

## 1.0 Description of the Smart Sampling Methodology

Smart Sampling couples economic decision analysis with geostatistics to guide site characterization and remediation efforts. The process begins with an initial set of data, collected during site characterization, which indicate the need for remediation. These data are used to design an initial excavation plan based on a map describing the *likelihood* or *probability* that the contaminant concentration exceeds an specified threshold (for example, define the excavation area such that the probability of exceeding the threshold after excavation is 5% or less).

Prior to implementing the initial excavation plan, a cost analysis is performed to determine the potential value of collecting additional data. The objective is to minimize the total project cost,  $C_{TOTAL}$ , defined as

$$C_{TOTAL} = \sum_i [C_{SAMPLING} + C_{REMEDICATION} + C_{FAILURE}]_i + C_{DESIGN}$$

for all remediation units  $i = 1, n$ .

$C_{SAMPLING}$  includes the cost of collecting, handling and obtaining the laboratory analytical results for the samples.

$C_{REMEDICATION}$  may include excavation, transportation and treatment or disposal costs.

$C_{FAILURE}$  is the cost associated with leaving behind contamination exceeding the regulatory compliance threshold. This may include the cost of bringing the equipment back to the site to remediate these areas, the cost of the additional remediation and verification sampling, and, possibly, fines (e.g., levied by the regulatory agency for negligence).

$C_{DESIGN}$  is the cost of designing the remediation plan based on the sampling results (whether performed in the classical manner, e.g., via hand contouring the concentration distribution, or using the Smart Sampling approach of geostatistical simulation, there is a cost associated with this phase of the remediation effort).

The advantage of the Smart Sampling approach is its ability to identify areas where the *expected cost of failure* is greater than the cost of taking additional samples. Smart Sampling techniques are used to not only identify these areas, but to determine the number and optimal locations within these areas where additional samples should be taken.

If additional samples are taken, the probability of exceedence map is revised, a new excavation plan is developed and the remediation costs are recomputed. This procedure is repeated (theoretically) until there is no significant potential reduction in the total cost. At this point, the last (final) remediation map is used as the design for the excavation. In practice, probably at most, only one additional sampling event would be carried out.

The primary component of the Smart Sampling methodology involves a technique called geostatistical simulation. Geostatistics is the study of data that exhibit spatial correlation. It is well known among earth scientists, that samples of contaminants, sediments, porosity, etc. tend to be more similar when the samples are closely spaced and less similar as the distance between the sample locations increases. Geostatistics provides a means of quantifying this spatial correlation

and also provides adaptations to classical regression techniques to take advantage of spatial correlation (Isaaks and Srivastava, 1989).

## 1.1 Variogram Estimation and Modeling

At the heart of geostatistical analysis is the measurement and modeling of the degree and type of spatial correlation. These operations are generally accomplished through the calculation of an experimental variogram and then the fitting of a model to that calculated variogram. The variogram is essentially a model of the increase in variability between sample locations as the distance between the samples increases. Although the terms variogram and semivariogram are often used interchangeably, the models of spatial variability used in the analyses are actually semivariograms, defined as

$$\gamma(\mathbf{x}, \mathbf{h}) = \frac{1}{2} E [(Z(\mathbf{x}) - Z(\mathbf{x}-\mathbf{h}))^2]$$

where E denotes expectation, Z is the variable of interest (e.g., Pu<sup>238</sup> concentration),  $\mathbf{x}$  is the spatial location and  $\mathbf{h}$  is the separation vector. The variogram equation is similar to the calculation of variance in classical statistics. In the calculation of variance, the mean is subtracted from each data point (each Z( $\mathbf{x}$ )), the differences are squared and then summed. In the variogram equation, the difference is taken between each data point and a data point a distance (h) away, the differences are squared and then summed. Thus, it measures the “variance of the differences” between field variables separated by a distance (h). It makes sense then, that the variogram value should be small for samples which are in close proximity and larger for samples further apart.

