Robert L. Johnson[1]

A Bayesian/Geostatistical Approach to the Design of Adaptive Sampling Programs

REFERENCE: R. L. Johnson, **"A Bayesian/Geostatistical Approach to the Design of Adaptive Sampling Programs"**, Geostatistics for Environmental and Geotechnical Applications, ASTM STP 1283, R. Mohan Srivastava, Shahrokh Rouhani, Marc V. Cromer, A. Ivan Johnson, Ed., American Society for Testing and Materials, Philadelphia, 1996.

ABSTRACT: Traditional approaches to the delineation of subsurface contamination extent are costly and time consuming. Recent advances in field screening technologies present the possibility for adaptive sampling programs---programs that adapt or change to reflect sample results generated in the field. A coupled Bayesian/geostatistical methodology can be used to guide adaptive sampling programs. A Bayesian approach quantitatively combines "soft" information regarding contaminant location with "hard" sampling results. Soft information can include historical information, non-intrusive geophysical survey data, preliminary transport modeling results, past experience with similar sites, etc. Soft information is used to build an initial conceptual image of where contamination is likely to be. As samples are collected and analyzed, indicator kriging is used to update the initial conceptual image. New sampling locations are selected to minimize the uncertainty associated with contaminant extent. An example is provided that illustrates the methodology.

KEYWORDS: adaptive sampling program, indicator kriging, Bayesian analysis, site characterization, sampling strategy

INTRODUCTION

Characterizing the nature and extent of contamination at hazardous waste sites is an expensive and time-consuming process that typically involves successive sampling programs. The total cost per sample can be prohibitive when sampling program mobilization costs, drilling or bore hole expenses, and sample analysis costs

[1] Staff engineer, Environmental Assessment Division, Argonne National Laboratory, Bldg. 900, 9700 S. Cass Ave., Argonne, IL 60439.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

are all included. For example, the Department of Energy (DOE) estimates that it will spend between $15 and $45 billion dollars for analytical services alone over the next 30 years to support environmental restoration activities at its facilities (DOE 1992).

One of the primary products of a site characterization study is an estimate of the extent of contamination. Traditional characterization methodologies rely on pre-planned sampling grids, off-site sample analyses, and multiple sampling programs to determine contamination extent. Adaptive sampling programs present the potential for substantial savings in the time and cost associated with characterizing the extent of contamination. Adaptive sampling programs rely on recent advances in field analytical methods (FAMs) to generate real-time information on the extent and level of contamination (McDonald et al. 1994). Adaptive sampling programs result in more cost effective characterizations by reducing the analytical costs per sample collected, by limiting the number of samples collected by strategically locating samples in response to field data, and finally by bringing characterization to closure in the course of one sampling program. Adaptive sampling programs can result in characterization cost savings on the order of 50% to 80% (Johnson, 1993).

Supporting adaptive sampling programs requires the ability to estimate the extent of contamination based on available information, to measure the uncertainty associated with those estimates, to determine the reduction of uncertainty one might expect from collecting additional samples, and to direct sample collection so that sample locations maximize information gained. Two key characteristics of contaminated sites must be taken into account. The first is that spatial autocorrelation is often present when samples are collected. The second is that there may be abundant "soft" information regarding the location and extent of contamination, even if little "hard" sample data are initially available. Soft data refers to information such as historical records, non-intrusive geophysical survey results, preliminary fate and transport modeling results, aerial photographs, past experience with similar sites, etc.

A number of geostatistical approaches to the design of sampling programs for characterizing hazardous waste sites have been proposed in the past. Early methods focused on minimizing some form of kriging variance (e.g., Olea 1984 and Rouhani 1985). More recent work has centered on stochastic conditional simulation techniques, Bayesian implementations of geostatistics and more complex decision rules (for example, Englund and Heravi 1992; McLaughlin et al. 1993; James and Gorelick 1994). In practice, site characterization sampling program designs tend to blend rigid sampling grids with selective sampling based on best engineering judgement. Typically there is little quantitative analysis to support the final sampling program design.

A combined Bayesian/geostatistical methodology is well suited to quantitative adaptive sampling program support. Bayesian analysis allows the quantitative integration of soft information with hard data. Geostatistical analysis provides a means for interpolating results from locations where hard data exists, to areas where it does not. This combined approach differs from other methodologies in the way uncertainty is handled, and the implementation of a distributed approach to Bayesian updating.

METHODOLOGY

Classical statistics estimates the most likely value for π, the probability of encountering contamination, by using hard sample data results. For example, if 20 random locations were sampled at a site and 5 of these samples returned contamination levels above an action threshold, then an unbiased estimator of the true probability of observing contamination above that threshold for any random location at the site would be the number of hits divided by the number of samples, or 0.25. In classical statistics one could carry the analysis one step further and develop confidence intervals around this estimator with some basic assumptions about the underlying probability distribution. Kriging provides similar results for individual points in space, accommodating spatial autocorrelation as well. Neither classical statistics nor geostatistics provide a means of quantitatively accommodating soft information in the analysis. For the design of sampling programs to characterize contamination extent and subsequent analysis of sampling program results, soft information often plays a crucial role.

A Bayesian approach differs from classical statistics by assuming that parameters (such as the presence of contamination at a node) are unknown initially, but have some known probability distribution called the prior probability density function (pdf). As additional information becomes available (such as results from new sampling locations), these prior pdfs can be updated quantitatively using Bayes' rule to produce posterior probability density functions:

$$P(X|Y) = P(X)P(Y|X) \tag{1}$$

P(X|Y) is the posterior pdf for X, P(X) is the prior pdf for X, and P(Y|X) reflects the probability distribution associated with observing what was observed given the prior distribution of X.

