WATER RESOURCES RESEARCH, VOL. 23, NO. 3, PAGES 496–500, MARCH 1987

Optimization of Sampling Locations for Variogram Calculations

A. W. WARRICK

Department of Soil and Water Science, The University of Arizona, Tucson

D. E. MYERS

Department of Mathematics, The University of Arizona, Tucson

A method is presented and demonstrated for optimizing the selection of sample locations for variogram estimation. It is assumed that the distribution of distance classes is decided a priori and the problem therefore is to closely approximate the preselected distribution, although the dispersion within individual classes can also be considered. All of the locations may be selected or points added to an existing set of sites or to those chosen on regular patterns. In the examples, the sum of squares characterizing the deviation from the desired distribution of couples is reduced by as much as 2 orders of magnitude between random and optimized points. The calculations may be carried out on a microcomputer. Criteria for what constitutes best estimators for variogram are discussed, but a study of variogram estimators is not the object of this paper.

INTRODUCTION

Geostatistics is applicable to a variety of problems in hydrology and soil science. The variogram is the key function which quantifies the interdependence of sampling locations; i.e., two samples from nearby locations tend to be more alike than two taken from widely separate locations.

One of the first steps in the application of geostatistics is the determination of the variogram. The variogram must, in general, be estimated from the data, and then a model is selected which satisfies the positive definite condition, as well as being compatible with the data in some appropriate sense, such as cross-validation. Whether the sample variogram is used or one of other proposed estimators, it is necessary to generate paired differences. The characteristics of the set of paired differences are crucial to the efficiency of the variogram estimator. In turn, these characteristics are strongly influenced by the sampling plan.

For any sampling exercise, the location of sites is a consideration. There are two obvious considerations for sample site selection: one pertains to the estimation of variogram and the second concerns the use of data for kriging but assumes the variogram has previously been determined. In the latter case, the kriging variance can be used to construct an objective function and that problem has attracted the interest of a number of authors [cf. *Burgess et al.*, 1981; *McBratney and Webster*, 1983]. We shall only consider the problem of variogram estimation and only the problem of site selection, rather than that of the derivation of alternative estimators. Our work extends that of *Russo* [1984], who discussed the optimization of location selection based on homogeneity within classes of a given lag tolerance. *Bresler and Green* [1982] discuss the use of randomly generated sample locations to ensure that the distribution of class numbers be as close to uniform as possible.

The objective of this paper is to develop a method to choose sample locations, optimized with respect to prespecified distributions of couples for the distance classes. A second criterion based on the dispersion of separation distances within each class is given. The scheme may be used either for choosing a complete set of points or to select additional points in order to augment an existing set of points or those of a specified pattern, such as along a transect or on a regular grid. The complete problem of what constitutes the best selection of points is not our objective; however, we will suggest appropriate criteria for that characterization. It is assumed that the ideal distribution is decided a priori. We are simply developing a scheme to meet prespecified constraints.

Paper number 6W4737.
0043-1397/87/006W-4737$05.00

REVIEW OF THEORY

Let $Z(x)$ denote the value of the characteristic or attribute of interest at the point x; x could be a point on a transect or in an area or in a volume. If $Z(x)$ is modeled by a random function satisfying the intrinsic hypothesis then the variogram is given by

$$\gamma(h) = (1/2)\ \mathrm{Var}\ [Z(x+h) - Z(x)]$$
$$= (1/2)E\{[Z(x+h) - Z(x)]^2\} \quad (1)$$

If $x_1, \cdots, x_N$ are N sample locations then the sample variogram is given by

$$\gamma^*(h) = [1/2n(h)] \sum_{i=1}^{n(h)} [Z(x_i + k) - Z(x_i)]^2 \quad (2)$$

where

$$|h| - \varepsilon \le |k| \le |h| + \varepsilon \quad (3)$$

$$\theta_h - \delta \le \theta_R \le \theta_h + \delta \quad (4)$$

with $|h|$ being the length of the vector h; θ_h the direction of h; 2ε the width of the distance class; and 2δ the width of the angle window. Finally, $n(h)$ is the number of pairs $(x_i + k, x_i)$ satisfying the distance and angle conditions.

