# SESSION 6
# QUANTITATIVE ASPECTS

Chairman & Session Organiser: Dr C D Brown
*Cranfield University, Silsoe, UK*

# Examining the spatial variation of environmental properties using geostatistics

Z L Frogbrook
*Department of Soil Science, University of Reading, Whiteknights, Reading, RG6 6DW, UK*
*Email: z.l.frogbrook@reading.ac.uk*

## ABSTRACT

Environmental properties vary across the surface of the earth and often at many different spatial scales. Over the last thirty years, geostatistics has been used successfully to analyse spatial data throughout the earth sciences. This paper gives a brief introduction to geostatistics. It describes the variogram, a model of the spatial variation, and kriging, a method of prediction. A case study shows how the spatial variation of soil organic matter can be explored geostatistically within a single field. The importance of sampling intensity is examined by sub-sampling a dense grid of data.

## SPATIAL VARIATION

Soil plays an important role in determining the fate of pesticides in the environment. Soil chemical, physical and biological properties influence the adsorption and degradation of pesticides. Soil properties, however, like many environmental properties vary more or less continuously across the earth's surface. The variation of these properties is complex because of the interaction of many different processes that operate at different spatial scales. Burrough (1983) suggested that the variation in soil properties occurs at all levels of resolution from millimetres to hundreds of kilometres. Figure 1 shows three possible levels of spatial variation superimposed on one another: one over hundreds of metres defined by two classes separated by a boundary (the steep slope), an intermediate scale of variation over tens of metres, and short scale variation over distances of less than a few metres. The latter appears to be unstructured and locally erratic, a feature often referred to as noise. It is important to remember, however, that what is observed as noise at one resolution can appear as structure at another and vice versa; it depends on the scale at which the variation is resolved (Oliver, 1999).

## SAMPLING AND ESTIMATION

Information on environmental properties is often restricted to observations on small areas or volumes of the survey area, i.e. a sample. To provide an overall view of the variation there is a need to predict values at unsampled sites. The two main methods that have been used for prediction are classification and interpolation.

Classification is the traditional approach to prediction where the mean value of a property is used as the predictor at all places within the class. Conventional farm management tends to use this approach and each field is treated as a distinct class for fertiliser applications, for example. Classification assumes that the classes account for all of the spatially correlated variation and that any remaining variation is random. In other words, it is assumed that there

is no statistical relation between values a given distance apart within the soil class. If this is the case then classification is a reasonable approach. The number of sample locations needed to estimate the mean values of the soil properties reliably for each class can be determined by conventional statistics for a given level of confidence (Webster and Oliver, 1990). If, as is usually so, there is spatially correlated variation remaining within the classes this represents variation that could be resolved but has not been.

Interpolation is an alternative to classification that expresses the continuity in the variation. This approach to prediction can provide the local detail that is often required for environmental management. All methods of interpolation assume implicitly that there is a statistical relation between values a given distance apart. Conventional methods of interpolation, such as inverse squared distance and nearest neighbour interpolation, provide no means of assessing this and there is a risk of predicting from data that are spatially uncorrelated. Geostatistics overcomes many of the weaknesses of traditional interpolation and has been used successfully to analyse spatial data throughout the earth sciences.

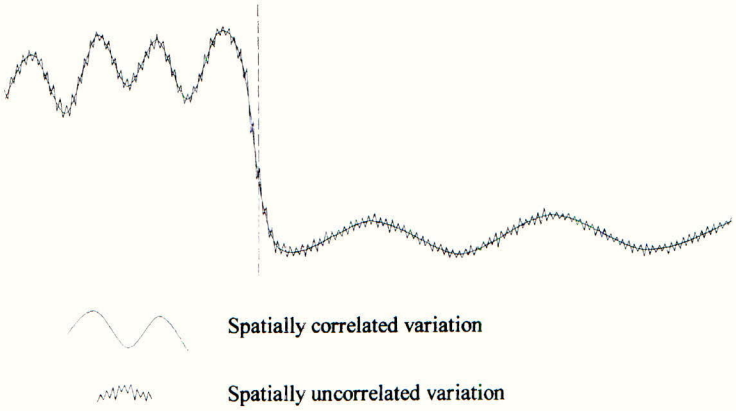


Figure 1. Different scales and types of spatial variation.