From a Bayesian perspective, a two parameter beta distribution Be(α,β) is a conjugate prior in the context of Bernoulli trials and the binomial distribution (Lee 1989). Be(α,β) ranges between zero and one, and can assume a variety of shapes depending on the values of α and β. For a random variable π that follows a beta distribution, the expected value of π is given by:

$$E(\pi) = \frac{\alpha}{(\alpha + \beta)} \tag{2}$$

where:

α,β = parameters associated with the beta pdf for π.

The variance of π is given by:

$$Var(\pi) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \qquad (3)$$

Binomial distributions provide the probability of observing a specified number of successes within a specified number of trials. Conjugate priors are priors that retain their same underlying pdf after the application of Bayes rule. In the case of a binomial trial with an unknown underlying probability π of seeing a success in any given trial, if X successes are obtained in N trials, a prior for π of the form Be(α,β) becomes the posterior Be(α+X,β+N-X). N functions as the total amount of additional information supplied to the prior. As N grows large, E(π) approaches the classical maximum likelihood estimator for π, X/N, and the Var(π) decreases monotonically.

When one considers only the presence or absence of contamination above some threshold, environmental sampling resembles a binomial trial---N samples collected, X of which encounter contamination above the threshold. The primary difference is that environmental samples are not independent, as required in a traditional binomial sampling sequence. Sample values, even at an indicator level, are spatially autocorrelated. The issue is how to update a prior beta distribution at a given point in space with results from samples nearby that is consistent with the derivation of beta distributions as conjugate priors for binomial distributions and that recognized their spatial autocorrelation.

Two pieces of information are required from the set of samples: N^*_{x0}, the total amount of information represented by the set of samples appropriate for that point in space, x_0, and p^*_{x0}, the expected probability of encountering contamination at x_0 based on the samples' results. Indicator kriging provides an means for deriving these two pieces of information. An unbiased estimator of p^* at x_0 is given by:

$$p^*_{x_0} = \sum_{i=1}^{N} w_i Z(x_i) \qquad (4)$$

where

x_i = locations where samples have been collected;
$Z(x_i)$ = 0 or 1, depending on whether the sample at x_i encountered contamination below or above the threshold;
w_i = kriging weights.

The set of kriging weights, **w**, can be derived by solving the following set of simultaneous linear equations:

$$\sum_{i=1}^{N} C_{ij} w_i + w_{N+1} = D_{j0} \qquad for\ j = 1,....,N \tag{5}$$

$$\sum_{i=1}^{N} w_i = 1 \tag{6}$$

where

C_{ij} = covariance between sample locations x_i and x_j;
D_{j0} = covariance between sample location x_j and the point where the interpolation is taking place, x_0.

N^* at x_0 can be tied to N, the number of samples taken, through the following relationship:

$$N^*_{x_0} = \frac{2C_{00}}{Var_{estim}} - 1 \tag{7}$$

$$Var_{estim} = C_{00} - (\sum_{i=1}^{N} w_i D_{i0} + \mu) \tag{8}$$

where

Var_{estim} = the estimation variance associated with the interpolation of p^* at location x_0;
C_{00} = the variance of the indicator values;
μ = the average of the indicator values for the sample locations involved in the updating.

Equation (7) is heuristically based. When the sampled locations are all "distant" from the point of interest (i.e., greater than the spatial autocorrelation range), N^* goes to zero, implying that the sampled locations contribute no information at the point of interest. As a sampled location comes close to the point of interest, N^* goes to infinity, indicating that the sample information has specified the probability at the point of interest exactly.

The methodology begins by defining a uniform grid over the region of interest. Grid nodes are designated as Decision Points (DPs). At each DP, a pdf based on the two parameter beta distribution $Be(\alpha,\beta)$ is defined. The beta pdf associated with each DP describes the probability of encountering contamination above a pre-selected threshold level at that DP. Initial values for α and β are selected to represent a synthesis of any soft information available for a site, using equations (2) and (3). In the unlikely case where no information is available, a "non-informative" prior can be selected that sets α and β equal to one.

Updating the set of decision points with hard sampling data requires knowledge of the variogram or covariance function for the site. Because the values of p and N at x are independent of C, the primary covariance function parameters of concern are its shape, or functional form, and its range. If sufficient hard data exist, one can estimate the covariance function from an experimental variogram analysis.

A simple measure of the uncertainty associated with contamination extent is to categorize decision points as either "clean", "contaminated", or state uncertain at a given certainty level, where the probability of contamination being present at any given decision point is based on equation (2) using the posterior beta pdf parameters that are associated with that decision point. For example, if one wishes to be 90% certain that the classification is correct when a decision point is classified as either clean or contaminated, then decision points with $E(\pi)$ ranging between 0.1 and 0.9 would be classified as state uncertain. This definition of uncertainty parallels the use of uncertainty by the EPA in its Data Quality Objectives approach to decision-making.

This method for handling uncertainty also leads naturally to measures of benefit one might expect from additional data collection. For example, one might wish to sample those locations that would be expected to maximize the number of decision points classified as "contaminated" at a given certainty level, or as "clean", or to minimize the number classified as state uncertain.

EXAMPLE APPLICATION

A simple example illustrates this methodology in action. Figure 1 provides a plan view of a hypothetical site with surface soil contamination. The site contains a waste lagoon that was breached during a storm. The owner's property is bounded by two secondary roads. The area shaded in grey indicates where surface soil contamination actually exists (7 940 m^2)---an area unknown to the site owner. The owner acknowledges that contamination exists, and that portions of the site will require remedial action. The purpose of the characterization effort is to determine the extent of contamination so that the soils can be removed and treated off-site.

