Estimation in Spatial Models with Censored Response

Thais S. Barbosa[a], Victor H. Lachos[a*], Larissa A. Matos[a] and Marcos O. Prates[b]

[a] Department of Statistics, Campinas State University, Campinas, São Paulo, Brazil
[b] Department of Statistics, Minas Gerais Federal University, Belo Horizonte, Minas Gerais, Brazil

Abstract

Spatial environmental data can be subject to some upper and lower limits of detection (LOD), below or above which the measures are not quantifiable. As a result, the responses are either left or right censored. Historically, the most common practice for analysis of such data has been to replace the censored observations with some function of the limit of detection (LOD/2, 2LOD), or through data augmentation, by using Markov chain Monte Carlo methods. In this paper, we propose an exact estimation procedure to obtain the maximum likelihood estimates of the fixed effects and variance components, using a stochastic approximation of the EM algorithm, the SAEM algorithm (Delyon *et al.*, 1999). This approach permits easy and fast estimation of the parameters of spatial linear models when censoring is present. As a byproduct, predictions of unobservable values of the response variable are possible. The proposed algorithm is applied to a spatial dataset of depths of a geological horizon that contains both left- and right-censored data. We also use simulation to investigate the small sample properties of predictions and parameter estimates and the robustness of the SAEM algorithm. In this simulation study comparisons are made between inferences based on the censored data and inferences based on complete data obtained by a crude/ad hoc imputation method (LOD/2, 2LOD). The results show that differences in inference between the two approaches can be substantial.

Key words: Censored data, Geostatistical data, SAEM Algorithm, Limit of Detection (LOD).

1. Introduction

Spatial data are common in ecology, environmental health, mining, hydrology and epidemiology, where sampling units are geographical areas or spatially located individuals. Analysis of spatial data is challenged by the spatial correlation among the observations. Some methods, like those proposed by Besag (1974); Besag *et al.* (1991); Leroux *et al.* (1999); Rodrigues & Assunção (2012) and other are able to capture spatial dependence for areal data, while Gaussian process regression with exponential, Gaussian, Matérn, spherical, powered exponential and others covariance functions can capture spatial dependence in a geostatistical setting (Cressie, 1993; Banerjee *et al.*, 2004). An additional complication is that spatial data are subject to upper or lower detection limits below and above which they are not quantifiable. For example, environmental (spatial) monitoring of different variables often involves left-censored observations falling below the

*Address for correspondence: Víctor Hugo Lachos Dávila, Departamento de Estatística, IMECC, Universidade Estadual de Campinas, CEP 13083-859, Campinas, São Paulo, Brazil. E-mail: hlachos@ime.unicamp.br.

minimum limit of detection (LOD) of the instruments used to quantify them. The proportion of censored data in these studies may not be small, so the use of crude/ad hoc methods, such as substituting a threshold value or some arbitrary point like a midpoint between zero and cutoff for detection (LOD/2), might lead to biased estimates of fixed effects and variance components (Fridley & Dixon, 2007).

As an alternative to crude imputation methods, De Oliveira (2005) introduced a Bayesian approach for inference and spatial prediction based on censored data while Fridley & Dixon (2007) incorporated the spatial correlation through an unobserved latent spatial process. Rathbun (2006) applied the Robbins & Monro (1951) stochastic approximation algorithm to estimate the parameters of a spatial regression model with left-censored observations. This algorithm uses importance sampling to obtain conditional simulations of left-censored observations. Toscas (2010) proposed a modification of the Bayesian approach of De Oliveira (2005) to correct the Bayesian bias of estimation and prediction of spatially correlated left-censored observations. From a likelihood-based perspective, a few proposals can be found in the literature. For instance, Militino & Ugarte (1999) developed an EM-type algorithm for maximum likelihood (ML) estimation in censored spatial data. However this approach suffers from several drawbacks that restrict its applicability. For instance, De Oliveira (2005) noted that this ML approach does not provide a means to estimate the correlation structure in the data and hence assumes it is known. Additionally, this approach does not account for the different amount of information contained in the exact and censored observations.

As mentioned, a typical algorithm for ML estimation in models involving spatial censored linear (SCL) models is the EM algorithm (see, for instance, Militino & Ugarte, 1999). However, in some cases EM-type algorithms are not appropriate due to the computational difficulty in the E-step, which involves the computation of expected quantities that cannot be obtained analytically and must be calculated using stochastic simulation. To deal with this problem, Delyon *et al.* (1999) proposed a stochastic approximation version of the EM algorithm, the so-called SAEM algorithm. This algorithm consists of replacing the E-step by a stochastic approximation obtained using simulated data, while the M-step remains unchanged. In the framework of spatial models, Jank (2006) showed that the computational effort of SAEM is much smaller and reaches convergence in just a fraction of the simulation size when compared to Monte Carlo EM (MCEM). This is due the memory effect contained in the SAEM method, in which the previous simulations are considered in the computation of the posterior ones. In this paper, we develop a full likelihood approach for SCL models, including the implementation of the SAEM algorithm for ML estimation with the likelihood function and predictions of unobservable values of the response as a byproduct. In the simulation study comparisons are made between inferences based on the censored data through the SAEM algorithm and inferences based on complete data obtained by a crude/ad hoc imputation method (LOD/2, 2LOD). As expected, the differences in inference between the two approaches can be substantial.

The rest of the paper is organized as follows. Section 2 gives a brief description of the spatial linear model, including an outline of the SAEM algorithm. Section 3 proposes the SCL model and shows how to get the ML estimates through the SAEM algorithm. In Section 4, we provide a brief sketch of prediction for models with incomplete data, and also develop a method pertinent to the SCL model. The method is illustrated in Section 5 with the analysis of a data set from depths of a geological horizon and in Section 6 by empirical studies. Section 7 concludes with a short discussion of issues raised by our study and some possible directions for a future research.

2. Preliminaries

2.1. The Spatial linear model

As in De Bastiani *et al.* (2014), we consider a Gaussian stochastic process $\{Z(\mathbf{s}), \mathbf{s} \in \mathbb{D}\}$, where $\mathbb{D}$ is a subset of $\mathbb{R}^d$, the d-dimensional Euclidean space. It supposes that data $Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)$ of this process are observed at known sites (locations) $\mathbf{s}_i$, for $i = 1, \ldots, n$, where $\mathbf{s}_i$ is a d-dimensional vector of spatial site coordinates, and generated from the model,

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \tag{1}$$

where both the deterministic term $\mu(\mathbf{s}_i)$ and the stochastic term $\epsilon(\mathbf{s}_i)$ may depend on the spatial location at which $Z(\mathbf{s}_i)$ is observed. We assume that the stochastic errors have zero mean, $E\{\epsilon(\mathbf{s}_i)\} = 0$, and that variation between spatial points is determined by a covariance function $C(\mathbf{s}_i, \mathbf{s}_j) = \text{Cov}\{\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)\}$. Suppose that for some known functions of $\mathbf{s}_i$, $x_1(\mathbf{s}_i), \ldots, x_p(\mathbf{s}_i)$, the mean of the stochastic process is

$$\mu(\mathbf{s}_i) = \sum_{j=1}^{p} x_j(\mathbf{s}_i)\beta_j, \tag{2}$$

where $\beta_1, \ldots, \beta_p$ are unknown parameters to be estimated. In addition, each family of covariance functions $C(\mathbf{s}_i, \mathbf{s}_j)$, is fully specified by a q-dimensional parameter vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q)^\top$. We use the following notations: $Z(\mathbf{s}_i) = Z_i$, $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$, $x_{ij} = x_j(\mathbf{s}_i)$, $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})$, $\mathbf{X}$ as the $n \times p$ matrix with ith row $\mathbf{x}_i^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\epsilon_i = \epsilon(\mathbf{s}_i)$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Thus, $\mu(\mathbf{s}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ and then $Z_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$. Equivalently, in matrix notation, we have the spatial linear model

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where $E\{\boldsymbol{\epsilon}\} = \mathbf{0}$ and the scale matrix of $\boldsymbol{\epsilon}$, is $\boldsymbol{\Sigma} = [C(\mathbf{s}_i, \mathbf{s}_j)] = \tau^2\mathbf{I}_n + \sigma^2\mathbf{R}(\rho)$. We assume that $\boldsymbol{\Sigma}$ is non singular and that $\mathbf{X}$ has full rank. The parameter τ^2 can be viewed as a measurement error variance or a nugget effect, σ^2 is defined as the sill, $\mathbf{R} = \mathbf{R}(\rho) = [r_{ij}]$, is an $n \times n$ symmetric matrix with diagonal elements $r_{ii} = 1$, for $i = 1, \ldots, n$ and ρ is a function of the range of the model. In general, $\mathbf{R}$ depends on the Euclidean distance $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ between the points $\mathbf{s}_i$ and $\mathbf{s}_j$. This parametric form occurs for several isotropic processes, for instance, the Matérn family of correlation functions $\mathbf{R}$ is defined by

$$\mathbf{R}(\rho) = \mathbf{R}(\rho, d_{ij}) = \begin{cases} \dfrac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\dfrac{d_{ij}}{\rho}\right)^{\kappa} K_\kappa(d_{ij}/\rho), & d_{ij} > 0, \\ 1, & d_{ij} = 0, \end{cases}$$

where $\rho > 0$; $K_\kappa(u) = \frac{1}{2}\int_0^\infty x^{\kappa-1}e^{-\frac{1}{2}u(x+x^{-1})}dx$ is the modified Bessel function of the third kind of order κ; (see, Gradshtejn & Ryzhik, 1965), with $\kappa > 0$ fixed. The Gaussian covariance function is a special case when $\kappa \to \infty$ and its correlation function is given by

$$\mathbf{R}(\rho) = \mathbf{R}(\rho, d_{ij}) = \begin{cases} \exp\left\{-\left(\dfrac{d_{ij}}{\rho}\right)^2\right\}, & d_{ij} > 0, \\ 1, & d_{ij} = 0. \end{cases}$$

