Modelos Lineares Generalizados para Dados Espaciais (MLGDE)

Kally Chung
Prof. Paulo Justiniano Ribeiro Jr
Prof. Wagner Bonat

04 de Fevereiro de 2020

UFPR - PPGMNE/LEG
I CidWeek

Table of contents

Introdução

Introdução

- Dados geoestatísticos gaussianos: contínuos e de distribuição simétrica.
  - Krigagem,
  - Método da verossimilhança.
- Dados geoestatísticos não gaussianos: dados binários, contagem, contínuos com cauda pesada, contínuos assimétricos, entre outros.
  - Modelo Linear Generalizado Espacial (MLGE) - Gotway & Stroup, 1997 [4],
  - Modelo Linear Generalizado Misto (MLGM) - Bonat & RibeiroJr, 2015 [1],
  - Modelo de Regressão de Cópula Gaussiana (MRCG) - Masarotto & Varin, 2017 [8]

.

Modelo Geral

Modelo Geral

- Considere N observações, n_β parâmetros de regressão e n_d parâmetros de dispersão.
- Modelo geral, conforme Liang & Zeger, 1986 [7]:

$$E[\mathbb{Y}] = \mu = g^{-1}(\mathbb{X}\beta),$$
$$Var[\mathbb{Y}] = C = V^{1/2}\Omega V^{1/2}$$

- onde $g(.)$ é função de ligação,
- $V_{N\times N}$ é a matriz de variância,
- $\Omega_{N\times N}$ é a matriz de covariância.

Modelo Linear Generalizado para Dados Espaciais

MLGDE

- Definição de $C = V^{1/2}\Omega V^{1/2}$:

$$V(p) = diag(v(p)) = diag(\mu^p),$$
$$\Omega(\tau) = \Omega(\tau_0, \tau_1, \tau_2) = \tau_0 R(\tau_1) + \tau_2 I.$$

- onde $v(p) = \mu^p$ é função de variância, da família Tweedie, (Jørgensen, 1987 [6]),
- $\tau = (\tau_0, \tau_1, \tau_2)$ é a contribuição $\tau_0 \geq 0$, o alcance $\tau_1 \geq 0$ e o efeito pepita $\tau_2 \geq 0$ (Diggle & Ribeiro Jr (2007) [2]),
- $R(d_{ij}, \tau_1)_{N\times N}$ é definida pela função de correlação espacial ρ.
  - Exponencial ou Matern $\kappa = 0.5$:

$$\rho(d_{ij}, \tau_1) = exp\left(-\frac{d_{ij}}{\tau_1}\right).$$

  - Para mais funções de correlação espacial, verificar Diggle & Ribeiro Jr (2007) [2].

Estimação (Holst & Jørgensen, 2015 [5])

- Seja a estimação dos parâmetros $\theta = (\beta, \lambda) = (\beta, p, \tau_0, \tau_1, \tau_2)$.
- Problema: resolver

$$\varphi = (\varphi_\beta, \varphi_\lambda) = \begin{cases} \varphi_\beta = D^T C(\mathbb{Y} - \mu) = 0 \\ \varphi_{\lambda_i} = tr(W_{\lambda_i}(rr^T - C)) = 0 \quad , \lambda_i = p, \tau_0, \tau_1, \tau_2 \end{cases}$$

- onde $D_{N \times n_\beta} = \nabla_\beta \mu$,
- $W_{\lambda_i} = C^{-1} \dfrac{\partial C}{\partial \lambda_i} C^{-1}$,
- resíduo $r = \mathbb{Y} - \mu$.

Algoritmo Chaser(Holst & Jørgensen, 2010 [5])

- Método Iterativo baseado no Fisher Scoring

$$\beta^{(i+1)} = \beta^{(i)} - S_\beta^{-1}(\beta^{(i)}, \lambda^{(i)})\ \varphi_\beta(\beta^{(i)}, \lambda^{(i)})$$
$$\lambda^{(i+1)} = \lambda^{(i)} - S_\lambda^{-1}(\beta^{(i+1)}, \lambda^{(i)})\ \varphi_\lambda(\beta^{(i+1)}, \lambda^{(i)})$$

- onde $S_\beta = -D^T C^{-1} D$,
- $S_{\lambda_{i,j}} = -tr\left(C^{-1}\dfrac{\partial C}{\partial \lambda_i} C^{-1}\dfrac{\partial C}{\partial \lambda_j}\right)$, com $\lambda_i = p, \tau_0, \tau_1, \tau_2$.

Correção de Viés(Holst & Jørgensen, 2015 [5])

- A equação quasi-score $\varphi_{\lambda_i} = tr(W_{\lambda_i}(rr^T - C))$ é viesada se os parâmetros de β são desconhecidos.
- Para fazer a correção de viés,

$$b_{\lambda_i} = -tr(J_\beta^{\lambda_i} J_\beta^{-1}) = -tr\left(J_\beta^{\lambda_i} \frac{\partial J_\beta}{\partial \lambda_i}\right)$$

$$\text{onde } J_\beta^{-1} = S_\beta^{-1} V_\beta S_\beta^{-T},$$

$$V_\beta = Var[\varphi_\beta] = D^T C^{-1} D.$$

- Desenvolvendo algebricamente b_{λ_i}, tem-se

$$b_{\lambda_i} = -tr(D^T W_{\lambda_i} D S_\beta^{-T}).$$

- Fazendo a correção de viés em φ_λ, tem-se:

$$\begin{aligned}\check{\varphi}_{\lambda_i}(\beta, \lambda) &= \varphi(\beta, \lambda) + b_{\lambda_i}(\beta, \lambda) \\ &= tr(W_{\lambda_i}(rr^T - C)) - tr(D^T W_{\lambda_i} D S_\beta^{-T}).\end{aligned}$$

Reparametrização

- Fazendo

$$\Omega = \tau_0 \left(\rho(\tau_1) + \frac{\tau_2}{\tau_0} I \right) = \tau_0 (\rho(\tau_1) + \tau_2^* I) = \tau_0 \Delta,$$

$$\text{com } \Delta = \begin{bmatrix} 1+\tau_2^* & \rho(d_{12},\tau_1) & \cdots & \rho(d_{1N},\tau_1) \\ \rho(d_{21},\tau_1) & 1+\tau_2^* & \cdots & \rho(d_{2N},\tau_1) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(d_{N1},\tau_1) & \rho(d_{N2},\tau_1) & \cdots & 1+\tau_2^* \end{bmatrix},$$

- considera-se a seguinte reparametrização:

$$\gamma = (\gamma_0, \gamma_1, \gamma_2) = (ln\ \tau_0, ln\ \tau_1, ln\ \tau_2^*) = \left(ln\ \tau_0, ln\ \tau_1, ln\ \frac{\tau_2}{\tau_0} \right)$$

