European Journal of Soil Science, December 2006, **57**, 787–799 doi: 10.1111/j.1365-2389.2005.00768.x

On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML

R. M. Lark[a], B. R. Cullis[b] & S. J. Welham[c]
[a]*Silsoe Research Institute, Wrest Park, Silsoe, Bedford MK45 4HS, UK,* [b]*New South Wales Agricultural Research Institute, Wagga Wagga, NSW 2650, Australia, and* [c]*Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK*

Summary

Geostatistical estimates of a soil property by kriging are equivalent to the best linear unbiased predictions (BLUPs). Universal kriging is BLUP with a fixed-effect model that is some linear function of spatial co-ordinates, or more generally a linear function of some other secondary predictor variable when it is called kriging with external drift. A problem in universal kriging is to find a spatial variance model for the random variation, since empirical variograms estimated from the data by method-of-moments will be affected by both the random variation and that variation represented by the fixed effects.

The geostatistical model of spatial variation is a special case of the linear mixed model where our data are modelled as the additive combination of fixed effects (e.g. the unknown mean, coefficients of a trend model), random effects (the spatially dependent random variation in the geostatistical context) and independent random error (nugget variation in geostatistics). Statisticians use residual maximum likelihood (REML) to estimate variance parameters, i.e. to obtain the variogram in a geostatistical context. REML estimates are consistent (they converge in probability to the parameters that are estimated) with less bias than both maximum likelihood estimates and method-of-moment estimates obtained from residuals of a fitted trend. If the estimate of the random effects variance model is inserted into the BLUP we have the empirical BLUP or E-BLUP. Despite representing the state of the art for prediction from a linear mixed model in statistics, the REML–E-BLUP has not been widely used in soil science, and in most studies reported in the soils literature the variogram is estimated with methods that are seriously biased if the fixed-effect structure is more complex than just an unknown constant mean (ordinary kriging). In this paper we describe the REML–E-BLUP and illustrate the method with some data on soil water content that exhibit a pronounced spatial trend.

Introduction

Soil scientists must often base their advice to land managers on predictions of the values of soil properties, at points or over blocks of land, that are predicted from a few observations. Geostatistical methods give predictions of minimum and known variance. Our predictions are therefore in some sense optimal, and we can quantify their attendant uncertainty. The basis of geostatistical estimation is the assumption that a datum on a variable z at location $\mathbf{s}$ is a realization of a random field $Z(\mathbf{s})$. The random function is assumed to be intrinsically stationary, that is to say

$$\mathrm{E}[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})] = 0,\ \forall \mathbf{s} \in \Re^2, \tag{1}$$

Correspondence: R. M. Lark, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK. E-mail: murray.lark@bbsrc.ac.uk
Received 13 December 2004; revised version accepted 1 September 2005

and

$$\mathrm{E}[\{Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\}^2] = 2\gamma(\mathbf{h}),\ \forall \mathbf{s} \in \Re^2, \tag{2}$$

where $\mathbf{h}$ is a lag vector, the separation between two locations in two-dimensional real space ($\Re^2$), and $\gamma(\mathbf{h})$ is the variogram, a function of the lag only.

It is often reasonable to assume that soil variation is a realization of a random process, but there may be components of variation that cannot be treated in this way. For example, soil properties influenced by topography may show a pronounced trend across a study region, and this is not consistent with the first part of the intrinsic hypothesis in Equation (1). In such circumstances the experimental variogram estimated from the data may appear to be unbounded.

A trend may be represented by some polynomial function of co-ordinates, and it may often be possible to treat soil variation as composed of such a trend with superimposed random

© 2006 Silsoe Research Institute

variation. We can sometimes generalize this model to treat soil variation as a function of any smoothly varying quantity that is known at all sites (an external drift variable), and superimposed random variation. This is advantageous when the external drift variable is relatively cheap to measure, such as a topographic index derived from a digital elevation model, or a remotely sensed measurement of the land surface.

The problem for the soil scientist is first, to estimate the parameters of the two components of this model of soil variation in an unbiased way and, second, to combine the two components into a prediction of the soil variable at an unsampled site. The state of the art for this problem in frequentist statistics is the empirical best linear unbiased predictor (E-BLUP) based on residual maximum likelihood (REML) estimation of the spatial variance model (Patterson & Thompson, 1971). The best linear unbiased predictor (BLUP) of some variable is computed from a linear mixed model – i.e. a model that contains fixed and random effects. If the random effect is assumed to be an intrinsically stationary random field then the BLUP is equivalent to kriging (Stein, 1999), and if normality can be assumed then the BLUP is the conditional expectation of the variable at the prediction site (conditional on our data).

More specifically, if the fixed effect is just an unknown mean then the BLUP is equivalent to ordinary kriging; if the fixed effect is some linear combination of predictors (such as spatial coordinates) then the BLUP is equivalent to universal kriging (UK). The problem for implementation of a BLUP is that it requires a variance model. A common solution in the ordinary kriging case is to estimate the variogram empirically for different lag intervals with a method-of-moments estimator such as that of Matheron (1962), then to fit a variogram function (e.g. Webster & Oliver, 2001). It is widely recognized that the main practical obstacle to the implementation of UK is how to obtain a variogram for the random variation (Cressie, 1993; Goovaerts, 1997) since the method-of-moments estimator will be influenced by both the spatial trend and the variation that we treat as random. The E-BLUP solution is to estimate the parameters of the variogram model by residual maximum likelihood (REML), and to use these to derive the (co)variances which are 'plugged in' to obtain the E-BLUP.

Our reason for writing this paper is to illustrate the REML–E-BLUP for spatial prediction of soil properties in the presence of a trend (or by incorporation of some other external drift variable). The soil science literature contains various examples of the E-BLUP, usually under the title universal kriging or regression kriging. However, the problem of how to acquire the variance model is commonly addressed either on the basis of *ad hoc* assumptions that lack generality, or by methods that are generally biased. This matters, because this bias will invalidate the estimation variances that we compute for the predictions. A principal advantage of geostatistics is that the uncertainty attendant on a prediction is known. Further, the variance model may be used to guide further sampling or a new sampling exercise in a cognate landscape, and if the variance is biased then the sampling scheme may not match our requirements. Given this, the virtues of the unified REML–E-BLUP framework do not appear to have received the attention that they merit from soil scientists. In the following section, we first review how the E-BLUP is currently used in the soil science community, and we then present and illustrate the REML–E-BLUP scheme.

Theory

E-BLUP in soil science, and the problem of estimating variance parameters

Universal kriging (UK) was proposed by Matheron (1969). In UK we consider the following model for $Z(\mathbf{s})$:

$$Z(\mathbf{s}) = \sum_{k=0}^{K} \beta_k f_k(\mathbf{s}) + \eta(\mathbf{s}), \tag{3}$$

where the $f_k(\mathbf{s})$, $k = 0, 1, \ldots, K$ is a specified set of functions of the co-ordinate vectors (such as polynomials) and the β_k are coefficients which must be estimated. The term $\eta(\mathbf{s})$ is a random field with zero mean and with a variogram $\gamma_\eta(\mathbf{h})$. The UK estimate at some target site $\mathbf{s}_\mathrm{p}$, based on N observations at neighbouring sites, is

$$\widehat{Z}(\mathbf{s}_\mathrm{p}) = \sum_{i=1}^{N} \lambda_i z(\mathbf{s}_i), \tag{4}$$

where the λ_i are kriging weights. These weights are optimized to minimize the kriging variance, subject to the unbiasedness constraints for $k = 0, \ldots, K$:

$$\sum_{i=1}^{N} \lambda_i f_k(\mathbf{s}_i) = f_k(\mathbf{s}_\mathrm{p}). \tag{5}$$

Webster & Oliver (2001) provide more detail. The UK estimate is the best linear unbiased predictor (BLUP) of $Z(\mathbf{s}_\mathrm{p})$ based on the trend model $f_k(\mathbf{s})$, $k = 0, 1, \ldots, K$ (Stein, 1999). However, in order to obtain the E-BLUP we must know $\gamma_\eta(\mathbf{h})$. This variogram cannot be estimated directly from data with the method-of-moments estimators commonly used for ordinary kriging (e.g. Matheron, 1962) since the random variation is only one component of the variable that we observe (Cressie, 1993). In practice, $\gamma_\eta(\mathbf{h})$ is sometimes estimated on the assumption that at short lags the effect of the trend is negligible and that at longer lags we can restrict the pair comparisons used to estimate the variogram – $z(\mathbf{s}_i)$, $z(\mathbf{s}_i + \mathbf{h})$ – to those for which

$$\sum_{k=0}^{K} \sum_{i=1}^{N} \lambda_i \beta_k f_k(\mathbf{s}_i) \approx \sum_{k=0}^{K} \sum_{i=1}^{N} \lambda_i \beta_k f_k(\mathbf{s}_i + \mathbf{h}). \tag{6}$$

So, for example, if inspection of the data suggested the presence of a linear trend in a particular direction, then we might estimate the variogram only from pair comparisons that are

approximately perpendicular to the trend. This, of course, entails the assumption that $\gamma_\eta(\mathbf{h})$ is isotropic. This approach is advocated in a geostatistics text widely used by soil scientists (Goovaerts, 1997). An example of its application to soil data is presented by Meul & Van Meirvenne (2003).