The exponential covariance is also a special case of Matérn Family, which corresponds to $\kappa = 1/2$ and can be written more simply as

$$\mathbf{R}(\rho) = \mathbf{R}(\rho, d_{ij}) = \begin{cases} \exp\left\{-\left(\dfrac{d_{ij}}{\rho}\right)\right\}, & d_{ij} > 0, \\ 1, & d_{ij} = 0. \end{cases}$$

In classic geostatistics, the spherical family is also widely used, which has correlation function given by

$$\mathbf{R}(\rho) = \mathbf{R}(\rho, d_{ij}) = \begin{cases} 1 - \dfrac{3}{2}(d_{ij}/\rho) + \dfrac{1}{2}(d_{ij}/\rho)^3, & 0 \leq d_{ij} \leq \rho, \\ 0, & d_{ij} > \rho, \end{cases}$$

where $\rho > 0$ is a single parameter with the dimensions of distance. One qualitative difference between this and the Matérn family is that it has a finite range i.e., $\mathbf{R}(\rho) = 0$ for sufficiently large d_{ij}, namely $d_{ij} > \rho$. The spherical family lacks flexibility by comparison with the two-parameter Matérn class. Also, $\mathbf{R}(\rho)$ is only once differentiable at $d_{ij} = \rho$, which causes technical difficulties with maximum likelihood estimation.

2.2. The EM and SAEM algorithms

In models with missing and censored data, the EM algorithm (Dempster *et al.*, 1977) has became established as the most popular tool to obtain the ML estimates of model parameters. Define $\mathbf{Z}_{\rm com} = (\mathbf{Z}_{\rm m}, \mathbf{Z}_{\rm obs})$, where $\mathbf{Z}_{\rm m}$ denotes the missing data and $\mathbf{Z}_{\rm obs}$ the observed data. This iterative algorithm maximizes the complete log-likelihood function $\ell_c(\boldsymbol{\theta};\, \mathbf{Z}_{\rm com})$ at each step, converging quickly to a stationary point of the observed likelihood $(\ell(\boldsymbol{\theta};\, \mathbf{Z}_{\rm obs}))$ under mild regularity conditions (Wu, 1983; Vaida, 2005).

The EM algorithm proceeds in two simple steps, ***E-Step:*** Replace the observed likelihood by the complete likelihood and compute its conditional expectation $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}\left\{\ell_c(\boldsymbol{\theta};\, \mathbf{Z}_{\rm com})|\widehat{\boldsymbol{\theta}}^{(k)}, \mathbf{Z}_{\rm obs}\right\}$, where $\widehat{\boldsymbol{\theta}}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at the k-th iteration;
M-Step: Maximize $Q(\theta|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ to obtain $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

However, in some applications of the EM algorithm, the E-step cannot be obtained analytically, so it has to be calculated using simulations. Wei & Tanner (1990) proposed the Monte Carlo EM (MCEM) algorithm in which the E-step is replaced by a Monte Carlo approximation based on a large number of independent simulations of the missing data. But this simple solution is computationally expensive, given the need to generate a large number of independent simulations of the missing data for a good approximation. Thus, in order to reduce the amount of required simulations compared to the MCEM algorithm, the SAEM algorithm proposed by Delyon *et al.* (1999) replaces the E-step of the EM algorithm by a stochastic approximation procedure, while the maximization step remains unchanged. Besides having good theoretical properties, the SAEM estimates the population parameters accurately, converging to the global maxima of the ML estimates under quite general conditions (Allassonnière *et al.*, 2010; Delyon *et al.*, 1999; Kuhn & Lavielle, 2004). At each iteration, the SAEM algorithm successively simulates missing data with the conditional distribution, and updates the unknown parameters of the model. Thus, at iteration *k*, the SAEM proceeds as follows

E-Step:

- Simulation: Draw $(\mathbf{q}^{(\ell,k)})$, $\ell = 1, \ldots, m$ from the conditional distribution $f(\mathbf{q}|\boldsymbol{\theta}^{(k-1)}, \mathbf{Z}_i)$.

- Stochastic Approximation: Update the $Q(\theta|\widehat{\theta}^{(k)})$ function as

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \approx Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \ell_c(\theta|\mathbf{Z}_{\text{obs}}, \mathbf{q}^{(\ell,k)}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) \right]. \tag{4}$$

M-Step:

- Maximization: Update $\widehat{\boldsymbol{\theta}}^{(k)}$ as $\widehat{\boldsymbol{\theta}}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$,

where δ_k is a smoothness parameter (Kuhn & Lavielle, 2004), i.e., a decreasing sequence of positive numbers such that $\sum_{k=1}^{\infty} \delta_k = \infty$ and $\sum_{k=1}^{\infty} \delta_k^2 < \infty$. Note that, for the SAEM algorithm, the E-Step coincides with the MCEM algorithm, but only a small number of simulations m (suggested to be $m \leq 20$) is necessary. This is possible because unlike the traditional EM algorithm and its variants, the SAEM algorithm uses not only the current simulation of the missing data at iteration *k* denoted by $(\mathbf{q}^{(\ell,k)})$, $\ell = 1, \ldots, m$ but some or all previous simulations, where this 'memory' property is set by the smoothing parameter δ_k.

Note that in Equation (4), if the smoothing parameter δ_k is equal to 1 for all k, the SAEM algorithm will have 'no memory', and will be equivalent to the MCEM algorithm. The SAEM with no memory will converge quickly (convergence in distribution) to a solution neighborhood, but the algorithm with memory will converge slowly (almost sure convergence) to the ML solution. We suggest the following choice of the smoothing parameter

$$\delta_k = \begin{cases} 1, & \text{for} \quad 1 \leq k \leq cW, \\ \frac{1}{k - cW}, & \text{for} \quad cW + 1 \leq k \leq W, \end{cases}$$

where W is the maximum number of iterations, and c a cutoff point ($0 \leq c \leq 1$) which determines the percentage of initial iterations with no memory. For example, if $c = 0$, the algorithm will have memory for all iterations, and hence will converge slowly to the ML estimates. If $c = 1$, the algorithm will have no memory, and so will converge quickly to a solution neighborhood. For the first case, W would need to be large in order to achieve the ML estimates. For the second, the algorithm will put out a Markov Chain where after applying a *burn in* and *thinning*, the mean of the chain observations can be a reasonably estimated.

A number between 0 and 1 ($0 < c < 1$) will assure an initial convergence in distribution to a solution neighborhood for the first cW iterations and an almost sure convergence for the rest of the iterations. Hence, this combination will lead us to a fast algorithm with good estimates. To implement SAEM, the user must fix several constants matching the number of total iterations W and the cut point c that defines the start of the smoothing step of the SAEM algorithm. However those parameters will vary depending on the model and the data. To determine those constants, a graphical approach is recommended to monitor the convergence of the estimates for all the parameters, and if possible, to monitor the difference (relative difference) between two successive evaluations of the log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{Z}_{obs})$, given by $||\ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{Z}_{obs}) - \ell(\boldsymbol{\theta}^{(k)}|\mathbf{Z}_{obs})||$ or $||\ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{Z}_{obs})/\ell(\boldsymbol{\theta}^{(k)}|\mathbf{Z}_{obs}) - 1||$, respectively.

3. The spatial linear model with censored response

We will consider the Gaussian model with a linear specification for the spatial trend, which allows the inclusion of a polynomial trend surface or, more generally, spatially referenced covariates. From (3), the spatial linear model is defined by

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{X}$ and $\boldsymbol{\beta}$ are as defined in (3) and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$. Moreover, we assume that the response Z_i is not fully observed for all i. Thus, let the observed data for the ith area be $(\mathbf{V}_i, C_i)$, where $\mathbf{V}_i$ represents the vector of uncensored reading ($\mathbf{V}_i = V_{0i}$) or the vectors of censoring level ($\mathbf{V}_i = (V_{1i}, V_{2i})$) and C_i the censoring indicators such that

$$C_i = \begin{cases} 1 & if \quad V_{1i} \leq Z_i \leq V_{2i}\,, \\ 0 & if \qquad Z_i = V_{0i}\,. \end{cases}$$

Note that if $Z_i \in (-\infty, V_{2i}]$, then we get a left censored SCL model (Toscas, 2010) and if $Z_i \in [V_{1i}, \infty)$ then we get a right censored SCL model. The model defined in (5)-(6), will be called the spatial censored linear (SCL) model.

3.1. The log-likelihood function

Classic inference of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top)^\top$, with $\boldsymbol{\alpha} = (\nu^2, \rho)^\top$, is based on the marginal distribution of $\mathbf{Z}$. For complete data, we have marginally that $\mathbf{Z} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Psi}$, with $\boldsymbol{\Psi} = (\nu^2\mathbf{I}_n + \mathbf{R}(\rho))$ and $\nu^2 = \tau^2/\sigma^2$. To compute the likelihood function associated with model (5)-(6), the first step is to treat separately the observed and censored components of Z_i.

