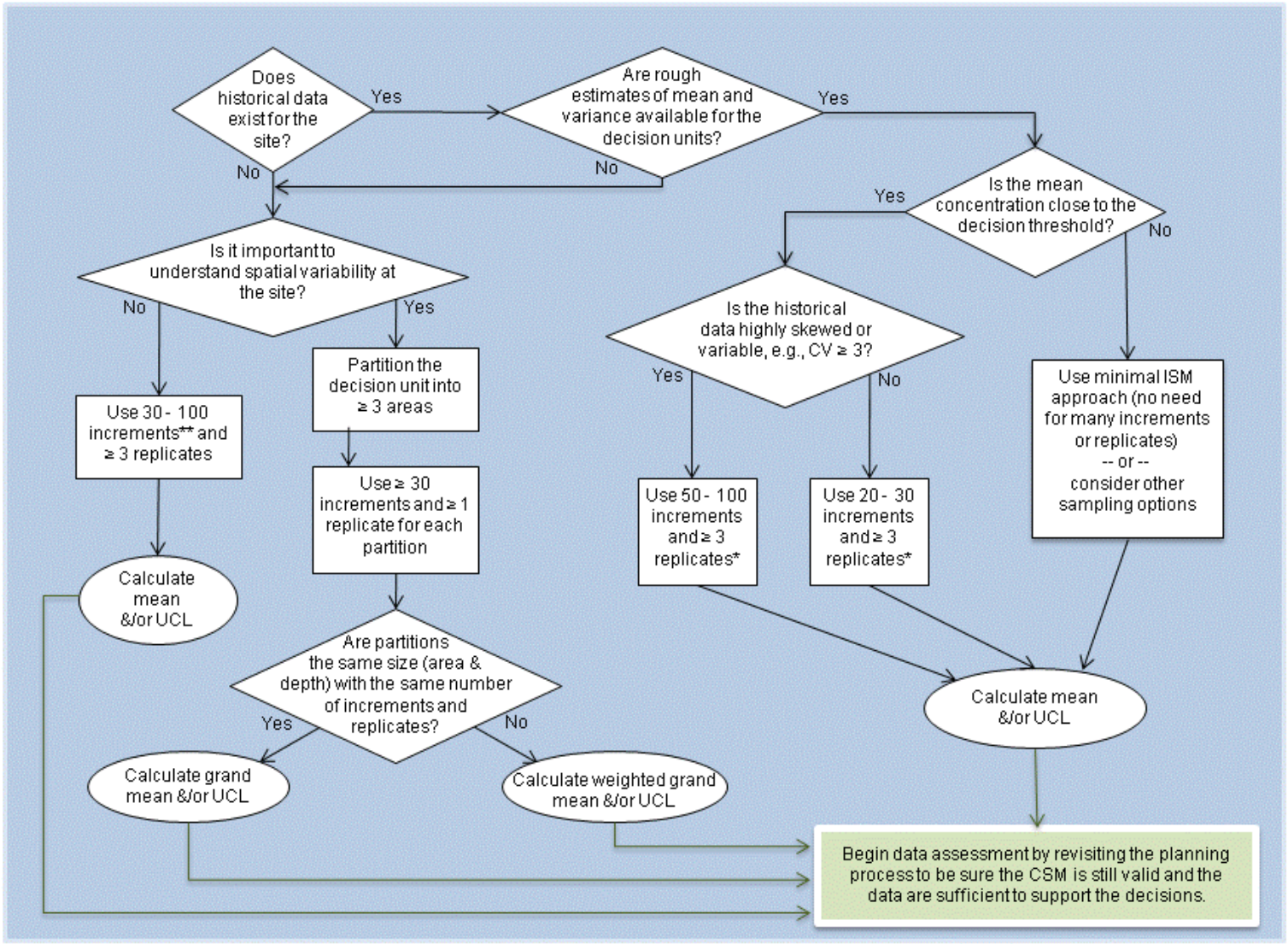
4. STATISTICAL SAMPLING DESIGNS FOR ISM

This section summarizes results of simulation studies used to evaluate the performance of ISM in estimating the mean under various conditions. Conclusions from these studies are discussed, and recommendations for ISM sampling based on this evaluation are presented in Figure 4-1. The recommendations for ISM sampling design, number of increments, and number of replicates as shown in Figure 4-1 come from the simulation studies discussed in this section and presented in more detail in Appendix A.

As mentioned in Section 3, a variety of sampling designs should be considered during systematic planning. To determine which will fully meet the project objectives most efficiently, it is necessary to have some idea of how many samples will be required as part of the design. Figure 4-1 provides guidance on the number of increments and replicates to collect within a DU if incremental sampling is selected as a sampling strategy. Within each DU, the pattern for spatial collection of the increments is not specified in this figure but is discussed in Section 4.3. Methods for estimating the mean concentration present at the site based on incremental samples (as depicted in the ovals in this figure) are also not presented in this figure but are discussed in detail in Section 4.2.



*Collecting more than 3 replicates will increase certainty in estimate of mean and UCL and is recommended in these cases. More than 10 have diminishing value.
**The number of increments depends on heterogeneity (highly variable sites require more increments) and on size (a small site may require fewer increments).

Figure 4-1. ISM decision tree.

After data are collected and reviewed, it is important to revisit the outcomes of the systematic planning process to ensure that the data meet the project objectives. The guidance offered in Figure 4-1 is meant as general guidance for the number of replicates and increments necessary to achieve particular objectives. These recommendations are likely to provide sufficient information to meet most basic objectives relating to comparison of an estimated mean or UCL for the mean to a decision threshold. However, the project team must consider whether or not the project objectives have been fully satisfied by the data collected. If the data are not satisfactory for decision making, further consideration and revision of the systematic planning process and outcomes are necessary.

4.1 Factors that Complicate Estimating the Mean Concentration

ISM sampling produces an estimate of the mean contaminant concentration in soil within a specified volume (i.e., a DU). As with any estimate derived from sampling, ISM results are subject to error, the components of which are described in Section 2.5. Understanding error introduced by sampling is squarely in the domain of statistical analysis. Rigorous statistical analysis regarding the extent to which various ISM sampling strategies provide accurate estimates of the mean contaminant concentration have not yet been published. This information is necessary to understand how factors such as number of increments, number of replicates, and contaminant distributions across the site influence the reliability of ISM estimates of mean contaminant concentration. An evaluation of the reliability of ISM based on statistical principles is vital to widespread acceptance of this sampling method for regulatory purposes.

Statistical evaluation of ISM is a new area. Thorough evaluation of ISM is a substantial undertaking, well beyond the scope of this document. Thus, the findings presented here should be viewed as the best available to date but incomplete in terms of addressing all of the points and questions that might be asked. It is also important to note that analyses described in this report have focused on the extent to which ISM field samples represent the true mean of the DU, assuming that the concentration within those samples can be measured with complete accuracy. Statistical evaluation of subsampling methods in the laboratory is also important (see example in Gerlach and Nocerino 2003) but is not addressed due to time and resource constraints.

> The statistical analysis presented in this document evaluates how ISM *field* sampling procedures may influence the error in the estimate of the mean concentration.

Data on chemical concentrations in environmental media present challenges for estimating the mean concentration. This problem applies to both ISM and discrete sampling. If a DU is perfectly homogenous, meaning that the contaminant exists in the same concentration everywhere across the DU, developing a sampling strategy to accurately estimate the concentration is simple. For that case, all sampling approaches, from a single discrete sample to the most expansive set of ISM samples, would yield the same average concentration (within the limits of laboratory error), and thus any can provide a reliable estimate of the mean. Unfortunately, this ideal situation is never encountered in soils. Site concentrations typically exhibit some degree of heterogeneity, and the greater the heterogeneity, the more difficult it is to accurately estimate the mean concentration through sampling. As discussed in the next section, this difficulty gives rise to error in the estimation of the mean, and different sampling approaches yield different values for the mean. This error can be managed so that reliable estimates of the

mean can be produced, but management requires an understanding of how the number of discrete samples, or the number increments and replicates in ISM sampling, affects estimates of the mean. Simulation studies to develop this understanding are described later in this section.

4.1.1 Skewness and Dispersion

The coefficient of variation, geometric standard deviation, and coefficient of skewness are all measures of dispersion of a distribution.

Both the skewness (asymmetry) and dispersion (spread) in the data can affect the confidence in estimates of the mean. Since it is common for environmental data to exhibit positive skewness (i.e, a longer right tail) or a wide range of concentrations, one challenge for sampling designs is to represent the upper and lower tails of the distribution in the proper proportion, thereby yielding a relatively precise estimate of the mean. For data sets generated with discrete sampling, graphical and exploratory data analysis techniques are commonly used to assess the degree of skewness and dispersion. For example, by plotting the data using histograms and probability plots, the distribution shape and the presence of multiple populations may become apparent. This assessment can be further supplemented by a statistical analysis of the goodness-of-fit (GOF) to normal, lognormal, or gamma distributions. Summary statistics can also be informative and readily calculated with both free and commercial statistics software, including (a) coefficient of skewness; (b) the ratio of the standard deviation (SD) divided by the arithmetic mean—referred to as the "coefficient of variation" or "relative standard deviation" (RSD); and (c) geometric standard deviation (GSD), comparable to the coefficient of variation (CV) (see footnotes of Table 4-1) and used specifically with lognormal distributions.

Table 4-1. Data dispersion in terms of CV and GSD

CV[a] (unitless)	GSD[b] (unitless)	Variability/ dispersion
≤1.5	≤3	Low
1.5–≤3	3–≤4.5	Medium
>3	>4.5	High

[a] Coefficient of variation (CV) = standard deviation (SD)/mean.

[b] Geometric standard deviation (GSD) = exp[sqrt(ln(CV^2 + 1))] for lognormal distributions.

For convenience in this document, the degree of dispersion of the concentration distribution in a DU is classified in terms "low," "medium," and "high," as shown in Table 4-1. These categories can be used to guide the selection of methods used to calculate the UCL in the mean, as discussed in Section 4.2.