The $\gamma^*(h)$ is an unbiased estimator of $\gamma(h)$ for each h, but, as is noted by *Cressie and Hawkins* [1980] and *Armstrong and Delfiner* [1980], it is not robust. Since it is not our intention to derive a new estimator but rather to enhance known estimators by appropriate sampling patterns, we will focus on the sample variogram. The results are equally relevant for others that have been proposed.

Davis and Borgman [1978, 1982] obtained a central limit property for $\gamma^*(h)$ assuming fourth-order properties which sug-

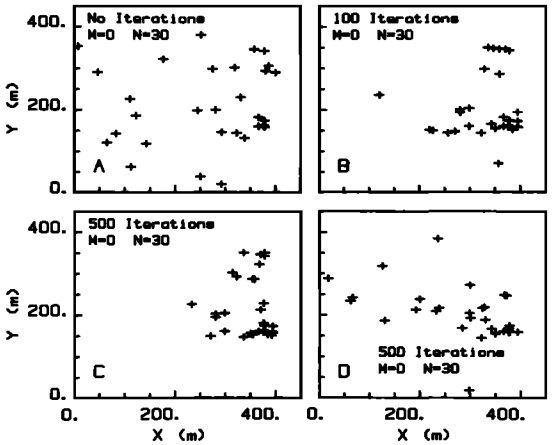

Fig. 1. Distribution of points initially (*a*), after 100 iterations (*b*), and after 500 iterations (*c*) for example 1. Also presented are results for example 3 based on minimum moments about class mean results (*d*) after starting with Figure 1*b*.

gests that $n(h)$ should be as large as possible. The problem is complicated by the fact that $\gamma(h)$ must be estimated for many values of h and not just one; i.e., it is the function that must be estimated and not just the values of $\gamma(h)$ at a finite number of points. Without additional multivariate distributional assumptions for $Z(x)$, statistical inference is not possible. It is known that the behavior of $\gamma(h)$ is most critical for $|h|$ small; hence $n(h)$ should be large for $|h|$ small. In addition to a large number of h's, in general, a number of angle classes must be considered. After these choices of h and of directional classes are made, the site locations must be chosen. One would like to make the selection to optimize one or more desirable characteristics, such as the confidence level and confidence width for estimates of $\gamma(h)$. In general, this is not possible without additional distributional assumptions and it does not guarantee optimal estimation of the variogram function. We will consider then conditions which are clearly necessary or desirable or both.

Suppose then that N locations are to be selected, then there are $N(N-1)/2$ pairs. In general, the number of pairs for short lag distances is small as is the number for large lag distances. The greatest number of couples occurs at approximately half the maximum separation distance. Moreover, there may be excessive dispersion within each distance-angle class, resulting in excessive averaging. Without loss of generality assume distance classes are defined by $0, h_1, \cdots, h_{NC}$, whre NC is the number of classes. For the isotropic case the h_i's are simply distances. For the class h_i, let f_i denote the number of pairs; i.e., $f_i = n(h_i)$. It is immediate that

$$\sum_{i=1}^{NC} f_i = N(N-1)/2 \tag{5}$$

hence we can only apportion the pairs between classes in some optimal way.

We can now identify a number of properties that are necessary or desirable but which, in general, are in conflict, as follows.

1. For each distance-angle class, the number of pairs should be as large as possible, particularly for short distances.
2. The average of the distances in each class should be close to the plotted lag.
3. The variance of the distances in each class should be small.
4. The average of the angles in each class should be close to the plotted angle.
5. The variance of the angles in each class should be small.

It should be noted that a regular grid will tend to ensure conditions 2–5, but N will have to be large to satisfy condition 1. An alternative choice, random selection of the sample locations does not ensure any of the five conditions.

Suppose $f_1^*, \cdots, f_{NC}^*$ is a prespecified distribution and we wish to obtain $f_1, \cdots, f_{NC}$ close to this distribution while satisfying 2–5 to some degree. For example, we might take all $f_i^* = N(N-1)/(2NC)$. With this in mind, we define SS to be minimized:

$$SS = a\sum_{i=1}^{NC} w_i(f_i^* - f_i)^2 + b\sum_{i=1}^{NC} m_{1i} + c\sum_{i=1}^{NC} m_{2i} \tag{6}$$

where $w_1, \cdots, w_{NC}$; a, b, and c are user-selected weighting coefficients; the m_{1i} are absolute or second moments of the distance classes; and the m_{2i} are the absolute or second moments for the angle classes. For the most part, the isotropic case is considered, and c is taken as 0. In addition, we will concentrate on the case $b = 0$.

