Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

Uncertainty management through stochastic modelling

Mario E. Rossi
GeoSystems International
479 Cascadita Terrace,
Milpitas, CA, 95035, USA
Email: mrossi@geosysint.com

Abstract

Quantifying uncertainty when dealing with predictions about water contamination, both for volume and for contained pollutants, has been recognized as critical in a thorough contamination assessment study.

The reason is that errors in the water sampling process, characterizations of the geometry of the contaminant plume or body, interpolation of the sampled values across a contaminated field, and characterizations of the physical processes that predict fate and transport are always present. It is not possible to obtain a realistic assessment of contamination without accepting first, and modelling later, the intrinsic uncertainty carried over the whole process.

This paper proposes a rational stochastic methodology, based on geostatistical conditional simulations, that allows assessment and modelling of the ever-present uncertainty in contaminant modelling. This assessment is then translated into risk levels, allowing for a decision-making process that is based on levels of uncertainty. This process utilizes the concept of Minimum Desirable Loss to make decisions. The paper describes an example drawn from a real-life case study in northern California, USA.

1.0 Introduction

Characterization of water or groundwater contamination, as a normal component of typical Feasibility Study/Remedial Investigation (FS/RI) work in the United States, requires a major time-consuming effort and economic commitment. This investment is orientated towards initially mitigating and in the future minimizing

Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

negative consequences of contamination on human population and/or prevailing ecosystem. The results of such mitigating effort is measured in terms of human health or eco-health risk assessment, and, in the case of groundwater, it follows mostly the U.S. Environmental Protection Agency's (EPA) guidelines.

Given the imprecise information handled, the overall uncertainty in the processes involved, and the often socio-political conflicts of interest that arise, it is reasonable to require that health and ecosystem risk assessments carry a measure of the level of uncertainty involved. This uncertainty is called in this paper "technical risk", as opposed to health or ecosystem risk.

There are multiple factors that contribute to the "technical" risk, including, but not limited, to:

- Errors in initial sample collection; the uncertainties here are related to the sampling techniques used, and the relationship between sampling methods and the heterogeneity of the contaminants being sampled, Gy[1].

- Errors related to sample preparation and analysis;

- Errors stemming from inadequate handling of data, including computerized mismanagement; these errors can be minimized using appropriate data handling and data quality objectives protocols.

- Errors related to overly simplified or inappropriate data evaluation and modelling techniques, statistical analysis, etc. These include ignoring or overlooking significant sources of spatial and natural data variability.

- Errors related to the level of contaminant being sampled and analyzed. In most instances, the EPA's required acceptable contaminant levels are very close to the Method Detection Limits (MDLs) for each particular contaminant. This introduces a significant technical challenge for the laboratories to analyze the samples, since the accuracy and precision of the analysis is questionable near the MDL. A concept called Practical Quantitation Limit (PQL) has been proposed to overcome this problem, see for example Gibbons[2].

Many of the "traditional" statistical techniques used to analyze groundwater contamination data is based on stringent assumptions about statistical distributions, lack of spatial correlation, independence among the samples considered, etc. These are all requirements of Gaussian-based statistical techniques often used, for example including Analysis of Variance (ANOVA), Cochran's approximation to the Behrens-Fisher t-test, etc., see among others Gibbons[2], Gilbert[3], EPA[4].

This paper proposes the use of geostatistical conditional simulations and the concept of Loss Functions as tools to adequately model the uncertainty involved in water pollution assessment. In essence spatial stochastic simulations, these tools have developed in recent years into the preferred toolbox for uncertainty modelling and spatial data analysis in several fields, most notably mining, petroleum, and environmental applications.

Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

2.0 Geostatistical Conditional Simulations

For a general background of the theory of geostatistical conditional simulations the reader is referred to Goovaerts[5] or Journel[6]. The simulations are used to build models that reproduce the full histogram and modeled measures of spatial continuity of the original, conditioning data. Therefore, they honor the spatial characteristics of the contaminant plume or any other spatial variable as represented by the three-dimensional sample (conditioning) data. In addition, it is possible to extend the use of these spatial statistical tools to the time dimension, see for example Rossi and Posa[7].