Parâmetros Iniciais

- Para β inicial, considera-se o usual modelo linear generalizado (MLG),
- Para p inicial, utiliza-se $p = 0$ se os dados são contínuos ou $p = 1$ se os dados são de contagem,
- Considera-se γ_2 empiricamente como 20% da dispersão encontrada no MLG.
- Calcula-se a função $\varphi_\lambda(\beta, \gamma_1)$ e considera-se γ_1^{Inic} tal que $\varphi_\lambda(\beta, \gamma_1^{Inic}) = 0$.
- Encontrado o valor de γ_1^{Inic}, calcula-se

$$\hat{\gamma_0} = \begin{cases} ln\left(\dfrac{r^T \Delta^{-1} r}{N}\right) & \text{, sem correção de viés} \\ ln\left(\dfrac{r^T \Delta^{-1} r}{N - n_\beta}\right) & \text{, caso contrário} \end{cases}.$$

Erro Padrão da Estimação

- Seja $\hat{\theta} = (\hat{\beta}, \hat{\lambda})$ a estimativa de θ, então a distribuição assintótica de $\hat{\theta}$ é

$$\hat{\theta} \sim N(\theta, J_\theta^{-1}),$$

- em que $J_\theta^{-1} = S_\theta^{-1} V_\theta S_\theta^{-T}$,
- $S_\theta = \begin{bmatrix} S_\beta & S_{\beta,\lambda} \\ S_{\lambda,\beta} & S_\lambda \end{bmatrix} = \begin{bmatrix} E[\nabla_\beta \varphi_\beta(\beta,\lambda)] & E[\nabla_\lambda \varphi_\beta(\beta,\lambda)] \\ E[\nabla_\beta \varphi_\lambda(\beta,\lambda)] & E[\nabla_\lambda \varphi_\lambda(\beta,\lambda)] \end{bmatrix}$,
- $V_\theta = \begin{bmatrix} V_\beta & V_{\beta,\lambda} \\ V_{\lambda,\beta} & V_\lambda \end{bmatrix} = \begin{bmatrix} V_\beta & V_{\lambda,\beta}^T \\ V_{\lambda,\beta} & V_\lambda \end{bmatrix}$.
- Para calcular o erro padrão EP_θ,

$$EP_\theta = \sqrt{diag(J_\theta^{-1})}.$$

Predição (Gotway & Stroup [4])

- Seja $\mathbb{Y} = (Y_1(s_1), Y_2(s_2), \ldots, Y_N(s_N))^T$ a variável resposta das observações nas localidades $s_1, s_2, \ldots, s_N$.
- Predizer os valores para $\mathbb{Y}_l = (Y(l_1), Y(l_2), \ldots, Y(l_{n_u}))^T$ dos n_u locais não observados $l_1, l_2, \ldots, l_{n_u}$.
- Para a predição, usa-se o estimador tipo krige através de *MLGE*, dado por

$$\hat{\mathbb{Y}}_l = \hat{\mu}(l) + C_{l,s} C_s^{-1} (\mathbb{Y} - \hat{\mu}(s))$$

- onde $Var \begin{bmatrix} \mathbb{Y} \\ \mathbb{Y}_l \end{bmatrix} = \begin{bmatrix} C_s & C_{s,l} \\ C_{l,s} & C_l \end{bmatrix} = \begin{bmatrix} C_s & C_{l,s}^T \\ C_{l,s} & C_l \end{bmatrix}$,
- $\hat{\mu}(s)$ e $\hat{\mu}(l)$ são os valores preditos pelos parâmetros β's da regressão.

Análise de Dados

Conjunto de dados - Rongelap ([3])

- Medições da contaminação residual de césio de testes nucleares no Atol Rongelap, um atol das Ilhas Ralik, pertencente às Ilhas Marshall, na Micronésia.

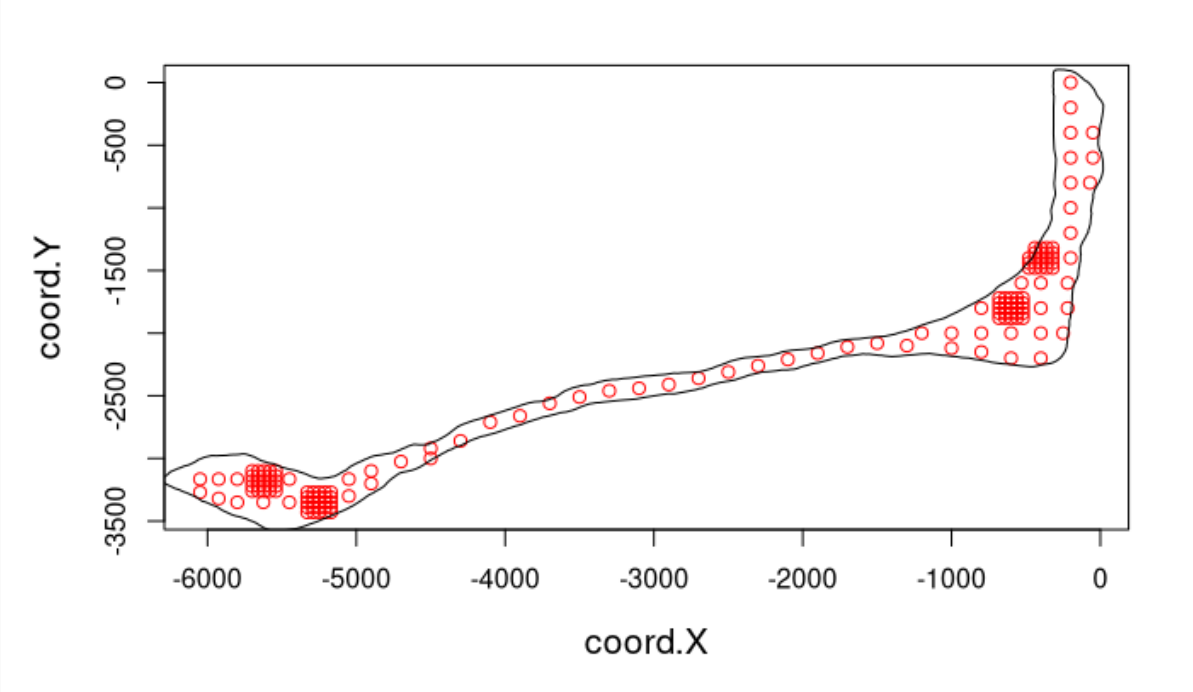

Figure 1: Mapeamento dos 157 locais de medições do resíduo césio ao longo do Atol. Existem 4 regiões da ilha com maior quantidade de medições.

