Geostatistics Troia '92, A. Soares, cd., Kluwer Academic Publishers, Dordreeht, pp. 613-624

CONDITIONAL SIMULATION: PRACTICAL APPLICATION FOR SAMPLING DESIGN OPTIMIZATION

EVAN J. ENGLUND
U. S. Environmental Protection Agency, Environmental Monitoring Systems Laboratory, Las Vegas, Nevada, 89193, U.S.A.

NASER HERAVI
University of Nevada - Las Vegas, Harry Reid Center for Environmental Studies, 4505 S. Maryland Parkway, Las Vegas, Nevada, 89154, U.S.A.

ABSTRACT

Detailed spatial models generated by conditional simulation provide a powerful tool for case-specific optimization of sampling designs. The entire process of sampling, estimation, and decision can be simulated on such a model by a Monte-Carlo approach. Optimization can be based on economic functions or on decision quality constraints rather than simple minimization of estimation variance. Efficient algorithms and 32-bit desktop computers make simulation feasible for routine use. A design solution based on conditional simulation will approach the true optimum only to the degree that the simulations accurately reflect the relevant real-world characteristics. The quality of a simulation depends on a number of factors including the number of conditioning data, the accuracy of the variogram model, and the use of data transformations. The method is illustrated with two examples - one based on the well-known Walker Lake model and the other on an actual case study involving remediation of contaminated soils. For practical purposes, the method appears to be accurate, precise, and robust.

INTRODUCTION

Many authors have applied geostatistics to the problem of spatial sampling design, including Barnes (1989), Burgess et.al. (1981), and Olea (1975). Most such applications involve some variation on minimization of the kriging variance to determine ideal sampling patterns, to identify the best locations for one or more proposed additional samples, or to find the minimum number of samples needed to attain a specified maximum' level of error. This approach, however, does not incorporate economics (Srivastava, 1987), nor does it readily permit evaluation of the consequences of decision-making with uncertainty. Rendu (1980) presented a parametric approach to computation of opportunity costs resulting from estimation errors for the normal and log-normal cases.

This paper describes an alternative approach using conditional simulation (Journel, 1974) and presents two examples which illustrate and attempt to validate the method. Although

computer-intensive, the conditional simulation approach can deal with true optimization when an economic objective function is available. A wealth of intermediate results permit the investigator to better understand the consequences of the available choices.

THE ENVIRONMENTAL REMEDIATION SCENARIO

Both examples illustrate how conditional simulation might be applied in a remediation situation. The problem involves sampling contaminated soils to select portions of the site with concentrations in excess of a regulatory standard which require remediation. We assume: an initial data set from a representative sub-area of the site; a specified action level applied to a remediation unit (RU) of a specified size appropriate to the remediation method (e.g., removal by front-end loader); and an economic objective function where total cost equals the sum of sampling cost plus remediation costs for all selected RU'S, plus the cost of residual contamination in all non-selected RU'S.

The decision rule is that if the estimated value of an RU exceeds the action level, it will be remediated. Unit RU remediation cost is constant; unit RU non-remediation cost is proportional to concentration; and the two unit costs are defined to be equal at the action level. The objective is to estimate the number of samples in a single-phase campaign which would result in the lowest total project cost. This defines the optimal sampling density for the remainder of the site. We will use the following notation:

- N_o The optimum number of samples, where the expected value of the total project cost is a minimum.
- N_e An estimate of N_o from the conditional simulation design procedure.
- n_c The number of initial samples used to condition a simulation (step 2, below).
- n_s The number of simulated samples taken in one iteration of the design procedure (step 4, below).
- n The number of samples to be taken in an actual field sampling campaign.

THE SAMPLING DESIGN PROCEDURE

The procedure is a Monte Carlo resampling scheme which simulates the remediation operation, including data collection, interpolation, and selection.

1) Estimate the variogram model from the initial (conditioning) data set.

2) Generate a detailed simulated site model which is consistent with both the conditioning data and the variogram model. The site model is a dense array of possible sample values.

3) Compute "true" RU values from the site model. Each RU value is the mean of all simulated values within it.

4) Input a value for n_s (selected by trial and error). Generate n_s sample locations by dividing the site model into n_s equal cells, and selecting one random location within each cell. Draw the corresponding simulated sample values from the site model.

5) Estimate mean RU concentrations by block kriging using the variogram model estimated in step 1.

6) Apply the decision rule to the estimated RU values. Evaluate the decisions by comparison to the RU values from step 3. Compute total project cost and other relevant statistics such as numbers of false positives and false negatives, the total quantity of contaminant remediated, etc.

