Optimizing Surface Soil Cleanup Designs

Introduction
This section will discuss design optimization in general and will present tables of designs that have been shown to be effective for certain cleanup scenarios.

A "cleanup design" is considered to include a combination of a sampling scheme and an explicit quantitative decision rule that, together, determine what soil, if any, will be cleaned.

To satisfy multiple stakeholders, including the EPA, the operator or responsible party, and the public, a design should meet a number of criteria:

- Remediation decisions should be adequately protective, based on reasonable exposure scenarios, risk models, and risk computation.
- The process should aim at minimizing the cleanup costs required to achieve the necessary level of protection.
- The approach should be statistically and geostatistically sound, and scientifically defensible.
- The approach should be simple. The design should be something that can be selected and implemented without special expertise in statistics or geostatistics.

Managing Uncertainty in Scientifically Defensible Decisions
Scientifically defensible soil cleanup requires the quantitative prediction of outcomes. We want to design a cleanup process to meet a specified objective, and know that it will work as designed.

Generally, uncertainty in the decision process means that the predicted outcome will take the form of a probability distribution, so it becomes necessary to take that into account when establishing cleanup objectives. To design an effective cleanup, we need quantitative models of all of the sources of variability that create uncertainty in the decision making process. Decision uncertainty can be managed by varying the number of samples and analytical measurements, by controlling sample preparation and analytical measurement errors, and by varying the decision rule to compensate for errors.

Spatial Variability and Area Averages
Just as air quality standards are expressed as time averages ("eight-hour averages" or "24-hour averages"), soil cleanup standards or target levels must be expressed in terms of area (or volume) averages. The time average in an air quality standard represents a volume or mass of air that is linked to an exposure of concern. Although an eight-hour average at an air monitoring station does not measure any individual's actual exposure, it provides a plausible estimate of exposures that are likely to occur. The linkage between exposure, risk, and a standard is essential to the defensibility of a standard.

Soil, like air, is best treated conceptually as a continuous medium. Although soil is made up of discrete particles, the characteristics of soil occur only in the aggregate. Soil is not a "population" of soil individuals that can be sampled statistically by selecting a random subset of individuals. A typical soil sample collected for site characterization or cleanup is likely to be a vertical cylindrical core a few centimeters in diameter. Assuming that the depth of the core is the same as the depth of the surface soil layer being investigated, we can consider the area represented by the core diameter to be a shorthand notation for the volume of soil in the sample. Similarly, any references to an "area average" in this section may be assumed to refer to the volume of soil in the surface soil layer within the specified area.

Every chemical concentration is an average by definition. All of the quality assurance procedures developed for sample handling, preparation, and analysis are designed to ensure that the analytical measurement of the final sample aliquot is very close to the true mean concentration of the original sample volume. Here we introduce the geostatistical concept of sample *support:* The mass of sample included within a volume specified by size, shape, and orientation, for which an average concentration is measured or estimated. The term can be used loosely to refer to the approximate scale or resolution of a measurement or estimate, but implies a strict geometric definition of scale or resolution. A support can be thought of as a moving window smoother that eliminates all of the variability at smaller supports. The term support applies not just to samples, but any area or volume over which an average is taken.

If multiple samples are taken in an area, the set of measurements can be considered as a statistical sample of the area. However, the population represented by the sample statistics is a hypothetical population: The set of all possible samples of the same support. The problem with statistical sampling of soils is that no support is unique. If one chooses a different support, such as a different core diameter, then the population represented by the statistical sample is different even though the sampled area remains the same. With one exception, all statistical parameters such as maximum, minimum, median, 95th percentile, standard deviation, and skewness are likely to change when the sample support changes. The arithmetic mean is the only statistical parameter that remains constant over an area regardless of the sample support. Smaller supports generally lead to more variable statistical distributions.

While a typical soil core might be 2 centimeters in diameter, it could just as easily be 2 millimeters, 20 centimeters, 2 meters, or 20 meters. There is no inherent theoretical reason to choose one over another. Each is equally valid. In practice, however, each is not equally useful. If we are concerned with locating contaminated areas of a particular size, sample supports that are too large will fail to resolve relevant detail, potentially missing areas of concern. On the other hand, sample supports that are too small will introduce irrelevant detail that obscures the larger pattern we seek.

Samples with large or small supports provide different information. It is critical in soil cleanup design to recognize the importance of spatial variability and scale, and to make necessary allowances for scale changes in the decision making process.

There are two ways to deal with spatial variability and support in soil cleanup design. One is to adopt the geostatistical framework, where the variogram provides a continuous quantitative model of spatial variability at all scales. In principle, this the best approach because the variogram leads to kriging estimation. Kriging uses the information available from a sample, in combination with other samples, to make estimates over a relatively large spatial area. By squeezing the most out of the data, geostatistics can use less of it, leading to more efficient sampling designs. The problem with the geostatistical approach is that it is complex and often difficult to implement without specialized and costly expertise. Thus the geostatistical approach may not always be worthwhile, especially for relatively small cleanup projects.

The alternative approach is to treat the problem as a classical statistical stratification exercise. The site is subdivided into spatial strata such as exposure units (EU's), which in turn may be stratified into remediation units (RU's) which are sampled at some sample support.   Spatial variability and support are dealt with by a simple nested variance model:

$$\mathrm{var}_{s:RU} + \mathrm{var}_{RU:EU} + \mathrm{var}_{EU:site} = \mathrm{var}_{s:site}$$

The variance of the samples within a remediation unit plus the variance of remediation units within an exposure unit plus the variance of exposure units within the site equals the total variance of samples within the site. The nested variance model breaks up the continuous variogram into a set of fixed supports. In this design process, variability within a stratum is assumed to be random. No interpolation is used. The arithmetic mean of random samples (or a composite of random samples) is used to estimate a stratum mean.   The mean of a larger stratum is the arithmetic mean of the smaller strata within it.

The nested variance approach leads to simple cook-book designs that do not require statistical or geostatistical analysis of the data during the cleanup process. This approach is used in developing the generic designs described below.

Cleanup Objectives – Area, Risk, and Confidence

Perhaps the most important decision in designing a cleanup operation is choosing the right objective. The obvious objective of a cleanup is to clean up contaminated soil. It is also obvious that we generally clean up contamination because it poses some sort of risk; so, risk management is at least an implicit objective. For design purposes, we need to be much more explicit. A cleanup process should remove some contamination and thus reduce risk. What is needed is a precise quantitative answer to "how clean is clean enough?"

The first step in answering that question is to specify a target threshold concentration: "We want to remove all of the soil with contaminant concentrations greater than 10 ppm." As discussed above, however, it is not meaningful to specify concentration without specifying support. We need to decide over what area to apply the threshold: one square millimeter, one square meter, one square kilometer, or…?

As with air quality standards, it makes sense to choose areas that are plausibly linked to human exposure, such as ¼ acre for residential exposure or 1 acre for industrial exposure. It is inappropriate to apply target thresholds to individual sample cores for the same reason we don't apply air quality standards to one-second averages… because they contribute a small fraction to an individual's total exposure. Conversely, it is also inappropriate to apply the cleanup level to areas of hundreds or thousands of acres because this "dilutes" risk by averaging the risks to both exposed and unexposed individuals.

We apply exposure unit thresholds to the average concentration of exposure units without any concern about how the concentration may vary within smaller supports. This follows directly from two assumptions: risk is directly proportional to concentration; and, exposure probabilities are everywhere equal. The latter assumption applies to hypothetical future exposures, but is not applicable when, say, cleaning up an existing residence where actual high-exposure areas, such as play areas or gardens, can be identified.

