AIIM

Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202

Centimeter

Inches

MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

1 of 1

Conf-9306158--2

A BAYESIAN APPROACH TO CONTAMINANT PLUME DELINEATION

R. L. Johnson

Environmental Assessment
and Information Sciences Division
Argonne National Laboratory
Argonne, Illinois

ABSTRACT

Fundamental to accurate modeling of groundwater contaminant transport is an understanding of the initial extent of contamination. The initial location of contamination must typically be inferred from relatively little "hard" sampling data, and related "soft" information. A combination of Bayesian analysis and spatial statistics is used to generate initial estimates of contamination based both on "soft" information (i.e., prior modeling results, past experience, etc.) and "hard" data (i.e., sampling results). A Beta distribution is used to represent the probability that contamination is present for at any grid node. For every node, prior Beta probability density functions are revised into posterior distributions as additional data become available. The result is a node by node probabilistic representation of plume extent that can be updated as new information is generated, that indicates where additional sampling would have the greatest impact on the uncertainty associated with plume extent, and that can serve as the basis for Monte Carlo analysis of contaminant transport. This approach is demonstrated with information from a subsurface chromium plume beneath the chemical waste landfill at Sandia National Laboratory, Albuquerque, New Mexico.

INTRODUCTION

Uncertainty is introduced into numerical transport modeling at a variety of levels. Model selection, parameter identification, boundary and initial condition specification---all contribute to the overall level of uncertainty attached to the final modeling results. This paper specifically addresses one source of uncertainty, that associated with the specification of initial contaminant extent, and shows how a combination of Bayesian analysis and spatial statistics can be used to quantify the initial level of uncertainty associated with contaminant extent, and identify sampling

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

locations that would have the greatest impact in reducing that uncertainty. The methodology developed provides a node by node probabilistic representation of plume extent that can serve as the basis for Monte Carlo analysis of contaminant transport.

METHODOLOGY

Contamination events generally possess two key characteristics. First, they are spatial processes that are dominated statistically by spatial correlation structures. This simply means that the results from one sampling location will look very similar to results from samples taken close by, but will have little connection to samples separated by greater distances. Second, there typically exists substantial "soft" information regarding probable plume extent, even if little hard sample data is available. This information might be from past experience with similar sites, from historical records documenting the release size, from non-intrusive geophysical surveys, or from preliminary modeling results. Such "soft" information is often just as important in defining the initial state of contamination as hard sample data.

The problem is quantitatively integrating "hard" sample data with soft information in a manner that is consistent, that preserves hard sampling data, and that accounts for spatial correlation of hard sampling results. Suppose for the region of interest one defines a three-dimensional nodal grid, and at each node assigns a probability density function to the probability π that contamination exists at that node. For various reasons that will become clear, a good choice for this probability density function would be a two parameter Beta distribution, Be(α, β):

$$Be(\alpha,\beta) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1} \qquad (0 \le \pi \le 1;\ \alpha,\beta > 0) \tag{1}$$

Beta distributions range between 0 and 1, and can assume a variety of shapes depending on the choice of α and β. Figure 1 shows three shapes corresponding to three different pairs of parameters. According to Lee, (1989), Beta distributions have as mean:

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

and variance:

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

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

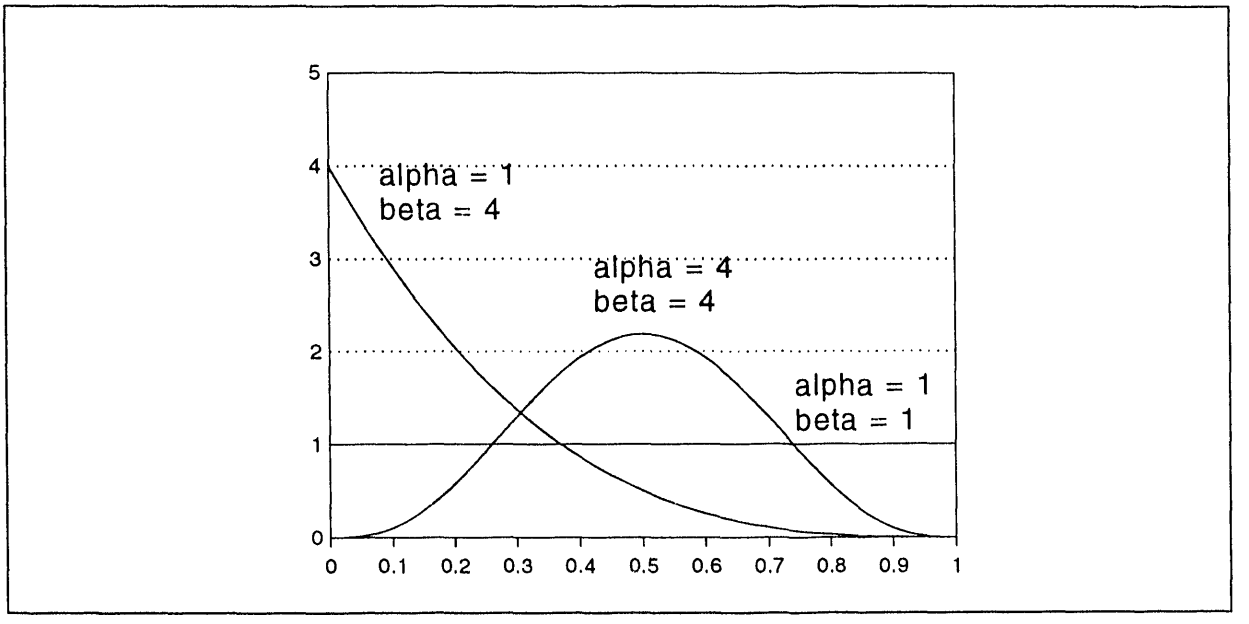

Figure 1. Examples of Beta Distributions

function (pdf). As additional information becomes available regarding the actual values of parameters, Bayesian analysis provides a quantitative means for merging the new information with the prior pdf to produce what is known as the posterior pdf via Bayes rule:

$$P(X \mid Y) \propto P(X)P(Y \mid X) \tag{4}$$

This states that the probability of X given Y is proportional to the product of the probability of X and the probability of Y given X. In (4), P(X) functions as the prior pdf, while $P(X \mid Y)$ is the posterior pdf.

From a Bayesian perspective, Beta distributions are conjugate priors in the context of Bernoulli trials and the binomial distribution (Lee, 1989). Binomial distributions provide the probability of observing a given number of successes within a specified number of trials. Conjugate priors are priors that retain their same underlying pdf after Bayesian updating. In the case of a binomial trial where X successes were obtained in N trials, a prior for π of the form $Be(\alpha, \beta)$ would become $Be(\alpha + X, \beta + N - X)$. N here functions as the total additional information supplied to the prior. As N grows large, $E(\pi)$ approaches the classical maximum likelihood estimator for π, X/N, and $Var(\pi)$ is steadily reduced.

When one considers only the presence or absence of contamination, environmental sampling resembles a binomial trial---N number of samples taken, of which a certain fraction actually encounter contamination. There are two main

differences, however, between Bayesian analysis as derived for binomial trials and Bayesian analysis used for updating nodal grid pdfs. The first is that for a nodal grid, the probability of contamination at a given node becomes either zero or one once that point has been sampled. The second is that environmental samples are not independent. For samples that are spatially close, one must assume that some spatial autocorrelation exists. For a given node with prior pdf $Be(\alpha_0, \beta_0)$ and a set of sample locations where hard information exists regarding the presence or absence of contamination, the issue is how to update the nodal prior pdf in a way that is consistent with the derivation of Beta distributions as conjugate priors for binomial trials. The approach taken in this paper employs indicator kriging. Indicator kriging as described by Journel (1983) provides a non-parametric estimate of the probability of contamination at a location x_0 given a set of sampled points with binary outcomes--zero if clean, one if contaminated. Indicator kriging assigns an "optimal" set of weights to sampled data to construct a linearly interpolated value:

$$Z(x_0) = \sum_{i=1}^{N} w_i Z(x_i) \tag{5}$$

where $Z(x_0)$ is the interpolated estimate at x_0, w_i are the kriging weights used, and $Z(x_i)$ are the observed indicator values at locations x_1 through x_N.