Let $\mathbf{Z}^o$ be the n^o-vector of observed outcomes and $\mathbf{Z}^c$ be the n^c-vector of censored observations with $(n = n^o + n^c)$ such that $C_i = 0$ for all elements in $\mathbf{Z}^o$, and $C_i = 1$ for all elements in $\mathbf{Z}^c$. After reordering, $\mathbf{Z}$, $\mathbf{V}$, $\mathbf{X}$, and $\boldsymbol{\Sigma}$ can be partitioned as follows:

$$\mathbf{Z} = vec(\mathbf{Z}^o, \mathbf{Z}^c),\ \mathbf{V} = vec(\mathbf{V}^o, \mathbf{V}^c),\ \mathbf{X}^\top = (\mathbf{X}^o, \mathbf{X}^c) \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{oo} & \boldsymbol{\Sigma}^{oc} \\ \boldsymbol{\Sigma}^{co} & \boldsymbol{\Sigma}^{cc} \end{pmatrix},$$

where $vec(.)$ denotes the function which stacks vectors or matrices having the same number of columns. Then, we have $\mathbf{Z}^o \sim N_{n^o}(\mathbf{X}^o\boldsymbol{\beta}, \boldsymbol{\Sigma}^{oo})$, $\mathbf{Z}^c|\mathbf{Z}^o \sim N_{n^c}(\boldsymbol{\mu}, \mathbf{S})$, where $\boldsymbol{\mu} = \mathbf{X}^c\boldsymbol{\beta} + \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}(\mathbf{Z}^o - \mathbf{X}^o\boldsymbol{\beta})$ and $\mathbf{S} = \boldsymbol{\Sigma}^{cc} - \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{oc}$. Now, let $\Phi_n(\mathbf{u}; \mathbf{a}, \mathbf{A})$ and $\phi_n(\mathbf{u}; \mathbf{a}, \mathbf{A})$ be the cdf (left tail) and pdf, respectively, of $N_n(\mathbf{a}, \mathbf{A})$ computed at vector $\mathbf{u}$. From Vaida & Liu (2009) and Jacqmin-Gadda *et al.* (2000), the likelihood function (using conditional probability arguments) is given by:

$$L(\boldsymbol{\theta}) = f(\mathbf{Z}|\boldsymbol{\theta}) = \phi_{n^o}(\mathbf{Z}^o; \mathbf{X}^o\boldsymbol{\beta}, \boldsymbol{\Sigma}^{oo})(\Phi_{n^c}(\mathbf{V}_2^c; \boldsymbol{\mu}, \mathbf{S}) - \Phi_{n^c}(\mathbf{V}_1^c; \boldsymbol{\mu}, \mathbf{S})), \tag{6}$$

which can be evaluated without much computational burden through the routine *mvtnorm()* available in R (see Genz *et al.*, 2008; R Development Core Team, 2015).

The log-likelihood function for the observed data is used to compute different model selection criteria, such as:

$$AIC = 2\,m - 2\,\ell_{max} \text{ and } BIC = m\,log\,n - 2\,\ell_{max},$$

where m is the number of model parameters and ℓ_{max} is the maximized log-likelihood value.

3.2. The SAEM algorithm for censored spatial data

In this section, we propose the SAEM algorithm by considering $\mathbf{Z}$ as missing data to update (M-step) all the parameters involved in the model. First, we parameterize to $\nu^2 = \tau^2/\sigma^2$ and write $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Psi}$, with $\boldsymbol{\Psi} = \nu^2\mathbf{I}_n + \mathbf{R}(\rho)$. Now, let $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$, $\mathbf{V} = (V_1, \ldots, V_n)$ and $\mathbf{C} = (C_1, \ldots, C_n)$, and (V_i, C_i) is observed for the ith subject. In the estimation procedure, $\mathbf{V}$ and $\mathbf{C}$ are treated as hypothetical

missing data, and augmented with the observed dataset $\mathbf{Z}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{Z}^\top)^\top$. Hence, the complete-data log-likelihood function is given by:

$$\ell_c(\boldsymbol{\theta}|\mathbf{Z}_c) \quad \propto \quad -\frac{1}{2}\left[\log(|\boldsymbol{\Psi}|) + n\log(\sigma^2) + \frac{1}{\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Psi}^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})\right] + cte, \tag{7}$$

with cte being a constant independent of the parameter vector $\boldsymbol{\theta}$. Given the current estimate $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, the E-step calculates the conditional expectation of the complete data log-likelihood function given by:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \quad = \quad E[\ell_c(\boldsymbol{\theta}|\mathbf{Z}_c)|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}] = -\frac{1}{2}\left[\log(\boldsymbol{\Psi}) + n\log(\sigma^2) + \frac{1}{\sigma^2}\widehat{A}^{(k)}\right],$$

where $\widehat{A}^{(k)} = tr\left(\widehat{\mathbf{ZZ}^\top}^{(k)}\boldsymbol{\Psi}^{-1}\right) - 2\widehat{\mathbf{Z}}^{(k)\top}\Psi^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta}$.

It is clear that the E-step reduces only to the computation of

$$\widehat{\mathbf{ZZ}^\top}^{(k)} = E\{\mathbf{ZZ}^\top|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}\} \text{ and } \widehat{\mathbf{Z}}^{(k)} = E\{\mathbf{Z}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}\}. \tag{8}$$

In the traditional EM algorithm, we should now evaluate the conditional expectations. As there are no closed forms for them, we have to introduce two intermediate steps, the simulation and approximation steps. In the simulation, we generate samples from the conditional distributions through the Gibbs sampling algorithm, according to the following scheme

Step E-1 (Sampling). Sample $\mathbf{Z}^c$ from a truncated normal distribution of the form $TN_{n^c}(\boldsymbol{\mu}, \mathbf{S}; \mathbb{A}^c)$, with $\mathbb{A}^c = \{\mathbf{Z}^c = (Z_1^c, \ldots, Z_{n^c}^c)^\top | V_{11}^c \leq Z_1^c \leq V_{21}^c, \ldots, V_{1n^c}^c \leq Z_{n^c}^c \leq V_{2n^c}^c\}$, $\boldsymbol{\mu} = \mathbf{X}^c\boldsymbol{\beta} + \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}(\mathbf{Z}^o - \mathbf{X}^o\boldsymbol{\beta})$ and $\mathbf{S} = \boldsymbol{\Sigma}^{cc} - \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{oc}$. Here $TN_n(.;\mathbb{A})$ denotes the n-variate truncated normal distribution on the interval $\mathbb{A}$, where $\mathbb{A} = A_1 \times \ldots \times A_n$.

Thus, the new observation $\mathbf{Z}^{(k,l)} = (Z_{i1}^{(k,l)}, \ldots, Z_{in^c}^{(k,l)}, Z_{n_i^c+1}, \ldots, Z_n)$ is a sample generated for the n^c censored cases and the observed values (uncensored cases), for $l = 1, \ldots, M$.

Step E-2 (Stochastic Approximation). Since we have the sequence $\mathbf{Z}^{(k,l)}$, at the k-th iteration, we replace the conditional expectations in (8) by the following stochastic approximations:

$$\widehat{\mathbf{ZZ}^\top}^{(k)} \quad = \quad \widehat{\mathbf{ZZ}^\top}^{(k-1)} + \delta_k\left[\frac{1}{M}\sum_{l=1}^{M}\widehat{\mathbf{Z}}^{(k,l)}\widehat{\mathbf{Z}}^{(k,l)\top} - \widehat{\mathbf{ZZ}^\top}^{(k-1)}\right], \tag{9}$$

$$\widehat{\mathbf{Z}}^{(k)} \quad = \quad \widehat{\mathbf{Z}}^{(k-1)} + \delta_k\left[\frac{1}{M}\sum_{l=1}^{M}\widehat{\mathbf{Z}}^{(k,l)} - \widehat{\mathbf{Z}}^{(k-1)}\right]. \tag{10}$$

The conditional maximization (CM) then conditionally maximizes $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ and obtains a new estimate $\widehat{\boldsymbol{\theta}}^{(k+1)}$, as follows:

Step CM

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(k+1)} &= \left(\mathbf{X}^{\top}\widehat{\boldsymbol{\Psi}}^{-1(k)}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\widehat{\boldsymbol{\Psi}}^{-1(k)}\widehat{\mathbf{Z}}^{(k)},\\
\widehat{\sigma^2}^{(k+1)} &= \frac{1}{n}\left[tr\left(\widehat{\mathbf{Z}\mathbf{Z}^{\top}}^{(k)}\widehat{\boldsymbol{\Psi}}^{-1(k)}\right) - 2\widehat{\mathbf{Z}^{\top}}^{(k)}\widehat{\Psi}^{-1(k)}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)} + \widehat{\boldsymbol{\beta}}^{\top(k+1)}\mathbf{X}^{\top}\widehat{\boldsymbol{\Psi}}^{-1(k)}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)}\right]\\
\widehat{\boldsymbol{\alpha}}^{(k+1)} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^{+}\times\mathbb{R}^{+}}{\operatorname{argmax}}\left(-\frac{1}{2}\log(|\boldsymbol{\Psi}|) - \frac{1}{2\widehat{\sigma^2}^{(k+1)}}\left[tr\left(\widehat{\mathbf{Z}\mathbf{Z}^{\top}}^{(k)}\boldsymbol{\Psi}^{-1}\right)\right.\right.\\
&- \left.\left.2\widehat{\mathbf{Z}^{\top}}^{(k)}\Psi^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)} + \widehat{\boldsymbol{\beta}}^{\top(k+1)}\mathbf{X}^{\top}\boldsymbol{\Psi}^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)}\right]\right),
\end{aligned}
\tag{11}
$$

with $\boldsymbol{\alpha} = (\nu^2, \rho)^{\top}$. Note that $\widehat{\tau}^2$ can be recovered using that $\widehat{\tau}^{2(k+1)} = \widehat{\nu}^{2(k+1)}\widehat{\sigma}^{2(k+1)}$. The more efficient CM step (11) can be easily accomplished by using, for instance, the optim routine in R software. This process is iterated until some distance between two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{Z})$ in Subsection 3.1, such as $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})|$ or $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1|$, becomes small enough.