## GEOSTATISTICS

Geostatistics has been well documented in several texts (e.g. Goovaerts, 1997; Isaaks & Srivastava, 1989; Webster & Oliver, 2001) and only a brief introduction is given here.

The *variogram* is the central tool of geostatistics. It describes the spatial structure of the variation by measuring the degree of correlation between sampling points a given distance apart. This measure is based on the notion that samples are more similar at nearby locations than at distant ones. The standard formula for computing the variogram is:

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

where $\hat{\gamma}(\mathbf{h})$ is the estimated semivariance, $z(\mathbf{x}_i)$ and $z(\mathbf{x}_i + \mathbf{h})$ are the measured values of $Z$ at any two places $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{h}$ separated by $\mathbf{h}$, a vector having both distance and direction known as the lag, and $M(\mathbf{h})$ is the number of paired comparisons at that lag. By changing $\mathbf{h}$,

an ordered set of values is obtained and this is the experimental variogram. The semivariance is computed for discrete values of **h**, but the experimental variogram represents the regional variogram, which is continuous. The latter is represented by authorized mathematical functions fitted to the experimental values.

Figure 2 illustrates some of the main features of variograms. In the first example, Figure 2a, the semivariances increase initially and then reach an upper limit, or bound. This maximum is known as the *sill* variance and it estimates the *a priori* variance of the process. The distance at which the sill is reached is the *range*, denoting the limit of spatial correlation (or spatial dependence); sampling locations separated by distances greater than this are spatially uncorrelated. The variogram in Figure 2b appears to increase indefinitely; it is unbounded. It suggests that the full extent of the spatial variation has not been encompassed at the scale of investigation. Both variograms meet the ordinate at a positive value, known as the *nugget variance*. This is common when the experimental semivariances are extrapolated to the origin. For continuous properties the nugget variance encompasses any measurement error and spatial variation occurring within the shortest sampling interval, i.e. the unresolved spatial variation. The latter is usually the larger of these (Oliver & Frogbrook, 1998). Sometimes the variogram appears to be flat, it is pure nugget, Figure 2c. For continuous properties this usually means that the sampling has failed to resolve the spatial variation present and that all of the spatial structure is contained within the smallest sampling interval. The spatial structure can only be identified by more intensive sampling.

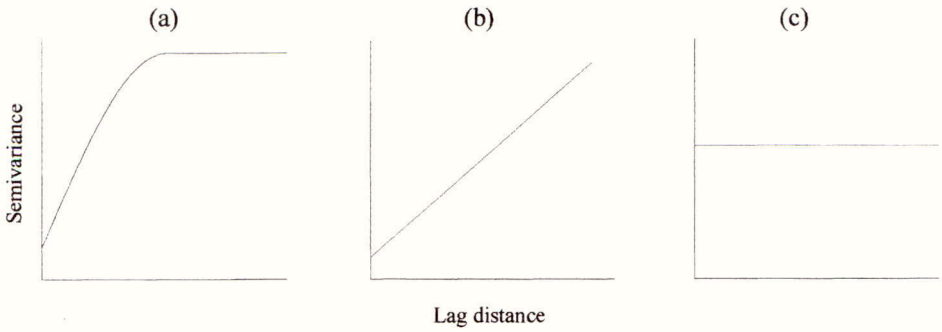


Figure 2. Forms of variogram (a) bounded (b) unbounded (c) pure nugget.