Comparação de MLG com MLGDE

- Conjunto de dados usado: Rongelap.
- Parâmetros estimados: $\theta = (\beta, \lambda) = (\beta_0, \tau_0)$ com outros parâmetros de dispersão $p = 1, \tau_1 = 1, \tau_2 = 0$ fixados.
- O parâmetro de potência $p = 1$ na família Tweedie indica a variância da distribuição de probabilidade Poisson.
- Quando $\tau_2 = 0$, então a matriz de correlação é $\rho = I$, indicando que os dados são independentes.

Table 1: Estimativas e erros padrões dos parâmetros de β_0 e de τ_0 no *MLG* e *MLGDE* nas duas primeiras linhas e os valores de quasi-score nas duas últimas linhas. O erro padrão de τ_0 não é informado no sumário da função *glm* de R.

	MLG		*MLGDE com corr.*	
	Estim.	E.Padrão	Estim.	E.Padrão
β_0	2.0140	0.0283	2.0140	0.0283
τ_0	378.815	NA	378.8142	47.4193
φ_β			$-2.884e-09$	
φ_λ			$9.516e-13$	

Reparametrização

- Conjunto de dados usado: Rongelap.
- Considerou-se a estimação de todos os parâmetros: $\theta = (\beta, \lambda) = (\beta, p, \tau_0, \tau_1, \tau_2)$.

Table 2: Estimativas e erros padrões de parâmetros obtido com o *MLGDE*, tanto com a correção de viés e sem.

	MLGDE sem corr.		*MLGDE com corr.*	
	Estim.	E.Padrão	Estim.	E.Padrão
β_0	1.9770	0.0670	1.9757	0.0770
p	1.7321	0.3613	1.7472	0.3324
τ_0	0.3720	1.0271	0.3627	0.9269
τ_1	312.6222	234.5627	408.5757	306.2242
τ_2	0.7766	2.2867	0.7020	1.9089
τ_1 inicial	56.5830		59.7487	

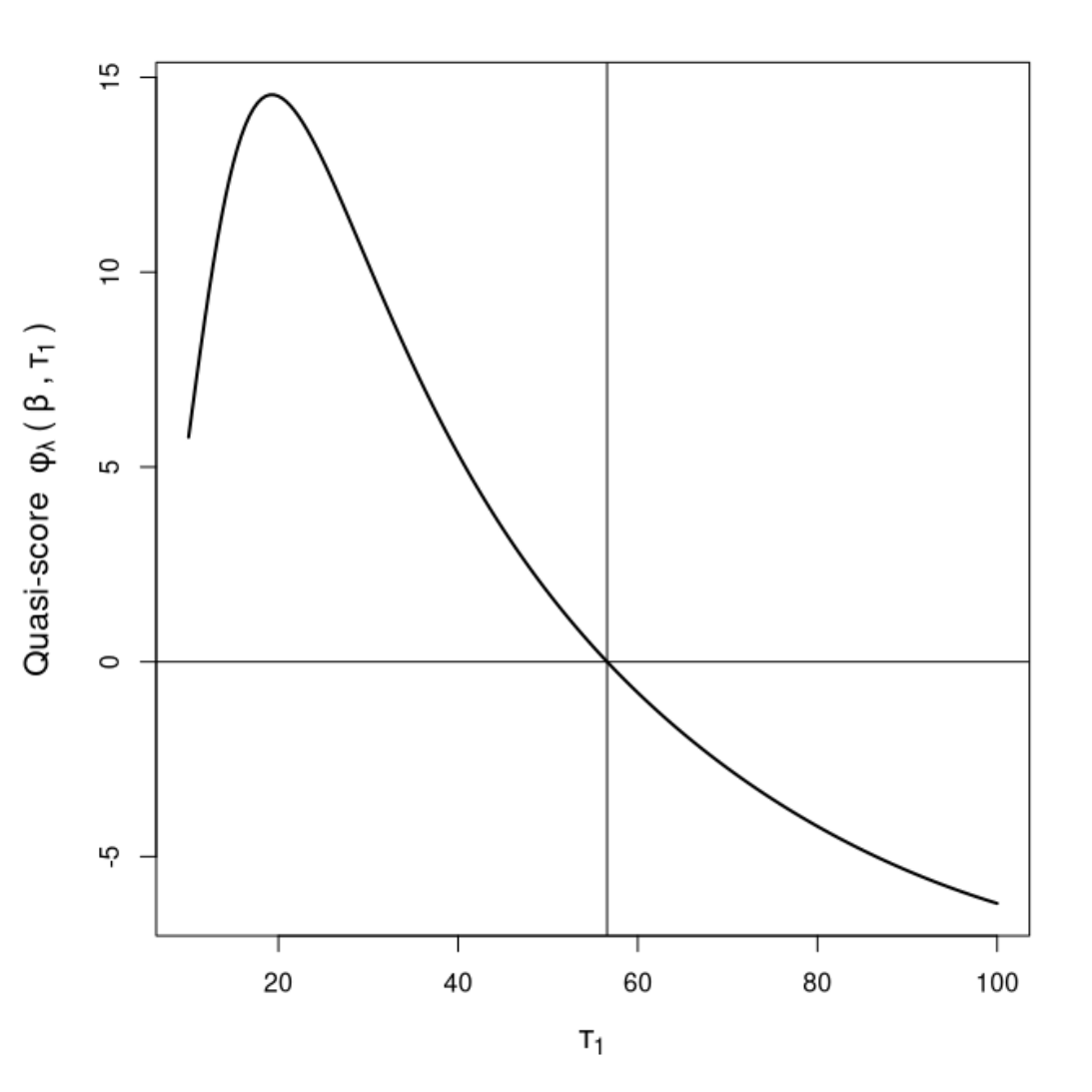

Figure 2: Função usada para determinar τ_1 inicial para o conjunto de dados Rongelap, sem correção de viés. Reforça-se que a função é similar para *MLGDE* com a correção de viés.