The case $w_i = 1/a$, all f_i^*'s are the same, and $b = c = 0$ is the same as the criterion given by *Bresler and Green* [1982]. The case where $a = 0$, $b = 1$, and $c = 0$ is very nearly the criterion used by *Russo* [1984].

We can equally well assume that M sites have already been selected; $0 < M < N$ and hence consider this option. Note that no a priori assumptions are made about the variogram type nor about the range of dependence although this information might be used in choosing the h_i's and the desired distribution for the number of pairs. While we have focused on the use of the sample variogram, conditions 2–5 are still relevant for all of the other variogram estimators that have been proposed.

Examples and Calculations

For our first two examples we will choose $a = 4[N(N-1)]^{-2}$, $w_i = 1$, and $b = c = 0$ in (2). A procedure to find a minimum value of SS is as follows.

1. Specify M fixed points at $x_1, x_2, \cdots, x_M$ and the total number of points N.
2. Specify the number of classes of couples NC along with

TABLE 1. Distribution of Couple Separation Distances for Examples 1–3

Class, m	Iterations								
	Example 1			Example 2			Example 3		
	0	100	500	0	100	500	0	100	500
0–20	8	31	40	2	11	17	31	25	25
20–40	12	47	45	6	15	18	47	25	29
40–60	20	32	41	10	13	17	32	30	30
60–80	18	32	42	21	12	19	32	33	46
80–100	27	36	42	17	33	32	36	28	51
100–120	20	41	44	45	61	51	41	30	25
120–140	32	31	44	21	43	31	31	27	28
140–160	23	39	41	42	45	46	39	36	28
160–180	44	41	41	19	23	23	41	35	25
180–200	30	50	42	24	22	26	50	42	25
>200	201	55	13	228	157	155	55	124	123
SS	0.026	0.0038	0.0002	0.037	0.027	0.020	1.0	0.886	0.80
m_{avg}	5.1	4.9	5.2	5.5	5.5	5.4	4.9	4.3	3.9

$N = 30$. In each case, $f_i^* = 43.5$ for every class.

the class limits $h_1, h_2, \cdots, h_{NC}$ and the desired number f_i for each class or the fraction of the total in each class ($f_i/[N(N-1)/2]$).

3. Choose $N - M$ random points giving $x_{M+1}, \cdots, x_N$ (note each x_i defines a point in one, two, or three-dimensions).
4. Calculate SS from (2).
5. Choose a substitute point x^* randomly.
6. Calculate SS^* by (2), where x^* is substituted for any x_i, with $M < i \leq N$.
7. If $SS^* < SS$, then substitute x^* for x_i and set SS equal to SS^*. Then either stop or return to step 5 and search for a still smaller SS.
8. If $SS^* > SS$, then return to step 6 and substitute x^* for another x_i or return to step 5 and choose another trial x^*.

Steps 1 and 2 are problem specifications and 5–8 are iterative. The process is concluded when the f_i no longer change significantly or a specified number of iterations has been made. Whether an absolute minimum is reached is not critical as any reduction in SS results in the actual distribution being closer to the specified. The SS may or may not approach zero. The use of random points could be constrained, if desired, to random points within finite blocks or on a given grid network, for example.