As discussed below, the distribution of the contaminant distribution in the DU is different from the distribution of DU means that is characterized by ISM sampling. Table 4-1 provides categories of dispersion for the contaminant distribution throughout the DU rather than the distribution of the DU means. For data sets generated with ISM, fewer exploratory data analysis options are recommended due to the relatively small number of samples. For example, one would not generate a histogram or perform a GOF test on a data set consisting of three replicates. Nevertheless, summary statistics of replicates can provide a measure of the precision in the estimates of the mean, which can be a useful diagnostic for evaluating laboratory DQOs (see Section 4.3.4.4). The mean and variance of the ISM samples can also used to calculate the UCL for

the grand mean. However, as discussed in Section 4.3, the RSD statistic does not serve as a reliable performance metric of a UCL calculation because the true DU mean is never known. In addition, simulations demonstrate that, for data sets in which the sample mean is less than the true mean, the likelihood that the UCL also underestimates the mean increases as the sample RSD decreases, due to the positive correlation between the estimated mean and estimated variance.

4.1.2 Spatial Scale, Mixtures, and Autocorrelation

It is important to recognize that the extent of heterogeneity can vary depending on how DUs are defined. In fact, one way to manage the difficulty of estimating the mean when greater heterogeneity is present is to designate DUs based on anticipated concentrations, defining DUs in such a way as to minimize the concentration variability within each. Other approaches for creating DUs, such as designating DUs according to anticipated exposure patterns (i.e., to correspond with exposure units), could result in greater heterogeneity within the DUs but may be appropriate for risk assessment.

Sampling designs, including designations of sampling units and decisions units, may need to accommodate multiple contaminants with different spatial patterns.

Heterogeneity may be different between contaminants being characterized within the same DU. Different sources or release mechanisms, as well as different transport mechanisms, can lead to differing degrees of heterogeneity among chemicals that need to be addressed through a single sampling plan. This fact can complicate decisions regarding the appropriate sampling approach. In general, the sampling strategy must be designed to accommodate the contaminant expected to have the greatest heterogeneity in order for good estimates of the mean to be obtained for all contaminants of interest.

For both discrete and ISM approaches, random sampling yields unbiased parameter estimates, even when a contaminant exhibits high spatial autocorrelation.

Yet another potential complicating factor is spatial relationships. For most sites, contaminants in soil exhibit some degree of positive spatial autocorrelation, mea
concentration reduces as the distance between sample locations de
that strong autocorrelation can reduce the effective statistical sa
number of samples needed to achieve acceptable decision errors) b
some redundant information (Cressie 1993). In statistical terms,
assumption that observations are independent. ISM confidenc
sampling of a site with high spatial autocorrelation can be too narrow, resulting in a higher frequency of decision errors. Spatial autocorrelation may also introduce bias in estimates of the mean and variance (and corresponding calculations of confidence intervals), depending on the sampling protocol. Random sampling strategies yield unbiased parameter estimates, whereas sampling that is targeted towards areas of suspected high or low concentrations can introduce redundancies that result in inaccurate calculations of confidence intervals and inaccurate estimation of decision errors. For targeted (nonrandom) sampling, the direction of the bias is generally towards overestimation of the mean since suspected source areas may be intentionally oversampled relative to the rest of the site. Nonrandom sampling of sites where contaminants exhibit positive spatial autocorrelation is an issue that applies to discrete as well as ISM sampling. With discrete sampling, spatial weighting methods are sometimes used to reduce the sampling bias. For ISM, spatial weighting methods do not apply since no information is retained from the individual increments collected throughout the DU. Nevertheless, since most ISM sampling protocols incorporate some variation of random sampling and a relatively large number

of increments (i.e., n ≥ 30), spatial autocorrelation is unlikely to impact the statistical performance metrics of ISM (Section 4.3). See Appendix A.3 for an example and additional discussion of this factor.

4.2 Uncertainty in Estimates of the Decision Unit Mean

Even the most comprehensive sampling protocols introduce some degree of sampling error. Therefore, one challenge in developing sampling designs is to balance the potential for decision errors against the practical constraints of site investigations, including having incomplete information about potential source locations, as well as time and budget limitations. The objective of ISM is to provide a reliable estimate of the average (i.e., arithmetic mean) contaminant concentration in a DU, recognizing that any individual ISM sample may over- or underestimate the mean to some degree. This sampling error may be attributed to a variety of factors. A principal objective of systematic planning of most sampling designs is to minimize the major sources of error in both the field and the laboratory. In practice, the estimated variance is often viewed as an overall measure that includes the contribution of many sources of error. Just as with discrete sampling, the estimated variance can be used to quantify a UCL for the mean for ISM samples and the same UCL equations apply. This section describes important concepts relevant to characterizing variance in ISM sampling. Section 4.3 builds from these concepts by presenting the results of simulation studies that examine the performance of alternative ISM sampling strategies applied to a wide range of theoretical site conditions.

4.2.1 One ISM Result

For sites where there is a regulatory requirement to calculate a UCL, at least three replicates should be collected within a DU. For sites where there is no regulatory requirement to calculate a UCL, it is important to understand the potential for decision errors if a decision is to be informed by a single ISM result. Two critical components to a decision error are the likelihood of underestimating the mean and the magnitude of the underestimation.

Each ISM sample provides an estimate of the true mean—the actual average concentration within the DU. As such, the distribution of ISM results is related to but conceptually different from the distribution of discrete samples. The two approaches share the same grand mean but can be expected to have different estimates of variance. For ISM, the mean of replicates is analogous to repeated trials of discrete sampling (i.e., the mean of the means, or the "grand mean"), and the standard deviation is analogous to the standard error for the mean in discrete sampling. Even the most comprehensive sampling protocols will introduce some degree of sampling error, and it is possible that a single ISM sample result can be well above or well below the true mean. The magnitude of the under- or overestimate depends on the overall heterogeneity of the underlying distribution, increasing as the heterogeneity increases. Figure 4-2 illustrates the probability and magnitude of underestimation of a single ISM sample of n=30 increments collected from DUs with underlying lognormal distributions with CVs ranging 1.0–3.0. The following observations are noted:

- A single ISM sample will underestimate the mean more than 50% of the time for all positively skewed distributions.

A single ISM result is likely to underestimate the mean more than 50% of the time for most distributions; the likelihood of a decision error increases as the variance in the distribution increases and the difference between the action level and true mean decreases.

- The magnitude of the underestimation depends on the degree of variability, as represented by the CV.
- Approximately one-third of the sampling events with a single ISM sample (n = 30 increments) will underestimate the mean by up to 10% for CV = 1 and 20% for CV = 2. For example, if the true mean is 400 ppm, approximately one out of every three ISM samples (n = 30) will yield an estimated mean <360 ppm for CV = 1, and <320 ppm for CV = 2.
- For a distribution with greater dispersion (i.e., CV = 3), approximately one quarter of the sampling events will yield a single ISM result that underestimates the mean by 30%–60%. For example, if the true mean is 400 ppm and CV = 3, approximately one out of every four ISM samples (n = 30) will yield a sample mean 160–280 ppm.

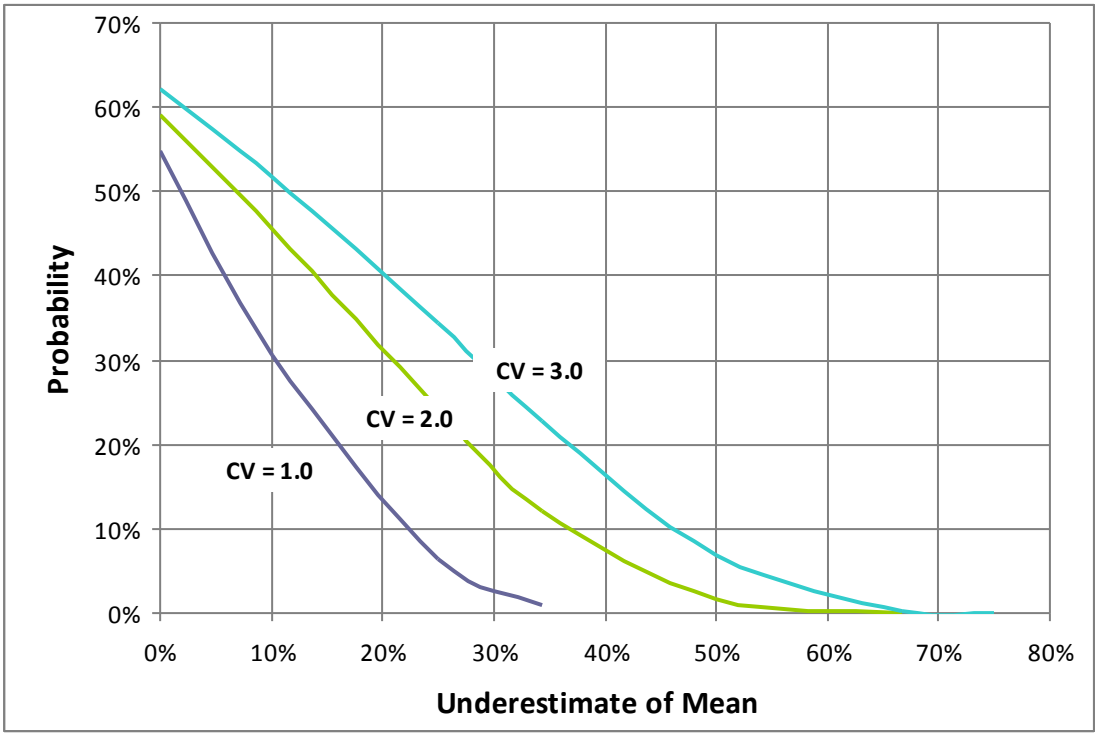


Figure 4-2. Examples of the probability and magnitude of underestimation of the mean from a single ISM sample.

The same issues with underestimation apply when discrete sampling is used to estimate the mean. As heterogeneity of the soil concentrations increases and sample size decreases, the potential magnitude of error in the estimate increases. Consider what would happen if you sent crews out to the same DU 100 times to collect an ISM sample of 30 increments or a series of discrete samples with which to calculate a mean concentration. If the separated estimates of the mean from these sampling events were plotted as a histogram, they might fit distributions shown in Figure 4-3. The top panel shows estimates of the mean that are normally distributed around the true mean of 100. Given that it is a normal distribution, the estimated mean of approximately half of the sampling efforts is below the true mean and half of the efforts produced an estimated mean above the true mean. The spread of the distribution gives an indication of how far away from the true mean some of the estimates were (i.e., an indication of the potential magnitude of

error). As the top panel illustrates, although both distributions are unbiased (centered at the true mean), variability in estimates of the mean are generally less for ISM than for comparable discrete samples due to differences in number of samples collected.[2] The lower panel in Figure 4-3 shows that the potential magnitude of error increases as the estimates of the mean becomes skewed due to heterogeneity.

