Robust predictive process approximation of spatial fields

Mariana del Pilar Lizarazo Osorio [1] and Thaís C. O. Fonseca[2]

Abstract

In this work we propose a robust approximation for spatial random fields. The method is based on a scale mixture of Gaussian processes and attains scalability for large datasets by projecting the original process into a lower dimensional space through knot based techniques. In particular, this paper considers the predictive approach to approximate a parent process. This proposal allows for more reliable predictions and, as a result, better representations of the parent process based on predictive process approximations. The predictive approach is based on krigging ideas and krigging predictors are well known to be affected by outliers as they are obtained as linear combination of observations. As a result, it is expected that the predictive process approximation might be affected by outliers and in this case, the approximation might not be a good representation of the original parent process. In this paper we propose a remedy to this issue by allowing the predictive process to be more robust to outlying observations. Furthermore, the predictive process formulation depend on the choice of knots which proved to be of major importance in the estimation of parameters of interest. We investigate the effects of assuming a particular set of knots for the predictive approximation. In addition, we investigate how these choices might influenciate outlier identification.

Keywords: Large data; Spatial statistics; Predictive processes; Nonstationarity.

[1]Department of Statistics, Universidade Federal do Rio de Janeiro

[2](to whom correspondence should be addressed) Department of Statistics, Universidade Federal do Rio de Janeiro, Ilha do Fundão, thais@im.ufrj.br

1 Introduction

With the increase of high-resolution geocoded data, the big n problem became crucial in the spatial and spatiotemporal setup. For instance, if Gaussianity is assumed, large covariance matrices need to be inverted in the inference procedure and computational effort is of cubic order on the number of locations. This limitation become even more important in the case of spatiotemporal or multivariate data. Few time records over space may lead to huge matrices of covariance, making the inference for unknown parameters not feasible. Thus, a compromise between complexity and parsimony is called for in this context.

Several approaches were proposed to overcome the computational limitation imposed by the Gaussian assumption. Vecchia (1988) and Stein et al. (2004) proposes the use of conditional distributions where $\mathbf{Z}$, the vector of observations in n locations, is partitioned into subvectors $\mathbf{Z}_1, ..., \mathbf{Z}_b$ of possibly different lengths, then $p(\mathbf{z}|\boldsymbol{\phi}) = p(\mathbf{z}_1|\boldsymbol{\phi}) \prod_{j=1}^{b} p(\mathbf{z}_j|\mathbf{z}_{(j-1)}, \boldsymbol{\phi})$, where $\mathbf{Z}_{(j)} = (\mathbf{Z}_1, ..., \mathbf{Z}_j)$. In particular, if $\mathbf{S}_{(j)}$ is some subvector of $\mathbf{Z}_{(j)}$ then $p(\mathbf{z}|\boldsymbol{\phi}) \approx p(\mathbf{z}_1|\boldsymbol{\phi}) \prod_{j=1}^{b} p(\mathbf{z}_j|\mathbf{s}_{(j-1)}, \boldsymbol{\phi})$. To use this approximation, one need to order the observations in some manner, and it is suggested to use the rank of the projections of the observations along some axis. Jones and Zhang (1997) use this approach to space-time processes and they suggest defining the nearest neighbors based in a preliminary estimate of the space-time correlation. More recently, Datta et al. (2015) use this ideia to define a new spatial processes called the Nearest Neighbor Gaussian Process which is defined over a fix reference set, using a parent process. In this set-up it is necessary to define a order for the spatial locations and a finite reference set. The authors suggest the ordering as presented in Vecchia (1988). Besides, they consider two reference sets: the observational and a randomly placed locations.

Furrer et al. (2006) proposed the tapering approach which sets the covariance function to zero beyond a certain range. The new covariance function is defined by the product of two functions, the original covariance and the taper function. The selected taper function should be exactaly zero from a certain range and it should preserves the original behaviour

at the origin. The authors analyzed the effect of tapering using the infill asymptotic theory which is based on the number of locations increasing within the fixed spatial domain. The proposal considers $K_{tap}(x, y) = K(x, y)K_\theta(x, y)$ and the Schur product K_{tap} preserves some of the shape of K and it is identically zero outside a fixed range. A possible objection to tapering is that it may not be effective for a spatial covariance with long range correlations.

Fuentes (2007) also presents an approximation for the likelihood of Gaussian processes but her approach is restricted to the spectral domain. The spectral density for the corresponding lattice process is written as a sum with infinite terms which is truncated after 2N terms achieving dimension reduction.

A different approach considers low-rank models which achieve the computational feasibility by writing the spatial component as a linear combination of spatial basis functions. The models differ in the parametrization and the basis function used. For example, commom choices are the Fourier basis function and bisquere functions. A particular case of these models is the Gaussian Predictive model (Banerjee et al., 2008), in this case the basis function is parameterized according to a parent process. Gaussian predictive processes represent the original process in a lower dimension in which inference is feasible. The method considers the projection of the original process onto a space defined by a set of locations called knots. Thus, the original big n problem is transformed in a problem of dimension m with $m << n$. In general terms, a new process is defined as a linear combination of observations at a fixed set of knots. In particular, the weights are defined by the kriging interpolation based on the points in the set of knots. Thus, this interpolation defines a new process for each location in the spatial domain of interest.

However, notice that predictions are usually highly affected by outlying observations. An outlier may have a strong effect in the prediction of its neighboors when the observed value for the process at this location is much higher or lower than expected for that region in space. Chilès and Delfiner (1999) comment that, in applied settings, even small changes in some regions in space might cause large differences between the predicted and observed

process. Observations in these regions should not be discarted as this might cause bias in the estimation of parameters and predictions (Chilès and Delfiner, 1999, page 221). In that context, it is expected that predictive processes as defined by Banerjee et al. (2008) will be highly affected by outliers. In particular, the predictive process are based upon linear combination of observations at a set of knots and linear combinations are well known to be highly affected by outliers.

In fact, the traditional kriging predictor is well known to be affected by outliers and several papers have proposed robust alternatives or modifications of usual kriging predictor. Fournier and Furrer (2005) proposed a model to robustify the kriging predictor by defining the spatial process as a mixture of a spatial process and a contamination process. In this proposal each site has a correponding contamination variable which indicated whether the site was contaminated or not. The optimal predictor in this case depends on weights which will be affected by the contamination variables. However, the predictor is unfiasible in practice and an approximation is considered.