The responsible regulator wants all contaminated soils identified and removed. The regulator wants to ensure that the sampling program is designed so that soils that are contaminated are not erroneously classified as clean. The owner will have to pay for the characterization, excavation and remediation of all soils believed to be contaminated. The owner wants to avoid remediating soils that are actually clean, and also to minimize his characterization costs. After negotiations, the regulator agrees to tolerate a 20% chance that a soil volume identified as clean is actually contaminated. The owner will be responsible for removing and remediating all areas that have greater than 20% chance of contamination being present.

There is no initial hard sampling data for this site. The available soft information includes the location of the lagoon, scattered survey points from which a terrain model can be built to indicate the probable direction of overland flow and hence contaminant migration, the location of a utility building on site that would have been a barrier to flow, and the location of roads with embankments that would have also blocked flow. This soft information is used to construct the initial conceptual image of where contamination likely is, and where it likely is not.

A grid is superimposed over the site that consists of 625 decision points (Figure 2). At each decision point, a beta distribution is defined, with parameters selected to reflect the soft information. For decision points that are in the building, α is set equal to zero and β to a very large number to reflect the fact that the interior of the building is known clean. For decision points within the lagoon, α is set equal to a very large number and β equal to zero, to reflect that fact that the lagoon is known to be contaminated. For the balance of the decision points, α and β are set to values less than 0.5, with their relative sizes selected so that equation (2) reflects the initial probability of the presence of contamination.

Figure 3 shows the grey-scale representation of the initial conceptual model once the beta distribution parameter values have been selected, along with a set of terrain contours based on the available survey points. As is obvious from Figure 3, the initial conceptual image is faithful to the location of the lagoon, building, and land surface contours. Based on this initial conceptual model, without any sampling, the owner would have to clean up 34 440 m^2 of soil, more than four times what is actually contaminated.

Before the adaptive sampling program can begin, the methodology requires a covariance function. At the outset there is no hard data upon which to base a covariance function choice. If the covariance function were selected to honor the initial conceptualization, a range of approximately 200 meters would used. The larger the assumed range, however, the fewer the samples that would be required to characterize the site. As a conservative start, for this example an isotropic exponential covariance function is assumed with range 50 meters.

A traditional sampling program for a site such as this would probably rely on a preplanned, regular sampling grid. As a point of comparison for the subsequent adaptive sampling examples, Figure 4 shows an example preplanned sampling program based on a triangular grid pattern. The gray-shaded surface contained in Figure 4 shows the results when the a non-informative initial conceptual model is updated with the information that would have been derived from this sampling program. The underlying beta distribution parameters for each decision point were set to $\alpha = \beta = 0.1$. In this scenario, the 14 samples result in classifying 23 230 m^2 of soil as requiring remedial action. This captures 87% of the soils actually contaminated, and includes 16 230 m^2 of uncontaminated soil.

If one uses the initial conceptual model shown in Figure 3, and updates it with the results from the sampling program shown in Figure 4, one obtains a different interpretation of the site. Figure 5 shows the results graphically. Using an initial conceptual model that reflects what is known at the outset about the site results in classifying 22 000 m^2 of soil as requiring remedial action. This captures more than 98% of the soil actually contaminated, and includes 14 190 m^2 of uncontaminated soil.

If one incorporates the underlying soft information available for the site, as displayed in Figure 3, and then selects 14 sampling locations that maximize the area that would be classified as clean at the 80% certainty level, then one obtains the preplanned sampling program shown in Figure 6. The sampling locations were selected sequentially, with the selection of the next location conditioned on the expected sampling results from the already selected locations. Figure 6 also shows the results from updating the underlying conceptual model with the results that would actually have been obtained from this preplanned sampling program. These 14 samples result in classifying 15 120 m^2 of soil as requiring remedial action. This captures more than 97% of the soils that is actually contaminated, and includes 7 395 m^2 of uncontaminated soil. Each sample reclassified, on average, 1 380 m^2 of soil as clean.

An adaptive sampling program at this site, driven by the objective of maximizing the area classified as clean at the 80% certainty level, would initially follow the same course as the preplanned program shown in Figure 6. The reason is that for the fourteen samples collected as part of the preplanned sampling program, all encountered what was expected---no contamination. In the case of an adaptive sampling program, however, one has the option of continuing sampling until the goals of the program have been met. Figure 7 shows the locations of an additional 14 samples for this site, along with the results from updating the underlying conceptual model with their results. The additional 14 samples reduced the area classified as requiring remedial action to 10 070 m^2. This included 96% of the soil actually contaminated, and 2 460 m^2 of uncontaminated soils. Each sample reclassified, on average, 350 square meters of soil, a significantly smaller amount than obtained from the first 14 sampled. There are two reasons for this: first, there is simply less area available for reclassification to clean. The second is that the sampling has begun to encounter the unexpected---contaminated soil.

CONCLUSIONS

Adaptive sampling programs provide the opportunity for significant cost savings during the characterization of a hazardous waste site. The challenge for adaptive sampling programs is providing real-time sampling program support that both incorporates the typically significant amounts of soft information available, and that accounts for the spatial autocorrelation that is omnipresent. A joint Bayesian analysis/indicator geostastistical method can be used to guide the selection of sampling locations, to estimate the extent of contamination based on available data, and to determine the expected benefits to be gained from additional sampling.

The example provided illustrates how the addition of soft information to the design of a sampling program can result in a more directed sampling strategy. When the ability to guide the program while in the field is added, the potential for cost savings is great.

ACKNOWLEDGEMENTS

The work presented in this paper was funded through the Mixed Waste Landfill Integrated Demonstration, funded by the Office of Technology Development, Office of Environmental Restoration and Waste Management, U.S. Department of Energy through contract W-31-109-ENG-38.

REFERENCES

Department of Energy, Analytical Services Program Five-Year Plan, Laboratory Management Division, Office of Environmental Restoration and Waste Management, Washington, D.C., January 29, 1992.

