European Journal of Soil Science, December 2004, **55**, 799–813 doi: 10.1111/j.1365-2389.2004.00637.x

Model-based analysis using REML for inference from systematically sampled data on soil

R. M. LARK[a] & B. R. CULLIS[b]

[a]*Silsoe Research Institute, Wrest Park, Silsoe, Bedford MK45 4HS, UK, and* [b]*New South Wales Agricultural Research Institute, Wagga Wagga, NSW 2650, Australia*

Summary

The general linear model encompasses statistical methods such as regression and analysis of variance (ANOVA) which are commonly used by soil scientists. The standard ordinary least squares (OLS) method for estimating the parameters of the general linear model is a design-based method that requires that the data have been collected according to an appropriate randomized sample design. Soil data are often obtained by systematic sampling on transects or grids, so OLS methods are not appropriate.

Parameters of the general linear model can be estimated from systematically sampled data by model-based methods. Parameters of a model of the covariance structure of the error are estimated, then used to estimate the remaining parameters of the model with known variance. Residual maximum likelihood (REML) is the best way to estimate the variance parameters since it is unbiased. We present the REML solution to this problem. We then demonstrate how REML can be used to estimate parameters for regression and ANOVA-type models using data from two systematic surveys of soil.

We compare an efficient, gradient-based implementation of REML (ASReml) with an implementation that uses simulated annealing. In general the results were very similar; where they differed the error covariance model had a spherical variogram function which can have local optima in its likelihood function. The simulated annealing results were better than the gradient method in this case because simulated annealing is good at escaping local optima.

Introduction

A common task for the soil scientist is to make inferences about soil properties in a region (e.g. to estimate regional means, to compare means between subregions such as soil map units or to fit a regression line for predicting a property from a readily measured variable). In order to do this we must sample the soil, and when the property of interest has been determined for each sample unit make an appropriate statistical analysis. Linear models are commonly used for this purpose. As we discuss below in more detail, these include familiar statistical analyses such as linear regression and the analysis of variance. There are two general methods for fitting these models and making inferences with them. These are ordinary least squares and model-based methods. In this paper we draw the attention of soil scientists to circumstances in which the ordinary least squares approach should not be used, and describe and exemplify model-based analysis.

Ordinary least squares

Ordinary least squares methods form the cornerstone of most introductory statistics texts and courses for environmental scientists. They are widely used in soil science, but not always appropriately. The methods are not appropriate unless the data have been collected by design-based sampling in which any possible sample unit is a member of an underlying population. The whole population constitutes all the soil in the region of interest. Since sample units are generally very small by comparison with the region we usually assume that the population is infinite. The aim of sampling and inference in the design-based approach is to obtain estimates of the underlying parameters of the population. The key idea is that the probability that a particular sample unit is included in the sample is determined by the sample design and so is known (de Gruijter & ter Braak, 1990; Papritz & Webster, 1995). Simple random sampling and stratified random sampling are well-known

This paper is one of the series 'Mathematical Geology – Series for Students' sponsored by the International Association of Mathematical Geology in association with the *European Journal of Soil Science*.

Correspondence: R. M. Lark. E-mail: murray.lark@bbsrc.ac.uk

Received 30 September 2003; revised version accepted 19 January 2004

examples of the design. In the former all sample units are selected at random and independently of each other. In the latter sample units are selected at random and independently within each stratum. This allows us to proceed on the assumption that the sample data are independent random variables regardless of any underlying spatial pattern in the variable (de Gruijter & ter Braak, 1990). This assumption allows us to use ordinary least squares for estimation and inference.

Such inference is generally done with the general linear model. This has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observed values of a soil variable y, $\mathbf{X}$ is an $n \times p$ 'design matrix' of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of model coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors. In ordinary least squares we assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, i.e. since $\mathbf{I}$ is the identity matrix, that the errors are independently and identically distributed (iid) random variables of mean zero and variance σ^2.

If the first column of $\mathbf{X}$ is a vector of 1s and columns 2 to p contain observations of $p-1$ continuous variables then Equation (1) is equivalent to the multiple linear regression with the elements of $\boldsymbol{\beta}$ equal to the regression coefficients. This model may be used to predict the variable y from measurements of more rapidly or cheaply measured variables, as in the widely used pedotransfer functions. Webster (1997) has discussed regression in more detail and explained the limited circumstances in which it is appropriate.

The design matrix may contain p dummy variables or indicators, each corresponding to one of a set of p comprehensive, mutually exclusive classes (such as soil map units or land use classes). In this case Equation (1) is equivalent to the one-way analysis of variance (ANOVA) model. This model may be used to make inferences about the differences between the classes, and to derive predictors for use at sites within the classes where the classes are, for example, soil map units or physiographic classes.

Continuous predictors and indicators may be combined in a single design matrix to give more complex models.

Under the assumption that the variables in $\boldsymbol{\varepsilon}$ are random and iid, an ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{2}$$

with $\mathbf{C}$, the covariance matrix of the coefficients, estimated by

$$\widehat{\mathbf{C}} = \widehat{\sigma}^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \tag{3}$$

where $\widehat{\sigma}^2$, the estimate of the error variance, is obtained by

$$\widehat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right). \tag{4}$$

In the design-based approach the justification of the assumption that the errors are mutually independent is based on the sample design, in which units are selected at random and independent of each other. If the error variables are from a normally distributed process then the OLS estimate of the model coefficients in Equation (2) is equivalent to the maximum likelihood estimate. An explanation of the derivation of the OLS estimator of $\boldsymbol{\beta}$ and its relation to the maximum likelihood estimator is given in the Appendix.

Model-based analysis

Model-based analysis is an alternative to OLS. In model-based analysis (exemplified by geostatistics) we assume that the variable is a realization of a random process. In geostatistics the process is distributed in a space of one, two or three dimensions and is the random function $Z(\mathbf{x})$. The actual value of the variable at a particular location, $z(\mathbf{x})$, is a realization of this random function. For model-based inference we must postulate an underlying model. Usually this model describes the spatial dependence of values of a realization of the random function at different locations in terms of the spatial separation between the locations. This is done on the basis of knowledge and experience, and evidence from the data. We must also make some assumptions about the model, e.g. that the random function is ergodic, or second-order stationary or intrinsically stationary (see Webster, 2000).

The model provides the basis for treating our observations as outcomes of random variables with particular properties. There is no reason in model-based sampling and analysis why samples should be drawn at random and independently of each other, and they usually are not. This is because we do not rely on the sampling to justify treating sample units as independent. We specifically do not treat the units as independent, but rather we assume that they exhibit a spatial dependence that is characterized by the model. When we make estimates and inferences from observations in model-based analysis we do so on the basis of the model that describes this spatial structure.

When we have not sampled with an appropriate randomized design we must use a model-based method for analysis, and we may not use OLS. If we did use OLS to analyse such data our estimates of the parameters in $\boldsymbol{\beta}$ are unbiased, but their variance is not obtained correctly with Equation (3). Systematic sampling, by definition, does not give rise to independent observations. For a specified sampling scheme (e.g. sampling on a square grid with 100-m intervals) once the location of the origin of the grid and its orientation are determined, then all the sample points have been specified. While there may be good reasons for randomizing the position of the origin and the orientation of the grid (to avoid biases) the sample design no longer allows us to treat our data as independently drawn, as if in design-based sampling.

This is an important point, because all too often investigators analyse systematically sampled data with OLS methods, as if the units had been drawn independently and at random.

Soil is often sampled systematically on grids or transects in survey and there are good reasons why this is so. In reconnaissance studies of a new region it is rational to sample regularly on transects since it covers the range of environmental variation efficiently (Brink *et al.*, 1982), as well as being simple to implement (Cochran, 1977). Such transects may generate data that we wish to explore using linear modelling to decide, for example, whether regression of observed soil properties on remote sensor data gives adequate predictions in the particular landscape, or whether particular physiographic units differ significantly with respect to soil properties. As another example, data might have been collected with the primary aim of mapping a variable by kriging, and the application of linear modelling arises subsequently, or is of secondary importance. Given the cost of field sampling and soil analysis we should make full use of the data that accrue.