7) Repeat steps 4 -6, drawing alternate random sets of **ns** samples. The number of repetitions necessary to obtain an adequate estimate of the distribution and expected value of total project cost for **ns** samples will depend on the variability of the results.

8) Repeat steps 4 -7 for additional values of **ns,** until a cost curve is obtained from which **Ne** can be determined.

This procedure is not subtle; it relies heavily on brute-force computing. Fortunately, the current generation of desktop computers is capable of running this software. Most of the examples in this paper were run on a 486/50 PC. Typical run times for an entire design sequence, steps 1 - 8 above, were under two hours.

Conditional Simulation

The conditional simulation procedure in step 2 generates a spatial array of simulated values consistent with (conditioned by) known values at measured locations, and with a given variogram model. In this paper, the sequential gaussian simulation (SGSIM) algorithm, (Deutsch and Journel, 1992) is used. This algorithm simulates nodes on a grid in random sequence by first estimating the value at the selected node by kriging with a local neighborhood of conditioning data, and then adding a random component from a normal distribution with zero mean and the kriging standard deviation. Once simulated, values are added to the conditioning set for use in simulating additional nodes.

THE QUALITY OF A SAMPLING DESIGN

How good is the **Ne** obtained by this procedure? If the procedure is valid and practical, it should be unbiased, precise and robust. That is, the mean **Ne** should be close to **No, Ne** should converge to **No** as n_c increases, and **Ne** should be insensitive to small changes in conditioning data, variogram model, or random seed.

CASE STUDY 1: WALKER LAKE

The model

The Walker Lake model used here has been described previously (Englund, 1990, Isaaks and Srivastava, 1989). It contains 19800 values on a 110x18O grid (Figure 1). The data,

arbitrarily assigned units of parts per million (ppm), are a reasonable surrogate for contaminated soils. They are highly skewed, discontinuous, and contain a spike of zero values analogous to non-detects. The log histogram is shown in Figure 2 (with zero values set equal to 1/2 the smallest non-zero value) and the exhaustive variogram in Figure 3.

The decision rule and costs

A secondary grid of 198 RU means (Figure 4) is derived from the original model. The decision rule is to remediate an RU when its estimated value exceeds 100 ppm. The cost of remediating any RU is $10,000; the cost of not remediating any RU is proportional to the actual concentration of contaminant in the RU, whether or not it exceeds 100 ppm. The unit sample cost is $500.

The optimum No

No was obtained by running the sampling design procedure on the Walker Lake model. Several initial values for **ns** were selected, and the model was resampled 1250 times for each **ns.** The process was repeated with additional values of **ns** to "zero in" on the minimum cost. The results are shown in Figure 5; No is approximately 256. This value is itself an estimate subject to some uncertainty. Note that the curve is quite flat in the range 200 - 350. This is helpful in practice, because small errors in estimating No will have only minor impact on total cost. The variability among the 1250 individual cost outcomes for each **ns** emphasizes that a correct estimate of No only guarantees the best *expected* value. The actual outcome from any particular sample of size **n = No** might be quite different.

The simulation Procedure

The SGSIM algorithm was used with

Figure 1. Shaded map of the Walker Lake model. Darker shades represent higher values.

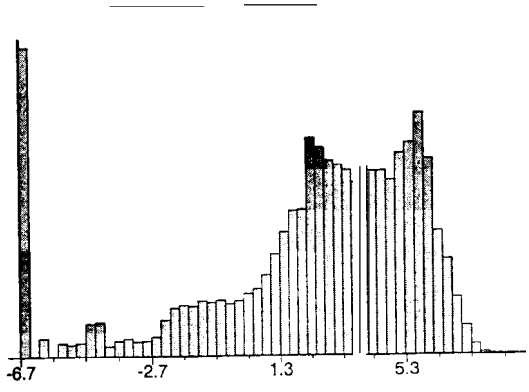

Figure 2. Histogram of log-transformed Walker Lake model.

