Geoderma 173-174 (2012) 231–240

Contents lists available at SciVerse ScienceDirect

Geoderma

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

# Quality measures for soil surveys by lognormal kriging

R.M. Lark [a,*], D.J. Lapworth [b]

[a] British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, UK
[b] British Geological Survey, Maclean Building, Wallingford, Oxfordshire, OX10 8BB, UK

ARTICLE INFO

*Article history:*
Received 22 July 2011
Received in revised form 8 December 2011
Accepted 10 December 2011
Available online 4 February 2012

*Keywords:*
Lognormal kriging
Prediction interval
Sampling
Geostatistics

ABSTRACT

If we know the variogram of a random variable then we can compute the prediction error variances (kriging variances) for kriged estimates of the variable at unsampled sites from sampling grids of different design and density. In this way the kriging variance is a useful pre-survey measure of the quality of statistical predictions, which can be used to design sampling schemes to achieve target quality requirements at minimal cost. However, many soil properties are lognormally distributed, and must be transformed to logarithms before geostatistical analysis. The predicted values on the log scale are then back-transformed. It is possible to compute the prediction error variance for a prediction by this lognormal kriging procedure. However, it does not depend only on the variogram of the variable and the sampling configuration, but also on the conditional mean of the prediction. We therefore cannot use the kriging variance directly as a pre-survey measure of quality for geostatistical surveys of lognormal variables. In this paper we present an alternative. First we show how the limits of a prediction interval for a variable predicted by lognormal kriging can be expressed as dimensionless quantities, proportions of the unknown median of the conditional distribution. This scaled prediction interval can be used as a presurvey quality measure since it depends only on the sampling configuration and the variogram of the log-transformed variable. Second, we show how a similar scaled prediction interval can be computed for the median value of a lognormal variable across a block, in the case of block kriging. This approach is then illustrated using variograms of lognormally distributed data on concentration of elements in the soils of a part of eastern England.



## 1. Introduction

There is a growing awareness of the need to manage the soil sustainably, and as a result regulatory frameworks have been developed to ensure that soil quality is maintained (e.g. European Commission, 2006). As Bone et al. (2010) observe, the assessment of soil quality is challenging, potentially costly and prone to uncertainty because of the variability of soil material. It is therefore important that sampling schemes for soil assessment are carefully designed. De Gruijter et al. (2006) discuss how sampling can be planned so that questions about the soil are answered satisfactorily and efficiently. It is necessary to make best use of costly field and laboratory effort, and the results from sampling and analysis must be sufficiently precise to meet the end-user's requirements. De Gruijter et al. (2006) emphasize the importance of clearly identifying what these requirements are before the survey is planned. For example, the *target quantity* that we want to know might be the mean value of some variable across a region of interest, and an estimate of this is usually best achieved by an appropriately designed probability sample which entails randomization. If, alternatively, the user wants a set of local predictions (perhaps presented as a contour map) then this requires a more or less regular array of sample locations, and appropriate model-based statistical analyses. Having identified the nature of the question that sampling is to answer, we must also have some idea of how reliable the answer must be. This can be expressed by what de Gruijter et al. (2006) call *quality measures*.

A quality measure is a measure of the precision of an estimate from sample data. Once we have some data we can compute estimates of target quantities from them (e.g. means), and associated quality measures (e.g. confidence intervals). These are post-survey quality measures, which tell us, and users of the information, how well we have done. What we require for planning sampling are pre-survey quality measures, which tell us how well we can expect to do given a certain survey effort. Usually we can only approximate pre-survey quality measures (they may depend on estimates of values such as the variance of the target quantity in the population of interest that we can only approximate before sampling). Such pre-survey quality measures may be the expected width of the confidence interval for a target quantity, or the statistical power with which we can detect a change in the soil (Brus and Noij, 2008; de Gruijter et al., 2006; Lark, 2009).

Ideally we identify a quality measure that is appropriate for a particular sampling problem, and which can be approximated, pre-survey, from available information. We also ask the data user to

* Corresponding author. Tel.: +44 115 9363026.
*E-mail address:* mlark@nerc.ac.uk (R.M. Lark).

0016-7061/$ – see front matter 
doi:10.1016/j.geoderma.2011.12.008

specify values of the quality measure that are acceptable for their purposes. It should then be possible to plan a sampling campaign that will return information of suitable quality at acceptable minimal cost, or to show the user that this is not possible, and that it is necessary either to increase the budget to permit the collection of more samples or to accept that less precise estimates will be possible than originally hoped.

The kriging methods introduced to soil science by Burgess and Webster (1980), and further developments of these, are routinely used to produce local predictions of soil properties when such predictions are the required outcome from a soil inventory. Geostatistical methods are model-based in that they invoke an underlying random variable that is held to be realized in observed data, rather than depending on randomized sampling. The spatial dependence of this random variable is modelled by the variogram function. Local predictions are obtained as weighted averages of neighbouring observations of the variable, the weights being selected to minimize the expected squared error of the predictions. This quantity, called the kriging variance, is reported along with the prediction. It is a useful quality measure. Note that local predictions by point kriging are made on the original quasi-point support of the data — our observations are made on soil cores or similar specimens that are of very small dimensions by comparison to the region under study. As an alternative to point kriging we may estimate the mean value of the target variable over some region or block, which may be a regular rectilinear panel or an irregular region such as a field or similar management unit. This is called block kriging.