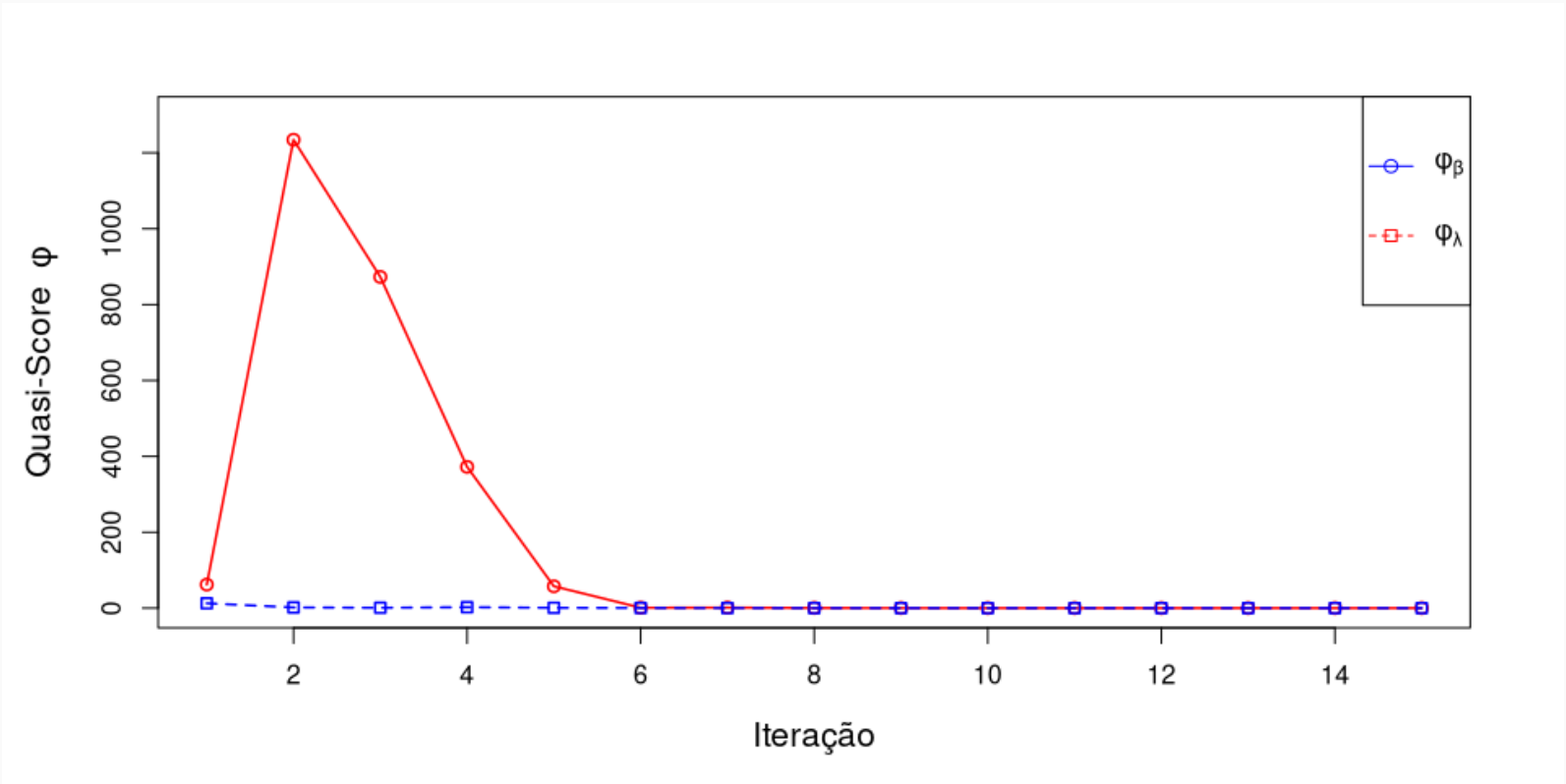

Figure 3: Valores de quasi-score ao longo das iterações do algoritmo Chaser no caso de *MLGDE* sem correção de viés. Observe que φ_β é estável ao longo das iterações, enquanto φ_λ cresce e depois cresce e estabiliza. Este comportamento ocorre também para o caso de *MLGDE* com correção.

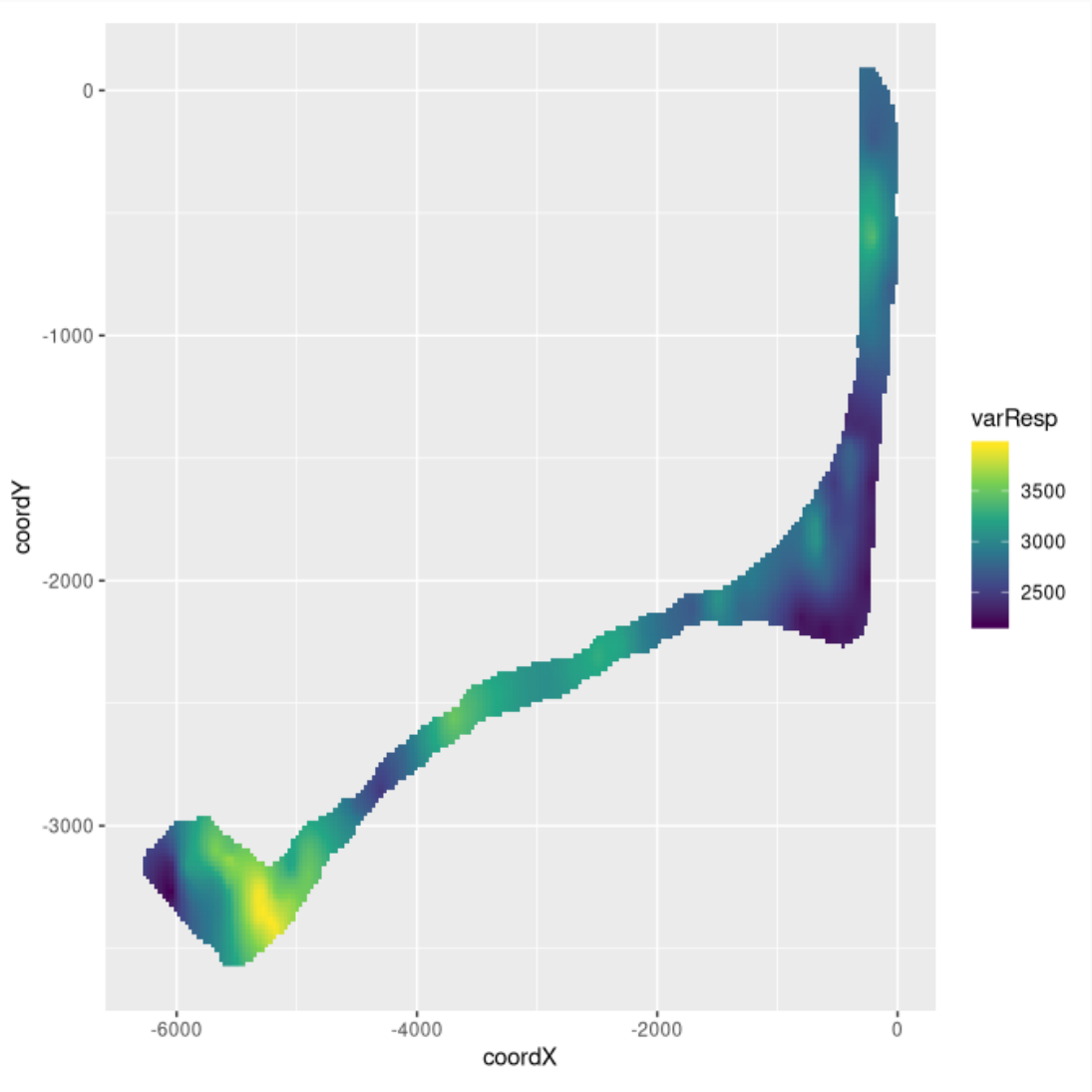

Figure 4: Mapa do atol Rongelap ilustrado por meio dos valores preditos, obtidos com a estimativa sem a correção de viés. O mapa predito com os parâmetros estimados com a correção de viés é muito similar a este apresentado.

