Adaptive Schemes for Geostatistical Sampling and Survey

B.P. Marchant and R.M. Lark

Silsoe Research Institute, Wrest Park, Silsoe, Bedfordshire, MK45 4HS, U.K.

Abstract: A large sampling effort is required to produce an accurate geostatistical map, and the extraction and analysis of each sample is often expensive. The effectiveness of a particular sampling scheme is dependent upon the spatial variability of the quantity being measured. This spatial variability is unknown when sampling commences and therefore conventional, single phase, sampling schemes are often inefficient. Adaptive schemes sample in several phases, the design of later phases incorporating information about the spatial variability derived from earlier phases. The authors compare the precision of variogram estimates from an adaptive scheme with those from a conventional transect method. The adaptive scheme selects sampling points that minimize an approximation to the variogram uncertainty. Although this approximation is biased, it is a suitable relative measure for optimizing the sampling scheme. The precision attained by a 100 point transect scheme is matched by an 80 point adaptive scheme when estimating both long- and short-range variograms from simulated data.

Keywords: *Optimal sampling, variogram uncertainty, spatial simulated annealing, adaptive*

1. INTRODUCTION

1.1. Motivation

Geostatistical analysis is a powerful technique for interpolating the distribution of a spatial variable, such as the concentration of an ore, pollutant or soil nutrient, over a region. It not only yields expected values of the variable across the region but also assigns a level of uncertainty to the estimate at each point (Burgess and Webster, 1980). Thus informed decisions can be made as to whether the extraction of the ore is viable or the pollutant should be cleaned up. A disadvantage of geostatistics is the large sampling effort required (Webster and Oliver, 1992). This can make an accurate geostatistical survey very costly to complete. Therefore methods of improving sampling efficiency, in terms of the number of sampling points required to achieve a certain precision of estimate, are of great value to practitioners in a number of different application areas.

Traditionally, sampling for geostatistical analysis is carried out in a single phase. The positions of the sample points may be chosen based upon the investigator's intuition and experience of sampling similar environments. Often, however, the sampling scheme will be designed with convenience in mind, without considering the specific properties of the variable being measured and the resulting effects upon the usefulness of the sampling scheme.

Standard geostatistical analysis is split into two parts. The values recorded at the sampling locations are first used to determine the structure of the spatial correlation. This information is presented as a variogram which is then used in a process known as kriging to interpolate values of the variable at unsampled locations within the field (Burgess and Webster, 1980).

The effectiveness of a sampling scheme, for both variogram estimation and kriging, is highly dependent upon the actual spatial correlation (Webster and Oliver, 2001, Chapter 5). If the variable is only correlated over a short range then there is little to gain from investigating the level of variation over much longer lag distances. Therefore, the sampling scheme should contain a large proportion of closely spaced pairs in order to derive the nature of the short-range variation. Conversely, widely spaced sampling points are required for a variable with a long range of spatial correlation in order to determine the nature of the variation up to this range. The uncertainty of the kriged estimates of the variable are determined by both the sampling scheme and the spatial variability of the variable (Webster and Oliver, 2001, Chapter 8).

The characteristics of the variogram are not known before sampling commences, so it is not possible to design an efficient sampling scheme at this stage. In this paper we discuss the potential for using phased or adaptive sampling schemes. By dividing the sampling process into distinct phases it is possible to use the information derived from earlier phases to aid the choice of sampling positions in later phases. Thus the sampling scheme can be optimized in a manner that reflects the properties of the variable under investigation.

1.2. Previous studies

The majority of previous studies of iterative procedures for spatial sampling have concentrated upon kriging rather than variogram estimation. McBratney et al. (1981) showed that once the variogram has been estimated it is possible to design the most cost-effective sampling grid for kriging, generating a map that meets precision requirements without oversampling. Van Groenigen et al. (1997) demonstrated the usefulness of a phased sampling scheme. Having calculated an initial kriged estimate of the variable they assessed the uncertainty of the estimate across the region. Further sampling phases concentrated upon the locations where the risk of error was greatest.

Optimal sampling schemes for kriging generally minimize the kriging variance in some manner. This variance is calculated as a by-product of the kriging procedure. Similarly, when the variogram is calculated by a maximum likelihood method, a measure of the uncertainty associated with the variogram estimate is generated. Lark (2002) designed sampling schemes that minimized this uncertainly. The variogram is more commonly calculated by the method of moments. There is not a generally accepted expression for the uncertainty of variograms estimated by this method.

