MULTIPLE SOURCES OF SPATIAL VARIATION AND HOW TO DEAL WITH THEM.

P.A. Burrough
Instituut voor Ruimtelijke Onderzoek
University of Utrecht,
Postbox 80.115, 3508 TC Utrecht,
The Netherlands

ABSTRACT

Conventional methods of thematic mapping often assume implicitly that only one major pattern can be recognized at any given scale of mapping. Conventional thematic map representations model spatial units by 'homogeneous' units or polygons representing the various components of the pattern being mapped. Interpolation methods allow gradual variation within spatial units to be mapped but they commonly also ignore the problems that arise from multiscale sources of variation. Observed natural variation may be caused by a number of separate spatial processes operating with various weights (intensities) over a range of scales. This paper reviews some ways in which theoretical multiscale models, complex semivariograms, robust methods and sampling strategies can be applied to the problem of multiple sources of spatial variation.

INTRODUCTION

The search for quick, cheap, simple, reliable and universal ways with which to capture and describe the spatial variation of attributes of the natural environment is a current major research activity. There are many ways to describe and map the spatial variation of soil, vegetation, landform, groundwater or pollution. Some researchers follow the well-worn paths of tried and tested methods while others strike out through thorny, mathematically difficult terrain. In spite of many, local near successes, and many global failures, the search for useful, reliable methods of spatial analysis continues unabated across all disciplines whose object it is to study the spatial variation of attributes of the earth's surface. Considering the costs involved in collecting and analysing spatial data, and the implications for landuse planning decisions of incorporating poor or incorrect data in geographical information systems, it is crucially important for data users to know how spatial data have been modelled, and what the limitations of these models are. One limitation that is frequently overlooked when choosing an interpolation method is the presence of important variation at several scales which may confound or reduce the success of the chosen spatial modelling technique.

Methods for spatial analysis

The two basic approaches to mapping the spatial distribution of any given attribute, or regionalized variable (Matheron 1971) are summarized in Table 1. In the first approach one has total coverage of an area, usually with remotely sensed imagery (aerial photos or digital scanned images) of an attribute or attributes that are thought to be correlated with the required environmental property. In the second approach one samples the property of interest directly at certain locations from which a model of the spatial variation is created by interpolation.

Table 1. Basic approaches to mapping

Whole area approach

- Many observations of cheap, possibly relevant data.
- Divide area into regular units (pixels) or into 'natural' units
- Devise and use hierarchical classification schemes
- Discover relations between attribute values of pixels or class means of 'natural' units and attribute of interest.

Point sampling approach.

- Choose sampling strategy (regular grid, stratified random, etc.)
- Choose and apply interpolation method (global, local, etc.).
- Map isolines

MATHEMATICAL MODELS OF SPATIAL VARIATION

The classificatory, choropleth map model approach relies on the model

$$Z(x) = \mu + \alpha_j + \epsilon \qquad (1)$$

Where $Z(x)$ is the value of attribute Z at point x, μ is the general mean of the area in question, α_j is the deviation between the mean of class j and μ, and ϵ is the residual variation, usually assumed in the first instance to be a normally distributed Gaussian noise function having zero mean and variance σ^2. The weakness of this model is revealed every time an area is remapped at a larger scale, thereby 'discovering' spatial structure in what was previously regarded as spatially unstructured and uncorrelated 'noise'. As this process of remapping at larger and larger scales can continue endlessly, the success of this mapping approach depends greatly on the balance between the different kinds and scales of spatial variation present. The universal nature of this problem is revealed by studies that show that irrespective of map scale, the distribution of boundaries on thematic choropleth maps over a wide range of scales can be modelled satisfactorally by a Poisson distribution

$$P(x) = 1 - \exp(-\lambda x) \qquad (2)$$

or related functions such as the Gamma distribution or the Weibull function (Burgess and Webster 1984, Burrough 1986)

Short-range variation in digital imagery.

The presence of short-range variation in digital imagery is usually considered a nuisance that needs to be removed. If the source of the noise is known, many techniques exist for its removal (e.g. destriping LANDSAT images). If the source is unknown, but local, simple digital filter techniques exist for mechanistic removal of the unwanted noise (c.f. Rosenfeld and Kak 1976). Statistical methods of image analysis, recently reviewed by Ripley (1986) also assume that at the chosen observation scale a clear signal is waiting to be cleaned up (see also Besag 1987).

Methods of interpolation.

