# Recent Achievements in Geostatistical Analysis of Soil

R. WEBSTER

AFRC Institute of Arable Crops Research, Rothamsted Experimental Station,
Harpenden, Hertfordshire /ENGLAND/

## Historial development

There are many situations in which soil scientists wish to predict the conditions of the soil in a spatial sense. They can sample only a very small fraction of the soil mantle, yet in principle they may wish to know what the soil is like everywhere. And they may wish to estimate the average values of soil properties in blocks of land very much bigger than any individual sample. Since the soil is not the same everywhere these aims can be achieved only by some form of local estimation. Traditionally this has been attempted by first classifying the soil and then predicting separately for each class from data for that class. The approach has undoubtedly had its successes, but those successes have depended on surveyors' flair and their eye for country. If the relation between physiography and soil is weak or obscured by the vegetation or land management then the soil pattern is likely to be revealed only by intensive routine and tedious sampling. It may also happen that the visible features on which the soil is classified do not relate to the properties that one wishes to predict, and so one must resort again to intensive sampling in order to interpolate with confidence. This is increasingly the case as the scale becomes larger.

Viewed statistically the traditional procedure whereby the soil mantle of a region is subdivided by boundaries into distinct classes is fairly simple. For each class there is a mean value for that class with more or less variation about it. Formally we can write this as a model

$$z_{ij} = \mu + a_j + \varepsilon_{ij} \qquad /1/$$

in which $z_{ij}$ is the value of the soil property Z at any place i in class j, $\mu$ is the general mean of the property, $a_j$ is the difference between $\mu$ and the mean of the jth class, $\mu_j$ and $\varepsilon_{ij}$ is a random term with variance $\sigma_j^2$, which may be assumed to have some particular distribution, If we have $n_j$ sample data for class j the average of those data will estimate the mean value of the class $\mu_j$, with a variance $\sigma_j^2/n_j$, the square root of which is the familiar standard error. We can in principle improve that estimate to any extent we like simply by increasing $n_j$, the size of the sample. If we wish to predict the value at some unsampled point our best estimate will also be $\mu_j$, though now with a variance $\sigma_j^2 + \sigma_j^2/n_j$. The error is now determined very largely by the within-class variance, $\sigma_j^2$. However much we increase the sample we cannot diminish the estimation variance to less than $\sigma_j^2$. And so the quality of the classification sets a ceiling on precision of prediction, and this is the statistical reason for surveyors' devoting so much attention to the quality of their soil classifications.

This model defined by equation /1/ formed the basis of our attempts at quantitative prediction of soil properties in the 1960s /WEBSTER and BECKETT, 1968, 1970; BECKETT and WEBSTER, 1971a/ and of our interpretation of other people's assessment of soil classifications /BECKETT and WEBSTER, 1971b/. The latter showed that soil classifications varied from the moderately successful to the useless. Nevertheless, there seemed always to be large residual within-class variances, and from our experience we know that this variance was not entirely haphazard. The soil did not vary in a wholly random way nor did it change abruptly at boundaries as implied by the model. Instead there seemed to be a structure within classes and more or less gradual change across many boundaries. Classification did not, indeed could not, recognize these, and predictions made no use of this specifically local knowledge. We were not making the most of information that was there for the asking. A further disadvantage of classifications is that they tend to be made once and for all. The approach is very inflexible. Only two kinds of estimate are possible: one for the whole class and the other for individual points; and the estimates within a class are the same in both cases.

It was against this background that I sought a new approach to the problem, and I began by adapting the methods of time series analysis /WEBSTER, 1973; WEBSTER and CUANALO, 1975/. In one dimension, at least, time and space are analogous. KOZLOVSKII and SOROKHINA /1976/ did similarly. However, the theory of spatial statistics and its application to estimation were already well advanced in the mining industry, mainly as the result of work by MATHERON /1965, 1971/ and his colleagues in France, and it was at this point that I realised that many of our problems in soil survey were soluble in principle. It is this theory, the Theory of Regionalized Variables, that has enabled us to improve the basis of spatial prediction in soil science and to achieve the advances that we have made in geostatistics recently. And so I devote the remainder of this paper to the theory and its application.

## The Statistical nature of soil variation

As above, to make progress we had to take a new view of the distribution of soil over the land. To accord with our intuition this must embrace both continuity and randomness. And so we replace the model of equation /1/ with another, changing the notation somewhat. A soil property, Z, is assumed to be distributed continuously in space and to take values $z(x_i)$ at places $x_i$, $i = 1, 2, \ldots, \infty$, where x denotes the spatial coordinates in one, two or three dimensions according to context. The variation in Z may have two components, one deterministic and the other stochastic, and we can represent these by the following model:

$$z(x) = \sum_{k=0}^{N} a_k f_k(x) + \varepsilon(x) \qquad /2/$$

The first term on the right-hand side of the equation represents the deterministic component in which $f_k(x)$ are known functions of x and the $a_k$ are unknown coefficients. The quantity $\varepsilon(x)$ is a random term that is defined below. The deterministic component may be global; i.e. it may be a general trend over the whole region being studied, or it may be local, in which case it is often referred to as drift.

In practice it is usually found that the first term can be ignored because the stochastic component dominates at the scale of the investigation, and in only very few studies of soil /e.g. WEBSTER and BURGESS, 1980/ have investigators identified any significant drift. As a result a regionalized soil property can be regarded as a realization of a random process, and equation /2/ simplifies to

$$z(x) = \mu_V + \varepsilon(x) \qquad /3/$$

where $\mu_V$ is the mean value of Z, $E[z(x)]$. The random term has the following properties. Its expectation is zero:

$$E[\varepsilon(x)] = 0 \qquad /4/$$

and its variance is such that for any two places x and x + h separated by the lag vector h, which has both distance and direction,

$$\mathrm{var}[\varepsilon(x) - \varepsilon(x+h)] = E[\{\varepsilon(x) - \varepsilon(x+h)\}^2] = 2\gamma(h) \qquad /5/$$

In other words, the variance of Z is structured in a way that depends on the separation of any two sites and not on their absolute positions. With the mean constant equations /4/ and /5/ are equivalent to

$$E[z(x) - z(x+h)] = 0 \qquad /6/$$

and

$$\mathrm{var}[z(x) - z(x+h)] = E[\{z(x) - z(x+h)\}^2] = 2\gamma(h). \qquad /7/$$

The assumption of equations /6/ and /7/ constitute MATHERON's Intrinsic Hypothesis, which forms the basis of much practical geostatistics. The quantity $\gamma$ is known as the semi-variance: it is half the variance of the difference between the values at the two sites. As equation /7/ shows, its value depends on h, and the function that relates the two is the semi-variogram, increasingly known as just the variogram.