Model-based analyses should therefore be used to analyse data from systematic samples, but they may also be used to analyse data from randomized samples since this can bring gains in efficiency if there is spatial dependence between the sample points. This is well established in the analysis of designed experiments (Gilmour *et al.*, 1997).

Here we draw the attention of soil scientists to the model-based methods that exist for fitting general linear models to data. In particular we discuss the fitting of these models by residual maximum likelihood (REML). There is a rich literature on model-based analysis of designed experiments (e.g. Gilmour *et al.*, 1997), but rather less on analysis of sample data where the primary objective is not geostatistical mapping. Below we describe how general linear models with a model of the spatial dependence of the error variation are fitted using REML. In a subsequent section we demonstrate the method using some data on soil.

Theory

In this paper we treat the combined effects of all sources of variation in $\mathbf{y}$ that are not accounted for by the fixed effects in $\boldsymbol{\beta}$ as a random variate $\boldsymbol{\eta}$, so

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}. \qquad (5)$$

By contrast to the ordinary least squares case, Equation (1), we do not assume that the elements of $\boldsymbol{\eta}$ are iid, but rather we specify $\boldsymbol{\eta} \sim \mathcal{N}(0,\mathbf{V})$. The matrix $\mathbf{V}$ is a variance–covariance matrix of the error variables. Our model-based analysis consists in finding an appropriate parametric form for this matrix, estimating these parameters and then using the estimate of $\mathbf{V}$ to obtain an estimate of $\boldsymbol{\beta}$.

We assume that the covariance matrix $\mathbf{V}$ is positive definite, that is to say its structure implies that all combinations of the n variables will have a positive variance. In the case of spatial variables we may usefully assume that the error is a second-order stationary regionalized variable, and so that the structure of $\mathbf{V}$ is determined entirely by the spatial distribution of the sample sites. Thus

$$\mathbf{V}_{i,j} = C(\mathbf{x}_i - \mathbf{x}_j), \qquad (6)$$

where C is a covariance function of the vector $\mathbf{x}_i - \mathbf{x}_j$, the separation (lag) between observations at $\mathbf{x}_i$ and $\mathbf{x}_j$. In this discussion we simplify further by assuming that $\mathbf{V}_{i,j}$ depends only on the lag distance, and not on the direction (i.e. that the variation is isotropic). When this assumption is not tenable then a more complex covariance function can describe dependence of the covariance on both the lag distance and the direction.

The covariance function must be an authorized function so that $\mathbf{V}$ is always positive definite. When the functional form of C has been chosen the covariance matrix may then be characterized by a vector $\boldsymbol{\theta}$ of q variance parameters including the variance, σ^2, and additional parameters that describe the spatial dependence. As an example, the isotropic exponential covariance function is

$$\begin{aligned} \mathbf{V}_{i,j} &= \sigma^2 s \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right), \quad & i \neq j \\ & \sigma^2, & i = j, \end{aligned} \qquad (7)$$

where $|\mathbf{x}_i - \mathbf{x}_j|$ denotes the (scalar) lag distance between the two locations, a is a distance parameter, and s is a second parameter, the spatial dependence, i.e. the proportion of the variance that has a spatial structure as defined by the exponential function. In geostatistical terms $\sigma^2(1-s)$ is equal to the nugget variance. A variable with a covariance function defined in Equation (7) has the familiar exponential variogram

$$\gamma(|\mathbf{x}_i - \mathbf{x}_j|) = c_0 + c\left\{1 - \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right)\right\}, \qquad (8)$$

where c_0 and c are the nugget and spatially structured variance components, so that the spatial dependence is

$$s = \frac{c}{c_0 + c}. \qquad (9)$$

By incorporating the nugget effect into our model of $\mathbf{V}$ we can model sources of error that are spatially structured, but also components of error that appear spatially uncorrelated (at least at the scales of our sampling). In this case q, the number of variance parameters in $\boldsymbol{\theta}$, is 3, and they are σ^2, s and a, so that $\boldsymbol{\theta} \equiv [\sigma^2, s, a]$.

When a covariance matrix $\mathbf{V}$ has been estimated then we may estimate $\boldsymbol{\beta}$ by inserting the estimate into the generalized least squares equations:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}, \qquad (10)$$

with the covariance matrix

$$\widehat{\mathbf{C}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}. \qquad (11)$$

In summary, to fit the model in Equation (5) we first estimate the variance subset of parameters $\boldsymbol{\theta}$, and then use these to estimate the remaining parameters $\boldsymbol{\beta}$, which are fixed effects. Investigators sometimes do this in spatial modelling by finding the OLS estimate of $\boldsymbol{\beta}$ using Equation (2), then estimating and modelling a variogram of the residuals in the usual way. The variogram of the OLS residuals then provides parameters in $\boldsymbol{\theta}$ for matrix $\mathbf{V}$. The problem with this approach is that the variance parameters obtained are biased (Cressie, 1993). Another approach is to find a maximum likelihood (ML) estimate for the joint set of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. If we can assume that our data are from a normal distribution then the log-likelihood function is

$$l(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}) = \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \quad (12)$$

where the elements of $\mathbf{V}$ depend on variance parameters. This is discussed in more detail in the Appendix. Estimates of all the parameters are found that maximize this likelihood function. Cook & Pocock (1983) and Mardia & Marshall (1984) describe this procedure in the context of multiple linear regression. The ML estimates of the variance parameters $\boldsymbol{\theta}$ are subject to bias, however. The ML estimate of the variance depends on $\boldsymbol{\beta}$, the elements of which are called nuisance parameters in this context. Stuart *et al.* (1999) point out that this dependence introduces the bias. The ML estimate of a parameter that is a function of another, i.e. $\tau(\eta)$, is $\tau(\widehat{\eta})$ where $\widehat{\eta}$ is the ML estimate of η. If $\widehat{\eta}$ is unbiased then in general $\tau(\eta)$ will be biased since

$$\mathrm{E}[\tau(\widehat{\eta})] \neq \tau(\mathrm{E}[\widehat{\eta}]), \quad (13)$$

see Equation (18.28) of Stuart *et al.* (1999).

This bias could be avoided if the dependence of the parameters $\boldsymbol{\theta}$ on the nuisance parameters in $\boldsymbol{\beta}$ could be removed. One does this by defining a new likelihood function in which the variance parameters $\boldsymbol{\theta}$ are variables but the likelihood is conditional on the parameters in $\boldsymbol{\beta}$. The parameters in $\boldsymbol{\theta}$ are then estimated by maximization of this likelihood. We may then compute the matrix $\mathbf{V}$ and insert it into the generalized least squares equations, Equations (10) and (11), to obtain an estimate of $\boldsymbol{\beta}$. This is the basis of the method of residual maximum likelihood (REML) introduced by Patterson & Thompson (1971).

Smyth & Verbyla (1996) give an illuminating presentation of REML for the linear model. One can interpret REML as an exact conditional likelihood where the conditioning is on a statistic wholly sufficient for $\boldsymbol{\beta}$, i.e. it summarizes all the information about the parameters $\boldsymbol{\beta}$ contained in the data. The estimate of $\boldsymbol{\theta}$ obtained by maximizing this conditional likelihood does not depend on the nuisance parameters. In the case of a general linear model a statistic $\mathbf{t}$ can be defined:

$$\mathbf{t} \equiv \mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}, \quad (14)$$

where $\mathbf{D}$ is any non-singular $p \times p$ matrix function of $\boldsymbol{\theta}$. Thus defined, $\mathbf{t}$ is completely sufficient for $\boldsymbol{\beta}$. The REML estimate of $\boldsymbol{\theta}$ maximizes the likelihood conditioned on $\mathbf{t}$. Following Smyth & Verbyla's (1996) presentation, Stuart *et al.* (1999) present the conditional log-likelihood function for the general linear model as

$$l\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}\right) = \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{I}-\mathbf{Q})\mathbf{y}, \quad (15)$$

where

$$\mathbf{Q} \equiv \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}. \quad (16)$$

Having found parameters $\boldsymbol{\theta}$ that maximize the conditional log-likelihood, we may substitute the estimated covariance matrix $\mathbf{V}$ into Equations (10) and (11) to obtain the REML estimates of the parameters $\boldsymbol{\beta}$ and their covariance matrix.