Conjunto de dados - CTC ([9])

- O indicador da capacidade da troca de Cátions (CTC) é importante pois mede a qualidade do solo e auxilia na decisão de quais produtos usar no solo antes de um plantio.

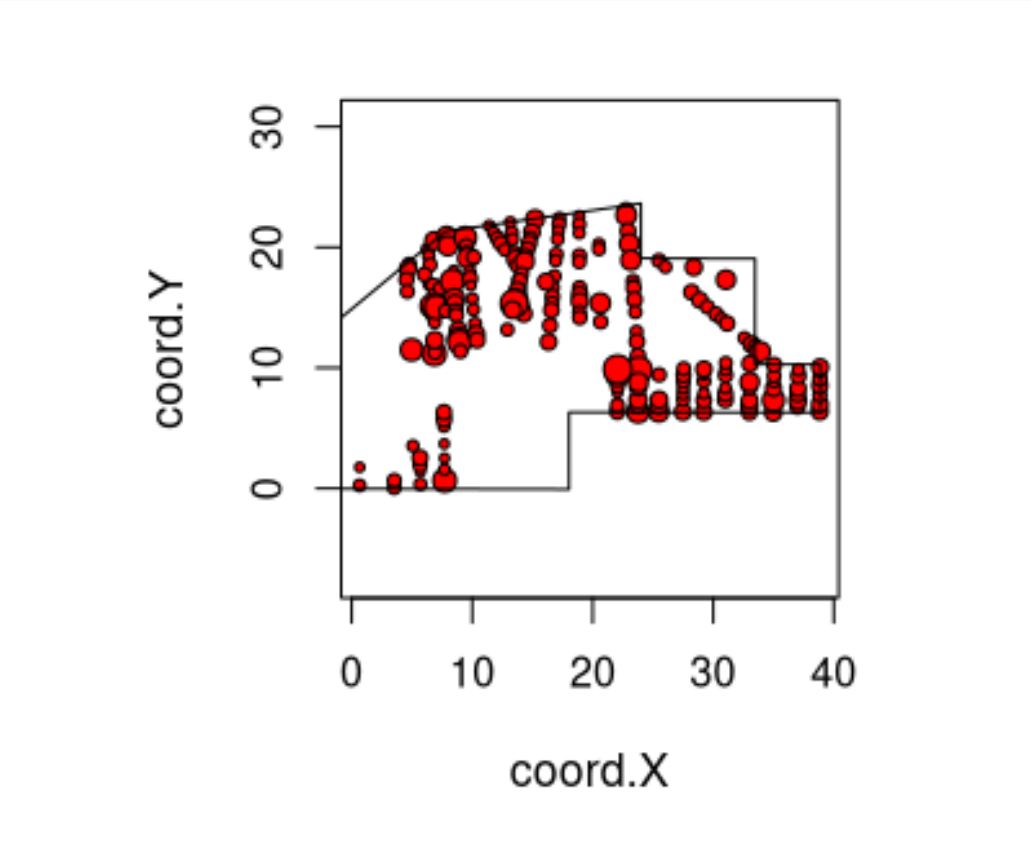

Figure 5: Mapa do local das 212 medidas do conjunto de dados de CTC, além de ter uma noção do valor do indicador a partir do raio do círculo.

Comparação da Verossimilhança com MLGDE

- Conjunto de dados usado: *CTC*.
- Tem-se os parâmetros $\theta = (\beta, \lambda) = (\beta_0, \tau_0, \tau_1)$ com outros parâmetros de dispersão $p = 0, \tau_2 = 0$ fixados.
- Com $p = 0$, usa-se a variância da distribuição gaussiana, conforme a família Tweedie.

Table 3: Estimativas e erros padrões de parâmetros entre *MLGDE* sem a correção de viés e a inferência por meio do método da máxima verossimilhança nas três primeiras linhas, além do valor inicial de τ_1 encontrado para o algoritmo Chaser na última linha.

	Máxima Veros.		*MLGDE sem corr.*	
	Estim.	E.Padrão	Estim.	E.Padrão
β_0	2.9349	NA	2.9349	0.1173
τ_0	1.9201	NA	1.9201	0.1972
τ_1	0.4343	NA	0.4343	0.0939
τ_1 inicial			0.4323	

Table 4: Estimativas e erros padrões de parâmetros entre *MLGDE* e a inferência por meio do método da verossimilhança restrita nas três linhas iniciais e o valor inicial de τ_1 encontrado para o algoritmo Chaser na última linha.

	Veros. Restrita		*MLGDE com corr.*	
	Estim.	E.Padrão	Estim.	E.Padrão
β_0	2.9355	NA	2.9355	0.1189
τ_0	1.9380	NA	1.9380	0.1997
τ_1	0.4469	NA	0.4469	0.0951
τ_1 inicial			0.4464	