By honoring the histogram, the model correctly represents the proportion of high and low values, the mean, the variance, and other spatial statistical characteristics of the data. By honoring the semi variogram (or any other more robust measure of spatial continuity), it correctly portrays the spatial complexity of the contaminant plume, and the connectivity of low and high contaminant zones. These are fundamental variables that need to be considered in order to improve predictions and diminish predictive uncertainty. When several simulated images are obtained, then it can be said that a model of uncertainty has been obtained.

Conditional simulations are built on fine grids, as fine as possible given the hardware available, so that they correspond to approximately the support size of the original samples. A reasonable grid for the simulation could be 3m by 3m by 3m, as is used in the case study presented here. The vertical resolution of the grid should be a function of the support data, typically the size of the screened interval. Larger grid sizes may still be used sometimes because of the amount of computer time and hard disk space involved. Such a fine grid is possible because of the random aspects of the algorithms used (conditional simulations are Monte-Carlo-based techniques). In building a conditional simulation model, many of the conditions and requirements of linear and non-linear estimations apply, most importantly regarding stationarity decisions. Shifts in the geologic or hydrogeologic setting requires the separation of the data into different populations, as would contamination barriers or boundaries. Thorough knowledge of the behavior of extreme and outlier values in the sampled population is required. Issues such as limiting the maximum simulated grade should be carefully considered.

The simulation method itself should be decided based on the type of contamination, quantity and quality of available samples, possibility of using "soft" or fuzzy information, and the desired output. The first decision is whether to use a parametric or non-parametric approach. Examples of each are the Sequential Gaussian (Isaaks[8]) and Sequential Indicator (Alabert[9]) simulations. The latter is more complicated, based on multiple indicator kriging techniques (Journel[6]), and requires definition of several indicator cutoffs. The former is simpler and quicker, although more restrictive in its basic assumptions. For contaminant concentrations, any available geologic or hydrogeologic criteria can and should be used. These can take the form of "soft" or imprecise information, including prior probabilities in a Bayesian sense (Alabert[9]). As with any estimation exercise, spatial continuity measures, such as correlograms, should be estimated and modeled. This is a particularly important point, since often it is claimed that there

Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

is insufficient samples to obtain a good estimate of the spatial continuity of contaminants. Potentially a weak link in spatial stochastic modelling, there are a number of alternative that ultimately depend on the experience of the scientist, prior knowledge about the site, and other parameters that are often considered. These include censoring data (or what to do with non-detects, usually a high proportion of the population), allowed minimum and maximum data values and simulated values, number of conditioning data to be used, search distances, anisotropies, etc.

When a number of these conditional simulations have been run and checked, then, for each cell defined in the grid, there are a set of equi-probable (by construction) contaminant values available. These contaminant are interpreted to describe the model of uncertainty for that cell, generally arranged as a posterior cumulative conditional probability curve. Preferably, a large number of simulations are needed to describe this curve better; however, and due to practical limitations, a much smaller number, perhaps as small as 10 simulations, can be used as an initial approximation. When there is significant conditioning information, these simulated values for each cell will not vary much, meaning that the probable cell value is known with a good degree of certainty. The opposite occurs when the cell has few samples nearby.

The stochastic model of uncertainty developed for each point within the area of interest can be described as (Journel[6]):

$$F(z;\underline{x}|(n)) = Prob\ \{Z(\underline{x}) \leq z|(n),\ \alpha=1,\ldots,n\} \qquad (1)$$

Here $F(z;\underline{x}|(n))$ represents the cumulative conditional distribution frequency curve for each vector $\underline{x}$ of the simulated grid, obtained using the (n), $\alpha=1,\ldots,n$ conditioning samples, and it provides the probability of that point in the grid of being above (or below) any contaminant value z.