These weights are "optimal" in the sense that they provide unbiased linear interpolations with minimum interpolation error. Ordinary indicator kriging derives these optimal weights via a set of linear equations:

$$\sum_{i=1}^{N} C_{ij} w_i + w_{N+1} = D_{j0} \qquad \textit{for } j = 1, \ldots N \tag{6}$$

$$\sum_{i=1}^{N} w_i = 1 \tag{7}$$

where C_{ij} is the covariance between points x_i and x_j, the **w** are the kriging weights, and D_{j0} is the covariance between sampled point x_j and the point where interpolation is taking place, x_0.

Indicator kriging has several advantages over ordinary kriging. Indicator kriging typically relies on a more robust estimated variogram than ordinary kriging, partly due to the fact that it is immune to outlying sample values. Indicator kriging can be applied to data that are only qualitative, such as might be generated by field screening sampling technologies. Finally, indicator kriging makes no assumptions about the distribution of contamination, or the global stationarity of the contamination event.

To update a prior Beta distribution at location x_0 that has not been sampled with information from neighboring sampled locations requires estimates of N^*, the

total amount of information available at x_0, and X^*, the portion of N^* that represents contaminated "hits". One way to measure N^* is to use the estimation error providedby indicator kriging:

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

Here μ is the average of the x_i involved in the interpolation. Var_{estim}, the estimation error, varies between 0, when the value at x_0 is known with certainty, up to $2*C_{00}$, which implies that the sampled points provide no information at x_0. N^* can be related to Var_{estim} by:

$$N^* = \frac{2C_{00}}{Var_{estim}} \tag{9}$$

and X^* simply becomes $Z(x_0)*N$. Note that N^* and X^* do not depend on the actual values of Var_{estim} and C_{00}, but rather on their relative sizes.

Figure 2 shows how one builds and updates a conceptual image of contaminant location using this technique. The first step is to capture existing soft data regarding contaminant location in the initial choices for α and β. The parameters α and β can be chosen initially for each node so that for every node, $E(\pi)$ matches one's estimate of the likelihood of contamination at that node, and $Var(\pi)$ reflects one's confidence in that estimate. Higher levels of confidence require larger α and β. In the unfortunate case where no soft data exists, one can begin by assuming "non-informative priors" for each grid node. Examples include Be(1, 1), which is uniform on the interval [0, 1] with mean 0.5, or Be(0, 0), which also has mean 0.5, but concentrates its density near 0 and 1.

Updating the nodal grid with hard sampling data requires knowledge of the covariance function or variogram for the hard data. Because the derivations of both X^* and N^* are independent of C_{00}, the two main determinants of the variogram are its functional form, and its range. If sufficient hard data exist, a variogram can be estimated directly from the data. If hard data are negligible, a good rule of thumb is to select the variogram's range equal to the expected width of contamination. It is possible to select anisotropic variograms that reflect a contaminant plume's spatial orientation or preferential flow directions. The variogram itself can be treated in a Bayesian fashion, following a process similar to that described by Massman and Freeze (1989).

Evaluating the need for additional sampling depends, of course, on the relative impact additional sampling information would have on overall modeling uncertainty. One simple measure is the number of nodal points whose expected probability of contamination falls above or below specified certainty levels. Since each node has assigned a probability density function corresponding to the probability of contamination being present, synthetic contaminant plumes can be constructed by

