An Application of Geostatistical Tools to Characterize an Environmental Site

M. Pyrcz
University of Alberta (mpyrcz@gpu.srv.ualberta.ca)

This paper and companion poster stem from a case study that focuses on the application of geostatistical tools to characterize the spatial distribution of lead (Pb) concentration at an environmental site. The work presented herein builds on the master's thesis of E. H. Isaaks. The study is based in the city of Dallas, Texas, where a lead smelter was situated. The data have been "scaled" and no regulatory requirements have been used.

This paper lays out the steps and tools that are effective in quantifying uncertainty and making decisions in site characterization. This paper is not an exhaustive study of the tools available; assumptions and simplifications have been accepted for brevity. For example, the decision making section of this paper utilizes a simple model. The actual construction of an economic model to describe the cost of clean up versus the cost of possibly leaving a section with unsafe lead concentrations would be an advanced topic.

Introduction

Geostatistical methods for modeling variables in space are widely accepted and applied in ore reserve and reservoir modeling (Deutsch, 1999, Goovaerts, 1997, Journel, 1989, Isaaks and Srivastava, 1989). These methods are finding application in a wide variety of fields, including contaminated site investigations. E.H. Isaaks conducted an investigation of an environmental site in his Master's Thesis (Isaaks, 1984).

The lead smelter data set is useful for illustrating stochastic modeling. In this work, I have revisited the environmental site and worked through a typical modeling work flow. This case study was prepared as a teaching tool. Exploratory data analysis, spatial correlation, kriging, simulation and decision making are illustrated.

Exploratory Data Analysis

Exploratory data analysis involves gaining an appreciation for the nature of the available data. As a logical beginning point, this step focuses on "getting a feel" for the data. Observations made during this stage are the foundation for subsequent analysis and modeling.

The location map of the Pb soil concentrations is shown in Figure 1. The sampling campaign appears biased towards the center where the smelter was located. An assessment of data clustering is required for the derivation of a representative histogram. In this example there is only one regionalized variable, lead concentration from soil samples, and therefore there are no multivariate relationships to explore.

The color coding of the samples, permits us to observe trends in the data. There is a localized region of high Pb levels in the center of the study space represented by a few

samples. Also, the data is sampled in an irregular pattern with a large band being unsampled in the first quadrant of the sample space. As mentioned in E.H. Isaaks's Master's Thesis, this region is coincident with the Trinity River Flood Plain.

The distribution of the samples is presented in the histogram of the equal-weighted data, see Figure 2. The distribution appears to be bimodal. The values under 1000 PPM approximate a log-Gaussian shape, which could represent naturally occurring deposition. The values over 1000 PPM appear to be disconnected from the background levels indicating possibly a deposition event. While being valuable, the equal-weighted data histogram can be misleading since it is not representative of the entire sample space.

Declustering must be used to weight the data with respect to area and remove the sampling bias. Cell declustering was chosen because it is robust and it does not require a well defined boundary to the study area. The declustered mean was calculated with varying cell sizes, see Figure 3, and the cell size that yielded the minimum declustered mean was chosen. The decision to accept the minimizing cell size was based on the observation that the areas of high Pb concentration are over-represented. From the plot of the declustered means versus cell size, the optimum cell size can be estimated at about 5100 feet. An area-representative histogram is generated with the declustered weights, see Figure 4.

The declustered histogram reports a slight drop in the mean and standard deviation. This drop in mean would be expected since the minimizing cell size was chosen. The reduction in standard deviation is due to the shifting of weight from the high extremes to lower values, which are closer to the mean.

Analysis of Spatial Correlation

We must calculate, interpret, and build a model of the spatial correlation of Pb concentration. The resulting model plays a significant role in our assessment of uncertainty, volume variance relations, and decision making. The model will rely on some subjective analysis and interpretation of calculated spatial statistics, and should be cross validated and refined. For example, the nugget effect may be difficult to determine and yet the results are sensitive to the nugget effect. A model with a too high nugget effect will create greater uncertainty than is realistic.

By mapping the semivariogram $\gamma(h)$ values in all directions and up to half the range of the data, 6000 feet, the principle directions of continuity can be interpreted, see Figure 5. As can be seen in Figure 5, anisotropy is not pronounced. The principle directions can be identified by trial and error variogram calculations, and with the aid of ancillary data, and geologic knowledge.

If one was to contour the $\gamma(h)$ values, a slightly ellipsoidal pattern is interpreted. An estimate of the major direction of continuity could be made as North 45 degrees West. The slight anisotropy could be attributed to wind dependent deposition, which would be expected with airborne particulate emissions. A variogram was calculated and modeled in order to represent correlation of lead values in all directions. The direction of major continuity was chosen as North 45 West and the direction of minor continuity was set orthogonal to the major. The resulting directional variograms are shown in Figure 6.

As previously stated the data is bimodal, with a low frequency of high values of lead (PPM) existing in localized areas. There is a suspicion of separate spatial trends between

the high and low values. This would be expected if the low levels of lead were inherent to the soil or background levels while the higher concentrations were due to recent industrial processes. In order to isolate separate trends between the high and low values, three separate indicator variograms where created at the 0.1, 0.5 and 0.9 quantiles, see Figure 7.

There is a slight decrease of continuity with increasing concentration. The results are noisy due to the small number of transitions, especially in the high and low values. The assumption will be made for the rest of the study that the data is best modeled as one stationary group. Indicator techniques would have captured the separate spatial relationships of the background and anomalous lead levels. In a more comprehensive study, greater efforts could be made to refine the variograms and isolate the differing spatial correlation of the separate lead levels.

In Isaaks' Masters Thesis, he explains that the band of low Pb concentrations running southeast is caused by a flood plain. Periodic flooding has interfered with the soil through erosion, deposition and leaching. Perhaps the floodway is interfering with a clear understanding of the lead correlation. Variogram calculation inherently assumes stationarity. If the sample space was broken up into regions, for example, flood channel and remainder, this may be a more correct stationarity assumption. The current method blends the spatial correlation of two distinct sample spaces. This problem is minimized by the fact that the flood plain is sparsely sampled in comparison with the remainder of the space. Therefore for this study, the entire space will be assumed to be stationary for (1) brevity and (2) practicality, since there is not enough information within the flood plain to explore separate statistics.