Strategies for optimizing sampling schemes for the method of moments, have primarily been guided by common sense ‘rules of thumb’ rather than direct minimization of a measure of the variogram uncertainty. For example, Warrick and Myers (1987) generated sampling schemes where the distribution of the distances between pairs of sampling points conformed to a prescribed distribution. Thus they ensured that the level of variability of the variable was measured over a range of different lag distances. This method is somewhat unsatisfactory. It does not account for how the ideal distribution of separation distances varies according to the actual variogram, nor does it ensure that the information obtained from different lag pairs is not itself highly correlated.

Recent studies have attempted to identify more rigorous criterion to optimize sampling schemes for variogram estimation (Bogaert and Russo, 1999; Muller and Zimmermann, 1999). However, the expressions used to determine variogram uncertainty in these studies are largely untested.

1.3. Overview of paper

The aim of this paper is to demonstrate the potential for using iterative sampling schemes. As an illustrative example we concentrate upon the estimation of the variogram by the method of moments.

The method of moments procedure is described in Section 2, and Section 3 discusses how a sampling scheme may be optimized for this purpose. Section 4 describes the development of an iterative sampling scheme for use in the field. The scheme is tested upon simulated data in Section 5.

2. ESTIMATING THE VARIOGRAM BY METHOD OF MOMENTS

The variogram expresses the variance of the difference between two observations of the variable as a function of the distance that separates them. It is most commonly estimated by the method of moments. This procedure has two parts. First the average variogram values for different separation distances are calculated directly from the data. This results in the experimental variogram. Then a mathematical function is fitted to this experimental variogram.

2.1. The experimental variogram

The experimental variogram is calculated by allocating the pairs of sampling points to different bins based upon their separation distance. The estimate of the variogram at separation distance h is then given by (Webster and Oliver, 2001, Chapter 5):

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left\{ z_1^i - z_2^i \right\}^2 \qquad (1)$$

where $N(h)$ is the number of pairs separated by distance h and $z_1^i \equiv z(x_1^i)$ is the first value of the ith pair and $z_2^i \equiv z(x_2^i)$ the second value of the ith pair. If the sampling points are irregularly located then the bins may have to be extended to contain pairs separated by h plus or minus some tolerance. The experimental variogram is presented as a plot of $\gamma(h)$ against h.

2.2. Fitting a variogram model

The variogram must be expressed as a mathematical function before being used for kriging. This is typically achieved by fitting a suitable function to the experimental variogram. Authorized functions are described by Webster and Oliver (2001, Chapter 6). Each function is defined in terms of a small number of parameters that are selected to best-fit the function to the experimental variogram. In this study we use the spherical function which is defined by:

$$\gamma(h) = \begin{array}{l} c_0 + c_1\left(\dfrac{3h}{2a} - \dfrac{h^3}{2a^3}\right), \text{for } 0 < h \le a \\ c_0 + c_1, \quad \text{for } h > a \end{array} \qquad (2)$$

This function is expressed in terms of three parameters, namely; a the range of spatial correlation, c_0 the nugget effect and c_1 the sill value.

A number of different methods exist for selecting values for these parameters. Some practitioners do so by eye, plotting variogram functions with different parameter values on top of the experimental variogram and then selecting the best match. Cressie (1985) describes three mathematical techniques for fitting the parameter values. Herein we favour the more rigorous generalized least squares (GLS) for fitting the parameter values. The GLS technique also accounts for the accuracy of each point of the experimental variogram estimate, and the correlation between each point.

3. OPTIMISING SAMPLING SCHEMES FOR VARIOGRAM ESTIMATION

3.1. Choosing a fitness function

In order to optimize a sampling scheme it is necessary to have some fitness function that the optimal scheme minimizes. For example, when designing a sampling scheme for kriging, McBratney et al. (1981) suggested minimizing the kriging variance.

Recently Pardo-Iguzquiza and Dowd (2001) suggested a method by which the uncertainty of GLS variogram parameter estimates may be approximated. For a known underlying variogram function and any particular sampling configuration, this technique yields a variance-covariance matrix of the variogram parameter estimates. These formulae are cumbersome and are not repeated herein, however, for full details of the method see Pardo-Iguzquiza and Dowd (2001).

