Prediction Intervals for Integrals of Some Types of Non-Gaussian Random Fields: A Semiparametric Bootstrap Approach

Victor De Oliveira[*] Bazoumana Kone[†]

Abstract
This work proposes a method to construct prediction intervals for integrals of non-Gaussian random fields over bounded regions (called block averages in the geostatistical literature). The method uses a semiparametric approach that does not require distributional assumptions, but only parametric assumptions about the mean and covariance functions of the random field. The resulting semiparametric bootstrap prediction interval overcomes some drawbacks of the commonly used plug-in block kriging prediction interval: the former has better coverage probability properties than the later since it accounts for the uncertainty from parameter estimation, and does not rely on the assumption of Gaussianity. The method is illustrated in the prediction of block averages of cadmium traces in a potentially contaminated region in Switzerland.

Key Words: Block average, Geostatistics, Kriging, Spatial average

1. Introduction

In this work we consider the problem of constructing prediction intervals for integrals of random fields over bounded regions (also called block averages in the geostatistical literature), based on observations at a finite set of sampling locations. This problem is of importance in many earth sciences, such as hydrology, mining and pollution assessment, where interest centers on spatial averages rather than on ensemble averages. Previous approaches to this problem have assumed, explicitly or implicitly, that the random field is Gaussian. But often the variables of interest display markedly non-Gaussian features, so there is a need for methods that do not rely on Gaussianity. We propose here one such method based on the bootstrap.

The suggestion of using bootstrap in geostatistical problems was first posed by Solow (1985), who proposed it to estimate kriging variances, and later Cressie (1993) expanded and outlined several possible approaches in generic terms. For the problem of constructing prediction intervals for the value of a random field at a single location, different bootstrap variants were proposed by Wang and Wall (2003), Sjöstedt-de Luna and Young (2003) and De Oliveira and Rui (2009). The latter two articles proposed parametric bootstrap calibration approaches that are applicable for, respectively, Gaussian and log-Gaussian random fields. Following Cressie (1993), Schelin and Sjöstedt-de Luna (2010) proposed a semiparametric bootstrap approach that does not require distributional assumptions, but only assumptions about the second-order structure of the random field.

The problem of prediction of an integral of a random field over a bounded region has been considered extensively in the literature, for instance, by Cressie (1993), Chilès and Delfiner (1999), Cressie (2006), De Oliveira (2006) and Gotway and Young (2007). But the problem of constructing prediction intervals for integrals has been much less studied. The common approach is to use the so-called block kriging prediction interval computed from estimated covariance parameters; this is called the plug-in (or estimative) approach. This approach has two potentially serious drawbacks. The first, common to all plug-in prediction intervals, is that

[*] Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX 78249, USA, victor.deoliveira@utsa.edu
[†] PPD, Austin, TX 78744, USA, bazoumana2000@yahoo.fr

it does not account for the uncertainty from parameter estimation, and as a result the coverage probability of these plug-in prediction intervals tend to be smaller than the intended (nominal) coverage probability. The second drawback is that the block kriging prediction interval is derived under the assumption that the random field is Gaussian, and as a result the coverage probability of these intervals may not be close to the intended (nominal) coverage probability when the random field is not Gaussian. For the case of Gaussian random fields, De Oliveira and Kone (2015) proposed a method to construct prediction intervals based on bootstrap calibration that overcomes the first drawback. It was shown there that bootstrap calibrated prediction intervals have better coverage properties than plug-in block kriging prediction intervals. We propose here a method to construct bootstrap prediction intervals for some types of non-Gaussian random fields that aims at overcoming *both* drawbacks.

In this work we adapt the semiparametric bootstrap approach proposed by Schelin and Sjöstedt-de Luna (2010) for constructing prediction intervals for values at single locations in some types of non-Gaussian random fields, to the construction of prediction intervals for integrals over bounded regions. In addition, the method is extended to the cases of random fields with non-constant mean function and when the data contain measurement error. The proposed methodology is semiparametric in the sense that parametric assumptions are made about the mean and covariance functions of the random field, where the former is assumed linear in the regression parameters and the latter does not depend on the mean function, but no assumptions are made about the distributions of the random field. The construction of the prediction intervals uses the so-called hybrid bootstrap method (Shao and Tu, 1995), where some key quantiles are estimated by semiparametric bootstrap. Finally, the proposed methodology is applied to the construction of prediction intervals for spatial averages of cadmium traces in a potentially contaminated region in Switzerland.

2. Problem Formulation and Model Description

Consider the random field $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ representing the spatial variation of a quantity of interest that varies continuously over the region of interest $D \subset \mathbb{R}^2$. It is assumed that D is compact and $|D| > 0$, where $|D|$ denotes the area of D (or more precisely its Lebesgue measure), and $Z(\cdot)$ is an L^2 random field, i.e., $E\{Z^2(\mathbf{s})\} < \infty$ for all $\mathbf{s} \in D$. No assumptions are made about the family of finite-dimensional distributions of $Z(\cdot)$, except for second-order assumptions. Specifically, the mean and covariance functions of the random field are assumed to be given by

$$E\{Z(\mathbf{s})\} = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}) =: \mu(\mathbf{s}) \qquad \text{and} \qquad \text{cov}\{Z(\mathbf{s}), Z(\mathbf{u})\} = \sigma^2 K_{\phi}(\mathbf{s}, \mathbf{u}), \tag{1}$$

where $\boldsymbol{f}(\mathbf{s}) = (f_1(\mathbf{s}), \ldots, f_p(\mathbf{s}))'$ are known location-dependent covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ are unknown regression parameters, $\sigma^2 = \text{var}\{Z(\mathbf{s})\} > 0$ is unknown, $K_{\phi}(\mathbf{s}, \mathbf{u})$ is a correlation function in $\mathbb{R}^2$ that is continuous on $D \times D$, and $\boldsymbol{\phi}$ is an unknown correlation parameter. Examples of non-Gaussian random fields satisfying (1) include t random fields (Røislien and Omre, 2006) and Gaussian-Log-Gaussian random fields (Palacios and Steel, 2006).

