Generalized Linear Models for Spatial Data (GLMSD)

Kally Chung (chung.kally@gmail.com)
Prof. Paulo Justiniano Ribeiro Jr (paulojus@ufpr.br)
Prof. Wagner Bonat (wbonat@ufpr.br)

February 13th, 2020

UFPR - PPGMNE/LEG
VIII WPSM

Table of contents

1. Introduction
2. General Model
3. Generalized Linear Models for Spatial Data
4. Data Analysis
5. Discussion

Introduction

Introduction

- Gaussian spatial data: continuous and symmetrically distributed.
    - Kriging,
    - Maximum likelihood estimation.
- Non-Gaussian spatial data: bynary data, count data, continuous with long tail, continuos and asymmetric, among others.
    - Spatial Generalized Linear Model (SGLM) - Gotway & Stroup, 1997 [4],
    - Generalized Linear Mixed Model (GLMM) - Bonat & RibeiroJr, 2015 [1],
    - Gaussian Copula Regression Model (GCRM) - Masarotto & Varin, 2017 [7].

General Model

General Model

- Model based on moments:

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

- $g(.)$ is the link function,
- μ is the mean vector,
- $V_{N \times N}$ is the variance matrix,
- $\Omega_{N \times N}$ is the covariance matrix.

Generalized Linear Models for Spatial Data

GLMSD

- Construction of $C = V^{1/2}\Omega V^{1/2}$:

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

- $v(p) = \mu^p$ is the variance function described by Tweedie family, (Jørgensen, 1987 [6]),
- $\tau = (\tau_0, \tau_1, \tau_2)$ is the sill $\tau_0 \geq 0$, range $\tau_1 \geq 0$ and nugget $\tau_2 \geq 0$ (Diggle & Ribeiro Jr (2007) [2]),
- $R(d_{ij}, \tau_1)$ is the matrix defined by spatial correlation function ρ.
    - Exponential or Matérn $\kappa = 0.5$:

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

    - For more spatial correlation functions, check Diggle & Ribeiro Jr (2007) [2].

Estimação (Holst & Jørgensen, 2015 [5])

- Let $\theta = (\beta, \lambda) = (\beta, p, \tau_0, \tau_1, \tau_2)$ be the parameters vector.
- Problem: solve

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

- $D = \nabla_\beta \mu$,
- $W_{\lambda_i} = C^{-1} \dfrac{\partial C}{\partial \lambda_i} C^{-1}$,
- residue $r = \mathbb{Y} - \mu$.

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

- Iterative method based on Fisher Scoring

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

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

Bias Correction (Holst & Jørgensen, 2015 [5])

- Quasi-score function $\varphi_{\lambda_i} = tr(W_{\lambda_i}(rr^T - C))$ is biased for unknown regression parameter β.
- The bias correction term is given by,

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

where

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

- Correction bias in φ_λ, we have

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

Reparametrization

- Note that

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

where

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

- then the reparametrization is given by

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

Initial Parameters

- For the initial β, we use the usual generalized linear model (*GLM*).
- For p, consider $p = 0$ for the continuous data, or $p = 1$ for the count data.
- Let γ_2 be 20% of the dispersion observed in *GLM*, empirically.
- We determine $\varphi_\lambda(\beta, \gamma_1)$ and estimate the initial parameter γ_1^{Inic} such that $\varphi_\lambda(\beta, \gamma_1^{\text{Inic}}) = 0$.
- Considering γ_1^{Inic}, $\hat{\gamma}_0$ is defined by

$$\hat{\gamma}_0 = \begin{cases} \ln\left(\dfrac{r^T \Delta^{-1} r}{N}\right), & \text{without bias correction} \\ \ln\left(\dfrac{r^T \Delta^{-1} r}{N - n_\beta}\right), & \text{otherwise} \end{cases} .$$

Standard Error in the Estimation

- Let $\hat{\theta} = (\hat{\beta}, \hat{\lambda})$ be the estimate of θ.
- The asymptotic distribution of $\hat{\theta}$ is given by