The semi-variance is related to the spatial covariance and autocorrelation. At lag h the spatial covariance, C(h), is defined by

$$C(h) = E[\{z(x) - \mu\}\{z(x+h) - \mu\}] = E[z(x)\,z(x+h)] - \mu^2 \qquad /8/$$

where $\mu$ is the expectation of Z, $E[z(x)]$.

The semi-variance is thus

$$\gamma(h) = C(0) - C(h) \qquad /9/$$

where C(0) is the covariance at zero lag, or the a priori variance of the process. Since the autocorrelation is

$$\varrho(h) = \frac{C(h)}{C(0)} \qquad /10/$$

we have

$$\gamma(h) = C(0)\{1 - \varrho(h)\} \qquad /11/$$

These relations require stronger assumptions than those of the intrinsic hypothesis. In particular there must be a finite a priori variance, and the variable must be stationary in both the mean and variance. It often happens that this cannot be assumed, and in these circumstances the semi-variance and the variogram exist, whereas the covariance and autocorrelation do not. For this reason the variogram is the more useful tool for describing soil variation quantitatively, and it is the one that we use now in most geostatistical applications.

*Estimating and modelling the variograms*

The variogram of a soil property in some region is useful only if we can estimate it and find a function to describe it. The first is best achieved by sampling at regular intervals along transects or on a grid. Often, however, data which have perhaps been recorded for other purposes are irregularly scattered. In any event the general formula

$$\gamma(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \{z(x_i) - z(x_i + h)\}^2 \qquad /12/$$

where m(h) is the numbers of pairs of points separated by the vector h, provides the usual estimate of $\gamma(h)$. By varying h in both distance and direction we obtain the ordered set which constitutes the sample variogram. The plotted points in Figures 1, 3 and 7 are examples.

This stage is fairly straightforward. The variance needs to be reasonably stable, as it does in many other forms of statistical analysis, and data that are strongly skewed should be transformed to approximate normality. The sample must also be large enough for the true semi-variance to be estimated with adequate confidence. For transect surveys there should be at least 100 measurements and the variogram computed to no more than about 1/5 of the run. In two dimensions some 400 measurements are likely to be needed to estimate anisotropy adequately. An investigator also needs some prior knowledge of the spatial scale to ensure that sampling is sufficiently intensive. This will be clearer when we examine a few variograms.

Choosing functions to describe variograms and fitting them to the sample estimates can be more problematic. The function must represent the salient features of the variogram and must also be such as to return only non-negative variables. Technically it must be conditional negative semidefinite /CNSD/.

Most variograms have fairly simple forms, at least over the lags to which they are usually estimated. They may show some or all of the following features. There is an increasing part that rises with increasing lag distance from near zero. This may continue indenfinitely. Alternatively it may flatten more or less abruptly or rise to an asymptote. The semi-variance at zero lag is itself zero, but many variograms appear to approach some larger value on the ordinate as the lag distance approaches zero. This value is known as the "nugget variance". The term comes from gold mining and represents the chance occurrence of finding a gold nugget in a drill core. The variogram may be more complex, but if we can find models for the simple forms we can combine them to describe the more complex ones.

As it happens, the simple functions available to describe variograms can be divided into two main families depending on whether the variogram is bounded or not.

*Bounded models.* - These models are often known as transitive ones. The underlying idea is that they derive from overlapping zones of influence - transition zones. Their general form in one dimension or where variation is isotropic is:

$$\gamma(h) = c\,\{f(h)\} \qquad /13/$$

where c is the a priori variance and f(h) is a function of the lag distance that increases from 0 to 1, and h is now a scalar in distance only.
The two most commonly fitted functions in this group are the spherical model:

$$\begin{cases} f(h) = \dfrac{3}{2}\dfrac{h}{a} - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3 & \text{for } h \leq a \\ f(h) = 1 & \text{for } h > a \end{cases} \qquad /14/$$

and the exponential model:

$$f(h) = 1-\exp(-h/r) \qquad /15/$$

In these equations a and r are distance parameters that define the extent of the spatial dependence. The spherical model reaches its maximum or sill at $h = a$, the range of the model. The exponential function approaches its sill asymptotically and has no definite range, though for practical purposes the effective range is often taken as $a' = 3r$. The variograms illustrated in Figs. 1, 3 and 7 are all examples of these.

Other models belonging to this family include the bounded linear, circular and penta-spherical functions, all defined in McBRATNEY and WEBSTER /1986/. The first two must be used with caution: the first is valid in only one dimension and the second in only one and two dimensions.

*Unbounded models.* - These are models without an a priori variance: the variance increases indefinitely with increasing lag distance or at least appears to on the evidence available. A general equation for the isotropic variogram of this type is the power function:

$$\gamma(h) = bh^{\alpha} \qquad /16/$$

The theoretical origin of this model lies in the traces produced by Brownian Motion. Unconstrained Brownian motion produces traces in which the changes produced in successive steps are uncorrelated and whose variograms are linear as a consequence; i.e. $\alpha = 1$ and the gradient is b. If the successive changes are positively correlated then $\alpha < 1$; contrarily if the changes are negatively correlated then $\alpha < 1$. The value of $\alpha$ must, however, be between 0 and 2 with the limits excluded.

