CEMP-CE Manual 1110-1-4014	Department of the Army U.S. Army Corps of Engineers Washington, DC 20314-1000	EM 1110-1-4014 31 Jan 2008
	Environmental Quality ENVIRONMENTAL STATISTICS	
	Distribution Restriction Statement Approved for public release; distribution is unlimited	

DEPARTMENT OF THE ARMY
U.S. Army Corps of Engineers
Washington, D.C. 20314-1000

EM 1110-1-4014

CEMP-CE

Manual
No. 1110-1-4014

31 Jan 2008

Environmental Quality
ENVIRONMENTAL STATISTICS

1. **Purpose.** The primary purpose of this Engineer Manual (EM) is to provide practical guidance for statistical evaluations of environmental chemical data to ultimately improve the quality of decisions.

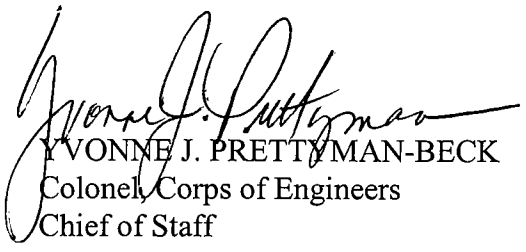
2. **Applicability.** This EM applies to all USACE commands having Civil Works and/or Military Programs hazardous, toxic, or radioactive waste (HTRW) project responsibilities.

3. **References.** References are provided in Appendix A.

4. **Distribution Statement**. Approved for public release, distribution is unlimited.

5. **Discussion.** This manual provides an overview of statistical methods that are applicable to the various life cycles of a typical environmental project. The manual explains basic statistical concepts and their application to environmental projects. The manual may be used as a desk-top reference, as it provides step-by-step instructions for conducting a variety of useful and common statistical tests for environmental data. However, it should be noted that the manual is not intended to replace statistical texts or electronic statistical software. It does not present derivations of statistical formulas or a comprehensive treatment of statistical concepts, but focuses on the application of select statistical methods.

FOR THE COMMANDER:



YVONNE J. PRETTYMAN-BECK
Colonel, Corps of Engineers
Chief of Staff

19 Appendices
App A - References
App B - Statistical Tables
App C - Sampling Strategies
App D - Descriptive Statistics
App E - Assumptions of Distribution
App F - Testing for Normality
App G - Detection Limits and Quantitation Limits
App H - Censored Data

DEPARTMENT OF THE ARMY
U.S. Army Corps of Engineers
Washington, DC 20314-1000

CEMP-CE

EM 1110-1-4014

Manual
No. 1110-1-4014

31 Jan 2008

Environmental Quality
ENVIRONMENTAL STATISTICS

Table of Contents

Subject **Page**

Subject **Page**

LIST OF TABLES

Subject **Page**

LIST OF FIGURES

CHAPTER 1
Introduction

1.1. Purpose. This document is intended to serve as a guide to project team members for the use of statistics in environmental decision-making.

1.2. Applicability. The U.S. Army Corps of Engineers (USACE) developed this document within the broader scope of Technical Project Planning (TPP), recognizing that understanding statistical evaluations can improve project planning and implementation at hazardous, toxic, and radioactive waste (HTRW) sites.

1.3. Distribution Statement. Approved for public release; distribution unlimited.

1.4. References. References are contained in Appendix A.

1.5. Introduction. This Manual's primary objective is to improve a decision-maker's understanding of common environmental statistical evaluations. The applicability of statistical tests and considerations is presented in the context of a typical environmental project life cycle. This document should serve as a first step in explaining statistical concepts and their application at HTRW sites. It is not intended to replace more robust statistical texts or electronic statistical software.

1.5.1. Statistics are applicable to environmental projects throughout their entire life cycle and yield defensible, cost-effective solutions to environmental questions. Statistics can be used to guide the selection of sampling locations, analyze large data sets, and verify that project objectives have been met. Statistics are of particular importance for quantifying the power and limitations of environmental data, specifically because these data are usually limited. It is not possible to collect and analyze every bit of an environmental medium (for example, soil, sediment, groundwater, or surface water) at a site; instead, a set of sample data is used to characterize the environmental medium as a whole.

1.5.2. This Manual is organized into four major Chapters, each associated with a stage in a typical Superfund project life cycle. These Chapters are supported by Appendices that provide detailed statistical or technical explanations of concepts or techniques used within the main sections.

1.5.3. The document is organized as follows:

Chapter 1	Introduction
Chapter 2	Preliminary Assessment (PA)/Site Investigation (SI)
Chapter 3	Remedial Investigation/Feasibility Study (RI/FS)
Chapter 4	Remedial Design (RD)/Remedial Action (RA)

Appendix A References
Appendix B Statistical Tables
Appendix C Sampling Strategies
Appendix D Descriptive Statistics
Appendix E Assumptions of Distribution
Appendix F Testing for Normality
Appendix G Detection Limits and Quantitation Limits
Appendix H Censored Data
Appendix I Identification and Handling of Outliers
Appendix J Graphical Tools
Appendix K Intervals and Limits
Appendix L Hypothesis Testing—Simple Cases
Appendix M Hypothesis Testing—Two-Population and General Cases
Appendix N Hypothesis Testing—Tests of Dispersion
Appendix O Measures of Correlation
Appendix P Comparing Laboratory and Field Data
Appendix Q Trend Analysis
Appendix R Geostatistics
Appendix S Geochemical Trend Analysis
Glossary

1.5.4. Statistical terms unfamiliar to some readers may be used in the four main chapters. When used for the first time, these terms will be printed in italics and footnoted. The footnote will direct the reader to the appropriate Appendix for a detailed explanation of the term. To demonstrate the types of statistical concepts necessary for the planning stages of environmental projects, concepts are presented in the context of *Comprehensive Emergency Response, Compensation, and Liability Act* (CERCLA) projects. The material is applicable to *Resource Conservation and Recovery Act* (RCRA) projects as well. The steps involved in the two programs are similar except for the use of different terminology and the applicable regulations. Table 1-1 presents a terminology crosswalk for the stages of CERCLA and RCRA investigations.