Finally, we need to decide exactly how we are going to deal with decision uncertainty in order to completely specify the design criteria. If correct risk management is the only consideration, the probability that the cleanup will fail because of decision error is part of the posterior risk of the cleanup process. Formally, the posterior risk, conditional on the cleanup process being implemented, equals the probability that the post-cleanup concentration will be x, times the risk if it is *x*, summed over all possible values of *x*. If the exposure unit threshold is equal to 1 ppm, a cleanup will satisfy the risk-based objective if the *expected* mean exposure unit concentration after cleanup is at or below 1 ppm.

Although the risk-based objective is completely defensible scientifically, it leads to the acceptance of relatively high failure rates at concentration levels just above the threshold. This is the most cost-effective approach to cleanup, but may pose problems when dealing with immediately affected stakeholders, as when cleaning existing residences, schoolyards, or parks. Rather than trying to convince people who are worried about failures that they should be looking only at the expected outcome, it may be more appropriate to incorporate reassurance into an alternate confidence-based objective: 95% certainty that the mean concentration of the exposure unit after cleanup will be at or below the threshold. This approach does not abandon the notion that risk is the primary concern. The exposure unit still provides a risk-based scale, although it might change from a hypothetical future residence to an actual one with real people. The confidence-based design results in a high probability that the risk will not exceed the threshold.

Cleanup designs for both risk-based and confidence-based objectives are included in the tables below.

Decision Rules

The basic form of a decision rule is:

- If (condition) then (action1); otherwise (action2).

Choosing the best decision rule is a critical part of designing a cleanup process. It is not sufficient to simply restate the objective in the form of a decision rule, as:

- If the mean concentration of an exposure unit is greater than the threshold level, then clean it; otherwise do nothing.

The rule above is not a useful decision rule because we cannot know the true concentration of the exposure unit. We must state the condition in terms of what we actually know, or can estimate, from sample data. Examples of possible practical decision rules include:

- If the mean concentration of the samples in an exposure unit is greater than a cleanup level (CL), then clean it; otherwise do nothing. Or:
- If the Student's t 95% upper confidence limit (UCL) of an exposure unit is greater than the CL, then clean it; otherwise do nothing. Or:
- If the ordinary kriging estimate of an exposure unit is greater than 0.5 * the CL, then clean it; otherwise do nothing. Or the more complex *iterative truncation rule*:
- If the estimated mean concentration of an exposure unit is greater than the CL, then flag the highest remaining estimated remediation unit for cleanup, recalculate the residual estimated exposure unit mean (with concentrations of flagged units set equal to zero), and repeat this rule; otherwise stop and clean up any flagged units.

The CL may be equal to the target threshold level, but in general it is simply another design variable that must be chosen along with the form of the rule and the sampling scheme.

Given a clear, quantitative cleanup objective, such as the risk-based or the confidence-based objective described earlier, any of the above decision rules (as well as numerous others) could accomplish the objective with suitable choices for the sampling scheme and the CL.

The rule using Student's t is an example of a statistical hypothesis test. Formal hypothesis tests provide useful decision rules because their performance is known when particular assumptions, such as normality, are met. Those assumptions are rarely, if ever, met in soil cleanup operations. Soil contaminant distributions are typically highly skewed. Log-transforms often make the results approximately normal, but decisions made on log-transforms are not necessarily valid decisions with respect to risk. The reason for this follows from the earlier discussion that related posterior risk to the expected mean concentration of an EU after cleanup. Statistics calculated on log-transformed data are not directly related to expected mean concentration. Therefore, it is not easy to evaluate the effect on risk of a decision based on log-transformed data. Similar problems occur with non-parametric statistical tests that compare medians or other percentiles. Percentiles of a sample distribution are a function of the particular sample support, and are not directly related to risk in an exposure unit.

The last paragraph notwithstanding, hypothesis tests are valid decision rules because they can generate yes-or-no decisions based on sample data. The primary reason they do not appear in the generic cleanup approach is to keep it as simple as possible.

The Generic Cleanup Procedure

The approach is a slightly simplified version of the method used at Piazza Road, (Ryti, et al., 1992). Piazza Road was a cooperative venture between EPA's Region 7, Office of Research and Development (ORD), and what was then the Quality Assurance Management Staff (QAMS) to field-test the current data quality objectives (DQO) guidance that was introduced in 1993. The Piazza Road cleanup design was a multi-scale iterative truncation method with specified EU and RU supports. The highest RU's within an EU were removed iteratively until the residual EU threshold fell below a risk-based cleanup level. Piazza Road demonstrated the importance of applying risk-based performance criteria at the appropriate risk management scale. The risk of concern was residential exposure, so the exposure unit was chosen to be 5000 square feet. Selective cleanup on smaller remediation units was desired in order to treat the minimum amount of soil required to bring the exposure unit below the threshold. A pilot study was conducted to evaluate sample variability within several typical EU's. The pilot data was used to choose the most effective RU size and sampling scheme in a manner similar to the simulation methods described below. This approach was shown to significantly reduce treatment costs while strictly controlling risk.

Assumptions

The generic approach is designed to be applicable to a variety of cleanup problems that have certain characteristics in common with the Piazza Road case:

- Direct exposure (contact, ingestion, and inhalation) to soil is the primary concern.
- The major health concern is chronic exposure to dispersed contamination.
- Surface soil is a layer of some specified constant thickness, such as the upper 6 inches, or a similar natural layer of relatively constant thickness.

- Contamination at depth is not a problem, or can be dealt with sequentially; i.e., by sampling the next lower layer after removing the surface layer.
- An individual sample (or composite increment) is a small diameter vertical core through the surface layer.
- Sampling, compositing, sub-sampling, and measurement are done according to sound sampling theory and good laboratory practice to adequately control total measurement error. (Gerlach and Nocerino, 2003)

Selecting a Cleanup Design

- Define the total clean-up area. This may be the entire site, or more likely, a sub-area of the site, such as a stratum or operable unit that has been identified based on site history and previous site investigations.
- Select the size of the exposure unit based on future use. Typically, this will be on the order of ¼ acre if a future residential exposure scenario is relevant, or one acre for a future industrial scenario. As discussed earlier, this is needed to focus the objective at scales large enough to be relevant to human exposure, but not so large as to obscure significant contamination through over-dilution.
- Choose the target threshold for mean EU concentration, and determine if it will be implemented as a risk-based or confidence based objective.
- Estimate the variability of samples within a typical EU in terms of log standard deviation. This will be used to help choose the design. The design tables include different designs for different levels of variability. Ideally this variability estimate should be based on data obtained during earlier phases of site investigation. Otherwise it may be estimated by analogy to other similar sites.
- Select the appropriate design from Table 2 or 3. The design will specify the size of an RU, how many samples and composite measurements should be taken within each RU, and the CL for the design.

Implementing the Cleanup Operation

- Sample all of the RU's in an EU, according to the design. Calculate the sample mean for each RU. If the design specifies only one sample or composite per RU, the measured value is taken as the RU mean.
- Make clean-up decisions using iterative truncation logic: If the mean EU concentration exceeds the CL, clean the highest measured RU (in case of a tie, choose either); repeat until the estimated residual mean EU is below the CL.

Note that the designs in Tables 2 and 3 do not specify the size of an EU. The designs depend only on the variability of samples within the EU, expressed as log standard deviation. At any particular site, if you reduced the exposure unit, say from one acre to ¼ acre, the standard deviation of samples within the smaller EU would also be reduced. This would likely lead to selecting a different design from the tables.

The generic method provides a standard of comparison for any proposed cleanup alternative. The test results below show how well the generic method performs in terms of decision outcomes under the assumed conditions. The performance of any proposed alternative should be demonstrated by similar testing to be better or cheaper than the generic approach.

The risk-based and confidence-based cleanup objectives have been defined strictly in terms of what will be considered a successful cleanup by the regulator. As long as a cleanup meets the objectives, the regulator should have no particular concern about the details of the design. The operator, or whoever is paying the cleanup costs, has the primary interest in choosing the optimal design, which can be defined as the lowest cost design that meets the objectives.

