EPA SUPERFUND POLICY ON EVALUATING ATTAINMENT OF CLEANUP STANDARDS IN SOIL
Methods for
the Mainte-
ance of
Cleanup Standards
Volume 1: Soil
1. INTRODUCTION

Congress revised the Superfund legislation in the Superfund Amendments and Reauthorization Act of 1986 (SARA). Among other provisions of SARA, section 121, Cleanup Standards, discusses criteria for selecting applicable and relevant or appropriate requirements (ARARs) and includes specific language that requires EPA cleanups to attain ARARs.

Neither SARA nor EPA regulations or guidances specify how to determine attainment or verify that the cleanup standards have been met. This document offers procedures that can be used to determine whether, after a remediation action, a site has attained an appropriate cleanup standard.

1.1 General Scope and Features of the Guidance Document

1.1.1 Purpose

This document describes methods for testing whether soil chemical concentrations at a site are statistically below a cleanup standard or ARAR. If it can be reasonably concluded that the remaining soil or treated soil at a site has concentrations that are statistically less than relevant cleanup standards then the site can be judged protective of human health and the environment. Figure 1.1 shows the steps involved in this evaluation which requires specification of attainment objectives, sampling protocols, and analysis methods.

For example, consider the situation where several samples were taken. The results indicate that one or two of the samples exceed the standard: How should this information be used to decide whether the standard has been attained? Some possible considerations include: the mean of those samples could be compared with the standard; the magnitude of the two sample values that are larger than the standard might be useful in making a decision; or the area where the two large sample values were obtained might provide some insight. The following factors are important in reaching the decision as to whether a cleanup standard has been attained:
CHAPTER 1: INTRODUCTION

• The spatial extent of the sampling and the size of the sample area;
• The number of samples taken;
• The strategy of taking samples; and
• The way the data are analyzed.

Simply to require that a Superfund site be cleaned until the soil concentration of a chemical is below 50 mg/kg is incomplete. Statements suggesting that the site will be remediated until the soil concentration of a chemical is 50 mg/kg reveal little in terms of the environmental results anticipated, the future exposure expected, the resultant risk to the local population, or the likelihood that substantial contamination will remain after a decision is made that the site has been fully remediated. A specific sampling and data analysis protocol must accompany the risk-based standard for the standard to be meaningful in terms of benefit or actual risk.

This document does not attempt to suggest which standards apply or when they apply (i.e., the "How clean is clean?" issue). Other Superfund guidance documents (e.g., USEPA, 1986c and USEPA, 1986d) perform that function.

1.1.2 Intended Audience and Use

Management/supervisory personnel will find the executive summary and introductory chapters useful. However, this manual is intended primarily for Agency personnel (primarily onsite coordinators and regional project managers), responsible parties, and their contractors who are involved with monitoring the progress of soils remediation at Superfund sites. Although selected introductory statistical concepts are reviewed, the document is directed toward readers that have had prior training or experience applying quantitative methods.

This document discusses data analysis and statistical methods for evaluating the effectiveness of Superfund remedial actions. However, there are many other technical aspects to this problem. Input from soil scientists, engineers, geologists, hydrologists, geochemists, and analytical chemists is essential. There must be dialogue among this group, including the statistician, so that each member understands and considers the point
Figure 2.1 A Statistical Perspective of the Sequence of Ground Water Monitoring Requirements Under RCRA

Notice that until contamination above a risk standard is documented (D) the null hypothesis is that the facility is clean. Once the facility has been proven to be in exceedance of a health criteria then the null hypothesis is that the facility is contaminated until proven otherwise (G).
CHAPTER 2: INTRODUCTION TO STATISTICAL CONCEPTS AND DECISIONS

When specifying simplified Superfund site cleanup objectives in consent decrees, records of decision, or work plans, it is extremely important to say that the site shall be cleaned up until the sampling program indicates with reasonable confidence that the concentrations of the contaminants at the entire site are statistically less than the cleanup standard. This prescription will result in the site being designated clean only after a situation similar to G is observed. However, attainment is often wrongly described by saying that concentrations at the site shall not exceed the cleanup standard. This second prescription can result in a situation similar to C being designated as clean.

As discussed in the introduction to this chapter, variation in sampling and lab analysis introduces uncertainty into the decision concerning the attainment of a cleanup standard. As a result of the uncertainty and the null/alternative hypothesis arrangement discussed above, the site can be determined clean when, in fact, it is not, resulting in a false positive decision (or Type I error). The converse of a false positive decision is a false negative decision (or Type II error), the mistake of saying the site needs additional cleanup when, in fact, it meets the standard. The Greek letter alpha (α) is used to represent the probability of a false positive decision and beta (β) is used to represent the probability of false negative decision. The definitions above are summarized in Table 2.1.