Residual maximum likelihood is now widely used in applied statistics when it is necessary to obtain unbiased estimates of variance components and OLS estimation is not possible. One such case is in the analysis of repeated measurements on a single set of experimental subjects. Here the errors will have a dependence induced by temporal autocorrelation. Webster & Payne (2002) discussed how REML can be used in these conditions, with an example from soil science.

Gilmour *et al.* (1995) present an algorithm for REML estimation in mixed modelling problems such as we have presented here. This is incorporated into the ASReml software (Gilmour *et al.*, 2002). This procedure uses a gradient method to find the maximum log-likelihood, and is fast. Gradient methods can have problems finding maximum log-likelihoods if the likelihood function is not smooth since they may find and stick at local optima and miss the global optimum. Spatial models with a spherical variogram for the random error do not have smooth likelihood functions (Warnes & Ripley, 1987). However, the spherical model is very commonly used in geostatistics to model bounded variograms which reach the sill variance at a finite range. For this reason, while we use ASReml to analyse our data in this paper, we also use a simulated annealing algorithm to find solutions to the maximum log-likelihood and compare its solutions to those from ASReml. Simulated annealing is a numerical method for optimization that is particularly suitable for maximizing functions that do not vary smoothly in parameter space, so it may be a suitable method to find REML solutions with a spherical spatial model.

In the remainder of this paper we show how REML may be used to estimate parameters of general linear models from data obtained by systematic sampling of the soil.

Materials and methods

Data

Two sets of data are used here. The first was collected in the Vale of the White Horse near Oxford in central England. The second was collected in the Swiss Jura.

Data set 1, Vale of the White Horse. These data were taken from those published by Burrough (1969). They were collected in his special Study Area 1, covering 1.26 km^2 in the Vale of the White Horse, central England (Burrough *et al.*, 1971). Jarvis (1973) surveyed and described the soils of the region. The soil comprises surface- and ground-water gleys formed over drift and solid parent materials – Lower Greensand, Gault Clay (Cretaceous) and Kimmeridge Clay (Jurassic). The area was sampled on a regular square grid, at 100-m intervals, with six columns and 21 rows giving a total of 126 sample sites.

At each site the soil was allocated to a predefined class (series) based on profile characteristics. Table 1 shows the series names used by Burrough (1969), the changes to the names or identifications reported by Burrough *et al.* (1971), and the modern correlatives of the series and the soil subgroups to which they belong in the current classification of the soils of England and Wales (Clayden & Hollis, 1984). We have retained the original series names for this study.

Several soil properties were recorded in the field or determined later in the laboratory from a bulked sample. We use data from the topsoil (defined as 0–18 cm), on clay content (determined by hand texturing with a few laboratory determinations for calibration), available potassium (K_{av}) and cation exchange capacity (CEC). Linear modelling was used to answer the question: What are the mean values of the properties within each soil series, and do these differ with respect to each of these properties?

Since the data were collected on a systematic grid, these questions cannot be answered correctly by OLS estimation.

Data set 2, Swiss Jura. These data are heavy metal concentrations in the topsoil of a region of the Swiss Jura, measured by Atteia *et al.* (1994) and analysed by the authors (Atteia *et al.*, 1994; Webster *et al.*, 1994; Goovaerts *et al.*, 1997). Measurements of the concentration of heavy metals in the soil were made on small cores of soil to depth 25 cm at 214 sites on a square grid of interval 250 m. Of the 214 data, four were excluded by Atteia *et al.* (1994) because the values were suspect. The land use and the underlying rock type was identified at each grid site.

We extracted a subset of 100 data from these 210, which are observations in eight columns of the original grid (excluding one point that was the only observation in the subset belonging to one of the rock types). The data on nickel and cobalt concentration were chosen for analysis.

The questions to be answered by linear modelling with these data are as follows.

1 What are the mean concentrations of the heavy metals in soils over each rock type, and do the rock types differ with respect to each of these properties?

2 Can cobalt concentration be predicted from nickel concentration? We accept that in most surveys both these variables would be measured, but the problem does exemplify a more general one of how to predict one continuous variable from another.

As with the data from the Vale of the White Horse, the sampling was systematic and so the assumptions of OLS are not met.

Exploratory analysis

Table 2 lists the summary statistics for the soil variables, and Figure 1 displays the histograms. The data on CEC showed mild positive skew, which was removed by transformation to square roots. The data on available potassium were strongly positively skewed. Although they remain skewed after transformation to natural logarithms, this appears to be due principally to two outlying values. The log-transformed data were used for further analysis, and the outliers were retained. The concentrations of cobalt and nickel in the Jura soils show very weak negative skew and do not require transformation.

Figure 2 shows classified post-plots of these data. In Figure 3(a) are the soil series (after Burrough, 1969) identified at each sampling site in data set 1, and Figure 3(b) shows the rock type identified at each sampling site from the Swiss Jura.

Estimating parameters of the linear model

As a preliminary step OLS estimates were made of model parameters. The residuals from these models,

$$\mathbf{z} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}, \tag{17}$$

were then computed, and their variogram was estimated by Matheron's (1962) method of moments estimator:

Table 1 Soil series in Burrough's Study Area 1 according to the original (Burrough, 1969) and revised (Burrough *et al.*, 1971) classifications and identifications, and the current correlatives and soil subgroups (Clayden & Hollis, 1984)

Series name			
Burrough (1969)	Burrough *et al.* (1971)	Current correlative	Current soil subgroup
Denchworth	Denchworth	Denchworth	7.12, Pelo-stagnogley
Fernham	Shellingford	Burlesdon	5.72, Stagnogleyic argillic brown earth
Fladbury	Fladbury	Fladbury	8.13, Pelo-alluvial gley
Mead	Mead	Thames	8.14, Pelo-calcareous alluvial gley
Uffington	Kingston	Kingston	7.11, Typical stagnogley

Table 2 Summary statistics for raw and transformed soil variables

	Clay /%	K_{av} /mg kg^{-1}	K_{av} /log(mg kg^{-1})	CEC /cmol$_e$ kg^{-1}	CEC /$\sqrt{}$(cmol$_e$ kg^{-1})	Co /mg kg^{-1}	Ni /mg kg^{-1}
Mean	42.60	201.54	5.04	25.54	5.00	9.08	33.98
Median	42.00	145.00	4.98	24.00	4.90	9.56	34.52
Variance	52.95	72.12×10^3	0.38	57.96	0.56	13.33	92.06
Skew	0.009	6.74	1.17	0.73	0.17	−0.22	−0.13

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{z(\mathbf{x}_i) - z(\mathbf{x}_i + h)\}^2, \qquad (18)$$

where $z(\mathbf{x}_i)$ is the residual at location $\mathbf{x}_i$, $z(\mathbf{x}_i + h)$ is the residual at a location separated from $\mathbf{x}_i$ by the lag distance h and there are $2N(h)$ pairs of residuals separated by the lag h. Note that we defined the lag as a scalar, ignoring any directional dependence of the covariance, because there were too few data to do otherwise.