In many situations such as in studies of soil fertility or pollution, it is impossible or impractical to obtain a complete overview using surrogate attributes and so the phenomenon of interest must be mapped using samples collected at point locations. The overall distribution of the variation of the phenomenon is then determined by interpolation. Methods of spatial interpolation (c.f. Agterberg 1982, Burrough 1986, Davis 1986, Lam 1983, Ripley 1981) adopt either a global or a local approach. Global methods, such as trend surface analysis, parallel choropleth map models in the sense that they attempt to 'explain' large amounts of spatial variation in terms of single structural units (complex polynomials). Just as with the choropleth map models, the 'noise' usually contains short-range spatially correlated variation. Local methods avoid these problems, but introduce others, such as how best to choose the local weighting function and how to select the most appropriate method of interpolation (e.g. smooth B-splines or moving weighted averages).

Optimal methods of interpolation (kriging).

The set of interpolation techniques collectively known as kriging recognise that spatial variation may be the result of structural, locally random but spatially correlated, and uncorrelated components. Information about these various components is used to compute the weights for local interpolation in such a way as to minimize the variance of the interpolation estimate. The basic model is:

$$Z(x) = m(x) + \epsilon'(x) + \epsilon'' \qquad (3)$$

in which the value of attribute Z at point x is modelled by m(x), a deterministic function describing the 'structural' component of variation, $\epsilon'(x)$ is a function describing the local, spatially correlated variation of Z, and ϵ'' is a random noise term. The essential steps in kriging (Journel and Huijbregts 1978, Webster 1985) are:

1. Sampling to determine the sample semivariogram
2. Fitting an appropriate model to the sample semivariogram
3. Using the semivariogram model to supply appropriate values of the weights with which to obtain estimates of the value of Z at unvisited points x0.

MULTIPLE SCALES OF VARIATION AND KRIGING

Kriging is a practical and a conceptual advance on previous methods of spatial interpolation because it allows 'non-structural' variation to be considered as being comprised of spatially correlated variation and random variation. The critical aspects of kriging, however, are the fundamental assumptions of the method and the choice and fitting of semivariogram models. In both instances, the type and nature of multiscale variation can be critically important.

The fundamental assumptions of kriging are contained in the intrinsic hypothesis of regionalized variable theory which regards spatial variation as the outcome of a random process with certain stationarity conditions. These are:

1. That the expected difference in the value of Z at any two places separated distance h is zero:

$$E[Z(x) - Z(x+h)] = 0 \qquad (4)$$

2. the variance of the differences depends on h and not on x, and is given by:

$$\begin{aligned} \mathrm{var}[Z(x) - Z(x+h)] &= E[\{Z(x) - Z(x+h)\}^2] \\ &= 2\gamma(h) \qquad (5) \end{aligned}$$

Clearly, these assumptions require that the spatial process in question operates over the whole of the area to which consideration is being given.

The semivariogram and semivariogram models.

The semivariogram displays the variation of semivariance with sample spacing, h. It is obtained by sampling and through the intrinsic hypothesis it is estimated by

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \cdot \sum_{i=1}^{n(h)} \{z(x_i) - z(x_i + h)\}^2 \qquad (6)$$

where n(h) is the number of pairs of observations with separation h.

Usually, the weights for interpolation are obtained by fitting a suitable model to the experimentally estimated semivariances. Two major classes of semivariogram model have been recognised: a) the transitive models; b) unbounded models.

Because of the variance of the estimate Z at any point can not be less than zero, the sample semivariogram cannot be modelled by any function that appears to fit the distribution of points. The following ***authorized models*** are recommended for use (McBratney and Webster 1986):

a) transitive models - i.e. models in which the semi-variance appears to reach a constant level (the sill) at a certain sample spacing or range:

linear model with sill	(1D only)
circular model	(1D, 2D)
spherical model	(2D, 3D)
gaussian model	(1D, 2D)
exponential model	(1D, 2D, 3D)

b) unbounded models - i.e. models in which the semivariance continues to increase with sample spacing:

linear model	(1D, 2D, 3D)
logarithmic model	(1D, 2D, 3D)
brownian fractal model	(1D, 2D, 3D)

Multiscale variation.