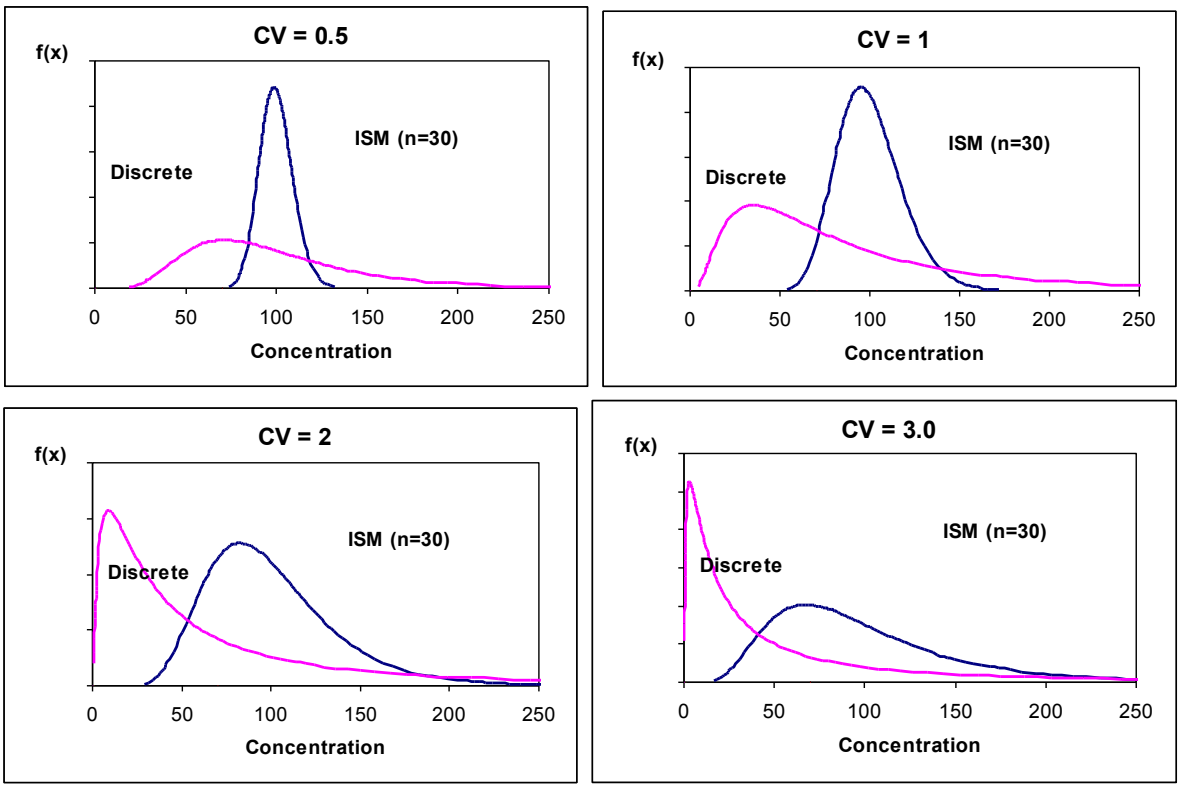


Figure 4-3. Dispersion of means from ISM (based on n = 30 increments) applied to a lognormal distribution (mean = 100) with CVs ranging 0.5–3.0.[3]

From a statistical standpoint, it should be noted that analysis of multiple ISM samples collected with the same sampling protocol (i.e., sampling method and number of increments) provides a direct measure of the variance in the mean. It is important to recognize that the distribution of replicate results is different from, but related to, the distribution of discrete results (X_i) obtained from the same population. As shown in Figure 4-3, both sampling types share the same estimate of the population mean (i.e., 100) but not the same variance. The variance of ISM samples composed of *n* increments is lower than the variance of discrete samples comprised of *n* discrete measurements. While this example is an oversimplification of the differences between ISM and discrete sampling, it

Discrete and ISM samples yield different distributions for the mean. They share the same (grand) mean but have different shapes and variances.

[2] Note that the distribution of ISM means (from repeated trials of one ISM sample) and means estimated from discrete samples would be expected to be equivalent if the number of discrete samples was the same as the number of increments.

[3] ISM generates a distribution of means that approaches normality, as predicted by the central limit theorem. However, the ISM distribution can also by asymmetric, and the asymmetry increases with increasing dispersion of the underlying distribution.

highlights an important statistical concept related to sampling from populations.

In practice, you can't send a crew out to sample the same DU 100 times and assess the variability. Instead, you typically have to obtain a reliable estimate of the mean through a single sampling exercise. Through understanding the concept of variability in estimates of the mean and the influence of heterogeneity, the limitation of basing a decision on a single ISM sample becomes apparent. There is no way to know whether any one estimate provided by a single sampling event is above or below the actual mean in the DU as well as the potential magnitude of the deviation from the actual mean without additional sampling data to assess heterogeneity of the concentrations within the DU.

By collecting multiple ISM samples within a DU (i.e., ≥3 replicates), we can obtain a direct measure of variability in sample means and calculate a UCL for the mean with an acceptable decision error rate.

Recognizing that variability and errors in estimates of the mean exist, regulatory agencies often require a 95% UCL to represent an EPC or to assess compliance with decision criteria. Just as with discrete sampling, the variance for replicate ISM samples can be used to estimate the standard error for the mean, which is one factor in the calculation of a UCL as discussed below. Similar to the difference in discrete and ISM variance estimates described above, the UCL calculated from ISM replicates is generally different (and lower) than the UCL calculated from discrete samples with typical sample sizes. In the case of ISM, the UCL can be thought of as a measure of the confidence in the estimate of the "grand mean," or the mean of the means given by replicate samples. In practice, it is expected that a typical ISM sampling protocol will consist of a relatively small number of replicates (e.g., three to five replicates). The small number of samples may have several implications on the performance of the ISM sampling effort, depending on the properties of the contaminant distribution at a site (e.g., heterogeneities, spatial patterns in the distribution, etc.).

With ISM, the UCL can be thought of as a measure of the confidence in the estimate of the "grand mean," or the overall mean of the individual means given by each replicate sample.

4.2.2 UCL Calculation Method

The concept of variability in estimates applies to UCLs as well as to the estimates of the means themselves. Several methods exist for calculating a UCL for estimates of the mean for a set of data. These methods often yield different answers for the same set of data. For example, if a 95% UCL[4] is estimated for a population 100 times, the 95% UCL will, on average, be greater than or equal to the true mean of the population 95 times. The ability of different methods to produce a value that meets the definition of a 95% UCL depends in part on the number of samples used to estimate the mean, as well as the distribution (e.g., normal, lognormal, gamma) and dispersion of the data. One method might generate 95% UCLs greater than or equal to the true mean for a population 95% of the time, while another 95% UCL method might generate estimates greater than the true mean only 80% of the time. In the latter case, although a 95% UCL method was used, that method did not perform up to the specified level for that population. Had more

[4] Note that throughout this document a UCL on a mean estimate is presented as 95% UCL. It is important to note that this is only an example of a UCL. It is possible to use a 90% UCL, 98% UCL, 99% UCL, etc. The specific UCL used should be determined by the project team during systematic planning.

samples been taken to estimate the mean or if the concentrations were distributed differently, the second method might have performed satisfactorily while the first method was deficient.

In practice, we cannot compare the performance of any UCL calculated at a site because the true mean within the DU is unknown. Similarly, there are no statistical calculations or diagnostics that can be used to compare the individual replicates or UCL to the unknown mean. These are limitations that apply to both discrete and ISM sampling. However, the likely performance of alternative UCL methods can be explored using simulation studies. Such studies have already been conducted by USEPA (2010b) to guide in the calculation of 95% UCLs for discrete sampling protocols. This type of performance evaluation has not been previously conducted for ISM sampling, so initial simulation studies were conducted in the development of this guidance, as summarized in Section 4.4.

In practice, the true mean is unknown, but with simulation we can define the mean. Simulation studies help guide the selection of a UCL method based on simulation-specific information, assumptions, and decision error criteria.

Three or more ISM samples are needed to calculate a 95% UCL. In theory, all of the UCL methods that are applied to discrete sampling results can also be applied to ISM. In practice, however, because fewer than eight replicate ISM samples are likely to be collected for a DU, fewer options are typically available to calculate a UCL compared with discrete sampling data. The small number of replicates precludes GOF evaluations as well as the use of methods that require more samples than typically collected in ISM sampling (USEPA 2010a). Therefore, the options for UCL calculations reduce to the set of methods that require only the parameter estimates themselves: mean and SD. Two candidate UCL equations that can accommodate ISM data sets and which are expected to "bracket" the range of UCLs that may be calculated from a data set are the Student's-*t* (representing the low end of the range) and Chebyshev (representing the high end of the range) UCLs as discussed in Section 4.2.2.1.

Two UCL calculation methods were evaluated for use with ISM samples:

- Student's *t* UCL
- Chebyshev UCL

The online version of this document contains a working calculator for these methods: http://www.itrcweb.org/ISM-1/4_2_2_UCL_Calculation_Method.html.

4.2.2.1 UCL of the mean based on Student's-t distribution

The following equation is used to calculate the one-sided (1–α) 100% UCL using the Student's-*t* approach:

$$UCL = \overline{X} + t_{(1-\alpha)(r-1)} \times \frac{S_{\overline{X}}}{\sqrt{r}}$$

where

$\overline{X}$ = arithmetic mean of all ISM samples
$S_{\overline{X}}$ = standard deviation of all ISM samples
r = number of ISM samples
t = $(1-\alpha)^{th}$ quantile of the Student's-*t* distribution with $(r-1)$ degrees of freedom

The Student's-*t* UCL is expected to provide valid 95% UCL values when the distribution of means is approximately normal. The central limit theorem (CLT, Casella and Berger 2001) provides support for the use of a Student's-*t* UCL for composite sampling as well as ISM sampling. The CLT is useful because it defines the distribution of the mean of the samples without having to know the exact underlying distribution of the data. The number of samples, *n*, and the shape of the distribution of the data are the two factors that most influence the accuracy of the approximation of the distribution of the mean. For approximately symmetric or slightly skewed distributions, a relatively small number of samples (e.g., n = 15) may be sufficient for the estimates of the mean to be approximately normally distributed as theorized by the CLT. If the population distribution is moderately skewed, a larger number of samples (e.g., n ≥ 30) is required to reliably invoke the CLT (Casella and Berger 2001). More highly skewed distributions require even larger numbers of samples. When the distribution of replicate samples is right-skewed instead of normal, the consequence of using the Student's-*t* UCL is that it will underestimate the true mean more often than desired.

