados espaço-temporais e explorar sua escalabilidade computacional e implementação. O procedimento de *covariance tapering* foi implementado para um modelo espaço-temporal, e um estudo de simulação foi desenvolvido para ilustrar as diferenças entre as metodologias. O estudo foi motivado por aplicação em análise de dados meteorológicos.

1 Introdução

Dados espaço-temporais completamente observados têm distribuições conjuntas que usualmente envolvem matrizes de covariância com número de linhas iguais ao tamanho do número de locais amostrados, vezes o número de períodos de tempo. Consequentemente, técnicas espaciais como a *krigagem* (Cressie, 1993) necessitam de alternativas escaláveis, para evitar a inversão de uma matriz com número de linhas e colunais iguais à dimensão dos dados. Isto foi feito utilizando o modelo espaço-temporal semi paramétrico de Ludwig et al. (2017) usando *covariance tapering* (Furrer et al., 2006) no componente espacial. Entre as aplicações possíveis de um modelo espaço-temporal com escalabilidade computacional encontra-se a análise dos dados meteorológicos da estações automáticas do INMET (Ministério da Agricultura, Pecuária e Abastecimento: Instituto Nacional de Meteorologia, 2011).

2 Metodologia

Seja $\mathbf{D} \subset \mathbb{R}^2$ um domínio bidimensional de interesse, e $T \subset [0, \infty]$ um intervalo de tempo de interesse. Um processo espaço-temporal é um conjunto de variáveis aleatórias indexadas em $\mathbf{s} \in \mathbf{D}$ e $t \in T$, ou seja:

$$\{Y(\mathbf{s}, t) : \mathbf{s} \in \mathbf{D}, t \in T\}.$$

Para o problema assumiu-se o modelo de Ludwig et al. (2017) com estrutura

$$Y(\mathbf{s},t) = \mu(\mathbf{s},t) + \eta(\mathbf{s},t) + \varepsilon(\mathbf{s},t),$$

em que $\mu(\mathbf{s},t)$ é uma função de média determinística, $\eta(\mathbf{s},t)$ é um processo espaço-temporal e $\varepsilon(\mathbf{s},t)$ um ruído branco. Sem perda de generalidade, considere a função média constante. Como $\varepsilon(\mathbf{s},t)$ é um ruído branco, assume-se média zero, variação constante $\text{Var}(\varepsilon(\mathbf{s},t)) = \sigma^2$ e correlação zero.

O processo espaço-temporal $\eta(\mathbf{s},t)$ possui média zero e função de covariância espaço-temporal $C(\mathbf{s},\mathbf{s}',t,t')$. Além disso, possui decomposição do tipo Karhunen-Loève (Gromenko and Kokoszka, 2013):

$$\eta(\mathbf{s},t) = \sum_{l=1}^{\infty} \xi_l(\mathbf{s})\varphi_l(t)$$

em que $\{\varphi_l(t)\}_{l=1}^{\infty}$ é uma sequência de funções temporais ortogonais determinísticas e $\{\xi_l(\mathbf{s})\}_{l=1}^{\infty}$ uma sequência de processos espaciais com média zero não correlacionados entre sí. Estes processos, além de estacionários, devem ser isotrópicos, isto é, com função de covariância dada por:

$$\text{Cov}(\xi_l(\mathbf{s}), \xi_l(\mathbf{s}')) = K_l(\mathbf{s},\mathbf{s}') = \lambda_l \rho_l(\|\mathbf{s}-\mathbf{s}'\|\,;\theta_l)$$

em que $\rho_l(\|\mathbf{s}-\mathbf{s}'\|\,;\theta_l)$ é uma função de correlação parametrizada por θ_l, $\|\cdot\|$ é a norma Euclidiana e $\text{Var}(\xi_l(\mathbf{s})) = \lambda_l$. Dessa maneira, a função de covariância espaço-temporal fica:

$$C(\mathbf{s},\mathbf{s}',t,t') = \sum_{l=1}^{\infty} K_l(\mathbf{s},\mathbf{s}')\varphi_l(t)\varphi_l(t').$$