All transitive models, with the exception of the exponential model, imply that the observed variation has been generated by a spatial process that operates at a definite scale, for example within overlapping blocks that have a definite size or scale. Under these circumstances the spatial model given by equation (3) describes the situation adequately. With the exponential model, and the unbounded models, however, it is implicit that variations can occur over a range of scales. The exponential model suggests that the overlapping blocks vary randomly in size; the unbounded models, particularly the fractal and the logarithmic model, suggest that spatial variation occurs at many scales. A semivariogram that approaches the origin parabolically may signify changing drift (i.e. change in the value of E[Z(x)] with x caused by local or regional trends - i.e. variation at another scale). Changing drift can be handled either by using a full structural analysis and universal kriging as described by Olea (1975), or by using intrinsic random functions of a higher order that the semivariogram to describe the spatial variation (Matheron 1973).

Choosing the correct semivariogram model is critical for kriging, yet little attention seems to have been paid to the physical grounds for choosing any particular model. There are several aspects of the problem. The first is the nature of the variation being studied - is it the result of a single, dominant process or the sum result of several superimposed processes? What kind of spatial distribution results from a given physical process? The second is the problem of sampling variation on the estimated semi-variogram - how much can the form of a semivariogram vary according to the sample of points used? The third is the problem of the choice and fitting of models, and whether that choice should be guided primarily by least-squared fit criteria or by using other criteria.

A simple multiscale model. Instead of considering that observed spatial variation is the result of structural, local randomly correlated and random components as expressed by equation (3), let us now assume that randomly correlated variation can exist at all scales. Mandelbrot's Brownian fractal model (Mandelbrot 1982) is the ideal embodiment of a model in which spatial variation occurs at all scales. The simple Brownian model has several draw- backs in practice, however; it assumes that variation occurs at all scales in a self-similar way, and that the roughness of the variation (the value of the D parameter) is the same at all scales. Consideration of real data suggests otherwise (Armstrong 1986, Burrough 1984). Real spatial processes (omitting special cases such as cloud formation) seem to lead to spatial patterns in which the fractal D value varies with location and with scale (Mark and Aronson 1984).

With this in mind, I developed a one-dimensional nested model of spatial variation that is an extension of equation (3), but within which the scales and the weights of the various components can be set independently (Burrough 1983). The value of Z at point x is now given by

$$Z(x) = \sum_{i=1}^{n} \{ \epsilon'_i(x) \} + \epsilon'' \qquad (7)$$

where the $\epsilon'(x)$ are a set of nested, spatially correlated random functions associated with scale i. As before, the ϵ'' term represents spatially uncorrelated random variation to take account of measurement errors and other essentially random, non-spatial sources of variation.

The model has since been programmed for interactive use as a personal computer 'game' and it allows the user to create one-dimensional displays of multiscale data by setting the ranges and weights of several nested random functions. The semivariogram is displayed together with the function (Figure 1). The computer game has proved invaluable for teaching students and others not familiar with spatial statistics how complex spatial variation can arise from nested random processes, and also for demonstrating the problems associated with under-sampling. The game allows transects from 20 to 600 points to be generated. Generating the same model several times for different transect lengths allows the user to see how an estimate of a semivariogram relies on sufficient samples.

If one can generate a transect from single random processes, it should be possible, in principle, to go the other way and to estimate the scales and weights of the contributing processes from the sample semivariogram. Simple geological transects gave good results (Burrough 1983), with the valuable by-product that the confidence limits and effective degrees of freedom of the fitted model could be calculated (Taylor and Burrough 1986; see also McBratney and Webster 1986). Alas, preliminary results of work with two-dimensional simulations suggest that decomposing multiscale two-dimensional patterns is not so straightforward.