The second approach to prediction in the presence of a trend has been adopted because it is difficult to obtain an estimate of the variogram $\gamma_\eta(\mathbf{h})$ for UK. Again, we use some specified function of spatial co-ordinates to model the deterministic trend. However, the coefficients, β_k, are estimated by ordinary least squares (OLS), and the estimated coefficients, $\widehat{\beta}_k$, are then used to compute residuals at each site where the target variable has been observed:

$$\widehat{\eta}(\mathbf{s}_i) = z(\mathbf{s}_i) - \sum_{k=0}^{K} \widehat{\beta}_k f_k(\mathbf{s}_i). \tag{7}$$

The variogram of these residuals is then estimated by a method-of-moments estimator such as that of Matheron (1962), and modelled with a continuous function of lag distance. At any target site, $\mathbf{s}_\mathrm{p}$, we may then obtain an estimate of the deterministic component of the target variable, $\Sigma_{k=0}^{K} \widehat{\beta}_k f_k(\mathbf{s}_\mathrm{p})$, then add to this an estimate of the random component that we obtain by kriging from the residuals, $\widehat{\eta}(\mathbf{s}_i)$. If we generalize this method so that the argument of the deterministic function may be some external drift variable, then we have one form of regression kriging, as used by Odeh *et al.* (1995). Regression kriging has been widely used in soil science – for example, by Terra *et al.* (2004) and Braimoh *et al.* (2005). Note that universal kriging and regression kriging are equivalent formally. In practice, universal kriging entails simultaneous estimation of the trend and spatially dependent random components of $Z(\mathbf{h})$, while in regression kriging the trend model is estimated as an initial step. This will lead to different predictions if the universal kriging is conducted in a local neighbourhood rather than on all the data.

The problem with this approach is that the variogram estimated from the residuals is a biased estimate of the variogram of the random component of spatial variation (Cressie, 1993). In short, this is because the point estimates of this variogram depend in a non-linear way on the estimates of the nuisance parameters, $\widehat{\beta}_k$, and so they are biased even though the $\widehat{\beta}_k$ are not.

This problem is illustrated in Figure 1. We simulated 5000 realizations of a standard normal random field at 100 points located on a regular linear transect of unit interval. This was done by Cholesky factorization of the covariance matrix (Cressie, 1993), the matrix having been computed for a linear variogram with a sill at a range of 40 units, and with zero nugget. Figure 1(a) shows the mean values of point estimates of this variogram at several lags, obtained from the realizations. The 95% confidence interval for each mean is also shown, and the continuous line is the specified variogram. The mean of the estimates of the variogram for any lag is close to the theoretical value, which always lies within the confidence interval. We then repeated the simulation, but added a simple linear trend to the simulated values so that the random variable at the ith position on the transect is

$$z(i) = \beta_0 + \beta_1 i + \eta(i), \tag{8}$$

where $\eta(i)$ is the random variable with a bounded linear variogram, and in this case $\beta_0 = 0$ and $\beta_1 = 5$. Once one realization of this process had been generated, the parameters of the linear trend, β_0 and β_1, were estimated by ordinary least squares. The estimates, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, were then used to compute the residuals, $z(i) - \widehat{\beta}_0 - \widehat{\beta}_1 i$, and the variogram was estimated from these. This was repeated for each realization. The results are shown in Figure 1(b). The bias at long lags is very apparent; note that by contrast the 95% confidence interval for the mean of each of the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ included the specified value.

This bias will have two consequences. First, the overall variability of the random variable is underestimated. Second, because the bias increases with lag (in fact the bias is quadratic with lag; Cressie, 1993) the experimental variogram, and models fitted to it, will not represent the spatial dependence of the random variation correctly.

A solution sometimes advanced for this problem (e.g. by Hengl *et al.*, 2004) is to use the variogram from the OLS residual to re-estimate the parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by generalized least squares (GLS). The variogram from the residuals of this model is then used in a further GLS estimation, and this is iterated until the estimates of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ no longer change. However, the point estimates of the variogram still depend on nuisance parameters, so this method does not remove the bias, although it may reduce it (Gambolati & Galeati, 1987). Kitanidis (1993) showed that the variogram can be used to estimate the generalized covariance function of data; i.e. the covariance function plus some arbitrary function of lag. In the case of a linear trend, the arbitrary function is a quadratic and Kitanidis showed that it does not affect the universal kriging predictions, or the prediction error variances, because of the unbiasedness constraints in the universal kriging equations with a linear trend model. However, this finding is of practical use only if the function fitted to point estimates of the variogram incorporates this arbitrary function, which requires that we correctly identify the variogram model and the order of the trend model. This will be difficult in practice, and would add to the uncertainty in the final variance model. For this reason, it is not surprising that Kitanidis (1993) concluded that 'a more objective and promising approach [than modelling experimental variograms of residuals]' is to use residual maximum likelihood, and it is to this that we now turn.

The REML–E-BLUP solution for prediction in the presence of a trend

Stein (1999) has argued for the use of residual maximum likelihood (REML) estimation in combination with the E-BLUP

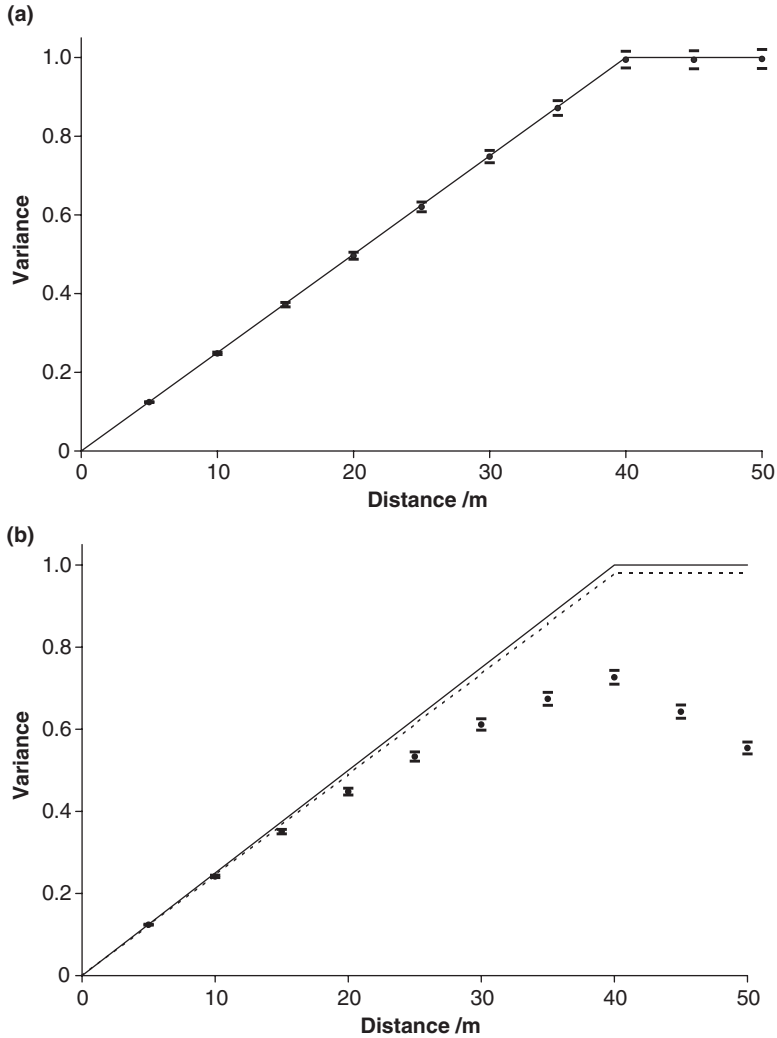

Figure 1 Mean (solid symbol) and 95% confidence interval (bars) for point estimates of (a) the experimental variogram of 5000 realizations of a random variable with a bounded linear variogram; the solid line is the specified variogram; (b) the corresponding results for the variogram of a variable with a linear trend superimposed on the same random process, the trend being estimated by ordinary least squares and subtracted before variogram estimation. The broken line on (b) shows the variogram with parameters equal to the mean of all REML estimates over the 5000 realizations.