In ISM sampling, the coverage of the Student's-*t* UCL also depends on the SD of the ISM replicates. The influence of the combination of factors for different sampling regimes can be difficult to anticipate. The simulation results in Section 4.3 demonstrate various performance metrics associated with the use of the Student's-*t* distribution for a wide range of plausible scenarios.

4.2.2.2 UCL of the mean based on Chebyshev inequality

The following equation is used to calculate the one-sided (1–α) 100% UCL using the Chebyshev approach:

$$UCL = \overline{X} + \left(\sqrt{1/\alpha - 1}\right) \times \frac{S_{\overline{X}}}{\sqrt{r}}$$

where

$\overline{X}$ = arithmetic mean of all ISM samples
$S_{\overline{X}}$ = standard deviation of all ISM samples
r = number of ISM samples

The Chebyshev is generally considered to be a conservative estimate of the UCL because it generally achieves or exceeds the desired coverage rates, even for nonnormal distributions. However, with small numbers of samples, the estimates of the mean and SD can be unstable, and the coverage is not guaranteed, especially when the contaminant distribution in the DU is highly heterogeneous. Each simulation discussed in Section 4.3 includes performance metrics for both the Student's-*t* and Chebyshev UCLs to illustrate conditions in which each may be favored.

4.2.3 Nondetects

While nondetects are relatively common for discrete sampling data sets due to spatial heterogeneities in the mechanisms of release of contaminants, it is less likely that an ISM result will be below the reporting limit because some small percentage of the increments are expected to capture source areas. In other words, while individual increments may be nondetect, it is unlikely that the mean of the increments (given by an ISM result) will be nondetect. Exceptions may include constituents that are below reporting limits under ambient (background) conditions and are unrelated to site activities or post-remediation confirmation samples. In both cases, so long as the reporting limits do not approach action levels, nondetect results should not introduce decision errors. If reporting limits approach action levels, users should consider alternative analytical procedures, revisions to the sampling design to characterize the DU, and other lines of evidence that concentrations might be below action levels. If replicate results include a mix of detects and nondetects, then the only option with small sample sizes is to apply substitution methods such as one-half reporting limits to ISM results and to qualify results as potentially biased due to the use of substitution methods. A variety of substitution methods may be applied and the consequences of those options should be explored.

4.3 Evaluating the Performance of Sampling Approaches

This section describes studies used to evaluate the performance of various ISM sampling strategies in providing accurate estimates of the mean and 95% UCL. Metrics used to evaluate performance and the approach used in the simulation studies are described.

4.3.1 Definitions of Performance Metrics

Performance metrics provide a way to systematically evaluate and compare various sampling strategies, including sampling pattern, statistical sample size (both number of increments and replicates), and UCL computation techniques. Collectively, these results can help to establish an optimal decision process for using ISM given a particular set of site conditions and decision criteria. The following four metrics are defined below and evaluated through simulation studies:

Performance metrics for a 95% UCL that can be evaluated when the true mean is known (or assumed):

- UCL coverage
- relative percent difference between UCL and true mean
- bias
- relative standard deviation of replicate means

- coverage of the UCL
- magnitude of UCL deviation from the mean (i.e., RPD between UCL and true mean)
- bias of the mean of the samples
- RSD of the estimates of the mean from each replicate ISM sample

4.3.1.1 Coverage and magnitude of UCL errors

In repeated trials, an appropriate 95% UCL should exceed or "cover" the true mean 95% of the time. In practice, we never know how well a 95% UCL has performed in terms of coverage[5] because the true mean is unknown. However, in simulation studies we have the opportunity to repeatedly evaluate a theoretical DU for which the mean is known and compute many UCLs. Accordingly, coverage is defined in this context as the percentage of the simulations for which the 95% UCL actually exceeds the true DU mean. As an example, Table 4-2 gives selected results for a simulation with 5000 trials where the mean and 95% UCL were calculated by sampling a lognormal distribution with mean = 100 and SD = 200. The 95% UCL for each trial was based on a Chebyshev equation applied to sample statistics for 3 replicates of 30 increments. The values from the UCL column are then compared to the true mean of 100. If the design were built to theoretically have 95% confidence that the true mean was less than the calculated UCL, then the ideal result from the 5000 iterations would be to find approximately 5% (i.e., 250 of 5000) of the UCL values are below the true mean. Figure 4-4 shows a histogram of the 5000 UCL values from this simulation where the y-axis represents the fraction of total iterations in each bin. In this example, the UCL histogram shows that approximately 5% of the UCL values are below the true mean. This exercise shows that the UCL coverage for this simplified scenario met the design criteria. It is interesting to note that the grand mean of 3 replicates underestimated the true mean nearly 60% of the time (in contrast to the 95% UCL underestimating the mean only about 5% of the time), exemplifying why the UCL is often used to protect against underestimation of the true mean.

For positively skewed distributions (e.g., lognormal), the mean of the ISM samples will underestimate the population mean >50% of the time, whereas the 95% UCL will typically underestimate <5% of the time.

Table 4-2. Example of UCL simulations

Trial	Mean	UCL	RPD
1	64.7	85.0	–15%
2	61.7	102.7	2.7%
3	100.7	105.2	5.2%
4	90.8	107.0	7.0%
⋮	⋮	⋮	⋮
4999	96.1	215.3	115.3%
5000	253.2	855.0	755.0%

RPD = [(UCL – 100)/100] × 100%.

[5] Note that this concept is completely separate and unrelated to that of spatial "coverage" as applied to areal representativeness of samples taken over a DU.

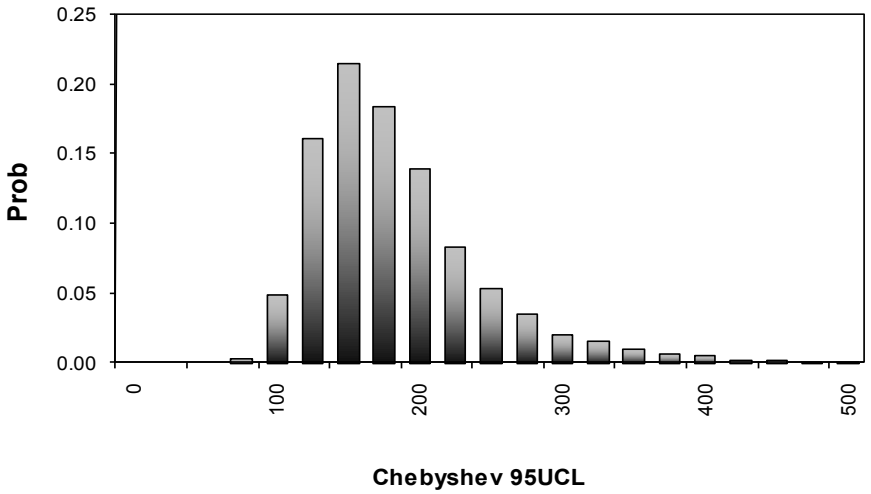


Figure 4-4. Histogram of calculated Chebyshev UCL values using 5000 trials of a lognormal distribution (mean = 100, SD = 200), 30 increments, and 3 replicates.
The "true mean" (100) is exceeded in approximately 95% of the trials.

The optimal methodology for calculating a UCL should provide adequate coverage of the mean and produce a UCL that is not unduly large. The magnitude of difference between the UCL and the true mean can be expressed as the RPD defined as follows:

$$RPD = [(95\%\ UCL - \mu)/\mu] \times 100\%$$

As shown in Table 4-2, RPD may be negative or positive depending on whether or not the UCL exceeds the true mean. RPD may be calculated for all UCL results or can be calculated for those UCLs that fall above (RPD_A) and below (RPD_B) the true site mean, separately. When used for just those UCLs that fall below the site mean, the RPDs reveal the magnitude of the potential underestimation. This calculation is particularly informative in situations where the coverage does not meet the specified 95% criteria.

Figure 4-5 illustrates examples of RPD_A and RPD_B for simulations using lognormal distributions with CV = 1 and CV = 4. Each simulation represents 5000 trials using 30 increments (m) and 2, 3, 5, or 7 replicates (r). Results for both the Chebyshev UCL and Student's-*t* UCL are given side by side. Error bars represent the 5^{th} and 95^{th} percentile RPD values, and the point in the center corresponds to the median. For example, for CV = 1 and r = 3, the Chebyshev UCL generally exceeds the true mean by less than 50% and underestimates by less than 10%. The deviation of the UCL using Student's-*t* is slightly lower for the overestimates and comparable for the underestimates. For CV = 4 and r = 3, the magnitude of the deviations increases for both the Chebyshev UCL (95^{th} percentile RPD_A of 214% and RPD_B of –23%) and Student's-*t* UCL (95^{th} percentile RPD_A of 160% and RPD_B of –25%). Information on coverage and RPD ranges can be combined to yield the following observations:

- Even for distributions with high variance (e.g., CV = 4, r = 3), the 95% UCL using either Chebyshev or Student's-*t* equations can be expected to yield values that exceed the true mean by no more than 150%–200% and underestimate by less than 25%;
- Student's-*t* UCL more frequently underestimates the true mean than does the Chebyshev UCL.
- The magnitude of the underestimate (RPD_B) will be comparable; however, the magnitude of the overestimate (RPD_A) will be greater for the Chebyshev UCL.

The Student's-*t* UCL and Chebyshev UCL provide estimates of the mean that, even for highly variable distributions, generally exceed the true mean by no more than 200% or underestimate the mean by no more than 25%.