In that context, we propose to extend the predictive approach of Banerjee et al. (2008) to robust settings. As a solution to the high influence of outlying observations in the approximation of large Gaussian processes we propose the use of heavy tailed processes which accommodate extreme observations in the sampling distribution as proposed by Palacios and Steel (2006) and extended by Fonseca and Steel (2011). These proposals are based on mixing a Gaussian process with a positive process allowing for heavier tails for the finite dimensional distributions of spatial data. This will allow for more reliable predictions and as a result better representations of the parent process based on predictive process approximations. Notice that this issue might be even more crucial in the case of large domains in which it is expected to find different behaviours across space. In such a case, a more flexible class allowing for spatial heterogeneity is called for.

In this work we follow the predictive approach in the non-Gaussian modeling of georeferenced data. We consider robust versions of predictive processes in order to obtain models

which are scalable to potentialy high dimensional spatial data and able to accommodate heterocedasticity in space and outliers.

Furthermore, the predictive process formulation depend on the choice of knots which proved to be of major importance in the estimation of parameters of interest. We investigate the effects of assuming a particular set of knots for the predictive approximation. Furthermore, the corrections proposed in the literature to improve variance estimation might lead to poor estimation of spatial correlation parameters. We also investigate this issue and discuss how outlier indentification in the non-Gaussian setup is affected by the choice of such knot sets and corrections.

In section 2 we describe the predictive approach to spatial process modeling as introduced by Banerjee et al. (2008), which is based on Gaussian processes. In section 3 we introduce the predictive approach based on non-Gaussian processes proposed in this work. Section 4 presents the design problem of choosing the knot set in the definition of a new process based on knot techniques. Further, we discuss the outlier detection procedure considered and how this might be related to the knot set choice. Section 5 presents a simulation experiment for Gaussian and non-Gaussian processes to investigate the effect of the design selected in the parameter estimation. In addition, we present results regarding the rate of outlier detection proposed for different designs for the knot set. Section 6 presents conclusions and future developments.

2 Predictive process modeling

Gaussian Predictive process modeling has the property of reducing the dimension of a original parent process from n to m, where $m << n$. The method is based on the ideas of kriging and projects the original data in a lower dimensional space defined by a set of knots. As follows we present the model, main limitations and remedies already proposed in the literature.

2.1 Predictive Gaussian process

Consider a stationary Gaussian process $\{\omega(s) : s \in D\}$ with mean function $\mu(s) = 0$ and covariance function $Cov(\omega(s_i), \omega(s_j)) = C(||s_i - s_j||)$ observed at locations $S = \{s_1, \ldots, s_n\}$. Thus, $\boldsymbol{\omega} = (\omega(s_1), \ldots, \omega(s_n))$ is such that $\boldsymbol{\omega} \sim N(\mu, \Sigma)$, $\mu_i = 0$ and $\Sigma_{ij} = C(||s_i - s_j||)$. Let $S^* = \{s_1^*, \ldots, s_m^*\}$, $m << n$ be a set of knots in D, which might or might not be a subset of S. Thus, from the parent process definition we obtain that $\boldsymbol{\omega}^* = (\omega(s_1^*), \ldots, \omega(s_m^*))^T \sim N_m(0, \boldsymbol{\Sigma}^*)$, with $\boldsymbol{\Sigma}_{ij}^* = C(||s_i^* - s_j^*||)$, $i, j = 1, \ldots, m$. Conditional on $\boldsymbol{\omega}^*$, the best linear predictor at a location s_0 is given by $E(\omega(s_0) \mid \boldsymbol{\omega}^*)$ which is

$$\tilde{\omega}(s_0) = c^T(s_0)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\omega}^*, \tag{1}$$

where $c^T(s_0) = [C(||s_0 - s_1^*||), \ldots, C(||s_0 - s_m^*||)]$. This interpolation is used by Banerjee et al. (2008) to define a new process called the predictive process. Let $\{\tilde{\omega}(s) : s \in D\}$ be a Gaussian process with 0 mean function and covariance function $\tilde{C}(s, s') = c^T(s)\boldsymbol{\Sigma}^{*-1}c(s')$. The predictive process representation for a variable of interest Z in a location s is given by

$$Z(s) = \mathbf{x}^T(s)\boldsymbol{\beta} + \tilde{\omega}(s) + \epsilon(s), \tag{2}$$

with covariance function $C(s, s') = c^T(s)\boldsymbol{\Sigma}^{*-1}c(s') + \tau^2 1_{s=s'}$, where $1_{s=s'}$ represents the indicator function, thats equal to 1 if $s = s'$. The process $\tilde{\omega}(s)$ is an orthogonal projection of $\omega(s)$ and is the best representation of the parent process (see Banerjee et al., 2008). Thus, for observed locations in $S = \{s_1, \ldots s_n\}$ the vector $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}(s_1), \ldots, \tilde{\omega}(s_n))$ is predicted by $\tilde{\boldsymbol{\omega}} = \mathbf{c}^T\boldsymbol{\Sigma}^{*-1}\boldsymbol{\omega}^*$, where $\mathbf{c}^T = [c^T(s_i)]_{i=1}^n$. The vector $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))$ is modeled as

$$\mathbf{Z} \mid \tilde{\boldsymbol{\omega}}, \boldsymbol{\beta}, \tau \sim N_n(\mathbf{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\omega}}, \tau^2 I_n) \tag{3}$$

$$\boldsymbol{\omega}^* \mid \boldsymbol{\theta} \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}^*). \tag{4}$$

The big n problem reduces to a problem of dimension m, that is, inverses and determinants required for likelihood computations and preditions depend on matrices of size m instead of n, by using the Sherman-Morrison-Woodbury identity. Notice that the model for