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

- $J_\theta^{-1} = S_\theta^{-1} V_\theta S_\theta^{-T}$,
- $S_\theta = \begin{bmatrix} S_\beta & S_{\beta,\lambda} \\ S_{\lambda,\beta} & S_\lambda \end{bmatrix} = \begin{bmatrix} E[\nabla_\beta \varphi_\beta(\beta,\lambda)] & E[\nabla_\lambda \varphi_\beta(\beta,\lambda)] \\ E[\nabla_\beta \varphi_\lambda(\beta,\lambda)] & E[\nabla_\lambda \varphi_\lambda(\beta,\lambda)] \end{bmatrix}$,
- $V_\theta = \begin{bmatrix} V_\beta & V_{\beta,\lambda} \\ V_{\lambda,\beta} & V_\lambda \end{bmatrix} = \begin{bmatrix} V_\beta & V_{\lambda,\beta}^T \\ V_{\lambda,\beta} & V_\lambda \end{bmatrix}$.
- Standard error SD_θ given by

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

Prediction (Gotway & Stroup [4])

- Let $\mathbb{Y} = (Y_1(s_1), Y_2(s_2), \ldots, Y_N(s_N))^T$ be the response variable of the observations at the locations $s_1, s_2, \ldots, s_N$.
- We want to predict the values for $\mathbb{Y}_l = (Y(l_1), Y(l_2), \ldots, Y(l_{n_u}))^T$ of n_u locations $l_1, l_2, \ldots, l_{n_u}$ not observed.
- For prediction, we use the krige estimator given by

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

- $Var \begin{bmatrix} \mathbb{Y} \\ \mathbb{Y}_l \end{bmatrix} = \begin{bmatrix} C_s & C_{s,l} \\ C_{l,s} & C_l \end{bmatrix} = \begin{bmatrix} C_s & C_{l,s}^T \\ C_{l,s} & C_l \end{bmatrix}$,
- $\hat{\mu}(s)$ and $\hat{\mu}(l)$ are the values predicted by the parameters β from regression.

Data Analysis

Dataset - Rongelap ([3])

- Cesium residual contamination measurements of nuclear tests at Rongelap Atoll on the Ralik Islands, part of the Marshall Islands, in Micronesia.

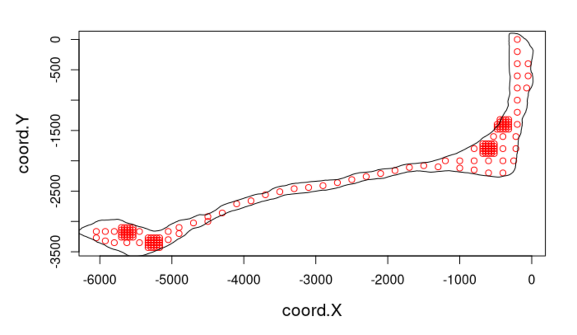


Figure 1: Mapping of the 157 locations of cesium residual measurements along the Atoll. Note that there are 4 regions in the island with a large amount of measurements.

Comparing *GLM* and *GLMSD*

- Dataset: Rongelap.
- Estimated parameters: $\theta = (\beta, \lambda) = (\beta_0, \tau_0)$ while the other dispersion parameters $p = 1, \tau_1 = 1, \tau_2 = 0$ is fixed.
- The power parameter $p = 1$ of the Tweedie family indicates the variance of the Poisson distribution.
- For $\tau_2 = 0$, we have the correlation matrix $\rho = I$, indicating that the data is independent.

Table 1: Estimates and standard errors of the parameters β_0 and τ_0 from *GLM* and *GLMSD* at the first two rows and the quasi-score values in the last two rows. The standard error of τ_0 is not informed from the summary of function ***glm*** of R.

	GLM		*GLMSD with corr.*	
	Estim.	Std.Error	Estim.	Std.Error
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			$9.516e-13$	

Reparametrization

- Dataset: Rongelap.
- We estimate the following parameters:
  $\theta = (\beta, \lambda) = (\beta, p, \tau_0, \tau_1, \tau_2)$.