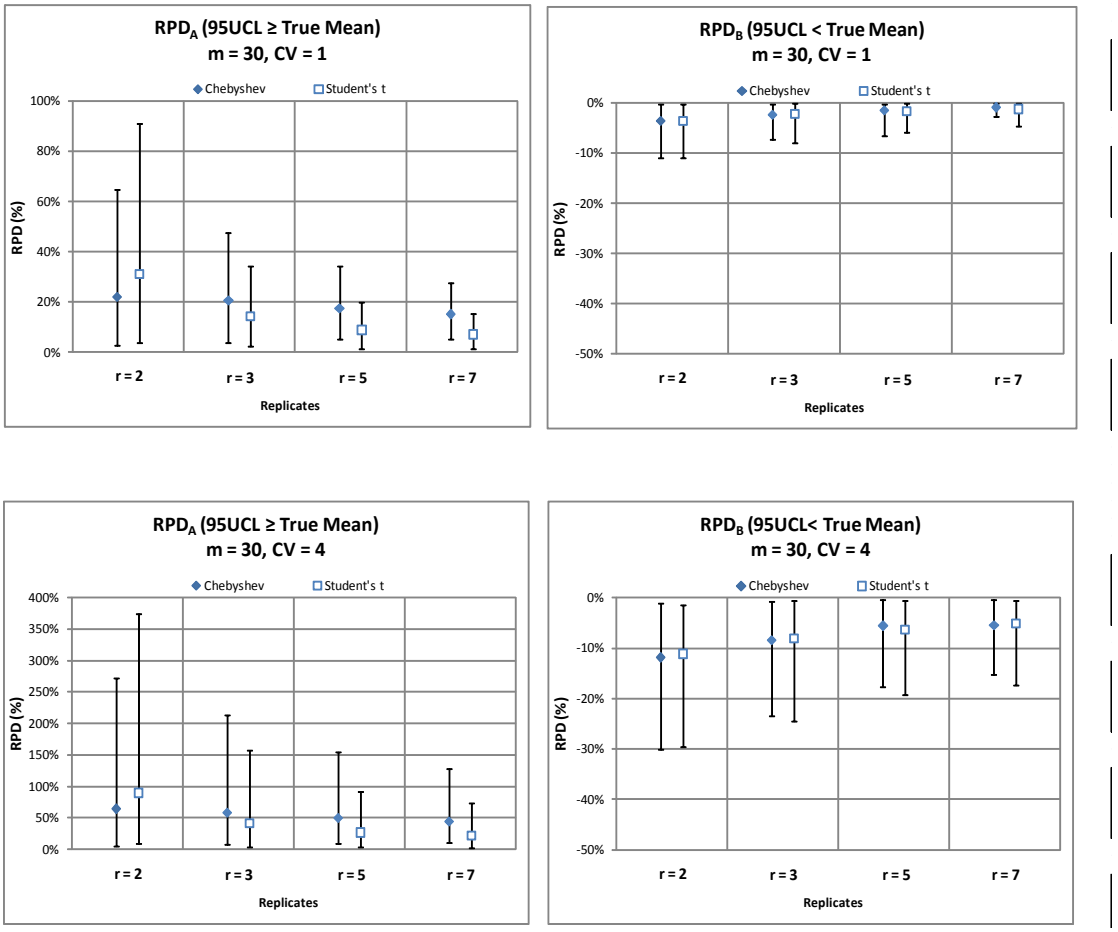


Figure 4-5. Range of overestimation (RPD_A) and underestimation (RPD_B) of 95% UCLs using Chebyshev and Student's-*t* calculation methods for ISM simulations with lognormal distributions (CV = 1 and CV = 4), 30 increments, and 2–7 replicates. Error bars represent 5th and 95th percentiles of 5000 trials.

ISM replicates tend to produce UCLs with smaller RPD_A and RPD_B than a corresponding data set of discrete samples. This desirable quality of ISM is due to the physical averaging of the

It is unlikely that one 95% UCL method excels at all performance metrics. In addition, performance can vary depending on site characteristics. Method selection requires balancing the importance of each metric.

individual increments. Therefore, ISM UCL values may provide reasonably reliable estimates of the site mean even when the desired 95% coverage is not achieved but RPD_B is minimal.

In general, all other conditions being the same, as the number of increments and replicates increases, the error is expected to decrease. This decrease in the standard error will be reflected by an improvement in bias, the coverage and RPD of the UCL. The influence of these components of the sampling design varies depending on characteristics of the population sampled (e.g., magnitude of DH, single or multiple populations) and the sampling method (e.g., systematic random sampling, random sampling with grid, or simple random sampling). The central concept governing the optimization of the sampling design is that while initial increases in the number of replicates and increments improve estimation, there are diminishing returns with increasing numbers of samples. At some point, increasing the number of samples is unlikely to yield an appreciable improvement in either the coverage of the UCL or the magnitude of the over/underestimate of the UCL as indicated by the RPD calculations.

4.3.1.2 Bias in estimated mean

"Bias" is defined here as a systematic over- or underestimation of the true site mean. Bias is generally introduced when the sampling strategy or sample collection method yields observations that are not truly independent or representative of site conditions. For example, use of a systematic sampling pattern that coincides with spatial trends in the data may produce a data set that disproportionately represents a range of concentrations; poor sample collection techniques may underrepresent actual soil characteristics.

Accuracy reflects a combination of precision (reproducibility) and bias (systematic over/underestimation). The RSD of replicate ISM means is a measure of precision.

4.3.1.3 Relative standard deviation of replicate samples

The reproducibility of ISM replicates collected from a DU can be evaluated in terms of RSD, also known as the coefficient of variation (CV), which is the SD divided by the mean. Because both RSD and CV are commonly used, both are used interchangeably here. Although included as a performance metric in the simulation studies, the RSD does not provide an indication of the accuracy of the estimate of the mean or 95% UCL. Figure 4-6 illustrates the distinction between reproducibility (or precision) vs. bias, which, taken together, represent accuracy. For example, a low RSD indicates the estimates are precise. The values might be reproducible but still yield a biased estimate of the mean and corresponding UCL.

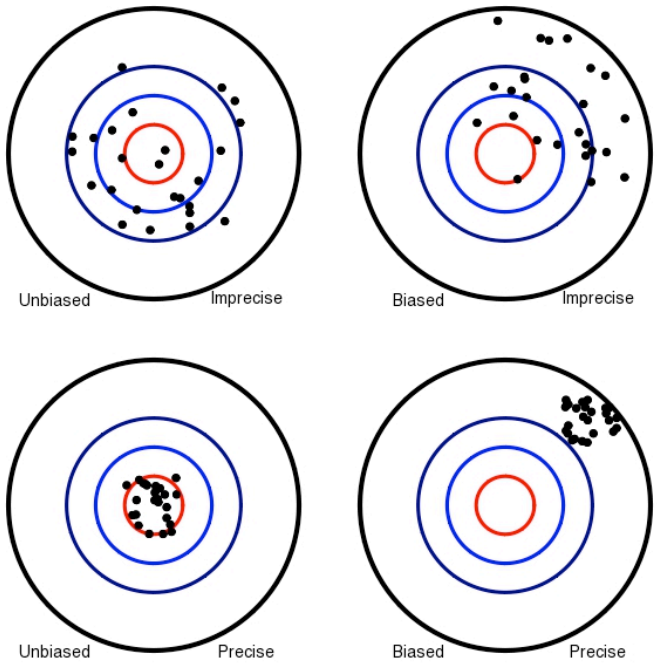


Figure 4-6. Four possible relationships between bias and precision.

4.3.2 Simulation Study Approach

A computer-based simulation is a numerical experiment in which a hypothetical site is sampled many times. The key utility of simulations is that the contaminant distribution can be specified so the population parameters are known. This is in contrast to actual sampling with ISM, in which the potential bias in results or coverage of the 95% UCL cannot be quantified since the true mean is not known. Simulation is a convenient tool to evaluate the performance of alternative ISM approaches based on comparisons to the true mean. Furthermore, a variety of different incremental sampling and statistical methods can be simultaneously applied to the same exact scenario to facilitate a comparison of sampling strategies. Each simulation followed this general five-step process:

Simulation studies were conducted to determine the performance of several aspects of ISM:

- number of increments and replicates
- sampling pattern
- heterogeneity and variability of concentrations

1 Define the population. This may be a probability distribution (e.g., lognormal or Gamma), a mixture of probability distributions, or a 2-D map of the concentration surface of a DU. For some scenarios, CH and DH may be explicitly defined, while for others the assumption is that the population variance represents the combination of both elements of heterogeneity and other sources of error.

2 Define an ISM sampling strategy. This step identifies the size and placement of DU, number of increments, sampling strategy (e.g., systematic random, random sampling within grid, simple random sampling; see Section 4.3.4.2 for more description), and number of replicates.

3 Implement a Monte Carlo analysis (MCA). Using MCA (described below), repeat the same ISM sampling strategy many times (e.g., 2000 iterations or more).

4 Calculate statistics. For each iteration of MCA, calculate the DU statistics, including the grand mean (i.e., mean of replicate samples), RSD of the replicate samples, bias in mean (i.e., estimated mean minus population mean), and 95% UCL using Student's-*t* UCL and Chebyshev UCL.

5 Evaluate performance metrics. In this step, the statistics are used to evaluate performance metrics, including coverage of 95% UCL, magnitude of UCL error, bias of the means, and RSD.

Using simulation, we can evaluate a variety of different statistical properties of ISM and determine if factors that can be controlled in the sampling design (e.g., number of increments, number of replicates, DU size, and use of multiple SUs) can be adjusted to achieve the sampling objectives. Furthermore, by running MCA simulations on a variety of different scenarios, we can develop an understanding of the alternative ISM sampling strategies under different conditions. For example, 30 increments and 3 replicates may be sufficient to obtain a reliable 95% UCL for a DU that is described well by a single probability distribution with relatively low DH, whereas greater numbers of samples may be needed for a DU with multiple overlapping contamination sources and relatively high CH and DH. Pitard (1993) highlights the value of summarizing such relationships with sampling nomographs, which are the "best available tool to quickly estimate

sampling errors, effectively design optimum sampling procedures, and find reasonable economical compromises."

Simulations can be used to determine the performance of ISM under very specific conditions and, therefore, the results cannot be expected to apply to all sites. Table 4-3 provides details regarding the range of conditions that have been investigated and summarized in this document.

Table 4-3. Summary of scenarios investigated with simulations

Condition	Levels
Increments	15–100
Replicates	2–5
Sampling method	Simple random sampling, random within grid, and systematic random
Sampling pattern	Entire DU and subdivided DU
Range of symmetry and dispersion	Normal data and multiple skewed data sets (lognormal and Gamma) with CV ranging 0.7–6
DU variability	Homogenous and multiple levels of heterogeneity
DU spatial patterns	Ranged from evenly distributed to localized elevated regions of differing sizes

A comprehensive review of the performance of discrete sampling methods for 95% UCL calculations already exists (USEPA 2010b) and was not evaluated here.

4.3.3 Objectives of the Simulation Studies

The objective of the simulation studies was to address several issues of practical importance in obtaining and using ISM data. As noted above, simulation studies have the unique advantage of evaluating the performance of ISM in estimating the mean under a variety of conditions where the right answer (i.e., the true mean) is known. Thus, they are the best, and in fact the only, way that the accuracy of ISM estimates of the mean concentration can be assessed.