In order to make our proposed algorithm more informative for the reader, in Figure 1 we present a flow diagram, which reports all the steps needed to implement the SAEM algorithm.

4. Prediction

In order to propose a strategy to generate predicted values from our SCL model, we used the plug-in approach proposed by De Oliveira (2005). Thus, let $\mathbf{Z}_{obs}$ be the observed response vector of dimension $n_{obs} \times 1$ and $\mathbf{y}_{pred}$ the corresponding $n_{pred} \times 1$ response vector over the future portion of time. Let $\bar{\mathbf{X}} = (\mathbf{X}_{obs}, \mathbf{X}_{pred})$ be the $(n_{obs} + n_{pred}) \times p$ design matrix corresponding to $\bar{\mathbf{Z}} = (\mathbf{Z}_{obs}^{\top}, \mathbf{Z}_{pred}^{\top})$.

To deal with the censored values existing in $\mathbf{Z}_{obs}$, we use the imputation procedure, by replacing the censored values with $\widehat{\mathbf{Z}} = E\{\mathbf{Z}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\}$ obtained from the SAEM algorithm. Therefore, when the censored values are imputed, a complete dataset, denoted by $\mathbf{Z}_{obs^*}$, is obtained. The reason to use the imputation procedure is that it avoids computing truncated conditional expectations of the multivariate normal distribution originated by the censoring scheme. Hence, we have that

$$
\bar{\mathbf{Z}}^{*} = \left(\mathbf{Z}_{obs^*}^{\top}, \mathbf{Z}_{pred}^{\top}\right)^{\top} \sim N_{n_{obs}+n_{pred}}\left(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}\right),
$$

where the matrix $\boldsymbol{\Sigma}$, can be represented by $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{obs^*,obs^*} & \boldsymbol{\Sigma}^{obs^*,pred} \\ \boldsymbol{\Sigma}^{pred,obs^*} & \boldsymbol{\Sigma}^{pred,pred} \end{pmatrix}$. As mentioned in De Oliveira (2005), the best linear predictor of $\mathbf{Z}_{pred}$ with respect to the minimum mean squared error (MSE) criterion is the conditional expectation of $\mathbf{Z}_{pred}$ given $\mathbf{Z}_{obs^*}$, which is given by:

$$
\widehat{\mathbf{Z}}_{pred}(\boldsymbol{\theta}) = \mathbf{X}_{pred}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{pred,obs^*}\boldsymbol{\Sigma}^{obs^*,obs^*-1}\left(\mathbf{Z}_{obs^*} - \mathbf{X}_{obs^*}\boldsymbol{\beta}\right). \tag{12}
$$

Therefore, $\mathbf{y}_{pred}$ can be estimated directly by substituting $\widehat{\boldsymbol{\theta}}$ into (12), leading to $\widehat{\mathbf{Z}_{pred}} = \widehat{\mathbf{Z}}_{pred}(\widehat{\boldsymbol{\theta}})$.

5. Application

To exemplify the method presented here, we consider a dataset previously analyzed by Dubrule & Kostov (1986) and De Oliveira (2005). The observations are placed over a region of about 9 by 5 km and represent depths of a geological horizon measured at 100 locations where 69 points are fully observed and 31

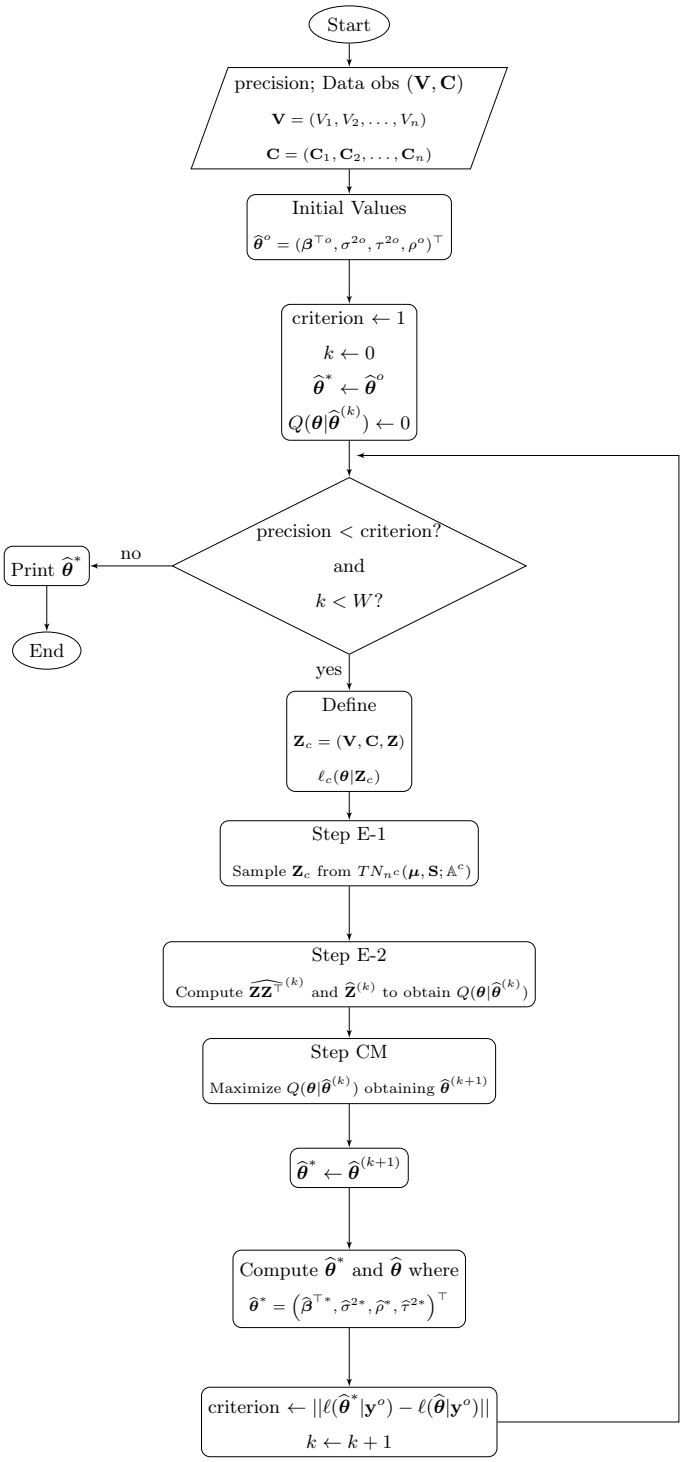

Figure 1: Flow diagram of the SAEM algorithm.

points are censored points, these are divided into left- and right- censored points. Figure 2 (left panel) shows a schematic description of the region that displays the sampling locations and the depth measurements. A

"●" represents an exact observed value, while "▼" represents a left-censored observation, and "▲" indicates a location with a right-censored observation. The numbers above the symbols are the observed value for $C_i = 0$ and the limit of detection for $C_i = 1$, e.g., $\overset{1000}{\blacktriangle}$ means that at location i, $V_i = [1000, \infty)$ is observed. The depth data were transformed and their original units remains unknown for confidentiality reasons. For additional details about this dataset we refer to De Oliveira (2005).

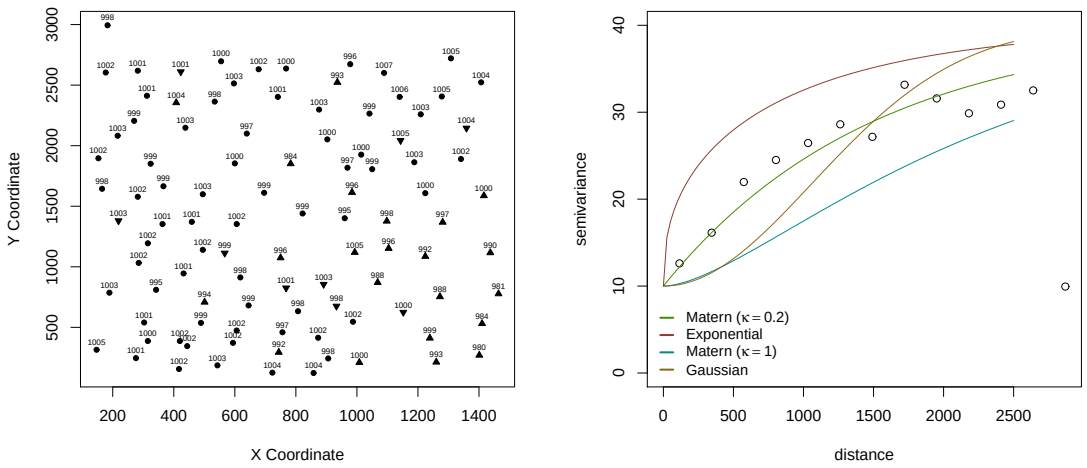

Figure 2: Depth dataset. (left) Depths of a geological horizon measured at each location. ● represents an exact observed value, ▼ represents a left censored value and ▲ represents a right censored value, the values above the symbols indicate observed value for uncensored points and limit of detection for censored points. (right) Plot of empirical variogram represented by ● and theoretical variogram represented by solid line.