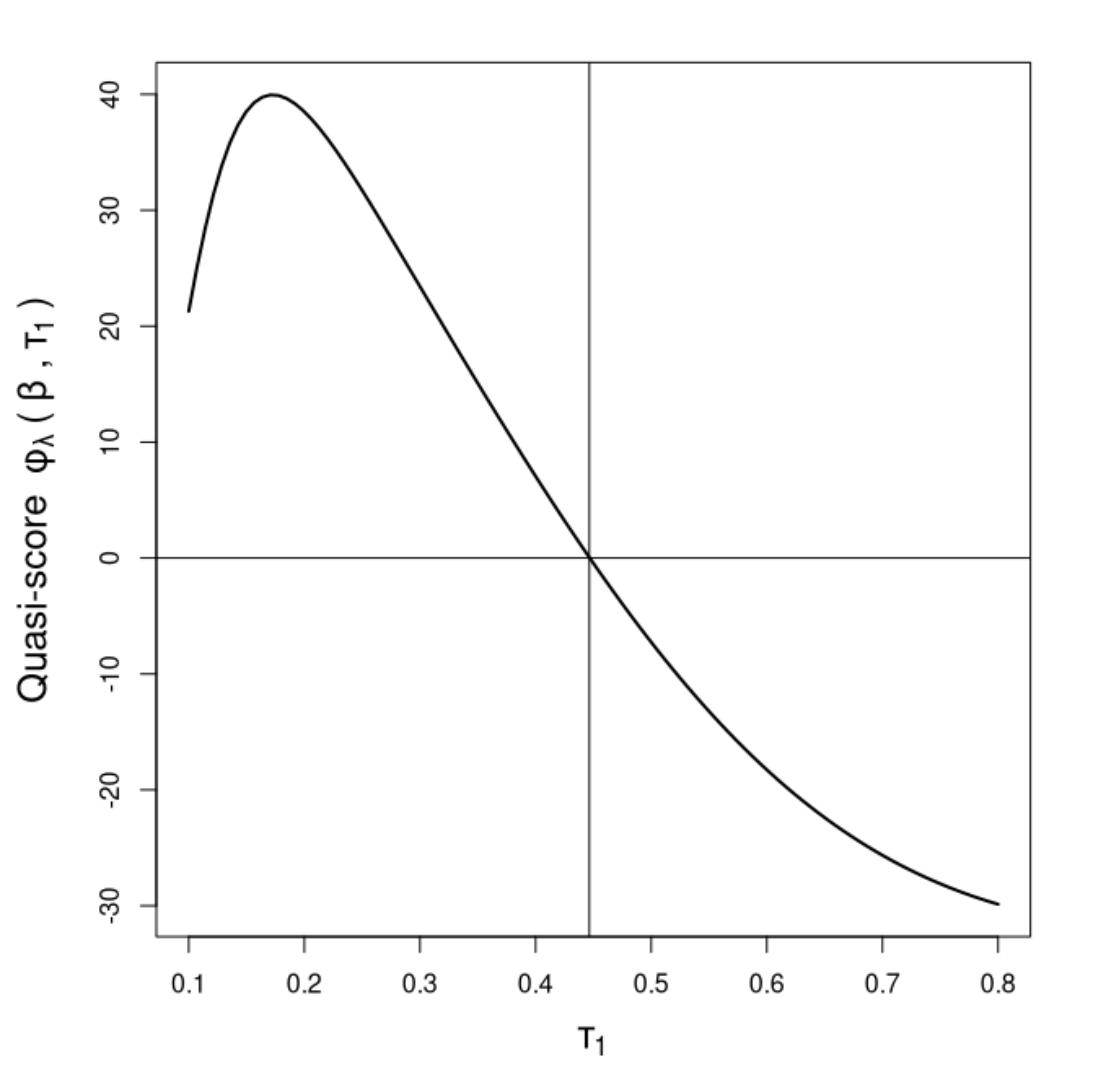

Figure 6: Função para obtenção de τ_1 inicial em *MLGDE* com a correção de viés, que é a raiz da função φ_λ. Esta função é similar também para a obtenção de τ_1 inicial em *MLGDE* com correção de viés.

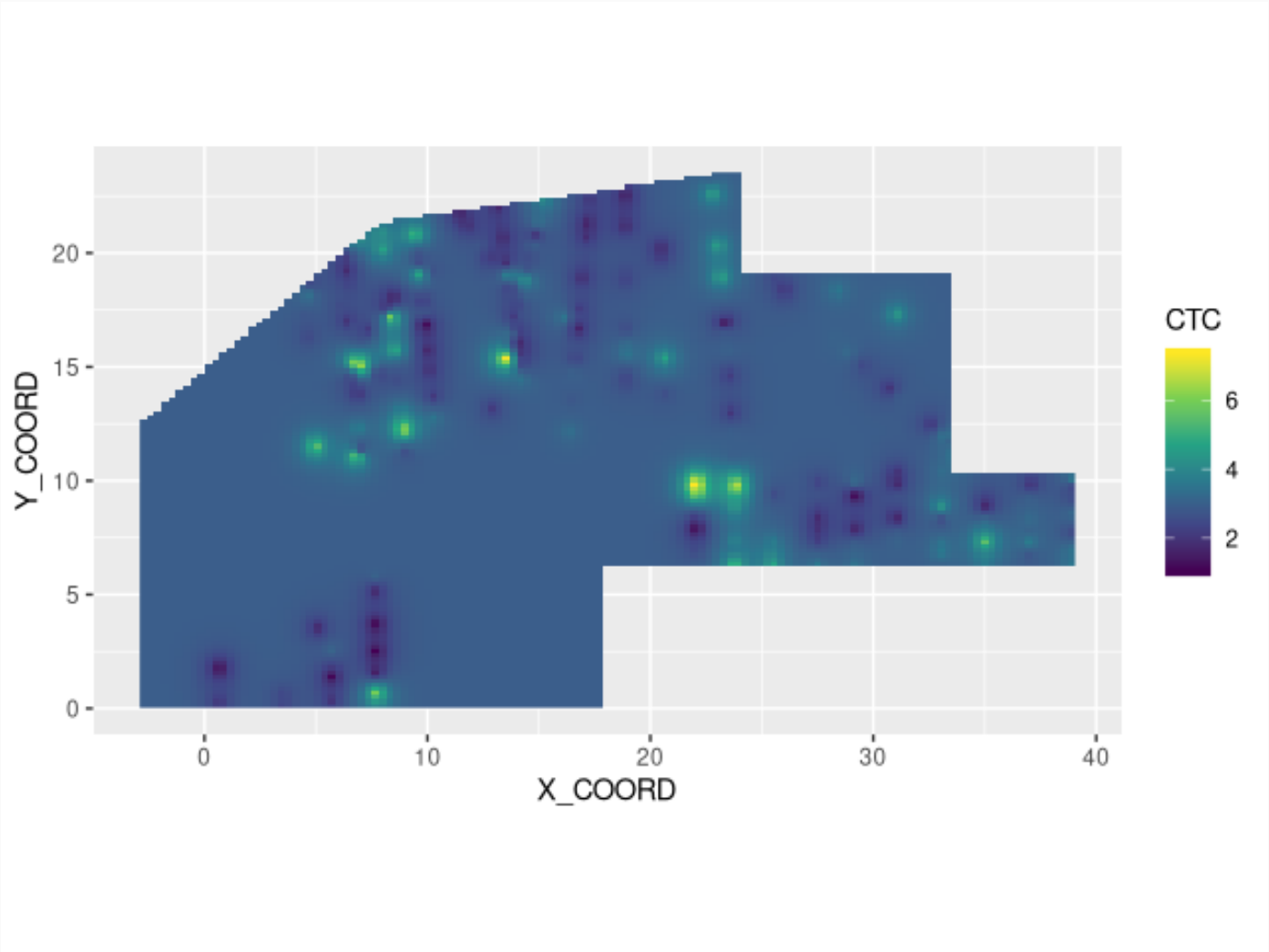

Figure 7: Mapa do *CTC* resultante dos valores preditos, a partir das estimativas da Tab. 4, isto é, com a correção de viés. O mapa predito a partir dos parâmetros obtidos com a estimação sem a correção de viés é muito similar a este apresentado.

Reparametrização

- Conjunto de dados usado: *CTC*.
- Tem-se os parâmetros $\theta = (\beta, \lambda) = (\beta_0, \tau_0, \tau_1, \tau_2)$ com o parâmetro de dispersão $p = 0$ fixado.