as a general scheme for spatial interpolation, and this is the approach that we present below. The model of spatial variation of data that we use in geostatistics is a special case of the linear mixed model. The linear mixed model comprises an additive combination of one or more fixed effects (such as a trend model), one or more random effects (which will be the spatially dependent random variables in a geostatistical context) and an independent random variable. If we are to estimate variance parameters for the random effects (a variogram model in geostatistics), then REML has the advantages of the likelihood framework (it is asymptotically efficient and consistent) along with an adjustment that reduces the bias found in maximum likelihood estimates of variance parameters. This is discussed by Cressie (1993).

The REML method is due to Patterson & Thompson (1971). Kitanidis (1987) used REML to estimate parameters of a spatial variance model, and Zimmerman & Zimmerman (1991) discussed REML estimation of variogram parameters and their use for ordinary kriging. More recently Gilmour *et al.* (2004) and Welham *et al.* (2004) have discussed the E-BLUP of linear mixed models, including the particular case of universal kriging. It is this state of the art that we describe below.

There are three elements in the computation of the E-BLUP of a variable, z, at an unsampled site, $\mathbf{s}_p$ (although the average information algorithm of Gilmour *et al.* (1995) allows these to be computed in a single step). The first is the estimation of a variance structure (e.g. a variogram) for some specified linear mixed model. This is then used to obtain estimates of the model coefficients that we need to form the E-BLUP. The estimated variance model is then used to compute the E-BLUP, $\tilde{z}_p$. These latter two steps are equivalent to the solution of the universal kriging equations and application of the resulting kriging weights, with the variance model estimated in the first step.

Estimation of the variance models. In a previous paper (Lark & Cullis, 2004), we described how REML allows the inference, from n observations, of the variance parameters of a linear mixed model. We may write the model as

$$\mathbf{z} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{9}$$

where the vector $\mathbf{z}$ contains our n observations, $\mathbf{X}$ is an $n \times p$ design matrix that associates each of the n observations with a value of each of the p fixed effects (such as terms of a polynomial trend model), and the vector $\boldsymbol{\tau}$ contains the p fixed-effect coefficients. The vector $\mathbf{u}$ contains q random effects, realizations of a variable u, that are associated with the n observations by the $n \times q$ design matrix $\mathbf{Z}$. Typically, with only one observation at each location, $q = n$ and $\mathbf{Z}$ is the identity matrix. Here we assume that u is the spatially dependent random variable and the term $\boldsymbol{\varepsilon}$ is a vector of independent random errors. These terms are independent of each other, thus

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2\xi\mathbf{G} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I} \end{bmatrix} \right), \tag{10}$$

where σ^2 is the variance of the independent error, ξ is the ratio of the variance of u to σ^2 and $\mathbf{G}$ is the correlation matrix of $\mathbf{u}$. Note that we are making an explicit assumption that the random terms are jointly Gaussian. The term $\boldsymbol{\varepsilon}$ represents both independent measurement errors and variation that arises from processes that are spatially dependent over shorter distances than separate our samples; this is the nugget in geostatistical terms. If we assume that $\mathbf{u}$ is drawn from a second-order stationary random process then the correlation matrix $\mathbf{G}$ will depend only on the relative locations of our observations given some specified correlation function $C(\cdot)$ with one or more parameters that characterize the spatial dependence:

$$\mathbf{G}_{i,j} = \mathrm{Corr}\ [\mathbf{u}(\mathbf{s}_i), \mathbf{u}(\mathbf{s}_j)] = C(\mathbf{s}_i - \mathbf{s}_j). \tag{11}$$

The correlation function may be the spherical or exponential function, familiar in geostatistics, although others may be used. For example, Stein (1999) argues for the Matérn function, which is particularly useful when variation is locally smooth. If we use the exponential function then there is a single distance parameter, a, that must be estimated and

$$C(\mathbf{s}_i - \mathbf{s}_j) = \exp\left\{ \frac{-|\mathbf{s}_i - \mathbf{s}_j|}{a} \right\}. \tag{12}$$

If we use the spherical correlation function then the distance parameter is the range, a, and

$$\begin{aligned} C(\mathbf{s}_i - \mathbf{s}_j) &= 1 - \frac{3|\mathbf{s}_i - \mathbf{s}_j|}{2a} + \frac{1}{2}\left(\frac{|\mathbf{s}_i - \mathbf{s}_j|}{a}\right)^3 \text{if } |\mathbf{s}_i - \mathbf{s}_j| < a \\ &= 0 \text{ otherwise.} \end{aligned} \tag{13}$$

Note that both these functions, as expressed, describe isotropic variation; the variogram depends only on the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$. The correlation function could be more complex with parameters that describe spatial anisotropy. The parameters of this function, which we represent by the vector $\boldsymbol{\theta}$, along with σ^2 and ξ, are estimated by REML. This removes dependence of the estimates on the fixed effects $\boldsymbol{\tau}$ which are nuisance parameters in this problem and which would increase the bias of estimates based on maximum likelihood or method-of-moments (Smyth & Verbyla, 1996).

The residual log-likelihood function, Equation (14) below, has the unknown terms σ^2, ξ and $\boldsymbol{\theta}$ as its arguments, conditional on the data $\mathbf{z}$. The residual log-likelihood is

$$\begin{aligned} \ell_{\mathrm{R}}(\sigma^2, \xi, \boldsymbol{\theta}|\mathbf{z}) = -\frac{1}{2}\{&\log|\mathbf{H}| + \log|\mathbf{X}^{\mathrm{T}}\mathbf{H}\mathbf{X}| + (n-p)\sigma^2 + \\ &\frac{1}{\sigma^2}\mathbf{z}^{\mathrm{T}}(\mathbf{I} - \mathbf{W}\mathbf{C}^{-1}\mathbf{W}^{\mathrm{T}})\mathbf{z}\}, \end{aligned} \tag{14}$$

where $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$ and $\mathbf{H} = \xi\mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}$. The estimates of σ^2, ξ and $\boldsymbol{\theta}$ that maximize $\ell_{\mathrm{R}}(\sigma^2, \xi, \boldsymbol{\theta}|\mathbf{z})$ are found numerically. The average information (AI) algorithm (Gilmour *et al.*, 1995) can do this efficiently. However, the spherical correlation function is not smooth (Mardia & Watkins, 1989) and this causes problems for the AI algorithm which uses gradient methods. Lark & Cullis (2004) used simulated annealing to find REML estimates in these circumstances, and that is what we have done here.

We revisited the simulation shown in Figure 1 and used REML to estimate the variogram of the random variation in the presence of the same trend. The dotted line in Figure 1(b) shows the variogram with the average value of the estimated parameters over all 5000 realizations. The 95% confidence interval for the mean range is 39.999–40.006 (specified value was 40), and for the sill variance 0.975–0.985 (specified value 1.0). The nugget variance was estimated as zero in all cases (the specified value). The range therefore appears to be estimated without bias, although there is a very small bias in the variance, but this is negligible by comparison with the bias in the experimental variogram of the OLS residuals.