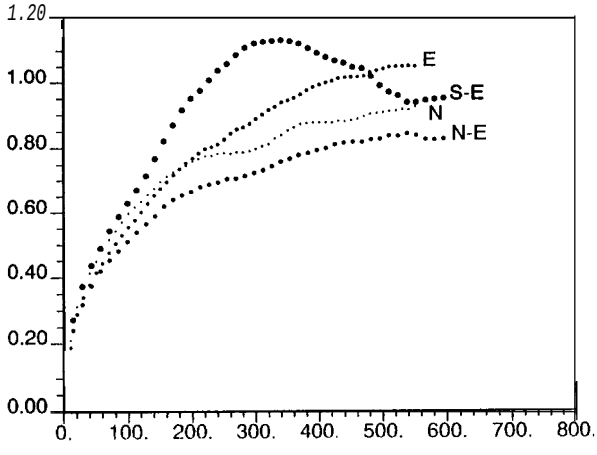

Figure 3. Directional normal-score variogram of the Walker Lake model.

conditioning data sets drawn from the Walker Lake model. Simulations were performed on normal-score transformed data, and back-transformed. The maximum possible simulated value was set at 10,000 (the actual Walker Lake Maximum is 3304). The "linear" option for the upper tail model was used, which results in a uniform distribution between the maximum conditioning value and the maximum simulated value. Variogram models were estimated manually from the transformed conditioning data.

Accuracy and precision

The accuracy and precision of N_e was evaluated on the Walker Lake model with the SGSIM algorithm by repeating the sampling design process for each of three values of n_c (50, 250, and 1250). Conditioning data sets were drawn independently from the Walker Lake model. Figure 6 shows N_o and the resulting N_e values plotted against n_c. For any specified n_c, there are four sources of variability in this method: A.

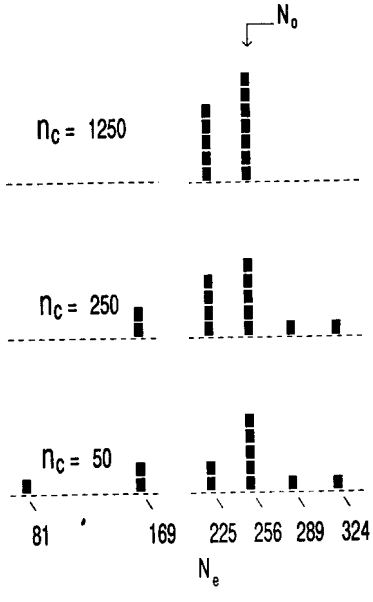

Figure 6. N_e vs. number of conditioning data for several simulations. N_o is approximately 256.

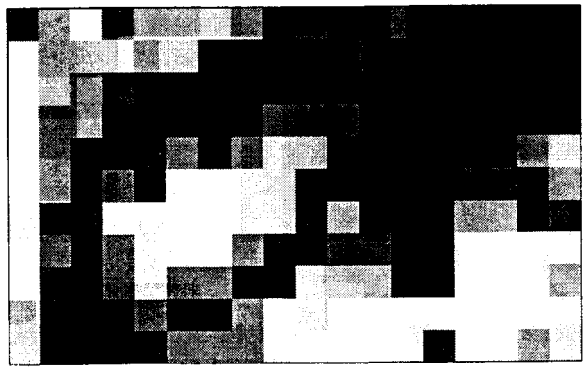

Figure 4. Shaded map of block means from the Walker Lake model.

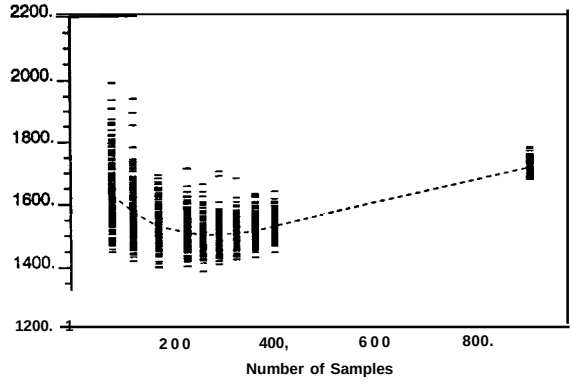

Figure 5. Cost vs. n_s for the Walker Lake model. The dotted line connects the means. N_o is approximately 256 samples.

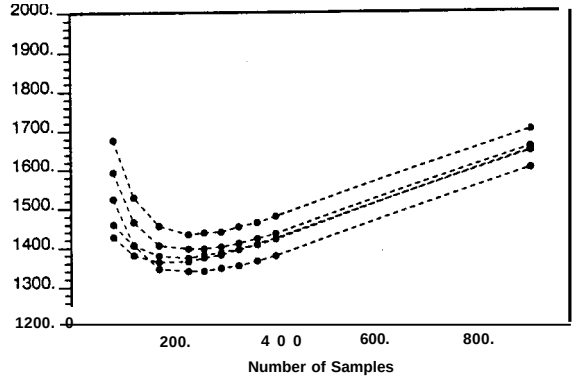

Figure 7. Cost curves from simulations based on a single set of 50 conditioning data, differing only in initial seed value.

variability among conditioning data sets; B. variability among estimated variogram models; C. variability among simulations due to different random seeds; and D. measurement variability in the sampling design process due to different random seeds and the number of iterations. Results for n_c =50 are summarized in Table 1, Case l, and will be discussed in more detail below. Possible sources of bias in this example are the simulation algorithm, the data transformation, and the method of estimating variogram models. However, the results in Figure 6 do not appear to show any bias of practical significance.