Englund, E.J. and N. Heravi, "Conditional Simulation: Practical Application for Sampling Design Optimization", Geostatistics Troia '92, A. Soares, ed., Kluwer Academic Publishers, Dordrecht, 1992, pp. 631-624.

James, B. R. and S. M. Gorelick, "When Enough is Enough: The Worth of Monitoring Data in Aquifer Remediation Design", Water Resources Research, Vol. 30, No. 12, December, 1994, pp.3499-3514.

Johnson, R. L., Adaptive Sampling Strategy Support for the Unlined Chromic Acid Pit, Chemical Waste Landfill, Sandia National Laboratories, Albuquerque, New Mexico, Argonne National Laboratory ANL/EAD/TM-2, Argonne National Laboratory, Argonne, IL, November, 1993.

Lee, P. M., Bayesian Statistics: An Introduction, Oxford University Press, New York, NY, 1989.

McDonald, W. C., M. D. Erickson, B. M. Abraham, and A. R. Robbat, "Developments and Applications of Field Mass Spectrometers", Environmental Science & Technology, Vol. 28, No. 7, 1994, pp. 336-343.

McLaughlin, L. B., L. B. Reid, S.-G. Li, and J. Hyman, "A Stochastic Method for Characterizing Ground-Water Contamination", Ground Water, Vol. 31, No. 2, 1993, pp. 237-249.

Olea, R. A., "Sampling Design Optimization for Spatial Functions", Mathematical Geology, Vol. 16, No. 4., 1984, pp. 369-392.

Rouhani, S., "Variance Reduction Analysis", Water Resources Research, Vol. 21, No. 6, June, 1985, pp. 837-846.