The observed data consist of possibly noisy measurements of the random field at distinct sampling locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$, say $\mathbf{Z}_{\text{obs}} = (Z_{1,\text{obs}}, \ldots, Z_{n,\text{obs}})'$, where

$$Z_{i,\text{obs}} = Z(\mathbf{s}_i) + \varepsilon_i \ , \quad i = 1, \ldots, n; \tag{2}$$

here $\{\varepsilon_i\}_{i=1}^{n}$ are i.i.d with mean zero and variance $\tau^2 \geq 0$ (the so-called *nugget effect*), representing measurement errors independently distributed of the random field $Z(\cdot)$. The model parameters are then the regression parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and covariance parameters $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi}, \tau^2) \in \Theta \subset \mathbb{R}^q$, $q \geq 3$.

The goal is to make inference about a spatial (weighted) average of the random field over a subregion of D of positive area, say $B \subseteq D$, also know as a block average in the geostatistical literature. This spatial average is the random variable defined by the stochastic integral

$$Z_B = \frac{1}{|B|} \int_B w(\mathbf{s}) Z(\mathbf{s}) d\mathbf{s}, \tag{3}$$

where $w(\cdot)$ is known, nonnegative and piecewise continuous on D; see Cramér and Leadbetter (1967) or De Oliveira and Kone (2015) for details and properties of this stochastic integral.

Let F be the joint distribution of $(\mathbf{Z}'_{\text{obs}}, Z_B)$, assumed to be compatible with (1) but otherwise unknown. For $\alpha \in (0,1)$ we are interested in the construction of *approximate* $100(1-\alpha)\%$ prediction intervals for Z_B, that is, we seek random intervals $\big(L(\mathbf{Z}_{\text{obs}}), U(\mathbf{Z}_{\text{obs}})\big)$ for which

$$P_F\{L(\mathbf{Z}_{\text{obs}}) \le Z_B \le U(\mathbf{Z}_{\text{obs}})\} \approx 1-\alpha, \qquad \text{for any } F \text{ compatible with (1).}$$

Our aim is to develop a distribution-free approach that works well for a wide variety of non-Gaussian random fields.

3. Distribution-free Plug-in Prediction Intervals

In all that follows the sampling design $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is fixed throughout and the dependence of many quantities on it is not made explicit. Also, it is assumed that the $n \times p$ matrix X with entries $(X)_{ij} = f_j(\mathbf{s}_i)$ has full rank ($= p < n$), and the $n \times n$ matrix Σ_θ with entries

$$(\Sigma_\theta)_{ij} = \sigma^2 K_\phi(\mathbf{s}_i, \mathbf{s}_j) + \tau^2 \mathbf{1}\{\mathbf{s}_i = \mathbf{s}_j\}, \tag{4}$$

is positive definite for all $\boldsymbol{\theta} \in \Theta$, where $\mathbf{1}\{A\}$ denotes the indicator function of event A.

The problem of predicting a spatial average based on point-referenced data has been considered extensively in the literature, for instance by Cressie (1993, 2006), Chilès and Delfiner (1999), De Oliveira (2006) and Gotway and Young (2007). The best linear unbiased predictor (BLUP) of Z_B based on the data $\mathbf{Z}_{\text{obs}}$ (also know as the block kriging predictor) and its mean squared prediction error are given, respectively, by

$$\begin{aligned} \hat{Z}_B(\boldsymbol{\theta}) &= \boldsymbol{\lambda}'_B(\boldsymbol{\theta}) \mathbf{Z}_{\text{obs}} \\ \hat{\sigma}^2_B(\boldsymbol{\theta}) &= \sigma^2 K_{BB}(\boldsymbol{\phi}) - 2\sigma^2 \boldsymbol{\lambda}'_B(\boldsymbol{\theta}) \mathbf{K}_B(\boldsymbol{\phi}) + \boldsymbol{\lambda}'_B(\boldsymbol{\theta}) \Sigma_\theta \boldsymbol{\lambda}_B(\boldsymbol{\theta}), \end{aligned} \tag{5}$$

where

$$\begin{aligned} \boldsymbol{\lambda}'_B(\boldsymbol{\theta}) &= \Big(\sigma^2 \mathbf{K}_B(\boldsymbol{\phi}) + X(X'\Sigma_\theta^{-1}X)^{-1}(\mathbf{f}_B - \sigma^2 X'\Sigma_\theta^{-1}\mathbf{K}_B(\boldsymbol{\phi}))\Big)' \Sigma_\theta^{-1} \\ \mathbf{f}_B &= \frac{1}{|B|}\Big(\int_B w(\mathbf{s}) f_1(\mathbf{s}) d\mathbf{s}, \ldots, \int_B w(\mathbf{s}) f_p(\mathbf{s}) d\mathbf{s}\Big)' \\ \mathbf{K}_B(\boldsymbol{\phi}) &= \frac{1}{|B|}\Big(\int_B w(\mathbf{u}) K_\phi(\mathbf{s}_1, \mathbf{u}) d\mathbf{u}, \ldots, \int_B w(\mathbf{u}) K_\phi(\mathbf{s}_n, \mathbf{u}) d\mathbf{u}\Big)' \\ K_{BB}(\boldsymbol{\phi}) &= \frac{1}{|B|^2} \int\int_{B \times B} w(\mathbf{s}) w(\mathbf{u}) K_\phi(\mathbf{s}, \mathbf{u}) d\mathbf{s} d\mathbf{u}; \end{aligned}$$

see Cressie (1993, Section 3.4.5) for details.