For the case study described here, sequential Gaussian simulations (SGS) were used (Isaaks [8]). SGS is based on a multiGaussian Random Function model assumption for the spatial process being simulated. The original data is first transformed into a Gaussian distribution using a Normal Scores transform, also known as "anamorphosis"; this process transforms any sample distribution into a univariate Gaussian distribution. Then, tests are performed to validate the required multiGaussian assumption made. In practice, the check is based on the relationship between indicator semi-variograms obtained from the transformed Gaussian samples and the theoretical Gaussian semi-variogram; this checks only the adequacy of the bivariate Gaussian distribution assumption, for specific details see Goovaerts[5]. This bivariate check is a necessary but not a sufficient condition for multiGaussianity; in practice, however, and after verifying the adequacy of the bivariate assumption, then a multiGaussian assumption is applied.

3.0 Loss Functions

Final recommendations in FS/RIs and final remediation decisions are typically based directly or indirectly on health risk assessments, which in turn are based on estimates of contamination, $z^*(\underline{x})$. Since the true contaminated value at each

location is not known, an error can and will likely occur. The loss function $L(e)$ (Journel[6], Rossi[10]) is a mathematical expression that attaches an economical value (impact or loss) to each possible error, measured in, for example, dollars. By applying a loss function to a set of equiprobable simulated grade values (a conditional probability distribution, as obtained by conditional simulations, Equation (1)), then the expected conditional loss can be found by:

$$E\{L(z^* - Z) \mid (n)\} = \int_{-\infty}^{+\infty} L(z^* - z) \cdot dF(z; \underline{x} \mid (n)) \qquad (2)$$

The minimum expected loss can then be found by simply calculating the conditional expected loss for all possible values of the estimates, and retaining the estimate that minimizes the expected loss. As described in Isaaks[8], the expected conditional loss is a step function whose value depends on the assumed costs of each bad decision, and the relative of costs of mis-classification. This implies that the expected conditional loss depends only on the *classification* of the estimate $z^*(\underline{x})$, not on the estimated value itself.

The Loss Function thus quantifies the consequences of false positives and false negatives, weighs the relative impact of each, the probability of each, and then suggests the minimum cost solution. For example, the loss incurred when a contaminated area is remediated when in fact it should not be is a direct function of the remediation costs incurred. If the same area is not remediated, and it fact it should have been, then the cost of the mistake will be a function of the consequences of that bad decision; if significant loss of health, quality of life, or life itself results, then the cost could be assumed to be infinite. Figure 1 shows an hypothetical Loss Function, where a false positive error incurs in unnecessary remediation costs, increasing linearly with the magnitude or the error, while a false negative error causes the Loss to increase exponentially with the absolute value of the error, and for certain errors it becomes infinite.

4.0 Case Study

The case study summarized here corresponds to an area within an old U.S. Army installation in northern California. There is a groundwater monitoring program in place, from which quarterly samples have been obtained since 1991 from more than 40 wells.

In one of the areas of concern in the installation, there was a need to determine the likelihood of metals contamination in groundwater. The aquifer's background values for a number of dissolved metals were obtained from upgradient wells and wells located laterally to groundwater flow. Significant spatial and temporal variations were observed in the background wells, yet they were used as reference for determining action levels for the monitoring program. The current case study describes the work done on Dissolved Selenium (Se), used here to demonstrate the application of the geostatistical assessment of technical risk. The Action Level (AL) chosen for Dissolved Selenium was 25μg/l.

Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

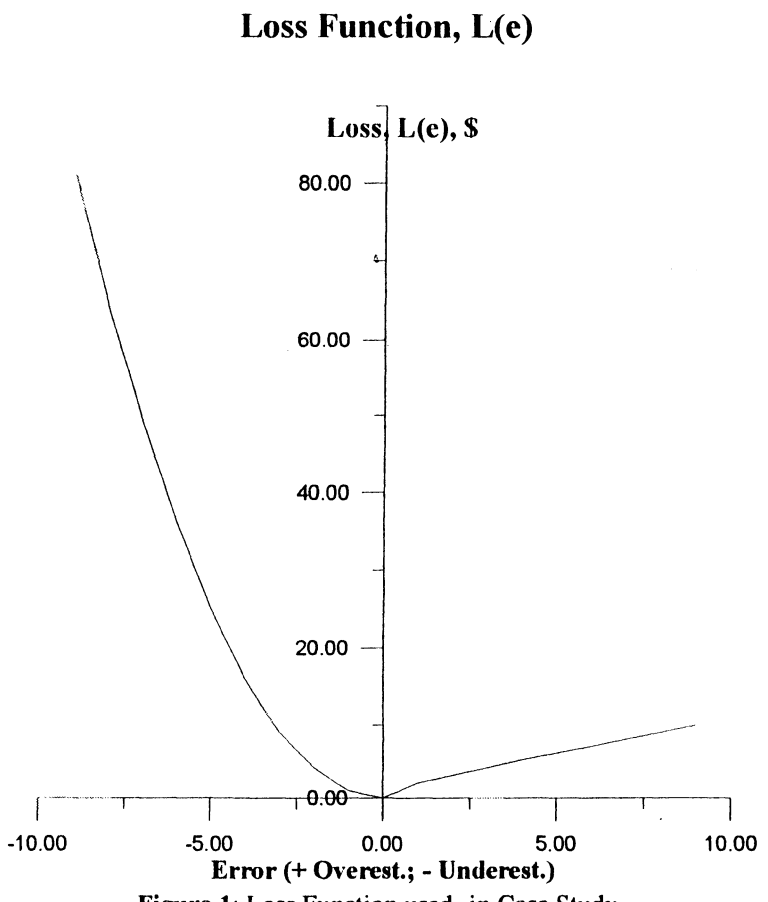

Figure 1: Loss Function used in Case Study.

Having determined the Action Level value for Se, then the next issue is determining when new samples have exceeded the AL, and how significant that is.

4.1 Developing the Uncertainty Model

The uncertainty model was developed using Sequential Gaussian conditional simulations (SGS), as described above. In order to develop the simulations, the following steps were completed:

a) Initial exploratory statistical data analysis was performed over the whole database. This included the background wells and the monitoring wells, as a unique population and also as separate background and "monitored" population.

b) The spatial continuity of Dissolved Selenium was modelled using correlograms, see for example Goovaerts [5]. The study was done in two

Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

and three dimensions, initially assuming a two-dimensional spatial grid; in this case the aquifer itself has little vertical extent. The third dimension was added using data from different quarters, i.e., time was the third dimension. At the end, a different correlogram model for each sampled quarter was obtained from all the data; from the background wells, only a omnidirectional (in three dimensions) sample correlogram was modelled. The study presented here uses the omnidirectional background wells correlogram. The model showed a nugget effect of 55% of the total variance, plus two structures, the first one with a range of approximately 12m, the second one with a range of approximately 30m.

c) The simulation grid for each quarter was defined on a 3x3m cell, and 10 SGS simulations were performed. The ten Se values obtained provide the model of uncertainty of Equation (1).

To visualize this model of uncertainty, multiple options are available. In this case it is most interesting to obtain the probability of each cell of exceeding the Action Level. That probability is described by Equation (1), where z=25µg/l. Figure 2 shows a greyscale map of those probabilities. Note that there is a north-east trending area of higher probability of exceeding (+60%) the 25µg/l. This is consistent with known possible sources, and the groundwater flow in the area. In addition, there are zones of higher probability of exceeding the Action Level value in the southern half of the mapped region. This area required further investigation, since there were fewer wells located within.