Table 5: Estimativas e erros padrões de parâmetros obtido com o *MLGDE*, tanto com a correção de viés e sem.

	MLGDE sem corr.		*MLGDE com corr.*	
	Estim.	E.Padrão	Estim.	E.Padrão
β_0	2.9517	0.1488	2.9540	0.1542
τ_0	1.0750	0.4661	1.0885	0.4452
τ_1	1.0385	0.5135	1.1006	0.5266
τ_2	0.8194	0.5000	0.8304	0.4749
τ_1 inicial	0.7364		0.7646	

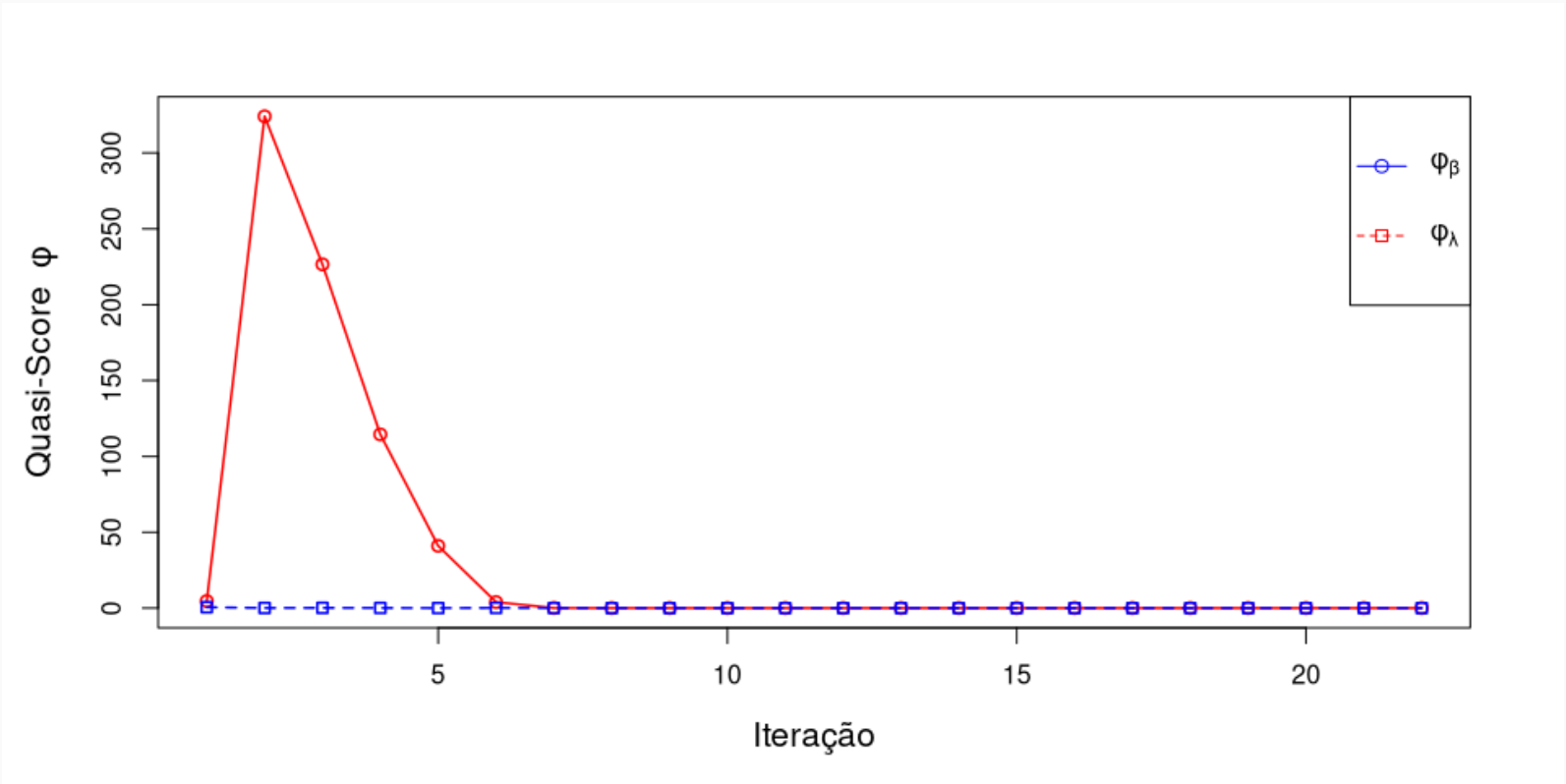

Figure 8: Valores de quasi-score ao longo das iterações do algoritmo Chaser no caso de *MLGDE* com correção de viés. Observe que φ_β é estável ao longo das iterações, enquanto φ_λ cresce e depois cresce e estabiliza. Este comportamento ocorre também para o caso de *MLGDE* sem correção.