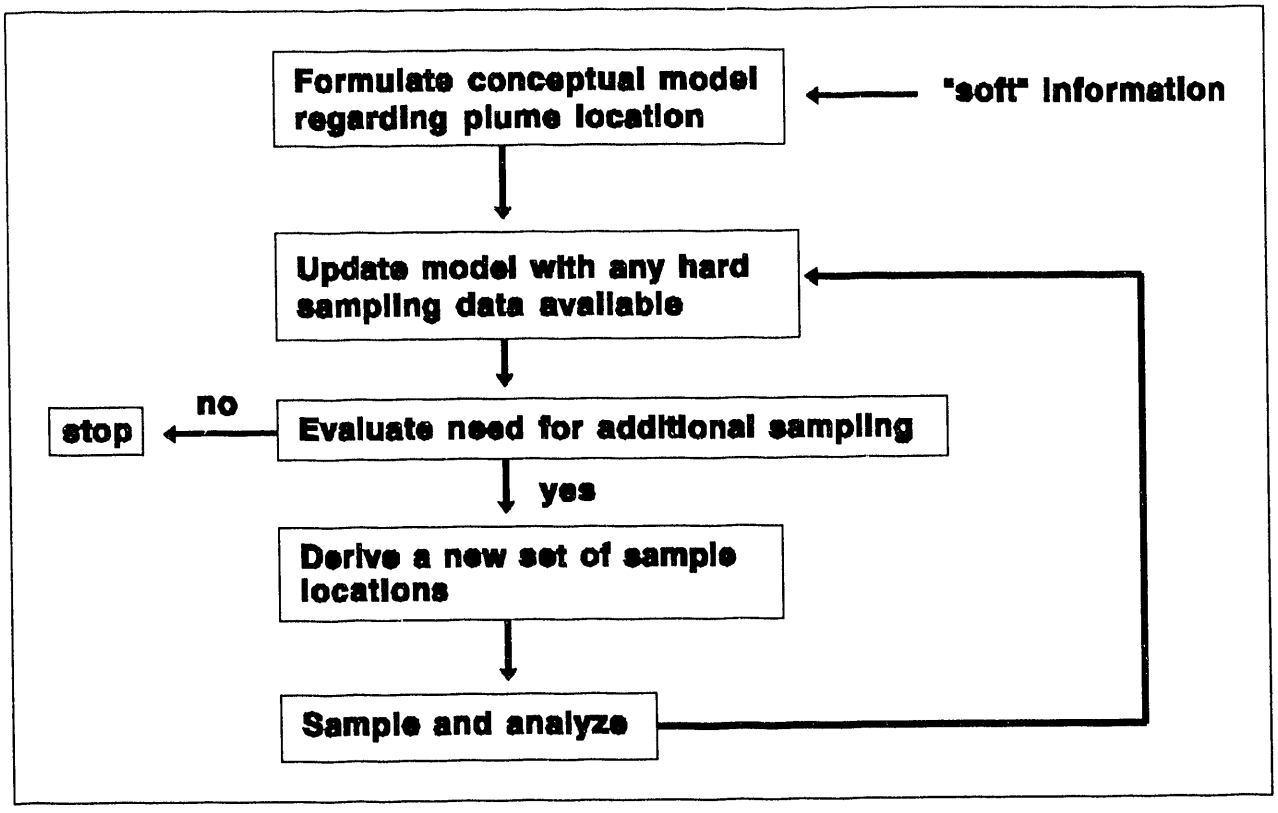

Figure 2. Logic Flow Diagram

Monte Carlo sampling from each pdf at each node. Alternatively, synthetic plumes can be constructed simply by selecting all nodal points whose expected probability of contamination is above a specified certainty level.

EXAMPLE

The U.S. Department of Energy (DOE), through its Office of Technology Development (OTD), supports the development and demonstration of emerging technologies for environmental assessment and remediation. One program, the Mixed Waste Landfill Integrated Demonstration (MWLID), demonstrates in-situ characterization technologies for landfills in arid environments that contain complex mixtures of metal, organic and radioactive wastes. In 1992, the MWLID focused its attention on an unlined chromic acid pit within Sandia National Laboratory's Chemical Waste Landfill. Because of a large but unknown amount of plating solution that was disposed of into the unlined chromic acid pit, a chromium plume now exists within the unsaturated zone beneath the pit. Past studies had indicated minimal sorption of chromium in the local soils, meaning that it moved freely with soil water.

