

# A General Method for the Efficient Selection of Sampling Locations for Problems in Environmental Restoration

RECEIVED  
MAY 15 1996  
STI

Brian M. Rutherford<sup>1</sup>

---

**Abstract.**—Problems in environmental restoration that involve detecting or monitoring contamination or site characterization often benefit from procedures that help select sampling or drilling locations for obtaining meaningful data that support the analysis. One example of this type of procedure is a spatial sampling program that will "automatically" (based on the implementation of a computer algorithm) guide an iterative investigation through the process of site characterization at a minimal cost to determine appropriate remediation activities. In order to be effective, such a procedure should translate site and modeling uncertainties into terms that facilitate comparison with regulations and should also provide a methodology that will lead to an efficient sampling plan over the course of the analysis. In this paper, a general framework is given that can accomplish these objectives and can be applied to a wide range of environmental restoration applications. The methodology is illustrated using an example where soil samples support the characterization of a chemical waste landfill area.

---

## INTRODUCTION

The objectives of environmental restoration projects often require data for detecting, monitoring, or estimating the extent of contamination. Spatial sampling procedures may be required to help locate soil sample or bore hole locations that will provide the most additional site information at the lowest cost. Two examples of problems that might benefit from this type of procedure follow:

- a) Find drilling locations for a monitoring system that will provide the information required to minimize the probability that a contaminant will cross a site boundary undetected; and
- b) Select samples to reduce as much as possible the area or volume of a 90% confidence region for a specific concentration contour.

---

<sup>1</sup> Senior Member Technical Staff, Sandia National Laboratories, Albuquerque, NM.

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED *ok*

MASTER

Two essential elements of procedures designed to accomplish these types of environmental restoration project objectives are: I) to translate site and modeling uncertainties into terms that permit comparison with regulations so that compliance and cleanup requirements can be assessed; and II) to provide methodology that can lead to an efficient sample plan throughout the course of the analyses. The concepts involved with these two elements of the procedure provide the basis for the remainder of this paper.

### TRANSLATING SITE AND MODELING UNCERTAINTIES INTO TERMS OF REGULATORY COMPLIANCE

Limited sampling and other factors contribute to site and modeling uncertainty. Most site characterization applications include uncertainty from a number of sources: contaminant values at locations that have yet to be sampled are unknown; flow and geostatistical models have uncertainty associated with them; and data available from samples collected at the site and "soft data" relevant to the site, collected through a number of alternative characterization methodologies have varying degrees of uncertainty associated with them.

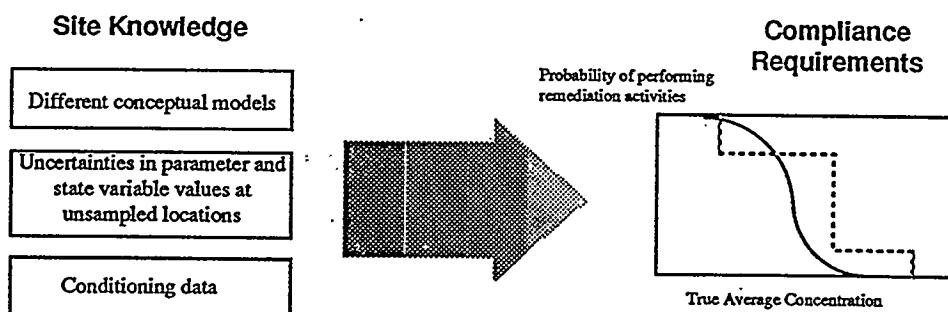


Figure 1

Recognizing these uncertainties as unavoidable, environmental regulations should be and often are stated probabilistically. Constraints on the allowable probability of exceeding a regulatory threshold, stated as a function of the (hypothetical unknown) true level of contamination, provides one example. In order to be effective, the procedure must translate site uncertainties into the probabilistic framework of these regulations. Figure 1 illustrates this element of the process for the (simpler than most) problem of characterizing average contamination levels. This will provide those responsible for demonstrating compliance of a site, or those responsible for analyzing cleanup alternatives and capabilities, with a way of making critical decisions.