5.1. Model specification and preliminary analysis

To illustrate our method, we propose to fit the model:

$$\log\{Z_i\} = \mu + \epsilon_i,$$

with different covariance function for the stochastic errors ϵ_i, $i = 1, \ldots, 100$. This application is based on left and right censoring, and the SAEM algorithm for censored data was implemented as described in Subsection 3.2. We choose a Monte Carlo sample size of $m = 20$, a maximum number of iterations $W = 150$ and a cutoff point $c = 0.2$. In our empirical studies, different choices of these arguments provided the same results for the SAEM estimates. The computational procedures were implemented using the R software (R Development Core Team, 2015) and the codes are available from the authors upon request.

In order to obtain information about the proper covariance function as well as the initial values of the parameters that will be used in the SAEM algorithm, in Figure 2 (right panel) we depict the theoretical variograms using exponential, Gaussian and Matérn family with $\kappa = 0.2$ and $\kappa = 1$ covariance functions, along with the empirical variogram. From this figure, it can be noted that the first plotted ordinate is approximately 10, suggesting a relative nugget variance i.e., depth has some measurement error variance. The rising curve of sample variogram ordinates, leveling out at a distance of around 1500, corresponds to a positive spatial correlation decaying with distance, and the sill is close to 30. This information suggests that the strength of spatial association is moderate. Besides, it can be seen that the theoretical variogram

of the exponential covariance function, is the closest to the empirical variogram. However, this does not necessarily indicate that the model using the exponential covariance function fits the data well, for this reason, we proceed with the estimation process via SAEM using also other covariances functions in order to find the best model fit for the $depth$ data. After this preliminary analysis, we use $\sigma^2 = 30$, $\tau^2 = 10$ and $\rho = 1500$ as initial values for the SAEM algorithm.

It is known that the use of the EM algorithm for this kind of model demands large computational effort. Thus, to compare the EM and SAEM algorithms we ran both algorithms. While the SAEM took 3 minutes to achieve convergence, the EM algorithm took around 15 hours, clearly showing improvement in the computation speed of the ML estimates.

5.2. Parameter estimation

Table 1 contains the ML estimates for the parameters of the four models, namely the Gaussian, exponential and Matérn (with κ = 0.20 and κ = 1) covariance, together with their corresponding log-likelihood (maximized) values and information criteria. Since the Gaussian and exponential covariance functions are special cases of the Matérn family and the variogram plot shown in Figure 2 presents similar behaviors for all theoretical variograms, we do not expect a large difference between the ML estimates. Thus, by looking at the values of the information criteria given in Table 1, we notice that these criteria and the log-likelihood slightly favor the SCL model with Gaussian covariance function. The most noteworthy difference between these results is the general increase in the estimate of ρ according the increase of κ. This arises because the range (ρ) and kappa (κ) are not orthogonal in their effects on the covariance structure and they tend to be strongly correlated. Consequently, the change in one of them implies a change in the other. With these considerations, we choose a SCL model with Gaussian covariance function and parameter estimates: $\widehat{\sigma}^2 = 9.68$, $\widehat{\rho} = 1421.92$ and $\widehat{\tau}^2 = 5.85$ and $\widehat{\mu} = 1002.07$. Note that $\widehat{\tau}^2$ is smaller than $\widehat{\sigma}^2$, which agrees with our intuition that depths can be measured with relatively small error.

Table 1: Depth dataset. ML estimation under Exponential, Gaussian and Matérn (with $\kappa = 0.2$ and $\kappa = 1$) covariance functions.

	$\widehat{\mu}$	$\widehat{\sigma}^2$	$\widehat{\rho}$	$\widehat{\tau}^2$	loglik	AIC	BIC
Matérn ($\kappa = 0.2$)	1000.91	4.26	241.45	2.71	-165.69	339.38	349.80
Exponential	1001.31	5.38	1179.01	5.19	-165.64	339.27	349.69
Matérn ($\kappa = 1$)	1001.64	8.88	1000.66	5.65	-165.76	339.52	349.94
Gaussian	**1002.07**	**9.68**	**1421.92**	**5.85**	**-165.37**	**338.74**	**349.16**

5.3. Prediction

One of the attractive features of the SAEM algorithm, is that it allows making prediction about unobserved location. In order to compare the SCL model under the four covariance functions described in the previous subsection, here we perform a cross-validation experiment. Various percentages of the data ($1\%, 5\%, 10\%, 20\%, 30\%$) were randomly set aside to use in the prediction stage of this experiment. To assess the prediction quality of the four models, we consider only the fully observed values to be predicted. Thus, the true observed values can be compared with the predicted values obtained via the plug-in approach described in Section 4. For instance, when 20% of data are taken, it means that 20 fully observed values are separated, with 31 censored observations and 49 fully observed values remaining, which will be used for estimation via the SAEM algorithm.

As a measure of prediction quality, the mean square prediction error (MSPE) is used to calculate the distance between the real value and the predicted value, which is defined by (see, Fridley & Dixon, 2007):

$$MSPE = \frac{\sum_{i=1}^{n}(Z_i - \hat{Z}_i)^2}{n}, \tag{13}$$

where Z_i is the observed value, $\hat{Z}_i$ is the predicted value and n is the number of samples to be predicted. Note that the best predicted model is determined by the lowest value of MSPE.

Table 2 presents the prediction results, and as expected, the bigger more data considered for prediction, the bigger the prediction error ir. However, the results are satisfactory and the Gaussian covariance function outperforms the other covariances (except for 1%) in term of prediction. This finding agrees with the previous analysis using model selection criteria.

Table 2: Depth dataset. The values in the table denote the MSPE for each percentage of the values set aside for prediction using Exponential, Gaussian and Matérn (with $\kappa = 0.2$ and $\kappa = 1$) covariance functions.

	1%	5%	10%	20%	30%
Matérn ($\kappa = 0.2$)	**1.99**	5.33	10.91	8.17	6.72
Exponential	2.28	5.58	11.11	8.22	6.73
Matérn ($\kappa = 1$)	2.58	5.58	11.36	8.13	6.14
Gaussian	2.76	**4.64**	**10.73**	**8.13**	**5.95**

Figure 3 presents the level maps corresponding to the predictive surface under the four covariance functions and the observed data. They are computed from the predictive distribution over a regular grid of 50×50. It can be seen from these maps that the predictive surface under the Matérn covariance function with $\kappa = 0.2$ provides better results in terms of predictions (near observed values). Note also that this result is not in agreement with the results given in the previous subsection, but is acceptable since the parameter κ is a smoothness parameter, then it is expected that the prediction using a Gaussian covariance function will generate smoothed predictions and be less sensitive to abrupt changes in the data.

6. Simulation Studies

In this section two simulation studies are conducted to examine the performance of the proposed method and to investigate the properties of the SAEM estimates and prediction accuracy. We consider a spatial linear model that contains both left- and right-censored observations as defined in (5)-(6) with an Exponential covariance function. We set $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2) = (1, 3, -2)$, $\sigma^2 = 2$, $\rho = 3$, $\tau^2 = 1$, and $\mathbf{x}_i^\top = (1, x_{i1}, x_{i2})$, where $x_{i1} \sim U(0, 1)$ and $x_{i2} \sim U(2, 5)$, for $i = 1, \ldots, n$, with n depending on the size of the grid adopted.

Three estimation methods were compared. First, the ML estimates were obtained using complete data by set V_{1i} and V_{2i} as the observed values Z_i, represented by *Uncens*. Second, the complete data are obtained by a crude/ad hoc imputation method $V_{1i} = LOD/2$ or $V_{2i} = 2LOD$. We call this procedure LOD*. Third, the SAEM estimates are obtained using the procedure described in this work for censored data, which is denoted by *Cens*.