Estimation of the fixed and random effects. Once we have obtained estimates of the variance parameters by REML, we can estimate the fixed and random effects by solution of the mixed model equation:

$$\mathbf{C}\begin{bmatrix} \widehat{\boldsymbol{\tau}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{z} \end{bmatrix}, \tag{15}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \xi^{-1}\mathbf{G}^{-1} \end{bmatrix}.$$

The solution of this equation returns us estimates of the fixed effects, $\widehat{\boldsymbol{\tau}}$, and predictions of the random effect, $\tilde{\mathbf{u}}$. We follow the convention of denoting an estimate of a fixed effect by $\hat{a}$, and a predictor of some random quantity (or a function of such a predictor such as the E-BLUP of $z(\mathbf{s})$) by $\tilde{a}$. The covariance matrix for the error of the estimates is

$$\mathrm{Cov}\begin{bmatrix} \widehat{\boldsymbol{\tau}} - \boldsymbol{\tau} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix} = \sigma^2\mathbf{C}^{-1} = \sigma^2\begin{bmatrix} \mathbf{C}^{1,1} & \mathbf{C}^{1,2} \\ \mathbf{C}^{2,1} & \mathbf{C}^{2,2} \end{bmatrix}. \tag{16}$$

Computation of the E-BLUP estimates. The next step is to form predictions at an unsampled site. Welham *et al.* (2004) and Gilmour *et al.* (2004) have recently shown how mixed

models should be used for prediction. We want a prediction at a new point $\mathbf{s}_p$, where the fixed effects are in the $p \times 1$ vector $\mathbf{x}_p$. The value that we predict is $f_p = \mathbf{x}_p^T \boldsymbol{\tau} + u_p$. This prediction is marginal on the error in our observations ($\boldsymbol{\varepsilon}$). That is to say, we assume that the term $\boldsymbol{\varepsilon}$ is due to measurement error, and so any estimate at some $\mathbf{s}_p$ that is one of our observation sites is not necessarily equal to our observed value at the site which is $z_p = f_p + \varepsilon_p$. This is by contrast to the usual ordinary kriging estimates which are conditional on $\boldsymbol{\varepsilon}$. At any site, $\mathbf{s}_p$, that is not one of our observation sites, the estimates marginal or conditional on $\boldsymbol{\varepsilon}$ are the same since the E-BLUP of ε_p is zero. For practical purposes the fact that our E-BLUP is marginal on $\boldsymbol{\varepsilon}$ is only of importance for our computation of the prediction variance, and we discuss this in more detail below.

The predicted value is:

$$\begin{aligned}\tilde{f}_p &= \mathbf{x}_p^T \hat{\boldsymbol{\tau}} + \tilde{u}_p \\ &= \mathbf{x}_p^T \hat{\boldsymbol{\tau}} + \mathbf{g}_{o,p}^T \mathbf{G}^{-1} \tilde{\mathbf{u}},\end{aligned} \tag{17}$$

where $\mathrm{Cov}[\mathbf{u}, u_p] = \xi\sigma^2 \mathbf{g}_{o,p}$. The vector $\mathbf{g}_{o,p}$ is therefore computed from our REML estimates of the variance parameters.

When we examine the expression for the E-BLUP in Equation (17), it is clear that the first term represents the prediction based on the polynomial trend model or external drift variables, that are included in the design matrix $\mathbf{X}$. The second term is the kriging estimate of the spatially dependent random effect.

Computation of the E-BLUP estimation variance. The estimation variance of the E-BLUP is

$$\mathrm{Var}\,[\tilde{f}_p - f_p] = \mathrm{Var}\left[\mathbf{x}_p^T(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) + \mathbf{g}_{o,p}^T \mathbf{G}^{-1} \tilde{\mathbf{u}} - u_p\right]. \tag{18}$$

It can be shown (see Appendix) that Equation (18) gives rise to the following formula for the estimation variance of the E-BLUP. The first term accounts for the uncertainty due to estimation error of the fixed and random effects, the second term is uncertainty of the interpolated random effect, that is a kriging variance.

$$\begin{aligned}\mathrm{Var}\,[\tilde{f}_p - f_p] = \sigma^2 &\left[\mathbf{x}_p, \mathbf{g}_{o,p}^T \mathbf{G}^{-1}\right]^T \mathbf{C}^{-1} \left[\mathbf{x}_p, \mathbf{g}_{o,p}^T \mathbf{G}^{-1}\right] + \\ &\xi\sigma^2 \left(g_{p,p} - \mathbf{g}_{o,p}^T \mathbf{G}^{-1} \mathbf{g}_{o,p}\right).\end{aligned} \tag{19}$$

This estimation variance is appropriate when all the independent error, $\boldsymbol{\varepsilon}$, can be attributed to measurement error. If this is the case, then Equation (19) is the mean square error of the E-BLUP as a predictor of the true value of the variable at $\mathbf{s}_p$. In practice this assumption is questionable for many soil properties. While measurement error is a component of the nugget variance, much is attributable to actual sources of variation in the soil that are only spatially dependent over short distances, and so do not appear correlated in our data. This variance should be accounted for when we quantify the uncertainty of the BLUP at $\mathbf{s}_p$. This can be handled in the context of the linear mixed model if we undertake duplicate measurements of the variable on material from the same site. These duplicate observations can be used to estimate the measurement error as a separate random component, not confounded with short-range spatial variation. In these circumstances, the design matrix for the random effect is not an identity matrix. When such information is not available then we cannot obtain an exact prediction variance since the component of $\boldsymbol{\varepsilon}$ that is attributable to measurement error is unknown. We can identify an upper bound on the prediction variance, by assuming that the measurement error is zero and so that all the nugget variance is due to actual variation of the soil. This is what is done in most geostatistical surveys by soil scientists, although it is rarely made explicit. In this case the prediction variance is

$$\begin{aligned}\mathrm{Var}\,[\tilde{f}_p - z_p] = \sigma^2 &\left\{\left[\mathbf{x}_p, \mathbf{g}_{o,p}^T \mathbf{G}^{-1}\right]^T \mathbf{C}^{-1} \left[\mathbf{x}_p, \mathbf{g}_{o,p}^T \mathbf{G}^{-1}\right] + \right. \\ &\left. \xi\left(g_{p,p} - \mathbf{g}_{o,p}^T \mathbf{G}^{-1} \mathbf{g}_{o,p}\right) + 1\right\},\end{aligned} \tag{20}$$

since, as in Equation (10), the two random terms in the model are independent.

Case study

The case study uses data on the gravimetric water content of the topsoil (0–200 mm) on Cashmore field at Silsoe in England (latitude 52°0.44′N, longitude 0°25.13′W). These data were collected on a 50-m square grid supplemented with additional sample sites to give a total of 100 observations. The data were collected in the spring of 1995 when the soil was at field capacity. The data are presented as a post-plot in Figure 2. This shows distinct (increasing) trends in water content from north to south and east to west. This trend is explicable from what we know about the soils of this field. A previous soil survey of the field (Lark *et al.*, 1998) showed that the north of the field overlies the Lower Greensand (Cretaceous sands). The Gault Clay (also Cretaceous) is downfaulted against the Lower Greensand with the boundary running approximately west–northwest to east–southeast in the southern half of the field. These solid formations are overlaid by superficial material. In the north of the field, the eastern area is dominated by coarse colluvium, while relatively heavy-textured drift is found to the west. The southwest of the field has fine loamy drift over the Gault, while the southeast includes Evesham series soil formed in swelling clay to the surface, and loamy textured alluvium. As a result of this, the driest soils are in the northeast of the field and the wettest in the southwest.

Table 1 presents summary statistics for these data, and a histogram is shown in Figure 3. Note that the data are somewhat skew, but the post-plot suggests that this may be largely attributable to the trend. The experimental variogram of the data is shown in Figure 4. Since there were only 100 data, we limited ourselves to the consideration of isotropic variograms.

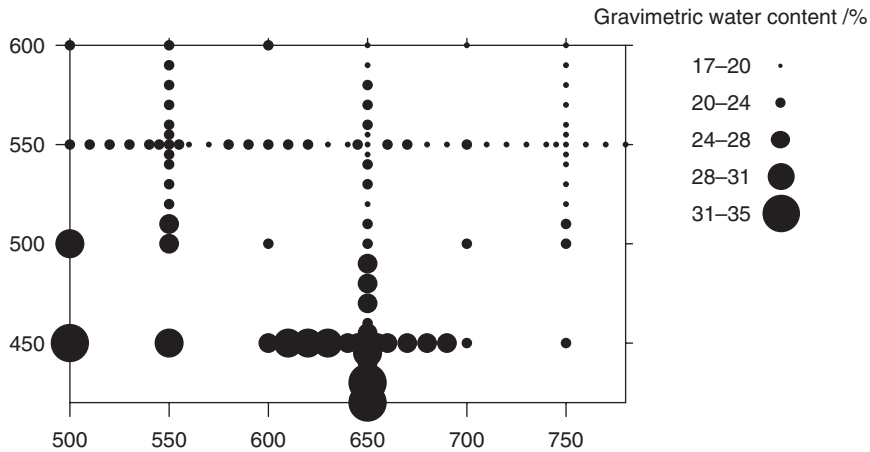

Figure 2 Post-plot of gravimetric water content at observation sites on Cashmore field. Co-ordinates are in metres relative to a local origin at UK Ordnance Survey 508000, 235000.