Kriging for Map Making

Kriging provides a smooth representation of the Pb concentrations, which is appropriate for trend visualization (Deutsch and Journel, 1998). Kriging also provides a measure of local uncertainty or estimation variance, which is based on the correlation model and the location of the estimate relative to the conditioning data. Moreover, kriging permits cross validation which helps choose between different implementation options.

Ordinary and simple kriging were performed with the previously shown variograms. The results were cross validated, which is a method of excluding data, estimating the data with adjacent data and then comparing the estimate to the actual value. The results are best displayed in scatter plots of the estimated values versus the actual data. The comparison between simple kriging and ordinary kriging is shown in Figure 8.

There is no appreciable difference in the cross validation of the two methods. If one were interested in estimates at the edge of the study area, where the data is sparse, then ordinary kriging would be preferable. The advantage of ordinary kriging is its moving window re-estimation of the mean at every point, while simple kriging assumes the mean is stationary. In this data set high values occur in the over- sampled center and low values occur in the under-sampled margins. Simple kriging would be appropriate if this trend was first modeled and removed before kriging or the estimates which are made in the sparsely sampled space were clipped. In this example, no attempt was made to remove the trend and this results in an artifact in the margins of the simple kriging map. The estimates increase towards the margins. Figure 9 shows this simple kriging artifact along the margin.

Ordinary kriging better handles the sparse data regions; the estimates approach the local mean, which is re-estimated at every point, from the conditioning. This can be seen in Figure 10.

The kriging variance is a valuable indication of the uncertainty associated with each kriging estimate. Kriging variance identifies areas of high uncertainty. This assessment may indicate the need for more data, or the need to clip estimates from the map, see Figure 11.

Indicator kriging is another useful form of kriging. Indicator kriging provides a least squares estimate of the conditional cumulative distribution function (ccdf) at a series of thresholds. The advantage of indicator kriging is that no assumptions are made concerning the local distributions of uncertainty. With careful selection of descriptive thresholds, indicator kriging can describe the distribution of uncertainty (represented by the ccdf) at every node.

In this study indicator kriging is used to obtain the percent probability of lead concentrations greater than a hypothetical regulatory threshold. The same results can be found with multiple realizations from sequential indicator simulation and sequential Gaussian simulation, which will be discussed in the next section.

The distributions of uncertainty can be used to develop various meaningful pointwise statistics. For example, the probability of contamination (assumed as lead concentrations over 500 PPM) may be inferred directly. The resulting probabilities of exceeding this threshold are shown in Figure 12.

From the simple and indicator kriging maps one may see basic data trends. There are several clusters of high estimates away from the main deposit that could be investigated further. In most cases these clusters of high estimates are caused by a single datum. Some of these isolated data are the result of a subsequent deposition of lead from small industry in the area. Also, the flood channel can be clearly seen in all of the kriging representations. An advantage in indicator kriging is that output data can be rapidly post processed to provide information such as probability of exceeding various contaminant thresholds. (See Figure 13)

Simulation for Mapping and Uncertainty Assessment

Sequential simulation is the sequential application of kriging with the addition of the missing variance (the kriging variance). Simulation ensures that the one and two point statistics of the conditioning data are honored. This study uses both sequential Gaussian simulation (SGSIM) and sequential indicator simulation (SISIM) in parallel in order to compare and cross validate.

Simulation assesses global and pointwise uncertainty through multiple realizations. The distribution of realizations at one point represents local uncertainty. These local distributions of uncertainty may be used to build accuracy plots, which are a method of model verification. Multiple realizations of the entire data space represent global or joint uncertainty.

Both Gaussian and indicator simulation provide a model of joint uncertainty, although indicator simulation does not require any assumptions about the shape of the pointwise distributions. The Gaussian assumption allows for the determination of any distribution

by two parameters, mean and variance. While the Gaussian assumption simplifies the algorithm, it also assumes all distributions are multigaussian. The Gaussian method has the benefit of exactly reproducing the one-point statistic of the conditioning data. The indicator techniques, as mentioned build the distributions by estimating probabilities at specific thresholds. Indicator methods can be thought of as non-parametric (Journel, 1989).

The accuracy plot is a scatter plot of the fraction of true values within a symmetrical confidence interval (about the median) of the local uncertainty distributions to the size fraction of the interval. Ideally a good model should have the fraction of true points within the interval equal to the fraction of the distribution in the interval. On the accuracy plot points on the line $(y = x)$ are precise and accurate, points above the line are accurate but not precise and points below the line are neither accurate nor precise. This tool estimates how well a model is estimating the local distributions of uncertainty (Deutsch, 1999).

Ordinary and simple kriging in the Gaussian approach were compared with accuracy plots. The accuracy plots show that the ordinary kriging yields a wider distribution of uncertainty than necessary. The model is generally accurate and not precise. The simple kriging results in more precise distributions of uncertainty, see Figure 14. It should be noted that accuracy plots and cross validation only check a model at the conditioning data. These methods would not explicitly indicate the model's performance in the sparsely sampled margin.

From Figure 14, it can be seen that simple kriging is the better estimator in this setting. By considering the similarity in the cross validation and the better performance in the accuracy test, simple kriging was chosen to be the method of estimating. Although, simple kriging is accepted with the condition that it should not be used to estimate in the regions in which the kriging variance is high (due to the trend not being modeled). It should be noted that even simulation with simple kriging will not adequately model at the margins since the kriging estimates are the means of the local distributions of uncertainty, which will be too high.

The two maps in Figure 15 show two equally possible SGSIM realizations from the data, and our variogram model. By producing multiple realizations insight is gained into the joint uncertainty of the data, that is, the uncertainty over a group estimates.

One method of accessing joint uncertainty would be to compile many equal probable simulations and then to build a distribution of a parameter of interest, such as area above a threshold concentration. Then one could look at the worst, best and most probable extent of contamination.

Simulations are also used for probability maps, e-type estimates and confidence intervals. These methods display pointwise uncertainty. From one hundred and one simulations, probability maps were produced, see Figure 16.

There are important differences between the probability maps derived through sequential Gaussian simulation and indicator simulation. There is a difference in the continuity of different lead levels. The indicator simulation was processed with the same variogram model used at all thresholds, which forced the continuity to be independent of magnitude. The Gaussian method relies on only one variogram and a normal transform. The Gaussian function is a maximum entropy function, and it minimizes unwarranted structural properties, which leads to maximum disconnectedness of extreme values.