The point estimates of the variogram were then plotted and examined to see if there was evidence for non-stationarity and

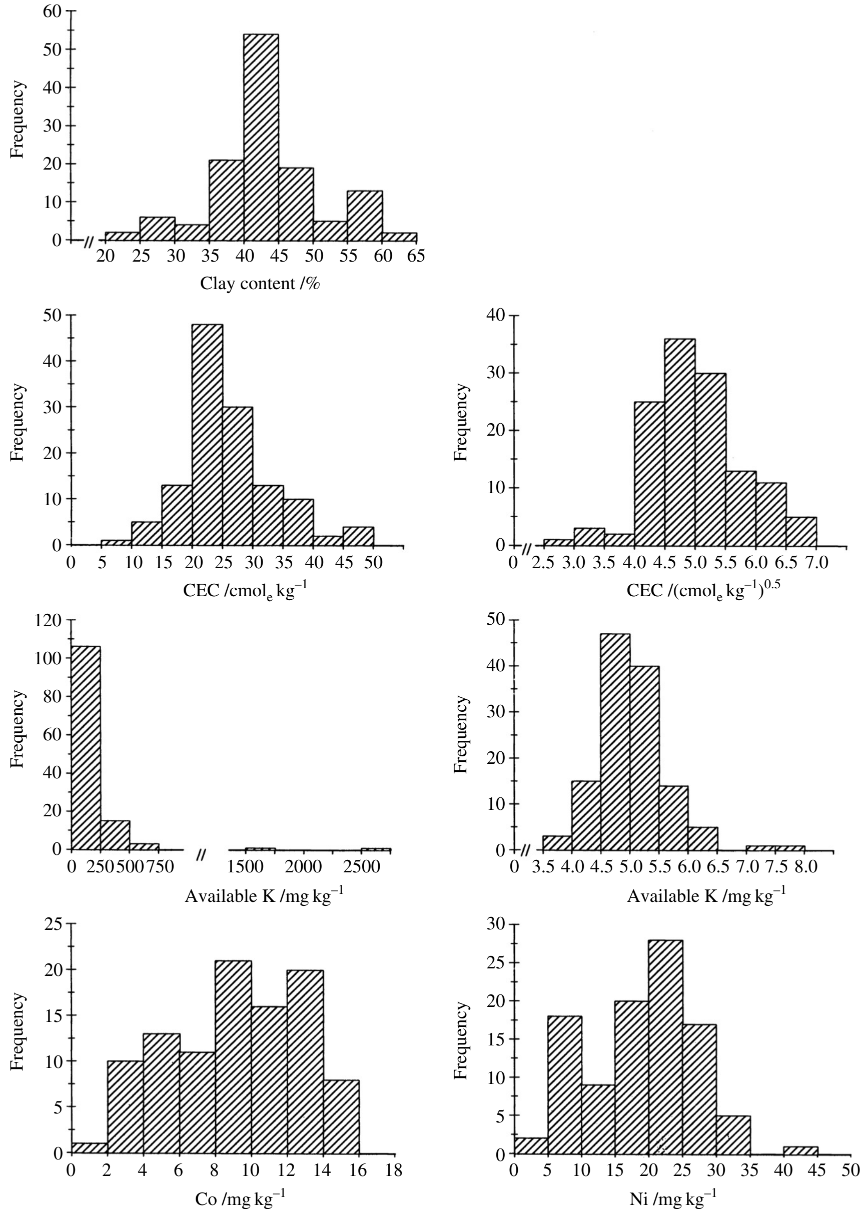


Figure 1 Histograms of raw and transformed variables.

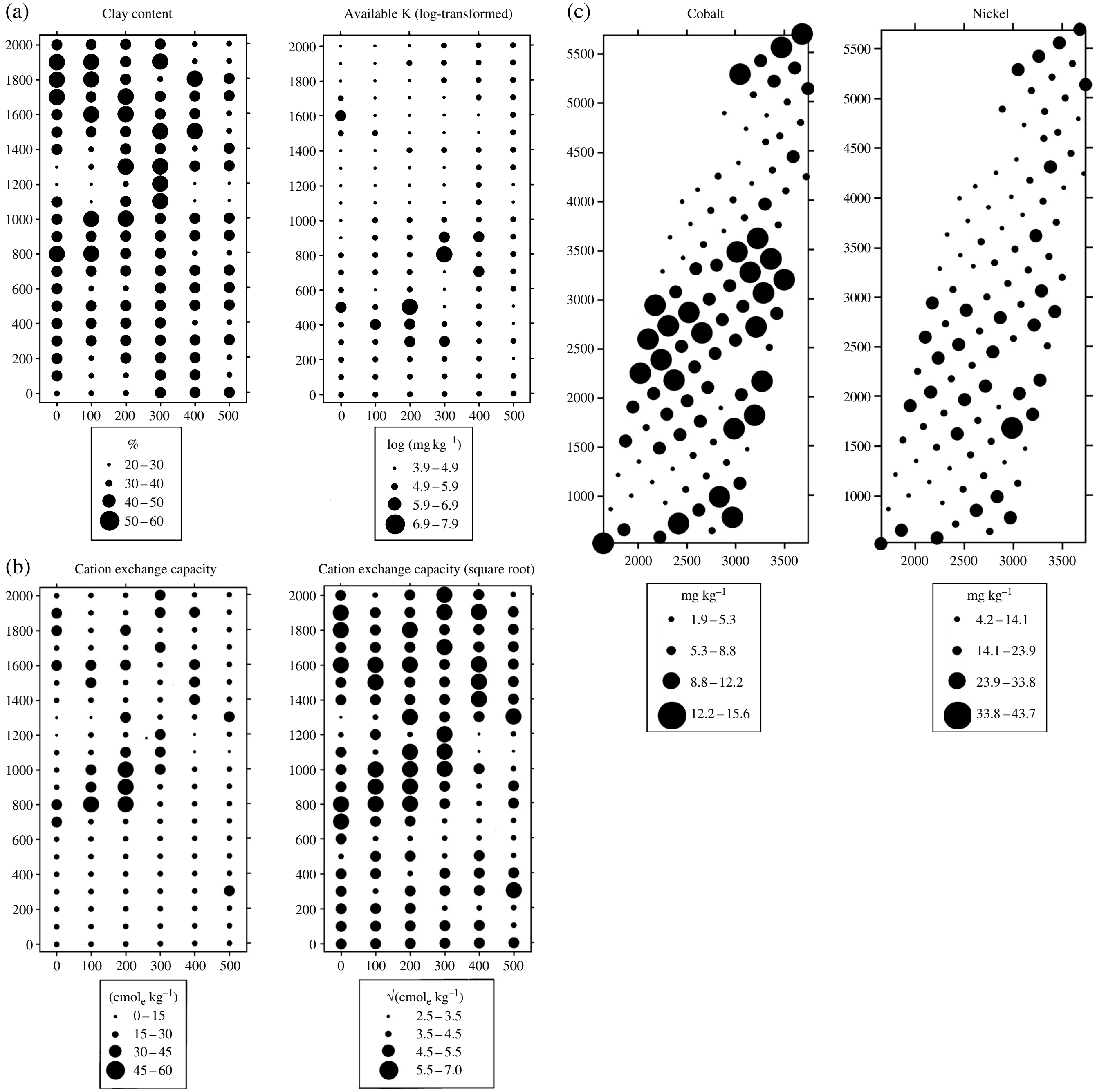


Figure 2 Classified post-plots of raw and transformed variables. Coordinates are in metres from a local origin.

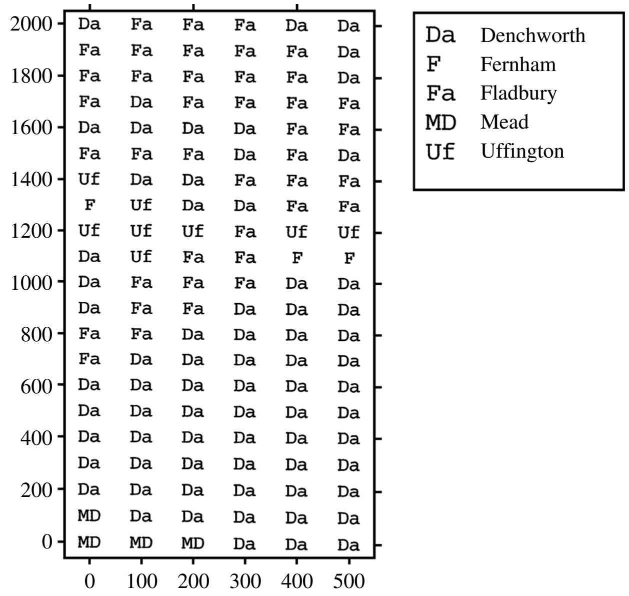
to obtain a guess of the *a priori* variance, spatial dependence and distance parameter of the error variable. These values were simply used to initialize the REML estimation since we know that the statistics of the OLS residuals will be biased.