If the covariance parameters $\boldsymbol{\theta}$ were known, then a tentative $100(1-\alpha)\%$ prediction interval for Z_B would be

$$\begin{aligned} I_B(\alpha, \boldsymbol{\theta}) &= \Big(\hat{Z}_B(\boldsymbol{\theta}) - \Phi^{-1}(1-\alpha/2)\hat{\sigma}_B(\boldsymbol{\theta}) \ , \ \hat{Z}_B(\boldsymbol{\theta}) + \Phi^{-1}(1-\alpha/2)\hat{\sigma}_B(\boldsymbol{\theta})\Big) \\ &= \big(L_B(\alpha, \boldsymbol{\theta}), U_B(\alpha, \boldsymbol{\theta})\big), \ \text{ say,} \end{aligned} \tag{6}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution (the dependence of $L_B(\cdot)$ and $U_B(\cdot)$ on the data is not made explicit to simplify notation). The coverage probability of this prediction interval is exactly $1-\alpha$ when $Z(\cdot)$ is Gaussian. But in practice the covariance parameters $\boldsymbol{\theta}$ are not known. The simplest and most common practical solution is to use the so-called plug-in (or estimative) prediction interval obtained by replacing $\boldsymbol{\theta}$ in (6) with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{Z}_{\text{obs}})$, an estimate obtained from the same data used for prediction. The resulting interval $I_B(\alpha, \hat{\boldsymbol{\theta}})$ is called the plug-in block kriging prediction interval. This solution has two potentially serious drawbacks.

First, the coverage probability of $I_B(\alpha, \hat{\boldsymbol{\theta}})$ differs from the nominal coverage probability (the one that holds when the true parameter values are used), because it does not take into account the sampling variability from parameter estimation. As a result, its actual coverage probability tends to be smaller than the nominal coverage probability, and the coverage probability error may range from negligible to substantial, depending on the data generating mechanism, observed data and true parameters. An approach to correct this drawback of plug-in prediction intervals is to use bootstrap. For Gaussian random fields, De Oliveira and Kone (2015) explored two bootstrap calibration strategies to calibrate $I_B(\alpha, \hat{\boldsymbol{\theta}})$ by adjusting its bounds $L_B(\alpha, \hat{\boldsymbol{\theta}})$ and $U_B(\alpha, \hat{\boldsymbol{\theta}})$ to $L_B^a(\alpha, \hat{\boldsymbol{\theta}})$ and $U_B^a(\alpha, \hat{\boldsymbol{\theta}})$ say, in such a way that the coverage probability of the calibrated prediction interval is *closer* to $1-\alpha$. It was shown through a simulation study that the coverage probability of bootstrap calibrated prediction intervals substantially improve upon that of plug-in block kriging prediction intervals.

Second, the construction of (6) relies on the assumption that the random field $Z(\cdot)$ is Gaussian, so for non-Gaussian random fields its coverage may be far from $1-\alpha$, even asymptotically. For instance, Schelin and Sjöstedt-de Luna (2010) showed, for random fields with constant mean and exponential covariance function observed on a regular sampling design in $\mathbb{R}$, that the asymptotic distribution (both infill and increasing domain) of $\hat{Z}_{\mathbf{s}_0}(\hat{\boldsymbol{\theta}}) - Z(\mathbf{s}_0)$ is not Gaussian when $Z(\cdot)$ is a non-Gaussian random field. If a random interval analogous to (6) could be constructed that is tailored to the particular non-Gaussian random field under study, then bootstrap calibration could also be used to adjust the bounds of such prediction interval. But this seems unfeasible for most non-Gaussian random fields since the joint distribution of $(\mathbf{Z}_{\text{obs}}', Z_B)$ is usually unknown for these models. For instance, this is the case for log-Gaussian random fields (arguably the class of non-Gaussian random field most commonly used in geostatistics), for which the sampling distributions of Z_B and $Z_B \mid \mathbf{Z}_{\text{obs}}$ are both unknown.

In this work we propose a distribution-free semiparametric bootstrap approach aimed at overcoming both of the aforementioned drawbacks. It is an adaptation of the method proposed by Schelin and Sjöstedt-de Luna (2010) to construct prediction intervals for $Z(\mathbf{s}_0)$, the value of the random field at a single location, to the construction of prediction intervals for Z_B. In addition, the method is extended to the case of random fields with non-constant mean function and when the data contain measurement error.