Variogram estimates typically exhibit somewhat ‘erratic’ behavior, i.e., empirical estimates are non-unique and exhibit a lot of scatter. Consequently, a smooth, analytical model is usually fit through the estimates as shown in Figure 1.

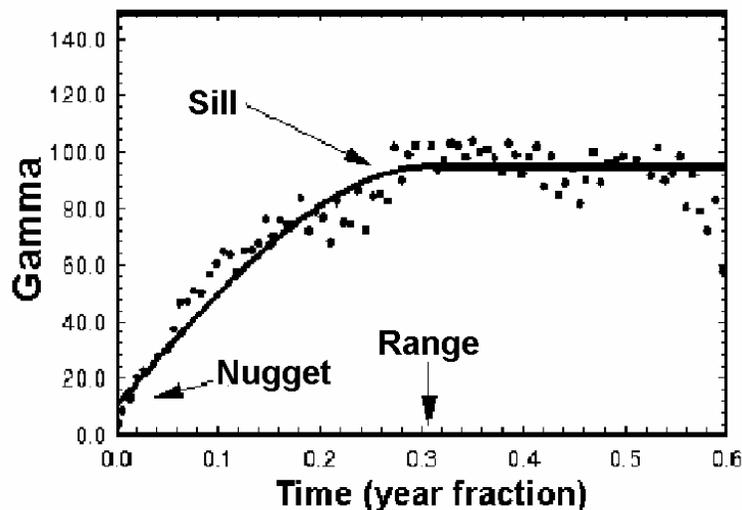


Figure 1 - Example of experimental semivariograms (points) overlain by analytical model (curve)

The variogram model is used to characterize spatial variability, or the manner in which field variables change from point to point in space. There are several important features to a variogram model. The “range” is the distance at which the variogram model becomes parallel with the x-axis (where the variables become completely uncorrelated). Often the term “correlation scale” is used to describe the distance over which the field variables are ‘significantly’ correlated. This is typically the distance represented by a parameter of the variogram model which results in the model reaching the “sill” value. The sill is the value of the variogram at the range distance. Theoretically, the value of the sill is equal to the variance of the data set. Intuitively, as the distance between sample locations decreases to zero, the amount of variability between samples should also decrease to zero. Often, variogram models do not go through the origin of the graph. In these cases, the variogram model y-intercept is known as the “nugget” value. The nugget effect generally represents sampling variability at a scale smaller than the smallest sample spacing and may also represent lack of precision in the measurements, or sampling error.

Variogram modeling is a fundamental component of geostatistical estimation and simulation techniques. Estimation is a linear interpolation technique which provides the “best estimate” in the minimum variance sense (i.e., it is developed by minimizing the variance between estimated and true values). Simulation, on the other hand, is designed to reproduce the variability in the sample data, while honoring the measured values. Both techniques are used to assign property values to unsampled locations within the site domain.