The variance parameters of the specific general linear model were then estimated by numerical minimization of the negative log-residual likelihood function $-l(\boldsymbol{\theta}|\widehat{\boldsymbol{\beta}},\boldsymbol{\beta})$ with respect to $\boldsymbol{\theta}$ where $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\beta}},\boldsymbol{\beta})$ is defined in Equation (15). This was done with the ASReml software, and secondly by the method of simulated annealing (Kirkpatrick *et al.*, 1983), which has been used elsewhere to obtain REML estimates of spatial variance parameters (Pardo-Igúzquiza, 1997).

A detailed discussion of simulated annealing is beyond the scope of this paper; more information is given by Kirkpatrick *et al.* (1983), Aarts & Korst (1989) and Press *et al.* (1992), but an outline of the method is given below.

Simulated annealing proceeds by random perturbation of an initial set of parameters with respect to which some objective

(a) Soil series



(b) Rock types

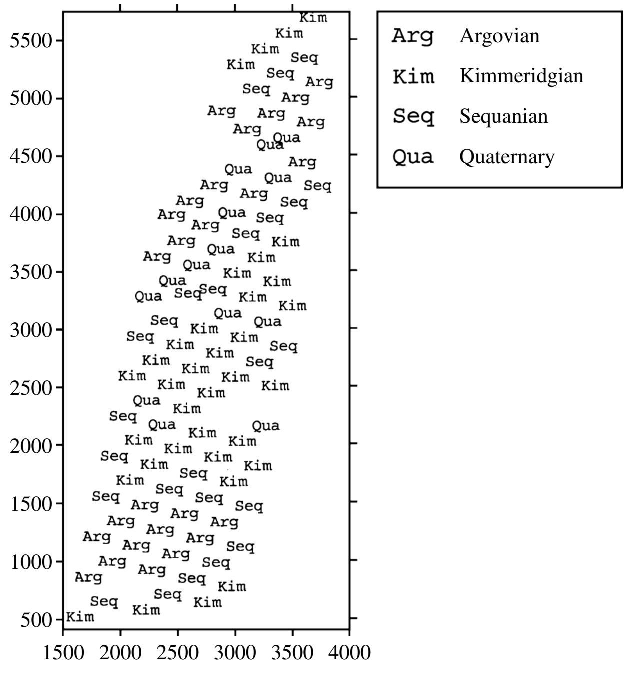


Figure 3 (a) Soil series identifications at each sample site in data set 1. (b) Rock type identified at each sample site in data set 2.

function is minimized. Perturbations that reduce this function are accepted, those that increase it are accepted or rejected randomly where the probability of acceptance, p_a, is determined by a function (the Metropolis criterion) that simulates the statistical mechanics of energy states in a molten metal. Thus, if the proposed change in the system results in a change in the objective function from f_i to f_j, where $f_j > f_i$, then the probability of acceptance of the change is

$$p_\mathrm{a} = \exp\left\{\frac{f_i - f_j}{\kappa}\right\}, \tag{19}$$

where κ is a parameter analogous to the temperature of the metal.

If the parameters are randomly perturbed many times at a fixed temperature, with changes accepted or rejected according to this criterion, then the system approaches a thermal equilibrium in which the distribution of values of the objective function is Boltzmann's distribution (Aarts & Korst, 1989). In simulated annealing the system is taken through many such sequences of perturbations, with the temperature parameter reduced at the end of each. To reduce the temperature reduces the probability of acceptance of a change in the system which results in a given increase in the objective function. The aim is to emulate the slow cooling of a molten metal that will cause it to 'anneal', i.e. to reach an energy state that is a global minimum – a regular crystalline solid. The particular advantage of simulated annealing as a method of optimization is that the Metropolis criterion allows the system in effect to jump over a barrier that could trap it at a solution that is only locally optimal. This is useful for the minimization of negative log-likelihoods with spatial covariance structures described by models such as the spherical function in Equation (20) below, since these may have local minima that can make other optimization methods non-robust (Warnes & Ripley, 1987; Ripley, 1988; Mardia & Watkins, 1989).

In this analysis a 'cooling schedule' was defined following Kirkpatrick *et al.* (1983). The initial temperature of the system, κ_1, was chosen so that the proportion of proposed changes accepted before the first reduction in temperature was in the range 0.90–0.99 and the new temperature of the system κ_{m+1} after the mth cooling step is $\alpha_c \kappa_m$ where $\alpha_c = 0.95$. The cooling step took place after a fixed number of perturbations of each parameter of the objective function. The algorithm kept track of the objective function at the most recent 20 cooling steps, and once the objective function had remained unchanged over this interval the algorithm stopped.

The initial values of the parameters in $\boldsymbol{\theta}$ were given to start the algorithm. These were the *a priori* variance and the spatial dependence and distance parameter for an exponential covariance function as described in Equation (7) above. The initial values were obtained by visual inspection of the variogram of the OLS residuals, but runs with other starting values were also tried to reduce the risk that the eventual solution was just a local optimum. Having obtained a solution for the exponential covariance function, the same procedure was followed with a spherical model:

$$\begin{aligned} \mathbf{V}_{i,j} &= \sigma^2 s\left\{1 - \mathrm{sph}\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right)\right\}, \quad i \neq j \\ &\quad \sigma^2, \qquad\qquad\qquad\qquad\qquad i = j, \end{aligned} \tag{20}$$

where

$$\mathrm{sph}\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right) \equiv \left\{\begin{matrix} \frac{3}{2}\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right) - \frac{1}{2}\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right)^3, & |\mathbf{x}_i - \mathbf{x}_j| \leq a \\ 1 & , \ |\mathbf{x}_i - \mathbf{x}_j| > a \end{matrix}\right\}. \tag{21}$$

The REML estimates of the parameters $\boldsymbol{\theta}$ were then selected for that covariance model (spherical or exponential) for which the negative log-residual likelihood was smaller. The variogram was then plotted on the same graph as the point estimates from the OLS residuals. Since the latter are biased the two graphs are not necessarily similar, but this should reveal any serious problems with the estimation.

We then obtained the REML estimates of the parameters $\boldsymbol{\beta}$ and their covariance matrix by substituting the REML estimate of $\mathbf{V}$ into Equations (10) and (11).

Inference

Having estimated the parameters of the linear model we may wish to test for significance. The tests, in these case studies, will be based on null hypotheses that are of the form either 'all soil series means are equal' or 'the regression coefficient is equal to zero'. In the latter case the hypothesis can be tested from the t ratio:

$$t = \frac{\widehat{\beta}}{\widehat{\sigma}_\beta}, \tag{22}$$

where $\widehat{\beta}$ is the estimate of the regression coefficient and $\widehat{\sigma}_\beta$ is its standard error extracted from the covariance matrix of the model parameters. This may be tested against Student's t statistic with $n-p$ degrees of freedom.

In the case of the ANOVA-type model we may test null hypotheses that particular contrasts or sets of contrasts among the elements of $\boldsymbol{\beta}$ are zero. To do this we form a matrix of contrasts, $\mathbf{L}$, in which each row corresponds to a single contrast. So, for example, to test the null hypothesis that the means of the first two of five classes are equal (a single contrast), we set $\mathbf{L} \equiv [1, -1, 0, 0, 0]$. Under this null hypothesis $\mathbf{L}\boldsymbol{\beta} = 0$. To test this we compute the Wald statistic:

$$\left(\mathbf{L}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \left(\mathbf{L}\widehat{\mathbf{C}}\mathbf{L}^{\mathrm{T}}\right)^{-1} \left(\mathbf{L}\widehat{\boldsymbol{\beta}}\right). \tag{23}$$

The central term is the inverse of the covariance matrix of the contrast or contrasts in $\mathbf{L}$. Under the null hypothesis this statistic is distributed as χ^2 with degrees of freedom equal to the rank of $\mathbf{L}$. To test the null hypothesis that all elements of $\boldsymbol{\beta}$ are equal (i.e. the class means in our example) $\mathbf{L}$ is a $p-1 \times p$ matrix, the rows of which contain a set of linearly independent contrasts so that the rank of $\mathbf{L}$ is $p-1$. One such matrix ($p=5$) is

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

A comment is necessary on the plausibility of the null hypotheses invoked in these tests of significance. Clay content of the soil is closely linked to the definition of the soil series, and so the null hypothesis that the means of the soil series are equal is clearly wrong from the point of view of soil science (Webster, 2001). Of course the linear modelling is still necessary to estimate the mean value of the property within each series and to attach confidence limits to it. In the case of the other properties the question of whether the soil series differ with respect to the property may be open, and the null hypothesis is of scientific interest. For example, the available potassium will depend in part on factors such as parent material and clay mineralogy (closely linked to the series definition) but also on farming activities and biological processes. If the latter factors dominate then the series may be indistinguishable with respect to this variable.