In the lead study the Gaussian representation may be appropriate since the lone outliers

represent isolated levels independent of the smelter, such as recent pollution from small industry. In the indicator example, with forced equal spatial continuity among all magnitudes, the results include artifacts of large contamination areas caused by the isolated outliers.

The series of Gaussian simulations were used to develop e-type and confidence interval maps, see Figure 17. These maps were calculated from the local distributions of uncertainty. The e-type estimate is the mean of the realizations at every point. The e-type estimates are similar to the kriging results because the mean of the posterior distribution is the minimal square error estimate, which is the kriging estimate.

In the area of contamination investigation, volume support plays an important role. The volume support size must be chosen prior to investigating a site. The variance between samples increases as the size of the sample decreases. The simulation model should be at an appropriate support size. To modify the simulated results to represent different volumes, or to scale up, one must consider the resulting loss of variance. This effect can be clearly seen in the Figure 18.

There is a smoothing effect in larger blocks due to the averaging of high and low values. There is a shift of all estimates toward the mean as the support size increases. This mechanism causes a decrease in variance with increasing support size.

Decision Making

Decision making follows the model of uncertainty in the contaminant concentrations. Various decision making tools are available. For example, the probability of exceeding a threshold could be directly used to assign areas to remediate. A decision could be made to clean areas with a greater than an assigned probability of exceeding a threshold. After this area has been designated, false-negative images (probability of wrongly calling an area clean) of the remaining area could be used to indicate areas in which more information should be gathered. The development of this method would be straightforward once the model had been developed.

A more complicated and optimum method would be to apply an economically derived decision making model to the model. A comprehensive study could be conducted on the cost of clean up for a unit area and the cost of leaving a contaminated unit area untreated. For the purpose of demonstration a simplified model is utilized in this study. The model developed for this study assumes a constant remediation cost and a linear negligence penalty with an intercept at the origin and a slope such that the cost of leaving a block with a concentration equal to 650 PPM is the same as the cost of remediation, see Figure 19. It should be noted that the derivation of a comprehensive economic model would require knowledge of the prevailing legal and political situation.

For the purpose of demonstrating this decision making process, the decision model was to minimize expected cost. Figure 20 shows the resulting output from the SelectPB program.

The image on the right has had all selected areas with a kriging variance above 0.75 removed. This was necessary because the simple kriging assigns higher estimates at the edges of the data space (since the appropriate trend was not removed previously) and the increasing uncertainty causes inflated local realizations that approach the global mean. The decision making model is sensitive to realizations with high local estimates and will

incorrectly mark regions for remediation contrary to a background understanding of the region.

In order to make informed decisions in the high variance areas additional sampling could be carried out and one could even set up a decision making problem on whether to collect more data. The benefit would be less uncertainty and the cost would be the sampling cost.

Complicated economic decision making models could be applied to the stochastic model with very little effort. The result is the economic optimum remediation limits, based on the decision model and the available data. In addition, procedures could be developed to transform the blocks into more manageable areas.

This paper has outlined the procedures, which would allow for the interpreting of spatial attributes of a environmental site and the techniques for applying simulated results and uncertainty directly into a decision making model. All steps were completed by using the Geostatistical Software Library (GSLIB).

References

[1] C. V. Deutsch. *Geostatistical Reservoir Modeling*. Edmonton, Alberta, 1999.

[2] C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, New York, 2nd edition, 1998.

[3] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.

[4] E. H. Isaaks. Risk qualified mapping for hazardous waste sites: A case study in distribution free geostatistics. Master's thesis, Stanford University, Stanford, CA, 1984.

[5] E. H. Isaaks and R. M. Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.

[6] A. G. Journel. *Fundamentals of Geostatistics in Five Lessons*. Volume 8 Short Course in Geology. American Geophysical Union, Washington, D. C., 1989.

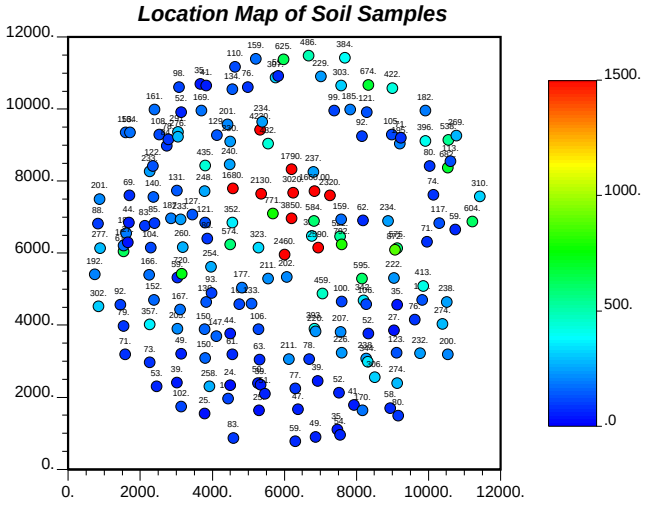

Figure 1: Location Map of Pb Data

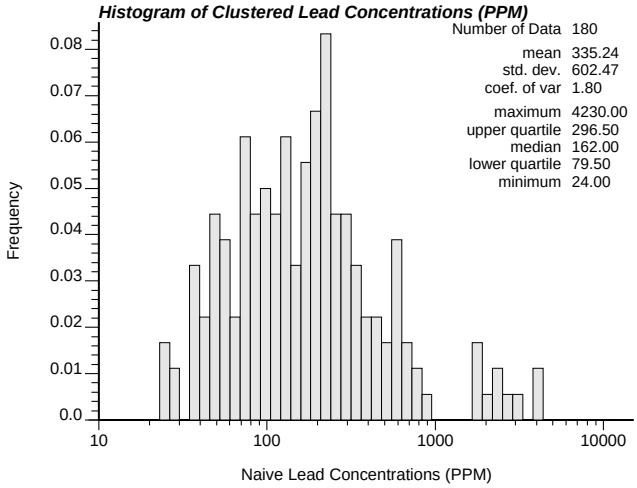

Figure 2: Equal weighted histogram of Pb data: Indicates a bimodal distribution with log-Gaussian shape for the < 1000 PPM data