Z is different from the parent process and as a result the inferences for the model parameters might be different. For instance, the variance for the process $w(s)$ is different from the process $\tilde{w}(s)$. In the parent process the spatial variance is $Var(w(s)) = C(0)$ while in the predictive process

$$Var(\tilde{w}(s)) = c^T(s)\mathbf{\Sigma}^{*-1}c(s) \tag{5}$$

Finley et al. (2009) proposed a solution to the bias problem in the variance estimation by modifying the predictive process to have the same variance as the original process. Furthermore, they propose a method to select knots which we investigate in more detail in our simulated study. The authors propose the process

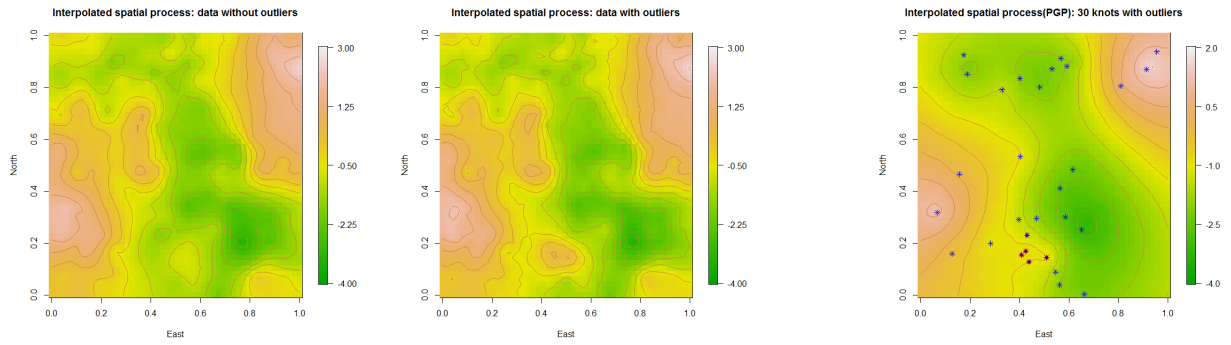
$$\ddot{w}(s) = \tilde{w}(s) + \tilde{\epsilon}(s), \tag{6}$$

which is called modified predictive process. The error term $\tilde{\epsilon}(s) \sim N(0, C(0) - c^T(s)\mathbf{\Sigma}^{*-1}c(s))$ and the variance for $\ddot{w}(s)$ is the same as the original process $w(s)$. In this work we investigate how this correction in the variance might affect estimation of other model parameters.

3 Nongaussian predictive process modeling

Usual kriging estimators are sensitive to aberrant observations in the data as they are obtained as a linear combination of sampled observations. Similarly, spatial predictions based on Gaussian processes may be highly affected by outliers. Thus, it is expected that the predictive process defined in (2) is potentially dependent on the presence of outliers in the data. Notice that if the outlier location is in S^* the whole process approximation will be affected by the outlying observation and the predictive approximation might not be a good representation of the original parent process. Figure 3 presents a simulated example in which the original process (Figure 3 (a)) is contaminated by outliers (Figure 3 (b)) and the predicted surface has a different pattern than the original parent process due to presence of outliers in the set of knots. In this example the low values in the bottom part of the region is not captured by the approximation, which smooths the low values and predicts only peaks in

that region. Examples of this kind motivates the idea that outlying observations should be considered in the sampling distribution and robust solutions should be proposed to best represent the original process. In this section we present a robust alternative to the predictive Gaussian approximation.



(a) Original surface. (b) Contaminated surface. (c) Gaussian predictive approximation.

Figure 1: Original process, interpolated surface for the data contaminated with outliers and predictive approximation based on a Gaussian process for the contaminated data.

3.1 Nongaussian model

The usual assumption in geostatistics is that the finite-dimensional distribution for n observations $Z(s_1), \ldots, Z(s_n)$ is multivariate Gaussian. However, Gaussian processes are very sensitive to extreme or outlier observations. In that context, Palacios and Steel (2006) proposed heavy tailed processes based on mixing Gaussian process with a positive process responsible for inflating the variance and accommodating extreme observations. Consider the spatial process

$$Z(s) = \mathbf{x}^T(s)\boldsymbol{\beta} + \frac{w(s)}{\lambda(s)^{1/2}} + \epsilon(s), \tag{7}$$

where $\omega(s)$ is a Gaussian process defined in $s \in D$, and it is independent of $\epsilon(s) \sim N(0, \tau^2)$. The process $\lambda(s)$ is the mixing process allowing for spatial heterogeneity. In matricial form

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2 I_n) \tag{8}$$

with $\mathbf{\Lambda} = diag(\boldsymbol{\lambda}), \quad \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. Integrating $\boldsymbol{\lambda}$, the finito dimensional distribution of $\mathbf{Z}$ has heavier tails than the Gaussian distribution. In order to obtain mean squared continuity the variable λ need to be spatially correlated (ver Palacios and Steel, 2006). Thus, the mixing process is modeled as

$$\mathbf{Z} \mid \boldsymbol{\omega}, \mathbf{\Lambda}, \boldsymbol{\beta}, \tau^2 \sim N_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{\Lambda}^{-1/2}\boldsymbol{\omega}, \tau^2 I_n) \tag{9}$$

$$\boldsymbol{\omega} \mid \sigma^2, a \sim N_n(\mathbf{0}, \mathbf{\Sigma}) \tag{10}$$

$$ln(\boldsymbol{\lambda}) \mid a, \nu \sim N_n\left(-\frac{\nu}{2}1_n, \nu\mathbf{R}\right) \tag{11}$$

with $\mathbf{\Sigma} = [C(s_i, s_j)]_{i,j=1}^n = \sigma^2\mathbf{R}$, $R_{ij} = Cor(Z(s_i), Z(s_j))$, for $i, j = 1, \ldots, n$ e 1_n is a unitary vector of size n and $\nu \in \mathbb{R}_+$. In this set-up the covariance between two points in space is

$$Cov(Z_i, Z_j) = \sigma^2 R_{ij} exp\{\nu(1 + (1/4)[R_{ij} - 1])\} \tag{12}$$

Let $\mathbf{Z} = (\mathbf{z}_o^T, \mathbf{z}_p^T)^T$, with $\mathbf{z}_o^T$the observations of the process Z and $\mathbf{z}_p^T$ the desired predictions in r non-observed locations. The predictive distribution is given by