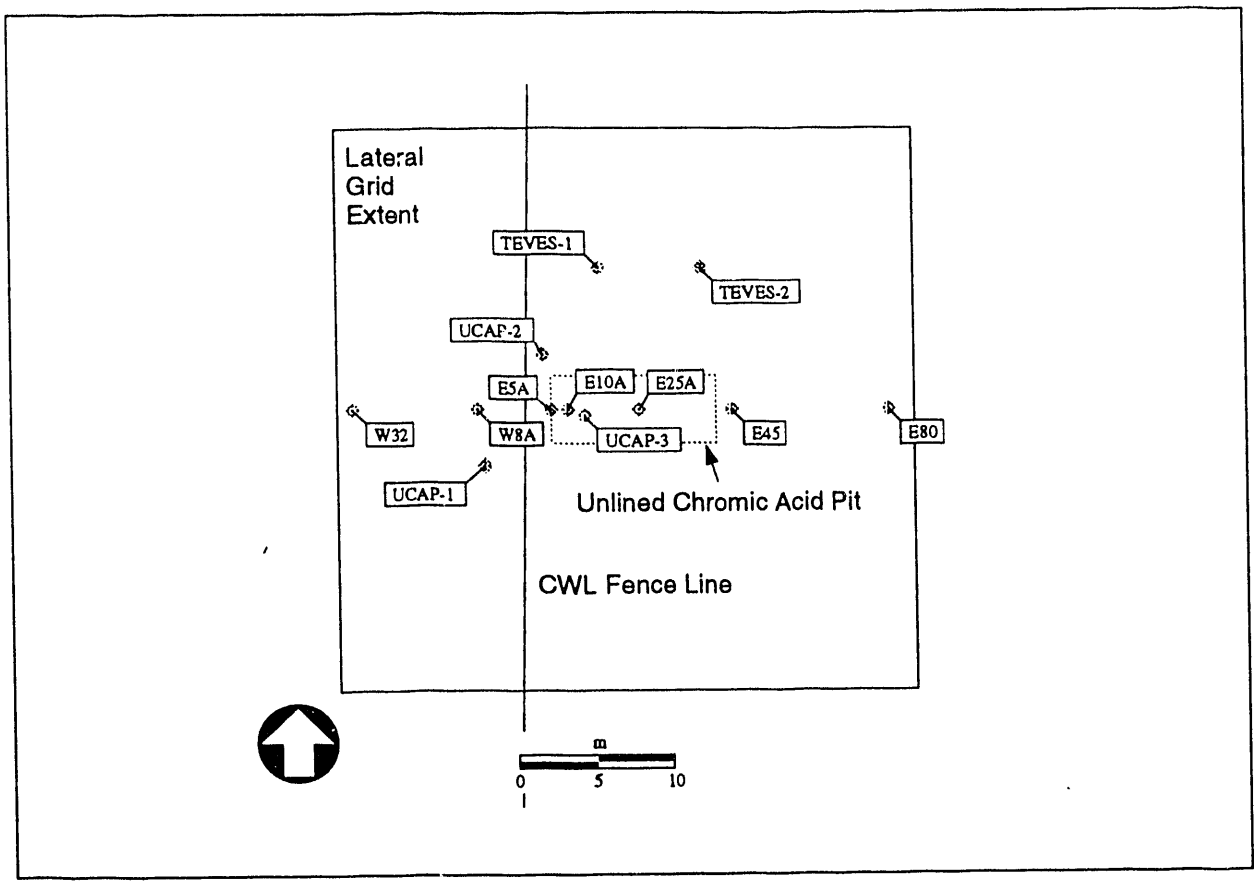

Figure 3. Unlined Chromic Acid Pit

Of primary concern is the travel time to the water table for this plume given base conditions, and various remediation alternatives. The current plume extent is crucial to determining plume travel times.