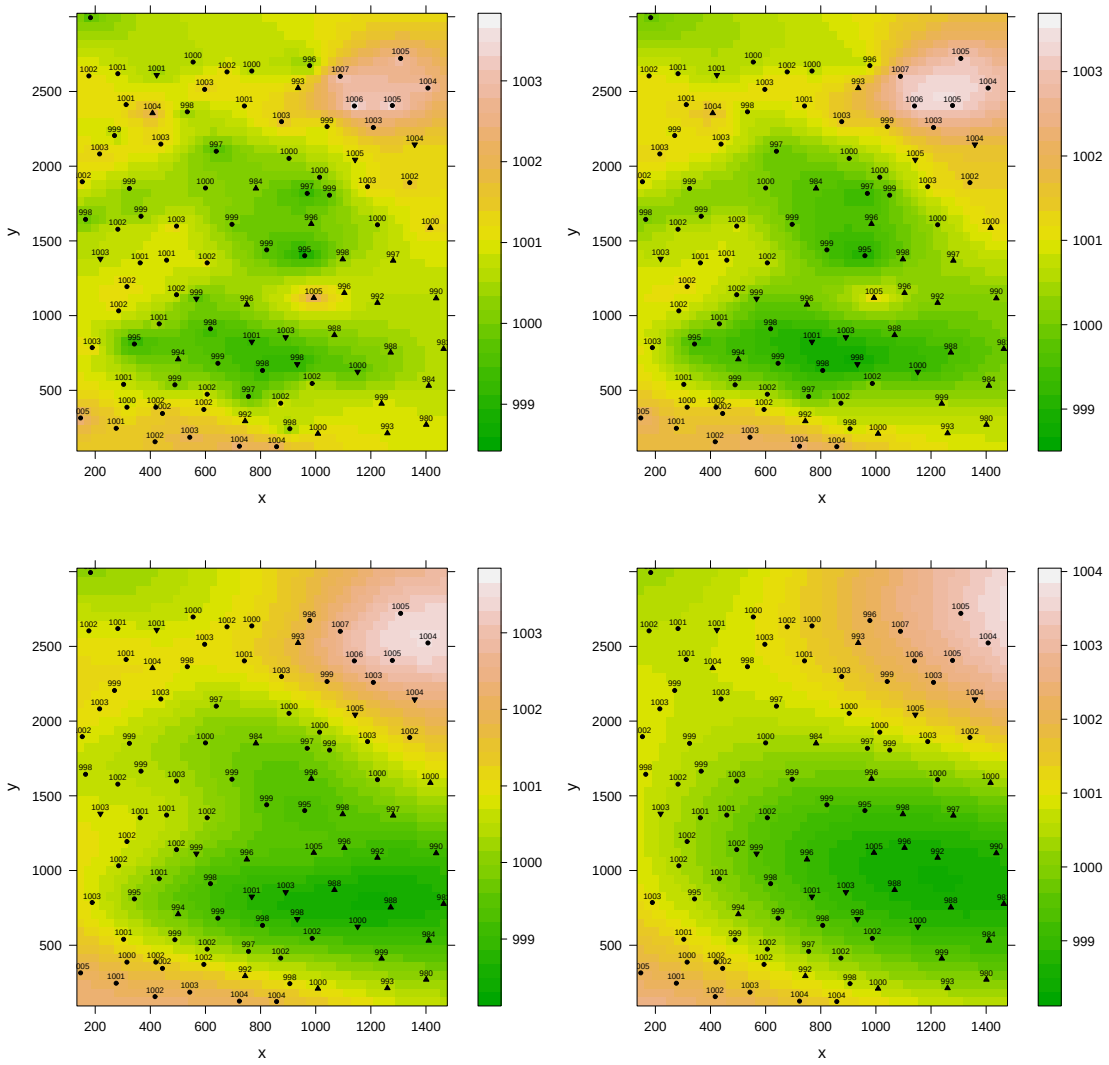

Figure 3: Depth dataset. Map of predicted depth based on the censored data. (top left) Using Matérn Covariance Function with $\kappa = 0.2$, (top right) using Exponential Covariance Function, (bottom right) using Matérn Covariance Function with $\kappa = 1$ and (bottom right) using Gaussian Covariance Function. The values above the symbols indicate observed value for uncensored points and limit of detection for censored points.

6.1. First Study

The first study is a Monte Carlo experiment, where the goal is show the best method of estimation in the presence of censoring. One hundred datasets were generated considering four grids of sizes $10, 15$ and 20 and three different levels of censoring proportions containing left- and right-censored observations $(5\%, 10\%, 20\%)$, which are combined on the left and on the right. Thus, the total level of censoring proportion varies between 10% and 40%.

From Figure 4 and Table 3, showing the results for a grid of size 20, we notice that on average the SAEM produced the closest estimates to the true values for all parameters and provided the best (maximized) log-likelihood and information criteria values. The largest discrepancy between the three methods is related to the estimation of the spatial variability, σ^2, where our method (*Cens*) presented average estimates almost equal to the real values, but produced more variability in the estimates than the other two methods. In

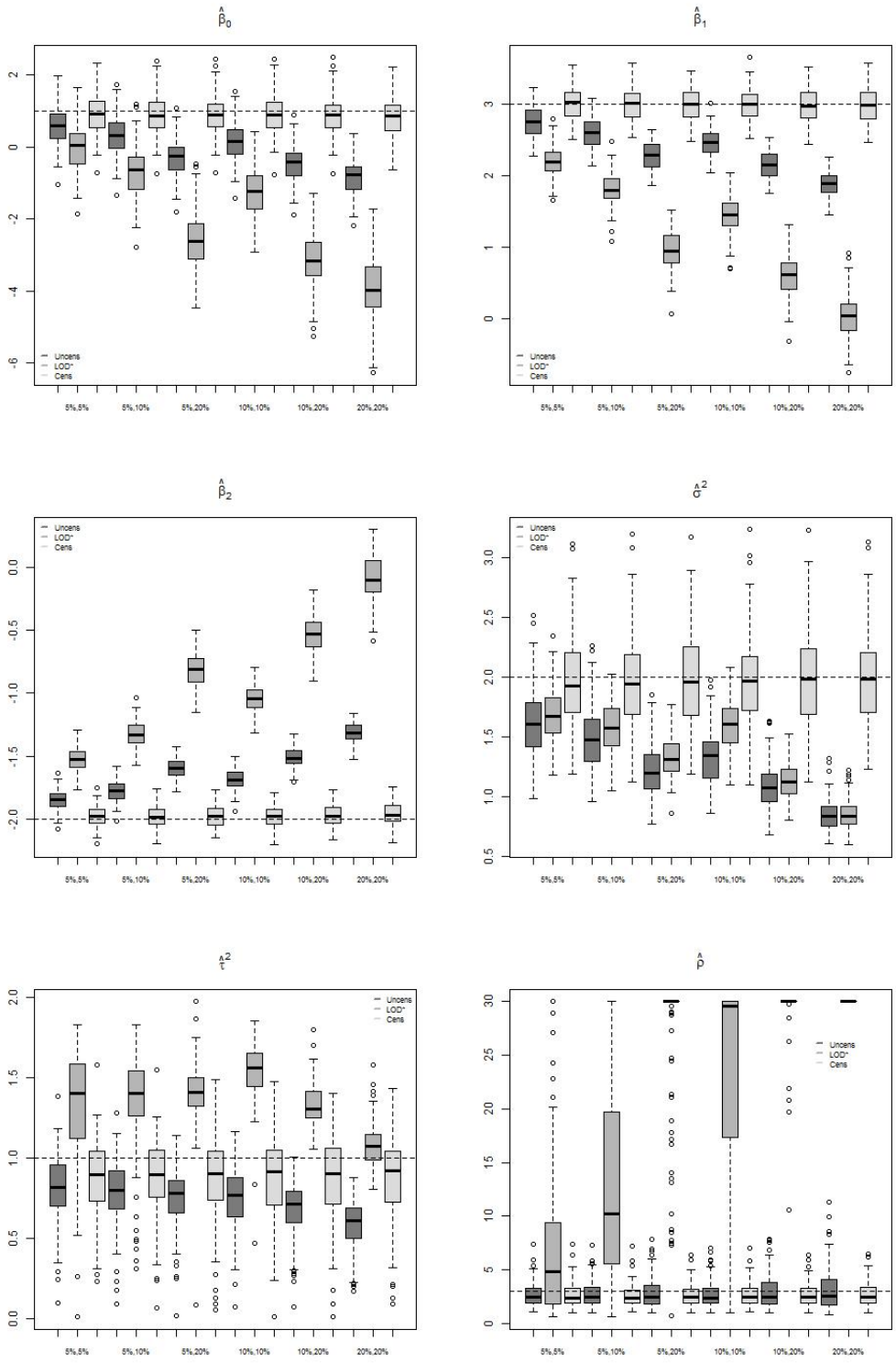

Figure 4: Simulation study 1. Boxplots of the parameter estimates (dotted line indicates the true value of the parameter) for grid of size 20.

addition, for the parameter ρ the LOD* method provides an average estimate distant to the true value, a

higher variability, several outliers and a slight asymmetry, while the other two methods provide estimates near the real values.

Table 3: Simulation study 1. The values in the table denote the information criteria and log-likelihood for each method considering a grid size of 20 and the combination for percentage of censuring level, where the percentages on left and right mean the percentage of data left- and right-censored data respectively.