Table 2: Estimates and standard errors of the parameters obtained from *GLMSD*, both with and without bias correction.

	GLMSD without corr.		*GLMSD with corr.*	
	Estim.	Std.Error	Estim.	Std.Error
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	

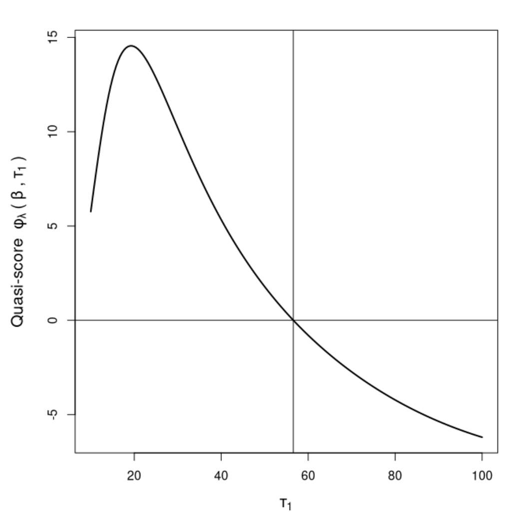


Figure 2: Function used to determine an initial τ_1 for the Rongelap dataset, without bias correction. We reinforce that the function is similar for *GLMSD* with bias correction.

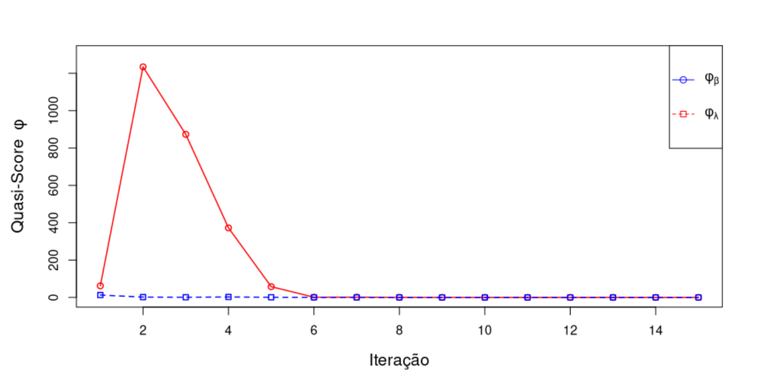


Figure 3: Quasi-score values along the iterations of the Chaser algorithm for *GLMSD* without bias correction. Notice that φ_β is stable along the iterations, while φ_λ increases, then decreases, and then stabilizies. This behaviour occurs for *GLMSD* with correction as well.

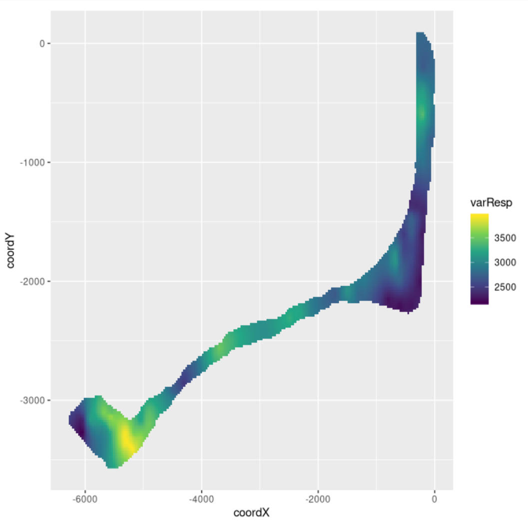


Figure 4: Rongelap Atoll map ilustrated by means of the predicted values, obtained from the estimation without bias correction. The predicted map with bias correction is similar to the one presented.

Dataset - CEC ([8])

- The Cátion exchange capacity (*CEC*) indicator is important because it measures the quality of soil and helps in the decision of which products to use in the soil before planting.