4. Bootstrap for Dependent Data

The bootstrap is a powerful methodology based on resampling for estimating sampling distributions of statistics. Although it was initially developed for situations with independent and identically distributed data, it was shortly extended to more general situations; see Efron and Tibshirani (1993) and Shao and Tu (1995) for extensive treatments. For situations involving dependent data, several resampling schemes are possible that aim at preserving the dependence of the observed data. The first called *block bootstrap* makes very few assumptions about the data generating mechanism. It involves dividing the data into blocks and resampling the blocks, which may be of different sizes and may or may not overlap; see Lahiri (2003) for an extensive treatment of different variants of this scheme. The second scheme called *semiparametric bootstrap* makes some parametric assumptions about the data generating mechanism. It involves fitting the proposed model using a distribution-free approach, and then resampling the residuals as if they were i.i.d. This scheme was first suggested in the geostatistical literature by Solow (1985) for the estimation of kriging variances, and later a more extensive description appeared in Cressie (1993, Section 7.3.2). Iranpanah, Mohammadzadeh and Taylor (2011) described both schemes and carried out a simulation experiment to compare their accuracy and efficiency in estimating the variance of several statistics, including that of the plug-in kriging predictor of $Z(\mathbf{s}_0)$. Their findings point to the superiority of the semiparametric bootstrap (SP) scheme over the moving block bootstrap (MBB) scheme. Nevertheless, it should be pointed out that their simulation study may not be totally fair. The true mean of the simulated data was used to simulate the bootstrap data in the SB scheme, while no such accommodation was made for the MBB scheme. This may partly explain the smaller biases and mean squared errors of the SP scheme.

Let F be the joint distribution of $(\mathbf{Z}'_{\text{obs}}, Z_B)$, assumed to be compatible with (1) but otherwise unknown. Let $R(\mathbf{Z}'_{\text{obs}}, Z_B)$ a random variable (root) that is a function of $(\mathbf{Z}'_{\text{obs}}, Z_B)$, with the property that for every $\mathbf{z}_{\text{obs}}$, $R(\mathbf{z}_{\text{obs}}, \cdot)$ is a strictly monotone function (say strictly decreasing). Two examples to be investigated in Section 5 are $R(\mathbf{Z}_{\text{obs}}, Z_B) = \hat{Z}_B(\hat{\boldsymbol{\theta}}) - Z_B$ and $R(\mathbf{Z}_{\text{obs}}, Z_B) = \hat{Z}_B(\hat{\boldsymbol{\theta}})/Z_B$, so they both depend on the observed data only through the plug-in block kriging predictor $\hat{Z}_B(\hat{\boldsymbol{\theta}})$. If $H(x)$ denotes the distribution function of $R(\mathbf{Z}'_{\text{obs}}, Z_B)$, then for $\alpha \in (0,1)$ and any F we have

$$\begin{aligned} 1-\alpha &= P_F\big(H^{-1}(\alpha/2) \le R(\mathbf{Z}_{\text{obs}}, Z_B) \le H^{-1}(1-\alpha/2)\big) \\ &= P_F\big(R^{-1}(\mathbf{Z}_{\text{obs}}, H^{-1}(1-\alpha/2)) \le Z_B \le R^{-1}(\mathbf{Z}_{\text{obs}}, H^{-1}(\alpha/2))\big), \end{aligned}$$

where $H^{-1}(\cdot)$ is the quantile function of H, and for any $a \in \mathbb{R}$, $R^{-1}(\mathbf{Z}_{\text{obs}}, a)$ is the solution in Z_B of the equation $R(\mathbf{Z}_{\text{obs}}, Z_B) = a$. If we replace in the above identity $H^{-1}(\alpha/2)$ and $H^{-1}(1-\alpha/2)$ with estimates, $\hat{H}^{-1}(\alpha/2)$ and $\hat{H}^{-1}(1-\alpha/2)$ say, then

$$\Big(R^{-1}\big(\mathbf{Z}_{\text{obs}}, \hat{H}^{-1}(1-\alpha/2)\big)\ ,\ R^{-1}\big(\mathbf{Z}_{\text{obs}}, \hat{H}^{-1}(\alpha/2)\big)\Big), \tag{7}$$

is an approximate $100(1-\alpha)\%$ prediction interval for Z_B; this is what Shao and Tu (1995, Section 4.1.5) call the hybrid bootstrap method for the construction of prediction intervals. The goal is to obtain the estimates $\hat{H}^{-1}(\alpha/2)$ and $\hat{H}^{-1}(1-\alpha/2)$ using semiparametric bootstrap.

4.1 Semiparametric Bootstrap

This bootstrap scheme makes only second-order assumptions about the random field, specifically those in (1), while its family of finite-dimensional distributions is left unspecified. The implementation relies on being able to express the variables' generating mechanism in terms of independent and identically distributed random variables, similarly as in common regression models and autoregressive time-series models. Let

$$\begin{aligned} \Psi_\theta &:= \text{var}(\mathbf{Z}'_{\text{obs}}, Z_B) \\ &= \begin{pmatrix} \Sigma_\theta & \sigma^2\mathbf{K}_B(\boldsymbol{\phi}) \\ \sigma^2\mathbf{K}'_B(\boldsymbol{\phi}) & \sigma^2 K_{BB}(\boldsymbol{\phi}) \end{pmatrix}, \end{aligned} \tag{8}$$

where Σ_θ is given in (4). Also, let L_θ and $\bar{L}_\theta$ be, respectively, the $n \times n$ and $(n+1)\times(n+1)$ lower triangular matrices from the Cholesky factorizations of Σ_θ and Ψ_θ, i.e., $\Sigma_\theta = L_\theta L'_\theta$ and