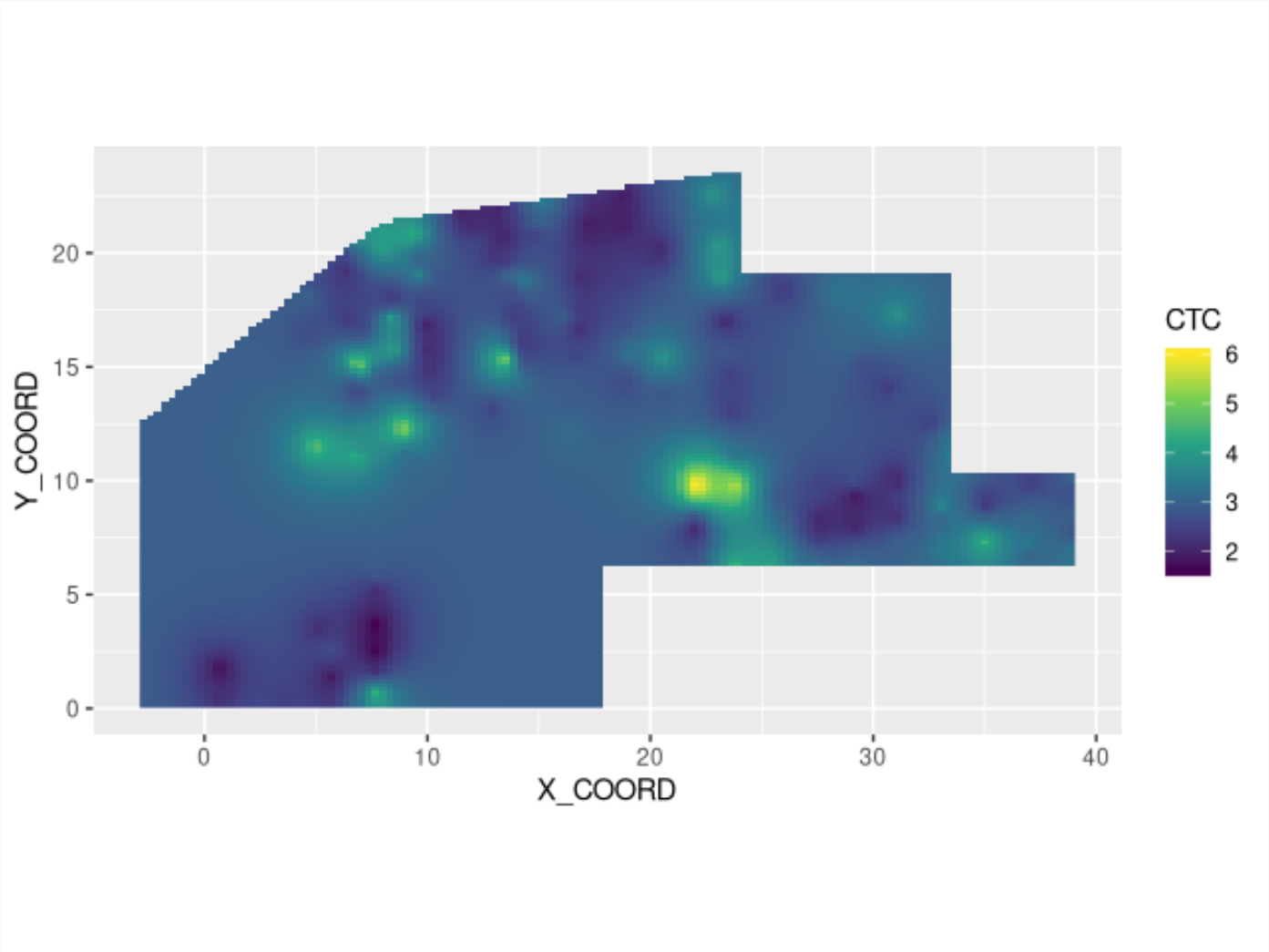

Figure 9: Mapa do *CTC* resultante dos valores preditos, a partir das estimativas da Tab. 5, isto é, com a correção de viés. O mapa predito a partir dos parâmetros obtidos com a estimação sem a correção de viés é muito similar a este apresentado.

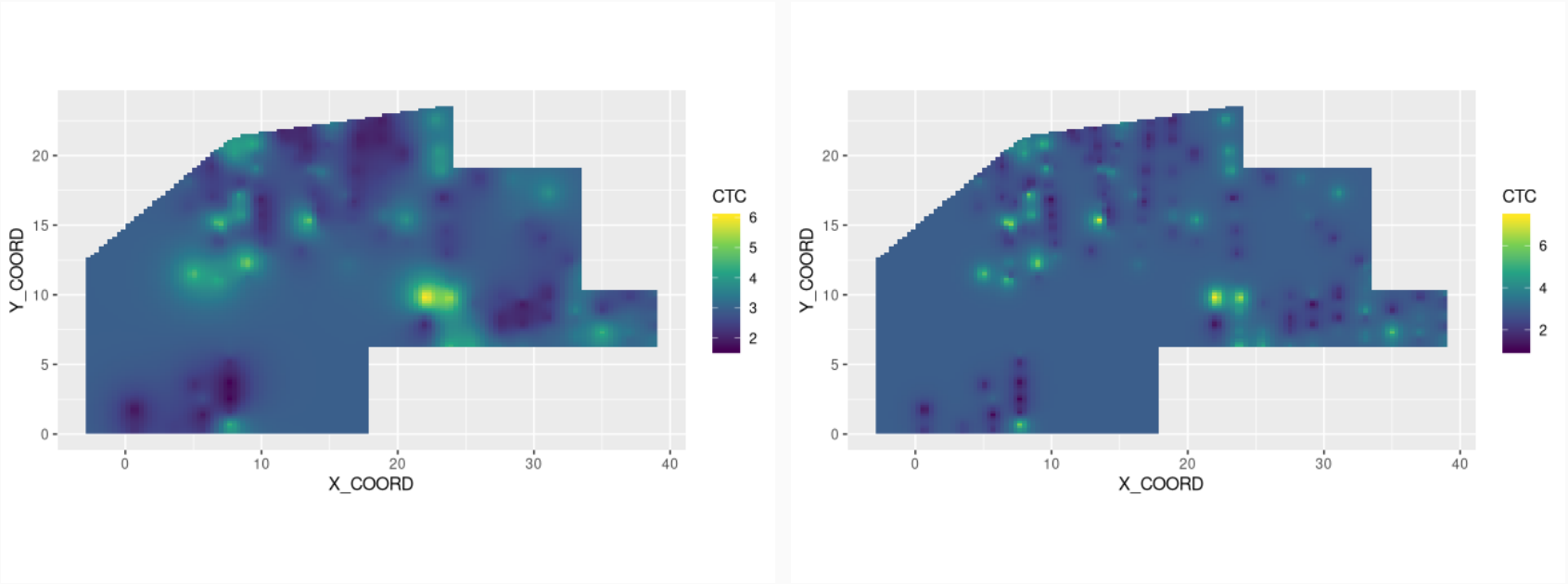

Figure 10: À esquerda, mapa do *CTC* predito com 3 parâmetros (τ_0, τ_1, τ_2) e à direita, mapa com 2 parâmetros (τ_0, τ_1).

Conclusão

Conclusão

- O objetivo da construção de *MLGDE* é a abrangência do modelo para lidar com dados independentes, gaussianos e, também não gaussianos.
- Além da abrangência, *MLGDE* se mostrou estável, preciso, além de eficiente.

Trabalhos futuros

- Análise de dados com dados binários,
- Testar outros métodos para fazer a estimação,
- Trabalhar no *MLGDE* caso multivariado.

References

[1] Wagner Hugo Bonat and Paulo Justiniano Ribeiro Jr. Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, 27(2):83–89, 2016.

[2] Peter J Diggle and Paulo Justiniano Ribeiro Jr. *Model-based Geostatistics*. Springer, New York, NY, 2007.

[3] Peter J Diggle, Rana Moyeed, and Jonathan A. Tawn. Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47:299–350, 1998.

[4] C A Gotway and W W Stroup. A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(2):157–178, 1997.

References ii

[5] René Holst and Bent Jørgensen. Generalized linear longitudinal mixed models with linear covariance structure and multiplicative random effects. *Chilean Journal of Statistics*, 6(1):15–36, 2015.

[6] Bent Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):127–162, 1987.

[7] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[8] Guido Masarotto and Cristiano Varin. Gaussian copula regression using r. *Journal of Statistical Software*, 77(8):1–26, 2017.

[9] Paulo Justiniano Ribeiro JR. Ctc dataset: http://www.leg.ufpr.br/geor/tutorials/da-tasets/ctc.dat, 2004. Accessed on 2019-11-28.

Obrigada!