$$p(\mathbf{z}_p \mid \mathbf{z}_o) = \int p(\mathbf{z}_p \mid \mathbf{z}_o, \lambda, \zeta) p(\lambda_p \mid \lambda_o, \zeta, \mathbf{z}_o) p(\lambda_o, \boldsymbol{\theta} \mid \mathbf{z}_o) d\lambda d\zeta \tag{13}$$

with $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_o^T, \boldsymbol{\lambda}_p^T)^T$, and $\zeta = (\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\theta}, \nu)$. And predictions might be obtained from the joint distribution of $(\mathbf{z}_p, \boldsymbol{\lambda}_p)$

$$ln(\boldsymbol{\lambda}_p) \mid \boldsymbol{\lambda}_o, \nu, \mathbf{z}_o \sim N_r\left(\mathbf{C}_{po}\mathbf{C}_{oo}^{-1}(ln\lambda_o + \frac{\nu}{2}1_n) - \frac{\nu}{2}1_r, v[\mathbf{C}_{po} - \mathbf{C}_{po}\mathbf{C}_{oo}^{-1}\mathbf{C}_{op}]\right) \tag{14}$$

$$\mathbf{z}_p \mid \mathbf{z}_o, \lambda, \zeta \sim N_r\left((\mathbf{X}_p - A\mathbf{X}_o)\boldsymbol{\beta} + \mathbf{A}\mathbf{z}_o, \sigma^2\left(\Lambda_p^{-\frac{1}{2}}\mathbf{C}_{pp}\Lambda_p^{-\frac{1}{2}} + \frac{\tau^2}{\sigma^2}\mathbf{I}_r - \Lambda_o^{-\frac{1}{2}}\mathbf{C}_{op}\Lambda_p^{-\frac{1}{2}}\right)\right) \tag{15}$$

with $A = \Lambda_p^{-\frac{1}{2}}\mathbf{C}_{po}\Lambda_o^{-\frac{1}{2}}\left[\Lambda_o^{-\frac{1}{2}}\mathbf{C}_{oo}\Lambda_o^{-\frac{1}{2}} + \frac{\tau^2}{\sigma^2}\mathbf{I}_n\right]^{-1}$ and $\mathbf{C}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{C}_{oo} & \mathbf{C}_{op} \\ \mathbf{C}_{po} & \mathbf{C}_{pp} \end{pmatrix}$.

3.2 Nongaussian Predictive process

Consider a set of knots $S^* = \{s_1^*, \ldots, s_m^*\}$, which might be or not in the sampled locations $S \in D$, and the non-Gaussian process (7). Let ω^* be predictions for the process $\omega(s)$ at

location in S^*. Also let λ^* be predictions for the variance process $\lambda(s)$ at location in S^*.

$$\boldsymbol{\omega}^* \mid \sigma^2, \boldsymbol{\theta} \sim N_m(\mathbf{0}, \sigma^2 \mathbf{R}^*) \tag{16}$$

$$ln(\lambda^*) \mid \boldsymbol{\theta}, \nu \sim N_m\left(-\frac{\nu}{2}1_m, \nu \mathbf{R}^*\right) \tag{17}$$

with, $\mathbf{R}^* = [Cor(s_i^*, s_j^*; \boldsymbol{\theta})]_{i,j=1}^m = \sigma^{-2}\mathbf{C}^*(\boldsymbol{\theta})$ and $R^T(s) = [Cor(s, s_1^*; \boldsymbol{\theta}), \ldots, Cor(s, s_m^*; \boldsymbol{\theta})]$. Thus, the predictive non-Gaussian process is defined as

$$Z(s) = \mathbf{x}^T(s)\boldsymbol{\beta} + \frac{\tilde{\omega}(s)}{\tilde{\lambda}^{1/2}(s)} + \epsilon(s), \tag{18}$$

with,

$$\begin{aligned} \tilde{\omega}(s) &= \mathbf{c}^T(s)\mathbf{C}^{*-1}\omega^* \\ ln(\tilde{\lambda}(s)) &= \frac{\nu}{2}\left[R^T(s)\mathbf{R}^{*-1}(\boldsymbol{\theta})1_m - 1_n\right] + R^T(s)\mathbf{R}^{*-1}(\boldsymbol{\theta})ln(\lambda^*) \end{aligned}$$

Proposition 3.1 *Consider the non-Gaussian predictive process as defined in (18). Thus, the covariance for the predictive processes $\tilde{\omega}$ and $\tilde{\lambda}$ are respectively given by*

$$Cov(\tilde{\omega}(s), \tilde{\omega}(s^{'})) = \mathbf{c}^T(s; \boldsymbol{\theta})\mathbf{C}^{*-1}(\boldsymbol{\theta})\mathbf{c}(s^{'}; \boldsymbol{\theta}) \tag{19}$$

$$Cov(ln(\tilde{\lambda}(s)), ln(\tilde{\lambda}(s^{'}))) = \nu R^T(s)\mathbf{R}^{*-1}R(s^{'}) \tag{20}$$

The model (18) is a predictive process and, as in the Gaussian case, induces a bias in the estimation of the marginal variance τ^2. Thus, we propose a correction for the variance as done in the Gaussian case (Finley et al., 2009).

Proposition 3.2 *The variance correction for $\tilde{\omega}\tilde{\lambda}^{-1/2}$ is*

$$\ddot{\omega}_i\ddot{\lambda}_i^{-1/2} = \tilde{\omega}_i\tilde{\lambda}_i^{-1/2} + \zeta_i, \quad i = 1, \ldots, n \tag{21}$$

onde $\zeta_i \sim N(0, \sigma_\zeta^2)$, com

$$\sigma_\zeta^2 = \sigma^2\left(exp\left\{\nu\right\} - R^T(s_i, s^*)\mathbf{R}^{*-1}R(s_i, s^*)exp\left\{\frac{\nu}{2}[1 + R^T(s_i, s^*)\mathbf{R}^{*-1}R(s_i, s^*)]\right\}\right)$$

4 Design for non-Gaussian modeling

The selection of points to be included in the set S^* is a crucial problem in the definition of the predictive process $w^*(s)$ as few points or points too close together might not represent the spatial behaviour in space. On the other hand, the selection of a too large subset S^* will result in a too large covariance matrix to be inverted.

