Geoderma 319 (2018) 89–99

Contents lists available at ScienceDirect

Geoderma

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

How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters?

R.M. Lark[1],*, B.P. Marchant

British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, UK

ARTICLE INFO

Keywords:
Spatial sampling
Prediction variance
Geostatistics

ABSTRACT

We use an expression for the error variance of geostatistical predictions, which includes the effect of uncertainty in the spatial covariance parameters, to examine the performance of sample designs in which a proportion of the total number of observations are distributed according to a spatial coverage design, and the remaining observations are added at supplementary close locations. This expression has been used in previous studies on numerical optimization of spatial sampling, the objective of this study was to use it to discover simple rules of thumb for practical geostatistical sampling. Results for a range of sample sizes and contrasting properties of the underlying random variables show that there is an improvement on adding just a few sample points and close pairs, and a rather slower increase in the prediction error variance as the proportion of sample points allocated in this way is increased above 10 to 20% of the total sample size. One may therefore propose a rule of thumb that, for a fixed sample size, 90% of sample sites are distributed according to a spatial coverage design, and 10% are then added at short distances from sites in the larger subset to support estimation of spatial covariance parameters.

1. Introduction

1.1. The problem and its motivation

How should we sample a variable in space to allow geostatistical prediction for an information system or mapping project? This is an important question for the application of geostatistics in soil science, particularly when limited resources are available to support soil sampling in the field and the analysis of sampled material in the laboratory. It is important because the sampling determines both the cost of the survey and the quality of the resulting predictions.

One of the first approaches to this question was made by McBratney et al. (1981) who showed that if the spatial covariance parameters (variogram parameters) of the target random variable are known, at least approximately or from a homologous setting, then one may identify the spacing of a square sample grid such that the kriging variance at the centre of a grid cell (where the point kriging variance takes its largest value) does not exceed some threshold. van Groenigen et al. (1999) demonstrated that spatial simulated annealing, a method for numerical optimization, can be used to find sampling designs in irregularly-shaped regions so as to minimize the mean or maximum kriging variance over that region. This approach will tend to produce a 'space-filling' or 'spatial-coverage' design, which can also be achieved by the methods of Walvoort et al. (2010).

The limitation of spatial-coverage designs for geostatistics, be these regular grids or space-filling designs in irregular regions, is that they do not provide information on spatial dependence over short intervals, and so the modelled spatial covariance at short lag distances is poorly constrained. The covariance at short distances is particularly influential on the kriging weights. While some early geostatistical studies in soil science used regular sampling grids (e.g. Burgess and Webster, 1980, Webster and Oliver, 1989) it was realized that it is necessary to include some observations within a sample array that are a short distance apart to support the estimation of spatial covariance parameters (e.g. Atteia et al., 1994; Cattle et al., 2002). However, we are not aware of an explicit analysis of the benefits of doing this in terms of the quality of final kriging predictions. Stein (1999), in a simple 1-D simulation with only 20 sample locations on a regular transect, showed that the likelihood function for spatial covariance parameters was very flat near the maximum, but that adding just three additional observations at finer intervals within the transect markedly reduced the uncertainty. In 2-D simulations with more realistic sample sizes Haskard (2007) supported this finding. She considered a total sample size of 100, but allocated either 10 or 20 of these points to clusters within an incomplete 10 × 10

* Corresponding author.
E-mail address: murray.lark@nottingham.ac.uk (R.M. Lark).

[1] Now at: School of Biosciences, University of Nottingham, Sutton Bonington Campus, Sutton Bonington, Leicestershire LE12 5RD, UK

https://doi.org/10.1016/j.geoderma.2017.12.022
Received 12 June 2017; Received in revised form 15 December 2017; Accepted 20 December 2017

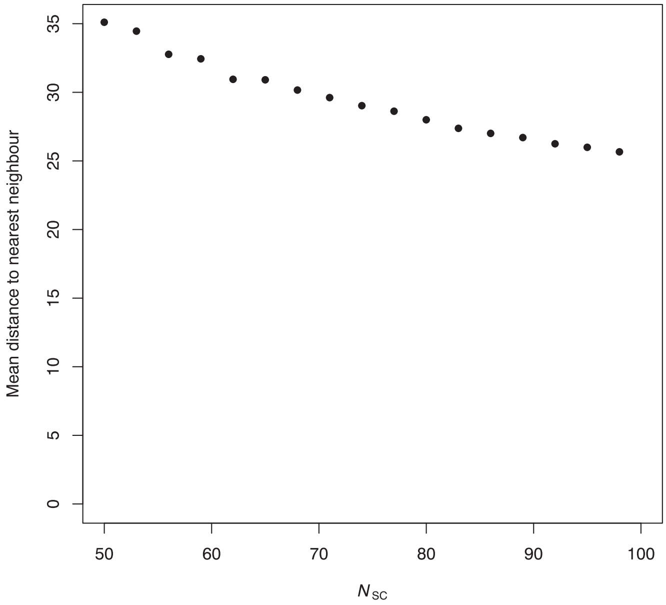

Fig. 1. Mean distance to nearest neighbour within a set of N_{SC} points in a spatial coverage sample in a 256 × 256-unit square region.

square grid. She found a marked reduction in the standard errors of spatial covariance parameters when using the sample array with 10 points in a cluster by comparison to the full 10 × 10 grid, and only a small additional benefit in using 20 of the 100 points in clusters.