The above functions all pass through the origin. Yet we have noted that many sample variograms seem to approach the origin at some limiting value greater than zero, the nugget variance. Similarly, although in defining the power function above we excluded ones with $\alpha = 0$ we have to recognize that some variograms appear to be flat; i.e. wholly nugget. These are defined formally using a Dirac function, $\delta(h)$, which takes the value of 1 when $h = 0$ and zero otherwise. The pure nugget variogram is then defined formally as

$$\gamma(h) = c_0 \{1 - \delta(h)\}, \qquad /17/$$

where $c_0$ is the variance as $h \to 0$.

We can then combine this with any of the above models of spatial dependence to describe variograms that appear to have nugget variances. Figures 3 and 7 show actual examples of such combinations.

Another combination that has proved valuable both in mining and soil science /McBRATNEY et al., 1982; WEBSTER and OLIVER, 1989a/ is the double spherical model. Fig. 1 is an example. It is the variogram of available cobalt in the topsoil of south-east Scotland. The two spherical components have distinctly different ranges. The larger at 15 km is almost certainly due to major changes in geology: the different Paloeozoic formations contain different amounts of copper. The component with the shorter range of about 3 km probably represents the farm-to-farm variation.

### *Fitting models*

With experience we may choose a model from the appearance of the sample variogram, or less often because we believe that a particular kind of model is appropriate. This model must then be fitted to the sample estimates. Techniques for fitting can range from full maximum likelihood estimation /e.g. KITANIDIS, 1983; MARDIA and MARSHALL, 1984/, which is generally regarded as the most reliable, to fitting by eye, which can scarcely be considered quantitative and has little to commend it. Unfortunately the maximum likelihood method is very demanding computationally and feasible only for fairly small samples, say up to 150. Surveys can produce many more data than that and must if anisotropic variation is to be described adequately.

A sound compromise is to fit models by weighted least squares estimation. The method is based on the assumption that the differences between the observed semi-variances and the fitted values are normally distributed and independent of one another. These assumptions are unlikely to hold exactly, but that might not matter. More seriously, the variance of $\hat{\gamma}$ depend on both the magnitude of the true values and on the numbers of pairs of comparisons, m(h) in equation /12/, used to estimate them. The former are unknown. The latter, however, are and they at least should be used to provide weights in the minimization. Further refinements are feasible, and some of these are described and discussed by CRESSIE /1985/ and McBRATNEY and WEBSTER /1986/.

Most of the models are non-linear in their parameters, and so a good computer program is needed for the fitting. At Rothamsted I use MLP, the Maximum Likelihood Program /ROSS, 1980/, for this purpose. The same algorithms are now embodied in the more widely available Genstat 5 /Genstat 5 Committee, 1987/ which was also written at Rothamsted.

We should realize that we generally choose a model for a variogram because it appears to fit well. The model does not necessarily represent a generating process. Nevertheless, if the model fits well it will serve empirically for estimation by kriging and designing sampling schemes, which I deal with next.

*Estimation*

The principal application of geostatistics is estimation. The theory was developed to meet the needs of mining, and it is now applied in many parts of the world to estimate the concentrations of metal in ore bodies and recoverable reserves. The techniques for doing it go under the general name of kriging, after D. G. KRIGE, a mining engineer, who developed some of them empirically for the South African goldfields /see KRIGE, 1966/.

My colleagues and I at Rothamsted and several other scientists have used the same techniques for estimating soil properties. Interestingly, here it is the desire to map distributions that has been the driving force. Soil scientists have found that simple kriging serves well to estimate properties of the soil. I give below the relevant equations, and then I present examples of these to mapping.

In simple kriging a kriged estimate is no more than a weighted average of the data. Suppose we wish to estimate the average value of a soil property Z in a block of land B. Denote this by $\hat{z}(B)$. Then

$$\hat{z}(B) = \sum_{i=1}^{n} \lambda_i \, z(x_i) \qquad /18/$$

where $\lambda_i$, $i = 1, 2, \ldots, n$ are weights.

We want $\hat{z}(B)$ to be an unbiased estimate, and so the weights are chosen so that

$$\sum_{i=1}^{n} \lambda_i = 1 \qquad /19/$$

The estimation variance, $E[\{z(B) - \hat{z}(B)\}^2]$, is given by

$$\sigma_k^2 = 2 \sum_{i=1}^{n} \lambda_i \, \bar{\gamma}(x_i,B) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i,x_j) - \bar{\gamma}(B,B), \qquad /20/$$

where $\gamma(x_i,x_j)$ is the semi-variance between sampling points i and j, $\bar{\gamma}(x_i,B)$ is the average semi-variance between the ith sampling point and the block B, and $\bar{\gamma}(B)$ is the average semi-variance of Z within the block, i.e. the within-block variance. This variance is minimized when

$$\sum_{i=1}^{n} \lambda_i \, \gamma(x_i,x_j) + \psi = \bar{\gamma}(x_i,B) \quad \text{for all } j, \qquad /21/$$

and this introduces the Lagrange multiplier, $\psi$, needed for the minimization. We thus have a set of n linear equations in n unknowns, the weights, plus an additional equation for $\psi$, and these are solved to find the weights. The estimation variance itself is obtained as a by-product as

$$\sigma_k^2 = \sum_{i=1}^{n} \lambda_i \, \gamma(x_i,B) + \psi - \bar{\gamma}(B,B). \qquad /22/$$

Simple kriging is thus a true statistical procedure for estimation. It gives the Best Linear Unbiased Estimate, and is sometimes known as BLUE therefore.

In soil survey we may wish to estimate the value of soil properties at points no bigger than the supports of the sample. In these cases B becomes a point, say $x_0$; the term $\bar{\gamma}(B,B)$ disappears from equation /20/ and $\bar{\gamma}(x_i,B)$ becomes the simple semi-variance between $x_i$ and $x_0$, $\gamma(x_i,x_0)$.

If there is any appreciable spatial dependence in the data then the kriging weights for sampling points nearest to the block being estimated are large, and usually only the nearest 16 to 20 carry sufficient weight to be of consequence. All others are negligible. Thus kriging is local, and this seems intuitively desirable. It also has important consequences for computing because it means that the matrices which have to be inverted to solve equations /21/ are never large.