This problem is known as a design problem. The points selected should not affect the estimation of parameters. In that context, a chosen set which is eficient for prediction might not be for parameter estimation (see Zimmerman, 2006). Several papers indicate ways to select these points. The methods are usually based on the minimization of a loss criteria, depending on the goals to be achieved in the research. For instance, Xia et al. (2006) proposed a method based on the maximization of the determinant of the Fisher information matrix. Fuentes (2007) considers entropy maximization in a Bayesian context in order to maximize the information obtained from data. Cressie (1993) uses the minimization of prediction variances to increase or decrease the size of a grid with focus on improving the prediction of new data.

In this work we are also interested in outlier detection. For this purpose, biased sampling in regions with larger variability might induce poor prediction with very large credibility intervals. While biased sampling in regions with lower variability might induce predictions with unrealistic small prediction intervals.

As follows we present a brief review of known methods of knots selection and present a proposal for the non-Gaussian geostatistic modeling.

4.1 Knots selection

Assume the number of knots is fixed and will be selected in $D \in \Re^d$. Some possible sampling choices take into consideration the spatial structure.

1. *Random grid:* This grid is a simple choice which does not consider any spatial information. In this case, all points have the same probability of being included in the knots set.

2. Regular grid: It is frequently the grid choice due to its simplicity. However, it is uneficient for some parameter estimation such as smoothness parameters as it does not allow observations very close together. An example of regular grid is the regular triangular grid (see Cressie, 1993, page 318).

3. *Finley's proposal:* Finley et al. (2009) proposed a minimization criteria for the predictive variance in predictive processes. The authors consider m to be known and construct the knot set to be as close as possible to the original process. Let $\boldsymbol{\theta}$ be the predictive variance of $w(s)$ conditional on the predictive process w^* in spatial points s^* which is given by

$$V_{\boldsymbol{\theta}}(s, s^*) = Var(w(s) \mid w^*(\cdot), s^*, \boldsymbol{\theta}) = C(s, s) - c^T(s; \boldsymbol{\theta}) C^{*-1}(\boldsymbol{\theta}) c(s^{'}; \boldsymbol{\theta}), \tag{22}$$

and measures how well $\tilde{w}(s)$ approximates $w(s)$. Based on this idea the authors proposed an algorithm to minimize the mean predictive variance for the observed locations by minimizing

$$V_{\boldsymbol{\theta}}(s^*) = \frac{1}{n} \sum_{i=1}^{n} Var(w(s_i) \mid w^*(\cdot), s^*, \boldsymbol{\theta}) \tag{23}$$

Thus the algorithm to select m knots is

   (a) Specify a set of all the possible locations that could be selected $S = s_1, \ldots s_N$ with $N > m$. Some examples of the possible choices for the set S are a regular grid or the observed locations, and others.

   (b) Specify an initial set with n_0 components. This set, could be randomly or deterministicly chosen.

   (c) In the step $t + 1$ compute $V_{\boldsymbol{\theta}}(s_i, s^*)$ for all $s_i \in S$. The point which achieves the minimum variance is included in the knots set.

(d) Repeat the the last item up to the selection of m points in the knot set.

4. *Diggle's proposal:* Diggle and Lophaven (2006) presents two different views of this problem, one is called retrospective and the other prospective. In the retrosprospective method assume that some data is available for parameter estimation. In a second step knots are chosen to minimize the predictive variance conditional on the estimated parameters. In the prospective case, the proposal considers the expected predictive variance to selec knots. They unite both goals of prediction and parameter estimation by proposing a modified regular grid.

   a) lattice plus close pairs $(k \times k, m, \alpha)$**:** This grid is regular with dimension $k \times k$ and f knots randomly selected in a circle of radius α centered in the f randomly chosen knots.

   b) lattice plus in-fill $(K \times k, m, k_0 \times k_0)$**:** Define a completely regular grid of dimension $k \times k$. Then m cells are randomly selected in the grid and in the selected cells a new grid is created with dimension $k_0 \times k_0$.

   According to the authors, the most effiecient grid is the *lattice plus in-fill*.

4.2 Knots selection for non-Gaussian models

In the non-Gaussian model (7) we need to specify the knots for the spatial process $\omega(s)$ and also the knots for the variance process $\lambda(s)$. We consider three approaches to non-Gaussian modeling.

i) Use a set of knots for ω and a different set for the λ process. This could include more information into the model. For example, if we use 40 knots, we could use two differents sets of dimention 40, one for each process, as well as use a different selection criteria of knots for the two processes. Note that the computational cost of this proposal is practically the same as that of using only a set of 40 knots.

ii) Since our interest is to model processes that have regions with high variability, we could think that a good design for the λ process should include a larger quantity of points in the region with more variance, to do so we could use a pilot sample to estimate the parameters of the model and select the locals with higher values of $\frac{\sigma^2}{\lambda_i}$, $i \in \{1, \ldots, n\}$.

iii) The idea in the item (ii) could have some problems, because selecting only points with large variability could make all the chosen points form a cluster, what harders or makes it impossible to apply technique such as block sampling. In order to solve this difficulty and allow the knots to consider not only locations with higher variability it is possible to use values with higher variance but also using point with lower variability. In this work the amount of points of higher variance will be equal to that of lower variance points.

5 Applications

5.1 Simulation experiment: Gaussian data

In this simulated study we consider 100 replicates generated from a spatial process with size $n = 100$ each. We intend to investigate the quality of parameter estimation when we consider different grid types (random, regular, Finley and Diggle). Furthermore, we also considered two versions of the model with corrected or not corrected marginal variance. Data was simulated from a stationary Gaussian process $\{Z(s) : s \in D\}$ with mean function $\mu(s) = \beta_1 x_1(s) + \beta_2 x_2(s)$ where covariates generated from $x_1 \sim N(3, 1)$, $x_2 \sim N(10, 2)$, and parameters $\beta_1 = 5$ and $\beta_2 = 3$. The covariance function considered was $C(s_i, s_j) = \sigma^2 exp\left(-\frac{||s_i - s_j||}{a}\right)$, with $\sigma^2 = 2$, $a = 2$. For all locations we considered $\tau^2 = 0.5$. The locations were randomly selected in a 10×10 square.