$\Psi_\theta = \bar{L}_\theta \bar{L}_\theta'$. If $\delta(\mathbf{s}) := Z(\mathbf{s}) - \mu(\mathbf{s})$ is the centered random field, then the data and spatial average can be decomposed as

$$\begin{aligned} \begin{pmatrix} \mathbf{Z}_{\text{obs}} \\ Z_B \end{pmatrix} &= \begin{pmatrix} X\boldsymbol{\beta} \\ \mu_B(\boldsymbol{\beta}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\zeta} \\ \delta_B \end{pmatrix} \\ &= \begin{pmatrix} X\boldsymbol{\beta} \\ \mu_B(\boldsymbol{\beta}) \end{pmatrix} + \bar{L}_\theta \boldsymbol{\epsilon}_{n+1}, \end{aligned} \tag{9}$$

where $\boldsymbol{\zeta} = (\delta(\mathbf{s}_1) + \varepsilon_1, \ldots, \delta(\mathbf{s}_n) + \varepsilon_n)'$, $\mu_B(\boldsymbol{\beta}) = \int_B w(\mathbf{s})\mu(\mathbf{s})d\mathbf{s}/|B|$, $\delta_B = \int_B w(\mathbf{s})\delta(\mathbf{s})d\mathbf{s}/|B|$, and $\boldsymbol{\epsilon}_{n+1} = \bar{L}_\theta^{-1}(\boldsymbol{\zeta}', \delta_B)'$. The components of $\boldsymbol{\epsilon}_{n+1}$ have mean zero, variance one, and are uncorrelated. In addition, we assume these components to be i.i.d. with distribution F_ϵ, say.

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Z}_{\text{obs}})$ and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{Z}_{\text{obs}})$ be distribution-free estimators of the regression and covariance parameters based on the observed data, for instance, $\hat{\boldsymbol{\theta}}$ may be the weighted least squares estimator (Cressie, 1993 Section 2.6.2) and $\hat{\boldsymbol{\beta}} = (X'\Sigma_{\hat{\theta}}^{-1}X)^{-1}X'\Sigma_{\hat{\theta}}^{-1}\mathbf{Z}_{\text{obs}}$. Define the $n \times 1$ vector of residuals by

$$\hat{\boldsymbol{\epsilon}}_n = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)' = L_{\hat{\theta}}^{-1}(\mathbf{Z}_{\text{obs}} - X\hat{\boldsymbol{\beta}}),$$

and the centered residuals by

$$\tilde{\epsilon}_i = \hat{\epsilon}_i - \frac{1}{n}\sum_{k=1}^{n} \hat{\epsilon}_k, \qquad i = 1, \ldots, n. \tag{10}$$

Then F_ϵ can be estimated by the empirical distribution function of the $\tilde{\epsilon}_i$s, namely, $\hat{F}_{\tilde{\epsilon}}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\tilde{\epsilon}_i \le x\}$. If $\boldsymbol{\epsilon}_{n+1}^* := (\epsilon_1^*, \ldots, \epsilon_n^*, \epsilon_{n+1}^*)' \overset{\text{iid}}{\sim} \hat{F}_{\tilde{\epsilon}}$, then the bootstrap data and spatial average are defined as

$$\begin{pmatrix} \mathbf{Z}_{\text{obs}}^* \\ Z_B^* \end{pmatrix} := \begin{pmatrix} X\hat{\boldsymbol{\beta}} \\ \mu_B(\hat{\boldsymbol{\beta}}) \end{pmatrix} + \bar{L}_{\hat{\theta}} \boldsymbol{\epsilon}_{n+1}^*. \tag{11}$$

By the above construction $(\mathbf{Z}_{\text{obs}}^{*'}, Z_B^*) \overset{\text{approx}}{\sim} F$, and hence the distribution of $R(\mathbf{Z}_{\text{obs}}, Z_B)$ can be approximated by that of $R(\mathbf{Z}_{\text{obs}}^*, Z_B^*)$. For the cases considered here $R(\mathbf{Z}_{\text{obs}}^*, Z_B^*)$ depends on $\mathbf{Z}_{\text{obs}}^*$ only through $\hat{Z}_B^*(\hat{\boldsymbol{\theta}}^*)$, where $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}(\mathbf{Z}_{\text{obs}}^*)$ and $\hat{Z}_B^*(\hat{\boldsymbol{\theta}}^*) = \boldsymbol{\lambda}_B'(\hat{\boldsymbol{\theta}}^*)\mathbf{Z}_{\text{obs}}^*$ are, respectively, the covariance parameters estimate and plug-in block kriging predictor based on the bootstrap data. By sampling from the bootstrap joint distribution of the data and spatial average independently a large number of times (say M times), $H(\cdot)$ can be estimated by the empirical distribution of the $R(\mathbf{Z}_{\text{obs}}^*, Z_B^*)$. We summarize all the steps in the following.

Algorithm. Let $B \subset D$, $\alpha \in (0, 1)$ and $M \in \mathbb{N}$ large. Then:

Step 1. Compute the parameter estimates $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Z}_{\text{obs}})$ and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{Z}_{\text{obs}})$, and the plug-in block kriging predictor $\hat{Z}_B(\hat{\boldsymbol{\theta}})$ in (5).

Step 2. Compute the Cholesky factors $L_{\hat{\theta}}$ and $\bar{L}_{\hat{\theta}}$ of, respectively, $\Sigma_{\hat{\theta}}$ and $\Psi_{\hat{\theta}}$ in (4) and (8).

