Centro Pi
Centro de Projetos
e Inovação IMPA

RELATÓRIO 7º WORKSHOP DE SOLUÇÕES MATEMÁTICAS PARA PROBLEMAS INDUSTRIAIS

GPP ESALQ / USP

Desafio de Mudança de Suporte

JÚLIO HOFFIMANN[1,*], YULIA PETROVA[1], J. EZEQUIEL S. SOTO[2], AND JOSÉ LUCAS SAFANELLI[3]

[1]*Centro Pi – IMPA, Rio de Janeiro, Brasil*
[2]*Instituto Tecgraf – PUC-Rio, Rio de Janeiro, Brasil*
[3]*Grupo de Políticas Públicas ESALQ/USP, Piracicaba, Brasil*
[*]*Contato: julio.hoffimann@impa.br*

24 de Setembro de 2021

Com a disseminação de tecnologias digitais na agricultura, um grande volume de dados têm sido produzido para auxiliar o monitoramento agrícola e tomadas de decisão em intervalos de tempo cada vez mais curtos envolvendo todos os municípios do território brasileiro. Nesse contexto, informações meteorológicas (ex: temperatura, precipitação, umidade) são cruciais para alimentar modelos de predição de safra e antecipar potenciais riscos na produção. No entanto, dados coletados em estações meteorológicas (pontos no mapa) precisam ser transferidos para municípios (polígonos no mapa) de forma eficiente para permitir tomadas de decisão em tempo real. O presente relatório tem como objetivo descrever o resultado de investigações realizadas para solucionar o desafio de mudança de suporte.

1. INTRODUÇÃO

O método mais comum para transferir dados de estações meteorológicas para municípios consiste em interpolar os dados em uma imagem de alta resolução e em seguida agregar os pixels dentro de cada município. Devido a etapa de interpolação (ou "rasterização") exaustiva, o método não é recomendado para aplicações em tempo real. Além disso, poucos são os estudos que tratam a mudança de suporte diretamente em geometrias esféricas, e muito menos em polígonos de estados e municípios do Brasil. Portanto, surgem as seguintes perguntas:

1. Seria possível contornar a etapa de interpolação e transferir diretamente os dados de estações para polígonos de geometria arbitrária na esfera?

2. Como integrar esses processos estocásticos em polígonos esféricos e obter uma quantificação de dispersão nas estimativas?

As soluções investigadas neste desafio possuem a capacidade de auxiliar o monitoramento de safras agrícolas de todo o território brasileiro, contribuindo assim com a proposição de ações imediatas que visam minimizar os riscos decorrentes de uma produção agrícola insatisfatória. Além disso, as soluções matemáticas exploradas são pertinentes para a comunidade de ciência de dados espacial.

A. Dados meteorológicos

Para explorar este desafio, propõe-se a utilização da rede de estações automáticas de monitoramento meteorológico do Instituto Nacional de Meteorologia (INMET), do Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Os dados de cerca de 500 estações meteorológicas automáticas podem ser consultados por interfaces programáticas em https://portal.inmet.gov.br/manual. As variáveis de interesse englobam as temperaturas máxima, mínima e média; precipitação acumulada; radiação solar acumulada; velocidade média do vento; e umidade relativa média, todas em escala temporal diária, as quais são imprescindíveis para alimentar modelos de simulação de cultura.

B. Dados geométricos

A extensão territorial da investigação engloba as estações automáticas do INMET e a malha de municípios dos estados de PA, MA, TO, PI, BA, MT, MS, GO, DF, MG, SP, PR, SC e RS. Propõe-se a utilização de geometrias do projeto https://gadm.org.

Ambos os dados podem ser facilmente acessados via interfaces programáticas através dos pacotes INMET.jl e GeoTables.jl para a linguagem de programação Julia.

2. FORMULAÇÃO DO PROBLEMA

Consideramos estações meteorológicas do INMET localizadas em pontos $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n \in \mathsf{S}^2$ na esfera e polígonos esféricos $\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_m \subset \mathsf{S}^2$ representando municípios do Brasil.

Definimos um processo estocástico $Z(\boldsymbol{s})$ de suporte pontual para cada instante de tempo e cada variável de interesse, e assumimos que esse processo é observado nas estações $z_i = Z(\boldsymbol{s}_i),\ i = 1, 2, \ldots, n$.