1.5.5. In the following Chapters of this document, major stages that require data gathering and evaluation are presented, and to the extent that statistical processes are applicable, examples are provided from case studies illustrating the application of those statistical processes. Some statistical elements may apply in more than one phase of the project life cycle. The Appendices provide detailed instructions on implementing the statistical processes.

1.5.6. The CERCLA project life cycle is not always linear. As information regarding a given site is gathered, additional questions may be raised about a previously unrecognized threat to human health or the environment. In that case, the process can repeat in whole or in part, creating a series of loops to previous portions of the cycle. In addition, at any point in the process, emergency activities (e.g., “time critical” remedial actions) may occur at earlier or later times in

the cycle. Finally, the process can terminate at the end of any given phase in a "no further action" determination.

Table 1-1.
Project Phase Crosswalk between CERCLA and RCRA

CERCLA Project Phase	RCRA Project Phase
Discovery and Notification	Permit Application
Preliminary Assessment	RCRA Facility Assessment
Site Investigation	Site Inspection
Hazard Ranking	Administrative Order
Remedial Investigation	RCRA Facility Investigation
Feasibility Study	Corrective Measures Study
Proposed Plan	Statement of Basis
Record of Decision	RCRA Permit
Remedial Design	Remedy Design
Remedial Action	Corrective Measures Implementation
Five Year Review	Monitoring/Annual Report
Closeout	Closure

1.5.7. The remedial action process under CERCLA is necessarily iterative and the same statistical tools can be employed repeatedly to address the original problem or newly identified issues at the site. For purposes of this text, however, we will assume a linear progression through an idealized project life cycle, consistent with the instructions contained in EM 200-1-2.

1.5.8. In the *Technical Project Planning Process*, the user is encouraged to identify the appropriate project phase for a given segment of work, then reference matching portions of this Manual for statistical guidance and methods appropriate to that phase.

1.6. Technical Project Planning and the Project Life Cycle. EPA QA/G-4 states, "EPA Order 5360.1 A2 [requires that] all EPA organizations (and organizations with extramural agreements with EPA) follow a systematic planning process to develop acceptance or performance criteria for the collection, evaluation, or use of environmental data." Similarly, ER 5-1-11 states, "Requirements for quality must be addressed during the planning phase of a project's life cycle, rather than waiting until the review or inspection stage." Thus, a systematic planning process of some sort is *required* for all HTRW projects involving the collection of data.

1.6.1. The EPA approach to systematic planning is described in detail in EPA QA/G-4 and is called the Data Quality Objectives (DQO) process. It is a seven-step process, which has as its goal the design of legally and scientifically defensible sampling strategies. The DQO guidance generally assumes that decision-making requires a probabilistic approach. Fundamental to the DQO process is identifying some statistic describing an environmental site that is compared via a statistical process to either a fixed threshold or risk-based value, or a statistical comparison of

some descriptive measure of data for two or more variables. The DQO process also incorporates statistical tools for estimating such things as the number of samples required to measure a site characteristic, spacing of sampling locations, and frequency of sampling. This permits data users to make decisions with specific degrees of statistical confidence.

1.6.2. The USACE TPP process is broader in scope, with the EPA's DQO process as one step within it, to the extent that probabilistic decision-making is appropriate to the goals of the project. The intent of the TPP process is to "get to closure" and to provide documentation of project decisions and project performance. The TPP process is useful for all sites, regardless of whether probabilistic decision-making is involved. It is highly flexible and promotes an approach that balances the size and complexity of a given site or problem with the level of effort involved in the planning process.

1.6.3. As described in EM 200-1-2, there are four phases to the TPP process, as follows.

1.6.3.1. *Identify the Current Project Phase.* The project manager establishes a project team to encompass all of the perspectives and skills required to take the project from beginning to end. The project manager briefs the team on client goals and existing site information and develops a conceptual model for the site. A broad, overall approach to the work is agreed upon, including an assessment of the most likely remedies or outcomes for the site. The work is broken down into clearly defined executable stages and the current stage of work is identified.

1.6.3.2. *Determine Data Needs.* Allowing all perspectives to be addressed, the team identifies the data required for each data user type (e.g., hydrogeologic, chemical, health and safety, risk assessment, engineering, etc.). The team reviews sources of existing information for availability, quality, and applicability to the current stage of work, and identifies data gaps that only new data can fill.

1.6.3.3. *Develop Data Collection Options.* With their respective needs defined, the team members decide on the best approach to obtain the required data. Usually, the team assesses a number of differing approaches and selects the approach that provides all of the requisite data with the best balance of available resources, measurement quality, and client risk tolerance. The TPP process clearly defines three data collection options: basic, optimum, and excessive. A basic sampling approach provides data applicable only to the current stage of work, whereas an optimum approach addresses both current data needs and anticipated future needs as well. An approach not focused on the specific data required to "get to closure" is excessive and should be avoided.

1.6.3.4. *Finalize the Data Collection Program.* At this point, the team encourages clients, regulators, the public, and in some cases other parties, to take part in the decision-making process. Specific DQO statements are prepared for each data user and data type and, to the extent that probabilistic decision-making is appropriate, the EPA's DQO guidance document (EPA

QA/G-4) is used and applied to these statements. From these DQO statements, scopes of work and other project controlling documents (PCDs) such as work plans, quality assurance (QA)/quality control (QC) plans, field sampling plans (FSPs), etc., are derived and cost estimates generated.