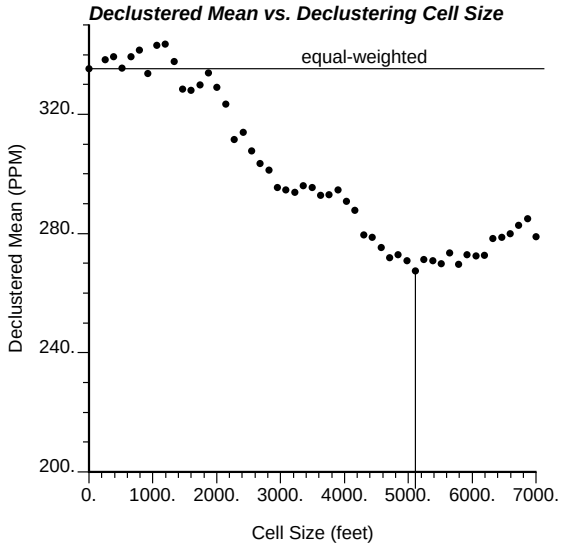

Figure 3: Declustered mean versus cell size: There is a bias toward sampling in high Pb area. The cell size that provides the lowest declustered mean is used to calculate the declustering weights.

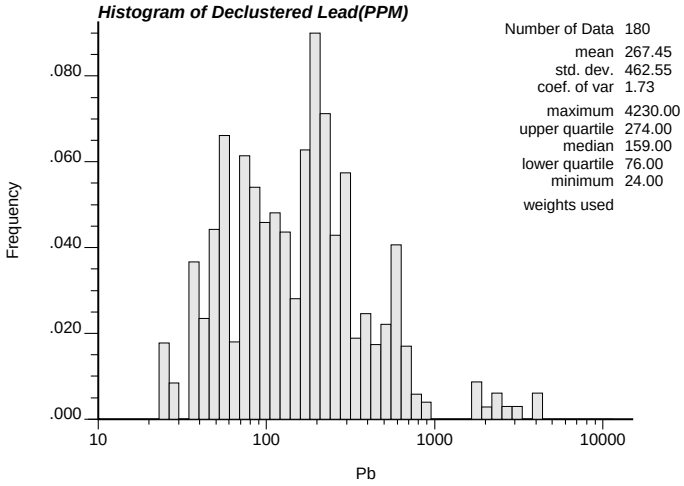

Figure 4: Declustered histogram: The histogram resulting from the application of declustering weights. This histogram is representative of the study area.

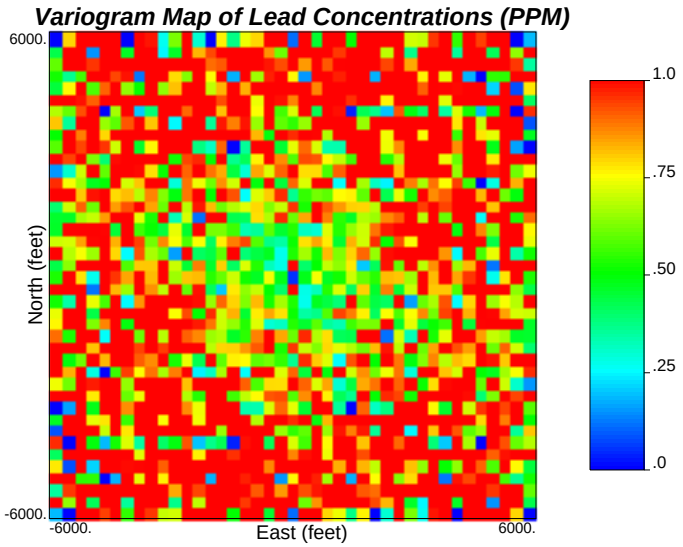

Figure 5: The $\gamma(h)$ Map of the soil samples: Indicates the correlation for all directions and for distances up to half the size of area interest.

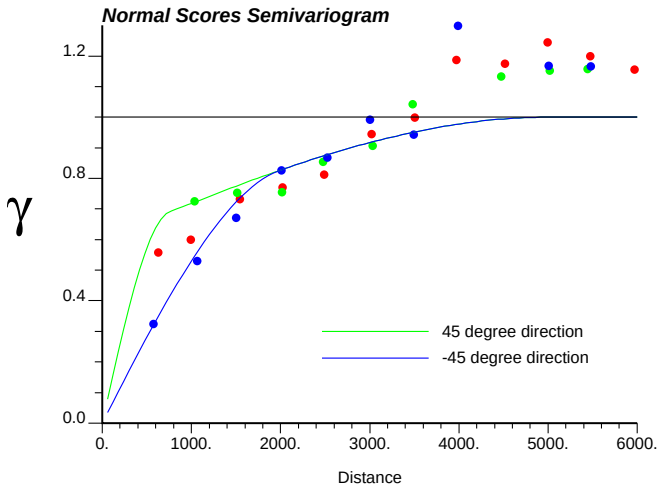

Figure 6: Semi-variograms of the soil samples: There is geometric anisotropy.

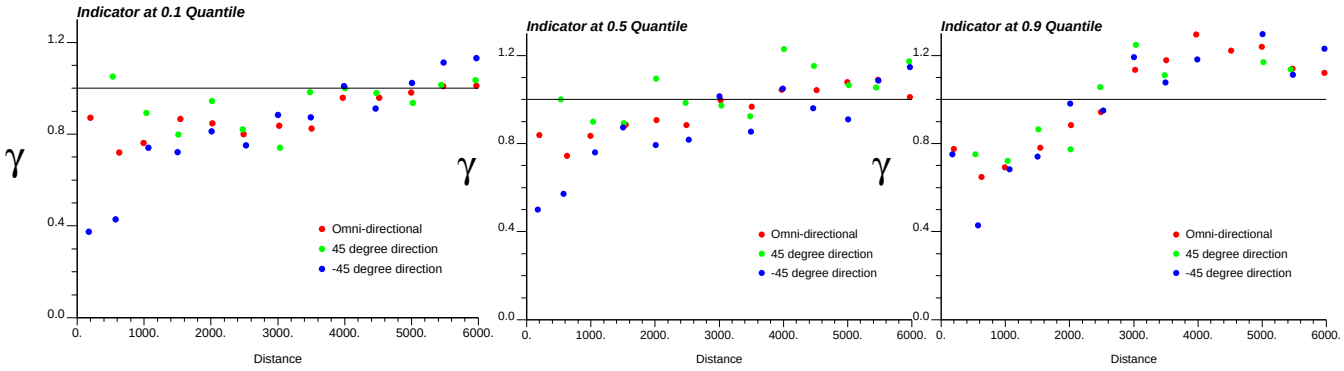

Figure 7: Indicator variograms of soil samples: A decrease in range of correlation is seen as the Pb values increase.