## **1.2 Random Field Estimation and Simulation**

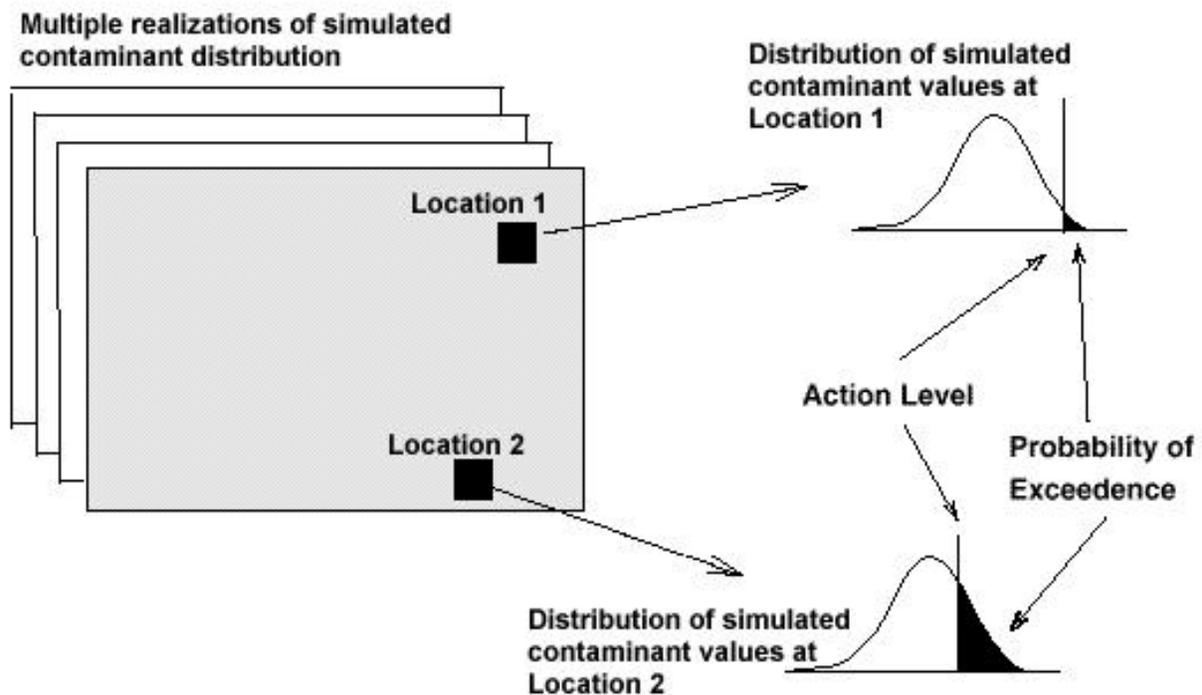
In the geostatistical approach, the variable of interest (e.g., contaminant concentration) is considered to be a spatial stochastic process, or random field. Estimation techniques are used to derive estimates of the variable at unsampled locations. Estimation techniques commonly used in the earth sciences include inverse distance-squared techniques, nearest neighbor polygons and kriging. All of these techniques can be classified as methods of interpolation, i.e., they all produce estimates of unknown concentrations at unsampled locations by interpolating from known values at other locations.

Of the three estimation techniques mentioned, only kriging exploits the model of spatial correlation derived through variogram modeling. Kriging is essentially the process of determining the expected value of concentration at a given location by calculating a weighted least-squares mean of the surrounding data points. The weights used in the least-squares estimation are calculated by using the model of spatial correlation given by the variogram. These weights account for the distance each data point is away from the location being estimated and the clustering of the data points (for example, a number of points all close to each other provide redundant information concerning concentration at the point being estimated and are weighted less than a single point an equal distance away in another direction). Because kriging is a minimum variance-based estimation technique, the concentration map derived from kriging will contain less variability than the actual sample data. This smoothing effect will ensure that the minimum and maximum values of the estimated map do not fall outside the bounds of the

minimum and maximum of the sample data. A kriged estimate of concentration along a transect would look very similar to the best guess drawn by hand.

Whereas estimation provides a *single best estimate* of the concentration at each location, simulation provides *multiple estimates* of the concentration field, all of which honor the available data. As noted previously, simulation is designed to reproduce the measured level of variability in the sample data for each map of the concentration field while honoring the measured data at their respective locations. Each equiprobable map of the contaminant distribution generated via simulation is termed a 'realization'. Based on the limited samples available, it is not possible to determine the exact underlying reality. Therefore, any one of the realizations are as equally plausible as any other. These multiple plausible renditions of the contamination distribution provide a means by which uncertainty in the contamination maps can be addressed. Specifically, the probability of exceeding a specified concentration level at any location can be estimated.

This is accomplished by, at each location, collecting all the estimated values of concentration across all realizations and developing a complementary cumulative distribution function (CCDF) from the ensemble of values. This concept is illustrated schematically in Figure 2.