The examples discussed earlier will be different (and generally more complex). Example a), for example, might have a vertical scale indicating the detection probability. The computation of these probabilities would depend on the assumed flow models and their parameters as well as the geological features of the site.

This translation of site uncertainties into a regulatory framework can be achieved through a two-step process. The first step is to convert site and modeling uncertainties into simulated realizations, each realization providing a possible map of site contamination. When taken collectively, the realizations reflect probabilistically the way that contaminant levels at the site might be configured. These realizations are then evaluated according to an objective function that calculates the compliance probability or risk.

The remainder of this section is partitioned into subsections that address these three components of this process

### **Factors Contributing to Site and Modeling Uncertainty**

At least five general sources might contribute to uncertainty in the estimated concentration levels of contaminants at a site:

- 1) Uncertainties in small scale hydrogeologic property values, either in parameter values or in state variables. These values are uncertain at locations that have not been sampled;
- 2) Uncertainty introduced by incomplete knowledge regarding the location and extent of large scale geological features such as stratigraphic units or zones of heavy fracturing;
- 3) Uncertainty concerning the extent and configuration of the contaminant source;
- 4) Uncertainty concerning the appropriate flow and transport models and in parameters associated with these models; and
- 5) Uncertainty in analytical results for samples already collected and analyzed.

The first step of the translation process is to convert site and modeling uncertainties into simulated realizations that differ from one another to an extent that reflects these uncertainties. Each source listed above must be taken into

account in the simulation process to reflect its impact on differences between simulated realizations.

Fields of hydrogeologic parameter values can be generated using stochastic simulation techniques conditioned on sample data where available. Replication of this simulation process can account for parameter uncertainty source (1). Uncertainties in state variables are often estimated through repeated application of the flow models using different sets of the hydrogeological and other parameter values (selected through simulation) for each replication and conditioned based on known values of state variables where applicable. Other parameter values may include the specification of large scale geological features, source (2) above, or contaminant source, strength and configuration, source (3). Probabilities for different contaminant sources and for different scenarios involving large scale geological features must be assessed (perhaps subjectively) and introduced into the simulation process accordingly.

Modeling uncertainty, source (4), can be accounted for by selecting among alternative models with probability of selection consistent with each model's likelihood of being applicable (again, likely to be a subjective decision). In most cases, the appropriate models are established, but parameters associated with these models must be specified. Their uncertainty distribution must be estimated, and the parameters generated randomly through Monte Carlo or perhaps alternative simulation techniques. The fifth source of uncertainty -- that resulting from inexact analytical procedures, can be accounted for by reducing the conditioning requirements of the simulation and consequently allowing differences between values obtained through sampling and those obtained through simulation at the same location.

Together, accommodating for these sources of uncertainty through the simulation process will, if performed correctly, provide an estimated probability space for site reality that is both unbiased and provides an accurate assessment of uncertainty based on present site knowledge. This discrete (by necessity) representation of the probability space is described next.

### **The Simulated Realizations**

The simulated realizations represent an intermediate step in the translation process. Each realization provides one possible map of site contaminant levels. At each grid location, the set of realizations yields a histogram of possible values. Figure 2 illustrates how the set of realizations can be interpreted in this way.