The unbounded form of the variogram also indicates the presence of a spatial trend.

We used the method of simulated annealing to find the REML estimates of variance parameters for four different models. We used this method because two of the models had a spherical correlation function for the random effect which is best estimated in this way (Lark & Cullis, 2004). First we fitted a simple linear trend model,

$$z = \beta_0 + \beta_1 x + \beta_2 y + u + \varepsilon, \tag{21}$$

where x and y are eastings and northings and the coefficients of the polynomial trend, β_0, β_1 and β_2, are the fixed effects. The term ε is an independent random variable of variance σ^2, and u is a random variable of variance $\xi\sigma^2$. The random effect, u, has a correlation function with a parameter a – the range of a spherical correlation function, Equation (13), or the distance parameter of an exponential correlation function, Equation (12). The variance parameters σ^2, ξ and a were estimated by REML. The fixed effects (β_0, β_1 and β_2) and random effects were then estimated with Equation (15). We fitted two models for a linear trend, one with a spherical and one with an exponential correlation function for the random effect. We then fitted quadratic trend models,

$$z = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + u + \varepsilon, \tag{22}$$

again specifying exponential and spherical correlation functions, and the same variance parameters as for Equation (21). The estimated variance parameters and fixed effects are shown in Table 2.

Table 1 Summary statistics for the data on gravimetric water content (%)

Mean	22.3
Median	21.3
Standard deviation	3.5
Skew	1.2

The likelihoods indicate that the exponential spatial variance model is preferred for the linear trend and the spherical variance model for the quadratic trend, although the difference is not large. Note that we can only compare two models on their residual likelihoods when their fixed-effect structure is the same. Thus we may use the residual likelihood to select the spherical or exponential variance model for a linear trend, but not to choose between a linear and a quadratic trend model. It is notable that the variance for the best-fitting linear trend model is large, and in fact the distance parameter is also large (247 m), indicating spatial dependence at distances up to 750 m, substantially larger than the dimensions of the area under study. The distance parameter for the spherical variance model is much smaller. In Figure 5, we show the residuals from the linear (Figure 5a) and quadratic (Figure 5b) trend models. There is a clear systematic structure to the former with the large negative

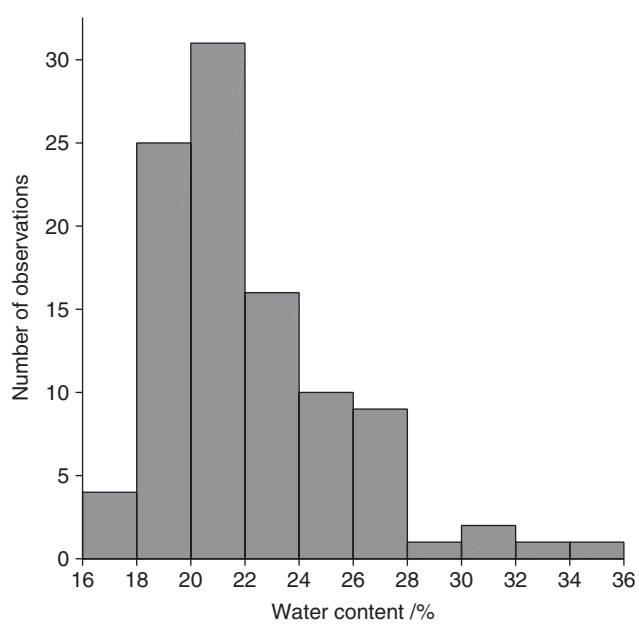

Figure 3 Histogram of observed soil water contents.

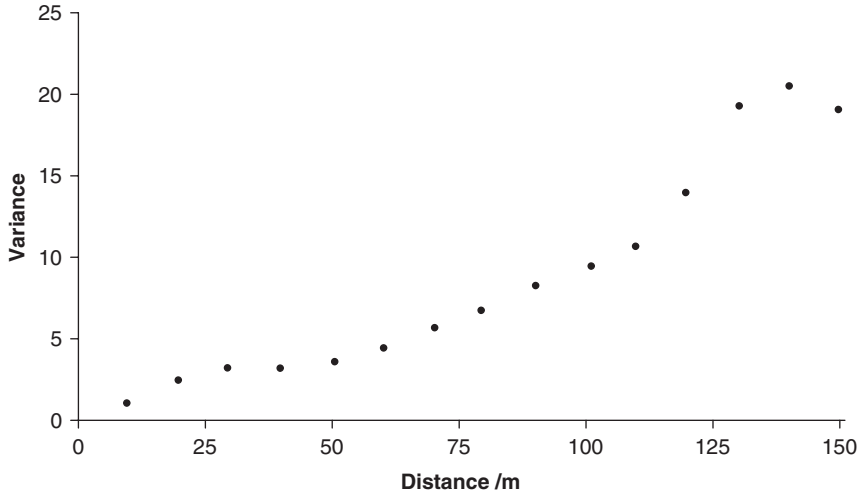

Figure 4 Experimental variogram of observed soil water contents.

values near the northern and southern boundaries of the field. This suggests that higher-order terms are required in the trend model. This is supported by the Wald statistics calculated to test the null hypotheses that each of the quadratic terms in Equation (22), β_3, β_4 and β_5, are zero. The Wald statistics were tested against the *F*-distribution and the denominator degrees of freedom were adjusted for small sample size after Kenward & Roger (1997). The Wald statistics were 0.32, 26.8 and 5.5 for β_3, β_4 and β_5, respectively. The null hypothesis may therefore be accepted for β_3 but rejected for β_4 and β_5 ($P = 1.6 \times 10^{-5}$ and $P = 0.002$, respectively). Figure 6 shows the experimental variograms (symbols) for the residuals from both trend models that we computed for exploratory purposes; they are of course biased. Since we are restricted to lags up to 150 m, in line with usual practice to minimize edge effects, the effect of the structure in the residuals from the linear trend model is not apparent.

Table 2 Results for fitting different trend models to soil water content on Cashmore field

Trend model: Variogram model:	Linear[a] Exponential	Linear[a] Spherical	Quadratic[b] Exponential	Quadratic[b] Spherical
Variance model				
σ^2	0.54	0.52	0.48	0.59
Parameter a/m[c]	247.2	75.5	18.8	35.3
ξ	24.0	5.67	3.0	1.94
Fixed effects				
Estimates				
β_0	71.9	67.6	226.4	223.6
β_1	−0.022	−0.021	−10.95	−10.79
β_2	−0.064	−0.060	−57.1	−56.3
β_3	–	–	0.211	0.183
β_4	–	–	4.211	4.114
β_5	–	–	1.186	1.228
Standard errors				
β_0	11.9	4.8	35.3	30.9
β_1	0.013	0.005	5.52	4.64
β_2	0.016	0.007	10.12	8.93
β_3	–	–	0.383	0.318
β_4	–	–	0.891	0.784
β_5	–	–	0.583	0.520
Log-likelihood	−85.62	−86.96	−70.37	−70.04

[a]As in Equation (21).
[b]As in Equation (22).
[c]Distance parameter of the exponential correlation function and range of the spherical correlation function.