Example 1

Assume a 400 × 400-m field. Assume the smallest reasonable sampling element is 2 × 2 m and $N = 30$ random samples are to be chosen. Assume that the desired class sizes are in 20 m intervals; i.e., the upper limits of the classes are $h_1 = 20$ and $h_2 = 40$ etc. Furthermore, assume 10 classes each containing equal number of couples are sought; i.e., $f_1^* = f_2^* = \cdots = f_{10}^* = (0.1)N(N-1)/2 = 43.5$, and the weights w_i are all set to unity. The initial random points are shown in Figure 1*a* and result in a sum of squares of $SS = 0.026$. Of the total $30(29)/2 = 435$ couples about $\frac{1}{2}$; in fact, 201 are outside of the last specified class limit of 200 m (see Table 1). Only 8 and 12 couples are in each of the 2 smallest classes. After 100 iterations the result is as in Figure 1*b* with $SS = 0.0038$. The points are selected closer together and give a very even distribution of class sizes. After 500 iterations the distribution of points appear visually about the same (Figure 1*c*) but the SS is reduced to only 0.0002. The distribution of couples is very uniform (Table 1) with a maximum of 45, a minimum of 40, and only 13 outside the 200-m limit. If repeated, we would anticipate a similar cluster, but centered at some other spot in the field. In fact, if we wished, we could move the centroid of the samples elsewhere in the field, and the separation classes would be unaffected. The spread of the cluster can be increased by increasing the largest class specified (compare example 4 which follows). The average of the absolute moments m_i about the mean of each class remained at about 5 (but was not used as a fitting criterion). If we repeated the exercise with a smaller maximum specified couple separation, the cluster would be more dense. (The calculation time for the 100 iterations was about 1 hour, 50 min on a Rainbow 100 in GW-BASIC; for the 500 it was nearly 7 hours.)

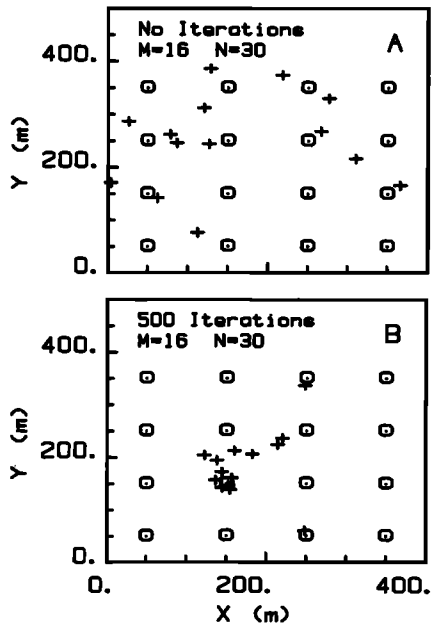

Fig. 2. Distribution of points initially (*a*) and after 500 iterations (*b*) for 14 random overlying a grid of 16 fixed locations (example 2).

Example 2

Assume the above field already has $M = 16$ samples regularly spaced with 1 sample/hectare; repeat as in example 1, with $M = 16$ and choosing $30 - 16 = 14$ new samples locations. The initial sampling pattern is shown as Figure 2*a* with the 16 fixed and 14 random sites. The initial value of SS is 0.037, slightly higher than for the random patterns. With the 16 existing locations, the uniform distribution cannot be met as before because many of the couple separation distances are already fixed at high values. At the end of 100 iterations, the random values are moved toward the center, and $SS = 0.027$. After 500 iterations the results are as in Figure 2*b* showing a cloud of random points chosen around the center, and the SS is reduced to 0.020. The distribution is somewhat sparse for the short distances, where $f_1 = 17$ compared to $f_1^* = 43.5$ etc. (Table 1). The classes which have small frequencies, of course, are balanced by classes with larger frequencies, and the number of couples for distances exceeding 200 m, which is already at 78 for the 16 fixed points alone, is now 155.

Example 3

In this case, we take $a = 0$ in (6) and optimize on the basis of minimum moments within classes of separation distances. The moments chosen are the absolute deviation about each class median, hopefully leading to separation values close to the middle of the class. A constraint was added that the minimum number of couples in any class should be greater than 25 or about one half of that sought in example 1. In order that the constraint be met, the starting point for example 3 was taken to be the results of example 1 after 100 iterations. (Runs were also made with both a and b nonzero, but the specifications of a minimum number of couples accepted in any class, and $a = 0$ is deemed more appropriate.)

The b was chosen such that the initial SS was 1. Of course, the initial average of absolute moments was 4.9, as in example 1. After 100 and 500 iterations, the average class moment was reduced to 4.3 and 3.9, respectively. The reduction from 4.9 to 3.9 is about 20%, which is roughly comparable to reductions in class standard deviations of 16 and 66% for two examples of *Russo* [1984, especially below equation (14*b*)]. Our criterion is more rigorous in that the moment is minimized about the class median and not a fluctuating class mean. Graphically, the results are in Figure 1*d*.