1.6.4. Table 1-2 provides a crosswalk between the EPA DQO Process and the USACE TPP process.

Table 1-2.
Crosswalk Between the TPP and DQO Processes

EPA's DQO Process		USACE TPP Process			
		Phase I	Phase II	Phase III	Phase IV
Step 1 **State the Problem**		Identify the Current Project	Determine Data Needs	Develop Data Collection Options	Finalize Data Collection Program
Step 2 **Identify the Decision**					
Step 3 **Identify Inputs to the Decision**				Develop Data Collection Options	
Step 4 **Define the Study Boundaries**					
Step 5 **Develop a Decision Rule**		Identify the Current Project			
Step 6 **Specify Limits on Decision Error**					
Step 7 **Optimize the Design**					Finalize Data Collection Program

1.6.5. Failure to apply, or to apply properly, the TPP process can result in a variety of negative consequences. Failure to properly plan for data collection may require more time and money to implement the work. Lack of planning may extend the time it takes to validate work because both objectives and verification methods may be unclear. Poor planning may create the need for extensive rework or remobilization. Finally, lack of advance planning can cause increases in legal risk to the client and to the USACE by increasing the potential for decision error. On the other hand, too great an emphasis on planning extends the planning cycle and the checking cycle, depleting the available resources.

1.7. Data Quality Objectives, Data Quality Indicators, and Measurement Quality Objectives. This paragraph provides a conceptual understanding of DQOs in the context of project planning for environmental investigations and remediations. The terminology is less important than the underlying concepts that support the decision-making process, as long as all parties possess a common understanding of that process. Project planners derive DQOs from scientific objectives, as well as social and economic objectives and the regulatory objectives of the environmental program under which the project is implemented. DQOs are technical, goal-oriented, qualitative, and quantitative statements derived from the planning process that clarify

study objectives, define the appropriate type of data, and specify tolerable levels of potential decision error. The DQO process typically uses statistics and is the basis for establishing the quality and quantity of data needed to support decisions. The DQO process does not establish specifications for data quality—called measurement quality objectives (MQOs)—or the mechanisms for measuring conformance to those specifications—called data quality indicators (DQIs). MQOs and DQIs are discussed in additional detail below.

1.7.1. *Data Quality*. Data quality depends on the integrity of each element in a series of events. It is critical to collect samples that are representative of the features of the environmental population being investigated in the study area. Representativeness depends on factors such as sample frequency, location, time of collection, and the nature of the sampled medium. Pretesting factors include sample containerization, preservation, transportation, and storage. Sample analysis factors generally include sample homogenization, sub-sampling, sample preparation (such as extraction and cleanup), as well as the instrumental analysis of the sample. The final steps of the process include data generation, reduction, and review.

1.7.1.1. Historically, attention has been focused primarily on the analytical component of data quality rather than on "total measurement system quality." Environmental decision-makers and practitioners tend to assume that data quality is primarily determined by the analytical methodology. For example, as fixed laboratory methods tend to be superior to field methods in terms of analytical uncertainty, data produced from field methods have been viewed to be too uncertain to support critical project decisions. However, defensible decisions are possible only when data quality encompasses total uncertainty rather than the uncertainty associated with only the analytical portion of the investigation. The value of data is limited less by the analytical procedures than by the quality of the *sampling design*[*] and the inherent variability of the environmental population of interest or condition being measured (the "field" component of variability). Because analytical uncertainty is typically small relative to field uncertainty, data quality usually depends more on sampling design than the quality of the individual test methods.

1.7.1.2. Table 1-3 summarizes sources or components of variability for environmental studies and how they are measured and controlled.

1.7.1.3. Regulators have also historically insisted on adhering to pre-approved analytical methods because of a perception that this ensures defensible data and that definitive data will be produced when EPA-approved analytical methods and QA/QC requirements are used. Though adequate data quality is often achieved using EPA-approved analytical methods, they are insufficient to ensure data of high quality. Efforts to improve data quality have primarily focused upon increasing laboratory oversight, rather than on developing mechanisms to manage the largest sources of uncertainty in data, which are issues related to sampling. Furthermore, prescriptive methods are scientifically feasible only when the sample matrices do not vary in any manner that

[*] Appendix C.

will affect the reliability of the analyses. As all analytical methods are potentially subject to chemical and physical interferences, given the variability and complexity of environmental matrices, it is unlikely that "one-size-fits-all" analytical methodologies are viable for all projects.

Table 1-3.
Variability in Environmental Studies

Source of Variability	Measurement Method	Control Methods
Analytical Variability		
Analytical instrumentation	Replicate measurements of instrumental standards (most common for inorganic analysis)	Regular preventive maintenance
Analytical method	Duplicate analytical spikes, lab-blind field duplicate samples	Use of standard methods documented as standard operating procedures; control of standards and reagents; control of instrument conditions
Sample preparation method	Duplicate control samples and matrix spike/matrix spike duplicates	Use of standard methods documented as standard operating procedures; control of standards and reagents; regular, close supervision
Analyst	Analyst demonstration of capability, blank spikes/performance evaluation (PE) samples	Inter-laboratory comparison studies; internal PE and auditing programs; analyst training; regular, close supervision
Field Variability		
Sampling equipment	Field blanks	Routine inspection and preventive maintenance; decontamination; selection of appropriate equipment for representative samples
Sampling method	Method-specific standard deviation of field duplicate results	Selection of appropriate methods for representative samples
Sampler	Inter- and intra sampler standard deviation of field replicate results	Independent auditing program; training; regular, close supervision
Matrix heterogeneity	Field duplicates or replicates, matrix specific standard deviation of field replicates, matrix spike duplicates	Effective field mixing of sample components; compositing
Sample selection	Site-wide or stratum-specific standard deviation of field replicate results	Representative sampling plan; sufficient number of samples; statistically-based sampling design

Note: Duplicates are separate aliquots of the same sample; replicates are a second sample from the same location.