Results

Figure 4 shows point estimates of the variograms of the OLS residuals for each of the models, and the (continuous) variograms estimated by REML. In general the point estimates and the continuous models are not very similar. This is not unexpected since the point estimates are determined from OLS residuals and so will be more or less biased. Table 3(a–c) presents results for the REML estimates of all model parameters, compared with those obtained by OLS. The Wald statistics were computed for the OLS analysis – as $(p-1) \times$ the variance ratio – for comparison with those obtained by REML in the case of the ANOVA-type analyses. One could compare the regression models obtained by OLS and REML by comparing the standard errors of the regression coefficients, and using the t ratio to test the null hypothesis that the coefficient is zero.

Note that, in each case, the OLS analysis provided stronger evidence against the null hypothesis than did REML. This will reflect the assumption of the OLS estimation that all n data are independent sources of information; the REML estimate recognizes that, because of spatial dependence, we have rather less information on which to base our inference. The effect of modelling the spatial structure of the error is two-fold. First the estimates of the error variance obtained by REML are larger than the OLS estimates in all cases. Second, the data are not all weighted equally in the computation of class means, i.e. the elements of $\widehat{\boldsymbol{\beta}}$ in Equation (10). For example, a datum in an isolated occurrence of one soil class in the sample will be given more weight in the computation of the class mean than would a datum surrounded by many sample sites within the same class.

The difference between the inferences from OLS and REML estimation depends on the strength of the spatial dependence revealed by the REML analysis. In the case of the clay content, compared between soil series in data set 1, the modelled spatial covariance structure has a short range compared with the grid interval, so the error at a location will be only weakly correlated with the errors at the nearest neighbouring sites. As a result the REML estimates of the class means and those obtained by OLS are similar, and the differences between the Wald statistics are smaller than for the other variables. Similarly in data set 2 the autocorrelation of error in the ANOVA-type model for the

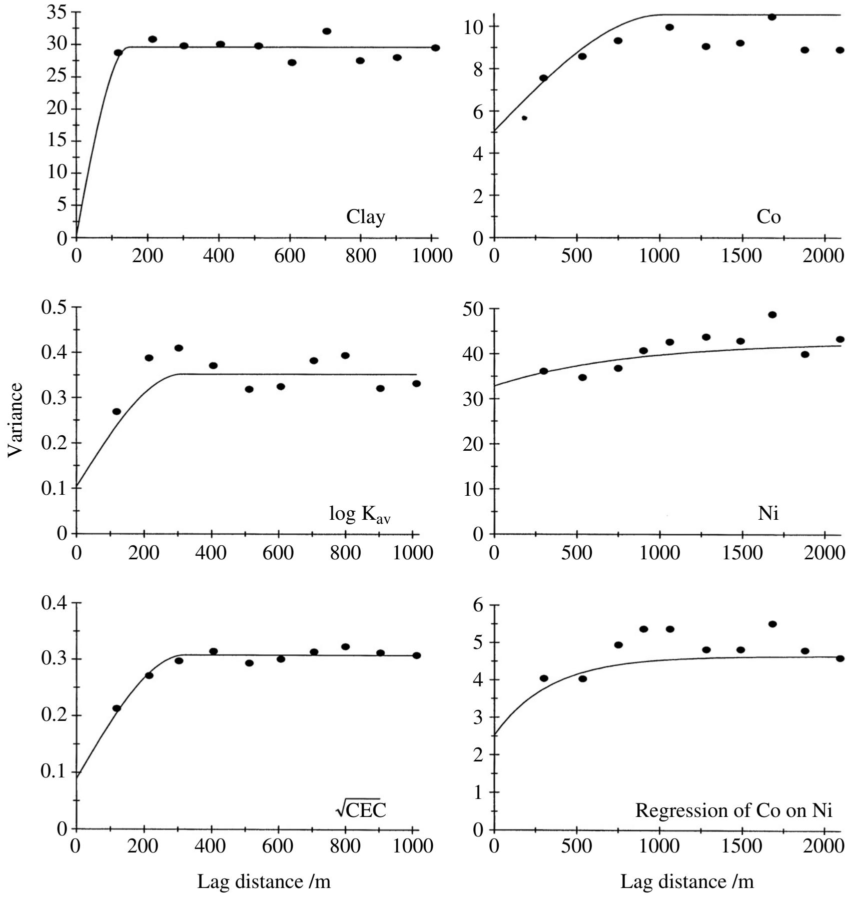


Figure 4 Point estimates of the variogram of OLS residuals and continuous variogram models obtained by REML for errors from general linear models.

Table 3a Results for linear models fitted by OLS and REML. ANOVA-type models for soil data from the Vale of the White Horse

Variable	Clay			K_{av}[a]			CEC[a]		
Estimator	OLS	REML[b]		OLS	REML		OLS	REML	
		SA	ASR		SA	ASR		SA	ASR
β parameters (Class means)									
Class									
Denchworth	42.45	42.62	42.62	5.21	5.10	5.10	4.80	4.89	4.89
Fernham	24.33	23.29	23.30	4.78	4.90	4.90	3.31	3.39	3.39
Fladbury	47.05	46.60	46.60	4.84	4.97	4.97	5.65	5.49	5.49
Mead	36.00	36.63	36.64	5.06	4.99	4.99	5.15	5.05	5.05
Uffington	31.00	31.54	31.51	4.58	4.87	4.87	4.09	3.87	3.87
θ parameters[c]									
Variance	29.54	29.63	29.63	0.34	0.35	0.35	0.29	0.31	0.31
Spatial dependence	–	1.00	0.95	–	0.70	0.70	–	0.71	0.71
Distance parameter /m	–	149.2	151.5	–	314.4	314.4	–	322.2	322.3
Negative log-likelihood	271.59	268.80	268.81	2.69	−13.98	−13.98	−8.38	−23.07	−23.07
Wald statistic	83.0	81.7	81.6	12.9	1.75	1.76	120.24	78.7	78.6
P-value	<0.001	<0.001	<0.001	0.012	0.782	0.780	<0.001	<0.001	<0.001

[a] The results are for potassium concentrations transformed to their natural logarithms and for CEC transformed to its square root.

[b] SA denotes REML estimates obtained by simulated annealing and ASR the ASReml results.

[c] The spherical model was selected in all cases – Equation (20).

Table 3b Results for linear models fitted by OLS and REML. ANOVA-type models for soil data from the Swiss Jura

Variable	Cobalt			Nickel			
Estimator	OLS	REML		OLS	REML		
		SA	ASR		SA	ASR[a]	ASR[b]
$\boldsymbol{\beta}$ parameters (Class means)							
Rock type							
Argovian	5.86	7.08	7.06	11.82	13.02	11.87	13.02
Kimmeridgian	11.03	10.23	10.24	24.31	23.98	24.30	23.98
Sequanian	10.20	10.25	10.25	20.70	20.45	20.28	20.45
Quaternary	8.17	8.85	8.84	17.17	18.61	17.34	18.60
$\boldsymbol{\theta}$ parameters[c]							
Variance	9.12	10.56	10.50	40.26	42.35	40.07	42.34
Spatial dependence	–	0.52	0.52	–	0.21	0.08	0.21
Distance parameter /m	–	1030	1028	–	1932	845	1942
Negative log-likelihood	160.4	156.7	156.7	231.7	229.1	231.3	229.1
Wald statistic	48.8	13.4	13.6	59.9	34.92	58.11	34.92
P-value	<0.001	0.004	0.004	<0.001	<0.001	<0.001	<0.001

[a] This solution was found from arbitrary initial values for the distance parameter (1500 m) and spatial dependence (0.5).