O problema de *mudança de suporte* [1] consiste em inferir estatísticas sobre o processo médio $Z_{\mathsf{P}_j}(\boldsymbol{s}) = \frac{1}{|\mathsf{P}_j|}\int_{\mathsf{P}_j} Z(\boldsymbol{x})dA$ em cada polígono P_j usando os dados pontuais de todas as

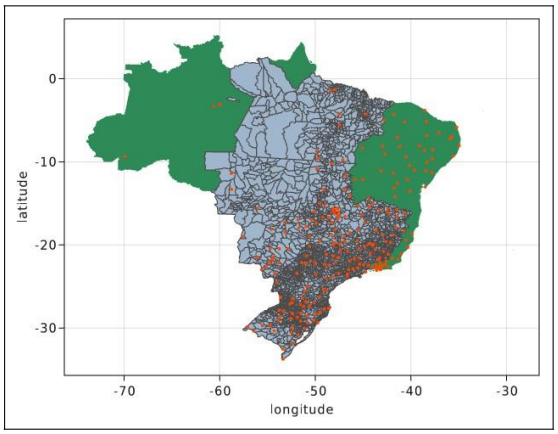

Fig. 1. Estações meteorológicas do sistema INMET em vermelho e geometrias de municípios brasileiros do projeto GADM em cinza.

estações $z_1, z_2, \ldots, z_n$. Neste conjunto de dados temos $n \approx 230$ e $m \approx 3700$.

Parte do desafio está no fato que estatísticas de dispersão como variância são função do suporte, neste caso polígonos esféricos de geometria complexa (não-convexa, com furos, etc.)

3. METODOLOGIA PROPOSTA

O estimador mais popular na geoestatística é o estimador linear

$$\hat{Z}_{\mathsf{P}_j}(s) = \lambda_1 Z(s_1) + \lambda_2 Z(s_2) + \cdots + \lambda_n Z(s_n)$$

com pesos $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)^\mathsf{T} \in \mathsf{R}^n$ calculados a partir de uma função covariância/variograma $\gamma \colon \mathsf{S}^2 \times \mathsf{S}^2 \to \mathsf{R}$:

$$\begin{bmatrix} \boldsymbol{G} & \mathbf{1} \\ \mathbf{1}^\mathsf{T} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ v \end{bmatrix} = \begin{bmatrix} g \\ 0 \end{bmatrix} \tag{1}$$

onde $G_{ij} = \gamma(s_i, s_j)$ são avaliações do variograma nas estações $1 \le i < j \le n$, $g_i = \frac{1}{|\mathsf{P}|} \int_{\mathsf{P}} \gamma(s_i, s) dA$ são integrais do variograma entre a estação i e o município com geometria P e v é o multiplicador de Lagrange do problema de minimização de variância $Var\left[Z_{\mathsf{P}_j}(s) - \hat{Z}_{\mathsf{P}_j}(s)\right]$ sujeito à restrição linear $\mathbf{1}^\mathsf{T}\boldsymbol{\lambda} + v = 0$.

A metodologia proposta consiste em calcular as integrais g_i para funções γ conhecidas e admissíveis na esfera, i.e. condicionalmente negativas. Dividimos o trabalho em duas principais investigações ao longo da semana do workshop:

- Integração numérica de uma função variograma admissível γ em polígonos arbitrários, também conhecido como o problema de *regularização espacial* na geoestatística [2].
- Definição de uma função covariância/variograma admissível na esfera, o problema de *interpolação em variedades*.

A. Regularização espacial

Assumindo que as posições das estações meteorológicas e as geometrias dos municípios são fixas (ou fixas por um longo período de tempo), é possível resolver o sistema linear da Equação 1 uma única vez para cada município e armazenar os vetores de pesos para uso em tempo real. Dado um vetor de observações z do processo nas estações, a estimativa e variância condicional em todos os polígonos municipais é dada por simples produtos matriciais

$$\hat{z} = \boldsymbol{\Lambda} z, \quad \boldsymbol{\sigma}^2 = \boldsymbol{\Lambda} g \tag{2}$$

onde $\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2 & \ldots & \boldsymbol{\lambda}_m \end{bmatrix}^\mathsf{T}$ é a matriz de pesos de tamanho $m \times n$ armazenada em disco.

Neste trabalho, as integrais g_i são aproximadas com o método de Monte Carlo:

$$g_i = \frac{1}{|\mathsf{P}|} \int_{\mathsf{P}} \gamma(s_i, s) dA \approx \frac{1}{N} \sum_{k=1}^{N} \gamma(s_i, s_k) \tag{3}$$

com pontos $s_k \in \mathsf{P}$, $k = 1, 2, \ldots, N$ no polígono de interesse, amostrados de uma distribuição "blue noise" ou "Poisson disk" para uma boa cobertura da geometria. O Algoritmo 1 de amostragem consiste em três principais etapas:

1. Triangulação do polígono P e cálculo de áreas dos triângulos para amostragem subsequente.
2. Amostragem homogênea onde triângulos são selecionados com probabilidade proporcional a área, e onde pontos são amostrados nos triângulos de forma uniforme usando coordenadas baricêntricas.
3. Rejeição de amostras próximas com o auxílio de uma árvore de busca e segundo uma distância mínima $\alpha > 0$.