The method of prediction embodied in geostatistics is known as *kriging*, which is a general term that embraces several types of prediction. The most commonly used form is *ordinary kriging*. Kriging is a method of local weighted moving averaging of the sample data within the neighbourhood of the point to be predicted. The weights depend on the variogram and the configuration of the sampling points, and are allocated in such a way as to minimize the kriging variance and to ensure that the estimates are unbiased. Kriging is optimal in this sense (Webster & Oliver, 2001). Kriged estimates can be made for points (punctual kriging) or over areas (block kriging). Punctual kriging is an exact interpolator and the kriged prediction at a sampling site is the observed value there and the kriging variance is zero. Block kriging results in smoother estimates and smaller kriging variances overall.

$$\hat{Z}(B) = \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i), \tag{2}$$

where $n$ usually represents the data points within the local neighbourhood, $A$, and is usually much less than the total number in the sample, and $\lambda_i$ are the weights. The weights are chosen to sum to 1 to avoid bias and to minimise the kriging variance:

$$\sum_{i=1}^{n} \lambda_i = 1. \tag{3}$$

The kriging variance of $\hat{Z}(B)$ is:

$$\begin{aligned} \sigma^2(B) &= \mathrm{E}\left[\left\{\hat{Z}(B) - Z(B)\right\}^2\right] \\ &= 2\sum_{i=1}^{n} \lambda_i \bar{\gamma}(\mathbf{x}_i, B) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) - \bar{\gamma}(B, B) \end{aligned} \tag{4}$$

where $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ is the semivariance between points $\mathbf{x}_i$ and $\mathbf{x}_j$, $\bar{\gamma}(\mathbf{x}_i, B)$ is the average semivariance between data point $\mathbf{x}_i$ and the block $B$, and $\bar{\gamma}(B, B)$ is the within block variance.

Some properties are sampled sparsely because they are expensive or difficult to measure and the predictions obtained from them are likely to be unreliable. In this situation, the predicted value might be improved by using the spatial relation, or coregionalization, between other better sampled properties that are cheaper or easier to measure. This procedure is known as *cokriging*. McBratney & Webster (1983) took this approach to improve the prediction of topsoil silt using the better sampled subsoil silt and sand. Frogbrook and Oliver (2001) used this to improve the predictions of soil organic matter content.

In agriculture many management decisions, such as the application of lime, are made using critical thresholds. For example, a farmer might decide that lime is needed only in areas of the field where the soil pH is 5.5 or less. One approach would be to map the kriged predictions of soil pH and from this map identify regions where an application of lime is required. However, the kriged values are estimates only and as such they are subject to error. In areas where the predictions are much less than or much greater than this threshold the decision as to whether to act is easy but when predictions are close to the threshold this decision is harder to make. To aid this decision an estimate is required of the probability that the threshold is, or is not, exceeded (Oliver *et al.*,1996). This can be determined by *indicator kriging* or *disjunctive kriging*. The former is a non-parametric approach to kriging where the data are transformed to a binary variable (indicator). Disjunctive kriging is also based on an indicator approach but allows all of the information of the original variable to be retained. Webster & Oliver (1989) used disjunctive kriging to determine areas where phosphate was required on Broom's Barn Experimental Station, England.

## SAMPLE DESIGN

To describe reliably the variation within a field, the sampling intensity should relate to the scale at which most of the spatial variation occurs. Otherwise the sampling might be more

intensive than necessary or, more seriously, too sparse to provide spatially correlated data for any method of interpolation. The sampling intensity might need to be different for different properties or for different regions. For example, Figure 3 shows that soil *A* varies over shorter distances than soil *B*; the peaks and troughs are closer together. To resolve the variation in soil *A* a smaller sampling interval would be required than that for soil *B*. If the sampling interval for *B* were used for *A*, as illustrated in Figure 3, the variation would appear as noise. However, for soil *B* the sampling intensity used for *A* would result in wasted effort.