**Figure 2 - Schematic of Complimentary Cumulative Distribution Function (CCDF) construction from an ensemble of concentration maps**

The kriged or simulated realizations can be processed to estimate the total amount of contaminant across the site. This calculation is accomplished by assuming that each measurement of concentration is representative of a volume or block of soil. Consequently, the kriged or simulated concentrations are also representative of the same volume of soil. For the purpose of

estimating the total amount of contaminant in a block, the contaminant is assumed to be evenly distributed throughout the block. The size of the blocks must be chosen with some consideration for the correlation scale of the process, i.e., the blocks should be sufficiently small to adequately represent the variability expressed by the variogram model, say, about  $1/5^{\text{th}}$  the range.

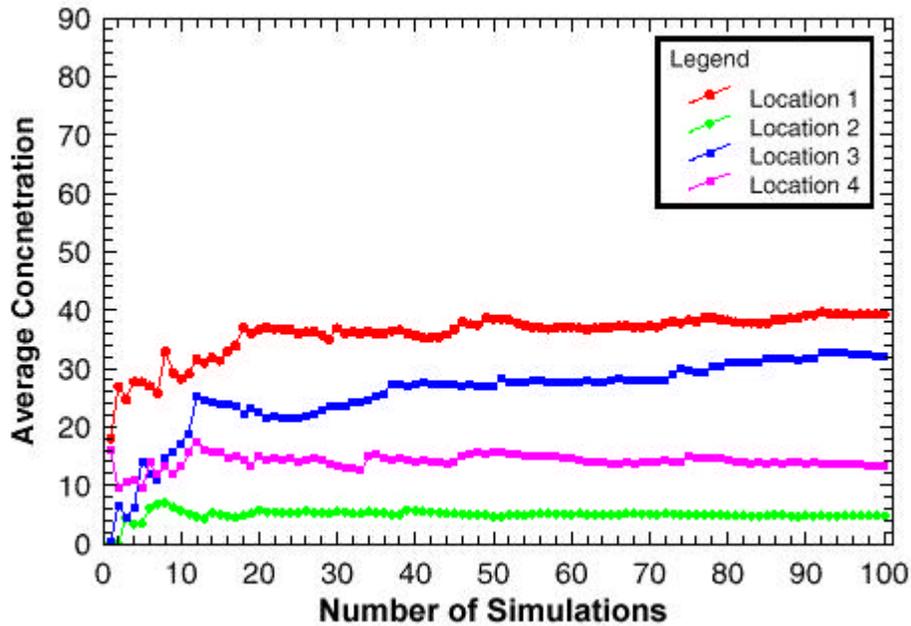
### 4.3 Choosing the Number of Simulations

One question that must be considered is how many realizations are required to adequately characterize the probability distribution (pdf) of contaminant concentrations (across all realizations) at each location. If the number of realizations is too few, the pdfs will not be meaningful, and it is unlikely that the appropriate remediation decision will be made (to remediate or not to remediate at each location).

The concept of a representative elementary volume (REV) discussed in ground water hydrology literature (Bear, 1972) can be used to illustrate how this question might be addressed. In the field of ground water hydrology, an REV is a volume of the porous media that is large enough to average out any discrete local effects of heterogeneity in the media, but small enough to not be affected by deterministic trends in the distribution of the property. A classic example of an REV is given by measuring porosity in a sandstone with a larger and larger sample. At the smallest sample size the porosity will be either 0.0 or 1.0 depending on whether or not the sampler encounters a sand grain or a void space. As the size of the sample increases, the fluctuations in the porosity value will diminish as the sampler begins to sample a representative volume (REV) of the porous media.