Algorithm 1. Amostragem blue noise

procedure BLUENOISE(P, α, ρ) — *d* Raio $\alpha > 0$ e $\rho > 0$.
  $\Delta \leftarrow$ *triangulate*(P) — *d* Triangulação do polígono
  $w \leftarrow$ *areas*(Δ) — *d* Áreas dos triângulos
  $K \leftarrow \left\lceil \frac{2|\mathsf{P}|}{\sqrt{3}} \frac{\rho^2}{\alpha} \right\rceil$ — *d* Valor esperado Poisson disk
  $n \leftarrow 0$ — *d* Número de amostras
  $s \leftarrow \{\}$ — *d* Conjunto de amostras
  while $n < K$ **do** — *d* Amostragem homogênea
    $\Delta_o \leftarrow$ *sample*(Δ, w) — *d* Amostra triangulo
    $s_o \leftarrow$ *sample*(Δ_o) — *d* Amostra ponto
    $s \leftarrow s \cup \{s_0\}$
    $n \leftarrow n + 1$
  $\mathsf{T} \leftarrow$ *kdtree*(s) — *d* Árvore de busca
  $s' \leftarrow$ *reject*(s, α, T) — *d* Rejeição de amostras próximas
  return s' — *d* Amostras blue noise

A Figura 2 ilustra a amostragem blue noise em função do parâmetro $\alpha > 0$. A amostragem apresenta várias propriedades teóricas interessantes que fogem do escopo deste relatório [3, 4].

B. Interpolação em variedades

O problema de interpolação em variedades engloba o caso particular de interpolação na esfera S^2. É possível utilizar resultados da literatura de equações diferenciais estocásticas para estabelecer uma função variograma admissível [5]. Em particular, é possível mostrar que a covariância Matérn é positiva semi-definida na esfera ao se trocar a distância Euclideana pela distância geodésica [6, 7]:

$$k(s_i, s_j) = \frac{\sigma^2}{C_\nu} \sum_{n=0}^{\infty} \left(\frac{2\nu}{\kappa^2} + n(n+1) \right)^{-(\nu+1)} \cdot c_n \cdot \mathsf{C}_n^{1/2} \left(\cos(\mathrm{dist}(s_i, s_j)) \right) \tag{4}$$

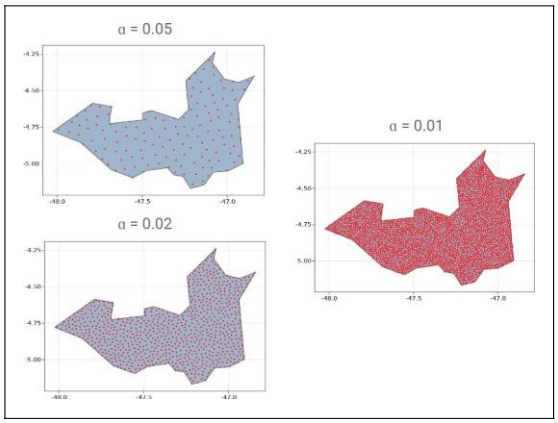

Fig. 2. Amostragem blue noise para diferenes parâmetros $\alpha > 0$. Quanto menor o parâmetro α, maior a densidade de pontos (em vermelho) no polígono P.

A Equação 4 é expressa em termos de uma série infinita de polinômios Gegenbauer C_b^{ν} ilustrados na Figura 3. A função covariância Matérn possui três parâmetros $\nu > 0$, $r > 0$, $\sigma^2 > 0$ que caracterizam a suavidade, comprimento de correlação e variância a priori, respectivamente.

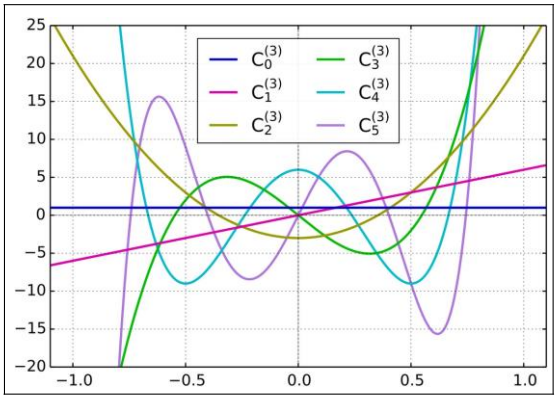

Fig. 3. Família de polinômios Gegenbauer.