1.7.1.4. The EPA has recently clarified its intended meaning of the term "data quality" in its broadest sense by defining it as "the totality of features and characteristics of data that bear on its ability to meet the stated or implied needs and expectations of the client." One must know how a data set is to be used to establish a relevant benchmark for judging whether the data qual-

ity is adequate. Linking data quality directly to their intended use provides a firm foundation for building a vocabulary that distinguishes the individual components of overall data quality.

1.7.2. *Data Quality Indicators*. DQIs are qualitative and quantitative descriptions of data quality attributes: the various properties of analytical data historically expressed as precision, accuracy, representativeness, comparability, and completeness. Collectively, these factors are called the PARCC parameters. These are discussed in detail in EPA guidance documentation. Because it is evaluated at the same time, an additional parameter often combined with the PARCC parameters is sensitivity, which is the ability of an analytical method or technology to reliably identify a compound in the sample medium.

1.7.2.1. Precision, accuracy, and sensitivity are quantitative properties of data directly measured through an appropriate analytical QC program. Representativeness is primarily a qualitative data quality indicator that is a function of the adequacy of the sampling design (for example, the number of samples and the manner in which samples were collected). Representativeness, in the context of an analytical measurement, can be inferred by examining factors such as duplicates/replicates, blanks, and sample collection procedures. Comparability is a qualitative measure that is critically important when *hypothesis testing*[*] involves comparing different populations, disparate in either space or time.

1.7.2.2. Completeness has been assigned an arbitrary goal of 80 to 100% based on the premise that decisions are still possible if a limited portion of the data are discarded (for example, because of quality control problems). However, the goal is based primarily on practical experience and is not mathematically based. Completeness should be evaluated in the context of project objectives.

1.7.2.3. In addition to these, selectivity is also a data quality indicator. “Selectivity” is the ability of an analytical method to identify the analyte of concern, e.g., the existence of other analytes in a sample or other interferences may mask the presence of the target analyte.

1.7.2.4. There may be more than one DQI for a single data quality attribute. For example, sensitivity is generally thought of in terms of detection, quantitation, or reporting limits, i.e., the lowest value that an analytical method can reliably detect or report. However, another important element of sensitivity is discrimination, the ability to distinguish between values to a given degree of precision. In other words, can the method tell the difference between values of 1 and 2 units, or only differences between 10 and 20 units? When developing DQIs, it is important to define them in terms of all the important attributes and assign specific numeric values to them as often as practicable.

[*] Appendices O and P.

1.7.3. *Measurement Quality Objectives.* MQOs are project-specific values assigned to DQIs derived from project-specific DQOs. MQOs are acceptance criteria for the DQIs and are derived by considering the level of measurement system performance needed to actually achieve project goals. MQOs are not intended to be technology- or method-specific. As with DQOs, MQOs specify *what* the level of data performance should be, but not *how* that level of data performance is to be achieved. A large part of the variability in environmental data stems from sampling considerations. MQOs should balance the relative contributions from analytical uncertainties and from sampling uncertainties. In many environmental media, matrix heterogeneity causes sampling variability to overwhelm analytical variability. Historically, the term MQO was restricted to the analytical side of the measurement process, but the broader concept of DQO (or decision confidence objectives) requires that sampling considerations be included. The importance of including both the sampling and analytical component of MQOs when assessing overall data quality cannot be overemphasized.

1.7.4. *Relationships Among Decision Goals, DQOs, MQOs, and QC Protocols.* During project planning, there should be a logical conceptual progression in the development of decision goals, DQOs, MQOs, and QC acceptance criteria. However, in practice, this will be a non-linear process.

1.7.4.1. As project planning develops, the following should be clearly presented:

1.7.4.1.1. General decision goals.

1.7.4.1.2. Technically expressed project goals (DQOs), and decision rules that will guide project decision-making.

1.7.4.1.3. Tolerable uncertainties for decisions.

1.7.4.1.4. Uncertainties that create decision errors.

1.7.4.1.5. Strategies for managing the uncertainties to achieve the desired tolerances for decision errors.

1.7.4.2. In the beginning of the project, program managers often set broad, non-technical goals. The next step is to translate these broad, non-technical goals into more technically oriented goals that can address specific considerations such as the following.

1.7.4.2.1. Regulations—what are the applicable environmental regulations?

1.7.4.2.2. Confidence in the outcome—how certain do we need to be by the end of the project that we have achieved goals such as risk reduction or regulatory compliance?

1.7.4.2.3. What are the constraints that need to be accommodated?

1.7.4.3. The next level of technical detail for data collection involves identifying DQIs and assigning to them project-specific MQOs that will be needed to achieve the project DQOs. At this point, the project team begins to consider in detail the options available for acquiring the needed measurements and selecting those that best meet the needs of the program. These decisions are documented in sampling and QC plans that specify the controls that will be used to ensure that MQOs are met and that any deviations are appropriately addressed.

1.7.4.4. Because sampling design and analytical strategy interact to influence the statistical confidence in final decisions, interaction among a statistician, a sampling expert, and an analytical chemist is critical for selecting a final strategy that can achieve project goals cost-effectively. The statistician is concerned with managing the overall variability of data, and with interpreting data with respect to the decisions being made. A statistician is a person having adequate familiarity with statistical concepts to correctly apply the required tests; this does not necessarily require a degree in statistics. The field sampling expert is responsible for implementing the sampling design while managing contributions to the sampling variability as actual sample locations are selected and as specimens are collected. The chemist is responsible for managing components of variability that stem from the analytical effort.