## Examples

Our appreciation of soil properties as regionalized variables satisfying the intrinsic hypothesis is recent. It is a product of the last ten or twelve years. So too is our experience of modelling the spatial variation. But already there are numerous examples in which the theory and techniques have been applied to estimate and interpolate soil properties optimally and to map them from sample data. I end this paper with three examples from our own recent experience at Rothamsted. The examples derive from my research with A. B. McBRATNEY, R. G. McLAREN, R. B. SPEIRS and I. M. BURAYMAH to whom I am grateful.

### *Cobalt in the soil of south east Scotland*

Deficiences of copper and cobalt can cause serious disorders, poor growth and even death in cattle and sheep. In south east Scotland cobalt deficiency is widespread, while locally there is deficiency of copper also. To advise farmers and alert them to the risks of these deficiencies the East of Scotland College of Agriculture samples the soil in the region and analyses it for these elements. Practice is to take a bulked sample of 20 randomly located cores of topsoil /0 to 20 cm/ from individual fields of 5 to 10 ha, and to measure the copper and cobalt extracted with mild reagents to give a value of the elements available to the plants and hence to the grazing livestock. More than 3500 samples of soil had been analysed when McBRATNEY et al. /1982/ analysed the accumulated data statistically. Here I present results just for cobalt in the eastern part of the region.

*Table 1*
Sample statistics of extractable cobalt in the topsoil of south east Scotland

| Index | Untransformed mg/kg | Transformed to $\log_{10}$ mg/kg |
|---|---|---|
| Mean | 0.271 | -0.613 |
| Standard deviation | 0.134 | 0.196 |
| Skewness | 1.990 | 0.126 |

Table 1 summarizes the statistics for available cobalt. The original measurements were strongly skewed and appeared to have a lognormal distribution /McBRATNEY et al., 1982/. The data were therefore transformed to their common logarithms to stabilize the variances. The variogram is shown in Fig. 1 with the sample values plotted as points. It is isotropic. The solid line is that of a double spherical model with nugget:

$$\begin{cases} \gamma(h) = c_0 + c_1 \left\{\frac{3}{2}\frac{h}{a_1} - \frac{1}{2}\left(\frac{h}{a_1}\right)^3\right\} + c_2 \left\{\frac{3}{2}\frac{h}{a_2} - \frac{1}{2}\left(\frac{h}{a_2}\right)^3\right\} & \text{for } 0 < h \leq a_1 \\ \gamma(h) = c_0 + c_1 + c_2 \left\{\frac{3}{2}\frac{h}{a_2} - \frac{1}{2}\left(\frac{h}{a_2}\right)^3\right\} & \text{for } 0_1 < h \leq a_2 \\ \gamma(h) = c_0 + c_1 + c_2 & \text{for } h > a_2 \\ \gamma(0) = 0 \end{cases} \qquad /23/$$

The coefficients are given in Fig. 1. As mentioned above the two spherical components with ranges of 3.4 km and 16.4 km represent two distinct sources of variation. The latter arises almost certainly from the major geological changes in the region. The former seems to represent farm-to-farm variation.

Using the variogram and the data the cobalt content of the soil was estimated as its common logarithm for 1 $km^2$ blocks at 0.5 km intervals on a square grid. Isarithms /"contours"/ were then threaded through the resulting figure field to produce the map, Fig. 2. A value of 0.25 mg Co/kg soil is regarded as critical for animal nutrition. A smaller concentration is likely to cause defficiency in livestock, and so on the map the part of the region where the estimated value is less than $\log_{10} 0.25 = -0.602$ is stip-

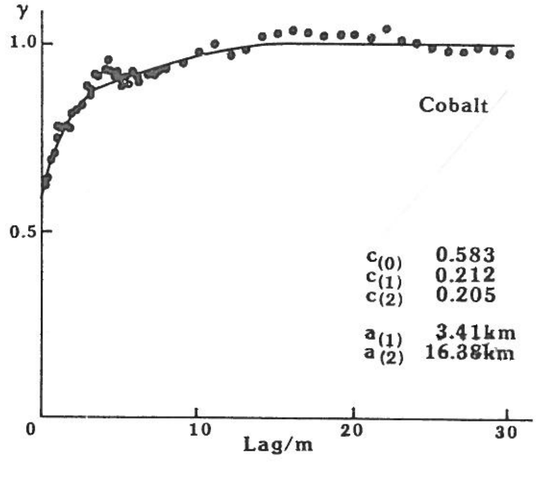


*Fig. 1*

Variogram of $\log_{10}$ cobalt in the topsoil of south east Scotland pled. WEBSTER and OLIVER /1989a,b/ have extended this study to map the probability that the true concentration is less than the critical value by disjunctive kriging /MATHERON, 1976/.

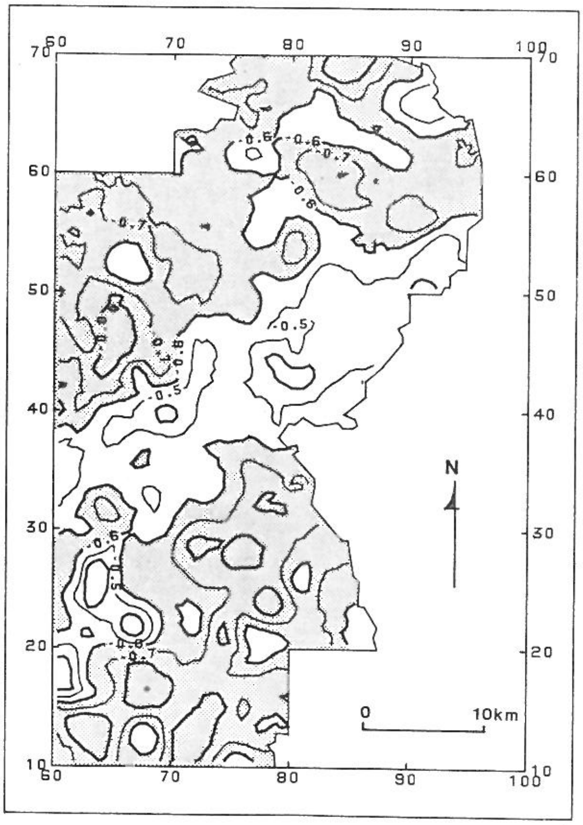


*Fig. 2*
Map of the amount of available cobalt in the topsoil of south east Scotland. Isarithms are in $\log_{10}$ (mg Co/kg soil). The stippled areas are judged deficient ($\log_{10}$ Co < -0.602)