Table 1. Summary sampling design results from simulations with n_c = 50.

Case	Sources of Variance	N_e Mean	N_e Variance
1	A,B,C,D	230	4126
2	C,D	221	668
3	A,C,D	320	1990

Sensitivity to random seeds

Each sampling design run has involved resampling a single site model, i.e., a single realization of the simulation algorithm generated by a single seed value for the random number generator. Would the results be more valid if a separate seed value and corresponding site model were used for each sampling draw? This question was addressed by generating several different site models conditioned by a single set of 50 sample data and the corresponding variogram model. Sampling design results from these site models are shown in Figure 7, and in Table 1, Case 2. Although cost curves for individual site models are higher or lower, reflecting different simulated distributions of contaminant, N_e is relatively constant. Resampling a single site model thus appears to be appropriate for estimating N_o (but note that this would not be true if the objective were to estimate absolute costs associated with N_o). Multiple simulations would add the variability among the curves to that already encountered about a single curve. This would require more iterations to obtain a solution, without necessarily improving accuracy. Nevertheless, any particular seed might lead to an extreme result - repeating the procedure on several different site models would be prudent.

Sensitivity to the conditioning data and the variogram model

The sampling design procedure was rerun for seven of the data sets with n_c = 50, values, using only a single variogram model. The results (Table 1, Case 3) incorporate all sources of variability except the variogram model. By subtraction, the variance components A and B are estimated to be 1322 and 2136, respectively. Although these variance estimates are based on only a few observations, they suggest that the variogram model component, B, is the most significant. Figure 8 illustrates the problem involved in estimating the variogram with sparse data, and reproducing it in conditional simulations. Figure 9 indicates that the method is somewhat more successful at reproducing the frequency distributions.

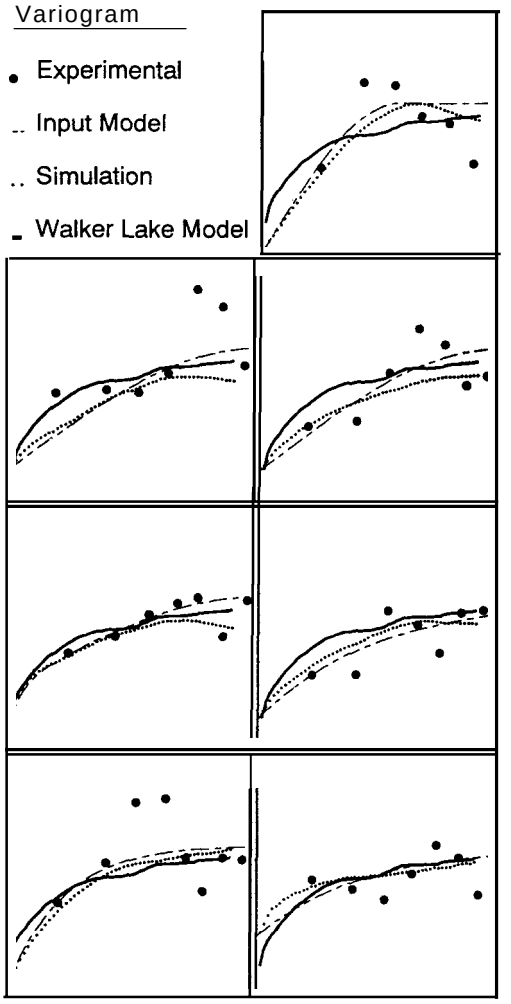

Figure 8. Variogram from seven simulations based on $n_c = 50$, compared with the corresponding experimental variogram, fitted models, and the exhaustive Walker Lake variogram.