The total cost of a design includes sampling and cleanup costs. The generic design tables included below provide an operator with several simple design options that will work. The operator will have to determine whether it is worthwhile to look for a lower cost alternative, and to demonstrate that it will work.

Evaluating Design Performance through Computer Simulations

Cleanup designs were evaluated through computer simulations of the sampling and decision making process. A design consists of a choosing the size of an RU, the number of composite samples in each RU, the number of sample increments in each composite, and a CL to be applied to the estimated residual EU mean during the iterative truncation process. Designs were tested at five different levels of assumed sample variability within an EU. Variability was assumed to be log-normal, with log standard deviations of 0.5, 0.75, 1.00, 1.25, or 1.50. The latter two represent quite high variability, corresponding to coefficients of variation (CV's) of approximately 2 and 3, respectively.

Total measurement error for all designs was assumed to have an RSD of 0.212, or 21.2%. This would correspond to independent sub sampling and analytical errors of 15% each, where the error variances are additive: $0.212 = \sqrt{\ } (0.15*0.15 + 0.15*0.15)$. (RSD and CV are mathematically identical measures of variability. CV typically denotes variability in a population or sample data set, while RSD usually refers to variability in the measurement process). The alternative designs section below discusses how a design can be adjusted when measurement variability is significantly different from this assumption.

Four choices of RU size, relative to EU size, were evaluated: EU area/RU area = 1, 1/4, 1/16, 1/64. Smaller RU sizes permit a more selective cleanup that tends to reduce the cleanup cost, but they also incur higher sampling costs. When EU area/RU area = 1, the cleanup is not selective – the decision to clean an EU is all or nothing.

For one of the chosen log standard deviations, each of the four RU sizes was evaluated over the following ranges of design parameters:

- 1 to 8 composite samples per RU. Assume 1 analytical measurement per composite sample.
- 1 to 16 sample increments per composite
- CL from 0.5 to 2.0 times the target EU threshold.

Because of computational constraints, only a selected subset of the possible composite and sample combinations was evaluated. The combinations provided a stepwise reduction in RU estimation variance. The search for acceptable designs began with the highest CL and the highest variance sampling scheme (lowest number of composites and samples). The design was tested by simulation. If it did not succeed, the next lowest sampling scheme was tested, and so on until all sample schemes had been tried. Then the CL was lowered and the process repeated. The first design to strictly satisfy the design objective was chosen for inclusion in the tables.

To simulate the decision process, a target threshold for EU mean concentration was set at 1.0. A series of 'true' EU concentrations was chosen to cover the range from 0.1 to 100 times the target threshold. The EU log standard deviation was partitioned into within-RU and between-RU components. 'True' RU means were randomly generated so that the mean of the true RU means equals the selected EU mean. Each RU was sampled randomly with a specified number of composite samples and sample increments. Measurement error was added to the mean of the samples to produce the RU estimated means.

Iterative truncation was performed on the RU estimated means until the estimated residual EU mean (assuming zero concentration after cleanup) was less than the current CL. The corresponding true residual EU mean was calculated to see how the cleanup actually worked. For the selected EU mean, the entire process was repeated 1000 times. For the risk-based objective, the average or expected performance was compared to the target threshold. For the confidence-based objective, the 95^{th} percentile of the 1000 true residual EU means was used. Finally, to evaluate the overall performance of the method, the entire simulation process was repeated for the chosen series of true EU concentrations.

Table 1 shows how the iterative truncation process works for a single EU with a true mean concentration of 3.0. The sampling scheme being tested was 1 composite, 6 samples per composite, and CL =1.0. If the true mean concentrations of the RU's were known, the best decision would be to clean the four highest RU's. Based on composite measurements, the first, second and fifth highest RU's were actually cleaned, resulting in a residual EU concentration of 1.09, slightly higher than the target threshold. In this case, the measured value of the fifth highest RU was nearly three times too high, so the benefit from cleaning it was overestimated. This kind of conditional bias is inherent in truncation decisions. On average, measured values will predict better results than you actually achieve. This is illustrated in the bottom row of Table 1, which shows the average results from 1000 repetitions of the decision process on 1000 different EU's, all with a true mean of 3.0. The average estimated residual concentration is 0.89, while the true residual is 1.02, nearly 15% higher. Ideally, you would have liked to have the estimated residual as close to 1.0 as possible. It is necessary to clean the next whole RU that drives the estimate at or below the CL, resulting in a slightly conservative estimated cleanup. Here it almost compensates for the conditional bias. Nevertheless, the design here fails the strict risk-based criterion that the residual EU mean should be at or below the threshold. In the design process we would move on to the next lower variance sampling scheme and try again.

Table1. Simulated cleanup decisions for 16 RU's in an EU. Underlined RU's should have been cleaned. RU's in italics were cleaned.

	Before Cleanup		After Cleanup	
	True RU	Est. RU	True RU	Est. RU
	4.65	3.91	4.65	3.91
	1.66	1.43	1.66	1.43
	3.73	*10.57*	*0.00*	*0.00*
	0.29	0.20	0.29	0.20
	0.32	0.31	0.32	0.31
	0.42	0.33	0.42	0.33
	0.12	0.09	0.12	0.09
	0.78	0.74	0.78	0.74
	2.12	2.13	2.12	2.13
	0.14	0.07	0.14	0.07
	0.39	0.40	0.39	0.40
	18.84	*17.11*	*0.00*	*0.00*
	4.30	4.11	4.30	4.11
	8.00	*8.37*	*0.00*	*0.00*
	1.54	1.19	1.54	1.19
	0.70	0.81	0.70	0.81
Single EU	**3.00**	**3.24**	**1.09**	**0.98**
1000 Reps	**3.00**	**3.00**	**1.02**	**0.89**

The third RU in Table 1 has a nearly three-fold estimation error that makes us think we are cleaning three times the contaminant that is actually there. Because that RU is only one of 16 cleanup decisions, the final cleanup only misses the target by about 10%.

Figure 1 shows the expected performance of the method over a range of EU concentrations. The steep line shows the initial true EU concentration and the horizontal line shows the target residual concentration of 1.0. Points show the cleanup performance in terms of true mean residual EU concentration. Overall, the performance here is very good, but this design is rejected because some points exceed the target threshold.

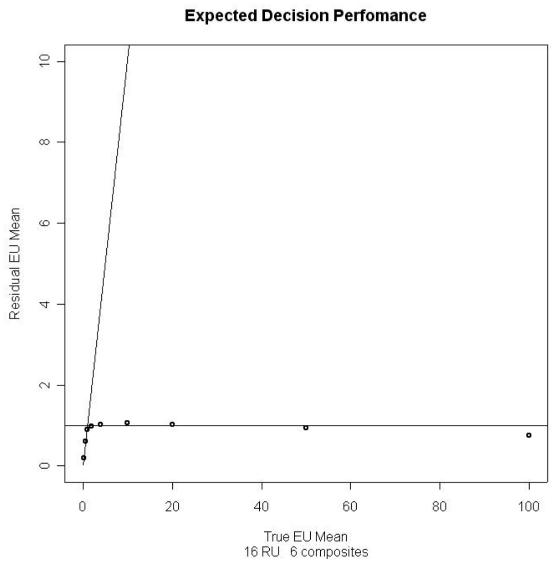

Figure 1 - Performance of the decision process for EU's with 16 RU's, each estimated by one 6-increment composite sample. The horizontal line represents target performance. Each point is the average of 1000 simulated EU cleanups.

Figures 2 -6 provide an alternative way of visualizing the performance of a design. The 1000 EU repetitions at each of 10 EU concentrations are displayed as if they were a large "site" with a total of 10,000 EU's. The EU concentrations are displayed as color-coded pixels on a site map. Figure 2 shows the EU concentrations before cleanup – the numbers in the white vertical column are the original concentrations. The numbers double as the color key for the subsequent after-cleanup maps: The number represents the upper concentration limit for the associated color.