Some of the simulation studies were directed to the basic design of an ISM sampling event (the number of increments) and the pattern in which the samples are taken within a DU. The accuracy of ISM estimates and 95% UCL coverage based on differing numbers of increments, replicates, and sampling patterns were evaluated with attention to bias and magnitude of error (i.e., RPD). Simulation studies evaluated different approaches for computing a 95% UCL using ISM data. Performance of sampling methods was evaluated in terms of coverage provided by a 95% UCL, as well as the extent of overestimation of the mean (RPD_A). Ideally, a calculation method yields a 95% UCL with adequate coverage without excessive overestimation of the mean.

Simulations conducted with hypothesized distributions (e.g., lognormal) did not attempt to distinguish between different sources of error (see Section 2.5) or real-site complexities associated with spatial patterns such as mixtures (i.e., multiple sources of contamination at a site) or hot spots (i.e., sources with elevated concentrations that occur in relatively small subareas across the DU). By sampling from a single lognormal distribution, the simulations do not explicitly address inherent heterogeneities (e.g., CH and both small- and large-scale DH). However, these simulations are particularly applicable for scenarios in which the contamination

is expected to be homogeneous through the DU (meaning that the mean and variance are the same in all subareas) and simple random sampling is applied. These simulations provide a convenient framework to begin to evaluate the performance of different UCL calculation methods with different sampling designs (i.e., numbers of increments and replicates) under a range of skewness and variance in the distribution. The simulations with maps extend the evaluation by exploring the effect of sampling methods (e.g., systematic or simple random) on bias in parameter estimates as well as the effect of DU heterogeneity on the performance of the 95% UCL.

4.3.4 Simulation Study Findings on ISM Performance

The following sections summarize conclusions from the simulation studies. Where possible, results are expressed in terms of the performance metrics outlined in the previous section.

4.3.4.1 Sample size (number of increments and replicates)

One option for reducing errors in sampling designs is to increase the sample size. For ISM, sample size can pertain to the mass per increment (i.e., sample support), number of increments (*n*), and number of replicates (*r*). Assuming a uniform mass per increment, several observations were made regarding the effects of increasing *n* and *r* on estimates of the mean (also see Appendix A, Table A-1):

- Increasing the *n* has a direct effect on the standard deviation of the replicates. Specifically, the central limit theorem suggests the standard deviation of the replicates (which is a measure of the standard error of the mean) reduces by a factor of the square root of *n*. For example, all other things being equal, if the SD of replicates is 4.0 with n = 30, doubling the increments to n = 60 would reduce SD by the square root of 2 (or 1.414) to approximately 2.8.

- Increasing *r* does not reduce the standard deviation of the replicates although it does improve the estimate of the SD by reducing the variability in the estimate. Likewise, increasing *r* reduces the standard error for the grand mean. Specifically, the standard error reduces by the square root of *r*.

- The overall reduction in the standard error for the (grand) mean is a function of the *total* mass collected and spatial area represented (i.e., increments × replicates), and this observation applies to parameter estimation with discrete sampling as well.

- Increasing the number of increments (n) or sample mass reduces the potential for errors in terms of both frequency and magnitude of underestimation of the mean.

Increasing the number of increments and/or replicates reduces the variability in ISM estimates of the mean.

- For nonnormal distributions, increasing *r* above 3 provides marginal return in terms of improving coverage of a UCL when the Chebyshev calculation method is used; however, increasing *r* does not improve coverage of the Student's-*t* UCL.

The difference between Chebyshev and Student's-*t* UCLs can sometimes lead to different decisions for a DU. While the Chebyshev method typically provides greater coverage, it also tends to have higher RPDs. Project teams must balance both properties of UCLs when deciding which method(s) to use.

- Increasing *r* reduces (i.e., improves) the RPD, meaning it will produce estimates of the 95% UCL closer to the DU mean. Therefore, increasing *r* may be an important sampling strategy when errors of either underestimation or overestimation of the mean can have significant consequences.

- Simulations produced varying results in terms of improvement in coverage by increasing the number of increments (n). In some simulations, increasing n produced little or no observable difference. In others, n twofold or more from typical increment numbers used in ISM resulted in marginal improvement. As with increasing replicates, increasing n decreases (i.e., improves) the RPD. The improvement in RPD performance is marginal when the underlying CV is small.

- Simulations showed that coverage provided by the two UCL calculation methods depends upon the degree of variance (or dispersion) of the contaminant distribution within the DU. A variety of statistics provide a measure of dispersion including the CV (i.e., SD normalized by the mean) and the geometric SD (specific to lognormal distributions). Table 4-4 summarizes findings grouped by CV (and GSD). Note that in this case, the CV reflects the SD of the increments divided by the mean and not the SD of the replicates divided by the mean. In practice, individual increments are typically not retained for analysis, so there may be no direct measure of the CV. If there is no site knowledge available to support an assumption about the degree of dispersion (i.e., low, medium, high) of increments, then the Chebyshev UCL may be the preferred calculation method because it is more likely to achieve the desired coverage than the Student's-*t* UCL. The CV (or SD) of the replicates is not a useful metric for determining which UCL method provides sufficient coverage.

Table 4-4. Likelihood that ISM achieves coverage depending on dispersion

UCL Method	Dispersion among individual increments		
	Low (CV <1.5 or GSD <3)	**Medium (1.5 < CV < 3 or 3 < GSD < 4.5)**	**High (CV >3 or GSD >4.5)**
Student's-*t*	Yes	No	No
Chebyshev	Yes	Yes	Maybe

Coefficient of variation (CV) = standard deviation (SD)/mean.
Geometric standard deviation (GSD) = exp[sqrt(ln(CV^2 + 1))] for lognormal distributions.

- The Chebyshev method always produces a higher 95% UCL than the Student's-*t* method for a given set of ISM data with r > 2. When both methods produce specified coverage, the Chebyshev consistently yields a higher RPD.

4.3.4.2 Effects of sampling pattern

Just as with discrete sampling, a variety of sampling methods may be implemented with ISM sampling. One of the more common approaches in ISM is systematic random sampling (a.k.a., systematic grid sampling [Gilbert 1987]), where the DU is divided in a grid pattern, a random sampling location is

Simple random sampling, systematic random sampling, and systematic grid sampling yield unbiased estimates of the mean. The systematic sampling patterns ensure relatively even spatial distribution of samples across the site and are generally easier to implement in the field.

identified within the first grid cell, and then samples (increments) are obtained from adjacent cells sequentially in a serpentine pattern using the same relative location within each cell (Figure 4-7). Another approach is random sampling within a grid (also called "stratified random sampling" [USEPA 1995b]), wherein samples are obtained sequentially from adjacent grid cells, but the location of the sample within each cell is random (Figure 4-8). A third approach is simple random sampling, where the samples are taken from random locations across the DU (without gridding) (Figure 4-9). Replicate ISM samples are collected with the same sampling method but

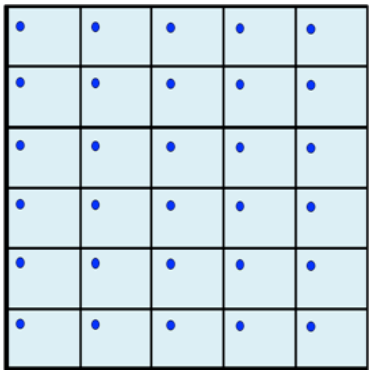


Figure 4-7. Systematic random sampling/ systematic grid sampling with a random start (Serpentine).

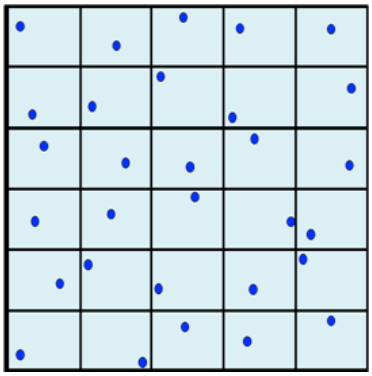


Figure 4-8. Random sampling within grids.

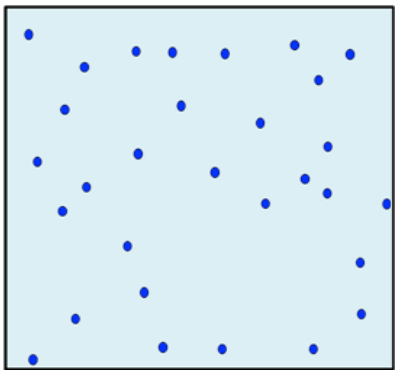


Figure 4-9. Simple random sampling within the entire DU.

not the same exact locations. Each sampling method has its strengths and weaknesses that should be considered when selecting the approach for a given site.

- If the site is relatively homogeneous, all three sampling patterns yield unbiased parameter estimates, but the magnitude of error in the mean may be higher with simple random sampling as compared with systematic random sampling. All three sampling patterns yield equivalent coverages.

- While all three sampling options are statistically defensible, collecting increments within the DU using simple random sampling is most likely to generate an unbiased estimate of the mean and variance according to statistical theory. From a practical standpoint, true random sampling is probably the most difficult to implement in the field and may leave large parts of the DU "uncovered," meaning without any increment sample locations. It should be noted that "random" does not mean wherever the sampling team feels like taking a sample: a formal approach to determining the random sample locations must be used.

- Systematic random sampling can avoid the appearance that areas are not adequately represented in the ISM samples. This approach is relatively straightforward to implement in the field. Theoretically, it is inferior to simple random sampling for obtaining unbiased estimates of the mean, especially if the contamination is distributed systematically so that areas of high or low concentrations are oversampled with the systematic design. Random sampling within a grid is in a sense a compromise approach, with elements of both simple random and systematic sampling.

4.3.4.3 Partitioning the DU

Partitioning the DU into multiple SUs is one way to characterize variability on a smaller spatial scale. This can be useful for both exposure assessment (e.g., assessing risks to multiple receptors with different sized exposure units) and remedial design (e.g., delineation of remediation units smaller than a DU).

When taken over the entire DU, replicates offer information on variability in the estimate of the mean provided by the ISM samples. They do not, however, provide any information on spatial variability of concentrations within the DU. Another approach is to divide the DU into multiple SUs and take one or more ISM samples from each. With this approach, ISM samples are not true replicates in that they are providing estimates of the mean for different subunits within the DU. Individually, they estimate the mean of a subarea, and collectively, they can be used to estimate the mean of the entire DU. Sampling designs with this method yield unbiased estimates of the mean.