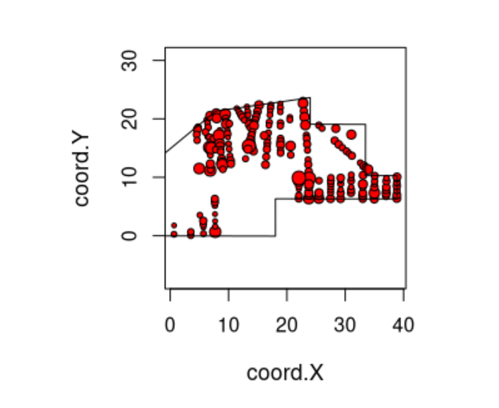


Figure 5: Map of the 212 locations at which *CEC* is measured. The circle radius reflect the indicator value.

Comparação da Verossimilhança com *GLMSD*

- Dataset: *CEC*.
- We have the parameters $\theta = (\beta, \lambda) = (\beta_0, \tau_0, \tau_1)$ with the other dispersion parameters $p = 0, \tau_2 = 0$ kept fixed.
- With $p = 0$, we use the Gaussian distribution variance, according to the Tweedie family.

Table 3: Estimates and standard errors of the parameters found by *GLMSD* without bias correction and the inference by means of a maximum likelihood estimation in the three first rows, in addition to the initial value of τ_1 found by the Chaser algorithm in the last row.

	MLE		*GLMSD without corr.*	
	Estim.	Std.Error	Estim.	Std.Error
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
Initial τ_1			0.4323	

Table 4: Estimates and standard errors of the parameters found by *GLMSD* and the inference by means of a restricted likelihood method in the first three rows, and the initial value of τ_1 found by the Chaser algorithm in the last row.

	Restricted Likelihood		*GLMSD with corr.*	
	Estim.	Std.Error	Estim.	Std.Error
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
Initial τ_1			0.4464	

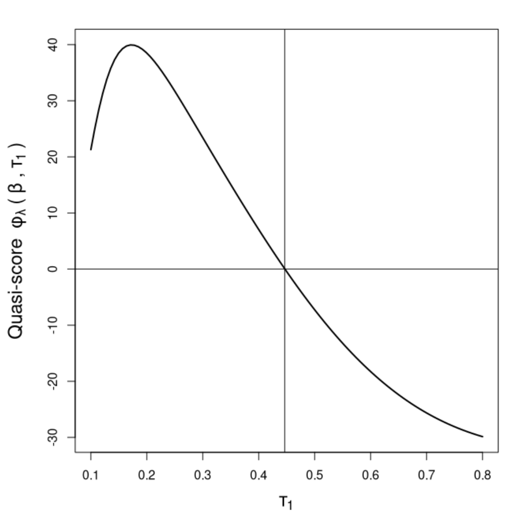


Figure 6: Function used to obtain the initial τ_1 inside the *GLMSD* with bias correction. This function doesn't vary much when there is no bias correction.

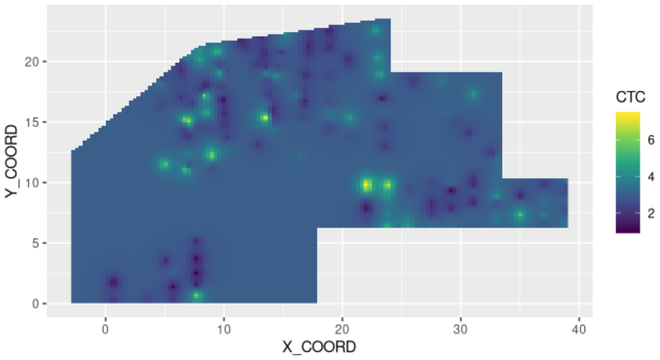


Figure 7: Map of the predicted *CEC* made with the estimates from Table 4, i.e., with bias correction. The predicted map without bias correction is very similar.

Reparametrization

- Dataset: *CCE*.
- We have the parameters $\theta = (\beta, \lambda) = (\beta_0, \tau_0, \tau_1, \tau_2)$ with the dispersion parameters $p = 0$ kept fixed.

Table 5: Estimates and standard errors of the parameters obtained from *GLMSD*, both with and without bias correction.