Censoring Level	Method	AIC	BIC	Loglik
$5\% - 5\%$	*Uncens*	1349.045	1372.994	-668.523
	LOD*	1425.575	1449.524	-706.788
	Cens	1281.700	1305.649	-634.850
$5\% - 10\%$	*Uncens*	1323.621	1347.570	-655.811
	LOD*	1378.012	1401.961	-683.006
	Cens	1207.654	1231.602	-597.827
$5\% - 20\%$	*Uncens*	1277.094	1301.043	-632.547
	LOD*	1329.427	1353.376	-658.713
	Cens	1068.242	1092.191	-528.121
$10\% - 10\%$	*Uncens*	1293.823	1317.772	-640.912
	LOD*	1377.956	1401.905	-682.978
	Cens	1165.320	1189.269	-576.660
$10\% - 20\%$	*Uncens*	1236.393	1260.342	-612.196
	LOD*	1297.024	1320.972	-642.512
	Cens	1025.583	1049.532	-506.792
$20\% - 20\%$	*Uncens*	1153.806	1177.755	-570.903
	LOD*	1207.488	1231.437	-597.744
	Cens	922.926	946.875	-455.463

6.2. Second Study

The second simulation study is constructed to compare the prediction error of the three methods. Here, 500 datasets of 225 observations were simulated from a regular 15×15 lattice with five units between nearest neighbors. Half of the simulated dataset (112 observations) was preserved for prediction purposes and the remaining half (113 observations) was used for estimation as well as for prediction. We consider 10% or 20% of censoring on each side (left- and right-) in order to have a moderate or high level of censoring proportion. Figure 5 illustrate the scheme used for the prediction. Since most of locations are surrounded by four observed locations, this scenario represents the best possible allocation for prediction.

In order to evaluate the accuracy of the prediction, we consider the MSPE measure described in Equation (13). Table 4 shows the number of times that each procedure (LOD*, *Uncens* and *Cens*) provides the lowest MSPE and consequently is most accurate in terms of predictions. It can be seen that the method presented in this paper not only produces better parameter estimates, but also better predictions, as expected. In addition, the *Uncens* procedure provides better results than the LOD* method in terms of prediction, but the opposite was expected, which can be explained because the prediction is obtained according to the observed neighborhood, and changing the data range of the response variable, when using the LOD* method, negatively affects the prediction, producing worse values for MSPE.

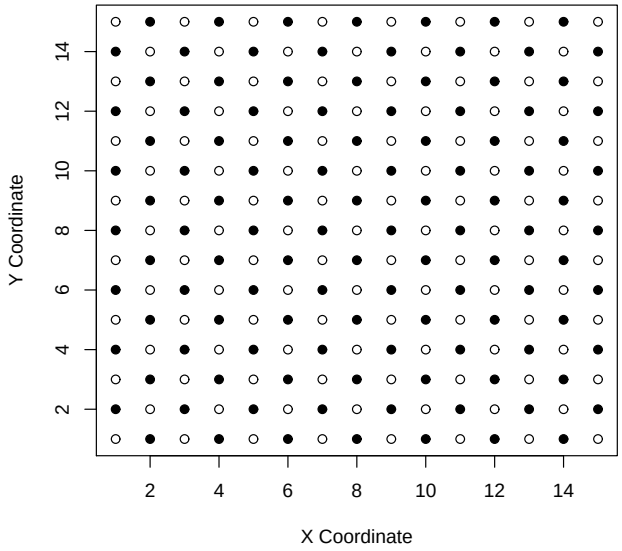

Figure 5: Layout of second simulation study. The ○ represents the sampling sites used for parameter estimation and the ● represents the locations used for comparison of predictions.

Table 4: Simulation study 2. The values in the table denote the number of times and the percentage that each method obtained the lowest MSPE.

Censoring level				
Left	Right	*Uncens*	LOD*	*Cens*
10%	10%	94 (18.8%)	2 (0.4%)	404 (80.8%)
10%	20%	28 (5.6%)	2 (0.4%)	470 (94.0%)
20%	10%	44 (8.8%)	0 (0.0%)	456 (91.2%)
20%	20%	17 (3.4%)	0 (0.0%)	483 (96.6%)

7. Conclusions

This work describes a likelihood-based approach to perform inference and prediction in Gaussian random fields based on spatial censored data. This is a generalization of Gaussian process regression since the traditional methods are not able to deal with censored values. We develop a stochastic approximation of the EM algorithm, called the SAEM algorithm, to obtain the maximum likelihood estimates of model parameters. For practical demonstration, the method is applied to a dataset of depths of a geological horizon that contains both left- and right-censored data and the proposed methods are implemented using the R software (codes available upon request from the first author), providing practitioners with a convenient tool for further applications in their domain. We also use simulation to investigate the properties of predictions and parameter estimates and the robustness of the SAEM algorithm. In this simulation study comparisons are made between inferences based on the censored data and inferences based on complete data obtained by a crude/ad hoc imputation method (LOD/2, 2LOD). We show that the differences in inference between the two approaches can be substantial. Moreover, the SAEM algorithm leads to an improvement in the

computation speed of the ML estimates, as opposed to the EM and Monte Carlo EM (MCEM) algorithms.

Future extensions of the work include the use of scale mixtures of normal distributions to accommodate heavy-tailed features, or the development of some diagnostics and tests for the model (De Bastiani *et al.*, 2014). Further extension would be to generalize the work of Diggle *et al.* (1998) by including censoring under a generalized linear mixed model (GLMM) framework. Another possible extension of the current work it is to include measurement errors in spatial models, as considered in Li *et al.* (2009).

Acknowledgments

Víctor H. Lachos acknowledges support from CNPq-Brazil (Grant 305054/2011-2) and FAPESP-Brazil (Grant 2014/02938-9). Thais Barbosa was supported by CAPES-Brazil. L. Matos acknowledges support from FAPESP-Brazil (Grant 2011/22063-9 and Grant 2015/05385-3). Marcos O. Prates acknowledges support from FAPEMIG-Brazil and CNPq-Brazil.

Appendix.

In this appendix we show complementary results of the first simulation study considering grid of sizes 10 and 15.

Table 5: Simulation study 1. The values in the table denote the information criteria and log-likelihood for each method considering the grid size of 10 and the combination for percentage of censoring level, where the percentages on the left and right mean the percentages of data left- and right-censored data respectively.

Censoring Level	Method	AIC	BIC	Loglik
$5\% - 5\%$	*Uncens*	341.952	357.583	-164.976
	LOD*	362.923	378.554	-175.462
	Cens	325.893	341.524	-156.946
$5\% - 10\%$	*Uncens*	334.887	350.518	-161.443
	LOD*	350.479	366.11	-169.24
	Cens	307.052	322.683	-147.526
$5\% - 20\%$	*Uncens*	321.801	337.432	-154.901
	LOD*	339.921	355.552	-163.96
	Cens	273.173	288.804	-130.587
$10\% - 10\%$	*Uncens*	327.612	343.243	-157.806
	LOD*	355.315	370.946	-171.658
	Cens	296.571	312.202	-142.285
$10\% - 20\%$	*Uncens*	311.997	327.628	-149.998
	LOD*	338.743	354.374	-163.372
	Cens	262.613	278.244	-125.307
$20\% - 20\%$	*Uncens*	293.545	309.176	-140.773
	LOD*	319.779	335.41	-153.89
	Cens	236.409	252.04	-112.204

Table 6: Simulation study 1. The values in the table denote the information criteria and log-likelihood for each method considering the grid size of 15 and the combination for percentage of censoring level, which the percentages on left and right means the percentages of left- and right-censored data respectively.

Censoring Level	Method	AIC	BIC	Loglik
$5\% - 5\%$	*Uncens*	764.09	784.586	-376.045
	LOD*	805.529	826.026	-396.765
	Cens	725.433	745.929	-356.716
$5\% - 10\%$	*Uncens*	751.494	771.99	-369.747
	LOD*	780.578	801.075	-384.289
	Cens	684.666	705.163	-336.333
$5\% - 20\%$	*Uncens*	726.897	747.394	-357.449
	LOD*	760.51	781.006	-374.255
	Cens	606.29	626.787	-297.145
$10\% - 10\%$	*Uncens*	734.788	755.285	-361.394
	LOD*	782.212	802.709	-385.106
	Cens	661.079	681.576	-324.54
$10\% - 20\%$	*Uncens*	704.059	724.556	-346.03
	LOD*	743.266	763.763	-365.633
	Cens	582.517	603.014	-285.259
$20\% - 20\%$	*Uncens*	657.094	677.591	-322.547
	LOD*	689.764	710.26	-338.882
	Cens	525.147	545.644	-256.574