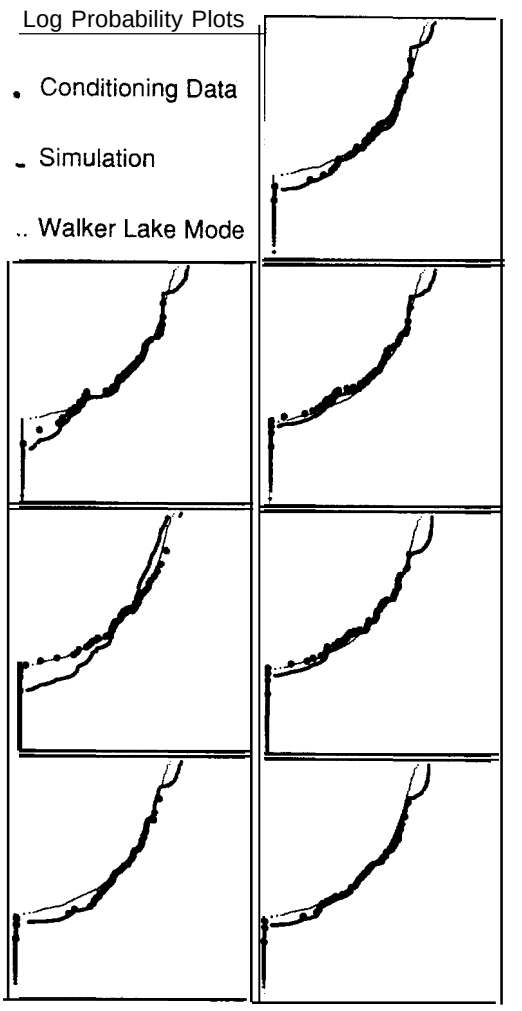

Figure 9. Cumulative probability plots for the seven simulations in Figure 8, the corresponding conditioning data, and the Walker Lake Model. The X-axis is ln(concentration).

CASE STUDY 2: PIAZZA ROAD

At the Piazza Road site in Missouri, dioxin-contaminated oil was sprayed as a dust suppressant on gravel roads. Soil in adjacent pasture land was also contaminated by drainage from the roads. The United States Environmental Protection Agency established a risk-based action level of 1 part per billion (ppb) over an exposure unit (EU) of 5000 sq. ft. (465 m^2) to a depth of 2 in. (5 cm). Soil from EU's estimated to exceed 1 ppb was to be removed and stored for future treatment (RU = EU). High remediation costs ($91,000 per EU) and relatively low sampling costs ($156 per sample) suggested that a more selective removal (RU < EU) might be cost-effective. In an extensive pilot study

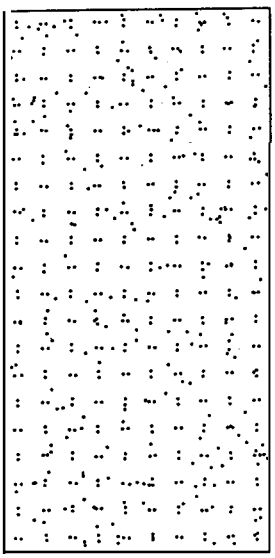

Figure 10. Map of 600 Piazza Road data locations.

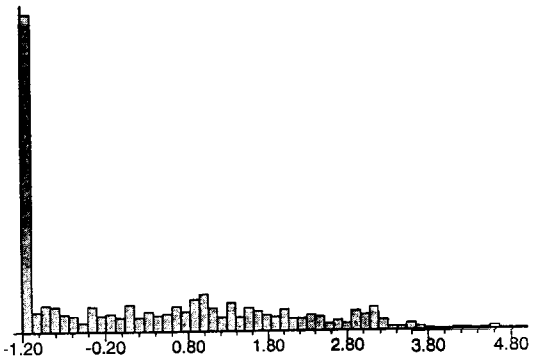

Figure 11. Log histogram of Piazza Road data.

on a representative portion of the site, 600 locations were sampled and analyzed over an area of four EU's to provide information on sampling, analytical, and spatial variability. Ryti, et al (1991) used this information to estimate the RU size and corresponding sampling grid which would minimize total sampling plus remediation costs, subject to pre-defined error limits. The RU was constrained to a minimum practical size of 10 x 10 feet. They coupled the size of the sampling grid to the RU, and recommended a 14 x 14 foot unit size (equivalent to N_e = 100).

The site remediation has since been completed, and the revised approach reduced total costs more than $5 million, after allowing for the $430,000 cost of the pilot study.