The REV concept can be applied to analyzing geostatistical simulations by noting that, at the basis of geostatistics, the ergodic hypothesis allows for the replacement of a spatially infinite sample by a large number of spatially finite, stochastically generated images. While this hypothesis is a basic tenet of geostatistics, the practical question of "what is large?" or at least "large enough" remains to be answered. In order to answer this question, the spatial averaging of the REV in ground water hydrology is replaced by a running average of the concentration (at selected locations) across an ever increasing number of realizations (see Figure 3, below). The calculated statistic is no longer a REV, but is now termed the representative number of realizations (RNR). If the average value stabilizes to a constant, representative value, then the number of realizations is deemed to be large enough.

The locations where the stability in the mean is to be evaluated are chosen with respect to the conditioning data. Generally, one should choose locations where kriging variance is high, i.e., where the distance from conditioning points is largest and, consequently, where the uncertainty is greatest. If stability at these locations is achieved, this provides assurance that at all other locations similar or better convergence will be observed. Within these areas, it may be advisable to select several points which cover the full range of concentrations, say low, midway and high concentration values. Additionally, one can plot the stability in the second moment, or the variance, of the distributions at these locations; for this statistic to become completely stabilized, a larger number of simulations will be required.



**Figure 3 - Example of the convergence of the mean of the probability distribution of contaminant concentrations with increasing number of realizations**

#### 4.4 Probability and Remediation Mapping

Probability mapping is defined here as the use of geostatistical simulation to estimate the probability of exceeding a specified level of a contaminant concentration at each location in the simulation domain. For example, suppose 100 realizations of a contaminant distribution are generated. If the action level is 25 pCi/g and 30 of the 100 realizations show concentrations greater than 25 at a given location, the probability of exceeding the action level at that location is 0.30, or 30%. The mean value at each location is the estimate derived from kriging.

Thus, probability maps can be generated by choosing an action level and determining, for each location in the simulation domain, the probability of exceeding that action level as described above. The resulting map could be contoured and areas having higher probabilities would be targeted for further sampling and remediation.

Once the probability map is generated, the ‘risk tolerance level’ is used to develop the remedial action plan. For example, if the regulating agency stipulates that no more than 5% of the samples shall exceed the target action level, then all “remediation units” (simulation blocks where the exceedence probability is estimated) exceeding 5% will be slated for remediation.

The remediation design could be developed as, for example, a green and red color map where red means yes, excavate and green means no, do not excavate. Superimposed on this map would be the grid lines from the sampling grid, as a means of ‘registering’ or orienting the remediation map. Additionally, the required depth of excavation can be printed in the center of each cell on the

remediation map. Such remediation maps can be produced for each 50' canal section on a single 8½ × 11 page, so it can be used directly in the field.

#### 4.5 Cost Versus Risk Decision Curves

Determining an 'acceptable' level of risk may depend, to some degree, on the cost consequences of reducing the risks. For example, if the regulatory agency and/or stakeholders are highly risk averse, then the only solution may be to remediate the entire site, possibly to an excessive degree, in order to ensure that *all* areas have a large margin of safety. On the other hand, the concept of *cost-effective* remediation is predicated on the realization that some finite level of risk must be accepted.

In order to aid the decision makers in this process, a set of cost versus risk curves can be developed from the remediation maps. Several different remediation maps can be developed, each with a different level of risk associated with it. By plotting the total remediation against its corresponding risk level, a curve can be developed showing the cost versus risk relationship. Several of these curves can then be generated, each for a different action level.

Examples of such curves are shown in Figure 4. These curves present an effective way to display the relationship between cost and the probability of remediation failure. These curves can also provide a focal point for discussion between the site owners, regulators and stakeholders concerning action levels, costs and risks associated with various future land-use scenarios.