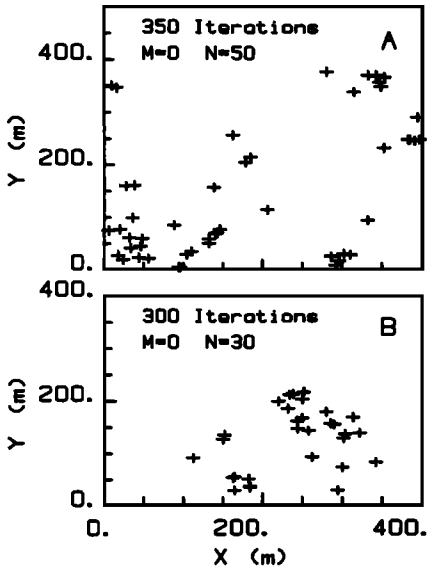

Fig. 3. Distribution of 50 points based on 30 classes out to 450 m separation (*a*) (example 4) and directional classes (*b*) (example 5).

Example 4

We choose 50 points at random, take $a = 1$, $b = 0$, and extend to 30 classes of 0–15, 15–30, $\cdots$, 435–450 more in line with the examples of *Bresler and Green* [1982, cf. their Figure 6]. The initial distribution given in Table 2 is close to their results for $N = 50$. The fitted results after 100 iterations are much closer to an idealized uniform distribution with a reduction in *SS* from 0.010 to 0.0016. Their result was the "best" realization of 27 runs based on a sum of squares similar to (6), with $a = 1$ and $b = 0$. The criterion of equal numbers for large separation classes led to a sorting of points out of the center and toward the edge of the field (see Figure 3*a*). The simulation was cut-off after 350 iterations for which the minimum couples in a class was 33 (see Table 2).

Example 5

As a final example, we choose 30 points at random and choose classes according to direction as well as separation distance. The couples are separated into horizontal and vertical with coarse windows of $\pm 45°$ so as to include all couples. The distance increments are specified as in example 1. The resulting pattern after 100 and 300 iterations are shown as Figure 3*b*. The points initially move toward two small clusters of 8 and 22 points, respectively. The *SS* goes from an original 0.31 to 0.0071 and 0.0049 after 100 and 300 iterations. After 300 iterations, the largest deviation is at the lowest class intervals where 12 and 8 couples were found for the horizontal and vertical compared to just under 22 for exactly 5% of the couples. The maximum couples in any size was 28, and only 5 couples were separated by more than 200 m.

DISCUSSION

Too often data are collected prior to determination of the intended statistical analysis. For some techniques it is sufficient to utilize random sampling with a large sample size. When applying geostatistics, however, it is not sufficient to simply choose a sample size, nor is random sampling sufficient to allow statistical inference on the variogram. The procedure given herein shows that reasonable criterion may be used to choose sample locations and to assure satisfying conditions designed to enhance the reliability of the sample variogram as an estimator while satisfying constraints on sample size. The algorithm may easily be programmed on a personal computer. Previously fixed sampling sites or specific regular patterns can be incorporated in the overall scheme.

TABLE 2. Distribution of Couple Separation Distances for Examples 4 ($N = 50$) and 5 ($N = 30$)

Class, m	Example 4		
	Iterations		
	0	100	350
0–15	1	14	33
15–30	10	36	36
30–45	35	30	37
45–60	35	38	37
60–75	52	42	40
75–90	44	43	43
90–105	48	51	43
105–120	51	46	45
120–135	64	42	41
135–150	57	43	41
150–165	75	45	37
165–180	66	43	37
180–195	68	46	35
195–210	63	46	39
210–225	71	43	46
225–240	56	39	47
240–255	61	51	35
255–270	62	47	36
270–285	59	49	41
285–300	43	43	42
300–315	51	45	37
315–330	37	48	39
330–345	32	47	44
345–360	19	38	45
360–375	17	39	44
375–390	15	33	42
390–405	13	48	36
405–420	3	33	41
420–435	7	20	41
435–450	5	20	40
>450	5	9	25
SS	0.010	0.0016	0.0003
	Example 5		
Class, m	Iterations		
	0	100	300
H, 0–20	1	7	12
V, 0–20	2	2	8
H, 20–40	2	18	22
V, 20–40	5	18	20
H, 40–60	5	25	18
V, 40–60	6	25	24
H, 60–80	7	29	23
V, 60–80	6	29	27
H, 80–100	11	33	22
V, 80–100	14	22	22
H, 100–120	17	28	25
V, 100–120	11	28	22
H, 120–140	12	21	20
V, 120–140	12	26	20
H, 140–160	13	26	26
V, 140–160	18	22	28
H, 160–180	14	20	27
V, 160–180	18	17	27
H, 180–200	18	15	17
V, 180–200	9	14	20
>200	234	15	5
SS	0.31	0.0071	0.0049