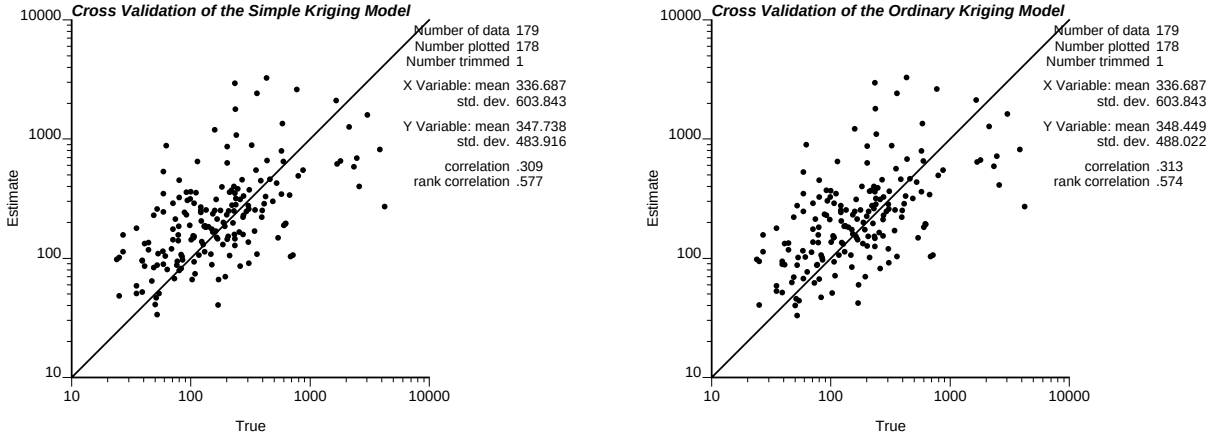

Figure 8: Cross validation: Results show no difference in estimate accuracy at the locations of the Pb data.

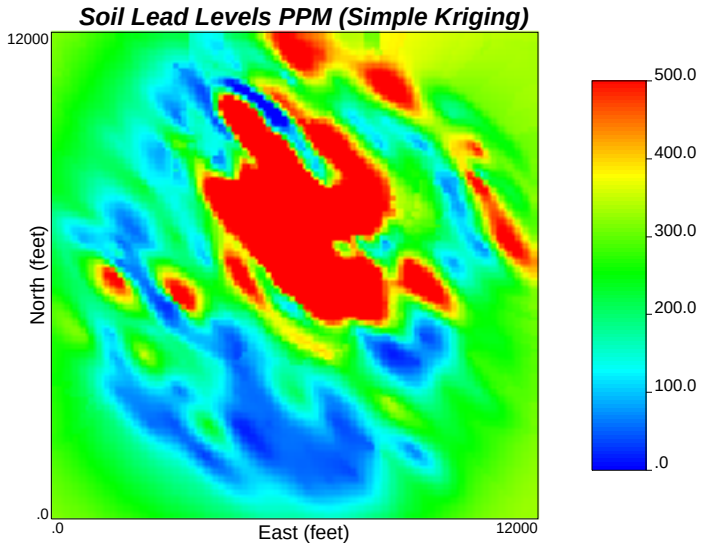

Figure 9: Simple kriging of the soil samples: The estimates in the margins approach the global mean because of data scarcity.

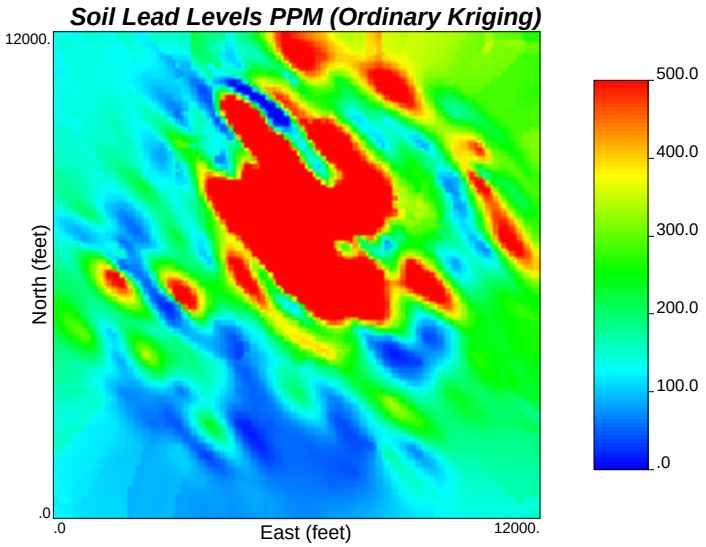

Figure 10: Ordinary kriging of the soil samples: The estimates in the margins approach the local mean.

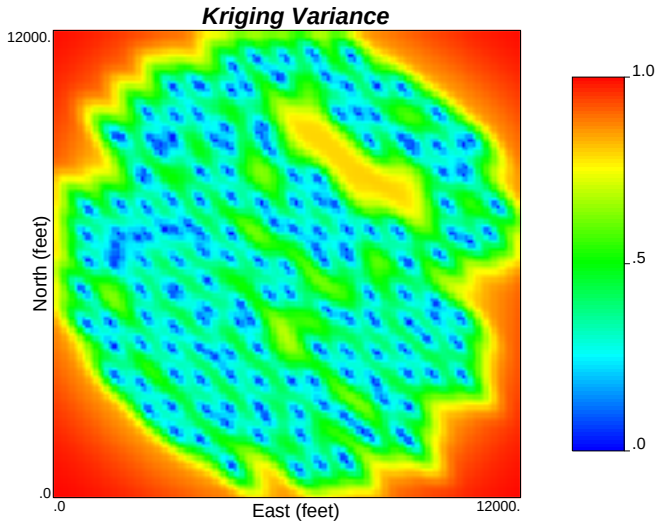

Figure 11: Kriging variance: A good measure of estimate certainty, based on the correlation model.