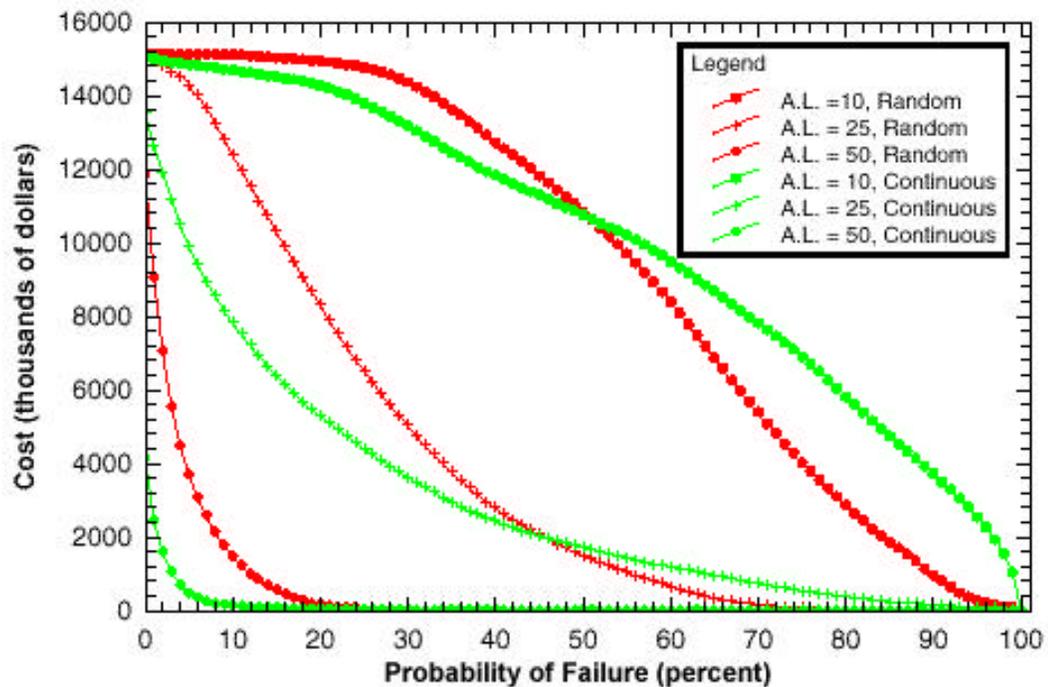


Figure 4 - Example of cost versus risk (probability of failure) curves.

## 4.6 Determining the Worth of Additional Sampling

Prior to finalizing the remediation design based on the samples collected and the probability maps generated, it is prudent to determine if acquisition of additional samples is likely to further reduce the total remediation costs. The key, as noted in Section 3, is identifying if there are any areas where the expected cost of failure is greater than the cost of taking additional samples. This expected cost of failure,  $E[C_{fail_i}]$ , is defined as

$$E[C_{fail_i}] = P_{fail_i} \times C_{fail_i}$$

where  $P_{fail_i}$  is the probability of failure in remediation unit “i” and  $C_{fail_i}$  is the cost of failure for that remediation unit (a remediation unit is a simulation gridblock). The cost associated with failure to achieve the required remedial action level includes the cost of bringing the equipment back to the site, excavating the area, transportation, treatment or disposal, and the cost of performing final verification sampling. It could, potentially, also include the cost of fines levied by the regulatory agency. The probability of failure is determined from the probability maps generated through geostatistical simulation and the action level being considered.

Thus, the expected cost of failure can be evaluated for zero, one, two ... up to N potential additional samples. Correspondingly, the cost of taking these additional samples can also be evaluated, and will likely be a linear function of the number of samples, being generally a fixed cost per sample. (Insert discussion about simulating the outcome and reestimating the excavation area as in the James and Gorelick paper.) Such a graph can be a valuable decision tool for the site operator by providing a defensible means of making cost-effective decisions about the remediation efforts as they are taking place. In the next section, we describe how these potential additional sampling locations can be determined.

## 4.7 Determining the Locations and Quantity of Additional Samples

The placement of additional samples should be optimized such that the information concerning a remediation decision is maximized. A number of techniques have been proposed for optimal location of future samples (Burgess, et al., 1981; Englund and Heravi, 1994; Kyriakidis, 1996) with reduction of kriging variance being the most popular (see Barnes, 1989 and Olea, 1984 for a review of kriging variances as a means of sample optimization).