Grid of size 10

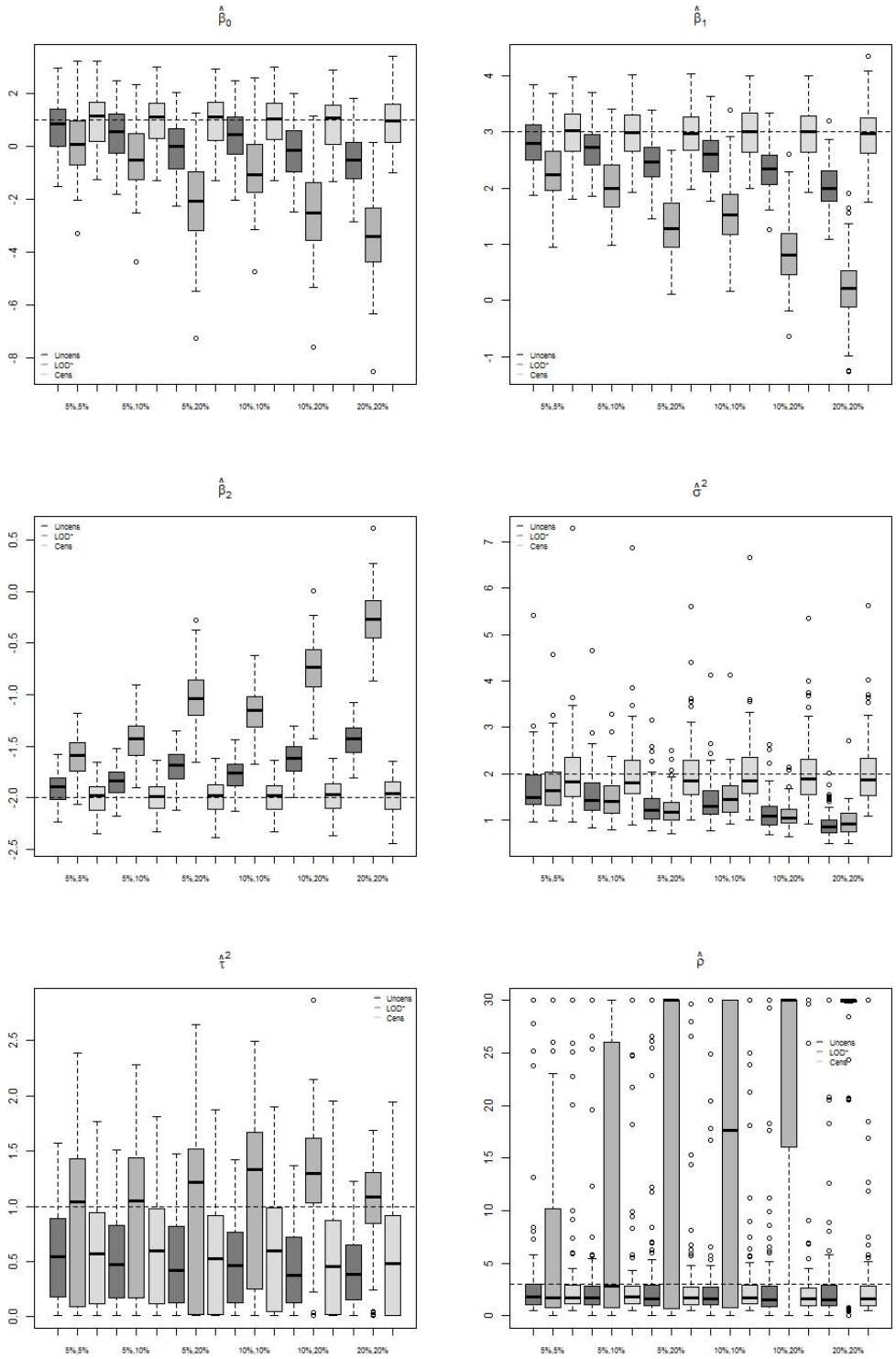

Figure 6: Simulation study 1. Boxplots of the parameter estimates (dotted line indicates the true value of the parameter) for grid of size 10.

Grid of size 15

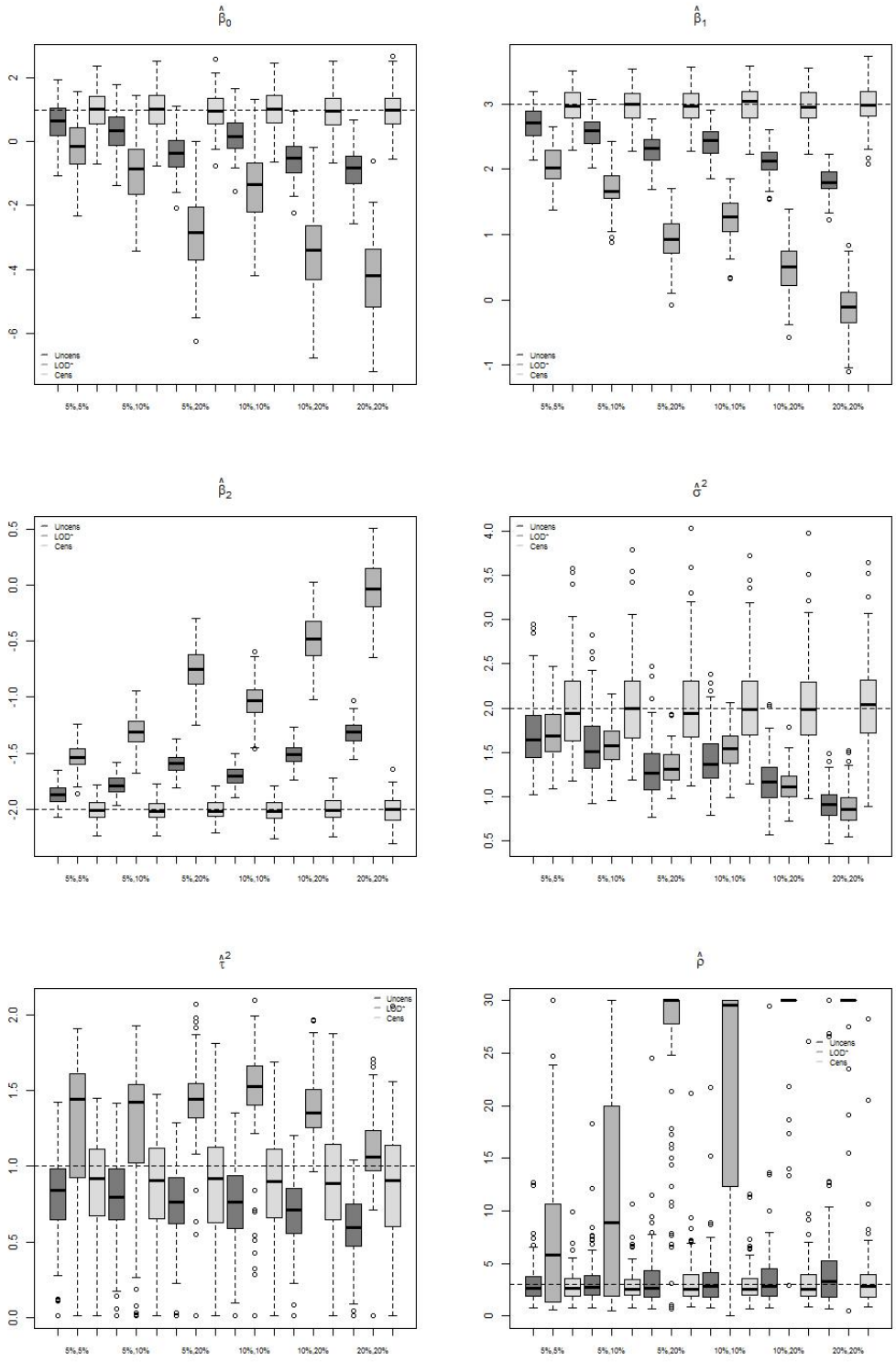

Figure 7: Simulation study 1. Boxplots of the parameter estimates (dotted line indicates the true value of the parameter) for grid of size 15.

References

Allassonnière, S., Kuhn, E., Trouvé, A. *et al.* (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, **16**(3), 641–678.

Banerjee, S., Carlin, P. B. & Gelfand, E. A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, New York.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice data systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 192–225.

Besag, J., York, J. & Mollie, A. (1991). Bayesian image restoration with two application in spatial statistics (with discussion). *Annals of the Institute Statistical Mathematics*, **43**, 1–59.

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

De Bastiani, F., de Aquino Cysneiros, A. H. M., Uribe-Opazo, M. A. & Galea, M. (2014). Influence diagnostics in elliptical spatial linear models. *TEST*, **24**(2), 322–340.

De Oliveira, V. (2005). Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics*, **14**(1), 95–115.

Delyon, B., Lavielle, M. & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, **27**(1), 94–128.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,*, **39**, 1–38.

Diggle, P. J., Tawn, J. & Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**(3), 299–350.

Dubrule, O. & Kostov, C. (1986). An interpolation method taking into account inequality constraints: I. methodology. *Mathematical Geology*, **18**(1), 33–51.

Fridley, B. L. & Dixon, P. (2007). Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics*, **18**, 107–123.

Genz, A., Bretz, F., Hothorn, T., Miwa, T., Mi, X., Leisch, F. & Scheipl, F. (2008). mvtnorm: Multivariate normal and t distribution. *R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm*.

Gradshtejn, I. S. & Ryzhik, I. M. (1965). *Table of integrals, series and products*. Academic Press.

Jacqmin-Gadda, H., Thiebaut, R., Chene, G. & Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, **1**, 355–368.

Jank, W. (2006). Implementing and diagnosing the stochastic approximation EM algorithm. *Journal of Computational and Graphical Statistics*, **15**(4), 803–829.

Kuhn, E. & Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, **8**, 115–131.

Leroux, B. G., Lei, X. & Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. E. Halloran & D. Berry, editors, *In Statistical Models in Epidemiology; the Environment and Clinical Trials*, pages 179–192. Springer–Verlag, New York.

Li, Y., Tang, H. & Lin, X. (2009). Spatial linear mixed models with covariate measurement errors. *Statistica Sinica*, **19**(3), 1077–1093.

Militino, A. F. & Ugarte, M. D. (1999). Analyzing censored spatial data. *Mathematical Geology*, **31**(5), 551–561.

R Development Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rathbun, S. L. (2006). Spatial prediction with left-censored observations. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**(3), 317–336.

Robbins, H. & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Rodrigues, E. C. & Assunção, R. (2012). Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, **109**(0), 88 – 102.

Toscas, P. J. (2010). Spatial modelling of left censored water quality data. *Environmetrics*, **21**(6), 632–644.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica*, **15**(3), 831–840.

Vaida, F. & Liu, L. (2009). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, **18**, 797–817.

Wei, G. C. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**(411), 699–704.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**(1), 95–103.