Para obter a predição de um ponto y não amostrado numa localização $\mathbf{s}_0$ e em um tempo t_0, é necessário utilizar a suposição em que assumimos o processo como Gaussiano. Dessa maneira, o melhor preditor linear ótimo não viesado de y_0 é a esperança condicionada nos pontos $\mathbf{y}$ conhecidos, dado por:

$$\hat{y}(\mathbf{s}_0,t_0) = \mu(\mathbf{s}_0,t_0) - \text{Cov}(\eta(\mathbf{s}_0,t_0),\eta)\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mu),$$

em que $\mathbf{y} \sim N(\mu,\boldsymbol{\Sigma})$, e $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{y},\mathbf{y})$ é a matriz de covariância calculada usando a função de covariância espaço-temporal definida anteriormente. Perceba que é necessário resolver um sistema linear para computar $\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mu)$, e o número de operações para este calculo é da ordem de $N^3 = (nm)^3$. Dados mensais de $n = 100$ lugares amostrados diariamente por quatro meses resultam em uma amostra de tamanho igual a $N = 1200$, uma matriz de covariância com 1200^2 entradas e um número de operações para resolver o sistema linear na ordem de 1200^3. Dessa forma, conforme aumentamos o tamanho da amostra, a complexidade computacional aumenta de maneira que seja necessário alternativas para lidar com a escalabilidade do problema.

Este projeto propôs-se à utilizar a metodologia desenvolvida em Furrer et al. (2006) baseada em funções *tapering*. A ideia é utilizar estas funções de maneira a tornar a matriz de covariância espaço-temporal esparsa, assim é possível utilizar técnicas, como por exemplo a decomposição de Choleski esparsa, que se utilizam dessa esparsidade para tornar a resolução do sistema computacionalmente mais eficiente.

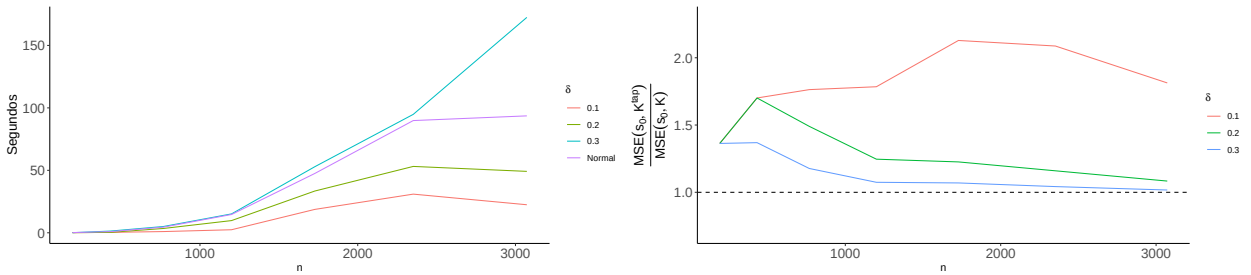
Além disso, o artigo de Furrer et al. (2006) mostra que, sob condições específicas, o erro quadrático médio da predição (MSE) utilizando o método de *tapering* em um ponto arbitrário $\mathbf{s}_0$ é assintoticamente equivalente ao MSE do modelo de covariância sem taper. Saliente-se que a metodologia apresentada em Furrer et al. (2006) é para processos espaciais apenas, mas pode ser estendido para o caso espaço-temporal (com número finito de autovalores) como será argumentado no estudo de simulação.

3 Resultados e Aplicação

3.1 Estudo de Simulação

Para verificar o ganho computacional utilizando o método utilizando o *taper* e sua equivalência assintótica do erro quadrático médio, foi realizado algumas simulações. Os fatores que variam são o tamanho da amostra n e o parâmetro de dependência δ, que é o parâmetro da função de *taper* responsável por introduzir esparsidade na matriz de covariância. O domínio espacial amostrado é $\mathbf{D} = [0, 1]^2$, particionado em uma malha regular cuja resolução está em função de n. Utilizado um tempo fixo $T = \{1, 2, \cdots, 12\}$, os valores de n utilizados foram 192, 432, 768, 1200, 1728, 2352 e 3072. Os valores de δ utilizados são 0.1, 0.2 e 0.3.

A Figura 1 (a) mostra as médias dos tempos de estimação dos parâmetros para cada interação dos fatores n e δ. Note que, conforme aumentamos a dimensão do problema, o tempo médio para estimação aumenta mais rapidamente para o método sem *taper* que os outros. Porém, note que para o valor de $\delta = 0.3$ o tempo é maior que o modelo sem *taper*. Conforme diminuímos δ, ou seja, introduzimos esparsidade na matriz, o tempo médio de estimação também é reduzido. Contudo, é necessário cuidado pois reduzir muito o parâmetro δ pode ocasionar em uma má predição dos dados, pois estaremos reduzindo demais δ ao ponto de não capturar nenhuma informação. Isto é evidenciado na Figura 1 (b), que mostra a razão do MSE utilizando o método de *taper* com o MSE do sem *taper*, já que para $\delta = 0.1$ temos um MSE quase 2 vezes maior que o do método sem *taper*. Além disso, na Figura 1 (b) observamos que conforme aumentamos o tamanho da amostra, a razão dos MSEs parece convergir para 1 para os casos de δ igual a 0.2 e 0.3. Isto é uma evidência de que temos equivalencia assintótica dos MSE utilizando o método com *taper*.