*Potassium content of the soil at Broom's Barn*

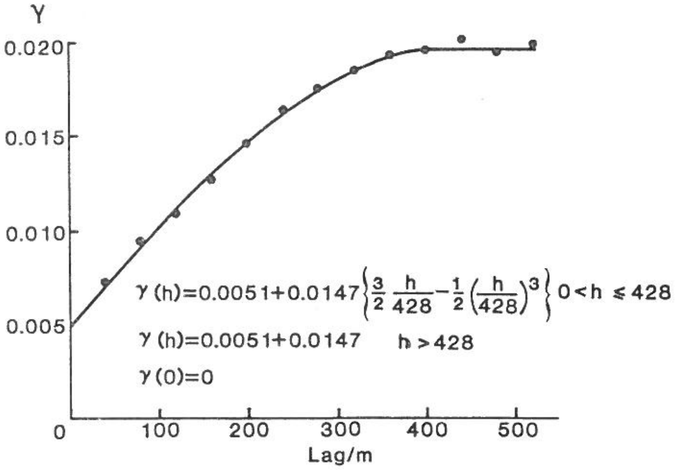
Broom's Barn Farm near Bury St Edmunds in Suffolk /eastern England/ covers some 77 ha of arable land. It was bought for research on sugar beet in 1959, and in the following year the fertility of its soil was assessed by sample survey. The topsoil was sampled by bulking 25 cores to 23 cm at random within 16 m x 16 m squares at 40 m intervals on a square grid. The soil in each square was then analysed for pH, available phosphorus, and exchangeable potassium, magnesium and sodium. Maps were then made to show the variation in these properties /DRAYCOTT et al., 1976/. Since then we have analysed the data geostatistically and mapped the distribution of pH, phosphorus and potassium /WEBSTER and McBRATNEY, 1987/ and the conditional probabilities that the true /but unknown/ values of these properties are less than the critical thresholds for good arable cropping /WEBSTER and OLIVER, 1989a,b/. Here I present the results of simple kriging of potassium.

*Table 2*
Sample statistics of exchangeable potassium in the topsoil at Broom's Barn

| Index | Untransformed mg/kg | Transformed to $\log_{10}$ mg/kg |
|---|---|---|
| Mean | 26.30 | 1.39 |
| Variance | 81.89 | 0.01811 |
| Skewness | 2.02 | 0.36 |

The sample statistics are given in Table 2. The original measurements were strongly skewed, and they were therefore transformed to their common logarithms to stabilize the variances. Fig. 3 shows the variogram of the transformed values with a spherical model fitted. The variation is isotropic.

As for cobalt, the variogram and data were used to estimate the values of $\log_{10}$ potassium on a fine grid. Fig. 4 is an isarithmic map of the 50 m x x 50 m block estimates. The results can also be represented as perspective diagrams. Fig. 5A is such a diagram of the same block estimates, and it shows a fairly smoothly varying surface. Fig. 5B is a perspective diagram of the punctual estimates. Striking in this diagram are the spikes at the sampling points. These illustrate the nugget effect. In punctual kriging the value estimated at a sampling point is the measured value there. In equation



$$\gamma(h)=0.0051+0.0147\left\{\frac{3}{2}\frac{h}{428}-\frac{1}{2}\left(\frac{h}{428}\right)^3\right\}\quad 0<h\leq 428$$

$$\gamma(h)=0.0051+0.0147\quad h>428$$

$$\gamma(0)=0$$

*Fig. 3*
Variogram of $\log_{10}$ exchangeable potassium at Broom's Barn

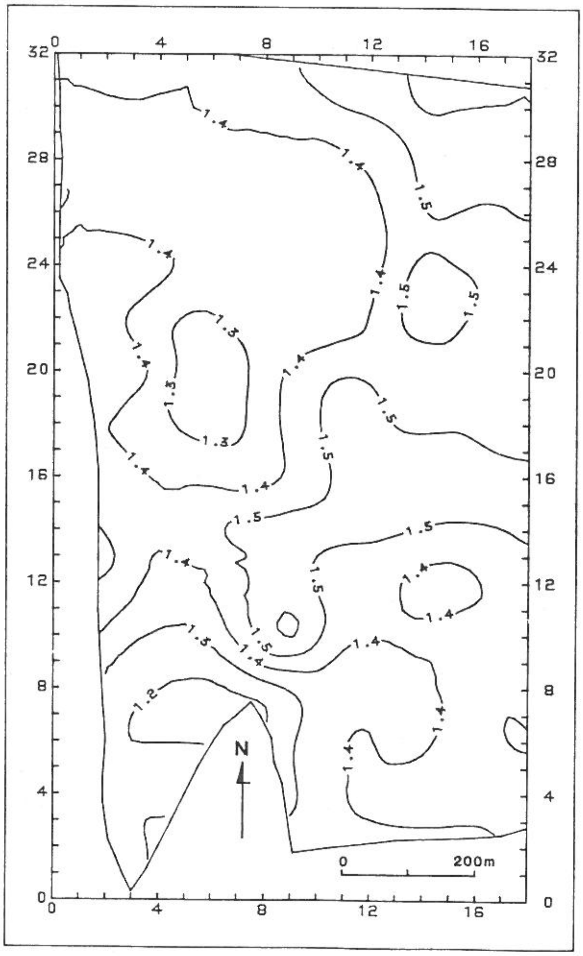


*Fig. 4*
Isarithmic map of $\log_{10}$ exchangeable potassium in the topsoil at Broom's Barn

/18/ the weight, $\lambda$, there is 1 and all the other weights are zero. And in equation /22/ the estimation variance is zero. Elsewhere the estimates are local averages. In the presence of a nugget variance, $c_0 = 0.0051$ in this case, there is a discontinuity in the interpolated surface, and Fig. 5B is an example of this.