Figures 2 and 4 illustrate the performance of two successful risk-based designs. Figure 3 is the result of a non-selective cleanup, where the RU is the same as the EU. In Figure 4, the cleanup was highly selective, with 16 RU's in each EU. The numbers in the white column are the average after cleanup concentrations of the EU's in the corresponding row. The more selective cleanup reduces the cleanup area from 55% to 39% of the total site area, but at the expense of much more sampling.

Figures 5 and 6 are similar to Figures 4 and 5, but show the performance of confidence-based cleanup designs.

Figure 2 - 10,000 simulated EU's representing one design test simulation, displayed as a hypothetical site map. Each horizontal band contains 1000 EU pixels. The initial EU concentrations are shown in the white column. Colors are considered to represent concentrations equal to or less than the initial concentration, so that after cleanup, all EU's equal to or les than 1.0 will be a shade of green, and all EU's greater than 1.0 will be a shade of red. The site mean is the mean of all EU's. The contaminated mean is the mean of all EU's with initial concentrations greater than or equal to 1.0.

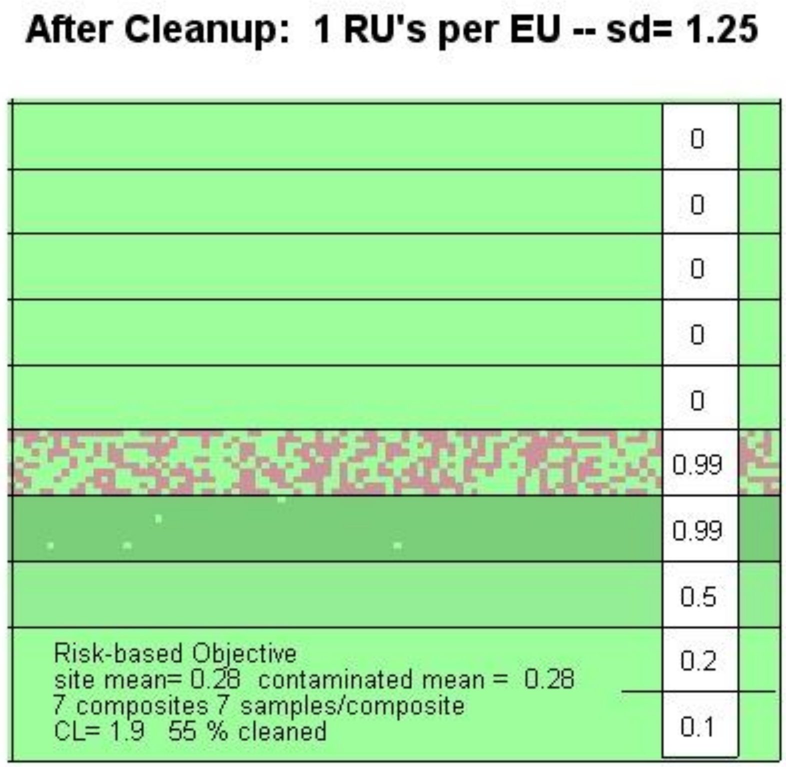

Figure 3 - The hypothetical site from Figure 2 after a successful risk-based cleanup with 1 RU per EU. The numbers in the white column are the average, or expected, post-cleanup concentrations. The criterion for a successful design is that none of these concentrations exceed 1.0. 1 RU per EU is the non-selective all-or-nothing option. This results in 100% cleanup of the higher concentration EU's, with 55% of the total site area cleaned.

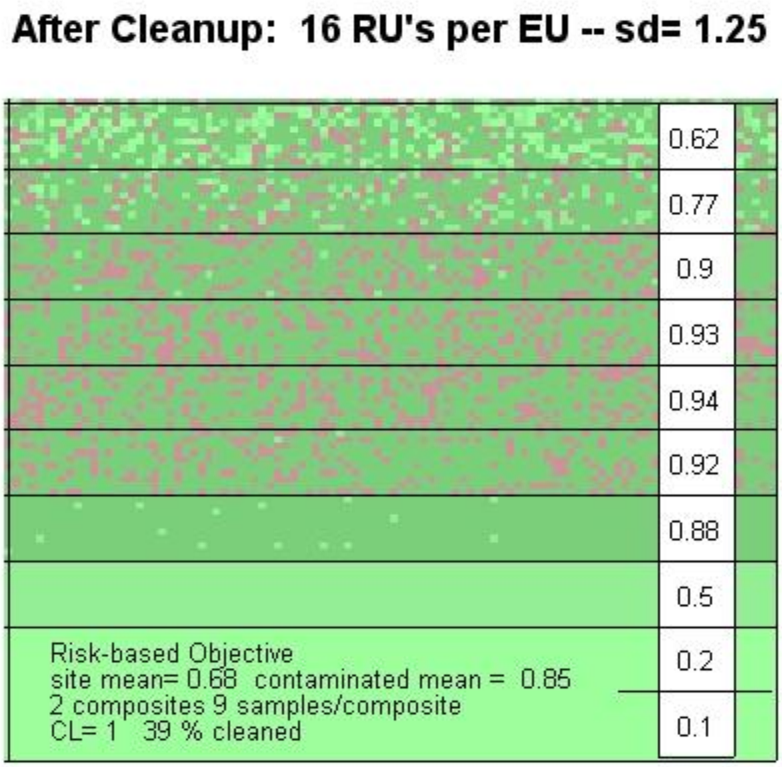

Figure 4 - The hypothetical site from Figure 2 after a successful risk-based cleanup with 16 RU's per EU. This is a selective cleanup where most contaminated EU's are partially cleaned. The total area cleaned drops to 39% of the site, while sampling costs increase.

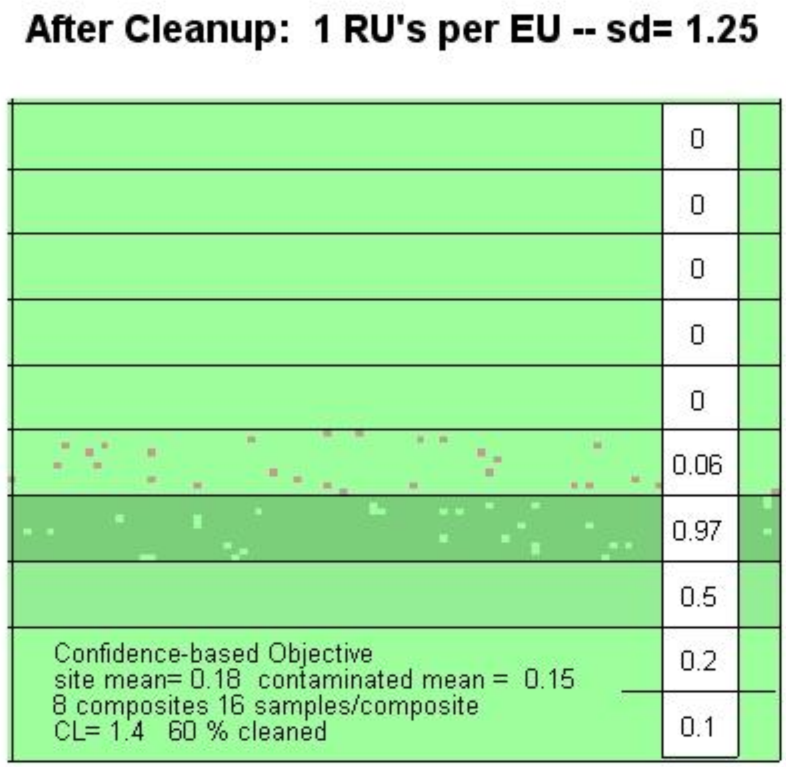

Figure 5 - The hypothetical site from Figure 2 after a successful confidence-based cleanup with 1 RU per EU.