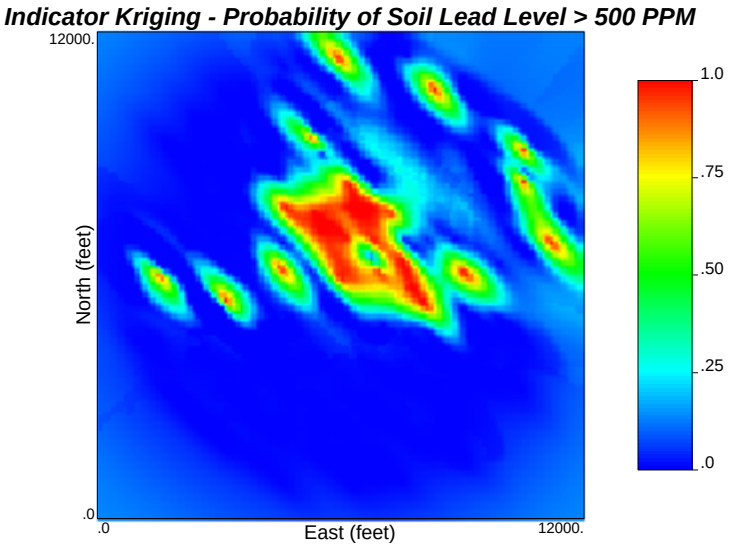

Figure 12: Probability of exceeding 500 PPM: Calculated directly from indicator kriging derived local distributions of uncertainty.

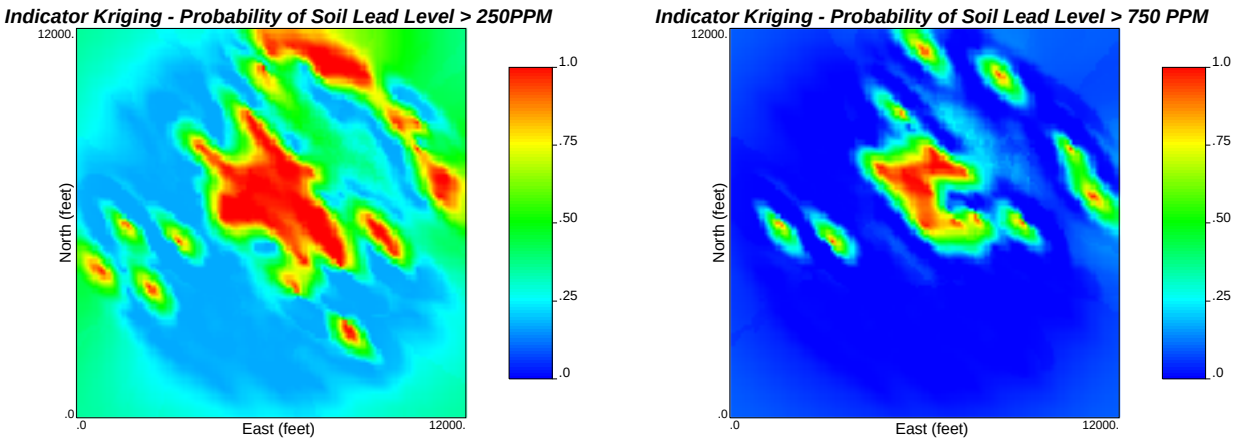

Figure 13: Probability maps with varying limits: Various thresholds recalculated from local distributions of uncertainty.

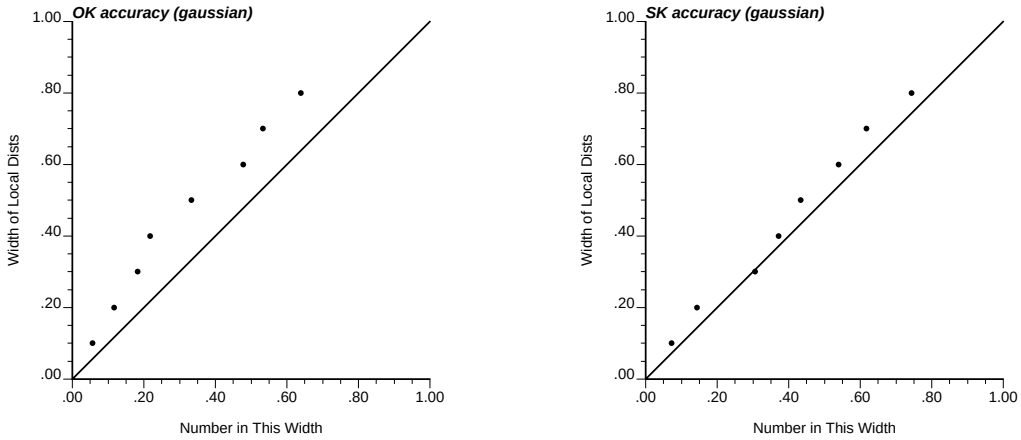

Figure 14: The accuracy plots: The local distributions of uncertainty derived from simple kriging are more precise than the ordinary kriging distributions.

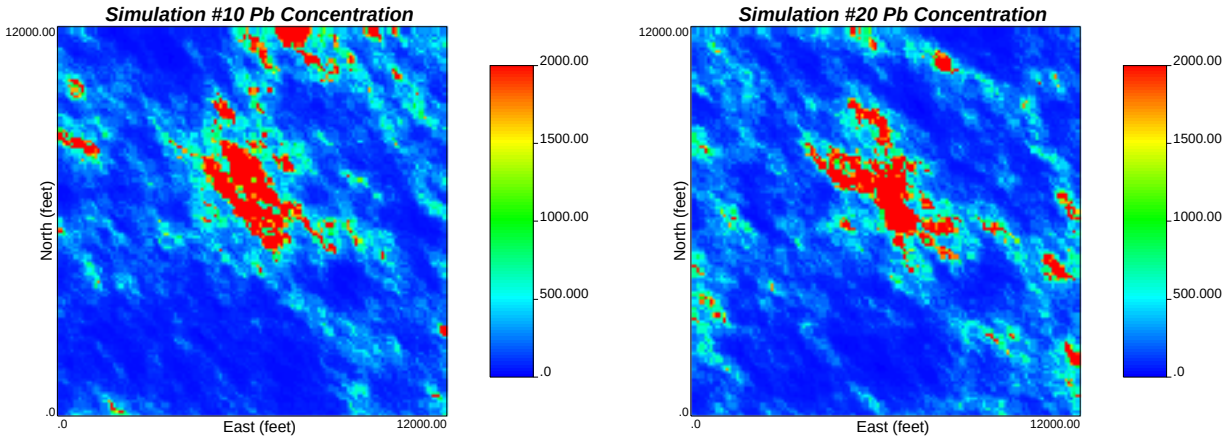

Figure 15: Simulated realizations: Equal probable realizations are used to model model joint uncertainty.