As described above and formalized in equation /22/ we obtain estimates of the kriging variance, and we can display these too. Figure 6A and 6B show these as perspective diagrams for block and punctual kriging, respectively. Notice how in general the estimation variances for punctual kriging are much larger than those for block kriging, though they are zero at the sampling points. The variances increase sharply beyond the limits of sampling around the margin of the farm. They are also large in the north west

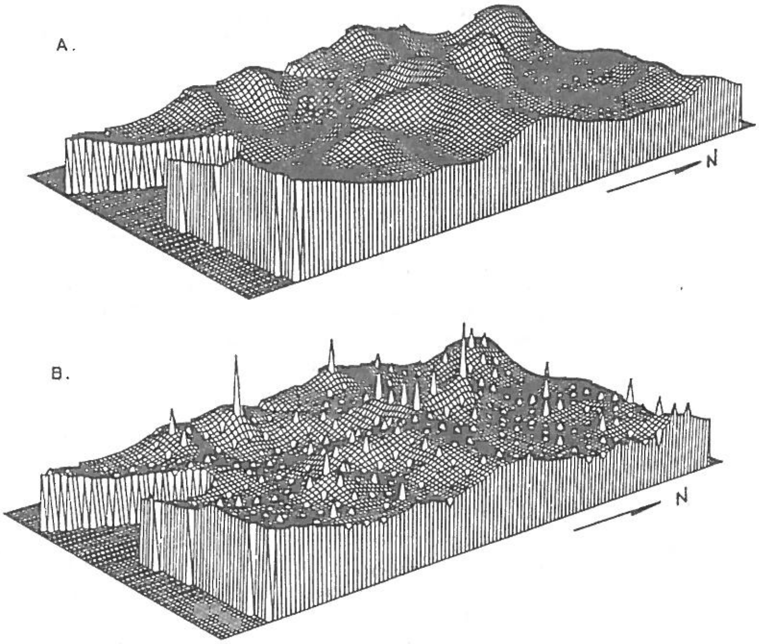


*Fig. 5*
Perspective diagrams of block estimates /A/ and punctual estimates /B/ of available potassium at Broom's Barn

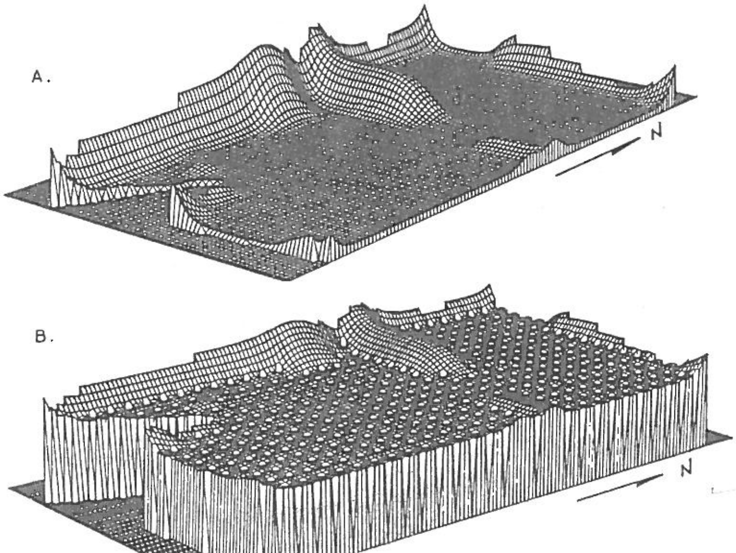


*Fig. 6*
Perspective diagrams of the estimation variances of block estimates /A/ and punctual estimates /B/ of available potassium at Broom's Barn

where the farm buildings and laboratories are, and along the access road from the east. And there is a small "hump" in the surface in the south west where data were missing from the grid.

### *Electrical conductivity in the soil of the Gezira*

The third example is from a survey of the experimental plots of land on the Research Station at Wad Medani in the Sudan Gezira. The Station was established in 1926 when extensive irrigation of the Gezira began using water from the Blue Nile. The standard practice on the cultivated land was to grow cotton, the principal cash crop, in rotation with wheat and vegetables. The climate is very dry, and all crops are irrigated by flooding.

As part of the experimental program two plots, each of 0.4 ha, were set aside in 1926 to receive constant treatment indefinitely. The one was left in its natural condition, unirrigated and growing sparse grass. The other was managed as if it were a standard commercial field growing cotton in rotation with wheat and vegetables. It was irrigated from the southern end and drained to the north. And it was cultivated lengthwise on all occasions.

In 1985 I. M. BURAYMAH surveyed the two plots to find out what effect the irrigation and cropping had had on the soil and in particular if the soil had become more salty or alkaline. He sampled it on a grid at 6.25 m intervals but with one quarter of the nodes omitted. The soil was sampled to 30 cm using a bucket auger of 8 cm diameter, taking five randomly chosen cores within a circle of 1 m radius around each node. The soil was then analysed for pH, electrical conductivity and the cations Na, Mg, and Ca, from which the sodium adsorption ratio /SAR/ was calculated.

BURAYMAH and WEBSTER /1989/ report the results, and here I present those for just electrical conductivity. Table 3 summarizes the sample statistics.