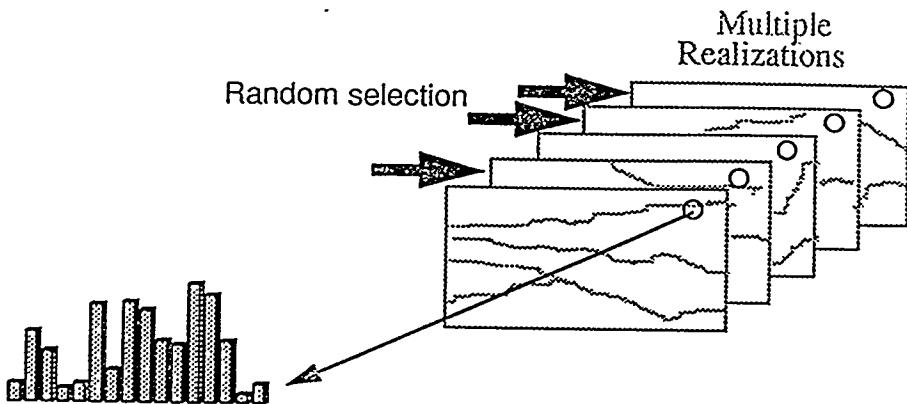


Figure 2

An appropriate interpretation of this set of realizations, analyzed pointwise, is that the probability of a contaminant value at this location, smaller than any level  $x$ , can be estimated by the ratio of the number of realizations that have a value less than  $x$  at this location, to the total number of realizations. This probabilistic interpretation can be extended to multiple locations or to the entire region. The probability that no point in the entire region exceeds the level  $y$ , for example, can be estimated by the ratio of the number of realizations where  $y$  is not exceeded at any point in the region to the total number of realizations.

This interpretation of the set of realizations is utilized in two ways in the methodology presented in this report. It provides justification for the probabilistic interpretation of the values indicating likelihood of compliance or risk functions that are compared to regulations and it provides a basis for the resampling procedure discussed in the subsection Determining an Efficient Sample Plan.

### The Objective Function

The objective function is a transformation that will compute a value indicating the level of compliance based on a complete specification of all site parameters and state variables. For the present problem, the objective function is used to transform the set of simulated realizations into a probabilistic expression that relates to regulatory compliance. Examples of the roles that objective functions might play in the example problems listed in the Introduction Section are:

- a) Provide an estimate of the probability that a contaminant will cross a site boundary undetected by a specific monitoring system within 10 years; and
- b) Provide a distribution showing probabilistically the spatial or volumetric extent of the 90% confidence region for a specific concentration contour.

## DETERMINING AN EFFICIENT SAMPLING PLAN

The process of sample location selection is often an iterative process. As new data are obtained, models are updated to provide more accurate estimates of site contamination levels or predictions of flow and transport and a reassessment of future data needs is performed. A second essential element of an effective spatial sampling procedure is that it utilizes data available at every sampling iteration and that models based on these data produce an efficient sample plan -- one where maximum relevant information is provided at the lowest possible cost. At each iteration, the next set of sample locations is selected from numerous potential supplemental sample sets. Methods for choosing sets to consider and methods for evaluating those that are considered must be carefully selected.

A number of applications have been reported in the hydrology and geology literature where selection of sampling or monitoring locations has been approached using probabilistic optimization techniques. Early applications, involving resource assessment generally in mining applications, were based on the assumption that the level of the state variables in the region (ore grades, in many applications) could be modeled as a Gaussian random field using regionalized variable theory. For these applications, the utility of any future sample set could be assessed prior to collection and analysis by estimating the reduction in kriging variance or block kriging variance that the set yielded.

Recently, environmental restoration applications addressing spatial sampling problems have appeared, primarily in the hydrology literature. These applications are often less straight-forward than those based on regionalized variable theory. Most applications are based on flow models that use as input, maps of hydrogeological parameters (porosity or hydraulic conductivity, for example). Another possible difference between environmental restoration problems and those of the earlier applications is the variety of objectives for different analyses. The objective function determines the utility of further sampling and that of sampling at specific sets of sample locations. The likelihood of a contaminant crossing a site boundary undetected by a specific monitoring system, for example, is an analysis objective that is not related to a reduction in the kriging variance in any obvious way. Consequently, the utility of additional sample or monitoring locations must be approached differently for this objective.

A number of the publications involving spatial sampling problems in environmental restoration focus on the development of methods that facilitate the assessment of state variable information (hydraulic head or contaminant levels, for example) provided by samples of hydrologic parameters. Another primary focus has been in the development of optimization techniques that approach these

specific objectives in an efficient manner. One article that provides an excellent review of accomplishments in these areas and provides numerous further references is Mc Kinney and Loucks (1992). The primary objective in the present paper is to develop a more general approach to optimization for spatial sampling problems that can be applied to a wide range of problems regardless of their modeling requirements and varying analysis objectives.