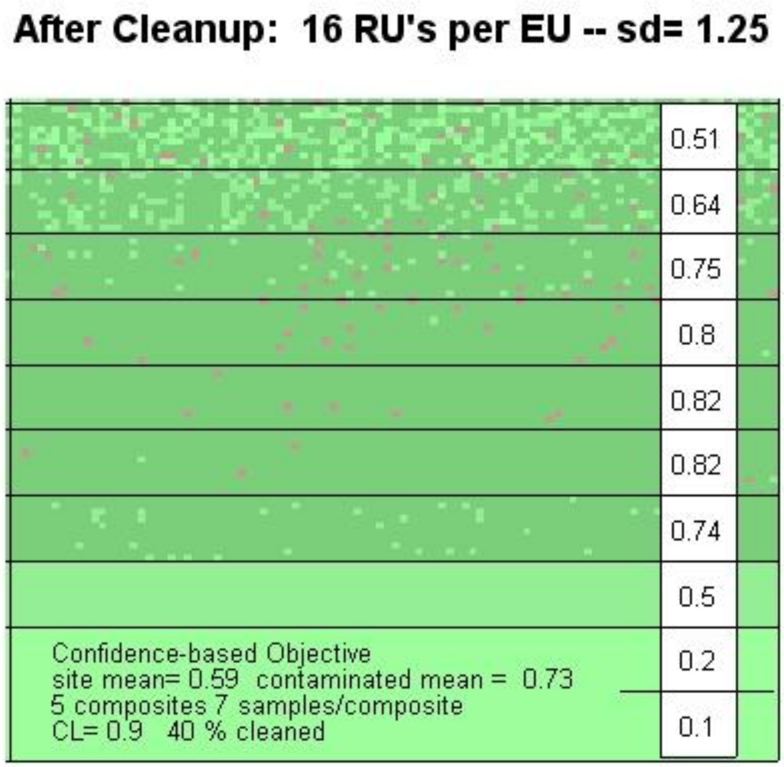

Figure 6 - The hypothetical site from Figure 2 after a successful confidence-based cleanup with 16 RU's per EU.

Cleanup Design Tables

Tables 2 and 3 contain successful risk-based and confidence-based designs, respectively. For a specified log standard deviation, each column represents one successful design. The first four rows describe the design, and the next two describe design performance for the hypothetical site. The last row lists the standard error of the RU mean, expressed as RSD. This combines sampling and measurement errors:

$$\text{RU mean RSD} = \sqrt{e^2/ncs + RUsd^2/(ncs*nsi)}, \text{ where}$$

e = measurement RSD, RUsd = within-RU log sample standard deviation,
ncs = number of composite samples, and nsi = number of sample increments /composite.

+

Table 2.
Generic designs -- Risk-based

Log Standard Deviation = 0.5

RU's per EU	1	4	16	64
Composites	1	2	2	2
Samples per Composite	4	9	9	9
Cleanup Level	1.9	1.2	1	0.9
Percent of Site Cleaned	55	51	48	47
Site mean	0.28	0.38	0.55	0.7
RU mean RSD	0.33	0.17	0.16	0.16

Log Standard Deviation = 0.75

RU's per EU	1	4	16	64
Composites	1	2	2	8
Samples per Composite	10	9	9	13
Cleanup Level	1.9	1.2	1	1
Percent of Site Cleaned	55	50	45	43
Site mean	0.28	0.41	0.61	0.74
RU mean RSD	0.32	0.2	0.17	0.08

Log Standard Deviation = 1.00

RU's per EU	1	4	16	64
Composites	5	1	1	7
Samples per Composite	5	7	12	7
Cleanup Level	1.9	1.2	1	1
Percent of Site Cleaned	55	48	42	39
Site mean	0.27	0.48	0.69	0.75
RU mean RSD	0.22	0.34	0.26	0.09

Log Standard Deviation = 1.25

RU's per EU	1	4	16	64
Composites	7	2	2	7
Samples per Composite	7	9	9	7
Cleanup Level	1.9	1.3	1	1
Percent of Site Cleaned	55	45	39	35
Site mean	0.27	0.5	0.68	0.75
RU mean RSD	0.2	0.26	0.21	0.1

Log Standard Deviation = 1.50

RU's per EU	1	4	16	64
Composites	5	2	1	5

Samples per Composite	**7**	**9**	**16**	**7**
Cleanup Level	**1.8**	**1.3**	**1**	**1**
Percent of Site Cleaned	**56**	**43**	**35**	**31**
Site mean	**0.28**	**0.54**	**0.72**	**0.76**
RU mean RSD	**0.27**	**0.29**	**0.28**	**0.13**

Table 3.
Generic designs -- Confidence-based

Log Standard Deviation = 0.5

RU's per EU	**1**	**4**	**16**	**64**
Composites	**5**	**7**	**2**	**2**
Samples per Composite	**7**	**7**	**9**	**9**
Cleanup Level	**1.6**	**0.95**	**0.85**	**0.8**
Percent of Site Cleaned	**60**	**55**	**51**	**49**
Site mean	**0.19**	**0.29**	**0.47**	**0.62**
RU mean RSD	**0.13**	**0.09**	**0.16**	**0.16**

Log Standard Deviation = 0.75

RU's per EU	1	**4**	**16**	**64**
Composites	**8**	**5**	**7**	**8**
Samples per Composite	**13**	**7**	**7**	**13**
Cleanup Level	**1.6**	**0.95**	**0.95**	**0.95**
Percent of Site Cleaned	**60**	**54**	**46**	**43**
Site mean	**0.19**	**0.3**	**0.55**	**0.71**
RU mean RSD	**0.11**	**0.13**	**0.1**	**0.08**

Log Standard Deviation = 1.0

RU's per EU	**1**	**4**	**16**	**64**
Composites	**7**	**5**	**5**	**7**
Samples per Composite	**7**	**5**	**7**	**7**
Cleanup Level	**1.4**	**0.9**	**0.9**	**0.95**
Percent of Site Cleaned	**60**	**53**	**44**	**40**
Site mean	**0.18**	**0.3**	**0.56**	**0.72**
RU mean RSD	**0.16**	**0.17**	**0.13**	**0.09**

Log Standard Deviation =1.25

RU's per EU	**1**	**4**	**16**	**64**
Composites	**8**	**7**	**5**	**7**

Samples per Composite	**16**	**9**	**7**	**7**
Cleanup Level	**1.4**	**0.95**	**0.9**	**0.95**
Percent of Site Cleaned	**60**	**51**	**40**	**36**
Site mean	**0.18**	**0.32**	**0.59**	**0.72**
RU mean RSD	**0.13**	**0.14**	**0.14**	**0.1**

Log Standard Deviation =1.50

RU's per EU	**1**	**4**	**16**	**64**
Composites	**8**	**7**	**8**	**5**
Samples per Composite	**11**	**7**	**13**	**7**
Cleanup Level	**1.2**	**0.9**	**0.95**	**0.9**
Percent of Site Cleaned	**62**	**49**	**36**	**32**
Site mean	**0.17**	**0.33**	**0.62**	**0.69**
RU mean RSD	**0.18**	**0.17**	**0.11**	**0.13**

Some trends in the tables are noteworthy. As expected, increasing the number of RU's in an EU makes the cleanup more selective, thus reducing the site area that needs to be cleaned. This is also reflected in the mean site concentration, which increases as the cleaned area decreases. Very significantly, the amount of reduction in cleanup area diminishes as the number of RU's increases.

Table 4 shows a simple cost analysis for the four risk-based designs from Table 2 where log standard deviation equals 1.25. Total cost is computed by multiplying unit EU cleanup cost by the average percentage cleaned, and adding sample and composite costs. A 4 RU design is almost certain to be more cost effective than a non-selective design. The most highly selective designs are not cost effective here except when unit cleanup costs are very high relative to unit sampling costs.

