Computational Statistics and Data Analysis 74 (2014) 125–141

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Spatial prediction in the presence of left-censoring

Lina Schelin, Sara Sjöstedt-de Luna *

Department of Mathematics and Mathematical Statistics, Umeå University, SE-90187 Umeå, Sweden

ARTICLE INFO

Article history:
Received 26 October 2012
Received in revised form 3 November 2013
Accepted 8 January 2014
Available online 20 January 2014

Keywords:
Kriging
Left-censoring
Minimum detection limit
Prediction
Spatial process

ABSTRACT

Environmental (spatial) monitoring of different variables often involves left-censored observations falling below the minimum detection limit (MDL) of the instruments used to quantify them. Several methods to predict the variables at new locations given left-censored observations of a stationary spatial process are compared. The methods use versions of kriging predictors, being the best linear unbiased predictors minimizing the mean squared prediction errors. A semi-naive method that determines imputed values at censored locations in an iterative algorithm together with variogram estimation is proposed. It is compared with a computationally intensive method relying on Gaussian assumptions, as well as with two distribution-free methods that impute the MDL or MDL divided by two at the locations with censored values. Their predictive performance is compared in a simulation study for both Gaussian and non-Gaussian processes and discussed in relation to the complexity of the methods from a user's perspective. The method relying on Gaussian assumptions performs, as expected, best not only for Gaussian processes, but also for other processes with symmetric marginal distributions. Some of the (semi-)naive methods also work well for these cases. For processes with skewed marginal distributions (semi-)naive methods work better. The main differences in predictive performance arise for small true values. For large true values no difference between methods is apparent.

1. Introduction

Spatial prediction methods generally assume that data are fully observed. However, in environmental monitoring, as well as in many other disciplines, the collected spatial data set often includes left-censored observations falling below the minimum detection limit (MDL) of the measuring device. Ways of handling this type of censoring are discussed, e.g. by Bernhardt et al. (2014) in the context of modeling survival data, when the covariates are left-censored. Some spatial prediction methods have also been proposed, ranging from rather naive distribution-free approaches to more sophisticated computer intensive model-based methods. The model-based spatial methods rely on Gaussian assumptions. For data sets with skewed distributions, which are frequently occurring in environmental sciences, the user has to find an appropriate Gaussian transformation of the data (if possible) in order to use the model-based methods. Naive methods do not rely on distributional assumptions and may thus be directly applied to data.

The naive methods are typically based on kriging predictors with plugged-in values at the censored locations. The kriging predictor is the best linear unbiased predictor, given the observed data, that minimizes the mean squared prediction error (see, e.g. Cressie, 1993). Values at censored locations are imputed by simple naive methods, e.g. assigning them the value of the MDL or MDL/2. When the studied variable is known to take values in $[l, \infty)$, where $l <$ MDL, values randomly (uniformly) selected on the interval $[l, \text{MDL})$ are sometimes used. The naive methods are easy to understand and implement and fast to compute, but can have difficulties to predict values below the MDL. Moreover, the spatial autocorrelation structure inherent in the data is not utilized when determining the imputed values at censored locations.

* Corresponding author. Tel.: +46 90 786 54 06; fax: +46 90 786 52 22.
E-mail addresses: lina.schelin@math.umu.se (L. Schelin), sara@math.umu.se (S. Sjöstedt-de Luna).

0167-9473/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2014.01.004

We focus on methods to predict a spatial process at new locations given that we have observed the process at n locations, and where some of the observations are left-censored. The spatial process is assumed to be stationary. We propose a *semi-naive method* designed for processes that take values in $[l, \infty)$, where $l <$ MDL is a known finite lower bound, often zero. Our method is based on the kriging predictor with imputed values at the censored locations. The method avoids distributional assumptions. The imputed values at censored locations are determined within an iterative algorithm that estimates the dependence structure (variogram) based on the observed uncensored data together with successively updated imputed values at the censored locations, using kriging prediction. The algorithm starts by setting all values at censored locations equal to the lower bound l. Hence, this algorithm enables the imputed values to be below the MDL and takes into account the spatial autocorrelation structure when determining them.

Model-based methods are feasible if the data satisfies the distributional assumptions and they may be able to provide predictive distributions including point estimates and their mean squared errors (MSEs) for new locations. Stein (1992) considered prediction and inference under the assumption that the observed data is a realization from a (transformed) truncated stationary Gaussian random field. The truncation point here corresponds to the MDL. Importance sampling is used to estimate predictive conditional distributions for the new locations. Maximum likelihood estimates of the parameters determining the mean and the covariance function are found through an approximation of the likelihood function. Rathbun (2006) used the main ideas of Stein (1992) but applied importance sampling both to find the maximum likelihood estimates and to predict at new locations. Rathbun's and Stein's methods are computationally demanding, especially if there are many censored values.