[b] Here the initial values were at the solution obtained by simulated annealing.

[c] The spherical model was selected in all cases – Equation (20).

Table 3c Results for linear models fitted by OLS and REML. Regression-type model for predicting cobalt from nickel

Estimator	OLS	REML	
		SA	ASR
$\boldsymbol{\beta}$ parameters			
Constant	1.93	2.45	2.24
Regression coefficient	0.375	0.348	0.358
$\boldsymbol{\theta}$ parameters			
Variance	4.46	4.64	4.45
Model type	–	Exponential	Spherical
Spatial dependence	–	0.46	0.55
Distance parameter /m	–	342	453
Negative log-likelihood	128.9	125.5	126.9
Standard error of coefficient	0.027	0.029	0.028
t ratio	14.07	11.92	12.79
P-value	<0.001	<0.001	<0.001

concentration of nickel in the soil is weaker than in the case of cobalt – the distance parameter of the variogram is relatively large but the spatial dependence ratio is small.

By contrast the error models for available potassium, CEC (data set 1) and cobalt (data set 2) show more than half the error variance spatially structured at medium to long distances. While the REML analysis still indicates that CEC differs significantly between soil series and cobalt between rock types, the evidence is weaker than in the OLS analysis.

In the case of the ANOVA-type model for available potassium, the OLS analysis indicates a significant difference between the soil series. However, the REML analysis provides insufficient evidence against the null hypothesis. The series means as estimated by REML are more similar than the OLS estimates. This and the (slightly) smaller estimate of the error variance results in a smaller Wald statistic.

The conclusion that the series do not differ with respect to (transformed) available potassium is plausible, since farm practice will have a larger effect on available potassium in the soil than will the pedogenetic processes on which the series are defined. Studies in the same region by random sampling to compare concentrations of available potassium in the soil of different physiographic units concluded that there was no difference (Webster & Beckett, 1968).

The conclusions that nickel and cobalt differ significantly between the rock types agrees with the conclusion of Webster *et al.* (1994), from spatial analysis, that the long-range variation of these two variables is determined by geology. It is therefore not surprising that there was a significant regression of cobalt concentration on that of nickel. This significant relationship might reflect various underlying factors. Its usefulness is independent of the scientific explanation since the proper use for such a model is for prediction, i.e. for predicting the cobalt concentration in soil samples from this same region where only the concentration of nickel has been determined by chemical analysis. It should not be mistaken for a functional relation between the variables, which should be estimated in other ways (Webster, 1997). Figure 5 displays the relation

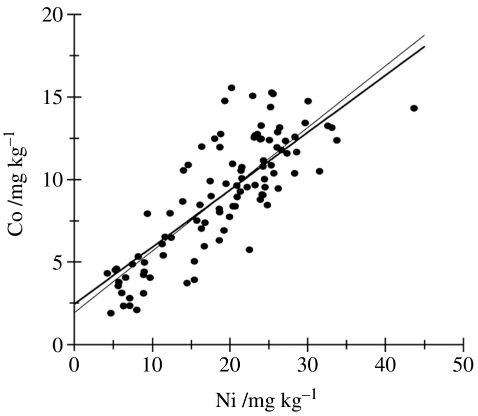


Figure 5 Scatter plot of cobalt against nickel concentration in the soil with regression lines of Co on Ni fitted by REML (heavy line) and OLS (fine line).

between the two variables, with the OLS and REML regression lines of Co on Ni.

In almost all cases the ASReml and simulated annealing REML estimates of the model parameters are very close or identical. An exception was in the analysis of the data on nickel compared between rock types. Here the ASReml program converged to a solution with a negative log-likelihood as small as the simulated annealing solution only when it was initiated from that solution. Initiating it from elsewhere resulted in solutions with a larger negative log-likelihood; one such is shown in Table 3b. These poorer solutions had small spatial dependence ratios, and so other parameter estimates were quite close to the OLS solutions. This problem is most likely due to the susceptibility of gradient-based minimization to the non-smooth behaviour of log-likelihood functions for models with a spherical spatial correlation structure.

Discussion and conclusions

Soil scientists using data from systematic samples should resist the temptation to analyse them with off-the-shelf OLS procedures such as ANOVA or regression. While OLS gives unbiased estimates of the parameters of a linear model, these are less efficiently estimated than by model-based methods, and the variance of the parameters is not correctly known. In these examples, using REML to estimate the error variance with an appropriate spatial model gave larger estimates of the variance and weaker evidence against the null hypothesis. In one case the REML analysis accepted a null hypothesis for an analysis where the OLS analysis provided evidence for rejection that would conventionally be regarded as significant. The advantage of REML over both OLS and ML estimation for systematically sampled data is that it provides unbiased estimates of the variance parameters, and so a sound basis for attaching confidence limits to estimates of model parameters. Since REML is now readily available its use for the analysis of systematically sampled data should become standard.

This study highlighted the possible problems of gradient methods for finding the REML estimates of model parameters when a spherical spatial covariance model is used. We have shown that simulated annealing might be preferable when this model is used, but it is much less efficient computationally. When using the spherical model and a gradient method you should compare the results of REML solutions from different initial values of the model parameters. If the results are very variable this indicates that local optima are causing problems.

While REML may be used to model the spatial covariance in a general linear model, this does not necessarily solve all the problems that a systematic sample might pose. If there is systematic variation in the soil properties of interest then this can be aliased with the systematic sampling, resulting in serious bias in the parameters, however they are estimated. Systematic variation in agricultural fields, for example, may arise from historical ‘ridge and furrow’ patterns, and other soils may exhibit systematic variation as in the gilgai patterns of eastern Australia and the polygonal or longitudinal patterns formed by ice wedges in periglacial conditions. Some of these systematic effects are obvious, but others might be fully revealed only after sophisticated analysis (e.g. McBratney & Webster, 1981). When the scientist is aware of systematic variation in the variables being studied systematic sampling is not necessarily precluded, but care must be taken to avoid bias. In particular, if the spatial frequency of the underlying pattern is ϕ then the spatial frequency of sampling in the direction of any periodic variation must be larger than the Nyquist frequency, 2ϕ.

Acknowledgements

We thank Professor P.A. Burrough for permission to use his data from the Vale of the White Horse. The soil data from the Swiss Jura were originally collected by Atteia *et al.* (1994) of the Swiss Federal Institute of Technology, Lausanne, and have been made available through the Internet by Dr P. Goovaerts.

This work was funded by the Biotechnology and Biological Sciences Research Council through its Competitive Strategic Grant to Silsoe Research Institute.

References

Aarts, E. & Korst, J. 1989. *Simulated Annealing and Boltzmann Machines – A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, New York.

Atteia, O., Dubois, J.-P. & Webster, R. 1994. Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, **86,** 315–327.

Brink, A.B.A., Partridge, T.C. & Williams, A.A.B. 1982. *Soil Survey for Engineering*. Oxford University Press, Oxford.

Burrough, P.A. 1969. *Studies in soil survey methodology*. DPhil thesis, University of Oxford.

Burrough, P.A., Beckett, P.H.T. & Jarvis, M.G. 1971. The relation between cost and utility in soil survey. *Journal of Soil Science*, **22,** 359–394.

Clayden, B. & Hollis, J.M. 1984. *Criteria for Differentiating Soil Series*. Technical Monograph 17, Soil Survey of England and Wales, Lawes Agricultural Trust, Harpenden.

Cochran, W.G. 1977. *Sampling Techniques*, 3rd edn. John Wiley & Sons, New York.

Cook, D.G. & Pocock, S.J. 1983. Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*, **39,** 361–371.

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

De Gruijter, J.J. & ter Braak, C.J.F. 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, **22,** 407–415.

Fisher, R.A. 1921. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, London, Series A*, **222,** 309–368.

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., Cullis, B.R. & Verbyla, A.P. 1997. Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics*, **2,** 269–293.

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.

Goovaerts, P., Webster, R. & Dubois, J.-P. 1997. Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics*, **4,** 31–48.

Jarvis, M.G. 1973. *Soils of the Wantage and Abingdon District*. Sheet 253, Soil Survey of England and Wales, Lawes Agricultural Trust, Harpenden.