Step 3. Compute the centered residuals $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$ in (10).

For $j = 1, \ldots, M$ do the following:

Step 4. Simulate $\boldsymbol{\epsilon}_{n+1}^{*(j)} = (\epsilon_1^{*(j)}, \ldots, \epsilon_n^{*(j)}, \epsilon_{n+1}^{*(j)})' \overset{\text{iid}}{\sim} \hat{F}_\epsilon$ and compute $(\mathbf{Z}_{\text{obs}}^{*(j)'}, Z_B^{*(j)})'$ as in (11).

Step 5. Compute $\hat{\boldsymbol{\theta}}^{*(j)} = \hat{\boldsymbol{\theta}}(\mathbf{Z}_{\text{obs}}^{*(j)})$ and $\hat{Z}_B^{*(j)}(\hat{\boldsymbol{\theta}}^{*(j)}) = \boldsymbol{\lambda}_B'(\hat{\boldsymbol{\theta}}^{*(j)})\mathbf{Z}_{\text{obs}}^{*(j)}$.

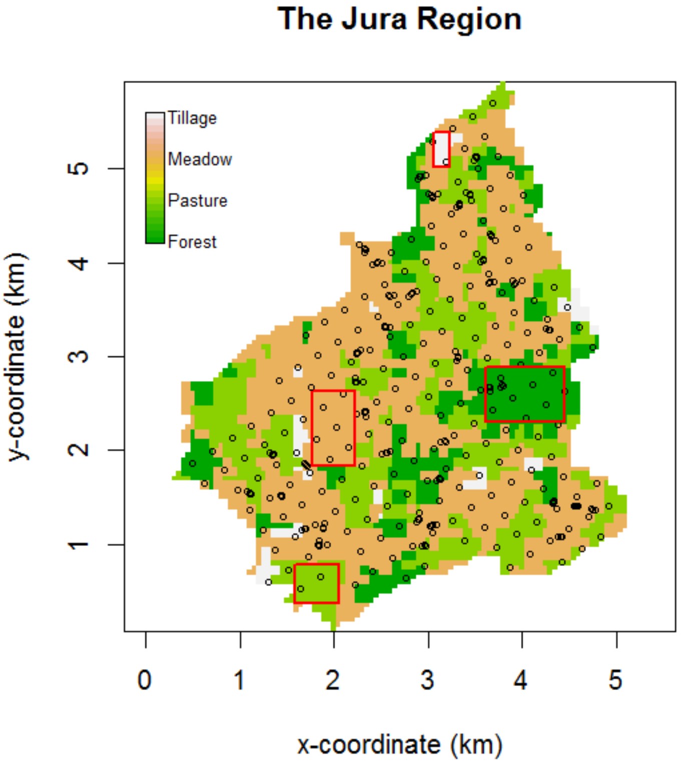

Figure 1: Map of the Jura region with the 359 sampling locations (○), land use type subregions (color-coded), and four rectangular blocks (red lines).

Step 6. Compute $R_j^* = R(\mathbf{Z}_{\text{obs}}^{*(j)}, Z_B^{*(j)})$ and order them from smallest to largest, say $R_{(1)}^* \leq R_{(2)}^* \leq \ldots \leq R_{(M)}^*$.

Step 7. An approximate $100(1-\alpha)\%$ prediction interval for Z_B is

$$\Big(R^{-1}\big(\mathbf{Z}_{\text{obs}}, R^*_{([M(1-\alpha/2)])}\big)\ ,\ R^{-1}\big(\mathbf{Z}_{\text{obs}}, R^*_{([M(\alpha/2)])}\big)\Big),$$

where $[a]$ denotes the integer part of a.

5. Example

We consider the problem of inference about contamination levels in a region of about 15 km^2 in the Swiss canton of Jura. A field survey that took place in 1992 collected measurements of traces in top soil of the heavy metals cadmium, chromium, cobalt, copper, lead, nickel and zinc at 359 locations scattered throughout the region. The region of interest includes the four land use types forest, meadow, pasture and tillage; see Figure 1. The sampling protocol and an initial analysis are described in Atteia et al. (1994), and the datasets and geostatistical analyses appear in Goovaerts (1997). In this section we analyze the cadmium (Cd) traces measured in parts per million (ppm).

Exploratory data analysis suggests that the mean of cadmium traces is constant throughout the region (not shown), and their histogram in Figure 2 (left) shows that the distribution of cadmium traces is not close to Gaussian. Figure 2 (right) plots the empirical semivariogram