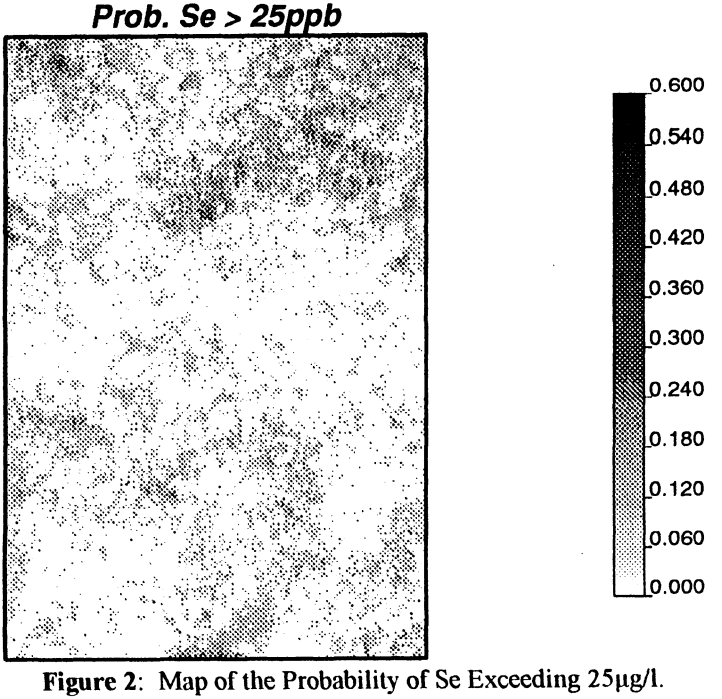

Figure 2: Map of the Probability of Se Exceeding 25µg/l.

4.2 Developing the Loss Function

The Loss Function applied to evaluate risk in this case was based on the following Equation:

$$Loss = Actual\ Cost - Potential\ Cost \quad (3)$$

The costs associated with each type of error are depicted in Figure 1, where it is shown that a false negative is significantly more costly than a false positive. This is because it is assumed that the costs of leaving contaminated areas without cleaning increases exponentially with the error magnitude, while the cost of cleaning groundwater in areas that did not needed to increases linearly with the error magnitude. Table 1 illustrates the form of the Loss Function used.

TABLE 1: Loss Function, in US$.

	True Se Value is Below AL	**True Se Value is Above Se AL**
Estimated Se Value is Below AL	A11 = 0	A13 = {c*|e|^2 - {a + b*|e|}
Estimated Se Value is Above AL	A12 = {a + b*|e|} - 0	A14 = 0

Applying the Loss Function of Table 1, it is possible to find out the actual economic losses for each simulated value in each cell of the study area. Compositing these losses according to Equation (2) result in a mapped "optimal loss" classification. These map is shown in Figure 3 for specific economic conditions, for the pre-determined Dissolved Se Action Level, 25µg/l. Contrary to Figure 2, here darker areas represent lower levels of risk.

Note that, depending on how linear the Loss Function in Table 1 is, there will be a more or less significant deviation of Figure 3 from what is depicted in Figure 2. In other words, the associated risk for each type and level of error may not be (and in general is not) directly proportional to the probability of making the error. Only if the Loss Function of Table 1 is linear that would be the case, and the Expected Value (or arithmetic average) of the conditional simulations would then provide directly a measure of risk; in this case, Figure 3 would be redundant.

4.3 Extension to Fate-and-Transport Problems

What has been described in Sections 4.1 and 4.2 correspond to a static condition, that is, the technical risk involved of making wrong decisions at a certain point in time. If there is a need to predict in the future how this risk would evolve under the predicted fate-and-transport conditions (usually modelled using a flow simulator), then it is immediate to extend these concepts, simply by repeating this process on a quarterly or annually basis.