(a) Comparação do tempo computacional entre os métodos com e sem *taper*. Foi calculado as médias dos tempos de estimação dos parâmetros para cada simulação com interação dos fatores n e δ. A curva roxa com legenda "Normal" é o modelo sem *taper*

(b) A razão entre os MSE dos modelos com e sem *taper*.

Figura 1: Estudo de simulação

3.2 Aplicação em Dados Meteorológicos

Os dados espaço-temporais utilizados na aplicação foram retirados das informações meteorológicas obtidas pelo INMET (Instituto Nacional de Meteorologia), um órgão do Ministério da Agricultura, Pecuária e Abastecimento. No endereço `http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas` estão disponíveis dados sobre estações meteorológicas automáticas (EMAs) que coletam, a cada 15 minuto, informações como: temperatura, umidade, pressão atmosférica, precipitação, direção e velocidade dos ventos, radiação solar. Foram utilizadas apenas informações referentes a precipitações (milímetros) em EMAs que possuíam dados completos de Janeiro de 2018 a Março de 2019, somando 38 estações distribuídas pelos estados de SP, MG e RJ.

Dois mapas dinâmicos dos dados de predição foram feitos, um utilizando o modelo espaço-temporal proposto em Ludwig et al. (2017), o outro aplicando metodologia de *taper* proposta neste trabalho. Foi feita uma transformação aplicando a raiz quadrada, dessa forma os dados ficam mais próximos de um processo Gaussiano. As Figuras 2 e 3 trazem mapas dinâmicos da predição para, respectivamente, o modelo sem e com *taper*.

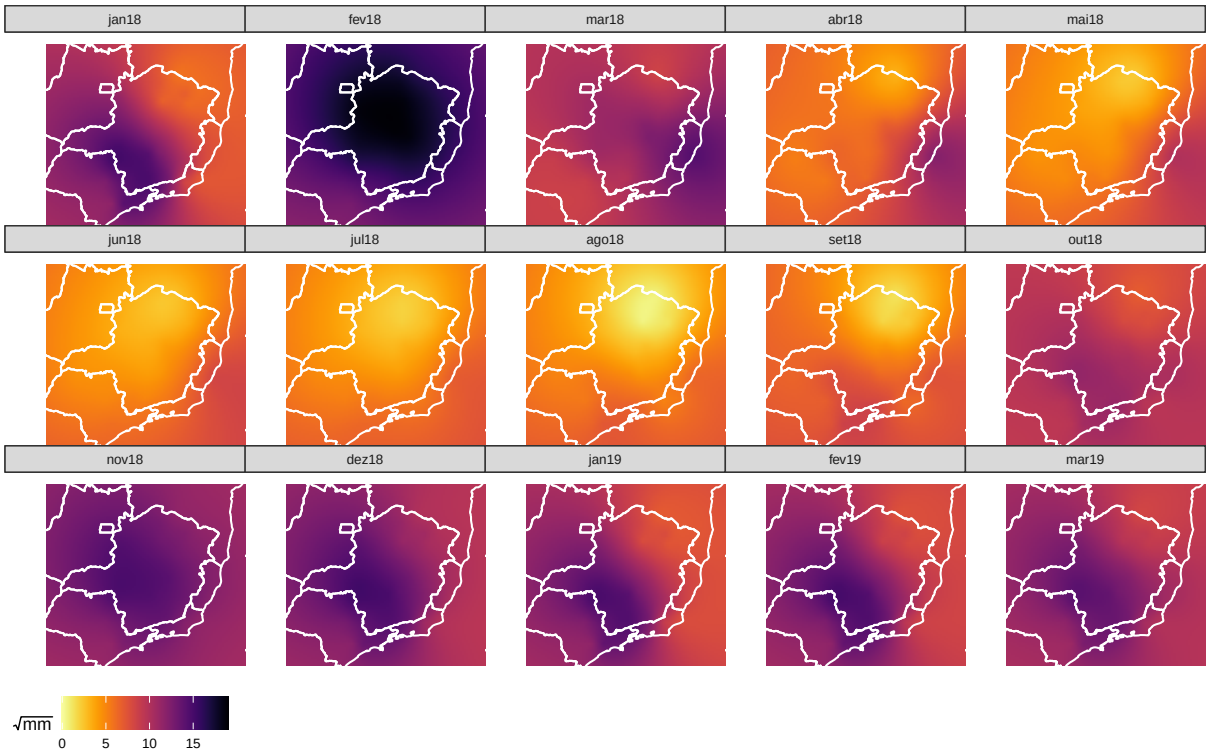


Figura 2: Mapas Dinâmicos da predição da raiz quadrada da precipitação usando o modelo espaço-temporal sem *taper*. Os períodos se dão entre Janeiro de 2018 a Março de 2019.