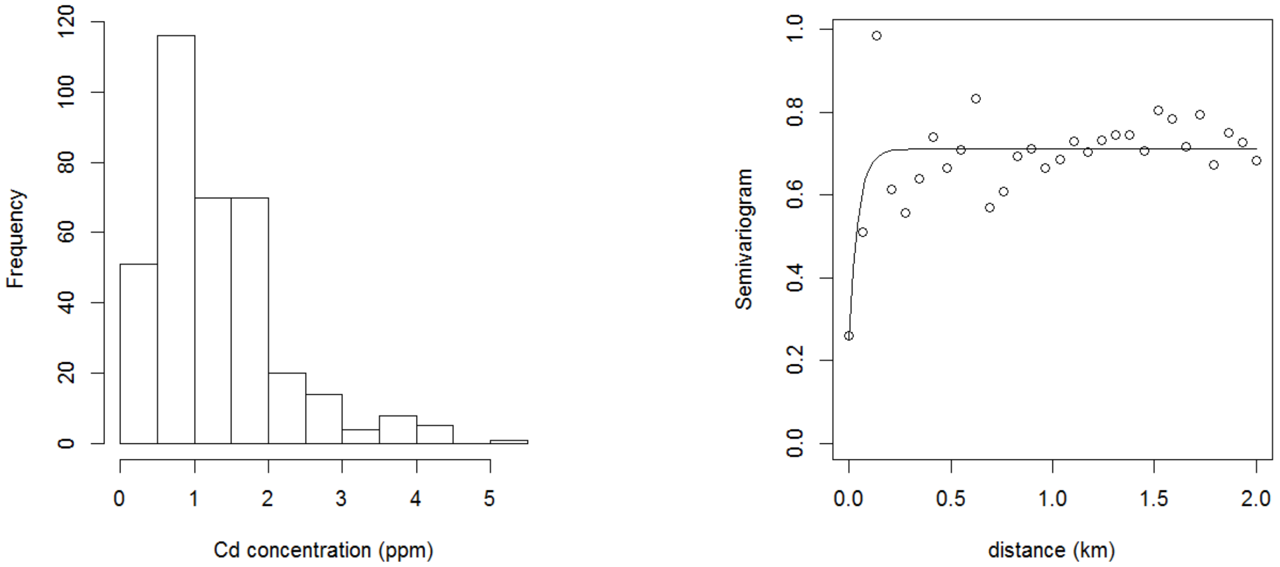

Figure 2: Left: Histogram of measured cadmium traces; Right: Empirical semivariogram (o) and fitted semivariogram (—) of measured cadmium traces.

of cadmium traces, which display an apparent discontinuity at the origin, interpreted as measurement error. Then, we assume the cadmium traces vary throughout the Jura region as an (unspecified) non-Gaussian random field with constant mean β and isotropic exponential covariogram $\sigma^2 \exp(-d/\phi)$, where $d \geq 0$ represents distance. The parameter estimates obtained by least squares are

$$\hat{\beta} = 1.290, \qquad \hat{\sigma}^2 = 0.476, \qquad \hat{\phi} = 0.039, \qquad \hat{\tau}^2 = 0.212,$$

and the fitted semivariogram is displayed in Figure 2 (right).

We construct 95% prediction intervals for spatial averages (3) with $w(\mathbf{s}) \equiv 1$ corresponding to the four blocks displayed in Figure 1, each of which is entirely contained in a land use type; the block coordinates appear in Table 1. For each of the aforementioned blocks we computed two 95% prediction intervals for Z_B based on the bootstrap algorithm described in Section 4.1 with $M = 3000$, using the roots $R_1(\mathbf{Z}_{\text{obs}}, Z_B) = \hat{Z}_B(\hat{\boldsymbol{\theta}}) - Z_B$ and $R_2(\mathbf{Z}_{\text{obs}}, Z_B) = \hat{Z}_B(\hat{\boldsymbol{\theta}})/Z_B$.

Table 2 displays the different prediction intervals. Except for block 4, the bootstrap prediction intervals obtained from the two roots are similar, and for all blocks the prediction intervals

Table 1: Coordinates of the rectangular blocks (in km) and their respective land use type.

Block	Block coordinates	Land use type
1	$[3.62, 4.45] \times [2.30, 2.88]$	Forest
2	$[1.77, 2.23] \times [1.84, 2.63]$	Meadow
3	$[1.58, 2.06] \times [0.38, 0.78]$	Pasture
4	$[3.06, 3.23] \times [5.02, 5.38]$	Tillage
Jura	$[0.2, 5.2] \times [0.2, 5.2]$	Mixed

Table 2: Semiparametric bootstrap 95% block prediction intervals for cadmium traces in four subregions using the roots $R_1(\mathbf{Z}_{\text{obs}}, Z_B) = \hat{Z}_B(\hat{\boldsymbol{\theta}}) - Z_B$ and $R_2(\mathbf{Z}_{\text{obs}}, Z_B) = \hat{Z}_B(\hat{\boldsymbol{\theta}})/Z_B$, and the plug-in 95% block prediction intervals obtained from (6). The numbers in square brackets are the lengths of the intervals.

Block	Land use type	$R_1(\mathbf{Z}_{\text{obs}}, Z_B)$	$R_2(\mathbf{Z}_{\text{obs}}, Z_B)$	Plug-in (6)
1	Forest	$(1.04, 1.54)$ [0.50]	$(1.06, 1.55)$ [0.49]	$(1.11, 1.47)$ [0.36]
2	Meadow	$(1.01, 1.59)$ [0.58]	$(1.06, 1.60)$ [0.54]	$(1.10, 1.52)$ [0.42]
3	Pasture	$(0.94, 1.65)$ [0.71]	$(1.02, 1.69)$ [0.67]	$(1.07, 1.62)$ [0.55]
4	Tillage	$(0.61, 1.65)$ [1.04]	$(0.83, 1.69)$ [0.86]	$(0.82, 1.64)$ [0.82]
Jura	Mixed			

obtained from $R_1(\mathbf{Z}_{\text{obs}}, Z_B)$ are slightly wider than those obtained from $R_2(\mathbf{Z}_{\text{obs}}, Z_B)$ [widths reported in square brackets]. These prediction intervals seem to have little sensitivity to the choice of root in this case. To assess the benefit of accounting for parameter uncertainty when constructing prediction intervals, we also computed the plug-in block kriging prediction intervals obtained from (6), which are also reported in Table 2. The plug-in prediction intervals differ substantially from the bootstrap prediction intervals for all blocks. Although for each block the bootstrap and plug-in prediction intervals are similarly centered, the bootstrap prediction intervals are wider than the plug-in prediction intervals, as expected since the former take into account the uncertainty from parameter estimation. When compared to the plug-in prediction intervals, the bootstrap prediction intervals obtained from $R_1(\mathbf{Z}_{\text{obs}}, Z_B)$ are between 22–39% wider than the corresponding plug-in prediction intervals. As a result, the coverage probability of these bootstrap prediction intervals is expected to be closer to 0.95 than that of the plug-in prediction intervals.