Table 4: Total Cleanup Costs (Sampling plus Cleanup)

Cleanup Cost	**RU per EU**			
in $ per EU	**1**	**4**	**16**	**64**
10,000	6,340	5,620	8,380	10,000
20,000	11,840	10,120	12,280	13,600
40,000	22,480	19,120	20,080	20,800
80,000	44,840	37,120	35,680	35,200

Based on designs from Table 2, log sd =1.25
Sampling costs assume $10/sample, $50/composite

Cleanup levels decrease as the number of RU's increases. Cleanup levels higher or lower than the target threshold introduce a bias into the decision process. This CL bias can compensate for other biases in the process. One source of decision bias is introduced by the decision rule itself, which cleans the next whole RU required to bring the EU estimate "at or below" the CL. On average, this biases cleanup decisions toward over-cleaning. A separate selection bias occurs because of estimation errors. Even if the estimation errors themselves are unbiased, the decisions become biased. Decision errors in both directions combine to fail to clean some higher concentration soil, while erroneously cleaning some lower concentration soils. The end result is a bias toward under-cleaning. With large RU's, the average of "at or below" is quite a bit below, so the over-cleaning bias dominates. With very small RU's, the over-cleaning bias becomes negligible, and the selection bias dominates. The changing CL's compensate for these shifting bias effects.

How robust are the designs?

What if we used only half as many composites as we should? What if measurement error is much worse than we think? What if the variability in an RU is much higher than we designed for? Good designs should fail gently. That is, fairly large errors in the design or in the design assumptions should result in relatively minor errors when a cleanup is actually implemented.

Figures 7-9 illustrate alternate outcomes for the selective risk-based cleanup of Figure 4, corresponding to the three questions in the previous paragraph. The results in Figures 7 and 9 would fail the strict criterion for inclusion of a design in Table 2 because some of the after-cleanup average concentrations exceed 1.0. That criterion is conservative, because it requires that the design succeed under a worst-case scenario – that all contaminated soil at the site has precisely the concentration at which the design performs worst. At an actual site with a range of contaminant concentrations, a risk-based design is successful if the expected after-cleanup concentration for all contaminated EU's is at or below the target threshold. In these examples, the expected after-cleanup concentration is indicated by the "contaminated mean", which ranges from 0.89 to 0.92. If these simulations represented actual sites, the cleanups would be successful in spite of the design flaws.

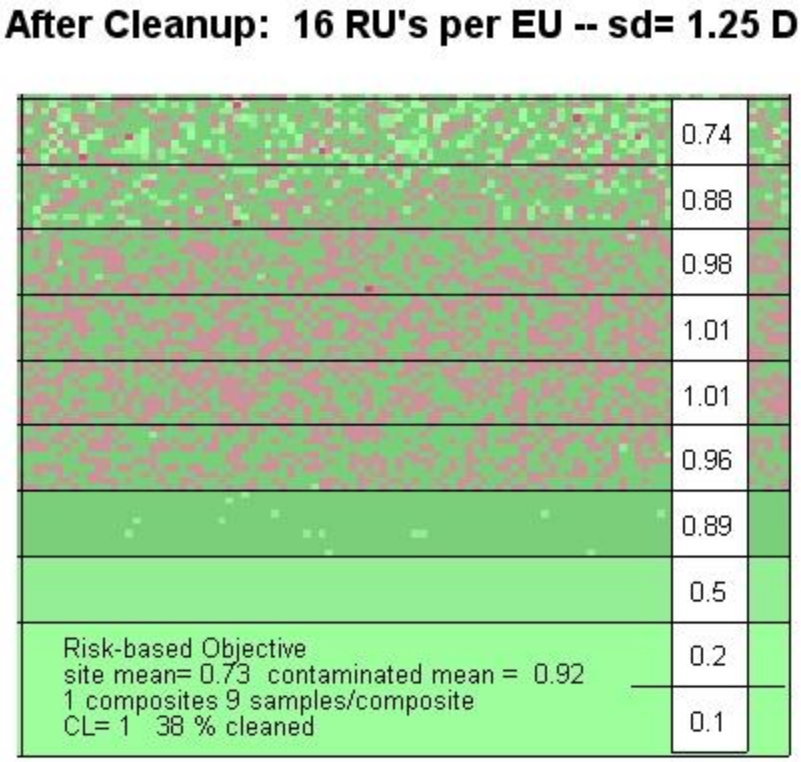

Figure 7 Test of robustness. The design here uses one-half the number of composites indicated in Table 2.

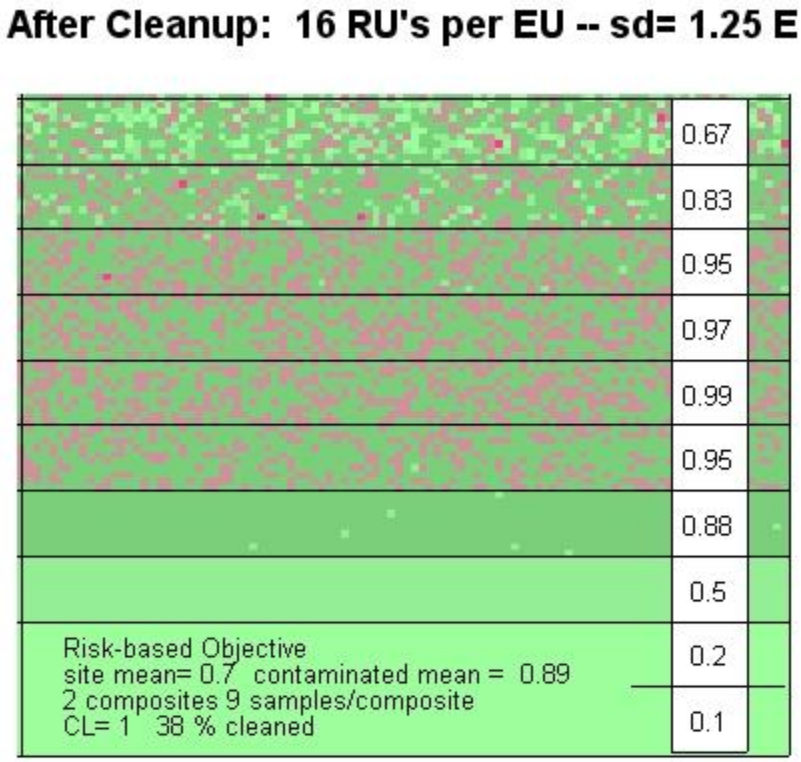

Figure 8 Test of robustness. The total measurement error in this cleanup simulation is twice as high as was assumed in generating Table 2 (RSD= 0.3 instead of 0.212).