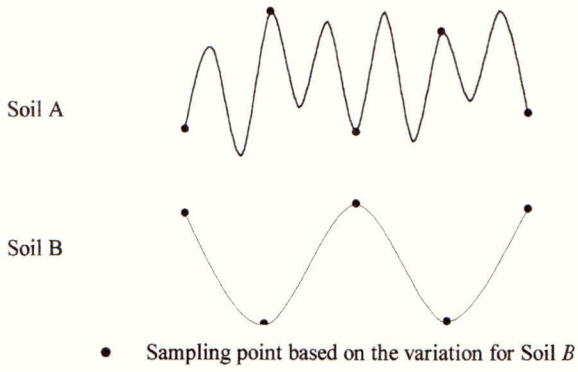


Figure 3. Location of sampling points for two soil types.

If there is no information on the likely scale of variation a nested survey may be a sensible starting point. This allows several magnitudes of spatial scale to be investigated in a single analysis. Price *et al.*, (2001) gives an example of a nested sampling design. If there is some prior knowledge of the scale of variation, transects or grid surveys can be used.

**Case Study**

The study site is an arable crop field at the Centre for Dairy Research (CEDAR), Berkshire, south central England. The soil in the field is predominately a clay loam. Soil samples were taken at the nodes of a 20-m square grid, giving a total of 160 sample points. The samples were analysed for soil organic matter content (OM).

The variogram is based on variances and, therefore, the statistical distribution of the data can affect the reliability of the semivariances. Geostatistics also assumes some degree of stationarity. When local trend or drift, or large scale variation exists in the study area the data may not be stationary. This violates the assumptions of geostatistics. An exploratory data analysis showed that the OM data were normally distributed with no evidence of trend.

The experimental variogram was computed and an exponential model fitted (Figure 4). The variogram shows a clear structure with a small nugget variance, which suggests that the sample design was adequate to identify the variation within the field. To examine the spatial variation of OM, predictions were made at 5 m intervals on a square grid by ordinary block kriging using blocks of 20 m by 20 m, which relate to management units. These values were mapped (Figure 4). The map shows that the variation is patchy, some areas have large values and other areas small values. Such patchy distributions are the transition feature that give rise to bounded variograms; the average extent of the patches generally relates to the range of the variogram.

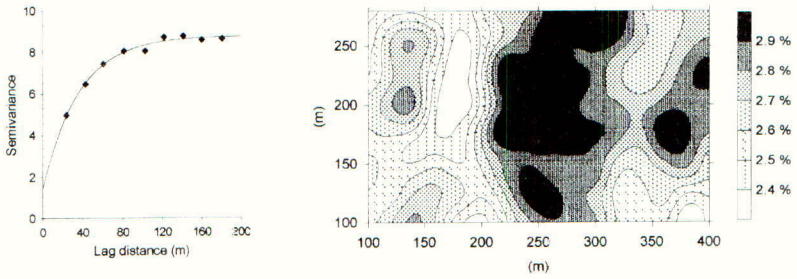


Figure 4. Variogram and map of kriged estimates for organic matter at CEDAR.

Sampling on a 20-m grid is likely to be too costly for most routine surveys. To explore the effect of the sampling interval, the original data on the 20-m grid were sub-sampled to produce data on 40-m (40 points), 60-m (24 points) and 100-m (8 points) grids. These subsets contain fewer than 100 data points, which is generally considered the minimum for producing a reliable variogram (Webster & Oliver, 1992). Variograms were computed to illustrate the effects of few data, and where possible models were fitted (Figure 5). These variograms show how the sampling intensity affects the form of the variogram. Even when a 40 m sampling interval is used there is a change in the structure of the variogram. For the 100-m subset the variogram has only one point and it is not possible to determine whether the data are spatially dependent.

The variogram and data for each subset were used for prediction by ordinary kriging. This could not be done for the 100-m subset and predictions were made using inverse squared distance. The maps of the predictions for OM are shown in Figure 5. They show how the detail in the variation is lost as sampling becomes increasingly sparse. This loss occurs even for data on the 40-m grid, although the main areas of small and large values are still evident. The map for predictions from data on the 60-m grid loses more detail and identifies only the patch of large values in the centre of the survey area. The map of predictions using data on the 100-m grid shows that the OM content is over- and under-estimated in many regions of the field and the original pattern is lost. The maps of kriging variance from data on the 20-m, 40-m and 60-m grids show that the values increase as the sampling interval increases; the predictions become less reliable (Figure 6).