The Pardo-Iguzquiza and Dowd (2001) method is divided into two parts. First the variance-covariance matrix of the experimental variogram estimates is derived. This derivation assumes that the variable is multivariate normal. These formulae are written in terms of the underlying variogram parameters and the procedure essentially averages the covariances of the variogram estimates from different pairs of points. The variance-covariance matrix of the fitted parameters is then approximated, to leading order, from this matrix.

A suitable fitness function may be expressed in terms of the entries of this matrix. An estimate of the underlying variogram is required to calculate such a fitness function. This estimate can be updated within an iterative scheme. The variogram estimate after phase '*n*' may be used to calculate the fitness function for phase '*n+1*'. The details of one such sampling scheme are described in Section 4.

The exact choice of fitness function is arbitrary. A previous study suggests using the determinant of the matrix but the justification is unclear (Muller and Zimmermann, 1999). Herein we minimize the variance of the estimates of the range parameter a. This choice is based upon the observation that perturbations of the range parameter are the primary influence on the configuration of the optimal sampling points. Therefore, the iterative scheme should operate smoothly if the range parameter estimate quickly converges to its actual value.

3.2. Spatial simulated annealing

Spatial simulated annealing (SSA) is a numerical technique for obtaining the configuration of sample points that minimize the chosen fitness function (van Groenigen, 1999). SSA optimizes at a point level, incorporating sample points from previous phases and honoring sampling constraints such as field boundaries, buildings or lakes. Thus it is very suitable for optimizing iterative sampling schemes.

The algorithm starts with a random array of new sampling points. The value of the fitness function is calculated for this initial configuration combined with any sampling locations from previous sampling phases. Random perturbations are applied to the new points in turn. If such a perturbation reduces the fitness function it is retained. A proportion of the perturbations that increase the fitness function are also retained at random, to avoid convergence to a local minima. The probability of retaining a particular perturbation that increases the fitness function is controlled by a parameter referred to as the temperature. A given increase in the fitness function is more likely when the temperature is large. As the algorithm progresses the temperature and the magnitude of the perturbations are reduced. The perturbation loop is repeated until either a given value of the fitness function is reached or no further reduction in the fitness function is being achieved. Full details of how to implement a SSA algorithm are given by van Groenigen (1999).

4. AN ITERATIVE SAMPLING SCHEME

This section describes the stages within an adaptive sampling scheme. The algorithms have been designed to run on a portable computer in the field, within a practical timescale.

4.1. Survey boundaries and obstacles

The position of the field boundaries and any obstacles where sampling cannot occur (e.g. lakes or buildings) are recorded using a GPS system. The approximate length and width of the field is derived from this data.

4.2. Initial sampling

The initial sampling phase occurs upon six transects, three of which are perpendicular to the other three. These transects are placed randomly within the field. Each consists of ten points and their length is half the length of the field.

4.3. Variogram fitting

The values from the initial phase of sampling are recorded, the experimental variogram is calculated and a variogram model is fitted by GLS.

4.4. Optimizing further sampling phases

Each further sampling phase consists of ten points. These are chosen using the SSA algorithm. The fitness function is the variance of 'a', as predicted by the method of Pardo-Iguzquiza and Dowd (2001). The variogram estimated from the previous sampling phase is used to calculate the fitness function for this phase. The loop of further sampling and variogram fitting is then repeated until the variogram uncertainty reaches a prescribed tolerance level.

5. TRIALS UPON SIMULATED DATA

In theory, the iterative sampling scheme described in Section 4 should yield an optimal sampling point geometry for variogram estimation and the approximate variogram uncertainty. However, little testing has been carried out upon the variogram uncertainty estimates. Therefore the first purpose of the trials upon simulated data is to determine the accuracy of these uncertainty estimates. Then the effectiveness of the iterative sampling scheme is compared to that resulting from adding further transects. A spherical variogram with $c_0 = 0.1$, $c_1 = 0.9$ and a variable range is used for all simulated examples herein.

5.1. Assessing variogram uncertainty

The variogram uncertainty measures were tested upon data sets simulated by LU decomposition (Deutsch and Journel, 1998). This simulation method was chosen because it generates multivariate Gaussian spatial variables. Thus the resulting simulated data sets should satisfy the statistical assumptions made by Pardo-Iguzquiza and Dowd (2001).