1.7.4.5. In summary, the conceptual progression starts with the project-specific decision goals, and then moves from broader, higher-level goals to narrow, more technically detailed articulations of data quality needs. Project decisions are translated into project-specific DQOs; then into project-specific MQOs; then into technology/method selection and development of a method-specific QC protocol that blends QA/QC needs of the technology with the QA/QC needs of the project. Then the process reverses. The data must be assessed against the project MQOs to document that data quality meets the decision-making needs of the project.

1.7.4.6. Figure 1-1 presents the life cycle in project planning. Figure 1-2 illustrates which guidance documents are useful in the planning phases of a project.

1.8. Statistics in Environmental Project Planning. The number of individual samples collected during a given study is called sample size and is generally designated by the statistic *n*. In order for decisions based on that sample to be meaningful in any scientific sense, the sample size has to be sufficiently large to account for the inherent variability in the characteristics measured. Sample size should be dependent on the variability in the measured condition but, in practice, is often limited by available resources.

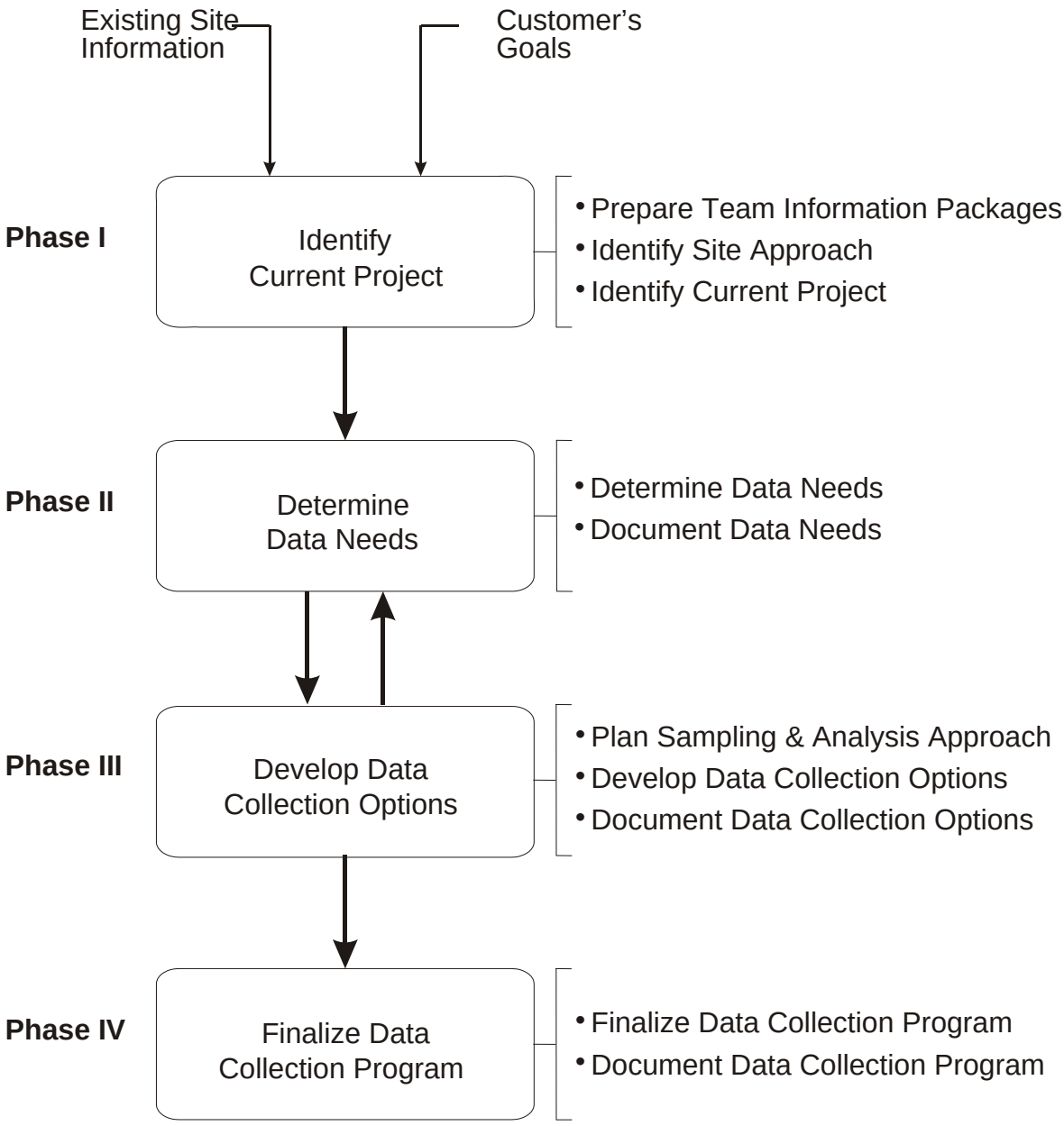


Figure 1-1. Project planning life cycle.