A sampling plan can be constructed utilizing the appropriate probability space generated using the procedures outlined above, along with optimization methods outlined in this section. A more detailed discussion of these methods is provided in Rutherford (1996). The sampling plan is based on an iterative procedure that requires evaluation of samples collected at each iteration of the process to provide new information for selecting sites to be used in the next iteration. Two issues arise in this process that will be outlined in the next two subsections. The final subsection provides an example of the entire process.

### **Assessing the Value of Sampling at Unsampled Locations**

As alternative sets of supplemental sample locations are considered, there is a need to evaluate the potential for improvement provided by each sample set prior to collection or analyses of the samples. This requires an estimation of the information available from each sample set and an evaluation of its likely impact on the objectives of the overall analysis. Assessment of the relative overall information provided by sampling at any set of sample locations depends on site and modeling information that has been translated to the simulated realizations. The effect of "realizing" values projected by the simulated realizations provides an indication of what might occur if that sample set was actually selected and analyzed.

Figure 2 shows how the resampling process might be implemented. For each hypothetical sample set, realizations (selected at random) are assumed to represent the true status of concentration at the site. Their values at the candidate supplemental sample locations can be submitted as conditioning data along with known sample information to formulate new models. These models can be used to generate a new set of simulated realizations that are then evaluated using the objective function to determine the impact of the new information provided. This process is repeated several times for each supplemental set of sample sites considered. Sets of sample locations that demonstrate only small modeling differences from one realization to the next, or those indicating significant differences to the models, but where the differences have little impact on the objective function, should be excluded from further consideration.

This resampling procedure can be very expensive in terms of computer time, especially when there are flow and transport models that must be reevaluated at each step. Consequently, effective procedures are required for selecting sets of sample sites for consideration.

### What is Meant by an Efficient Optimization Algorithm

There are several techniques available for selecting supplemental sample sites. Some techniques will yield an exact optimal solution; other, less computer intensive techniques, will yield a "near optimal" solution. The distinction between these solutions, and the considerations that might provide a preferred technique for any specific application, are outlined in this subsection.

Spatial sampling problems can be approached as a combinatorial optimization problem where a large but finite number of solutions are available. Each possible solution (a set of supplemental sample locations) can be compared through the objective function, to alternatives, using the resampling procedure.

The ideas involved in the sample set selection algorithm are either to find an exact optimal solution dismissing as many solutions as possible without a formal evaluation (recall that formal evaluation may require many executions of a computer intensive simulation involving a flow and transport code), or to find a near optimal solution using the fewest comparisons possible. These methods can be, and probably are, most effectively used in combination.

Simulated annealing and genetic algorithms are methods that compare solutions iteratively, providing a near optimal solution where "near" depends on the efficiency of the algorithm and the number of computational iterations involved. Branch and bound algorithms can be used to find an exact optimal solution. These algorithms derive their benefit (compared to an exhaustive search of every possible solution set) by dismissing large numbers of solutions through a single objective function comparison. Hybrid methods are available that utilize features of all of these random search methods.

Details concerning the concepts and methods involved in these algorithms can be found in references cited in Rutherford (1996). Details of how the procedures might best be implemented for spatial sampling or monitoring problems can also be found in that paper.

---

### 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.

## Chemical Waste Landfill Example

This example is provided to illustrate how these spatial sampling procedures might work. The procedure demonstrated here was implemented using a spatial sample tool (OPTMAS), developed at Sandia National Laboratories, for the purpose of locating soil sample locations to provide the maximum relevant site information. The site analyzed is a chemical waste landfill site at Kirtland Air Force Base in Albuquerque, New Mexico. The objective of this analysis (for purposes of illustration) is to determine the extent of a fixed contamination contour with a 90 percent level of certainty.