H, horizontal; V, vertical.

Whether an absolute minimum of the *SS* function (equation (6)) is found is a moot point, since any reduction more closely meets the desired specifications. The minimization procedure can likely be improved, but for the moment, a more pressing question is what constitutes a best choice of the "*SS*" or an appropriate alternative.

What is the best scheme for locating sampling sites? This problem does not have a simple solution, but some observations are possible. If the only purpose is to satisfy preset variogram couples, then a pattern can be generated to come very close to meeting the specifications, even for large class separation distances. The system will lead to a total set of points concentrated in area which is about equal to the largest class specified (see Figure 1*b*). If the maximum class separation is large, the points will be over most of the field (compare example 4). A combination of a coarse fixed grid and some random points (compare example 2) would seem to provide sufficient uniformity in the distribution of separation distances and at least sparse coverage of the overall field. Thus the data could also be used to interpolate for unsampled sites after modeling the variogram. The success of the directional search (example 5) suggests that patterns optimized for at least two directions can be set up with little extra effort. Debatably, a sound general strategy would be to use a fixed grid for half the points with the other half selected to give half vertical, half horizontal and class sizes specified up to about one half the maximum dimension of the field. Another possibility would be to choose locations in stages, but with sites randomly chosen for each stage, if the desire is to have at least a few sites in all parts of the field. The number of samples necessary can be gauged by dividing the total couples $N(N-1)/2$ by the number of couple classes to observe whether each group contains at least 30 couples or whatever is desired.

Determining the full consequences of the sampling pattern on the estimation of the variogram is an important problem, but beyond the scope of this study. What we have done here is to illustrate that we can meet reasonable specifications of separation groups.

Acknowledgments. Support was provided in part by Western Regional Research Project W-155. Technical Paper 4161, Arizona Agricultural Experiment Station, Tucson.

REFERENCES

Armstrong, M., and P. Delfiner, Towards a more robust variogram, *Rep. N-671*, Cent. de Geostat., Fontainebleau, France, 1980.

Bresler, E., and R. E. Green, Soil parameters and sampling scheme for characterizing soil hydraulic properties of a watershed, *Tech. Rep. 148*, 42 pp., Water Resour. Res. Cent., Univ. of Hawaii, Honolulu, 1982.

Burgess, T. M., R. Webster, and A. B. McBratney, Optimal interpolation and isarithmic mapping of soil properties, 6, Sampling strategy, *J. Soil Sci., 31*, 643–659, 1981.

Cressie, N., and D. Hawkins, Robust estimation of the variogram, *J. Int. Assoc. Math. Geol., 12*, 115–125, 1980.

Davis, B., and L. Borgman, Some exact sampling distributions for variogram estimates, *J. Int. Assoc. Math. Geol., 11*, 643–645, 1978.

Davis, B., and L. Borgman, A note on the asymptotic distribution of the sample variogram, *J. Int. Assoc. Math. Geol. 14*, 189–193, 1982.

McBratney, A. B., and R. Webster, How many observations are needed for regional estimations of soil properties?, *Soil Sci., 135*, 177–183, 1983.

Russo, D., Design of an optimal sampling network for estimating the variogram, *Soil Sci. Soc. Am. J., 48*, 708–716, 1984.

D. E. Myers, Department of Mathematics, Mathematics Building 89, The University of Arizona, Tucson, AZ 85721.

A. W. Warrick, Department of Soil and Water Science, The University of Arizona, 429 Shantz Building 38, Tucson, AZ 85721.

(Received October 27, 1986;
accepted December 10, 1986.)