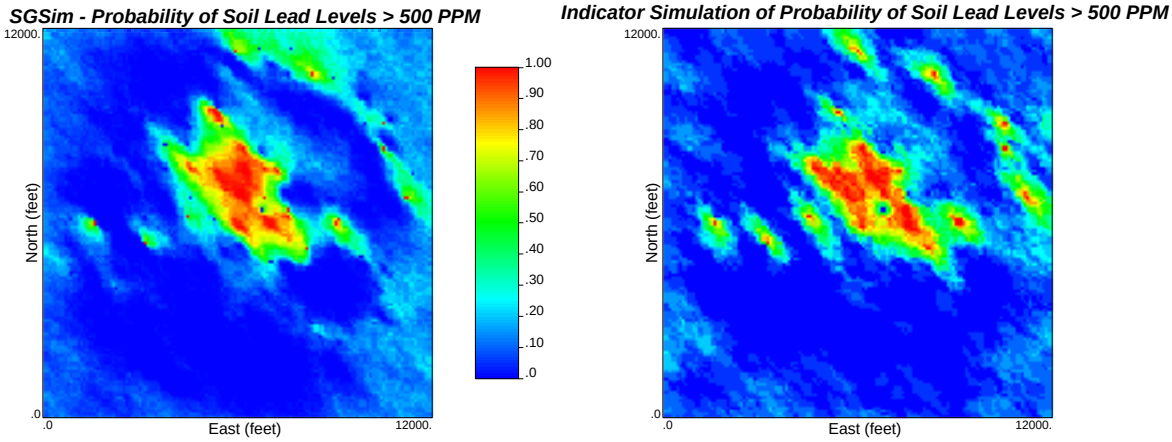

Figure 16: Probability maps from SGSIM and SISIM: The Gaussian assumption leads to the maximum disconnectedness of the extremes. Indicator simulation is not restrained by this assumption.

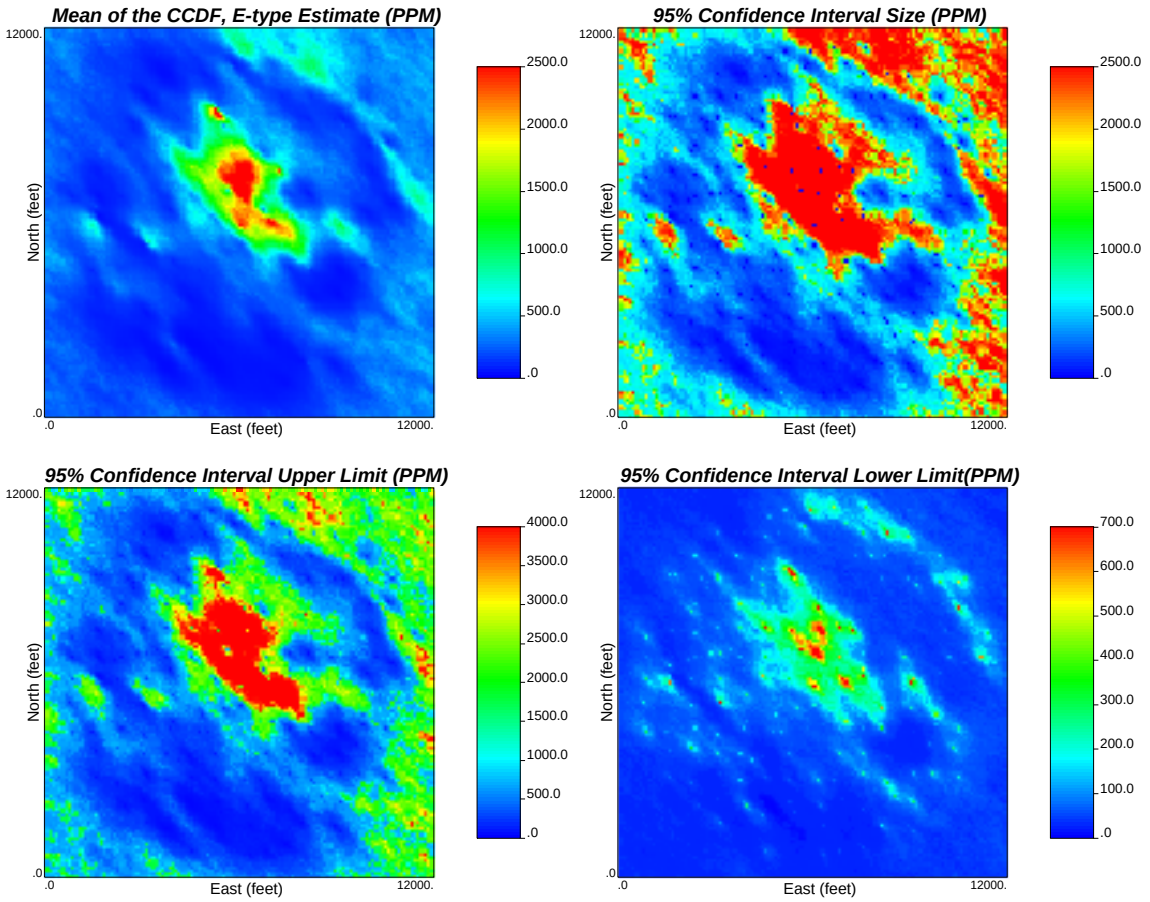

Figure 17: E-type and 95% confidence interval: Other useful statistics from local distributions of uncertainty.