Militino and Ugarte (1999) suggested an EM algorithm to estimate the true values of Gaussian spatial processes at locations with censored values. A linear transformation, designed from the dependence structure of the process, was applied to the spatially observed data yielding a transformed data set with approximately independent heteroscedastic errors. An EM algorithm for independent data was then used to estimate the unknown values at censored locations. These estimates were imputed at the censored locations and combined with the fully observed data to predict at new locations e.g. via kriging. The method is computationally less demanding than Rathbun's and Stein's methods, but requires that the dependence structure is known. Sedda et al. (2010) suggest an algorithm, relying on spatial simulated annealing to impute values at censored locations. The estimated values at censored locations are chosen through an iterative procedure with the goal to minimize errors in variogram and histogram fitting and kriging prediction. The form of the marginal distribution (histogram) and the functional form of the variogram need be decided in advance as well as tuning parameters.

Two papers within the Bayesian framework are Toscas (2010) and De Oliveira (2005). De Oliveira (2005) performs inference and prediction for Gaussian random fields, accounting for the different amount of information contained in exact and censored observations. Data augmentation and Markov chain Monte Carlo algorithms are used in the proposed approach. This work is slightly modified and evaluated by a simulation study in Toscas (2010).

Here we compare prediction performance of our semi-naive method with two naive methods (imputation with MDL and MDL/2). We also compare with Rathbun's model-based method that relies on Gaussian assumptions but does not require the covariance structure to be known in advance. Comparisons are made through a simulation study on Gaussian and lognormal spatial processes as well as non-Gaussian Laplace fields with asymmetric marginal distributions (Åberg and Podgórski, 2011; Bolin, 2013). In particular we study how the various methods perform with respect to the marginal distribution of the process (symmetric or skewed) and the size of the true values (below MDL, middle, or large values).

This article is organized as follows. In Section 2, we describe kriging prediction in more detail, which is the cornerstone in the spatial prediction methods discussed. In Section 3 and Appendix, we describe Rathbun's method and the naive and semi-naive methods. These methods are compared through a simulation study in Section 4. A real data example is considered in Section 5. Finally, in Section 6, concluding remarks are given.

2. Kriging prediction

Let $\left\{Z(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^d\right\}$ be a second-order stationary stochastic process with covariance function $C(\cdot; \boldsymbol{\theta})$. Here $\boldsymbol{\theta}$ denotes the parameters of the covariance function. We want to predict the value of the process, $Z(\mathbf{s}_p)$, at a new location $\mathbf{s}_p$, given the observed values $\mathbf{z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))^T$ at locations $\mathbf{s}_1, \dots, \mathbf{s}_n$. The ordinary kriging predictor, $\hat{Z}(\mathbf{s}_p, \boldsymbol{\theta}) = \boldsymbol{\lambda}(\boldsymbol{\theta})^T\mathbf{z}$, is a linear combination of the observed values. Given the dependence structure, the ordinary kriging weights $\boldsymbol{\lambda}(\boldsymbol{\theta}) = (\lambda_1(\boldsymbol{\theta}), \dots, \lambda_n(\boldsymbol{\theta}))^T$ are obtained by minimizing the mean squared prediction error $\sigma^2(\boldsymbol{\theta}) = E[(\hat{Z}(\mathbf{s}_p, \boldsymbol{\theta}) - Z(\mathbf{s}_p))^2]$ subject to the constraint $\sum_{i=1}^n \lambda_i(\boldsymbol{\theta}) = 1$. This constraint ensures that the estimator is unbiased. It turns out that the kriging weights are functions of the dependence structure solely;

$$\boldsymbol{\lambda}(\boldsymbol{\theta})^T = \left(\mathbf{c} + \mathbf{1}\frac{\left(1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{c}\right)}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}\right)^T \boldsymbol{\Sigma}^{-1},$$

where $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$, $\mathbf{c} = (C(\mathbf{s}_1 - \mathbf{s}_p; \boldsymbol{\theta}), \dots, C(\mathbf{s}_n - \mathbf{s}_p; \boldsymbol{\theta}))^T$ and where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{z}$, with element $C(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})$ in position (i, j). With these weights, the ordinary kriging variance becomes

$$\sigma^2(\boldsymbol{\theta}) = C(\mathbf{0}; \boldsymbol{\theta}) - \mathbf{c}^T\boldsymbol{\Sigma}^{-1}\mathbf{c} + \frac{\left(1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{c}\right)^2}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}.$$