Soil sampling and analysis are costly and time consuming, but unless they provide information that is reliable for the purpose then any effort is wasted. The results from this field illustrate the effect of increasingly sparse sample information on the reliability of the experimental variogram and accuracy of the predicted values. They show that maps made from such predictions can be an unreliable representation of the variation. Depending on the level of detail required, a 40-m or 60-m grid might be adequate for this field. For a different field, however, a different sampling intensity might be required to reflect the spatial scale of variation. Careful planning is required at the outset to ensure that the sample design is suitable for the study site. Some guidance on the spatial variation of soil properties can be gained from ancillary data, data that are cheaper to obtain but which are related to the soil in some way. This may include aerial photographs, electro-magnetic inductance (EMI) surveys or yield maps.

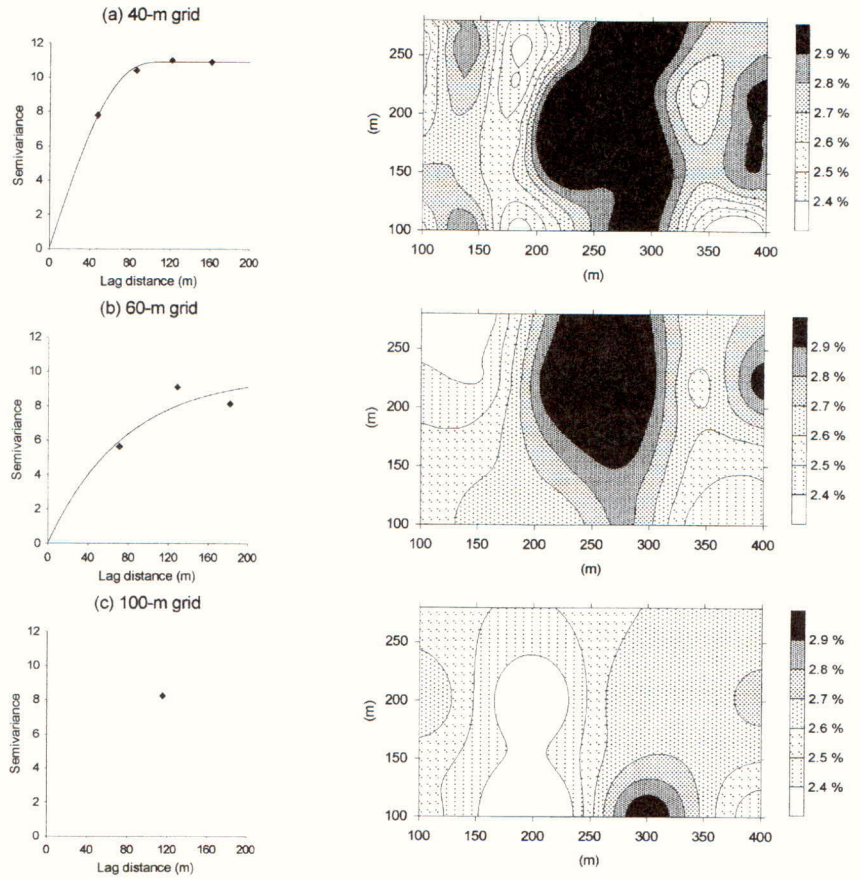


Figure 5. Variograms and maps for organic matter at CEDAR for the sub-sampled data.