Simple spatial-coverage sampling will not do to support geostatistical prediction, so how can appropriate designs be discovered? Zhu and Stein (2006) and Marchant and Lark (2007a,b) showed how to define an overall objective function for the quality of a sampling design, an expected mean square error of predictions, which accounts for the two sources of uncertainty in the empirical best-linear unbiased prediction (E-BLUP, equivalent to the kriging prediction in the general case with no covariates and the local mean assumed to be stationary). These two sources are the spatial variation of the target variable and the uncertainty in the maximum likelihood (ML) estimates of the spatial covariance parameters. More detail is provided in Section 1.2. The key point is that we do not assume that the spatial covariance parameters are known without error, but account for their uncertainty, which depends in part on the sampling design. Spatial simulated annealing can then be used to minimize the mean value, or the maximum value, of this objective function across a study area. The resulting designs resemble a spatial-coverage sample with some additional points at shorter distances.

These formal methods for optimization may be complex to implement. They require an approximation of the spatial covariance parameters of the target variable, or a specification of their joint prior distribution. In practice the scientist who is planning a survey may have a more-or-less fixed sample size to deploy, and simple rules of thumb may be more useful than complex procedures for optimization, which may also be computationally demanding. There are various rules of thumb in geostatistics which have been influential amongst practitioners. For example, it is generally advised to form empirical estimates of the variogram for lag distances no longer than $D/2$ where D is the maximum distance between observations (Journel and Huijbregts, 1978). Webster and Oliver (1992) suggest that at least 100 observations are required to obtain a reliable estimate of the variogram. Kerry et al. (2010) advise that a sampling grid for geostatistical prediction should have a spacing no coarser than half the range of spatial dependence of the target variable, and ideally one third to two fifths of the range.

The objective of this paper is to see whether it is possible to devise rules of thumb to plan a geostatistical soil survey *de novo*. Following the observations of Stein (1999) and Haskard (2007), and from the simulation results of Zhu and Stein (2006) and Marchant and Lark (2007a,b)), we propose that the rule for a geostatistical survey with N observations is to withhold some number of these (a short-distance subset), distribute the remaining N_{SC} according to a spatial-coverage design and then to insert each observation from the short-distance subset into the resulting regular array at some fixed short distance, but in a random direction, from a randomly selected site in the spatial-coverage subset. We examine a quality measure for the resulting surveys, the mean square error of prediction as computed by Marchant and Lark (2007a) which accounts both for the density of sampling around a prediction site and the uncertainty of the spatial variance parameters. The key question is whether a general recommendation can be made as to how many sample sites to reserve for the short-distance subset. Our study is therefore one in the spirit of 'innovization' (innovation by optimization), as discussed by Deb et al. (2014). The key idea of innovization is that one seeks to discover rules which a practitioner can implement which capture the key properties of solutions identified by formal optimization.

In the next section we review the calculation of the prediction error variance of the E-BLUP as proposed by Zhu and Stein (2006) and Marchant and Lark (2007a). Section 2 then sets out the sampling schemes and scenarios for which we evaluated this error variance. The scenarios correspond to random variables with a range of spatial covariance parameters. These include parameter sets selected from a Markov Chain Monte Carlo sample of parameters for the random effects in a linear mixed model for the variation of soil carbon content across a part of eastern lowland England with a range of contrasting land uses.

1.2. The mean-square error of the empirical best linear unbiased prediction

In this paper we consider the case of ordinary kriging, although the formulation of the problem extends to the more general best-linear unbiased prediction (BLUP) which includes universal kriging (or regression kriging in an approximately equivalent presentation). The ordinary kriging prediction of a variable, Z at a location $\mathbf{x}_0$, given q covariance parameters in $\boldsymbol{\theta}$ and n observations in $\mathbf{z} = (z\,(\mathbf{x}_1), z\,(\mathbf{x}_2), \ldots, z\,(\mathbf{x}_n))^{\mathrm{T}}$ can be written as

$$\widehat{Z}\,(\mathbf{x}_0|\theta) = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{z}, \qquad (1)$$

where $\boldsymbol{\lambda}$ is a vector of weights. The weights are obtained from the ordinary kriging equation

$$\mathbf{L} = \mathbf{A}^{-1}\mathbf{b}, \qquad (2)$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{C}, & \mathbf{1}_n \\ \mathbf{1}_n^{\mathrm{T}}, & 0 \end{bmatrix}$$

the matrix $\mathbf{C}$ is the covariance matrix of the n observations given their locations, $\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_n$ and the covariance function with parameters in $\boldsymbol{\theta}$, $C\,(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta})$; $\mathbf{1}_n$ is a vector length n of ones,

$$\mathbf{L} = \begin{bmatrix} \lambda \\ \psi \end{bmatrix},$$

where ψ is a Lagrange multiplier. If $\mathbf{c}$ is a vector of the covariances between the target location $\mathbf{x}_0$ and the observations, $\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_n$, then

$$\mathbf{b} = \begin{bmatrix} \mathbf{c} \\ 1 \end{bmatrix}.$$

In this formulation the expected square error of the prediction, the