Transactions on Ecology and the Environment vol 26, © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

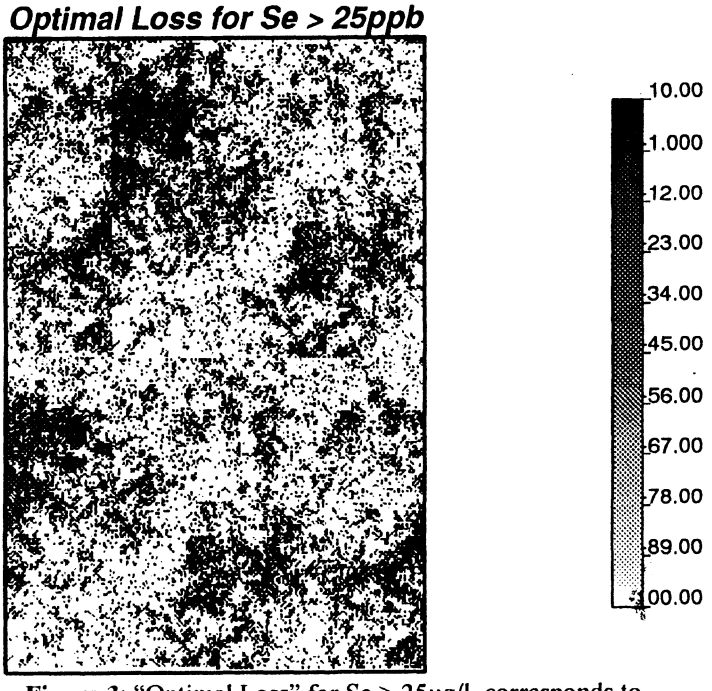

Figure 3: "Optimal Loss" for Se > 25μg/l, corresponds to Uncertainty Model of Figure 2 and Loss Function of Table 1.

Conclusions

When trying to model, make decisions, and eventually operate and monitor a groundwater remediation program, it is often difficult to accurately assess and predict a number of technical aspects of the problem. Most of these difficulties stem from intrinsic spatial and temporal variability, sampling inaccuracies, and procedural errors. These errors can and will cause mistakes in the decision-making process, with sometimes sever consequences. A method has been proposed here whereby the modelled errors are incorporated into the technical risk evaluation process through the use of stochastic conditional simulations, interpreted as models of uncertainty.

These models of uncertainty are then used to evaluate the consequences of all possible mistakes through the use of Loss Functions; clearly, the quality of the final product will depend on the virtues of the model of uncertainty, and the accurate reflection of costs through the Loss Function.

A major advantage of this method is its flexibility with respect to assessing costs, since in the formulation of the Loss Function there can be several types of costs included, such as the actual remediation operating costs, costs stemming from health risk assessments, more speculative socio-political costs, etc. The price to be paid for such flexibility is the more mathematically involved methodology, and the responsibility that results from actually stating more of the hidden assumptions that usually occur in any risk assessment process.

References

1. Gy, P.M., *Sampling of Particulate Materials: Theory and Practice*, 2nd. Ed., Elsevier Amsterdam, 1982.

2. Gibbons, R.D., *Statistical Methods for Groundwater Monitoring*, J. Wiley & Sons, Inc., New York, 1994.

3. Gilbert, R.O., *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York, 1987.

4. U.S. Environmental Protection Agency, *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities*, Interim Final Guidance, April 1989.

5. Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

6. Journel, A.G., *Fundamentals of Geostatistics in Five Lessons*, Stanford Center for Reservoir Forecasting, Stanford University, Stanford, CA, USA, 1988.

7. Rossi, M.E., and Posa, D., *Geostatistical Modeling of Dissolved Oxygen Distribution in Estuarine Systems*, Environmental Science and Technology, Vol. 25, No 3, pp. 474-481, January 1991.

8. Isaaks, E.H., *The Application of Monte Carlo methods to the Analysis of Spatially Correlated Data*, PhD Thesis, Stanford university, Stanford, CA, 1990.

9. Alabert, F.G., *Stochastic Imaging of Spatial Distributions using Hard and Soft Data*, MSc Thesis, Stanford University, Stanford, CA, 1987.

10. Rossi, M.E., *Grade Control using Conditional Simulations and Economic Optimization*, Proc. of the Third Annual International Association of Mathematical Geology Conference, V. Pawlosky G., ed., Barcelona, pp.1003-1008, September 1997.