For each Monte Carlo replication, the parameters were estimated from a Bayesian point of view using the original Gaussian model and the Banerjee et al. (2008) proposal with 20, 40 and 60 knots. The knot selection procedures were the random, Finley's and Diggle's.

The prior distributions for the parameters of interest are $\beta \sim N(0, \Sigma_\beta)$, $\sigma^2 \sim GI(\alpha, \delta)$, $\tau^2 \sim GI(\gamma, \xi)$ and $a \sim Ga\,(1, \phi/med(d))$, with $med(d)$ the distance median. The prior hyperparameters where chosen to result in vague prior distributions. The estimates were obtained using MCMC method and 200000 iterations, and convergence was verified with Geweke and Raftery and Lewis criteria (Raftery and Lewis, 1992).

Table 1 presents the relative Mean Squared Error (MSE) for the model parameters computed over the 100 replicates. To have a clearer presentation, we compared the estimated parameters with the estimates obtained for the Gaussian model using the complete dataset, which would be the best estimates we could obtain from data instead of using the true parameter values. From table 1 it is evident that the largest bias created by the predictive approximation are observed for the variance parameters σ^2 and τ^2 and the spatial range a. In addition, the percentual gain by increasing the number of knots from 20 to 60 is very small for both parameters. The computational time gain obtained for 20 knots is significant as the complete data takes 2765.715 seconds to run.

Table 2 presents the results for the correction in the global variance and different grids (Finley and Diggle) for knot selection. The use of correction with the Finley grid improved the estimation of τ^2, σ^2 and a. The best results however were obtained by the use of correction with the Diggles grid, except for the range parameter a. Notice that for this sample size (n=100) the use of Diggle or Finley grids leads to a large increase in the computational time for parameters estimation.

Table 1: MSE for model parameters and computational time for different number of knots (20,40 and 60) for the random sample design.

	Number of knots		
Parameter	**20**	**40**	**60**
β_1	0.0039	0.0022	0.0011
β_2	0.0012	0.0005	0.0003
τ^2	0.8590	0.4228	0.2228
σ^2	0.9012	0.4706	0.3350
alcance	4.0499	2.9488	2.1953
Time (sec)	**379.7637**	**692.3181**	**1268.411**

Table 2: MSE for model parameters and computational time for different number of knots (20,40 and 60). For different combination of proposals.

Correction	**Grid**	**Parameter**	**Number of knots**		
			20	**40**	**60**
No	Finley	β_1	0.0230	0.0232	0.0226
		β_2	0.0173	0.0169	0.0169
		τ^2	1.7572	1.2024	0.9476
		σ^2	1.4936	1.0140	0.9240
		a	1.6152	1.6888	1.7093
Time (sec)			**382.12**	**689.72**	**1289.33**
Yes	Finley	β_1	0.0281	0.0282	0.0282
		β_2	0.0197	0.0198	0.0198
		τ^2	1.3198	1.3065	1.2957
		σ^2	0.4790	0.4781	0.4771
		a	1.6274	1.5046	1.0210
Time (sec)			**2137.31**	**3508.68**	**5715.41**
No	Diggle	β_1	0.0240	0.0238	0.0231
		β_2	0.0167	0.0171	0.0169
		τ^2	1.9809	1.6201	1.4306
		σ^2	2.3586	1.7062	1.4614
		a	1.5726	1.5436	1.5450
Time (sec)			**386.75**	**716.07**	**1329.59**
Yes	Diggle	β_1	0.0281	0.0280	0.0281
		β_2	0.0198	0.0198	0.0198
		τ^2	1.2539	1.2284	1.2326
		σ^2	0.4621	0.4606	0.4622
		a	2.4861	1.8332	1.1605
Time (sec)			**1888.33**	**2841.502**	**4334.72**

5.2 Simulation experiment: Contaminated data

The same 100 replicates from subsection 5.1 was considered for data contamination. In each replicated dataset, 7 points were contaminated to be outlying observations. The contamination was performed by adding a random increment to 7 random selected points. Figure 5.2 (a) presents the spatial domain considered in this study and the outlier locations.

In this simulation different grid choices and corrections for the variance are investigated. In particular, for the knot set selection we considered the following proposals as discussed in subsection 4.2.

i) Random: 20, 40, 60 points randomly selected from the 100 locations. The same locations were used for both processes ($\lambda(s)$ and $\omega(s)$);

ii) Proposal 1: 20. 40, 60 points selected. In each scenarium half points were used in the variance process prediction and the other half for the ω process;

iii) Proposal 2: A pilot estimate of the variance $\frac{\hat{\sigma^2}}{\hat{\lambda}_i}$ was used to select half locations with the larger variances to be in the variance process prediction;

iv) Proposal 3: Similar to proposal 2, however, half points in the variance prediction set of knots were selected to have smaller variances.

v) Diggle: The same considered in the Gaussian case.

The prior distributions are the same as in the Gaussian simulation study. To complete the Bayesian model $v \sim GIG(0, c_1, c_2)$. As follows it is presented the Mean Squared Error for the model parameters computed over the 100 replicates. We compared the estimated parameters with the estimates obtained for the Non-Gaussian model using the complete dataset, which would be the best estimates we could obtain from data.

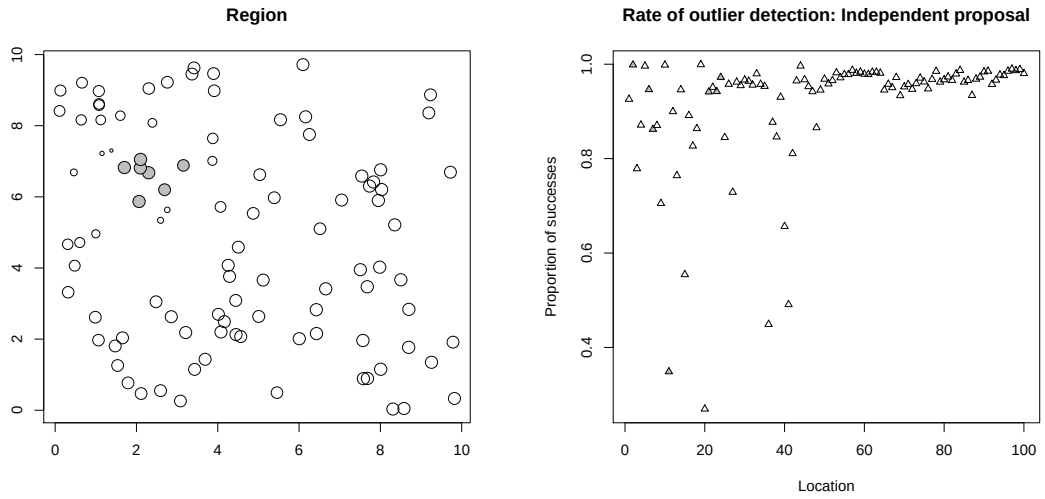
The results are presented in tables 3 and 4. The tables present also the rate of right outlier detection for each proposal. Note in table 3 that the MSE for β_1 and β_2 using the random or Diggle's proposal are very similar while the MSE for τ^2 using Diggle's proposal