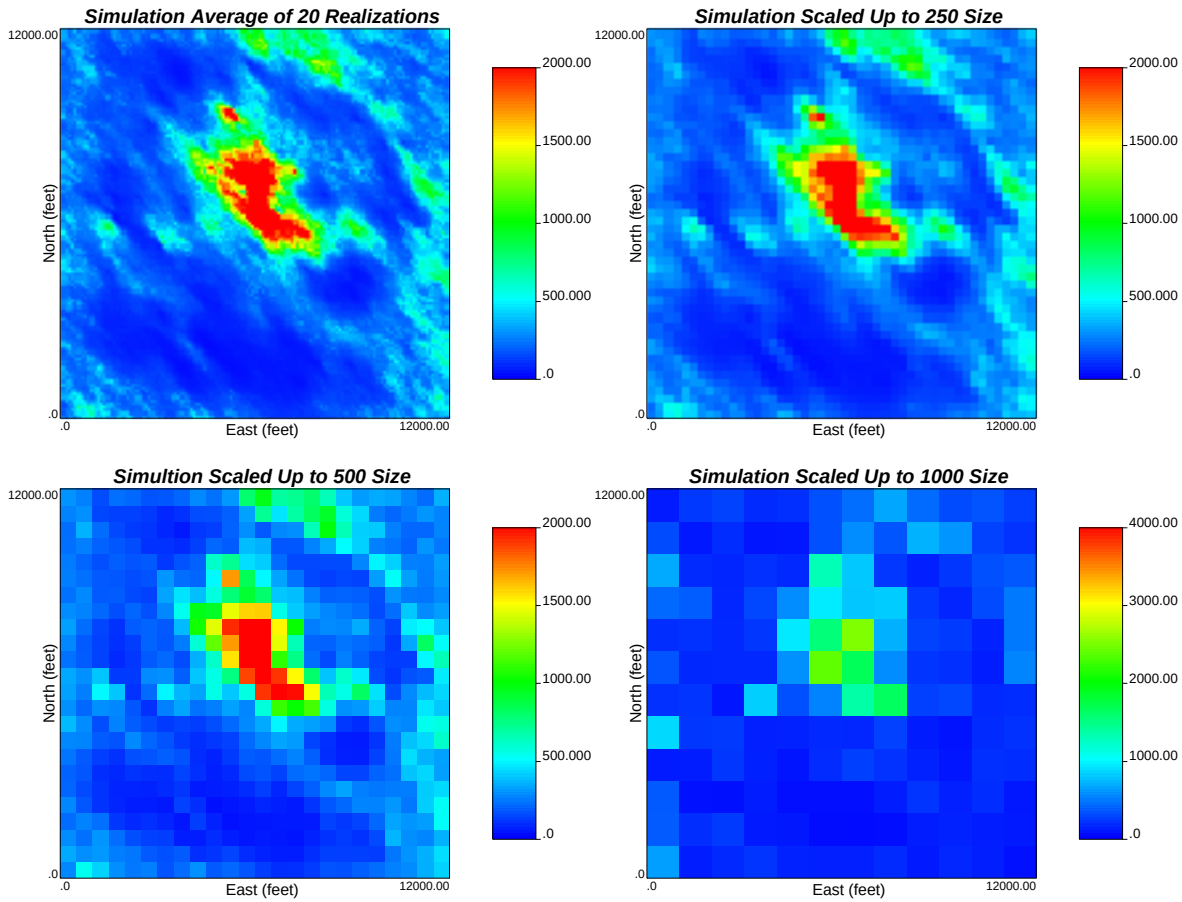

Figure 18: Volume support: Scaling up is straight forward when the linearity assumption is appropriate.

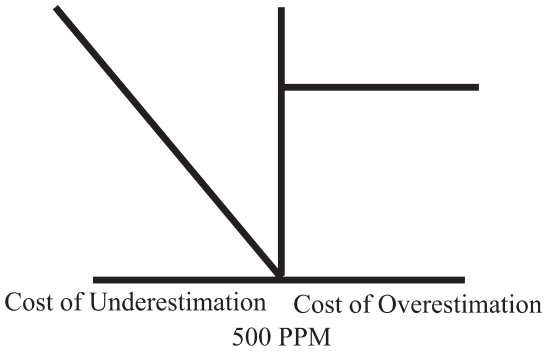

Figure 19: Economic model: The simple economic model used in the decision making exercise.

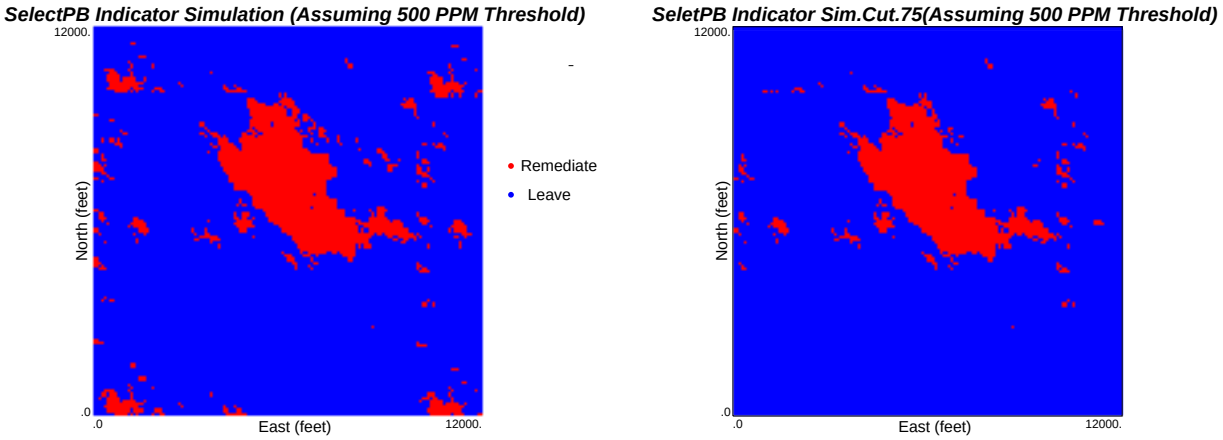

Figure 20: Decision making model results: The optimum remediation plan based on uncertainty and the decision making model.