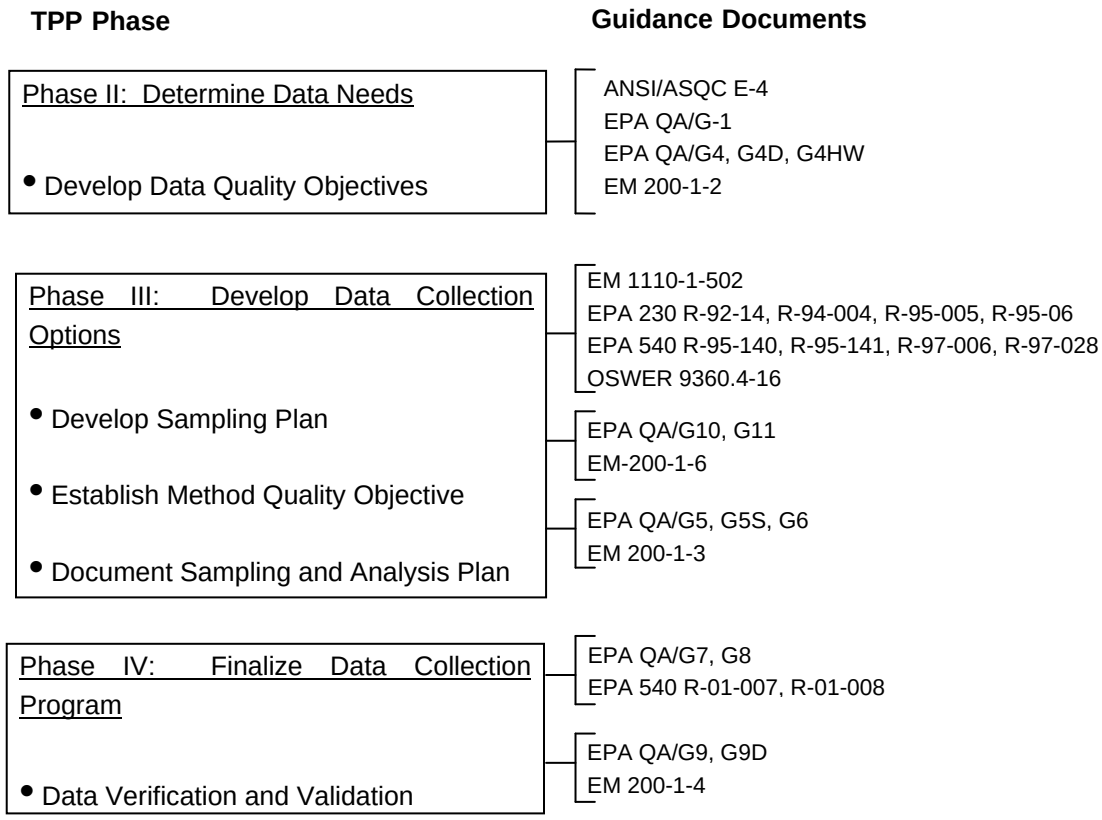


Figure 1-2. Guidance document life cycle.

1.8.1. A hypothetical illustration may be helpful in understanding this relationship. Let us suppose that a researcher wants to know the average concentration of a particular chemical constituent in the air of a sealed room. The constituent of interest is initially absent from the room and the researcher releases the chemical into the room from a port in the north wall of the room. Immediately after opening the port, a measurement taken along the south wall will not detect the presence of the chemical, while a sample taken adjacent to the port will display a high concentration. As the chemical disperses throughout the room via various physical processes, a single sample taken at any location in the room will not provide a representative value for the average concentration in the room as a whole. Even if a single sample were collected some time well after the release of the gas (i.e., after an equilibrium state of dispersion has been achieved), depending upon the physical characteristics of the chemical and the room, it may not be uniformly spread throughout the room. Thus, a sample taken at any single randomly selected location will not give a representative result for the room as a whole, or even necessarily a good approximation.

1.8.2. Only when the chemical is uniformly dispersed throughout the three dimensions of the room, and is held static in that condition, can a representative result be arrived at from a single sample. The analytical error or measurement uncertainty would also need to be negligible when analyzing the one sample. In all other cases, the true *population mean* (μ)[*] *(*the real average concentration for the room as a whole) must be approximated by averaging the results from a number of samples.

1.8.3. The greater the variability in the chemical concentration throughout the room is, the more individual samples will be required to formulate an accurate approximation of the true average. Therefore, as decision *confidence* requirements increase (i.e., as confidence increases toward 1 or 0 decision error tolerance), the number of samples required to correctly estimate any statistical parameter will also increase.

1.8.4. Variability is a measure of the degree of dispersion (or spread) for a set of values. The *sample variance*[†], s^2, and *sample standard deviation*, s, measure the spread of individual measurements or values about the *sample mean*[‡], $\bar{x}$. Some factors that may contribute to variability in environmental populations are the following.

1.8.4.1. Distance, direction, and elevation relative to point, area, or mobile population sources.

1.8.4.2. Non-uniform distribution of pollution in environmental media owing to topography, hydrogeology, meteorology, actions of tides, and biological, chemical, and physical redistribution mechanisms.

1.8.4.3. Diversity in species composition, sex, mobility, and preferred habitats of biota.

1.8.4.4. Variation in natural background levels over time and space.

1.8.4.5. Variable source emissions, flow rates, and dispersion parameters over time.

1.8.4.6. Accumulation or degradation of pollutants over time.

1.8.5. For a particular sampling plan where n measurements are taken for some contaminant of concern in a study area, a (sample) mean concentration ($\bar{x}$) and (sample) standard deviation (s) for the contaminant are calculated. The standard deviation measures the variability of the individual measurements. However, it is often the case that it is the variability of $\bar{x}$ itself that is of interest. The variability of the mean is often measured by the standard deviation of the sample

[*] Appendices C and D.
[†] Appendices D, E, and H.
[‡] Appendices C, D, E, F, G, and H.

mean, $s_{\bar{x}} = s/\sqrt{n}$. Those two sample values, $\bar{x}$ and $s_{\bar{x}}$, are used to estimate the interval (range) within which the true mean (μ) of the chemical concentration probably occurs, under the assumption that the individual concentrations exhibit a normal (bell-shaped) distribution.