Table 2.1 A diagrammatic explanation of false positive and false negative conclusions

<table>
<thead>
<tr>
<th>Decision based on the sample data is:</th>
<th>The true condition is:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Clean</td>
</tr>
<tr>
<td>Dirty</td>
<td>Dirty</td>
</tr>
<tr>
<td>Correct Power (1 - β)</td>
<td>False positive (Probability is α)</td>
</tr>
<tr>
<td>False negative (Probability is β)</td>
<td>Correct Certainty (1 - α)</td>
</tr>
</tbody>
</table>

2-5
CHAPTER 2: INTRODUCTION TO STATISTICAL CONCEPTS AND DECISIONS

- The false positive rate; and
- The sample size.

In summary, there are two important uses of power curves. The first is to further facilitate understanding of the concept that, although the site may actually be clean, a set of samples from the site can be obtained that suggest the site is dirty. The cleaner the site, the less chance of this happening. Conversely, a site may be dirty, but the particular set of samples suggest the site is clean. Again the dirtier the site, the less chance of this occurring. The chances of these errors are controlled by the position and shape of the power curve relative to the cleanup standard. Figures A.1 - A.4 illustrate several families of power curves. The ideal shape of a power curve is a step function that has a 1.0 probability of declaring the site clean whenever the true concentration is less than the cleanup standard and a zero probability of declaring the site clean when the concentration is greater than the cleanup standard.

The second use of a power curve is to help decide on an appropriate sample size for a sampling program. The lower the variability and the more samples taken, the closer the power curve will come to approaching the ideal step function described above. In addition, the trade-off between the false positive and negative rate influences the position of the power curve. Use the power curves in Appendix A to assist with the sample size determination process in one of two ways:

- Select the power curve desired for the statistical test and determine from this the sample size that is required; or
- Select the sample size to be collected and determine what the resulting power curve will be for the statistical procedure.

Chapters 6, 7, and 8 provide specific methods for making sample size determinations.

2.3 Attainment or Compliance Criteria

The characteristic of the chemical concentrations to be compared to the cleanup standard must be specified in order to define a statistical test to determine whether a sample area attains the cleanup standard. Such characteristics might be the mean.
concentration, the median, or the 95th percentile of the concentrations. In other words, it must be decided whether the cleanup standard is intended to be applied as a mean value such that the mean of the site must be below the cleanup standard or whether the cleanup standard is a high percentile value that must rarely be exceeded at only 5 or 10 percent of the site. Figure 2.4 illustrates these characteristics on three distributions. Section 3.5 offers a more detailed discussion of these parameters.

2.3.1 Mean

The location or general magnitude of a set of data is often characterized by the mean of the distribution. The mean of the concentration distribution is the value that corresponds to the "center" of the distribution in the sense of the "center of gravity." In determining the mean from a highly skewed lognormal distribution, small amounts of soil with concentrations far above the mean are balanced by large amounts of soil with concentrations close to, but below, the mean.

Whether the mean is a useful summary of the distribution depends on the characteristics of the sample area and the objectives of the cleanup. In a sample area with uniform contamination and very little spread or range in the concentration measurements, the mean will work well. If the spread in the data is large relative to the mean, the average conditions will not adequately reflect the most heavily contaminated parts of the population. If interest is in the average exposure or the chronic risk, the mean may be an appropriate parameter.

When using the mean, consideration should be given to the number of measurements that are likely to be recorded as below the detection limit. With many observations below the detection limit, the simple estimate of the population mean cannot be calculated (see the discussion in section 2.5.2).

2.3.2 Proportions or Percentiles

High percentiles or proportions pertain to the tail of a distribution and control against having large concentration values. The 50th percentile, the median, is often a useful alternative to the mean.
Throughout this document, the cleanup standard will be denoted by Cs.

3.5 Selection of the Statistical Parameter to Compare with the Cleanup Standard

3.5.1 Selection Criteria for the Mean, Median, and Upper Percentile

Criteria for selecting the parameter to use in the statistical assessment decision are:

- The criteria used to develop the risk-based standards, if known;
- The toxicological effect of the contaminant being measured (e.g., carcinogenic, systemic toxicant, developmental toxicant);
- The relative sample sizes required or the relative ease of calculation;
- The likelihood of concentration measurements below the cleanup standard; and
- The relative spread of the data.