Geostatistical prediction by kriging is based on a random statistical model of the variable of interest which is inferred from data. Our data are treated as a realization of the underlying random model (de Gruijter et al., 2006). In kriging our target quantity is a point or block value specific to the realization, the block value is the spatial mean of the variable over the block's extent. The measures of uncertainty (kriging variances) are derived over the model distribution conditional on the observations.

Given the variogram function, the kriging variance (point or block kriging) for some variable at a particular location depends only on the configuration of sample sites. This makes the kriging variance a useful pre-survey quality measure. If we have an estimate of the variogram, perhaps from a survey of a neighbouring region, we can identify a sample network which ensures that the kriging variances of local predictions fall within an acceptable range. This was demonstrated by McBratney et al. (1981), and their approach has been applied to the design of soil surveys (Di et al., 1989; van Groenigen et al., 1999) including cases where the kriging prediction includes an external drift modelled by covariates such as remote sensor data (Brus and Heuvelink, 2007). More recently this work has been developed for the optimization of spatial surveys including both variogram estimation and prediction by kriging (Marchant and Lark, 2006, 2007; Zhu and Stein, 2006). Note that sound inference from the variogram requires that it has been estimated reliably. When data are prone to including outliers then robust variogram estimators may be needed, and the resulting model must be validated (Lark, 2000).

It is commonly found that soil and other geochemical variables do not appear to be normally distributed (Allègre and Lewin, 1995; White et al., 1987). This is best judged by exploratory statistics, such as the coefficient of skewness, and histograms of the data (Webster and Oliver, 2007). In these circumstances the data should be transformed to a scale of measurement on which an underlying normally distributed random variable can plausibly be assumed. Geostatistical predictions can be obtained on this new scale and then back-transformed to the scale of measurement. This is called trans-Gaussian kriging (Cressie, 1993). A common case is lognormal kriging, when the data are transformed to logarithms. Let $Y$ be the normal variable obtained by transformation of our original variable, $Z$, to natural logarithms. The ordinary point kriging of $Y$ at location $x_0$ is the conditional mean of the variable $Y(x_0)$, conditional on the observed values used for prediction, the random model (variogram) the assumption of a fixed but unknown local mean of $Y$ and the assumption that $Y$ is a normal random variable (Stein, 1999). The conditional distribution of $Y(x_0)$ has variance $\sigma_K^2(x_0)$, the kriging variance. However, for scientific or practical purposes we generally require predictions on the original scale. The ordinary point kriging prediction of $Y$ is back-transformed to $Z$, the corresponding variable on the original scale of measurement by

$$\tilde{Z}(x_0) = \exp\left\{\tilde{Y}(x_0) + \frac{\sigma_K^2(x_0)}{2} - \psi(x_0)\right\}, \quad (1)$$

where $\psi(x_0)$ is a Lagrange multiplier obtained in the solution of the kriging equations and $\tilde{Z}$ and $\tilde{Y}$ denote the kriging predictions of the respective random variables. The prediction error variance on the original scale of measurement can be written as

$$\left[\exp\left\{2\mu_Y + \sigma_K^2(x_0)\right\}\right] \times \left[\exp\left\{\sigma_K^2(x_0)\right\} + \exp\left\{\mathrm{Var}\left(\tilde{Y}(x_0)\right)\right\} - 2\exp\left\{\mathrm{Cov}\left(Y(x_0), \tilde{Y}(x_0)\right)\right\}\right], \quad (2)$$

where $\mu_Y$ is the mean of $Y$ and $\mathrm{Var}(\cdot)$ and $\mathrm{Cov}(\cdot,\cdot)$ denote, respectively the variance and covariance of the terms in brackets (Cressie, 1993). The key property of this latter expression is that the variance of the prediction depends on the mean of the variable. For this reason, unlike ordinary point or block kriging on the untransformed data, we cannot express the kriging variance as a quality measure dependent only on the variogram and the sampling design. In the log-normal case the kriging variance is therefore only useful as a post-survey quality measure, and cannot be used to select among different sampling designs before we have sampled a particular region.

The aim of this paper is to explore and demonstrate alternative quality measures that could be used for pre-survey planning of sampling for lognormally distributed variables. Some approaches are proposed for ordinary point and block kriging, and then illustrated with soil data from a baseline geochemical survey of part of eastern England.

## 2. Theory

### 2.1. The proposed quality measures: standardized prediction intervals and quantiles

As seen above, the kriging variance of a normally distributed random variable is a useful pre-survey quality measure for a sampling scheme because it depends only on the variogram of the variable and the sampling configuration. By contrast the mean square prediction error (kriging variance) at location $x_0$ on the untransformed scale depends, *inter alia* on $\mathrm{E}[Z(x_0)]$ conditional on the observations, so will not serve as a pre-survey quality measure. In this paper we propose quality measures based on prediction intervals rather than variances. A prediction interval of some random quantity $X$, $(L_\alpha(X), U_\alpha(X))$, is an interval with an assigned probability $\alpha$ such that

$$\mathrm{Prob}[L_\alpha(X) < X < U_\alpha(X)] = 1 - \alpha. \quad (3)$$

In this paper we consider prediction intervals which are symmetric in the sense that

$$\mathrm{Prob}[L_\alpha(X) < X < \mathrm{median}(X)] = \mathrm{Prob}[\mathrm{median}(X) < X < U_\alpha(X)] = \frac{1-\alpha}{2}, \quad (4)$$

where $\mathrm{median}(\cdot)$ denotes the median of a random variable. We show how we can compute the upper and lower bounds of prediction intervals for target quantities that might be obtained by lognormal point