As with the cobalt and potassium concentrations in the two previous examples the data were strongly skewed and seemed to be lognormally distributed. We therefore worked on the values transformed to their common logarithms. The variograms for the two plots are shown in Fig. 7. Fig. 7A for the natural plot is virtually flat: it is almost wholly nugget variance at the working scale. The variogram for the irrigated plots is more interesting. It shows distinct spatial dependence, and it is also anisotropic: there is more variation across the plot than along it. This is shown by the different symbols for the semi-variances estimated for the different directions, as I have described elsewhere /WEBSTER, 1985/. In this instance the anisotropy has been treated as defined by the function $\Omega$:

$$\Omega(\theta) = \{A^2 \cos^2 (\theta - \Phi) + B^2 \sin^2 (\theta - \Phi)\}^{\frac{1}{2}} \qquad /24/$$

where $\theta$ denotes direction, $\Phi$ is the direction in which the distance parameter A is greatest and B is the distance parameter in the perpendicular direction. The best fitting model was exponential, defined by

$$\gamma(h, \theta) = c_0 + c_1[1 - \exp\{-h/\Omega(\theta)\}] \qquad /25/$$

The values of the coefficients are as follows:

$c_0 = 0.00654$; A = 150.0 m; $\theta = 1.30$ radians.
$c_1 = 0.00542$; B = 60.5 m;

*Table 3*

Sample statistics of electrical conductivity in the Sudan Gezira

| | Index | Electrical conductivity /EC/ $mS \cdot cm^{-1}$ | Transformed to $\log_{10}$ /EC/ |
|---|---|---|---|
| Natural plot | Mean | 0.524 | -0.284 |
| | Variance | 0.005230 | 0.003209 |
| | Skewness | 1.47 | 0.65 |
| Irrigated plot | Mean | 0.501 | -0.312 |
| | Variance | 0.01764 | 0.009934 |
| | Skewness | 2.10 | 1.03 |

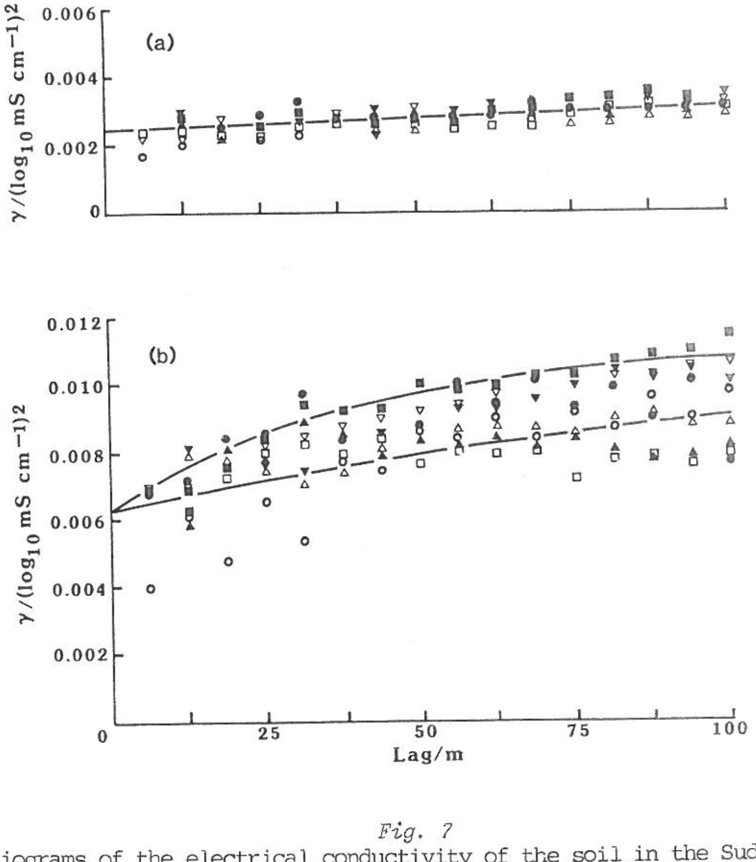


*Fig. 7*

Variograms of the electrical conductivity of the soil in the Sudan Gezira in the natural state /a/ and after some 60 years of irrigation and cultivation /b/

Using this model a surface was estimated by kriging over blocks of 10 m x 10 m. The result is shown in Fig. 8B. For comparison the surface of conductivity for the natural plot is also shown, Fig. 8A. As expected from its variogram the conductivity surface of the natural plots is almost flat. In contrast that of the irrigated plot shows distinct waves parallel to the long dimension of the plot and to the Blue Nile some 4 km away. This is the anistropy prominent in the variogram. The surface also has numerous small ridges almost perpendicular to the larger waves. These are not immediately evident in the variogram and were not modelled. Nevertheless the source of variation is evident in the estimates in direction 0. The semi-variance increases to a maximum at about 20 m, decreases and then increases again: there is repetition.

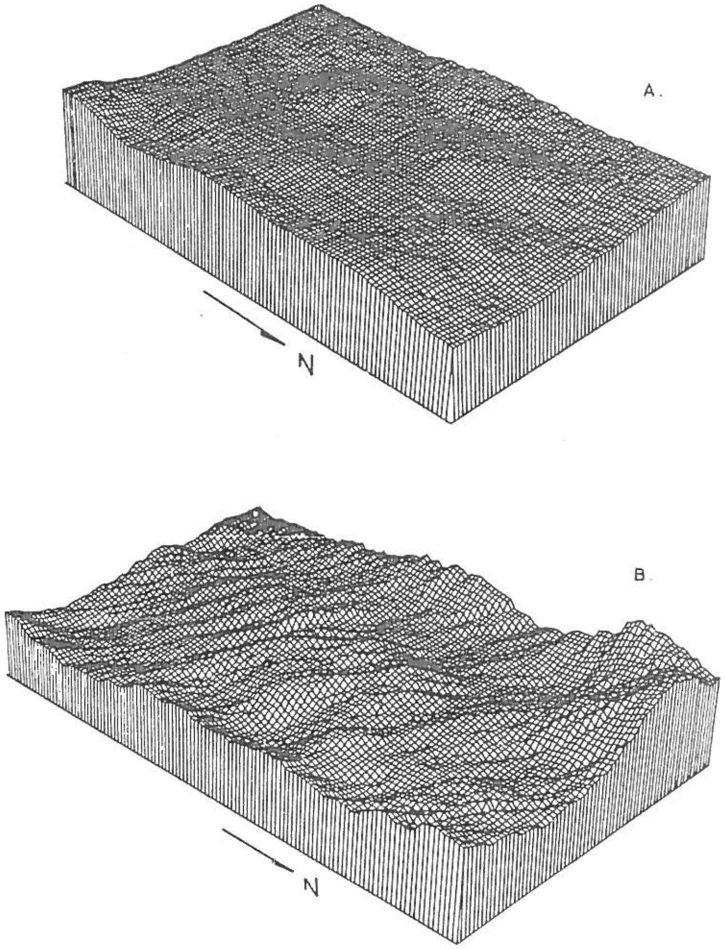


*Fig. 8*
Perspective diagrams of the electrical conductivity of the soil in the Sudan Gezira, in the natural state /A/, and after some 60 years of irrigation and cultivation /B/

As Table 3 shows irrigation, here with fairly pure water from the Blue Nile, has not increased the salinity of this soil. It has, however, increased the variability. This is still not serious in this instance, but on more salty land concentrating salts locally by irrigation could be very serious. Periodic geostatistical analysis of soil under irrigation could show whether variation was increasing, warn managers of trends and identify those regions where preventative action is needed before it becomes too late and requires more expensive remedial treatments.