Table 3.1 presents these criteria and when they support or contradict the use of the mean, upper percentile, and median. The median may offer a reasonable compromise because the median is the 50th percentile and a measure of central tendency. Table 3.2 illustrates the broad potential utility of the median.
### Table 3.1 Points to consider when trying to choose among the mean, high percentile, or median

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Points to Consider</th>
</tr>
</thead>
</table>
| **Mean**             | 1) Easy to calculate and estimate a confidence interval.  
2) Requires fewer samples than other parameters to achieve similar confidence.  
3) Useful when the cleanup standard has been based on consideration of carcinogenic or chronic health effects or long-term average exposure.  
4) Useful when the soil is uniform with little spread in the sample data.  
5) Not as useful when contamination exists in small areas within a larger area that is being sampled because the mean can be "diluted" or reduced by the inclusion of clean areas in the sample area.  
6) Not very representative of highly variable soils because the most heavily contaminated areas are not characterized by a mean.  
7) Not useful when there are a large proportion of less-than-detection-limit values. |
| **Upper Proportion/Percentile** | 1) Can be expressed in terms that have more meaning than tests of the mean. Volumes or areas can be expressed relative to the total volume or area of concern, and this can be a proportion of importance. For example, if no more than 10,000 m³ in a total volume of 1,000,000 m³ can exceed a cleanup standard, then this becomes a test to verify with reasonable confidence that no less than 99 percent of the site is below the cleanup standard.  
2) Will provide the best control of extreme values when data are highly variable.  
3) Some methods are unaffected by less-than-detection-limit values, as long as the detection limit is less than the cleanup standard.  
4) If the health effects of the contaminant are acute or worst-case effects, extreme concentrations are of concern and are best evaluated by ensuring that a large proportion of the site is below a cleanup standard. |
Table 3.1 Points to consider when trying to choose among the mean, high percentile, or median (continued)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Points to Consider</th>
</tr>
</thead>
</table>
| Upper Proportion/Percentile (continued) | 5) Similar to the mean, if contamination exists within a small area, but if the sampling program is conducted to include a much larger surrounding area with little contamination, the proportion will be affected or “diluted.”  

6) The proportion of the site that must be below the cleanup standard must be chosen.  

7) When statistical methods are used that require few assumptions, a larger sample size will be required than for tests based on the mean. |
| Median                  | 1) Has benefits over the mean because it is not as heavily influenced by outliers and highly variable data, and can be used with a large number of less-than-detection-limit values.  

2) Has many of the positive features of the mean, in particular its usefulness for evaluating cleanup standards based on carcinogenic or chronic health effects and long-term average exposure.  

3) Has positive features of the proportion, including its reliance on fewer assumptions.  

4) Retains some negative features of the mean in that the median will not control extreme values. |
CHAPSTER 3: SPECIFICATION OF ATTAINMENT OBJECTIVES

Table 3.2 Recommended parameters to test when comparing the cleanup standard to the average concentration of a chemical with chronic effects

<table>
<thead>
<tr>
<th>Data Variability</th>
<th>Proportion of the data with concentrations below the detection limit:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low (Perhaps &lt; 50%)</td>
</tr>
<tr>
<td></td>
<td>High (Perhaps &gt; 50%)</td>
</tr>
<tr>
<td>Large Coefficient of Variation</td>
<td>Mean (or Median)</td>
</tr>
<tr>
<td>(Perhaps cv &gt; .5)</td>
<td>Upper Percentile</td>
</tr>
<tr>
<td>Small Coefficient of Variation</td>
<td>Mean (or Median)</td>
</tr>
<tr>
<td>(Perhaps cv &lt; .5)</td>
<td>Median</td>
</tr>
</tbody>
</table>

3.5.2 Multiple Attainment Criteria

This guidance document addresses testing for a single parameter—the mean or a specified percentile of the distribution—that is below the cleanup standard. However, in some situations two or more parameters can be chosen. The sample area would be declared clean if all parameters were significantly less than the cleanup standard. For example, there may be interest in providing protection against excessive extreme and average concentrations. Therefore, the mean and an upper percentile can be tested using the rule that the sample area attains the cleanup standard if both parameters are below the cleanup standard. When testing both parameters, the number of samples collected will be either the number required for the test of the mean or the number required for the test of the percentile (whichever number is larger).

Other more complicated criteria may be used to assess the attainment of the cleanup criteria. Multiple criteria are established in the following examples. In each case it is desirable that:

- Most of the soil has concentrations below the cleanup standard and that the concentrations above the cleanup standard are not too large.