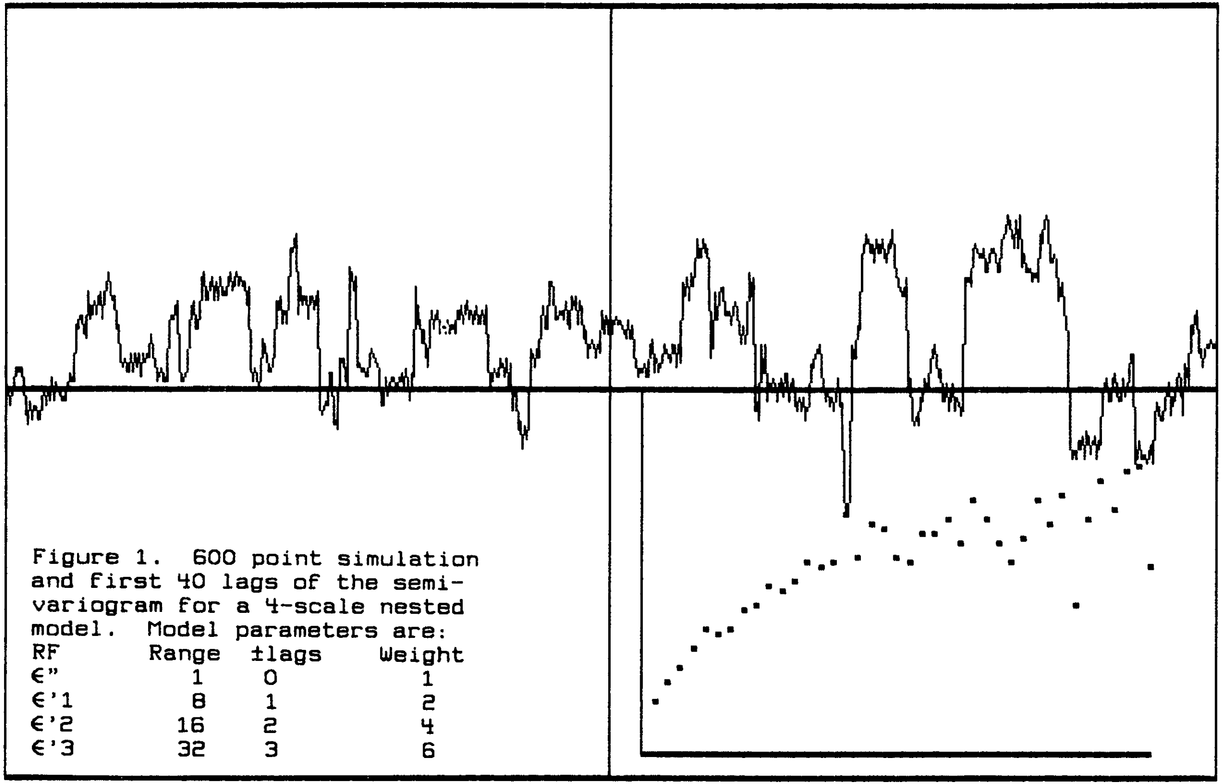

Figure 1. 600 point simulation and first 40 lags of the semi-variogram for a 4-scale nested model. Model parameters are:

RF	Range	±lags	Weight
∈”	1	0	1
∈’1	8	1	2
∈’2	16	2	4
∈’3	32	3	6

Complex multiscale models. The one-dimensional nested model is only authorized for work in one dimension, so the approach must be modified when working in two or more dimensions. An alternative to fitting a single, complex model is to choose several standard authorized models and to combine them to give an overall, complex model. The question then is on what grounds the separate models should be chosen. McBratney and Webster (1986) demonstrate the use of double models for semi-periodic soil variation in Australian gilgai, and for heavy metal concentration in soil in Scotland. In both cases they made use of their knowledge about the physical soil processes to guide their choice of the components of the model. As with all models, the investigator needs to strike a balance between goodness of fit to the data and parsimony. McBratney and Webster (1986) suggest that the choice between a single scale model and a multiscale model (or between two multiscale models) can be estimated by using Akaike's (1973) information criterion which is estimated by

$$\hat{A} = n \ln(R) + 2p \qquad (8)$$

where n is the number of observations, p is the number of estimated parameters and R is the residual sum of squares of the fitted model. The model with lowest $\hat{A}$ is the best. Here I should like to remark that it is possible that the best fitting model may not always make physical sense. For example, if a best-fitting semivariogram model returns an estimate of the nugget variance ϵ" that is considerably less than that known to be possible with the given laboratory technique, the results should be treated with caution.

Robust methods of estimating the semivariogram

When an essentially point process is superimposed upon a continuous process, estimates of the semivariogram obtained by equation (6) may be heavy tailed because the intrinsic hypothesis is locally invalid. McBratney and Webster (op cit.) cite this problem when mapping soil potassium over a cow pasture contaminated with faeces; we have noted similar problems in cracking clay soils in the Sudan and in soil pollution (Rang et al 1987). Cressie and Hawkins (1980) proposed robust methods to deal with the problem of heavy-tailed distributions; McBratney and Webster (op cit.) suggest that the robust methods are of most value when an underlying spatial process needs to be separated from the effects of a contaminating point process.