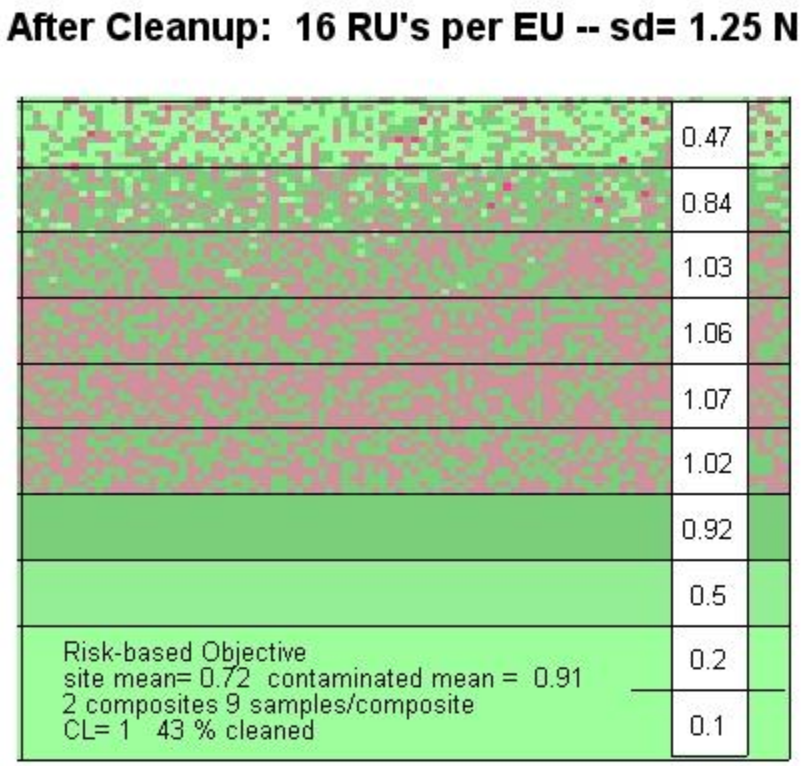

Figure 9 Test of robustness. Simulations for generating Table 2 calculated within RU and between RU variances as if EU variability were described by a zero-nugget variogram. Here the design is applied to a simulated site where 50% of EU variance is nugget (random noise).

Alternative Designs

The designs tabulated above are not optimal; they are only a subset of many possible successful designs. The planning team needs to decide if it is worth the time and effort to try to find a better solution (and to defend it to the regulators or the public). Some possible options for improvements are described below:

- If large areas of the site are expected to be either very dirty or very clean, an adaptive sampling or triage approach may be effective. Individual EU’s, or groups of EU’s could be screened by a preliminary composite sample. The initial decision rule would be three-fold: clean the highest concentration EU’s; take no further action on the lowest concentration EU’s; or initiate a selective iterative truncation design. This limits the high sampling cost associated with sampling every RU in an EU during a selective cleanup is limited to those EU’s where it will make the most difference.

- Field analytical methods can be used to provide real-time information for a single-pass cleanup. Sometimes this comes at the expense of increased measurement error. As discussed above,

  $$\text{RU mean RSD} = \sqrt{e^2/\text{ncs} + \text{RUsd}^2/(\text{ncs*nsi})}$$

  RU mean RSD quantifies the overall ability of the sampling scheme to estimate mean RU concentrations. Any combination of samples and composites where RU mean RSD is less than or equal to the tabulated value, will provide an acceptable design. The tabulated designs assume that e = 0.212. (In a QA/QC program, e can be measured directly by comparing field splits). If e for the existing measurement method differs from 0.212, a design can be adapted by changing ncs or nsi. Similarly, designs can be adapted to accommodate alternate methods, such as changing to field analytical methods, or modifying sub-sampling procedures
- If spatial "features" are larger than RU's, that is, if cleanup would likely occur as a group of contiguous RU's, then geostatistics will usually provide better results. The designs in Tables 2 and 3 assume no correlations among samples, RU's, or EU's. As a consequence, only samples within an RU can be used to estimate the RU, and every RU must be sampled. The geostatistical model allows using data near an RU to be used for estimating the RU mean, even if there are no samples in the RU itself. Through kriging, the information in a sample or composite is extended spatially to multiple RU's, so long as it continues to add value to an estimate. When this is valid it can significantly reduce the amount of sampling required.