- The principal advantage of subdividing the DU is that some information on heterogeneity in contaminant concentrations across the DU is obtained. If the DU fails the decision criterion (e.g., has a mean or 95% UCL concentration above a soil action limit), information will be available to indicate whether the problem exists across the DU or is confined. This information can guide redesignation of the DU and resampling to further delineate areas of elevated concentrations.

- If only one ISM sample is collected per SU, then it is important to understand that each result independently provides an estimate of the mean concentration within the respective SU. Just as a single ISM collected throughout the DU may over- or underestimate the mean by some magnitude (see Section 4.2.1), the information on heterogeneity at the scale of the SU is also subject to uncertainty. If greater certainty is needed at the scale of the SU, then additional increments and/or replicates should be collected at the scale of the SU.

- Collectively, the results from each SU can be used to estimate the mean and 95% UCL at the scale of the DU.

- Error estimates from partitioning a DU into SUs are larger than those from replicate data if the site is not homogeneous. Hence, 95% UCL estimates from a subdivided DU are as high as or higher than those obtained from replicate measurements collected across the DU (using the same number of total increments). The higher 95% UCLs improve coverage (generally attain 95% UCL) and increase the RPD_A. These increases occur if unknown spatial contaminant patterns are correlated with the partitions.

- It must be clearly understood by all that if the 95% UCL for the DU is below the action level, the entire DU passes, even if the ISM result for one or more of the partitioned areas is above the action level. Even with partitioning, the DU remains the unit over which a decision is made.

Note: “Row-column” is an additional sampling pattern proposed by Patil and Tallie (2001). This sampling pattern has not been widely discussed in the context of ISM and consequently was not explored in the simulation studies. However, this approach is discussed in the composite

sampling literature and has the potential advantage of providing spatial information on localized areas of high concentration (see "oversized DUs" in Section 4.4.4).

4.3.4.4 Relative standard deviation

The RSD was calculated from the set of simulated ISM results for each iteration of the Monte Carlo simulation. Collectively, patterns in 95% UCL coverage and other performance metrics can be evaluated for different ranges of RSD. The following were noted:

- Data sets with a high RSD are more likely to achieve specified coverage for 95% UCL of the (population) mean than data sets with low RSD. This effect is explained by the greater variability among replicates leading to higher 95% UCL values, resulting in better coverage.

- A low RSD may intuitively appear to ensure specified coverage by the 95% UCL or low bias in a single estimate of the mean. However, the opposite is in fact the case when the underlying distribution is positively skewed (e.g., lognormal, gamma). For situations in which the UCL or one replicate mean is less than the true mean, the underestimate increases as RSD decreases. This phenomenon reflects the "proportionality effect," whereby the mean and variance are expected to be positively correlated for positively skewed distributions (Goovaerts 1997). Therefore, when the mean is relatively low, so too is the SD. Taken together, there is a greater likelihood that the UCL exhibits insufficient coverage.

4.4 Areas for Further Study

Other potential uses of ISM samples not directly included in the simulation studies are included below. Some considerations are discussed, but additional simulations studies in the future might add further clarification and recommendations for these situations.

ISM is a recent addition to environmental sampling strategies, and there is still much to be learned about it. Areas with recognized information gaps include the following:

- combining information from multiple DUs after sampling
- extrapolating the information from one DU to another
- sampling very large DUs
- comparison of results from multiple sites

4.4.1 Combining DUs

On occasion, there might be a desire to combine information from multiple DUs into a single, larger area. There are two primary explanations for when this might occur:

- A site has areas with different conceptual models in terms of expected contamination, as could happen when there is, for example, a stream channel, a meadow, and a rocky out-cropping in an area that we would like to define as an exposure unit. Each of those areas might be investigated as a separate DU for site characterization but then combined to define a single exposure unit.

- For ecological and human health risk assessment, we might need to consider a variety of sizes of DUs to accommodate multiple receptor scenarios. For example, if the area of a

pocket mouse habitat is a quarter that of a muskrat, which is an eighth of that of an eagle, then we might need to sample in DUs of a size defined for pocket mice, but then combine DUs for the receptors with larger home ranges.

When these considerations are incorporated in the initial planning stages, they can be addressed by using a stratified sampling design. Within each strata, it may be appropriate to use ISM, but then one encounters the challenge of combining the ISM data from the strata into the larger DU. Conversely, this issue may also arise when ISM data are collected from multiple DUs and combined to estimate the mean in a single, larger DU. Whether preplanned or not, the same treatment of the data is appropriate.

When there are multiple samples in each stratum, the overall mean of the larger DU can be estimated using the following formulae. Let n_i represent the number of samples from region i, $x\text{-}bar_i$ represent the mean of the ISM samples from region i, s_i represent the SD of the replicate ISM samples from region i, and w_i represent the weight, i.e., the relative size associated with region i. Note that if all strata are of the same size, the w_i are equal, and these equations simplify to the more common calculation methods for the mean and standard deviation. The relative size is the percentage of the larger DU that is made up of region i. The weighted mean is thus:

$$\text{Weighted Mean} = \sum_i w_i \bar{x}_i$$

The standard error associated with the weighted mean is:

$$\text{Standard Error} = \sqrt{\sum_i w_i^2 \frac{s_i^2}{n_i}}$$

which has degrees of freedom approximated by the Welch-Satterthwaite approximation (Cochran, 1977):

$$df \approx \frac{\left(\sum_i \frac{w_i^2}{n_i} s_i^2\right)^2}{\sum_i \frac{\left(\frac{w_i^2}{n_i} s_i^2\right)^2}{n_i - 1}}$$

Table 4-5 provides a numerical example of this calculation, where data from two DUs are combined to derive a 95% UCL for a larger DU. In this example, an elementary school is divided into two DUs representing different play areas: DU1 is the kindergarten playground, and DU2 is the playground for older children. A maintenance worker has contact with both DUs, and a separate DU is constructed to reflect exposure of this worker.

Assume the concentrations of replicate results in DU1 and DU2 are as shown in Table 4-5, based on n = 30 increments per replicate:

Table 4-5. Summary statistics used to combine DUs

Playground area	Area (acres)	Sample statistics			95% UCL	
		Replicates	Mean	SD[a]	Student's-*t*	Chebychev
DU 1 (kindergarten)	0.25	25, 100, 140	88.3	58.4	187	235
DU 2 (older child)	0.50	5, 25, 305	111.7	167.7	394	534
Equal weight	0.75	25, 100, 140, 5, 25, 305	100	113	193	301

[a] SD = standard deviation.

The 95% UCLs for each DU are given for both the Student's-*t* and Chebyshev methods. Section 4.3.4 provides a discussion of different performance metrics for the UCL that can be used to determine which UCL method may be more likely to achieve the study objectives. Because the true mean for each DU is unknown, the RPD between the UCL and mean cannot be calculated. Figure 4-5 provides examples of the 5th, 50th, and 95th percentile RPDs of UCLs calculated with r = 3 replicates for lognormal distributions with CVs of 1 and 4 when the UCL exceeds the true mean. Recall that the CV in this context refers to the dispersion of the underlying distribution (e.g., distributions given by individual increments), not the distribution of means given by the ISM results. The mean of the ISM replicates can be assumed to approximate the mean of the underlying distribution, and the SD of the replicates can be assumed to approximate the standard error of the mean of the underlying distribution: $SE = SD/\sqrt{n}$. We can rearrange to solve for SD: $SD = SE \times \sqrt{n}$. So for n = 30, we can estimate SD of the underlying distribution by multiplying the SD of the ISM results by $\sqrt{30} = 5.5$. Therefore, the following are estimates of the SD and corresponding CV of the underlying distributions for each DU and the combination of DUs:

- CV of DU1 = SD/mean = (58.4 × 5.5)/88.3 = 3.6
- CV of DU2 = SD/mean = (167.7 × 5.5)/111.7 = 8.3
- CV of DU1 + DU2 (equally weighted) = SD/mean = (113 × 5.5)/100 = 6.2

For r = 3 replicates and CV = 4, Figure 4-5 suggests that the median RPD for both UCL methods is 90% and the 95th percentile is about 200% for Chebyshev and 150% for Student's-*t*. The magnitude of the RPDs is expected to be even more pronounced for CV = 8.

As summarized in Table 4-4, the coverage of the UCLs also depends on the CV of the underlying distribution. Both DUs appear to have high CVs (i.e., >3), and the Student's-*t* UCL is not expected to yield a coverage close to 95%, even if the number of replicates were increased. Therefore, Chebyshev UCL is expected to yield more reliable results (based on coverage).

If it is assumed that, on average, a maintenance worker spends equal time in DU1 and DU2, then the replicates from each DU can be weighted equally, yielding the results shown in the third row of Table 4-5. Alternatively, it may be assumed that a maintenance worker's exposure is proportional to the respective areas of each DU and the equations from Section 4.4.1 can be used to generate summary statistics for the combined area (0.75 acres). The weighting factors applied to each DU should sum to 1.0, which is achieved by dividing each area by the sum of the two areas:

- $w_1 = 0.25/0.75 = 0.33$
- $w_2 = 0.50/0.75 = 0.66$

Weighted mean = $\sum_i w_i \bar{x}_i = (0.33 \times 88.3) + (0.66 \times 111.7) = 103.9$

Standard error for mean (SE) = $\sqrt{\sum_i w_i^2 \frac{s_i^2}{n_i}} = \sqrt{(0.33)^2 \times \frac{(58.4)^2}{3} + (0.66)^2 \times \frac{(167.7)^2}{3}} = 65.5$

Degrees of freedom (df) = $\frac{\left(\sum_i w_i^2 \frac{s_i^2}{n_i}\right)^2}{\sum_i \frac{\left(w_i^2 \frac{s_i^2}{n_i}\right)^2}{n_i - 1}} = \frac{\left((0.33)^2 \times \frac{(58.4)^2}{3} + (0.66)^2 \times \frac{(167.7)^2}{3}\right)^2}{\frac{\left((0.33)^2 \times \frac{(58.4)^2}{3}\right)^2}{3-1} + \frac{\left((0.66)^2 \times \frac{(167.7)^2}{3}\right)^2}{3-1}} = 2.1$

Student's-*t* 95% UCL = $\bar{X} + t_{(1-\alpha)(df)} \times SE = 103.9 + 2.92 \times 65.5 = 295$

Chebyshev 95% UCL = $\bar{X} + \left(\sqrt{\frac{1}{\alpha} - 1}\right) \times SE = 103.9 + 4.36 \times 65.5 = 390$

The online version of this document contains a working calculator for the Weighted 95% UCL for a Combined DU from Several Smaller DUs:
http://www.itrcweb.org/ISM-1/4_4_1_Combining_DUs.html

This same methodology could be used to combine a surface DU with its corresponding subsurface DU. The only slight difference would be that the weight term, w_i, would reflect the proportion of the total soil volume within the DU.