6. Conclusions

This work proposes a semiparametric bootstrap approach for the construction of prediction intervals for integrals of random fields over bounded regions, that is applicable to a variety of non-Gaussian random fields. The methodology seeks to overcome the two drawbacks of plug-in block kriging prediction intervals discussed in Section 3. The main attractive of the proposed method is its semiparametric nature which does not require distributional assumptions, but only parametric assumptions about the mean and covariance functions. The analysis of the cadmium data in Section 5 illustrates the fact that the semiparametric bootstrap prediction intervals may be substantially wider than the plug-in block kriging prediction intervals, so the former are expected to have much better coverage properties than the latter.

The proposed methodology is applicable to many but not all non-Gaussian random fields. For instance, if $Z(\cdot) = \exp(Y(\cdot))$ where $Y(\cdot)$ is a Gaussian random field with mean function $\sum_{j=1}^{p} \alpha_j f_j(\mathbf{s}) - C(\mathbf{s}, \mathbf{s})/2$ and covariance function $C(\mathbf{s}, \mathbf{u})$, then

$$E\{Z(\mathbf{s})\} = \exp\Big(\sum_{j=1}^{p} \alpha_j f_j(\mathbf{s})\Big) =: \mu(\mathbf{s}) \quad \text{and} \quad \text{cov}\{Z(\mathbf{s}), Z(\mathbf{u})\} = \mu(\mathbf{s})\mu(\mathbf{u})\big(\exp\big(C(\mathbf{s}, \mathbf{u})\big) - 1\big),$$

so these do not satisfy (1): the mean function is not linear in the regression parameters and the covariance function depends on the mean function. Nevertheless, we conjecture that the semiparametric bootstrap approach may be extended to random field models such as this, but doing so would require the use of distribution-free methods to estimate the regression and covariance parameters *jointly*.

REFERENCES

Atteia, O., Dubois, J.-P. and Webster, R. (1994), "Geostatistical Analysis of Soil Contamination in the Swiss Jura," *Environmental Pollution*, 86, 315–327.

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

Cramér, H. and Leadbetter, M.R. (1967), *Stationary and Related Stochastic Processes.* Wiley.

Cressie, N. (2006), "Block Kriging for Lognormal Spatial Processes," *Mathematical Geology*, 38, 413–443.

Cressie, N.A.C. (1993), *Statistics for Spatial Data* (rev. ed.). Wiley.

De Oliveira, V. and Kone, B. (2015), "Prediction Intervals for Integrals of Gaussian Random Fields," *Computational Statistics and Data Analysis*, 83, 37–51.

De Oliveira, V. and Rui, C. (2009), "On Shortest Prediction Intervals in Log-Gaussian Random Fields," *Computational Statistics and Data Analysis*, 53, 4345–4357.

De Oliveira, V. (2006), "On Optimal Point and Block Prediction in Log-Gaussian Random Fields," *Scandinavian Journal of Statistics*, 33, 523–540.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap.* Chapman and Hall.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation.* Oxford University Press.

Gotway, C.A. and Young, L.J. (2007), "A Geostatistical Approach to Linking Geographically Aggregated Data From Different Sources," *Journal of Computational and Graphical Statistics*, 16, 1–21.

Iranpanah, N., Mohammadzadeh, M. and Taylor, C.C. (2011), "A Comparison of Block and Semi-parametric Bootstrap Methods for Variance Estimation in Spatial Statistics," *Computational Statistics and Data Analysis*, 55, 578–587.

Lahiri, S.N. (2003), *Resampling Methods for Dependent Data.* Springer-Verlag.

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

Ribeiro, P.J. and Diggle, P.J. (2001), "geoR: A Package for Geostatistical Analysis," *R-NEWS*, 1, 14–18.

Røislien, J. and Omre, H. (2006), "T-distributed Random Fields: A Parametric Model for Heavy-tailed Well-log Data," *Mathematical Geology*, 38, 821–849.

Schelin, L. and Sjöstedt-de Luna, S. (2010), "Kriging Prediction Intervals Based on Semiparametric Bootstrap," *Mathematical Geosciences*, 42, 985–1000.

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap.* Springer-Verlag.

Sjöstedt-de Luna, S. and Young, A. (2003), "The Bootstrap and Kriging Prediction Intervals," *Scandinavian Journal of Statistics*, 30, 175–192.

Solow, A.R. (1985), "Bootstrapping Correlated Data," *Mathematical Geology*, 17, 769–775.

Wang, F. and Wall, M.M. (2003), "Incorporating Parameter Uncertainty Into Prediction Intervals for Spatial Data Modeled via a Parametric Bootstrap," *Journal of Agricultural, Biological and Environmental Statistics*, 8, 296–309.