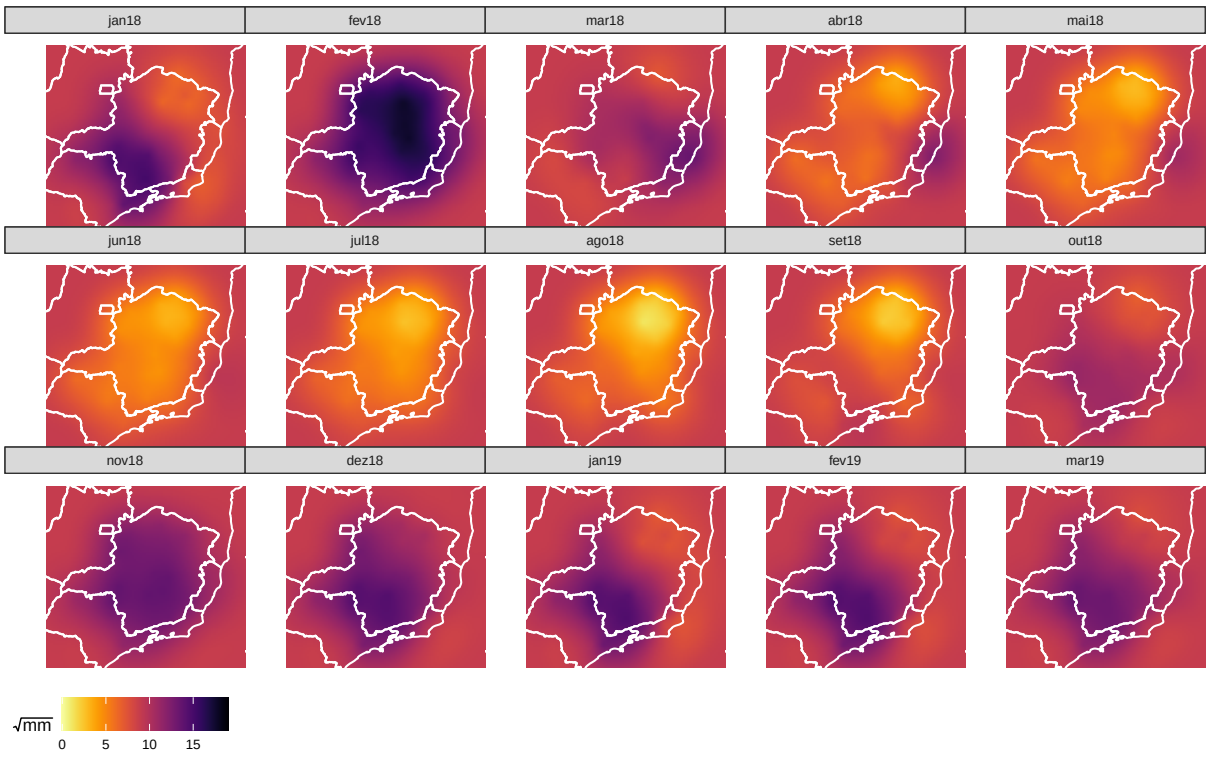


Figura 3: Mapas Dinâmicos da predição da raiz quadrada da precipitação usando o modelo espaço-temporal com *taper*. Os períodos se dão entre Janeiro de 2018 a Março de 2019.

Referências

Cressie, N. (1993). *Statistics for Spatial Data, 2nd edition*. Wiley, New York.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gromenko, O. and Kokoszka, P. (2013). Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Computational Statistics & Data Analysis*, 59:82–94.

Ludwig, G., Chu, T., Zhu, J., Wang, H., and Koehler, K. (2017). Static and roving sensor data fusion for spatio-temporal hazard mapping with application to occupational exposure assessment. *The Annals of Applied Statistics*, 11(1):139–160.

Ministério da Agricultura, Pecuária e Abastecimento: Instituto Nacional de Meteorologia (2011). *NOTA TÉCNICA No. 001/2011/SEGER/LAIME/CSC/INMET: Rede de Estações Meteorológicas Automáticas do INMET*.