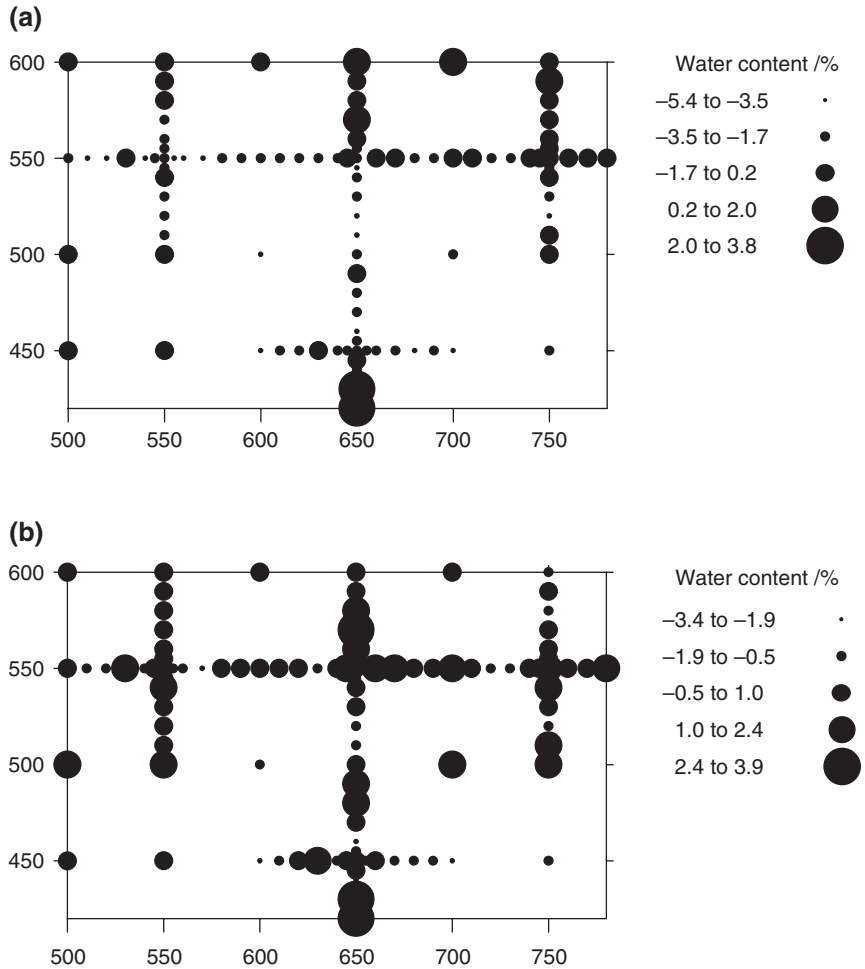

Figure 5 Post-plot of residuals from (a) linear trend model and (b) quadratic trend model.

Figure 6 also shows the REML estimates of the variograms of the random variation. The parameters of these are in Table 2. On each graph the solid line represents the variogram with the larger likelihood, and the broken line is the alternative with smaller likelihood. Note that the selected (spherical) and the alternative variogram for the random variation about the quadratic trend are very similar. The selected and alternative variogram for the linear trend model are very similar at lags up to 60 m, which will dominate the covariance matrix of these data, but diverge at longer lags. This probably reflects artefacts due to the poor fit of the linear trend (notably the large residuals from the trend near the northern and southern field margins).

We therefore selected the quadratic trend model, and the version estimated with a spherical variance model given that the log-likelihood was larger than for the exponential model.

Table 3 shows the summary statistics of the residuals from the selected trend model and Figure 7 their histogram. The residuals are symmetrically distributed, supporting the assumption of a Gaussian random process superimposed on a trend; the trend was responsible for the skew in the raw data.

We then computed the E-BLUPs of soil water content at points on a regular grid across the field, using Equation (17) with the estimated fixed and random effects for the selected model. We estimated the variance of the estimates using Equation (20) to map the uncertainty of the predictions. These are the variances of the estimates conditional on the uncorrelated random variation in the data, $\boldsymbol{\varepsilon}$. They are thus an upper bound on the error variance of the E-BLUP as a prediction of the true values, equivalent to the assumption that there is no measurement error so all the nugget variance is due to real soil variation.

These results are shown in Figure 8. This shows both the smooth trend from the driest soils in the northeast corner to the wettest in the southwest, described by the fixed effects of the model, but also short-range variation arising from the kriging estimate of the error at each target site. The estimation variance is smallest near the observation sites (an effect of the kriging component of the estimation).

Discussion

In this paper we have strongly advocated the REML–E-BLUP method. It is worth reflecting on its possible limitations.

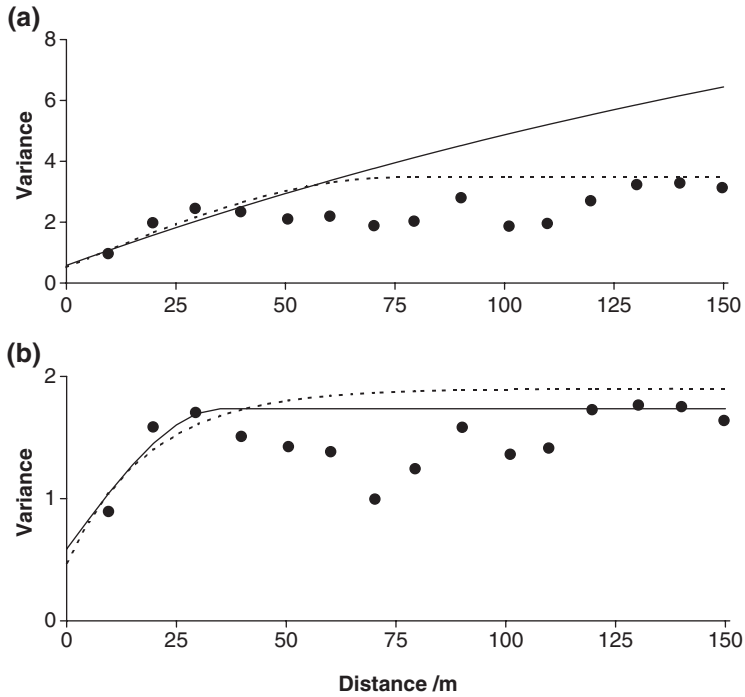

Figure 6 Experimental variogram (solid symbols) of residuals from (a) linear trend model and (b) quadratic trend model. The lines are the REML estimates (with parameters in Table 2). The solid line corresponds to the selected variogram, and the dotted line to the alternative for each trend model.

First, unlike regression kriging, REML estimation entails an explicit assumption that **u** and $\boldsymbol{\varepsilon}$ have a joint Gaussian distribution. Many soil variables cannot be plausibly regarded as normally distributed on the scales on which they are most naturally or conveniently measured. It is therefore necessary, as we have done here, to study the descriptive statistics, histograms and postplots of residuals from the trend models, and where necessary consider appropriate transforms. We can only evaluate the plausibility that the marginal distribution of the data is Gaussian in this way, and the assumption of a joint Gaussian distribution is not necessarily made safe. As noted by Pardo-Igúzquiza (1998a) we only have one realization of the full joint distribution of a variable, so the assumption is, in principle, unverifiable. He went on to argue that the assumption of a joint Gaussian distribution may be justified because it is the distribution of maximum entropy given the mean and covariance matrix. This rationale is in concordance with other recent developments in geostatistics (Christakos, 2000). Furthermore, Kitanidis (1985) showed, using simulated data, that likelihood methods to estimate spatial variance models are robust to departures from a normal distribution and actually performed better with non-normal data than did method-of-moments estimation of the experimental variogram. We accept that the assumption of a joint Gaussian distribution remains a problem for REML. Alternatives might be based on hierarchical generalized linear models (Lee & Nelder, 1996) or penalized quasi-likelihood (Breslow & Clayton, 1993) but these methods require fuller evaluation in a spatial context. Bayesian methods (Diggle *et al.*, 2003) may also offer greater robustness, but have their own limitations, notably the dependence on assumptions entailed in the prior distributions.

Table 3 Summary statistics for the residuals from the trend model for gravimetric water content (%)

Mean	0.13
Median	0.14
Standard deviation	1.21
Skew	0.08

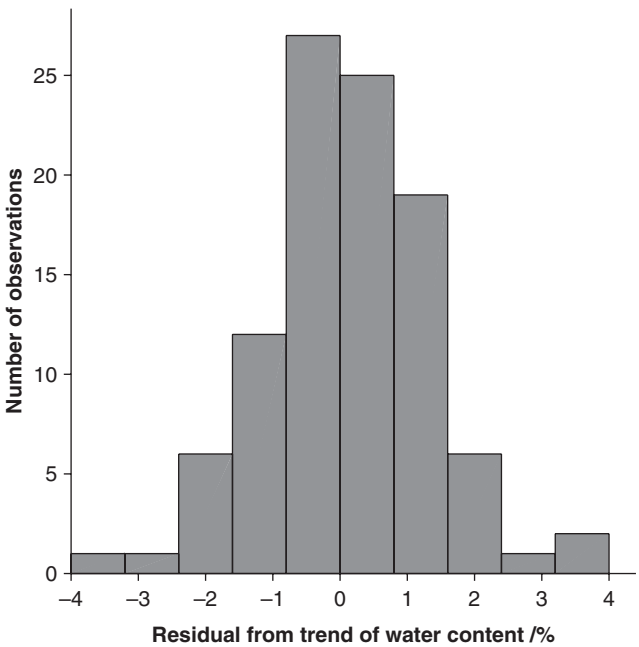

Figure 7 Histogram of residual of soil water contents from the quadratic trend model.