was much larger. Opposed to what was observed in the Gaussian study, the MSE for the range parameter was smaller for the Diggle's proposal with 20 knots, and larger for 40 and 60 knots which might be explained by the influence of λ in the range estimation.

Table 4 presents the results for proposals 1 and 2. Note that proposal 1 has larger MSE than the random proposal except for the range parameter when $m = 20$. For $m = 40$ and $m = 50$ the proposal 1 has the best MSE results. Proposal 2 compares favorably with the Diggle's proposal, however, it has a worse preformance than the random and proposal 1 choices.

The rate of right outlier detection was very similar in all proposals and knot set sizes. The rates were high, close to 95%, and are smaller only for points which are in a neighbourhood of outlying observations.

In summary, proposal 1 has the best performance in the knot set selection. This is probably due to the fact that this proposal includes different points in the two knot sets used for variance and process predictions. Regarding the knot set size, 20 locations was too small and MSE were quite large indicating that robustifying the model implies that larger knot sets are required for parameter estimation.



(a) (b)

Figure 2: Contaminated data experiment. (a) Spatial domain for data simulation. Grey circles represent the contaminated data locations. (b) Rate of outlier detection.

Table 3: Relative MSE for random and Diggle's proposals, computational time for parameter estimation and rate of outlier correct detection.

	Random			**Diggle**		
Parameter	**20**	**40**	**60**	**20**	**40**	**60**
β_1	0.0052	0.0087	0.0098	0.0225	0.0173	0.0140
β_2	0.0029	0.0023	0.0016	0.0051	0.0051	0.0036
τ^2	9.5431	3.9891	1.7654	18.8581	12.9365	9.3495
range	5.0925	0.2102	0.4163	1.9724	1.1146	1.4246
Time (Sec.)	1187.55	1840.68	3363.50	1281.24	2378.65	3338.64
Rate	0.94	0.97	0.96	0.93	0.93	0.93

Table 4: MSE for proposals 1 and 2.

	Proposta 1			**Proposta 2**		
Parameter	**20**	**40**	**50**	**20**	**40**	**60**
β_1	0.0069	0.0048	0.0027	0.0128	0.0131	0.0056
β_2	0.0045	0.0013	0.0017	0.0034	0.0037	0.0029
τ^2	10.8164	2.3155	1.0308	13.3828	5.5007	2.7252
range	0.8214	0.1716	0.2259	7.8679	0.7032	1.5482
Time (Sec.)	1418.06	2346.96	2866.23	1269.08	2192.50	3353.57
Rate	0.94	0.97	0.97	0.95	0.95	0.96

6 Conclusions

This work aims to robustify predictive approximations for Gaussian processes which are usually highly influenced by the presence of outiliers in the observed data. Indeed, if the chosen set of knots used in the approximation contains outliers the approximation for the parent process might not represent reality.

We have proposed the use of two different sets of knots, one for the Gaussian process approximation and another one for the variance process approximation. This approach has improved the approximation performance as illustrated in our simulated examples. Proposals 1 and 2 compares favorably with the Diggle's proposal for the knot set selection.

Furthermore, the rates of outlier detection are very large close to 95%. Thus, the model is able to robustify the parent process approximation and also allows for outlier detection.

The proposed approximation has limitations which are inherent of predictive approximations such as oversmoothing. This is not corrected by this approach. However, nearest neighbor processes might be considered for this purpose and also in this context the mixture ideas would be usufel to obtain better representations of the original process. This is topic of future research.

Appendix A: Algorithms

Posterior simulation

For the Non-Gaussian predictive process, the mixing latent variables $\lambda(s_i)$, $i = 1, \ldots, n$ are sampled in the MCMC algorithm using independent proposals as suggested in Palacios and Steel (2006).

1. **Sampling from posterior distribution of $\lambda(s_i)$:**

$$\begin{aligned} P(ln(\lambda^*_{(i)}) \mid ln(\lambda^*_{(-i)}), Z_{(i)}\beta, W_i, \tau 2, v) \propto \\ exp\left\{-\frac{1}{2}\left(ln(\tilde{\lambda}_{(i)})^T Diag(s^{-1}_{(i)}) ln(\tilde{\lambda}_{(i)}) - 2ln(\tilde{\lambda}_{(i)})^T Diag(s^{-1}_{(i)}) m_{(i)}\right)\right\} \\ exp\left\{-\frac{1}{2}\left(ln(\lambda_{(i)})^{*T} C^{-1}_{(i)} ln(\lambda_{(i)})^* - 2ln(\lambda_{(i)})^{*T} C^{-1}_{(i)} M_{(i)}\right)\right\}, \end{aligned} \tag{24}$$

with

$$\begin{aligned} M_{(i)} =& -\frac{v}{2}\mathbf{1}_{(i)} + R^*_{12} R^{*-1}_{22}[ln(\lambda_{(-i)})^* + \frac{v}{2}\mathbf{1}_{(-i)}] \\ C_{(i)} =& v(R^*_{11} - R^*_{12} R^{*-1}_{22} R^*_{21}) \end{aligned} \tag{25}$$

and

$$\begin{aligned} s_i =& 4log\left(\frac{1 + \eta_i\delta(\eta_i) + \eta^2}{[\eta_i + \delta(\eta_i)]^2}\right) \\ m_i =& log\left(\frac{\tilde{W}_i^2}{\tau^2}\frac{1 + \eta_i\delta(\eta_i) + \eta^2}{[\eta_i + \delta(\eta_i)]^4}\right), \quad \text{with,} \\ \eta_i =& \tau^{-1}(Z_i - X_i^{'}\beta) sign(\tilde{W}_i) \\ \delta(\eta_i) =& \frac{\phi(\eta_i)}{(\Phi\eta_i)} \end{aligned}$$

where ϕ and Φ are the density and cumulative distribution functions from the standard Gaussian, respectively.