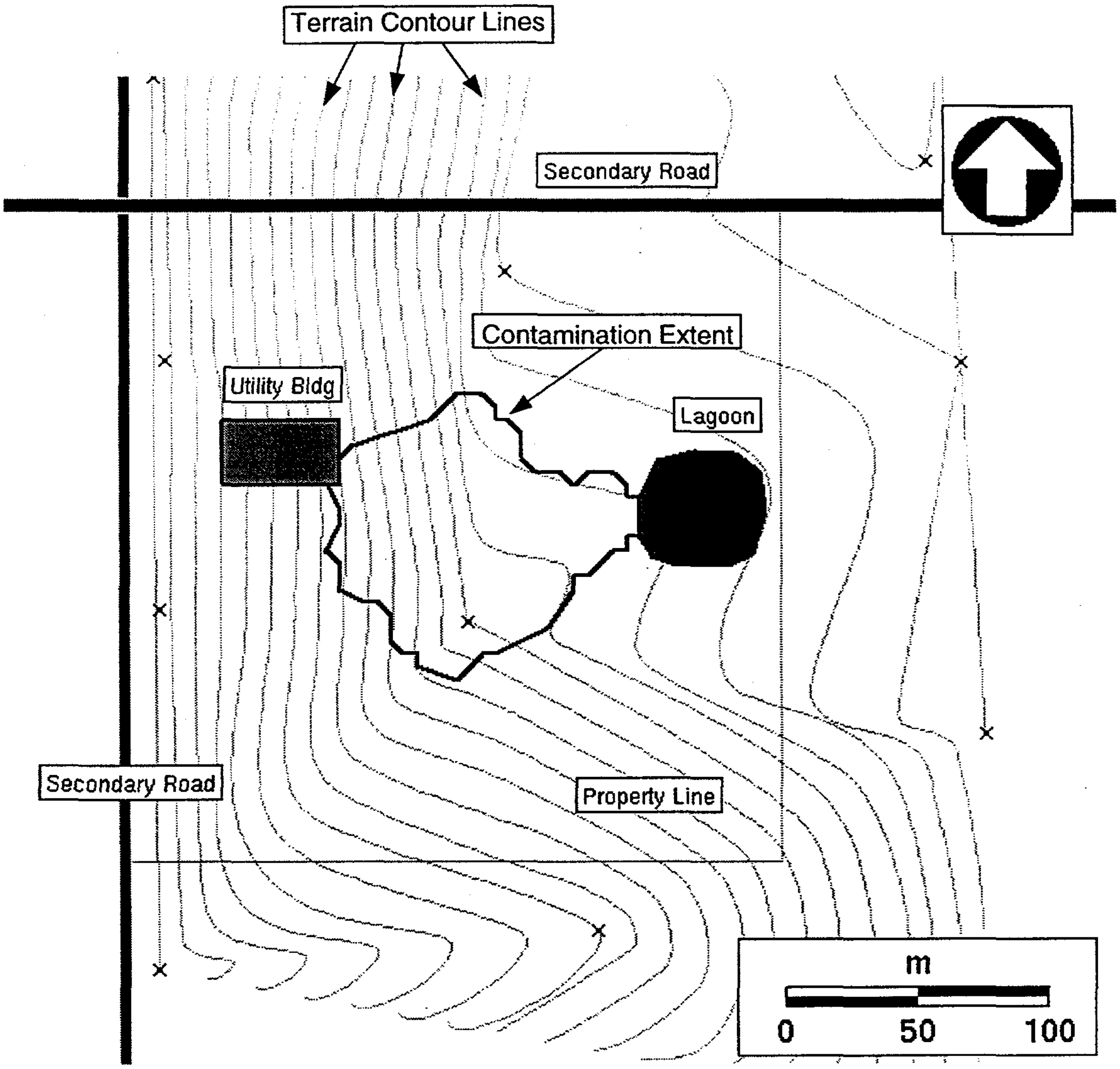

FIG. 1--Example Site

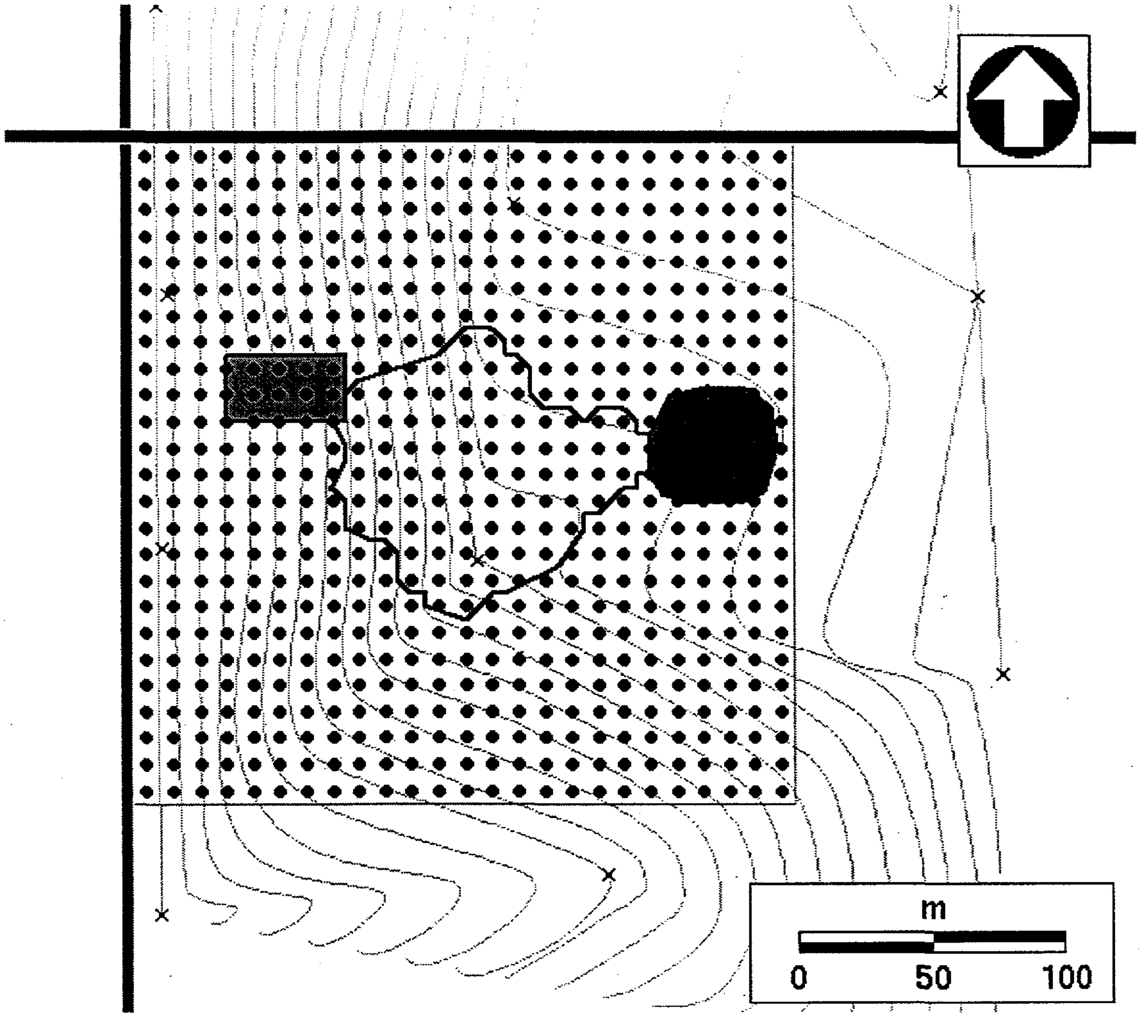

FIG. 2--Decision Point Grid