Figure 3 focuses on the unlined chromic acid pit, displaying the locations of various soil bores that have been completed to characterize the extent of contamination. These soil bores ranged between 20 and 36 meters deep, with samples taken at regular intervals. A regular grid was superimposed on the site consisting of 28,830 nodes---31 in the east-west direction, 31 in the north-south direction, and 30 deep, with a 1.2 meter separation between nodes. The demarcation in Figure 3 shows the lateral extent of the grid. Based on data from borings completed in the past decade, prior pdfs were assigned to each node. Beta distribution parameters α and β were set on the borders so that the probability of contamination there was unlikely, and uniformly to $\alpha = \beta = 0$ throughout the remainder of the nodal grid, for a probability of contamination equal to 0.5. The nodal grid was then updated first with 1987 sampling data from seven bores lying on

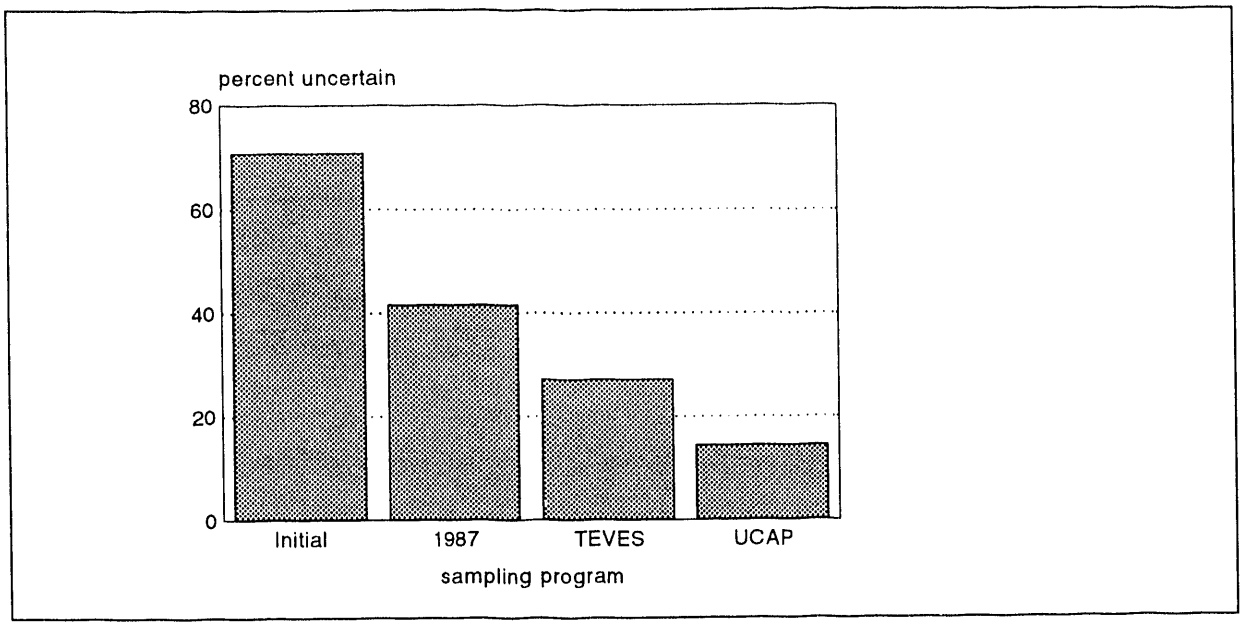

Figure 4. Nodal Classification as a Function of Sampling Program

an east-west transect of the pit, then with data from the two TEVES bores, and finally with data from the three UCAP bores. The variogram used for updating was determined from an analysis of the 1987 sampling data.

At each step of this process, the number of nodes with probabilities above and below specified certainty levels was determined. Figure 4 shows the impact each sampling program had on the classification of nodal points at the 60% certainty level. Nodes with a probability of contamination between 0.4 and 0.6 were classified as uncertain. Figure 5 shows the percent of nodal points with probability of contamination greater than or equal to different probability values based on sampling data from all three sampling programs. As is obvious from Figure 5, the certainty level selected for determining initial contaminant conditions has a strong bearing in this instance on the presumed initial extent of contamination.