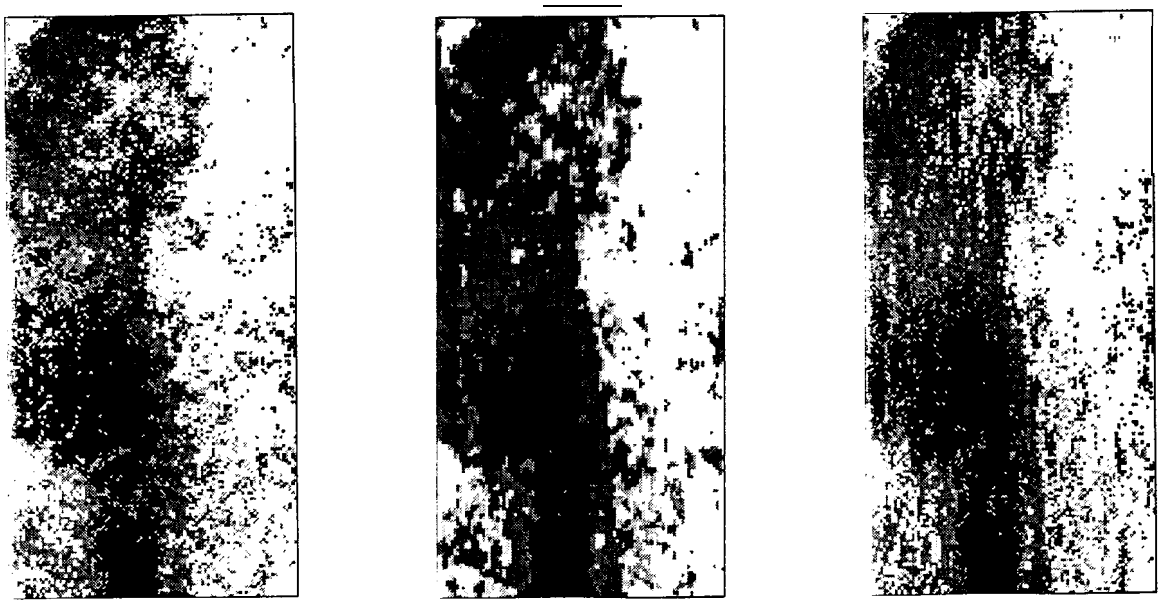

Figure 12. Three Piazza Road simulations based on 600 data.

Although successful in this case, the time and expense involved in such large pilot studies may be prohibitive in many cases. In this example, we will evaluate whether conditional simulations based on a much smaller pilot study could have provided useful results.

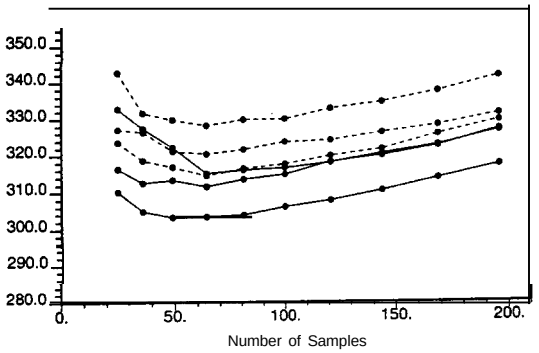

Figure 13. Cost optimization curves from Piazza Road simulations. Solid lines: n_c = 600; dashed lines: n_c = 60.

Optimization based on the full pilot data set

The 600 data locations are shown in Figure 10, and the histogram in Figure 11. The spike at 0.3 ppb represents non-detects arbitrarily set equal to the detection limit. The earlier results suggest that the main source of variability in the sampling design procedure is the variogram model. To evaluate the level of uncertainty, we simulated the pilot area, conditioned by the 600 data, with each of three alternative variogram models which represented a range of possible interpretations of the experimental variogram. Each of the three simulations (Figure 12) was used in a sampling design run, with the sampling and remediation costs listed above. The three optimization curves (solid lines, Figure 13) have nearly identical N_e values of 64, 64, and 49 samples.

Simulation and optimization with small subsets of the pilot data

The sampling design exercise was repeated with three independent subsets of 60 n_c data drawn at random from the 600 pilot data. Each variogram was estimated from the subset only. Differences among the three simulations (Figure 14) illustrate the increased