Uma boa aproximação da função de covariância k é obtida via truncamento da série nos $M = 50$ primeiros termos. A Figure 4 ilustra a covariância para diferentes comprimentos de correlação.

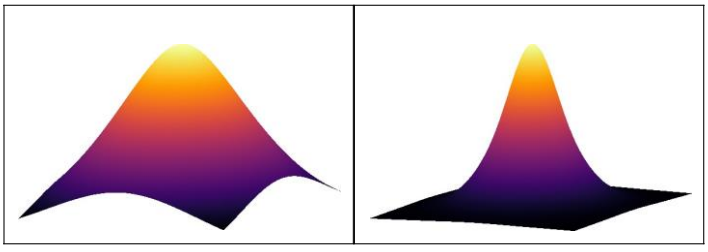

Fig. 4. Covariância Matérn para $\kappa = 0.25$ e $\kappa = 0.1$, comprimento de correlação.

4. RESULTADOS ALCANÇADOS

A etapa de regularização do variograma para todos os $\approx 4k$ polígonos municipais de interesse e todas as ≈ 200 estações meteorológicas foi realizada com sucesso em $< 20min$ de CPU. O resultado desse pré-processamento foi então utilizado para obter estimativas instantâneas ($\approx 50ms$) em todo território nacional sem etapas intermediárias de rasterização, como ilustrado na Figura 5.

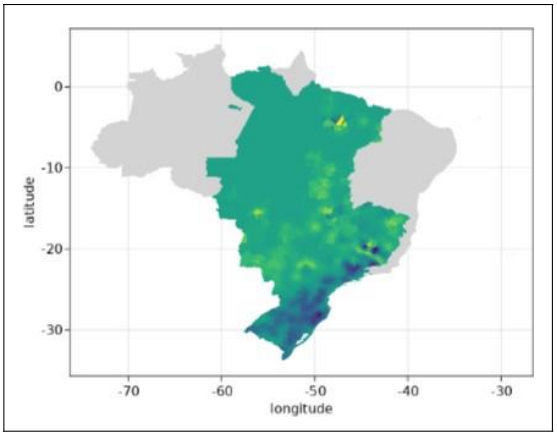

Fig. 5. Estimativas em $\approx 4k$ municípios brasileiros diretamente de ≈ 200 estações meteorológicas, sem etapa de rasterização intermediária.

Além disso, alguns resultados preliminares comparando a função covariância Matérn no plano versus na esfera foram alcançados.

5. PRÓXIMOS PASSOS

Os próximos passos do trabalho consistem em estudar melhor os parâmetros de aproximação da função covariância na esfera, como por exemplo o número de termos no truncamento da série infinita e o número de amostras Monte Carlo para os diferentes municípios brasileiros.

Em seguida, pretendemos realizar estudos comparativos de estimativas municipais no plano versus esfera, assim como definir algoritmos Bayesianos de ajuste de paramêtros. Esses algoritmos de ajuste vão permitir identificar a frequência mínima de pré-processamento para os diferentes processos físicos, i.e. períodos de estacionaridade temporal da covariância.

AGRADECIMENTOS

À GPP/Esalq pelo desafio proposto, pela sugestão do conjunto de dados e pelo patrocínio do evento. À coordenação do Centro Pi pela organização do evento.

REFERENCES

1. P. Renard, H. Demougeot-Renard, and R. Froidevaux, *Geostatistics for Environmental Applications* (Springer Berlin Heidelberg, 2005).
2. X. Emery, "Change-of-support models and computer programs for direct block-support simulation," Comput. & Geosci. **35**, 2047-2056 (2009).
3. A. Lagae and P. Dutré, "A comparison of methods for generating poisson disk distributions," Comput. Graph. Forum **27**, 114-129 (2008).
4. J. Bowers, R. Wang, L.-Y. Wei, and D. Maletz, "Parallel poisson disk sampling with spectrum analysis on surfaces," ACM Transactions on Graph. **29**, 1-10 (2010).

5. V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth, “Matérn gaussian processes on riemannian manifolds,” (2021).
6. C. Huang, H. Zhang, and S. M. Robeson, “On the validity of commonly used covariance and variogram functions on the sphere,” Math. Geosci. **43**, 721-733 (2011).
7. T. Gneiting, “Strictly and non-strictly positive definite functions on spheres,” Bernoulli. **19** (2013).