1.8.6. The relationship among variability, available resources (expressed as sample number, *n*), and decision confidence or lack of uncertainty is fundamental to the project planning process. In general, cost increases as the desired level of confidence or lack of uncertainty increases. Thus, balancing cost and confidence is a primary objective of the planning process. As illustrated in Figure 1-3, this can be depicted as a balance between cost and level of uncertainty: reducing uncertainty increases project costs. As the number of samples increases, the uncertainty decreases but the cost increases. As depicted in Figure 1-3, project planning is the fulcrum of a seesaw balancing cost and uncertainty.

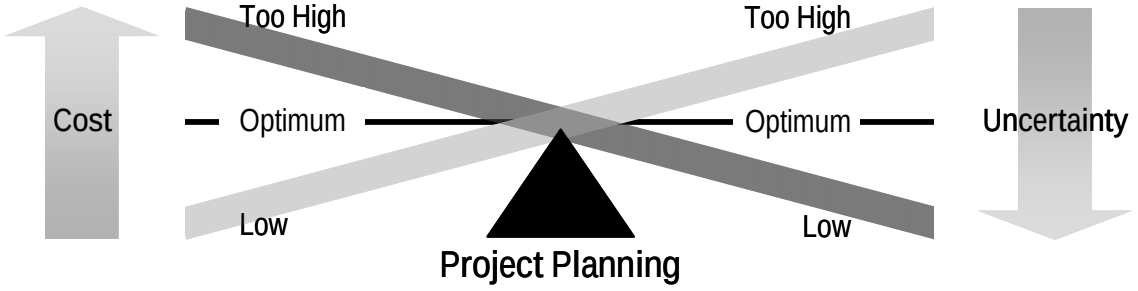


Figure 1-3. Balance between resources and certainty.

1.8.7. When dealing with regulators and clients, it is often beneficial to illustrate, in mathematical terms, the relationship among the project objectives, the desired confidence for decisions, and the cost of the project.

1.8.8. Figure 1-4 illustrates the relationship of factors that need to be considered in successful project planning.

1.8.9. The purpose of the project planning triad approach is managing total decision uncertainty. Total uncertainty may be viewed as the sum of analytical and field uncertainty. Analytical uncertainty is the portion that arises from variability and bias in the instrumental or analytical test method (as indicated in Table 1-3). Field uncertainty depends on factors such as the temporal and spatial variability of the target environmental population (Table 1-3). Field variability typically exceeds the analytical variability and primarily depends on the sampling design (e.g., the total number of samples, the sample mass, and the nature of field sampling and laboratory sub-sampling methods). In general, data produced by screening analytical methods will contain more analytical variability and bias than data produced by definitive methods. However, field analyses are less costly than laboratory analyses, so a greater number of field samples can be analyzed than laboratory samples for the same fixed cost. Thus, even though field analy-

ses typically contain higher analytical variability relative to laboratory analyses, a larger number of field samples can reduce the total variability more effectively than a smaller number of similarly collected laboratory samples. Field analytical methods should be scrutinized, however, because the total uncertainty does not depend on measurement precision (variability) alone; it also depends on a number of data quality elements such as analytical bias, sensitivity, and specificity (i.e., the ability to detect or quantify the analyte or contaminant of concern in the presence of other analytes or interferences in the sample).

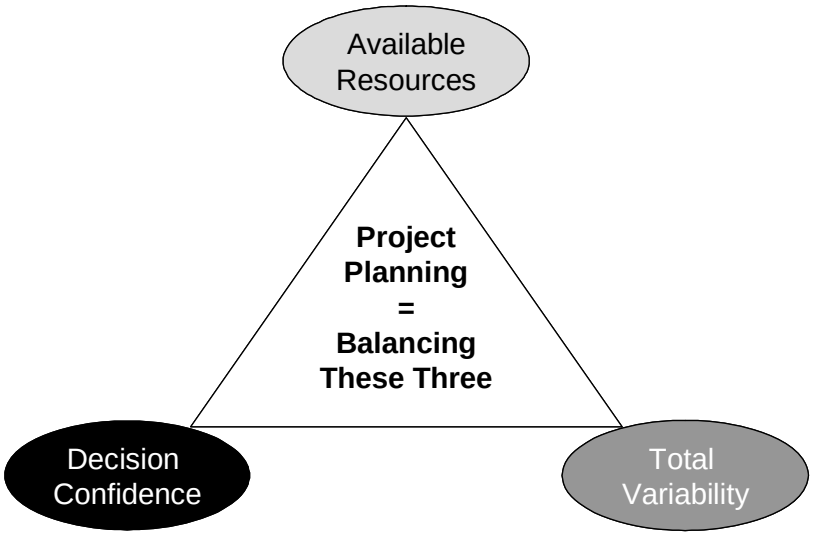


Figure 1-4. Project planning triad.

1.8.10. The triad approach also makes use of rapid turn-around times for field methods. Field methods have an advantage over laboratory methods in that they are capable of providing data to support decisions while mobilized in the field. For example, managers can modify sample locations on the basis of new information about the extent of contamination during a single mobilization. In contrast, fixed laboratory data packages are produced several weeks after sampling is complete. Remobilization may be necessary to resolve questions arising from laboratory results.

1.8.11. The triad approach is especially useful for statistical designs such as *adaptive sampling,*[*] *ranked set sampling*[*], and *systematic sampling*[*], as these designs often require larger numbers of samples. To successfully implement the approach, the capability of the field methods must be scrutinized with respect to project data quality and measurement objectives. For example, many field methods are not as sensitive or selective as laboratory methods. If the primary objective is to characterize contamination with respect to some fixed risk-based limit or cleanup goal, and the detection limit is greater than the decision limit, then comparisons of the field data

[*] Appendices C and D.

to the decision limit will not be viable. Comparisons of field and laboratory data during a pilot test phase to verify or establish correlation between two sets of results is a useful approach for evaluating and selecting field methodologies.