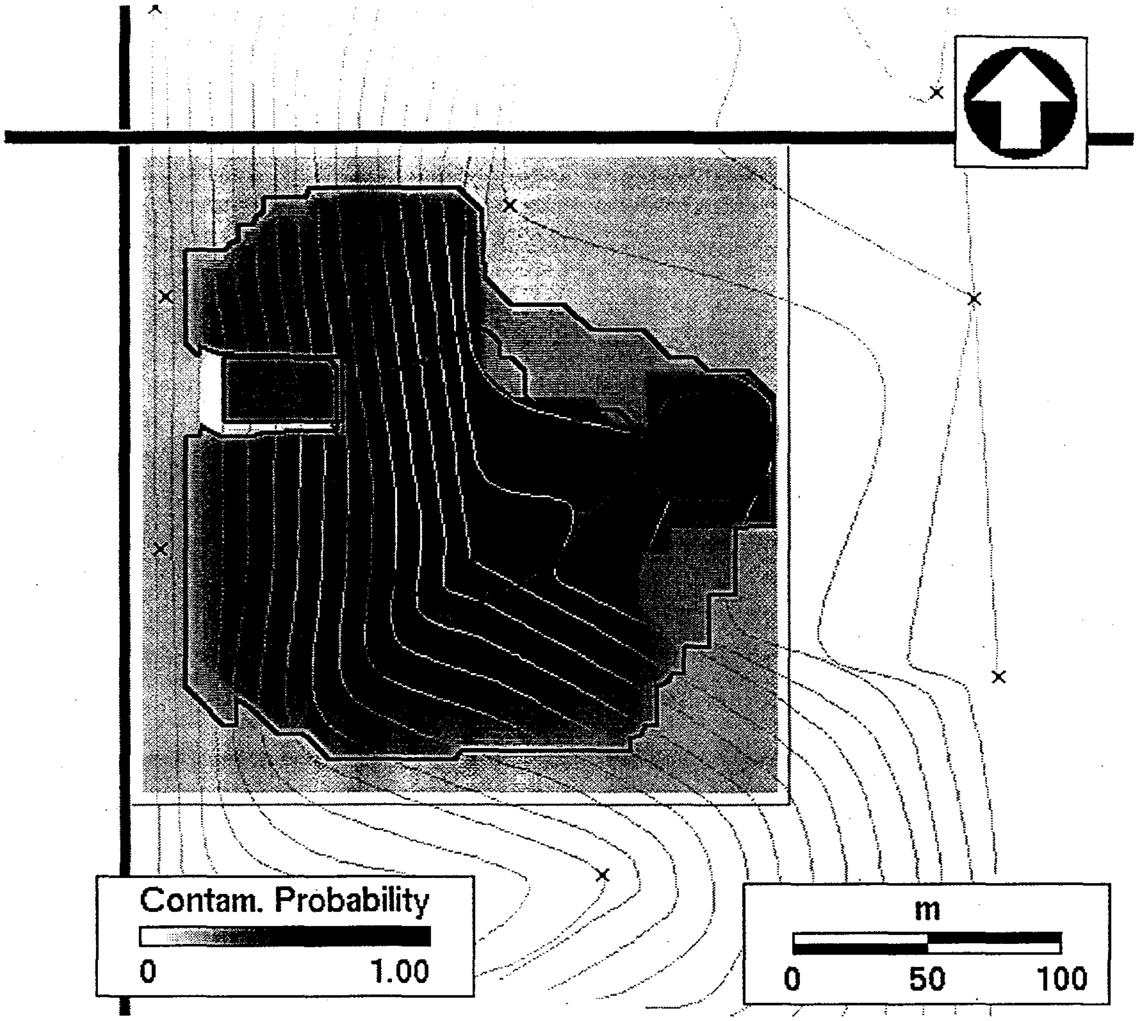

FIG. 3--Initial Conceptual Model

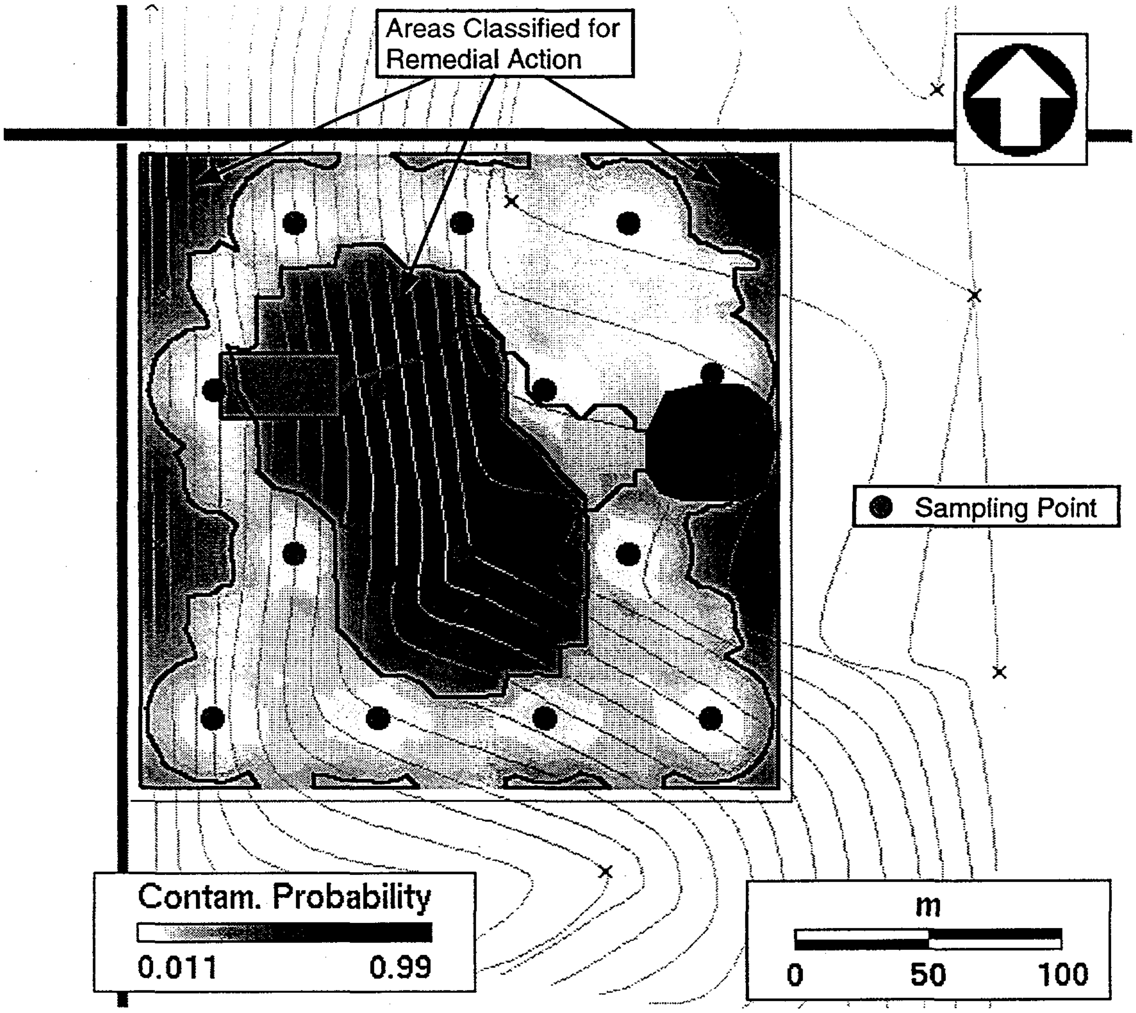

FIG. 4--Standard Sampling Program Results

