Computational Statistics and Data Analysis 83 (2015) 37–51

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Prediction intervals for integrals of Gaussian random fields

Victor De Oliveira [a,*], Bazoumana Kone [a,b]

[a] Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX 78249, USA
[b] PPD, Austin, TX 78744, USA

ARTICLE INFO

Article history:
Received 30 August 2013
Received in revised form 12 September 2014
Accepted 13 September 2014
Available online 20 September 2014

Keywords:
Block average
Bootstrap calibration
Change of support problem
Geostatistics
Kriging
Spatial average

ABSTRACT

Methodology is proposed for the construction of prediction intervals for integrals of Gaussian random fields over bounded regions (called block averages in the geostatistical literature) based on observations at a finite set of sampling locations. Two bootstrap calibration algorithms are proposed, termed indirect and direct, aimed at improving upon plug-in prediction intervals in terms of coverage probability. A simulation study is carried out that illustrates the effectiveness of both procedures, and these procedures are applied to estimate block averages of chromium traces in a potentially contaminated region in Switzerland.

1. Introduction

In this work we consider the problem of constructing prediction intervals for the integral of a spatially varying quantity over a bounded region (also called block average 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 often centers on spatial averages (rather than on ensemble averages). This was, in a mining context, a problem that D.G. Krige considered and that motivated G. Matheron to develop the geostatistical methodology named after him (kriging): estimating the average block ore-grade over a mining panel based on measurements of internal core-samples (Cressie, 1990; Chilès and Delfiner, 1999). As the data support are 'points' while the support of the quantity of interest is a region of positive area, this is an example of what is generically called a change of support problem; see Gotway and Young (2002) for an extensive review.

The problem of 'point 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 'interval prediction' has not received similar attention, and it could even be argued that it has not been adequately explored. When the model covariance parameters are *not* known, the common practice in the above works is to use a two-stage approach: the covariance parameters are first estimated and then prediction intervals are computed by treating these estimates as if they were the true covariance parameters. This is called the plug-in (or estimative) approach. It is by now well known that plug-in prediction intervals have coverage properties that differ from the nominal coverage properties and are often overly optimistic, having actual coverage probability smaller than the desired coverage

* Correspondence to: Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA. Tel.: +1 210 4586592.
E-mail address: victor.deoliveira@utsa.edu (V. De Oliveira).

http://dx.doi.org/10.1016/j.csda.2014.09.013
0167-9473/

probability. The main approaches to correct this drawback of plug-in prediction intervals are the Bayesian and bootstrap approaches. Both approaches have been explored for the case of inference about the quantity of interest at single locations, but similar approaches for the case of inference about spatial averages do not seem to have been explored, with the exception of the paper by Gelfand et al. (2001) who proposed a Bayesian approach. This work studies bootstrap calibration approaches.

A general idea for the construction of improved prediction intervals is to calibrate plug-in prediction intervals, namely, to adjust plug-in prediction limits in such a way that the resulting prediction interval has coverage probability closer to the desired coverage probability. Two variants of this general idea have been explored that differ on how the adjustment is made. In the first variant the adjusted limit is obtained by modifying the nominal coverage probability, a variant termed as *indirect* by Ueki and Fueda (2007). This variant was initially proposed by Cox (1975), and later studied further by Atwood (1984), Beran (1990), Escobar and Meeker (1999) and Lawless and Fredette (2005). In the second variant additive adjustments are made to plug-in prediction limits, a variant termed as *direct* by Ueki and Fueda (2007). This variant was studied by Barndorff-Nielsen and Cox (1994, 1996), Vidoni (1998) and Ueki and Fueda (2007). For both variants the adjustments can be computed either analytically (Cox, 1975; Atwood, 1984; Barndorff-Nielsen and Cox, 1996; Vidoni, 1998) or by simulation (Beran, 1990; Escobar and Meeker, 1999; Lawless and Fredette, 2005; Ueki and Fueda, 2007). Analytical adjustments are often complex and difficult to obtain, while simulation-based adjustments (also called bootstrap calibration) are usually more practically feasible. The simulation-based indirect calibration variant has been studied and applied for the construction of prediction intervals in random fields at single locations by Sjöstedt-de Luna and Young (2003) and De Oliveira and Rui (2009), but bootstrap calibration does not seem to have been studied for the construction of prediction intervals for spatial averages of random fields.

In this work we study the application of both indirect and direct bootstrap calibration strategies to the construction of prediction intervals for spatial averages of Gaussian random fields over bounded regions. We extend the indirect bootstrap calibration algorithm proposed by Sjöstedt-de Luna and Young (2003) for the construction of prediction intervals for the random field at *single* locations to the construction of prediction intervals for spatial averages over bounded regions. Also, we extend the direct bootstrap calibration algorithm proposed by Ueki and Fueda (2007) for i.i.d. data to the construction of prediction intervals for spatial averages, which relies on a 'predictive distribution' that only depends on the covariance parameters. A simulation study is carried out to illustrate the effectiveness of both types of calibrated prediction intervals at reducing the coverage probability error of plug-in prediction intervals. Finally, the proposed methodology is applied to the construction of prediction intervals for spatial averages of chromium traces in a potentially contaminated region in Switzerland.

2. Model and problem formulation

Consider the random field $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ representing the spatial variation of a quantity of interest, thought to vary 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 $\mathbf{s} \in D$. The mean and covariance functions of the random field are assumed to be given by

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

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, $\sigma^2 = \operatorname{var}\{Z(\mathbf{s})\} > 0$ is unknown, $K_\phi(\mathbf{s}, \mathbf{u})$ is a correlation function in $\mathbb{R}^2$ that is continuous on $D \times D$, and ϕ is an unknown correlation parameter.

The 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

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

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

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 given 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, and the integral is defined in mean square sense; its definition and some of its properties are given in the next section. For $\alpha \in (0, 1)$ we are interested in constructing $100(1-\alpha)\%$ 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

$$P_{\beta,\theta}\{L(\mathbf{Z}_{\text{obs}}) \leq Z_B \leq U(\mathbf{Z}_{\text{obs}})\} = 1 - \alpha, \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p, \ \boldsymbol{\theta} \in \Theta,$$

where $P_{\beta,\theta}\{\cdot\}$ refers to the joint distribution of $(\mathbf{Z}'_{\text{obs}}, Z_B)$ when the true parameter vector is $(\boldsymbol{\beta}', \boldsymbol{\theta})$.