Appendix B: Proofs of main results

Proof of proposition 3.1

As follows, we drop $\boldsymbol{\theta}$ from the notation for clearer exposition of results.

$$\begin{aligned} Cov(\tilde{\omega}(s), \tilde{\omega}(s^{'})) &= Cov\left(\mathbf{c}^T(s)\mathbf{C}^{*-1}\omega^*, \mathbf{c}^T(s^{'})\mathbf{C}^{*-1}\omega^*\right) \\ &= \mathbf{c}^T(s)\mathbf{C}^{*-1}Var(\omega^*)\mathbf{C}^{*-1}\mathbf{c}(s^{'}) \\ &= \mathbf{c}^T(s;)\mathbf{C}^{*-1}\mathbf{c}(s^{'}) \end{aligned}$$

$$\begin{aligned} &Cov(ln(\tilde{\lambda}(s)), ln(\tilde{\lambda}(s^{'}))) = \\ &= Cov\left(\frac{\nu}{2}\left[R^T(s)\mathbf{R}^{*-1}1_m - 1_n\right] + R^T(s)\mathbf{R}^{*-1}ln(\lambda^*), \frac{\nu}{2}\left[R^T(s^{'})\mathbf{R}^{*-1}1_m - 1_n\right] + R^T(s^{'})\mathbf{R}^{*-1}ln(\lambda^*)\right) \\ &= Cov(R^T(s)\mathbf{R}^{*-1}ln(\lambda^*), R^T(s^{'})\mathbf{R}^{*-1}ln(\lambda^*)) \\ &= R^T(s)\mathbf{R}^{*-1}Var(ln(\lambda^*))\mathbf{R}^{*-1}R(s^{'}) \\ &= \nu R^T(s)\mathbf{R}^{*-1}R(s^{'}) \end{aligned}$$

Proof of proposition 3.2

$$\begin{aligned} V(\omega_i\lambda_i^{-1/2}) &= V(\omega_i)E(\lambda_i^{-1}) = \sigma^2 exp\left\{\nu\right\} \\ V(\tilde{\omega}_i\tilde{\lambda}_i^{-1/2}) &= V(\tilde{\omega}_i)E(\tilde{\lambda}_i^{-1}) = \sigma^2 R^T(s_i, s^*)\,\mathbf{R}^{*-1}R(s_i, s^*)E(\tilde{\lambda}_i^{-1}) \end{aligned}$$

and, $E(\tilde{\lambda}_i^{-1}) = E(exp\{-ln\tilde{\lambda}_i\}) = M_{t=1}(-ln\tilde{\lambda}_i)$, where M_t is the moments generating function of the variavel $-ln\tilde{\lambda}_i$, which has a normal distribution with mean $\nu/2$ and variance $\nu R^T(s_i, s^*)\mathbf{R}^{*-1}R(s_i, s^*)$. Then, using the properties of the log-normal distribution, we have that

$$V(\tilde{\omega}_i\tilde{\lambda}_i^{-1/2}) = \sigma^2 R^T(s_i, s^*)\,\mathbf{R}^{*-1}R(s_i, s^*)exp\left\{\frac{\nu}{2}[1 + R^T(s_i, s^*)\mathbf{R}^{*-1}R(s_i, s^*)]\right\}.$$

□

Acknowledgments The work of Mariana del Pilar Lizarazo Osorio was supported by FAPERJ and the work of Thais C O Fonseca was supported in part by CNPq.

References

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). "Gaussian predictive process models for large spatial data sets." *Journal of the Royal Statistical Society Series B*, 70, 825–848.

Chilès, J.-P. and Delfiner, P. (1999). *Modeling Spatial Uncertainty*. New York: Wiley.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2015). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, , just-accepted.

Diggle, P. and Lophaven, S. (2006). "Bayesian Geostatistical Design." *Scandinavian Journal of Statistics*, 33, 1, 53–64.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). "Improving the performance of predictive process modeling for large datasets." *Computational Statistics and Data Analysis*, 53, 8, 2873 – 2884.

Fonseca, T. C. O. and Steel, M. F. J. (2011). "Non- Gaussian Spatiotemporal Modelling through Scale Mixing." *Biometrika*, 98, 4, 761–774.

Fournier, B. and Furrer, R. (2005). "Automatic mapping in the presence of substutive errors: a robust kriging approach." *Applied GIS*, 1, 12–1–12–16.

Fuentes, M. (2007). "Approximate Likelihood for Large Spaced Spatial Data." *Journal of the American Statistical Association*, 102, 477, 321–331.

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.

Jones, R. H. and Zhang, Y. (1997). *Models for Continuous Stationary Space-Time Processes*. New York: Springer.

Palacios, M. B. and Steel, M. F. J. (2006). "Non-Gaussian Bayesian Geostatistical Modeling." *Journal of the American Statistical Association*, 101, 474, 604–618.

Raftery, A. and Lewis, S. (1992). "How many iterations in the Gibbs sampler." *Bayesian statistics*, 4, 2, 763–773.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). "Approximating Likelihoods for Large Spatial Data Sets." *Journal of the Royal Statistical Society Series B*, 66, 275–296.

Vecchia, A. V. (1988). "Estimation and Model Identification for Continuous Spatial Processes." *Journal of the Royal Statistical Society. Series B (Methodological)*, 50, 2, 297–312.

Xia, G., Miranda, M. L., and Gelfand, A. E. (2006). "Approximately optimal spatial design approaches for environmental health data." *Environmetrics*, 17, 4, 363–385.

Zimmerman, D. L. (2006). "Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction." *Environmetrics*, 17, 6, 635–652.