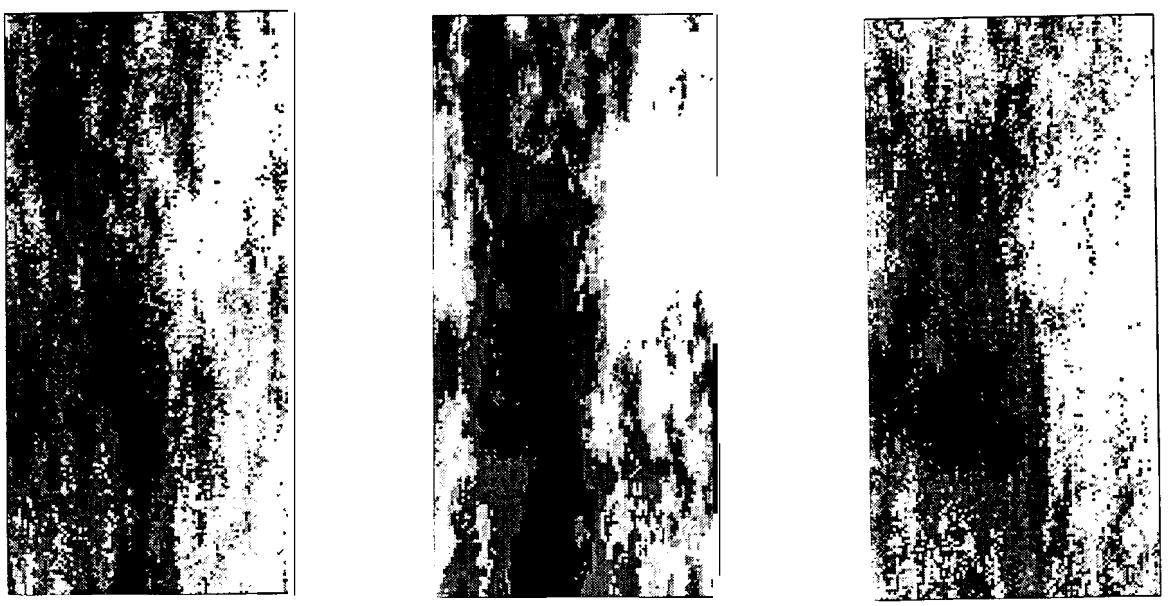

Figure 14. Three Piazza Road simulations based on subsets of 60 data.

variability resulting from fewer data. Nevertheless, the three corresponding cost optimization curves (dashed lines, Figure 13) provide results nearly identical to those obtained with all 600 data, with all three Ne values equal to 64 samples.

Evaluation of sampling alternatives

Suppose an alternative sampling approach such as a portable analyzer could provide data at one-fifth the cost ($31), but with an incremental loss of precision of 25'% (relative standard error) compared to the existing method. Would it be advantageous to change methods? If so, how many samples should be taken?. Figure 15 compares sampling designs for the two cases run on the same site model. The upper curve was run as before; the lower curve was run with lower sample cost and with a random error component added to each simulated sample (error from the existing method is already included in the site models via the variogram). Note that more complex, less favorable

Table 2. Comparison of Sampling Alternatives

Point (From Figure 15)			
Unit sample cost	$156	$31	$31
Number of samples		121	196
Total sampling cost	$9,984	$3,751	$6,076
Remediation cost	$288,789	$284,020	$283,811
Non-remediation cost	$51,268	$51,496	$49,688
Total cost	$350,040	$339,267	$339,555
% False positives	17.1	14.0	13.2
% False negatives	7.4	7.0	6.3
% of total contaminant remediated	97.7	97.6	97.7

error distributions could be used if appropriate. In spite of the added sampling error, the lower sample cost option has lower total costs, with the minimum occurring between 121 and 196 samples. The curve was so flat in this range that 200 iterations were insufficient to resolve the exact minimum. Much of the power of this procedure comes from the large amount of detail that can be obtained in addition to the cost curves. This is illustrated in Table 2 with data about the three points labeled on Figure 15. The reduced total cost from the lower sampling cost alternatives results from a combination of a lower total sampling cost and a lower remediation cost due to reduction of the false positive rate. Although fewer blocks are remediated, there is no significant change in the total amount of contaminant remediated. Evaluation of alternatives is not limited to sampling options. It would be just as easy, for example, to examine the trade-offs associated with changing to a larger remediation unit to obtain economies of scale in remediation.

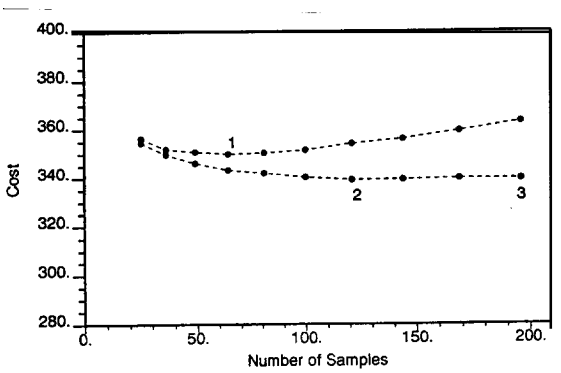

Figure 15. Cost optimization curves comparing two alternate sampling/analytical methods.