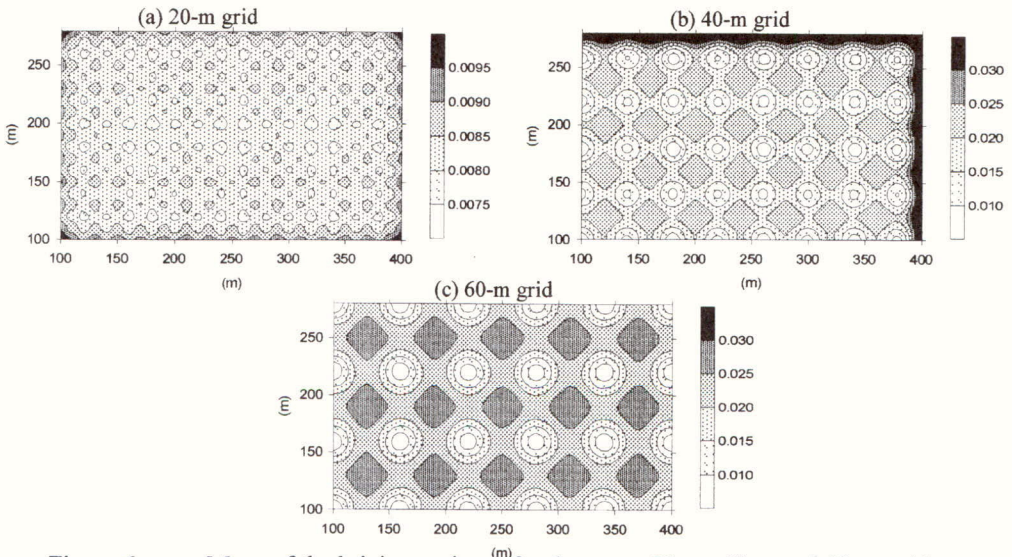


Figure 6. Maps of the kriging variance for data on a 20-m, 40-m and 60-m grid.

## CONCLUSION

Environmental properties, such as soil properties, vary in a complex way. Geostatistics provides a suite of methods that are suitable for analysing the spatial variation of these properties. The variogram describes the structure and spatial scale of the variation and kriging uses this model to predict the values of a property at unsampled places. An important part of any survey is the sample design. To describe the variation reliably the sampling intensity should relate to the scale of spatial variation that needs to be resolved. This might be different for different soil types or different landscapes. The case study illustrates the use of geostatistics and the effect on the predictions when sampling is too sparse.

Once the spatial variation within a study area in known it is possible that this information can be incorporated into pesticide models. This would allow the fate of pesticides to be predicted at unsampled locations across the field. In areas where properties vary considerably this would be a more accurate approach than using a single value for prediction.

## REFERENCES


Burrough P A (1983). Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science*, **34**: 577-597.

Frogbrook Z L; Oliver M A (2001). Comparing the spatial predictions of soil organic matter determined by two laboratory methods. *Soil Use and Management*. In press.

Goovaerts P (1997). *Geostatistics for natural resources evaluation.* Oxford University Press, New York.

Isaaks E H; Srivastava R M (1989). *An introduction to Applied Geostatistics*. Oxford University Press, Oxford.

McBratney A B; Webster R (1983). Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. *Journal of Soil Science*, **34**: 137-162.

Oliver M A (1999). Exploring soil spatial variation geostatistically. In: *Precision Agriculture '99, Proceedings of the 2nd European Conference on Precision Agriculture*, ed J.V. Stafford, pp. 3-17. Sheffield Academic Press, Sheffield.

Oliver M A; Frogbrook Z L (1998). Sampling to estimate soil nutrients for precision agriculture. The International Fertiliser Society. Proceedings No 417.

Oliver M A; Webster R; McGrath S P (1996). Disjunctive kriging for environmental management. *Environmetrics*. **7**: 333-358.

Price O R; Walker A; Wood M; Oliver M A (2001). Using geostatistics to evaluate the spatial patterns in pesticide/soil interactions at the field scale (These proceedings).

Webster R; Oliver M A (1989). Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. *Journal of Soil Science*, **40**: 497-512.

Webster R; Oliver M A (1990). *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press, Oxford.

Webster R; Oliver M A (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, **43**: 177-192.

Webster R; Oliver M A (2001). Geostatistics for Environmental Scientists. J. Wiley & Sons Chichester.