In the past few years, decision-based sample optimization has proven to be superior to kriging variance for locating in-fill samples. This result is not surprising given that kriging variance is based solely on the data locations, not on the data values. By incorporating the data values of the initial sampling into the optimization through consideration of the action level, the amount of information gained by the additional samples can be maximized.

Several techniques have been proposed for incorporating the action level into optimizing the placement of additional samples. The first technique is an intuitive approach suggested by Rautman et al. (1994) that simply targets the locations with median probability of exceeding the

action level (probabilities near 0.5). This approach is termed the "median probability" (MP) technique.

Another approach involves targeting the locations with median probability of exceedence and also accounting for the variability between simulations at those locations. One variant of this approach is implemented in the OPTMAS program (Knowlton, *et al.*, 1995) where a single additional sample is located along the median probability of exceedence contour line at the location of the highest simulation standard deviation. An extension of this approach is the "weighted standard deviation" (WSD) technique. Here, the simulation standard deviation at any location is multiplied by a weight between 0.0 and 1.0. The weight is a function of the probability of exceeding the action level. For locations with a probability of exceedence equal to 0.5, the weight is 1.0. The value of the weight tails off to 0.0 as the probability of failure approaches both 0.0 and 1.0.

A third technique is the "reference uncertainty" (RU) technique defined as

$$R(\mathbf{x}) = (C_{0.75} - C_{0.25}) / (1 + |C_{0.5} - C_{AL}|)$$

where C is concentration, the subscripts indicate the quantile of the distribution at location  $\mathbf{x}$ , and AL is the action level. The reference uncertainty value will increase both as the variability of the simulated concentrations at a location increases and as the difference between the simulated values and the action level decreases. The idea behind a reference uncertainty is presented in Kyriakidis (1996), where it was used to determine the remediation panels with the greatest uncertainty of exceeding the action level. These three techniques all incorporate the action level into determining locations for additional sampling.

All of the techniques discussed for determining the optimal locations of additional samples produce a continuous distribution of values from 1 to N, where N is equal to the number remediation "panels" in the *simulation* grid (i.e., the number of simulated grid blocks). The locations within the simulation grid can then be ranked from the maximum (1) to the minimum (N) value of the uncertainty measure. The locations of additional samples are then typically chosen based on the ranks computed for each location.

In previous applications of the Smart Sampling methodology the locations of follow-up samples were determined strictly on the rank of each uncertainty measure at a location without considering the proximity of other potential follow-up samples. It is important to consider the locations of potential follow-up samples *relative to each other* because many are likely to be found in close proximity. For example, if the locations with the three highest ranks for additional sampling are all next to each other, then obtaining a sample at the location with the highest rank will also provide information on the other two potential sample locations, decreasing the uncertainty at all three locations; it would not be necessary or cost effective to collect additional samples at the other two locations. Consideration of the proximity as well as the rank among potential follow up samples will ensure the most effective spread of additional sampling locations across the domain.

One possibility for optimizing the amount of information provided by the ensemble of potential additional samples is to conduct a second ranking process which takes into account the proximity

of the next potential sampling location to all previously chosen (higher-ranked) locations. For example, the location with the highest RU is the first potential additional sample location. The next highest RU value in the simulated grid may be very close to the first one. In this case, say if the distance between the two potential sampling locations is less than the correlation scale of the process, skip this location and go on to the next highest RU value and continue until a high RU value is found outside the “correlation range” of all previously selected potential sample locations. Repetition of this process will result in a matrix similar to the uncertainty ranking matrix, representing the preferential order for selecting additional samples that takes into account the correlation among the sample values.

After the ordering of potential additional samples is determined, the expected cost of remediation failure at each of these locations can be computed and cumulated so the cost versus probability of failure curves can be generated as described in Section 4.6 above.