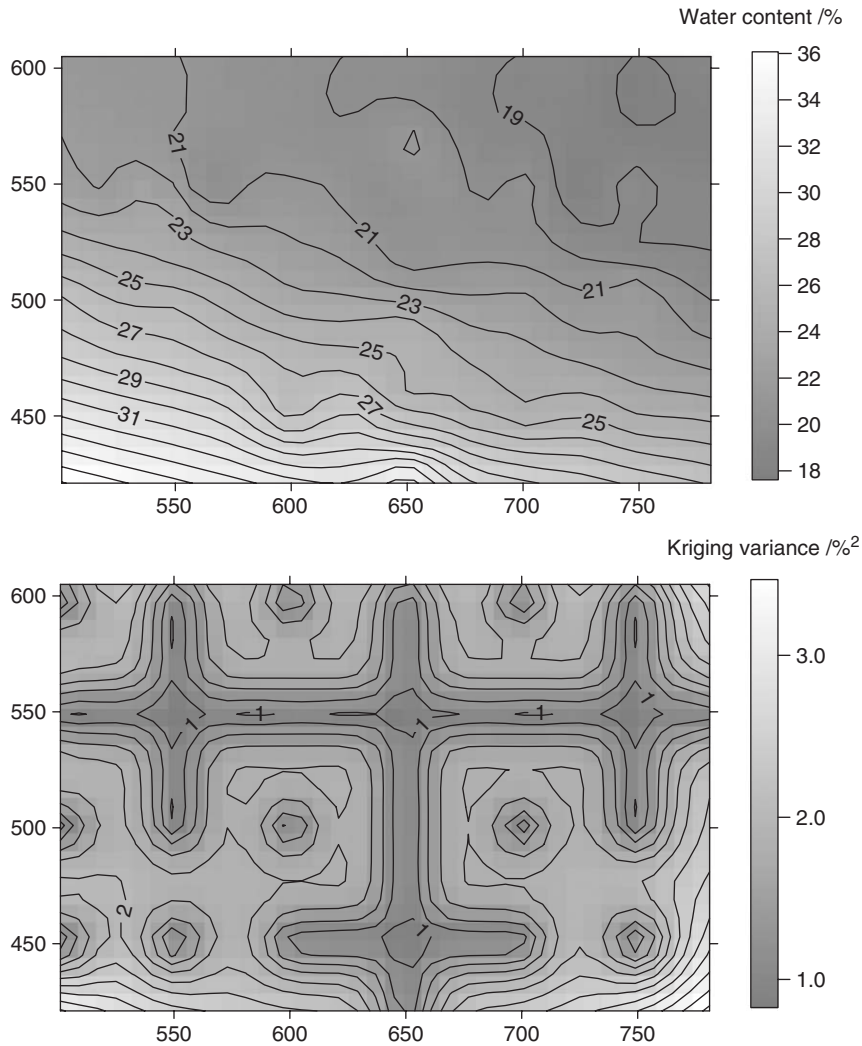

Figure 8 (a) BLUP of soil water content on Cashmore field; (b) variance of estimates.

Second, as noted above and described by Mardia & Watkins (1989), the likelihood function for a spherical variogram can be multimodal, because it is not differentiable at the range. This requires that we use methods to find the maximum likelihood that are less efficient than the AI algorithm. Stein (1999) argues for the more general use of Matérn correlation functions, or exponential functions which are both asymptotic, and doubts the general applicability of functions that cut off sharply at the range. However, spherical variograms have been widely used in soil science, and seem appropriate, for example, to describe variation that originates from pronounced and more or less uniformly spaced boundaries. Stein (1999) suggested the use of a squared spherical function for circumstances where a cut-off is required, and this should be investigated further.

Third, likelihood-based methods require numerical optimization. In the past this has been thought to limit the size of the data sets that can be analysed, but the AI algorithm of Gilmour *et al*. (1995) is efficient and can be run with larger data sets (a few thousand points). A problem remains for analysing large data sets with spherical variogram functions. It might be possible to combine the AI algorithm with simulated annealing, by using the latter just to find the distance parameter, and Stein *et al*. (2004) consider approximate methods for the analysis of large data sets based on likelihood.

A fourth consideration, that applies generally to E-BLUP including regression kriging, is that if data sets are small then the sampling error of the variance model may cause bias when it is "plugged in" to obtain the E-BLUP. A Bayesian approach to estimation has been suggested (Diggle *et al*., 2003) to account for the uncertainty of estimation of the variance parameters implicit in the computation of the prediction error variance.

Conclusions

The REML–E-BLUP procedure can be used for spatial prediction of soil properties which show a spatial trend, and has advantages, both theoretical and practical, over other methods that soil scientists have used to address this problem. Stein (1999) wrote

that 'REML and best linear unbiased prediction for [intrinsically stationary random functions] form a coherent conceptual package'. We see the practical advantages of this coherence in the case study where the computation of a variance model, the fixed effects and the predictions are done with no *ad hoc* decisions about how to approximate the variogram. The procedure is even more compact and efficient when we use the average information (AI) algorithm of Gilmour *et al.* (1995). We obtain estimates of all parameters of this model, based on REML estimation of the variance parameters (which are less biased than regression kriging approaches based on OLS), and we can use all data to estimate the variance parameters of the random effect. By predicting the soil property at unsampled sites with the E-BLUP we combine both the trend surface as a predictor with the best estimate of the local random contribution. We also obtain an estimate of the local prediction error variance.

For these reasons we would strongly advocate the use of REML–E-BLUP by soil scientists for problems in spatial prediction where a trend surface must be modelled or an external drift variable can be exploited. In fact, the approach may be used more generally. Stein (1999) casts doubt on the general suitability of the standard approach to the estimation of a variance model, based on estimates of the variogram for particular lags or lag classes. His particular concern is that the strong correlation between estimates of the variogram for different lags can lead to biased estimates of the parameters of fitted models of the variogram. While methods exist to circumvent this problem they entail a degree of approximation (since the correct weighted least squares solution depends on the variogram itself). We would therefore agree with Stein (1999) and Diggle *et al.* (2003) that the REML–E-BLUP approach is used for all spatial estimation problems unless data sets are very large.

Soil scientists can use REML–E-BLUP for their prediction problems with software from existing packages (although in most cases they should be cautious when using spherical correlation functions since these might not be reliably estimated for reasons described above). The ASReml package (Gilmour *et al.*, 2002) is very efficient, and the same algorithms are available in Genstat (Payne, 2003, Chapter 5) and the SAMM package for the S-PLUS environment, although with a smaller range of correlation functions. Similar facilities are available in the MIXED procedure of SAS and in geoR (Ribeiro & Diggle, 2001). Some Fortran code is described by Pardo-Igúzquiza (1998b) and is available from the website of the International Association for Mathematical Geology (http://www.iamg.org/CGEditor/index.htm).

Acknowledgements

The data from Cashmore field were collected as part of work funded by the UK Department for Environment, Food and Rural Affairs under project CE0204. Helen Wheeler's contribution to the field work is acknowledged. R.M.L.'s work was funded by the Biotechnology and Biological Sciences Research Council through its Competitive Strategic Grant to Silsoe Research Institute. We are grateful to two anonymous referees for comments on this paper.

References

Braimoh, A.K., Stein, A. & Vlek, P.L.G. 2005. Identification and mapping of associations among soil variables. *Soil Science*, **170,** 137–148.

Breslow, N.E. & Clayton, D.G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88,** 9–25.

Christakos, G. 2000. *Modern Spatiotemporal Geostatistics*. Oxford University Press, New York.

Cressie, N.A.C. 1993. *Statistics for Spatial Data*. John Wiley & Sons, New York.

Diggle, P.J., Ribeiro, P.J. & Christensen, O.F. 2003. An introduction to model-based geostatistics. In: *Lecture Notes in Statistics, Number 173: Spatial Statistics and Computational Methods* (ed. J. Möller), pp. 43–86. Springer-Verlag, New York.