These detailed results can help to

evaluate sampling alternatives even when it is not possible to quantify the costs associated with not remediating contaminated material. Examining the relationships among sampling cost, remediation cost, and decision quality from results such as those in Table 2 can effectively quantify the consequences of design choices.

A NOTE ON CONDITIONAL SIMULATION ALGORITHMS AND DATA TRANSFORMATIONS

The objective of this paper was to determine whether resampling a conditional simulation site model has potential as a sampling design tool, particularly in cases where conditioning data are sparse, and the population histogram and variogram are not well known. The SGSIM algorithm used in this study was selected because of its simplicity and efficiency, not because of any perceived theoretical superiority. Other conditional simulation algorithms, such as turning bands (Journel, 1974), LU decomposition (Davis, 1987), and frequency domain (Easley et al, 1991) could be expected to produce similar results. The indicator simulation algorithm (Alabert, 1987) might deal more easily with truncated data, but it is relatively cumbersome. Evaluation of alternative simulation methods is a topic for further investigation.

The cost model used here is particularly sensitive to the upper tail of the simulated distributions. If the simulated values are much too high, they will tend to exaggerate the cost of false negatives and over-estimate $\mathbf{N_c}$[1]; if the values are too low, they will under-estimate $\mathbf{N_e}$. Because simulated normal scores frequently extend beyond the range of the conditioning data, a model defining the upper tail must be subjectively chosen in order to back-transform the simulation. This choice cannot be avoided - it can be made implicitly as in a log transform, or explicitly, as required by the normal-score transform. Either way, it is a potential source of operator-induced bias in the procedure.

CONCLUSIONS

Monte-Carlo resampling of conditionally simulated site models is a powerful tool for the design of sampling programs. In the cases presented here, which involved less than ideal circumstances, it provided estimates of optimal sampling density that are sufficiently precise, unbiased and robust to be of considerable practical value. The fact that this method can be successfully performed on small computers with relatively small sets of conditioning data makes it potentially routinely applicable in remediation situations, even at small sites.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the work of Allen R. Sparks, who developed and assembled most of the software system used in this investigation.

[1]This was the case with the simulations in our original manuscript which were based on log-transformed data. The change to normal-score transforms was made at the suggestion of one of the reviewers.

REFERENCES

Alabert, F.G. (1987), Stochastic imaging of spatial distributions using hard and soft information, MSC thesis, Stanford University.

Barnes, R. (1989) “Sampling design for geologic site characterization”, in M. Armstrong (cd.), Geostatistics, Kluwer Academic Publishers, Dordrecht, pp. 809-822.

Burgess, T., Webster, R., and McBratney, A. (1981) “Optimal interpolation and isarithmic mapping of soil properties. IV Sampling strategy”, Journal of Soil Science 32, 643-659.

Davis, M. (1987) “Production of conditional simulations via the LU triangular decomposition of the covariance matrix”, Mathematical Geology 19, 91-98.

Deutsch, C., and Journel, A. (1992, in press) GSLIB: Geostatistical Software Library, oxford University Press, New York.

Easley, D., Bergman, L., and Weber, D. (1991) “Monitoring well placement using conditional simulation of hydraulic head”, Mathematical Geology 23, 1059-1080.

Englund, E. (1992) “A variance of geostatisticians”, Mathematical Geology 22, 417-455.

Isaaks, E. and Srivastava, R. M. (1989) An Introduction to Applied Geostatistics, Oxford University Press, New York.

.Journel, A. (1974), “Geostatistics for conditional simulation of ore bodies”, Economic Geology 69, 673-687.

Olea, R. (1975) Optimum Mapping Techniques Using Regionalized Variable Theory, Kansas Geological Survey, Lawrence.

Rendu, J.M. (1980) “Optimization of sampling policies: a geostatistical approach”, Fourth Joint Meeting MMIJ-AIME 1980, Tokyo.

Ryti, R., Neptune, D. and Groskinsky, B. (1991) “Superfund soil cleanup”, Environmental Testing and Analysis, Jan/Feb.

Srivastava, R.M. (1987) “Minimum variance or maximum profitability?”, CIM Bulletin, 80:63-68.

NOTICE

Although the research described in this article has been funded in part by the United States Environmental Protection Agency through Cooperative Agreement CR818526 to the Harry Reid Center for Environmental Studies, University of Nevada - Las Vegas, it has not been subjected to Agency review. Therefore it does not necessarily reflect the views of the Agency.