Various types of sampling scheme (regular and irregular, optimized and unoptimized) were trialed. In each case 1000 data sets were simulated upon the sampling scheme with a particular spherical variogram function. The experimental variogram and GLS fitted variogram parameters were calculated for each simulation. The variance-covariance matrices for both the experimental variogram and variogram parameter estimates were then compared with those predicted by Pardo-Iguzquiza and Dowd (2001).

The results of these trials were consistent for different underlying variograms and sampling schemes. In each case the variance-covariance matrix of the experimental variogram, resulting from simulated data, agreed almost exactly with the predicted matrix. This is to be expected since the expression for the variance-covariance matrix of the experimental variogram contained no approximations.

Pardo-Iguzquiza and Dowd's (2001) formulae also accurately predicted the variance of both the fitted nugget and sill parameters (and the covariance of the two). In every case however the formulae underestimated the variance of the range estimate, sometimes by up to four times (Figure 1). This error presumably results from the leading order approximation in the derivation of the parameter covariance matrix expression.

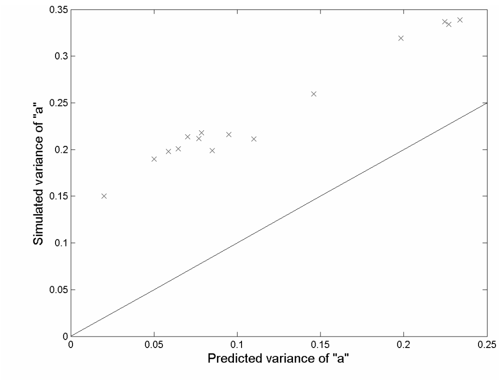

Figure 1. Comparison of predicted and simulated values for the variance of '*a*'. The prediction technique of Pardo-Iguzquiza and Dowd (2001) consistently underestimates the variance but the predicted value is highly correlated to the simulated value. The underlying variogram has *a=1.0*

Although the predicted variance values for the fitted range parameter are underestimated they do correlate with the simulated values. Therefore, they may be treated as a relative measure of variogram uncertainty.

5.2. Testing the iterative sampling scheme

The iterative sampling scheme was also tested upon data sets created by LU simulation (Deutsch and Journel, 1998). Each trial used the field boundaries seen in Figure 2. For each test an underlying variogram model was chosen. This was used to simulate the variable onto the transects of the initial sampling phase. A variogram model was fitted to this data and a further ten sampling points were optimized based upon the fitted parameter values. The simulated values of the first 60 points were retained and the values at the ten new points were generated. Then the loop of variogram fitting and optimizing was repeated.

After each sampling phase the effectiveness of the resulting sampling configuration was tested by generating 1000 realisations of the variable (with the underlying variogram structure) at the sample points. The variance of the resulting parameter estimates was compared to that from an additional transect rather than ten optimized points.

Figure 2 shows the sampling schemes that resulted from underlying variograms with short (*a=5*) and long (*a=18*) range. The sampling points for the short-range process are far more closely packed than for the long-range process. Also, the sample points have been positioned in order to minimize the correlation between the variogram estimates from different pairs.

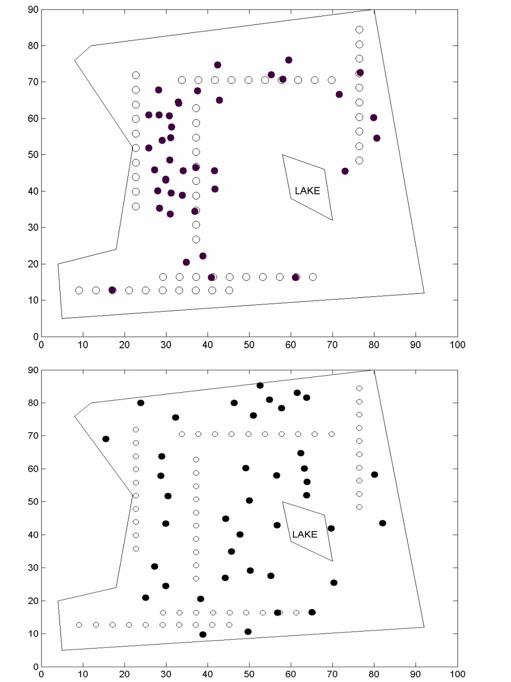

Figure 2. Iterative sampling schemes generated for underlying variograms with *a=5* (top) and *a=18* (bottom). Clear circles are transect points, black dots optimized points. The sampling points for the short-range process are far more closely packed than the long-range process

Figure 3 indicates superior performance of the iterative scheme, over further transects, for both underlying variograms. The precision of 80 point adaptive schemes is seen to match that of 100 points within transects. These findings were repeated for all underlying variograms up to a range of *a=30.* At this range the parameter estimates from the initial phase became unreliable, and the iterative scheme did not converge.