CONCLUSIONS

For many contaminated sites, soft information about the extent of contamination is often in greater abundance than hard sample data. This soft information can include historical records of the contamination event, experience with other similar sites, preliminary modeling results, etc. Soft data can be more important than hard sample data when delineating the potential boundaries of plumes. Using a combined Bayesian/geostatistical approach, the proposed methodology provides a means of quantitatively merging soft information regarding

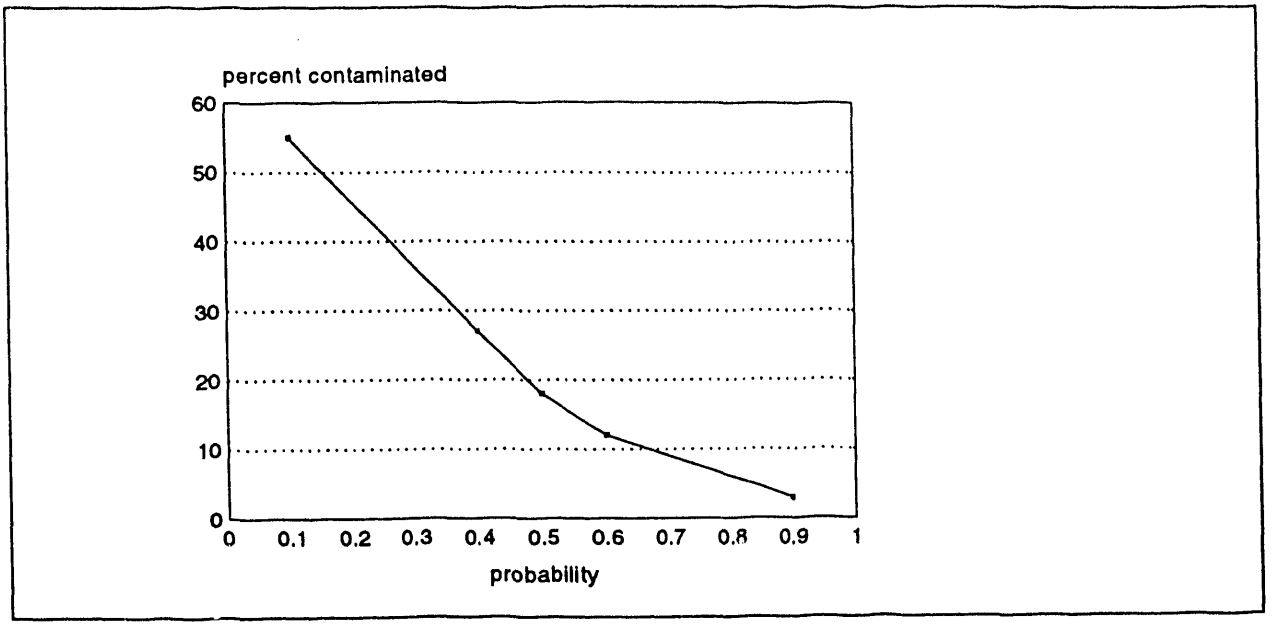

Figure 5. Nodal Classification as a Function of Certainty Level

contaminant location with hard sample data. The result is a nodal grid containing node by node pdf's that represent the probability of contamination at each node, and that can be further updated as additional hard sample data becomes available. This nodal grid is useful for determining where additional sampling would have the greatest impact on the uncertainty associated with plume extent, and can serve as the basis for Monte Carlo analysis of contaminant transport.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy, Office of Environmental Restoration and Waste Management, under contract W-31-109-ENG-38.

REFERENCES

Journel, A.G., 1983, "Nonparametric estimation of spatial distributions", Mathematical Geology, Vol. 15, pp.445-469

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

Massman, J. and R.A. Freeze, 1989, "Updating random hydraulic conductivity fields: A two-step procedure", Water Resources Research, Vol. 25, pp. 1763-1765.

DATE
FILMED
9/14/93

END