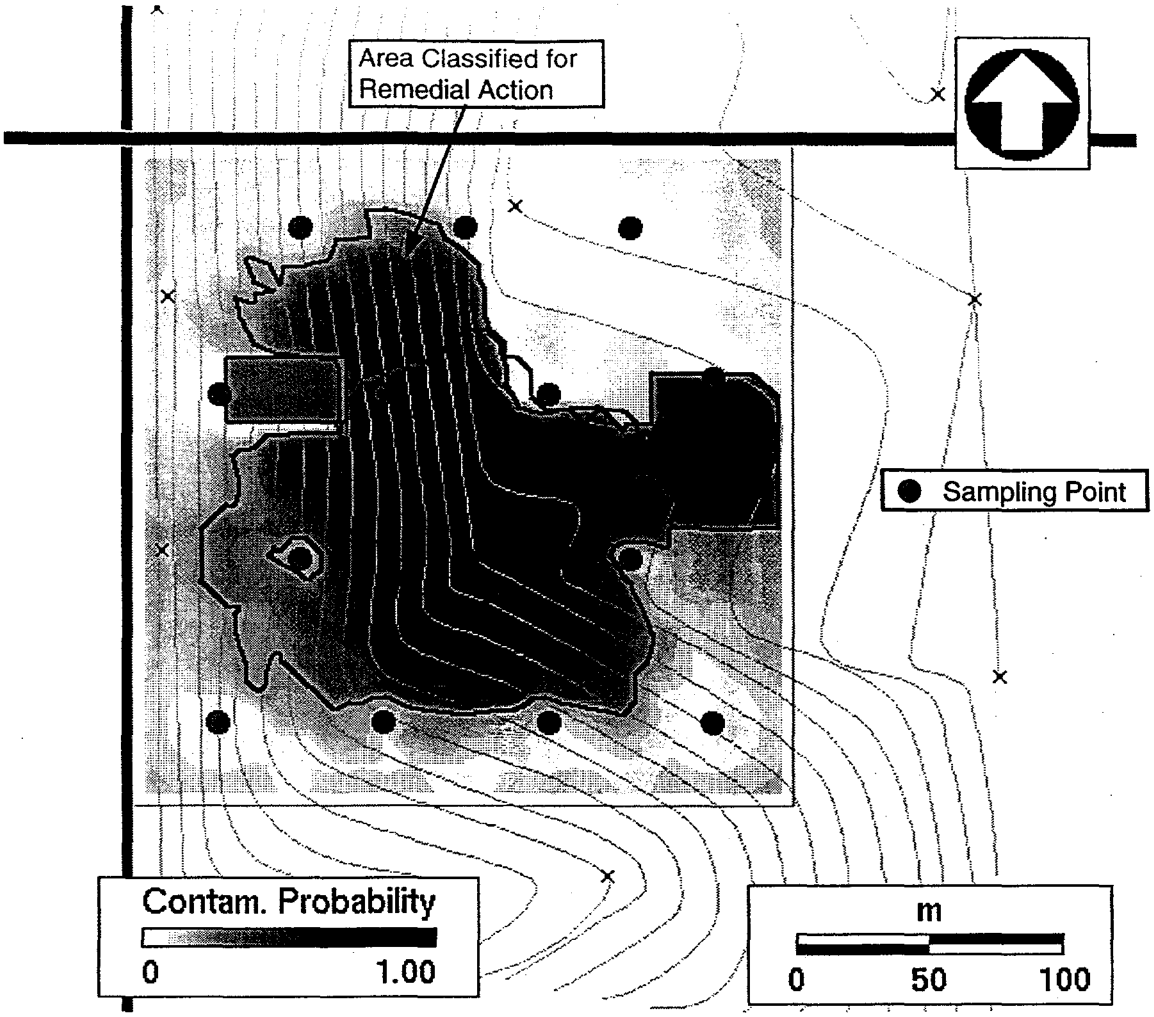

FIG. 5--Standard Sampling Grid with Initial Conceptual Model

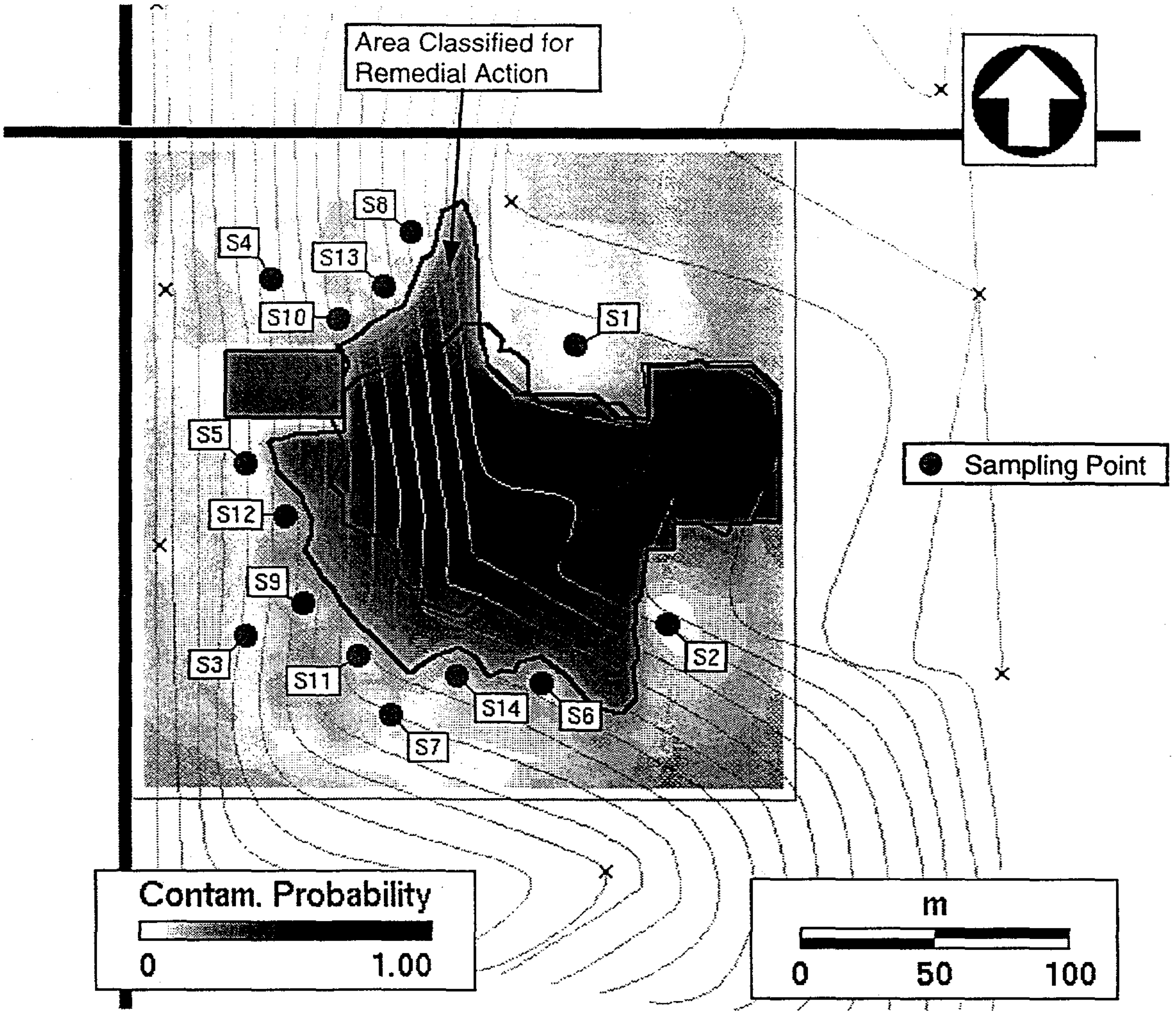