DISCUSSION AND CONCLUSIONS

Most natural patterns of variation contain contributions from processes operating at various scales. When a particular scale of variation is dominant and obvious, standard mapping techniques will often suffice. When several scales are important, it may be necessary to identify them before proceeding further, using all available knowledge about the processes in question in order to make sensible decisions.

Separation into 'natural' physiographic units may be a wise first move that can ensure that the basic assumptions of a mapping technique hold throughout a single area (e.g. see Burrough 1986). Knowledge of spatial processes and the patterns they are likely to create may also assist when choosing both simple and complex models. The definite choice of complex models and the estimation of relative weights and scales of variation is made difficult by uncertainties in the estimation of semivariograms.

One way to avoid capturing too many levels of spatial variation is by tailoring sample spacing before mapping. There is now considerable evidence (e.g. Oliver and Webster 1986, Webster 1985) that nested methods of sampling can provide useful estimates of the scales of spatial variation present in an area before mapping or sampling for the semivariogram commences.

REFERENCES

Agterberg, F.D. 1982. Recent developments in GeoMathematics. Geo-Processing 2, 1-32.

Akaike, H. 1973, Information theory and an extension of maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds. B.N. Petrov and F. Coaki) pp. 267-281, Akademia Kiado, Budapest.

Armstrong, A.C. 1986, On the fractal dimensions of some transient soil properties. J. Soil Sci. 37, 641-651.

Besag, J. 1987. On the statistical analysis of dirty pictures. J. Royal Statistical Soc. Section B. (in press).

Burgess, T.M. and Webster, R. 1984, Optimal sampling strategies for mapping soil types. I. Distribution of boundary spacings. J. Soil Sci. 32, 643-659.

Burrough, P.A.. 1983, Multi-scale sources of spatial variation in soil. II. A non-Brownian Fractal model and its application to soil survey. J. Soil Sci. 34,599-620.

Burrough, P.A. 1984, The application of Fractal ideas to geophysical phenomena. Bull. Inst. Mathematics and its Applications. 20, 36-42.

Cressie, N. and Hawkins, D.M. 1980. Robust estimation of the variogram. Math. Geology 12, 115-125.

Davis, J.C. 1986. Statistics and Data Analysis in Geology. Wiley (2nd. Edn).

Journel, A.J. and Huijbregts, Ch. J. 1978. Mining Geostatistics. Academic Press.

Lam, N. S., 1983. Spatial interpolation methods: a review. The American Cartographer 10, 129-149.

Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. Freeman, New York.

Mark, D.M. and Aronson, P.B., 1984. Scale-dependent fractal dimensions of topographic surfaces: an empirical investigation with application in geomorphology and computer mapping. *Mathematical Geology*, *16*, 671-83.

Matheron, G., 1971, The theory of regionalized variables and its applications. *Cahiers du Centre de Morphologie Mathêmatique de Fontainebleu,* No. 5, Paris.

Matheron, G. 1973. The intrinsic random functions and their applications. *Adv. Appl. Prob*. *5,* 439-468.

McBratney, A.B. and Webster, R. 1986, Choosing functions for semivariograms of soil properties and fitting them to sample estimates. *J. Soil Sci.* *37*, 617-639.

Olea, R.A. 1975, Optimum mapping techniques using regionalized variable theory. *Series on Spatial Analysis No. 2,* Kansas Geological Survey, Lawrence.

Oliver, M.A. and Webster, R., 1986. Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geographical Analysis* *18*, 227-242.

Rang, M.C., Ockx, J, Hazelhoff, L. and Burrough, P.A. 1987, Geostatistical methods for mapping environmental pollution. Paper presented Int. Symposium on Soil and Groundwater Pollution, Noorwijkerhout, The Netherlands, 30 March-2 April 1987.

Ripley, B. 1981. Spatial Statistics, Wiley, New York.

Ripley, B. 1986. Statistics, Images and Pattern Recognition. *Canadian J. Statistics* *14*(2), 83-111.

Rosenfeld, A. and Kak, A. 1976. Digital Picture Processing. Academic Press, New York.

Taylor, C.C. and Burrough, P.A., 1986. Multiscale sources of spatial variation in soil III. Improved methods for fitting the nested model to one-dimensional semivariograms. *Math. Geology* *18*, 811-821.

Webster, R. 1985. Quantitative Spatial Analysis of Soil in the field. Advances in Soil Science Volume 3, Springer-Verlag New York.