## Conclusions

The last ten years have seen enormous progress in the adaptation and application of advanced statistical theory to practical problems in the earth sciences. A few of us in soil science have been able to play our part. We have learned much about the statistical nature of soil variation, and now with our current understanding and the right tools we should be able to tackle many more problems in estimation, spatial prediction and mapping. There is a rich future for anyone prepared to get to grips with the subject.

## Summary

Spatial prediction of soil conditions relied for many years on using traditional methods of classification to stratify regions according to soil type and then applying classical statistical technique for estimation within the strata. It has had its successes and failings. The recent development of geostatistics has provided new tools for spatial prediction. In many instances soil properties are best regarded as realizations of random processes. Their spatial variation can be described adequately by the variogram assuming the intrinisic hypothesis of stationarity. Values of a soil property at unvisited sites or over larger blocks of land can be estimated without bias and with minimum variance by kriging.

The paper summarizes the underlying theory, presents the computational steps, and illustrates the procedures with results from surveys of cobalt in the soil of south east Scotland, with exchangeable potassium over an arable farm in England and the effect of irrigation on the electrical conductivity of the soil in the Sudan Gezira.

## References

BECKETT, P. H. T. and WEBSTER, R., 1971a. The development of a system of terrain evaluation over large areas. Royal Engineers' Journal. 85. 243-258.

BECKETT, P. H. T. and WEBSTER, R., 1971b. Soil variability: a review. Soils and Fertilizers. 34. 1-15.

BURAYMAH, I. M. and WEBSTER, R., 1989. Variation in soil properties caused by irrigation and cultivation. Soil and Tillage Research?. 11. 57-74.

CRESSIE, N., 1985. Fitting variogram models by weighted least squares. Mathematical Geology. 17. 563-585.

DRAYCOTT, A. P. et al., 1976. Changes in Broom's Barn Farm soils, 1960-1975. In: Rothamsted Experimental Station Report for 1976, Part 2. 33-52.

GENSTAT 5 COMMITTEE, 1987. Genstat 5 reference manual. Clarendon Press, Oxford.

KITANIDIS, P. K., 1983. Statistical estimation of polynomial generalized covariance functions and hydrological applications. Water Resources Research. 19. 909-921.

KOZLOVSKII, F. I. and SOROKHINA, N. P., 1976. The soil individual and elementary analysis of the soil-cover pattern. In: Soil combinations and their genesis. /Ed.: FRIDLAND, V. M./. 55-64. Amerind Publishing Co. New Delhi.

KRIGE, D. G., 1966. Two-dimensional weighted moving average trend surfaces for ore evaluation. Journal of the South African Institute for Mining and Metallurgy. 66. 13-38.

MARDIA, K. V. and MARSHALL, R. J., 1984. Maximum likelihood models for residual covariance in spatial regression. Biometrika. 71. 135-146.

MATHERON, G., 1965. Les variables régionalisées et leur estimation. Masson. Paris.

MATHERON, G., 1971. The theory of regionalized variables and its applications. Cahiers du Centre de Morphologie Mathématique. 5. Fontainebleau.

MATHERON, G., 1976. A simple substitute for the conditional expectation: the disjunctive kriging. In: Advanced geostatistics in the mining industry. /Eds.: GUARASCIO, M., DAVID, M. and HIUJBREGTS, C./ 221-236. Reidel. Dordrecht.

McBRATNEY, A. B. and WEBSTER, R., 1986. Choosing models for semi-variograms of soil properties and fitting them to sampling estimates. J. Soil Sci. 37. 617-639.

McBRATNEY, A. B. et al., 1982. Regional variation in extractable copper and cobalt in the topsoil of South-East Scotland. Agronomie. 2. 969-982.

ROSS, G. J. S., 1980. MLP Maximum Likelihood Program. Rothamsted Experimental Station. Harpenden.

WEBSTER, R., 1973. Automatic soil boundary location from transect data. Mathematical Geology. 5. 27-37.

WEBSTER, R., 1985. Quantitative spatial analysis of soil in the field. Advances in Soil Science. 3. 1-70.

WEBSTER, R. and BECKETT, P. H. T., 1968. Quality and usefulness of soil maps. Nature, London. 219. 680-682.

WEBSTER, R. and BECKETT, P. H. T., 1970. Terrain classification and evaluation using air photography: a review of recent work at Oxford. Photogrammetria. 26. 51-75.

WEBSTER, R. and BURGESS, T. M. 1980. Optimal interpolation and isarithmic mapping of soil properties. III. Changing drift and universal kriging. J. Soil Sci. 31. 505-524.

WEBSTER, R. and CUANALO de la C, H. E., 1975. Soil transect correlograms of North Oxfordshire and their interpretation. J. Soil Sci. 26. 176-194.

WEBSTER, R. and McBRATNEY, A. B., 1987. Mapping the soil fertility at Broom's Barn by simple kriging. Journal of the Science of Food and Agriculture. 38. 97-115.

WEBSTER, R. and OLIVER, M. A., 1989a. Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. J. Soil Sci. 40. /In press/.

WEBSTER, R. and OLIVER, M. A., 1989b. Disjunctive kriging in agriculture. In: Proceedings of the 3rd Geostatistics Congress. Reidel. Doredrecht. /In press/.