FIG. 6--Preplanned Sampling Program with Initial Conceptual Model

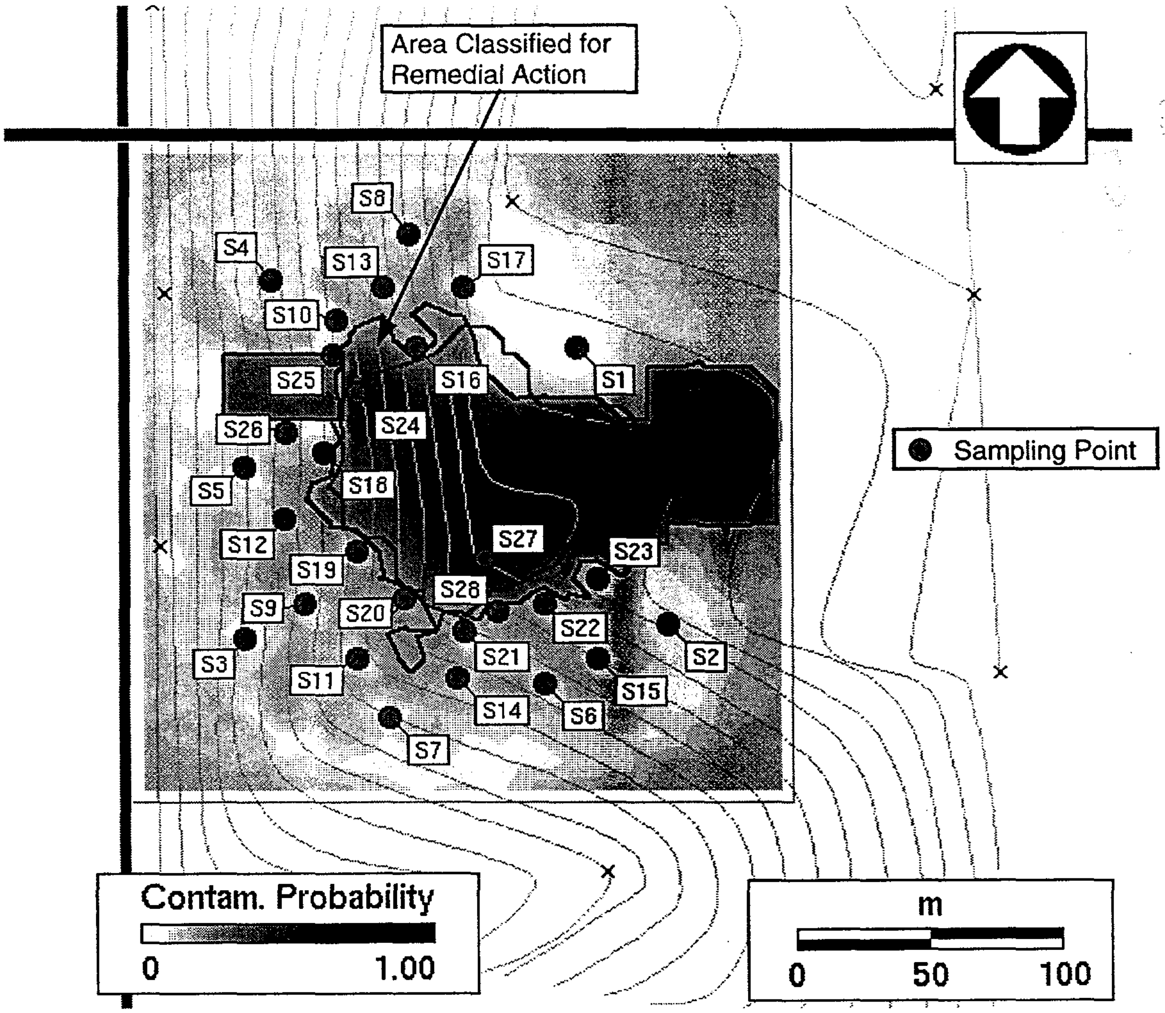

FIG. 7--Extended Adaptive Sampling Program