Figure 3 illustrates the present sample locations by the black dots. Clearly, previous sampling was performed according to some systematic prescription. The center, shape, and extent of the site is presumed to be unknown, but it is assumed that the contamination occurs throughout a contiguous region or, at least, that the contour associated with the outer boundary is the one of interest. This is a realistic problem, for example, in situations where the concern is to achieve a certain level of assurance that a contaminant is contained within site boundaries. It was further assumed that 15 additional sample locations were to be specified at each iteration throughout the analysis. (That is, 15 sample locations are required here. These samples will be taken and analyzed and the process will be repeated if necessary).

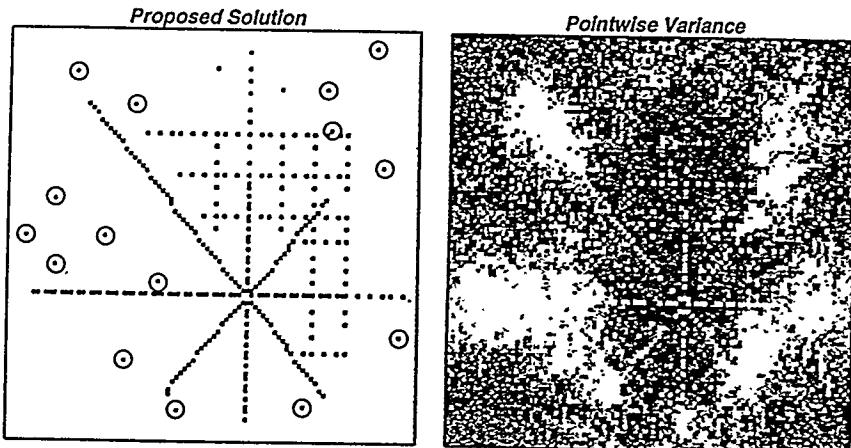


Figure 3

The approach taken to estimate the extent of these contours is the approach described in previous sections of this paper, but with a number of simplifications. First, the simulated realizations are generated using sequential Gaussian simulation (because it was readily available, but not necessarily the best approach to modeling a contaminated region). The set of realizations was used to compute the pointwise variance as illustrated in Figure 3 (the darker areas indicate higher variance). The

reduction in variance is the assumed objective function criterion for this example. Therefore, after the pointwise variance of the set of realizations has been determined, the realizations are no longer needed for the remainder of the analysis (until the next sampling iteration). This further simplification is possible because it is assumed for simplicity that the variance is reduced at, and around, sample locations in a way proportional to the reduction indicated by the previously simulated realizations. The details are provided in Rutherford (1996). This is a luxury that is not offered in the generic spatial sampling problem where new (hypothetical) realizations are required to evaluate each set of potential sample locations. Finally, it was assumed for reasons outside the scope of this paper that the sampling had to occur outside the 10th percentile bound for this contamination contour.

Alternative sample locations were selected using the hybrid sample selection algorithm mentioned above, and detailed in Rutherford (1996). For each sample set selected in the simulated annealing or genetic algorithm portions of the procedure, the objective function (total variance throughout the region) was computed. The sample locations were dismissed or retained as a candidate solution for the problem as prescribed by the hybrid sample site selection algorithm. The circled locations in Figure 3 present a near optimal solution to this problem. It is difficult to determine exactly how close they are to the true optimal solution without using an exact approach. For the present illustration, it suffices to say that the solution appears to satisfy the obvious desirable characteristics of a solution to the problem -- the samples are in areas (constrained as described earlier) where the variability is relatively high and they are spread out throughout this region so that the reduction in variance achieved by sampling at one location will have a minimal effect on the variance reduction achieved by sampling at another selected location.

This simplified example was chosen with a relatively intuitive solution to illustrate the process. The samples selected in a more realistic example might be more difficult to analyze intuitively.

## References

Mc Kinney D.C., and D.P. Loucks (1992). "Network Design for Predicting Groundwater Contamination", Water Resources Research, Vol. 28, No. 1.

Rutherford, B. (1996). "A General Approach to the Problem of Selecting Optimal Spatial Sampling or Monitoring Locations", white paper (in progress), Sandia National Laboratories, Albuquerque, NM 87106-0829