There are other considerations for combining DUs that may benefit from further study:

- a single ISM result in one of the DUs so that a SD cannot be calculated for that region
- the impact of very different numbers of increments in the DUs
- the impact of very different numbers of replicates in the DUs

4.4.2 Extrapolating from DUs

As discussed in Section 4.2, the motivation for collecting replicate ISM samples within a DU is to obtain an estimate of the variance in the mean, from which a UCL can be calculated. When a site includes many DUs, it may be tempting to extrapolate the estimate of the variance (or the CV) from one DU to another. However, we must first consider the extent to which the distributions may be comparable across DUs. Two related questions about the distribution should be considered:

- *Identically distributed*: Does our knowledge of potential sources suggest that similar contaminant distributions can be expected at the spatial scales represented by each DU? In effect, we would like to be able to assume that the distributions are approximately the same.

- *Normally distributed*: Estimates of the means and SDs will vary by random chance across DUs even if the distributions are the same and the same number of increments are used. Is it preferable to extrapolate estimates of the standard SD or CV?

Both questions require that we understand factors that might influence the relationship between the mean and SD of ISM replicate results within a given DU. Statistical theory suggests that we can expect the estimated mean and SD to be independent for normal distributions but positively correlated for positively skewed distributions (Goovaerts 1997). If the ISM mean and variance estimates are independent, this notion presents a challenge because we would have no reason to assume that the ratio of the SD to the mean (as represented by the CV) is the same. DUs with relatively high estimated means may have low SDs and vice versa. Instead of extrapolating the average CV across DUs, we would introduce less uncertainty by extrapolating the average SD. By contrast, if the parameters are correlated because of some asymmetry in the distribution of mean concentrations (despite the CLT, as described in Section 4.2), then it would be preferable to extrapolate the average CV. A priori knowledge about the distribution shape is unlikely, and this source of uncertainty cannot be fully addressed through simulation studies. Therefore, one must be very cautious in how information is extrapolated between DUs and how an extrapolation may ultimately introduce decision errors.

4.4.3 Comparing DUs

Simulation studies and case studies should be conducted to elucidate the advantages and disadvantages and the practical constraints for comparing DUs where some or all have ISM samples. Without such studies, recommendations on implementation of comparisons are not possible, but there are some general considerations that are clear without the aid of simulations.

4.4.3.1 Site-to-site comparisons

ISM data from one site can be compared to that from another. For example, sampling for two DUs may consist of 30 increments and 5 replicates each. Standard two-sample hypothesis tests (e.g., Wilcoxon Rank Sum or Gehan) can be applied to determine whether the differences are statistically significant under the assumption that the variances are the same. While similar in concept to distribution testing with discrete sampling data, some aspects of ISM data comparisons are unique. First, while it is not necessary for the DUs to have the same number of ISM replicates, it is likely that the number of replicates will be quite small (e.g., 3–5). Therefore, the decision errors may be higher with hypothesis testing using ISM data compared to discrete data. In addition, since the estimated SD of ISM replicate results is a function of the number of increments obtained from the DU, samples can be compared directly only if the same number of increments is collected. To conduct a hypothesis test for ISM data based on unequal increments, a statistical adjustment to the estimated SD may be appropriate to reduce the chance of violating the hypothesis test assumption of equal variance (see Appendix A).

4.4.3.2 Incremental to discrete sample comparisons

Occasionally, it may be desirable to consider comparing or combining discrete data and ISM data. Conceptually, this can only be done when specific conditions are met:

- The design for selecting the discrete samples is known (i.e., simple random sampling, adaptive cluster sampling, etc.), and the discrete sample set is representative of the entire DU (i.e., the sampling design was statistically based and not biased).
- The samples have been collected using the same collection method or methods similar enough to ensure equivalent particle size distributions between types of samples.
- The samples are representative of the same soil conditions (e.g., soil type, depth).
- The samples have been processed in a laboratory using the same sample preparation method or methods similar enough to ensure equivalent digestion and extraction of contaminants from the sample matrix for analysis.
- The samples have been analyzed in a laboratory using the same analytical method or methods similar enough to ensure equivalent analytic results.
- The quality of both data sets is understood (via data validation reports) such that it is known that the data are appropriate for the intended use.

One must be very cautious in how information is compared or combined between DUs since it is likely that one or more of these conditions will be violated to some degree, and in practice, there are no established methods for combining discrete and ISM data.

4.4.3.3 Site-to-background comparisons

For background screening approaches, summary statistics from discrete sample results (representing individual site measurements) are not directly comparable to summary statistics from ISM sample results, which represent mean estimates.

A common element of most site investigations is the comparison of the contaminant distribution in volumes of soil collected from a DU to the distribution in soil collected from a suitable background or reference area. In some cases, regional background values that represent upper-bound estimates may have been derived or endorsed by regulatory entities. Since these types of background values are derived from discrete samples, their information content is sufficiently different from that of an ISM sample to preclude a direct comparison of summary statistics. For example, a set of discrete sample results provides a measure of the distribution of concentrations in relatively small volumes of soil throughout the DU, whereas a set of ISM samples provides measure of the distribution of mean concentrations, each of which is an estimate of the population mean for the entire DU. Therefore, the SDs estimated from the samples represent very different properties of the contaminant distribution. Regional background levels are typically based on an upper-bound statistic, such as an upper percentile or an upper tolerance limit (UTL, i.e., a UCL for a percentile). The objective is to establish a threshold for point-by-point comparisons to each individual (discrete) site result. If no site result exceeds the threshold, one can be reasonably confident that the distribution is not elevated with respect to background. Similar to the discussion of comparisons with numerical *action level*s, it may not be possible to satisfy decision objectives with ISM when a numerical threshold is intended for comparison to discrete observations (i.e., maximum concentrations in small volumes) rather than estimates of

average concentrations. Discrete and ISM data sets have different characteristics, and statistical procedures for comparing DU ISM data with discrete background data, and vice versa, have not been well established.

An alternative background screening approach is to use hypothesis testing to compare the distributions, rather than screening against an upper-bound statistic. This alternative is often used because it is well established that there is a high likelihood with point-by-point screening that one or more site exceedances will be observed by random chance even if the distributions are exactly the same. Furthermore, the error rate increases with increasing numbers of samples for the site. The hypothesis testing approach allows for localized exceedances so long as the difference in the means (or upper tails) is not statistically significant.

For this document, comprehensive simulation studies were not conducted to evaluate the statistical performance of background comparison tests for ISM results (i.e., small number of samples, moderate asymmetry). Since tests are robust to moderate violations of assumptions of normality and equal variance, the fact that formal distribution testing cannot be conducted (see Section 4.1.1) is not expected to be a major limitation for background screening with ISM data. Instead, the two key challenges for ISM are achieving the desired statistical power of the tests (i.e., likelihood of detecting differences in the populations that exist) due to small number of samples and the inability to evaluate upper tails of the underlying distributions. Section 7.2.4 provides a detailed discussion of the assumptions associated with different hypothesis tests, highlighting why results of statistical tests can be misleading when the background and site data sets have fewer than five observations each. In addition, decision errors may be affected if the samples are collected with different sampling designs, including different number of increments/replicates, different sample masses, DU volume, sampling protocols, depth intervals, and sampling patterns. Therefore, the results of hypothesis tests applied to ISM data sets should be interpreted with caution until these limitations can be more thoroughly studied. If formal statistical tests are not used, simple graphical analysis (e.g., dot plots grouping ISM results by study area) may be informative as a semi-quantitative method for comparing background and site distributions.

> For background comparisons, graphical evaluations are preferred over formal statistical tests (e.g., hypothesis tests) because the performance of hypothesis tests has not been evaluated for small sample sizes (number of replicates) expected with most ISM sampling designs.

Comparison of site ISM data to background discrete data using either hypothesis testing or UTLs is not recommended because the variance is represented differently in ISM and discrete sampling. Comparison of an ISM estimate of the mean to a discrete sample collected from soil representing background is likely to lead to decision errors in which one incorrectly concludes that the contaminant distribution on site is consistent with background conditions.

4.4.4 Oversized DUs

Generally, DUs should be no larger than the exposure units used for risk assessment if risk assessment is likely to be needed for the site. However, this limit may be impractical under some circumstances. Examples might include an acute exposure scenario (e.g., single soil ingestion event for a small child) or ecological risk assessment for a species with a very small home range. In these situations, DUs are by necessity oversized, and the average concentration for the DU

provided by ISM offers only a crude approximation at best of concentrations that might exist for individual exposure units within the DU. Extrapolations of estimates of dispersion (e.g., SD or CV) across DUs to calculate a 95% UCL or other upper-bound statistic should be performed with caution, as discussed above (see Section 4.4.2).

Another approach is to use estimates of possible upper-end concentrations within a DU to evaluate potential "worst-case" situations, but the information to derive these estimates is limited due to the nature of ISM. This is not a new issue, and an analogous problem exists for composite samples. The literature for composite sampling contains a number of approaches for estimating high-end concentrations within the sampled area. The simplest of these is to multiply the mean value from the composite (or ISM sample) by the number of increments. This method represents the situation in which all of the contaminant is present in one of the increments. Given the number of increments in a standard ISM design, this approach is extraordinarily conservative and can yield quite high values. Other approaches that are less conservative include multiplying the average concentration by the square root of the number of increments or more complicated formulas (Barnett and Bown 2002). It would be advantageous to explore approaches to "decomposite" data in the context of ISM for situations in which the upper end of the concentration range within a DU is an important component of meeting site DQOs.

Note: A computationally equivalent approach is to use the average concentration but divide the soil criterion by the number of increments.

4.4.5 Explicitly Address Additional Sources of Error

Other sources of error, such as blending error and segregation and grouping error could be addressed through more complex simulation studies in the future. The impact of grinding on both the measurable concentration of metals in soil and on bioavailability also merits further study.