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. 1983. Optimization by simulated annealing. *Science*, **220,** 671–680.

Koch, K.-R. 1988. *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer-Verlag, New York.

Lark, R.M. 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, **51,** 717–728.

Mardia, K.V. & Marshall, R.J. 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71,** 135–146.

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. Mémoires du Bureau de Recherches Géologiques et Minières, Paris.

McBratney, A.B. & Webster, R. 1981. Detection of ridge and furrow pattern by spectral analysis of crop yield. *International Statistical Review*, **49,** 45–52.

Papritz, A. & Webster, R. 1995. Estimating temporal change in soil monitoring: I. Statistical theory. *European Journal of Soil Science*, **46,** 1–12.

Pardo-Igúzquiza, E. 1997. MLREML: A computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers and Geosciences*, **23,** 153–162.

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

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 1992. *Numerical Recipes (Fortran)*, 2nd edn. Cambridge University Press, Cambridge.

Ripley, B.D. 1988. *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

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.

Stuart, A., Ord, J.K. & Arnold, S. 1999. *Kendall's Advanced Theory of Statistics,* Volume 2A: *Classical Inference and the Linear Model*, 6th edn. Arnold, London.

Warnes, J.J. & Ripley, B.D. 1987. Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74,** 640–642.

Webster, R. 1997. Regression and functional relations. *European Journal of Soil Science*, **48,** 557–566.

Webster, R. 2000. Is soil variation random? *Geoderma*, **97,** 149–163.

Webster, R. 2001. Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52,** 331–340.

Webster, R. & Beckett, P.H.T. 1968. Quality and usefulness of soil maps. *Nature, London*, **219,** 680–682.

Webster, R. & Payne, R.W. 2002. Analysing repeated measurements in soil monitoring and experimentation. *European Journal of Soil Science*, **53,** 1–13.

Webster, R., Atteia, O. & Dubois, J.-P. 1994. Coregionalization of trace metals in the soil in the Swiss Jura. *European Journal of Soil Science*, **45,** 205–218.

Appendix

Least squares, generalized least squares and maximum likelihood estimation of $\boldsymbol{\beta}$ in the general linear model

In the general linear model the $n \times 1$ vector of observed values $\mathbf{y}$ is fitted by the vector $\mathbf{X}\boldsymbol{\beta}$. The vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is therefore a vector of errors. It represents a point in an n-dimensional space. If the coefficients in $\boldsymbol{\beta}$ are unbiased then the expected location of this point is at the origin.

The goodness of fit of a general linear model may be measured by the length of the error vector, its 'norm' in the usual terms of vector algebra. The simplest norm is the Euclidean distance, and the squared Euclidean norm of the error vector is equal to the sum of squared errors over the n observations used to fit the model. This may be written as

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{A1}$$

The ordinary least squares criterion for the fit of $\boldsymbol{\beta}$, due to Gauss, is to minimize the squared Euclidean norm of the error vector. Following the normal rules of matrix algebra Equation (A1) may be expanded to

$$S = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}. \tag{A2}$$

The minimization of S with respect to $\boldsymbol{\beta}$ is achieved by finding $\boldsymbol{\beta}$ so that the partial derivative of S with respect to $\boldsymbol{\beta}$ is zero. Following the normal rules for evaluating derivatives with respect to matrices (see, for example, Koch, 1988), we may write

$$\frac{\partial}{\partial \boldsymbol{\beta}} S = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}. \tag{A3}$$

Substituting the OLS estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in Equation (A3) sets the derivative to zero by definition, and so

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{A4}$$

If the errors are identically and independently distributed (iid) then we may think of the error vector as drawn from a population with a hyperspherical distribution around the origin (i.e. a cloud of vectors with the same distribution in all directions from the origin). This assumption is implicit in the use of the OLS criterion since this is based on the magnitude of the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and not on its direction.

If the error terms are identically distributed with variance σ^2 but with a correlation matrix $\mathbf{A}$ with non-zero terms off the diagonal (i.e. the errors are not independent) then the error vector is drawn from a population with a hyperellipsoidal shape, i.e. the cloud is elongated along a principal axis. Now if we wish to use the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ as a measure of the goodness of fit it is clear that we need to consider its direction as well as its length. An error vector of Euclidean length d along the principal axis of the distribution represents a better fit than an error vector that has the same length but which is perpendicular to the principal axis.

The solution to this problem is to use the generalized distance, described in standard texts on multivariate analysis. This is an alternative to the Euclidean norm. In effect the generalized norm defines an n-dimensional space within which error vectors with variance–covariance matrix $\mathbf{V}$ have a hyperspherical distribution. The squared generalized length of the error vector is

$$S_{\mathrm{g}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{A5}$$

where $\mathbf{V} = \sigma^2\mathbf{A}$. As with the ordinary least squares we may write the partial derivative of the squared generalized distance norm of the error vector with respect to the parameters in $\boldsymbol{\beta}$. This is

$$\frac{\partial}{\partial \boldsymbol{\beta}} S_{\mathrm{g}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}, \tag{A6}$$

and so the generalized least squares (GLS) estimate is obtained by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}. \tag{A7}$$

The OLS and GLS criteria, while intuitively appealing, are none the less somewhat arbitrary. A better theoretical foundation for estimating the parameters of the general linear model was provided by R.A. Fisher, with the likelihood concept (Fisher, 1921). Consider $\mathbf{y}$ to be a random variate drawn from some stochastic process with a probability density function $G\ (\mathbf{y}|\boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ is a vector of the parameters of the distribution. As a probability density function G is a function of the variate $\mathbf{y}$ conditional on its parameters $\boldsymbol{\zeta}$. It has various properties, for example

$$\int_{\mathbf{y}\in\mathbf{Y}} G\ (\mathbf{y}|\boldsymbol{\zeta})\mathrm{d}\mathbf{y} = 1, \tag{A8}$$

where $\mathbf{Y}$ is the space of all possible variates $\mathbf{y}$. The problem is to obtain an estimate of the parameters $\boldsymbol{\zeta}$ given a set of observations in $\mathbf{y}$. Fisher's proposal was to consider G as a function of the parameters $\boldsymbol{\zeta}$ conditional on the observations, and to estimate $\boldsymbol{\zeta}$ by $\widehat{\boldsymbol{\zeta}}$ that maximizes this function. He called it the likelihood. While it is obtained from a probability density function, the likelihood is not a probability since

$$\int_{\zeta\in\mathbf{Z}} G\ (\boldsymbol{\zeta}|\mathbf{y})\mathrm{d}\mathbf{y} \neq 1, \tag{A9}$$

where $\mathbf{Z}$ is the space of all possible parameter vectors.

If we assume that $\mathbf{y}$ arises from an underlying general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \tag{A10}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector, and $\boldsymbol{\eta}$ is drawn from a multivariate normal process, then the probability density function for $\mathbf{y}$ is given by the multivariate normal probability density function:

$$\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{V}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{A11}$$

This may be thought of as a likelihood function conditional on $\mathbf{y}$, $L(\mathbf{V}, \boldsymbol{\beta}|\mathbf{y})$. This likelihood function has the same general form as Equation (A9). It is therefore a function of the parameters to be estimated (in this case $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$), and is conditional on the observations, $\mathbf{y}$. For convenience we usually work with the log-likelihood:

$$l(\mathbf{V}, \boldsymbol{\beta}|\mathbf{y}) = -\frac{p}{2}\log 2\pi - \log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad \text{(A12)}$$

This log-likelihood function must be maximized with respect to $\boldsymbol{\beta}$ in order to find the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\beta}}$. Inspecting Equation (A12) we see immediately that its partial derivative with respect to $\boldsymbol{\beta}$ will yield the same estimate as by GLS in Equation (A7). Therefore the ML estimator, for a model with a multivariate normal error vector, is equivalent to the GLS estimator, and in turn this is equivalent to the OLS estimator when the errors are iid. In fact there are grounds for regarding an ML estimator as optimal even when the errors cannot be assumed to be from a multivariate normal process, that have been summarized elsewhere (e.g. Lark, 2000).