Software

The design simulations were done using R, a freeware statistical computing language. R can be downloaded from the site listed in the references. The code for producing a set of risk-based designs similar to those in Table 2 is provided below. Because of the variability inherent in simulations, results may not be identical between runs. Although R can be run under Windows, it is not point-and-click software; rather it requires the user to type commands. A long series of commands, like the code below, is executed by saving it as a text file with a ".R" extension, linking R to the appropriate file directory, and running it with the command:
source("filename.R").

```
####  List EU means to test. For each EU mean:
####    For each repetition:
####      Create a (lognomal) set of RU's in the EU. Normalize to EU mean.
####      Sample RU's. Remove highest RU until est EU < threshold T.
####      Calculate true residual EU mean
####    Average residual EU means over repetitions
####-----------------------------------------------
#### Initialize variables
trueeumean=c(.1,.2,.5,1,2,5,10,20,50,100) ### list of true EU means: the simulated 'site'
nummeans=length(trueeumean)
eusd=c(.5,.75,1,1.25,1.5) ### population log standard deviations in the EU
samprsd=c(0.212) ### measurement error RSD
runumber=c(1,4,16,64)  #### number of ru in an eu
T=c(1.9,1.8,1.7,1.6,1.5,1.4,1.3,1.2,1.1,1.0,.95,.90,.85,.80,.75) ## cleanup thresholds(estimated residual concentration)
tx=1
target=1 ## desired residual concentration
```

```
residualeumean=numeric()
estresidmean=numeric()
residualeumax=numeric()
fractioncleaned=numeric()
samprumean=numeric()
est=numeric()
true=est
reps=1000
residual=numeric()
estresid=numeric()
rucleaned=numeric()
cleaned=numeric()
notcleaned=numeric()
samptotal=0
count=1
sl=numeric()
prow=c(1,1,2,2,3,3)
pcol=c(1,2,1,2,1,2)
count.out=numeric()
ru.out=numeric()
comp.out=numeric()
samp.out=numeric()
eusd.out=numeric()
error.out=numeric()
threshold.out=numeric()
maxmean.out=numeric()
maxucl95.out=numeric()
cost.out=numeric()
mwin.out=numeric()
uwin.out=numeric()
site.out=numeric()
totalrsd.out=numeric()
cc=8
ss=16
source ("sort.data.frame.R") ###from:http://www.biostat.wustl.edu/archives/html/s-news/2004-09/msg00171.html
### plot 'before' site map
beforegrid=rep(trueeumean,1000);beforegrid=sort(beforegrid)
dim(beforegrid)=c(100,100)
ggrid=beforegrid
windows(width=5,height=5)
image(ggrid,col=c("palegreen1","palegreen2","palegreen3","pink3","violetred1","violetred2",
                "red","brown"),breaks=c(0,.201,.501,1.001,2.001,
                 5.001,10.001,20.001,100.001),xaxt="n",yaxt="n");title(main="Before Cleanup")
text(.03,.1,paste("max EU mean=",max(trueeumean),"  site mean=",mean(trueeumean)),pos=4,cex=.8)
text(.03,.05,paste("contaminated mean = ",round(mean(trueeumean[trueeumean>=1]),2)),pos=4,cex=.8)
polygon(c(.85,.95,.95,.85),c(0,0,1,1),col="white")
text(.9,((1:10)-.5)/10,paste(round(trueeumean,2)),cex=.8)
abline(h=(2:10)/10)
abline(v=0)
lines(c(.8,1),c(.1,.1))

for(rl in 1:length(runumber)) #### number of ru in eu
{
runum=runumber[rl]
if (runum==4) {tx=4} ### skip some of the higher cleanup thresholds that won't work in this case.
if (runum==16) {tx=6}
```

```
if (runum==64) {tx=9}

for (ml in 1:length(samprsd))  #### measurement error sd
{

for(sl in 1:length(eusd))  #### log sd
{

### calculate within ru and between ru variances. Assuming a linear variogram and square ru &
### eu shapes, within ru variance is proportional to side(ru)/side(eu)
svar=(eusd[sl]*eusd[sl])/sqrt(runum)
sampinru=sqrt(svar)
ruineu=sqrt(eusd[sl]*eusd[sl]-sampinru*sampinru)
#### generate a table of designs for this eu-ru combo. All combinations of composite and samples are considered.
#### total ru variance (tvar) is computed for each. When multiple designs have the same total number of samples,
#### the one with fewest composites is chosen. Sort designs by tvar,select highest tvar and any lower tvar design
#### where tvar is at least .02 lower than the last selected design.
ruvar=rep(svar,cc*ss)
errorvar=rep(samprsd[ml]*samprsd[ml],cc*ss)
composites=sort(rep(1:cc,ss))
samples=rep(1:ss,cc)
nsamps=samples*composites
tvar=ruvar/nsamps + errorvar/composites

tempdesign=data.frame(composites,samples,nsamps,tvar)
tempdesign=sort.data.frame(~nsamps+composites,tempdesign)
design=tempdesign[1,]

for (i in 2:length(tempdesign[,1])) ##retain n with fewest composites
{
  if (tempdesign[i,3]>tempdesign[(i-1),3]) {design=rbind(design,tempdesign[i,])}
}

tempdesign=design
design=tempdesign[1,]
flag=1
for(i in 2:length(tempdesign[,1]))  ## retain sufficiently decreasing variance designs
{

  if(tempdesign[flag,4]-tempdesign[i,4]>.02)
   {
   design=rbind(design,tempdesign[i,])
   flag=i
   }
}
### table of designs completed

hit=FALSE
for(tl in tx:length(T))  #### action threshold
{

if(hit) break

#### create within-RU population of samples normalized to mean=1
sampset=(exp(rnorm(10000,0,sampinru))); sampset=sampset/mean(sampset)
```

```
##pdf(file=paste(count,"-",eusd[sl],samprsd[ml],runum,T[tl],".pdf"),height=10,width=7.5)
##frame()
##par(mfrow=c(3,2),cex=.66)
for (cl in 1:length(design$nsamps)) #### test a design
{
compnum=design$composites[cl]
sampnum=design$samples[cl]
aftergrid=numeric()

for (i in 1:nummeans)  #### sample a particular EU mean
{

  for (j in 1:reps)   #### sample the EU a bunch of times
  {
    residual[j]=0
    rucleaned[j]=runum
    ####  create the RU means and normalize to the EU mean
    truerumean=exp(rnorm(runum,0,ruineu));truerumean=(truerumean/mean(truerumean))*trueeumean[i]

    for (k in 1:runum)
    {
    #### generate RU measurements: composite mean plus analytical error (RSD)
      samprumean[k]= truerumean[k]*mean(sample(sampset,sampnum*compnum))*rnorm(1,1,samprsd[ml]/sqrt(compnum))
      samptotal=samptotal+ samprumean[k]
    } #### next k ru's per eu
    srank=rank(samprumean)
    est=rep(0,runum)
    true=rep(0,runum)
    for (k in 1:runum)
    {
      #### cumulates the true and estimated EU residual assuming the k lowest-sample-ranked RU's are uncleaned
      #### retains the true EU residual for highest estimate (i.e.,minimum cleanup) that meets threshold
      est[k] = sum(samprumean[srank<=k])/runum
      true[k] = sum(truerumean[srank<=k])/runum
      if (est[k]<T[tl])
       {
       residual[j]=true[k];estresid[j]=est[k]
       rucleaned[j]=runum-k
       }
    } #### next k
    aftergrid=c(aftergrid,residual[j])
  } #### next j sampling repetition
  #### keep score for the particular EU mean
  residualeumean[i]=mean(residual)
  estresidmean[i]=mean(estresid)
  ##residualeumax[i]=max(residual)
  residual=sort(residual); residualeumax[i]=residual[round(.95*length(residual))] ##upper 95 percentile
  fractioncleaned[i]=mean(rucleaned)/runum
} #### next i eu mean

mwin.out[count]=0
uwin.out[count]=0
if(max(residualeumean)<=1) ### test residualeumax for confidence objective
{  ### begin output if design is successful
hit=TRUE
```

```
mwin.out[count]=1
s=samptotal/(runum*reps*nummeans)
count.out[count]=count
ru.out[count]=runum
comp.out[count]=design$composites[cl]
samp.out[count]=design$samples[cl]
eusd.out[count]=eusd[sl]
error.out[count]=samprsd[ml]
totalrsd.out[count]=round(sqrt(design$tvar[cl]),2)
threshold.out[count]=T[tl]
maxmean.out[count]=round(max(residualeumean),2)
maxucl95.out[count]=round(max(residualeumax),2)
cost.out[count]=round(mean(fractioncleaned)*100,0)
site.out[count]=round(mean(residualeumean),2)
### plot 'after' site map
dim(aftergrid)=c(100,100)
ggrid=aftergrid
windows(width=5,height=5)
image(ggrid,col=c("palegreen1","palegreen2","palegreen3","pink3","violetred1","violetred2",
        "red","brown"),breaks=c(0,.201,.501,1.001,2.001,
         5.001,10.001,20.001,100.001),xaxt="n",yaxt="n");title(main=paste("After Cleanup: ",runum,"RU's per EU –
        sd=",eusd[sl]))
text(.03,.16,"Risk-based Objective",pos=4,cex=.8) ### change risk to confidence if needed
text(.03,.12,paste("site mean=",site.out[count]," contaminated mean = ",round(mean(aftergrid[,31:100]),2)),pos=4,cex=.8)
text(.03,.08,paste(comp.out[count],"composites",samp.out[count],"samples/composite"),pos=4,cex=.8)
text(.03,.04,paste("CL=",threshold.out[count]," ",cost.out[count],"% cleaned"),pos=4,cex=.8)

polygon(c(.85,.95,.95,.85),c(0,0,1,1),col="white")
text(.9,(1:10-.5)/10,paste(round(residualeumean,2)),cex=.8)
abline(h=(2:10)/10)
abline(v=0)
lines(c(.8,1),c(.1,.1))

## alternative plot: performance graph
##oldpar=par(mfg=c(prow[cl],pcol[cl]))
##plot(trueeumean,residualeumean,type="p",xlab="Initial EU Mean",
##ylab="Residual EU Mean",main=paste("Generic Cleanup Performance"),
##lwd=2,ylim = range(0,10),xlim = range(0,max(trueeumean)))

##lines(c(0,target,1000000),c(0,target,target),lwd=2)
##text(10,9.7,paste(runum,"RU per EU"),pos=4,cex=1.2)
##text(10,9,paste(compnum,"samples/composite"),pos=4,cex=1.2)
##text(10,8,paste("EU log sd =",eusd[sl]),pos=4)
##text(10,7.5,paste("measurement error log sd =",samprsd[ml]),pos=4)
##text(10,7,paste("EU target concentration =",target),pos=4)
##text(10,6.5,paste("EU action level =",T[tl]),pos=4)
##text(10,5.5,paste("Circle: mean EU residual"),pos=4)
##text(10,5,paste("Plus: upper 95% EU residual"),pos=4)
##text(10,4,paste("Cleanup cost:",round(mean(fractioncleaned)*nummeans,2),"of",nummeans,"EU's cleaned"),pos=4)
##text(10,3.5,paste("maximum expected EU residual:",round(max(residualeumean),2)),pos=4)
##points(trueeumean,residualeumax,pch=3)
count=count+1
break
}  #### end output for successful design
}  #### next cl design
##dev.off()
```

```
##par(mfrow=c(1,1),mfg=c(1,1,1,1),cex=1)

}  #### next tl target action level

}  #### next sl eu log standard deviation

}  #### next ml measurement error

}  #### next rl number of ru's in an eu

gc.table=data.frame(ru.out,comp.out,samp.out,eusd.out,error.out,threshold.out,
        maxmean.out,maxucl95.out,cost.out,site.out,totalrsd.out)
filedate=date()
substring(filedate,14)="_"
substring(filedate,17)="_"
write.table(gc.table,file=paste("risk",filedate,".txt"))
```

References

Gerlach, R.W., and Nocerino, J. M., *Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples*, EPA/600/R-03/027, November 2003

R: A language and environment for statistical computing,R Development Core Team, Foundation for Statistical Computing, Vienna, Austria, 2004,{http://www.R-project.org}

(Ryti, R. T.; Neptune, D.; Groskinsky, B. Environmental Testing & Analysis 1992, January/February.)