Gambolati, G. & Galeati, G. 1987. Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels – comment. *Mathematical Geology*, **19,** 249–257.

Gilmour, A.R., Thompson, R. & Cullis, B.R. 1995. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51,** 1440–1450.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. & Thompson, R. 2002. *ASReml User Guide*, Release 1.0. VSN International, Hemel Hempstead.

Gilmour, A., Cullis, B., Welham, S., Gogel, B. & Thompson, R. 2004. An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis*, **44,** 571–586.

Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.

Hengl, T., Heuvelink, G.B.M. & Stein, A. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, **120,** 75–93.

Kenward, M.G. & Roger, J.H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53,** 983–997.

Kitanidis, P.K. 1985. Minimum-variance unbiased quadratic estimation of covariances of regionalised variables. *Journal of the International Association of Mathematical Geology*, **17,** 195–208.

Kitanidis, P.K. 1987. Parametric estimation of covariances of regionalized variables. *Water Resources Bulletin*, **23,** 557–567.

Kitanidis, P.K. 1993. Generalized covariance functions in estimation. *Mathematical Geology*, **25,** 525–540.

Lark, R.M. & Cullis, B.R. 2004. Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, **55,** 799–813.

Lark, R.M., Catt, J.A. & Stafford, J.V. 1998. Towards the explanation of within-field variability of yield of winter barley: soil series differences. *Journal of Agricultural Science*, **131,** 409–416.

Lee, Y. & Nelder, J.A. 1996. Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58,** 619–673.

Mardia, K.V. & Watkins, A.J. 1989. On multimodality of the likelihood in the spatial linear model. *Biometrika*, **76,** 289–295.

Matheron, G. 1962. *Traité de Géostatistique Appliqué*, Tome 1. Memoires du Bureau de Recherches Géologiques et Minières, Paris.

Matheron, G. 1969. *Le krigeage universel*. Cahiers du Centre de Morphologie Mathématique, No 1. Ecole des Mines de Paris, Fontainebleau.

Meul, M. & Van Meirvenne, M. 2003. Kriging soil texture under different types of nonstationarity. *Geoderma*, **112,** 217–233.

Odeh, I.O.A., McBratney, A.B. & Chittleborough, D.J. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, **67,** 215–226.

Pardo-Igúzquiza, E. 1998a. Maximum likelihood estimation of spatial covariance parameters. *Mathematical Geology*, **30,** 95–108.

Pardo-Igúzquiza, E. 1998b. Inference of spatial indicator covariance parameters by maximum likelihood using MLREML. *Computers and Geosciences*, **24,** 453–464.

Patterson, H.D. & Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58,** 545–554.

Payne, R.W. (ed.) 2003. *The Guide to Genstat, Part 2: Statistics*. VSN International, Oxford.

Ribeiro, P.J. & Diggle, P.J. 2001. geoR: a package for geostatistical analysis. *R-News* **1,** 15–18. At: http://cran.r-project.org/doc/Rnews

Smyth, G.K. & Verbyla, A.P. 1996. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58,** 565–572.

Stein, M.L. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

Stein, M.L., Chi, Z. & Welty, L.J. 2004. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, **66,** 275–296.

Terra, J.A., Shaw, J.N., Reeves, D.W., Raper, R.L., van Santen, E. & Mask, P.L. 2004. Soil carbon relationships with terrain attributes, electrical conductivity, and a soil survey in a coastal plain landscape. *Soil Science*, **169,** 819–831.

Webster, R. & Oliver, M.A. 2001. *Geostatistics for Environmental Scientists*. John Wiley & Sons, Chichester.

Welham, S., Cullis, B., Gogel, B., Gilmour, A. & Thompson, R. 2004. Prediction in linear mixed models. *Australian and New Zealand Journal of Statistics*, **46,** 325–347.

Zimmerman, D.L. & Zimmerman, M.B. 1991. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, **33,** 77–91.

Appendix

The estimation variance of the E-BLUP

As stated in Equation (18) we require $\mathrm{Var}\left[\mathbf{x}_\mathrm{p}^\mathrm{T}(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau})+\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\tilde{\mathbf{u}}-u_\mathrm{p}\right]$. We use

$$\widehat{\boldsymbol{\tau}}=(\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{z} \quad \text{and} \quad \tilde{\mathbf{u}}=\xi\mathbf{G}\mathbf{Z}^\mathrm{T}\mathbf{P}\mathbf{z},$$

with

$$\begin{bmatrix}\mathbf{C}^{1,1} & \mathbf{C}^{1,2}\\ \mathbf{C}^{2,1} & \mathbf{C}^{2,2}\end{bmatrix}=$$
$$\sigma^2\begin{bmatrix}(\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X})^{-1} & -\xi(\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{Z}\mathbf{G}\\ -\xi\mathbf{G}\mathbf{Z}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X})^{-1} & \xi\mathbf{G}-\xi^2\mathbf{G}\mathbf{Z}^\mathrm{T}\mathbf{P}\mathbf{Z}\mathbf{G}\end{bmatrix}$$

where

$$\mathbf{H}\equiv\xi\mathbf{Z}\mathbf{G}\mathbf{Z}^\mathrm{T}+\mathbf{I}, \quad \text{and}$$
$$\mathbf{P}\equiv\mathbf{H}^{-1}-\mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}.$$

The variances of the two terms in the square brackets in Equation (18) are, respectively,

$$\mathrm{Var}\left[\mathbf{x}_\mathrm{p}^\mathrm{T}(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau})\right]=\sigma^2\mathbf{x}_\mathrm{p}^\mathrm{T}\mathbf{C}^{1,1}\mathbf{x}_\mathrm{p}$$

and

$$\mathrm{Var}\left[\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\tilde{\mathbf{u}}-u_\mathrm{p}\right]=\xi\sigma^2\left\{g_{\mathrm{p,p}}-\xi\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{Z}^\mathrm{T}\mathbf{P}\mathbf{Z}\mathbf{g}_{\mathrm{o,p}}\right\},$$

where $\xi\sigma^2 g_{\mathrm{p,\ p}}=\mathrm{Var}[u_\mathrm{p}]$ (so $g_{\mathrm{p,p}}$ is usually 1). We may therefore write

$$\begin{aligned}\mathrm{Var}\left[\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\tilde{\mathbf{u}}-u_\mathrm{p}\right]&=\xi\sigma^2\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}(\mathbf{G}-\xi\mathbf{G}\mathbf{Z}^\mathrm{T}\mathbf{P}\mathbf{Z}\mathbf{G})\mathbf{G}^{-1}\mathbf{g}_{\mathrm{o,p}}+\\&\quad\xi\sigma^2\left(g_{\mathrm{p,p}}-\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\mathbf{g}_{\mathrm{o,p}}\right)\\&=\sigma^2\mathbf{C}^{2,2}+\xi\sigma^2\left(g_{\mathrm{p,p}}-\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\mathbf{g}_{\mathrm{o,p}}\right).\end{aligned}$$

The covariance of these terms is given by

$$\begin{aligned}\mathrm{Cov}\left[\mathbf{x}_\mathrm{p}^\mathrm{T}(\widehat{\tau}-\tau),\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\tilde{\mathbf{u}}-u_\mathrm{p}\right]&=-\sigma^2\xi\mathbf{x}_\mathrm{p}^\mathrm{T}(\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{H}^{-1}\mathbf{Z}\mathbf{g}_{\mathrm{o,p}}\\&=\sigma^2\mathbf{x}_\mathrm{p}^\mathrm{T}\mathbf{C}^{1,2}\mathbf{G}^{-1}\mathbf{g}_{\mathrm{o,p}}.\end{aligned}$$

Then we can write Equation (18) as

$$\begin{aligned}\mathrm{Var}\,[\tilde{f}_\mathrm{p}-f_\mathrm{p}]&=\sigma^2\left[\mathbf{x}_\mathrm{p},\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\right]^\mathrm{T}\mathbf{C}^{-1}\left[\mathbf{x}_\mathrm{p},\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\right]+\\&\quad\xi\sigma^2\left(g_{\mathrm{p,p}}-\mathbf{g}_{\mathrm{o,p}}^\mathrm{T}\mathbf{G}^{-1}\mathbf{g}_{\mathrm{o,p}}\right).\end{aligned}$$