6. CONCLUSIONS

This paper demonstrates that an iterative sampling scheme for variogram estimation can be more efficient than a conventional transect sampling scheme. Therefore, the use of such a scheme can lead to reduced sampling costs and increased accuracy of the estimated variogram. These results have been obtained from simulated data sets. The next stage of this study is to test the iterative sampling scheme in the field.

In the course of designing the iterative scheme, expressions for the uncertainty of variogram parameter estimates were tested. These expressions, suggested by Pardo-Iguzquiza and Dowd (2001), were found to consistently underestimate the uncertainty in estimates of the variogram range. Although the expressions were still adequate for optimization, accurate measures of variogram uncertainty are required so that it may be incorporated into the final kriging variance.

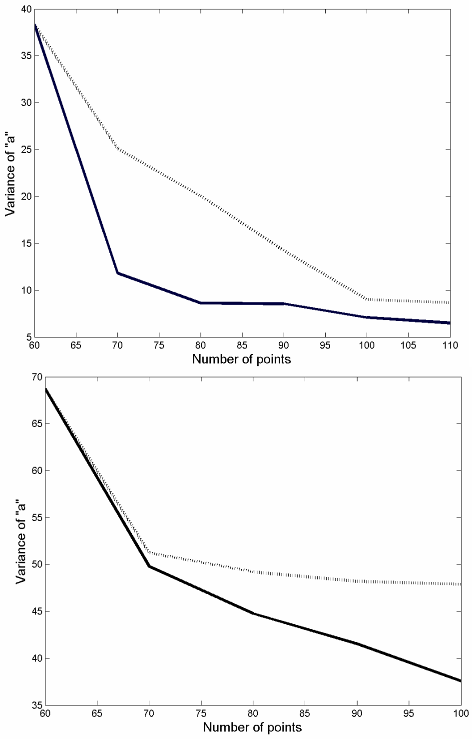

Figure 3. Comparison of the variance of the estimated value of '*a'* for the iterative schemes in Figure 2 (continuous line) and schemes consisting only of transects (dotted line). The underlying variograms have *a=5* (top) and *a=18* (bottom). In each case an adaptive scheme of 80 points is as precise as 100 transect points.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of the Douglas Bomford Trust. This work was carried out with funding from the Biotechnology and Biological Sciences Research Council (Project 204/D15335) and Home-Grown Cereals Authority (Project 2453).

8. REFERENCES

Bogaert P., and D. Russo, Optimal spatial sampling for the estimation of the variogram based on a least squares approach, *Water Resources Research*, 25(4), 1275-1289, 1999.

Burgess, T.M., and R. Webster, Optimal interpolation and isarithmic mapping of soil properties I: The semivariogram and punctual kriging, *Journal of Soil Science*, 31(2), 315-331, 1980.

Cressie, N., Fitting variogram models by weighted least squares, *Mathematical Geology*, 17(5), 563-586, 1985.

Deutsch, C.V., and A.G. Journel, *GSLIB: Geostatistical Software and User's Guide,* OUP*,* 1998.

Lark, R.M., Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood, *Geoderma*, 105(1-2), 49-80, 2002.

McBratney, A.B., R. Webster, and T.M. Burgess, The design of optimal sampling schemes for local estimation and mapping of regionalised variables, *Computers and Geosciences,* 7(4), 331-334, 1981.

Muller, W.G., and D.L Zimmerman, Optimal designs for variogram estimation, *Environmetrics,* 10(1), 23-27, 1999.

Pardo-Iguzquiza, E., and P. Dowd, Variance-covariance matrix of the experimental variogram: assessing variogram uncertainty, *Mathematical Geology*, 33(4), 397-419, 2001.

van Groenigen, J.W., *Constrained optimisation of spatial sampling*, ITC Publications, Enschede, 1999.

Warrick, A.W., and D.E. Myers, Optimisation of sampling locations for variogram calculations, *Water Resources Research*, 23(3), 496-500, 1987.

Webster, R., and M.A. Oliver, Sample adequately to estimate variograms of soil properties, *Journal of Soil Science,* 43(2), 177-192, 1992.

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