	GLMSD without corr.		*GLMSD with corr.*	
	Estim.	Std.Error	Estim.	Std.Error
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
Initial τ_1	0.7364		0.7646	

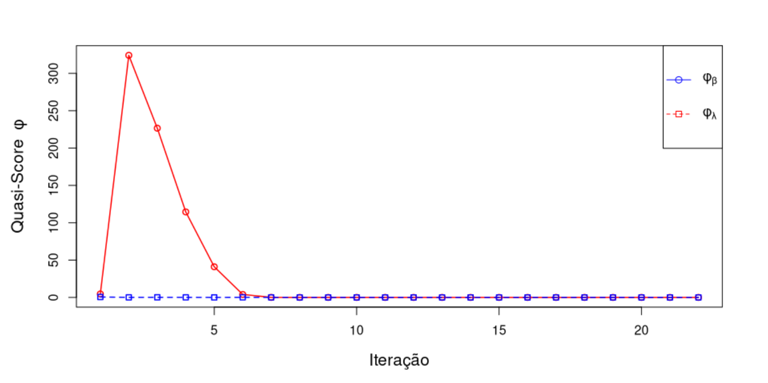


Figure 8: Quasi-score values along the Chaser algorithm iteration in the case of *GLMSD* with bias correction. Notice that φ_β is stable along the iteration, while φ_λ increases, then decreases, and then stabilizies. This behaviour occurs for *GLMSD* without correction as well.

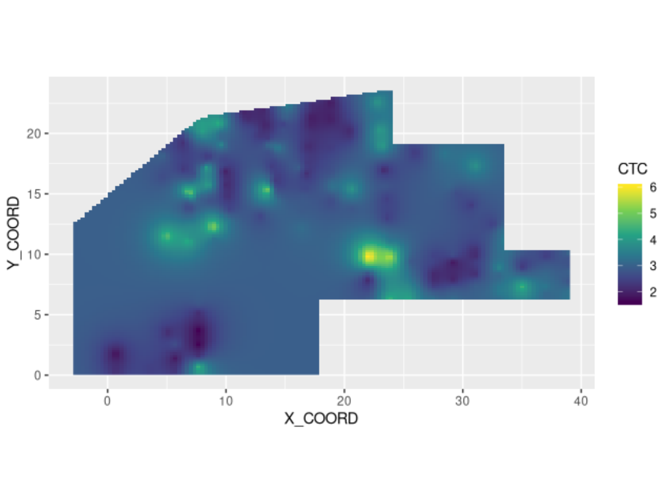
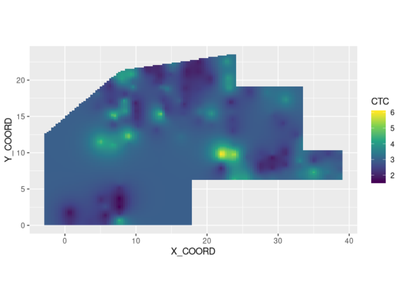


Figure 9: *CEC* map predited by the estimates of Tab. 5 with bias correction. The map predicted with estimates without bias correction in very similar to this presented map.



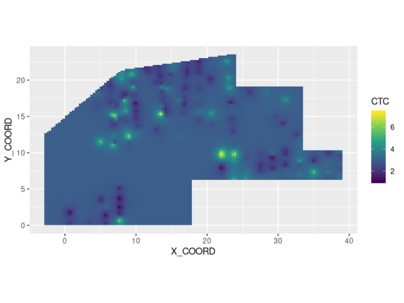


Figure 10: The *CEC* left map was predicted considering 3 dispersion parameters (τ_0, τ_1, τ_2) while right map is predicted by 2 dispersion parameters (τ_0, τ_1).

Discussion

Discussion

- We provide a method that handles wild variety of types response, independent, Gaussian and also non-Gaussian.
- Furthermore, *GLMSD* is stable, precise and efficient.

Future Work

- Data analyses with binary data,
- Consider another methods for the estimation,
- Working on multivariate *GLMSD*.

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.

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

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

Thank you!