1.8.12. The triad approach relies on thorough, systematic planning to articulate clear project goals and encourages negotiations among stakeholders to determine the desired decision confidence. A multidisciplinary technical team then determines what information is needed to meet those goals. A key feature of this planning is identifying what uncertainties could compromise decision confidence and allowing team members with appropriate sampling and analysis expertise to explore cost-effective strategies to minimize them. Often, the most cost-effective work strategy involves the second leg of the triad, which is using a dynamic work plan to make real-time decisions in the field. The third leg of the triad uses field analytical methods to generate real-time on-site measurements that support the dynamic work plan. Projects managed using these concepts have demonstrated cost savings of up to 50% over traditional approaches.

1.8.13. The contributions to the total variability (i.e., the total precision component of the uncertainty) can be expressed as a vector sum of an analytical component and sampling component of the variability (e.g., or as a ratio of the sampling to analytical variability, say 9:1). Although the analytical variability is minimized by conventional laboratory analyses, sampling variability is often not adequately addressed. Budget constraints invariably limit the number of laboratory analyses. A combination of high laboratory analysis costs and a poor sampling design often results in a low sampling density that is not very representative of the environmental population of interest. Field studies consistently find that the sampling design, rather than analytical considerations, predominately governs the total variability.

1.8.14. When analytical costs are lower, more samples can be analyzed, yielding more confidence in the representativeness of the data set (Phase 1). This is most effective if field methods are used to generate data and a dynamic work plan rapidly resolves any uncertainty about location and volume of contamination (for example, locate and delineate hot-spots in a single field mobilization). If the analytical data quality used to manage sampling uncertainty is less than what is eventually needed to make final project decisions, such as whether the site can be declared clean, more expensive definitive analyses may be performed on samples selected to refine the feature of interest (Phase 2). However, if the initial method produces data of sufficient rigor to support defensible decision-making, then additional, expensive analyses would be redundant and unnecessary.

1.8.15. In Phase 1, analytical uncertainty (variability) increases so that unit sample costs decrease, allowing a higher sampling density than with the conventional approach. As a result, sampling uncertainty (variability) decreases, lowering the overall uncertainty in data interpretation. Sampling uncertainty is further decreased if hot-spot removal reduces the variability in contaminant concentration and if representative sampling locations for more rigorous analysis are identified based on Phase 1 information. The vector representation of uncertainty for this ap-

proach indicates that the overall uncertainty in the data set for site decision-making will be much less than the overall uncertainty in the conventional method.

1.8.16. Data quality should be judged on whether both the sampling and the analytical uncertainties in the data sets support decision-making at the desired degree of decision confidence. However, relying solely on regulator-approved, definitive analytical methods, while ignoring sampling uncertainty, easily produces uncertain decisions.

1.8.17. When field analytical methods are used, the process and resulting data are often referred to as "field screening." The term is misleading when field methods are of adequate quality to satisfy project DQOs; field analyses are not necessarily "screening" or inferior to fixed-laboratory analyses in the context of the overall end use of the data. Here, alternate terminology is proposed to reflect current EPA guidance that both sampling and analytical uncertainties must be managed to assess data quality. We consider the two terms "effective data" and "decision-quality data," to be equivalent when describing data of known quality that are effective for making defensible primary project decisions, because both sampling and analytical uncertainties have been explicitly managed to the degree necessary to meet clearly defined project goals.

1.8.18. Primary project decisions are those decisions that drive resolution of the project, such as whether or not a site is contaminated and what subsequent actions, if any, will be taken. Therefore, contaminant data are usually the data sets of interest. But data sets can interact in complex ways, and are referred to as collaborative data sets. For example, a contaminant data set considered alone might not be effective for making project decisions, yet the same data set might be more effective when combined with other data or information to manage the remaining uncertainties. Ancillary data refers to data used to support many other project decisions that fall under worker health and safety monitoring, data that help in the understanding of fate and disposition of contaminants, and data that aid in decisions about the representativeness of environmental samples.

1.8.19. This decision-making paradigm and terminology embodies the central theme of systematic project planning, the management of decision uncertainty.

CHAPTER 2
Preliminary Assessment and Site Investigation (PA/SI)

Section I
Preliminary Assessment

2.1. Introduction. A Preliminary Assessment (PA) is initiated after a CERCLA site (or suspected site) is identified. Statistical evaluations are not typically conducted for a PA. The purpose of the PA is to determine if a site poses a potential threat to human health or the environment. EPA maintains a list of actual and potential hazardous substance releases requiring CERCLA response. The property owner or agent is obliged to perform a PA; for Federal facilities, a PA is required within 18 months of listing (57 FR 31758; 17 July 1992).

2.1.1. The PA process collects information from existing resources. Generally, PA data are qualitative rather than quantitative, and do not require statistical evaluation. In some instances, historical chemical data may be available, but the PA does not require that such data be statistically manipulated. The EPA evaluates the site information according to the Hazard Ranking System (HRS) as detailed in 55 FR 51531 (14 December 1990). HRS calculations do not have statistical components. Some examples of PA information necessary to the HRS are as follows.

2.1.1.1. Identification of wastes or waste sources.

2.1.1.2. Physical site conditions, such as precipitation rates, depth to groundwater, or distance to surface water bodies.

2.1.1.3. Workers or residents at a site.

2.1.1.4. Local population within a set radius of a site.

2.1.2. Based on the results of the HRS, a site may warrant further investigation or no further action. Though quantitative statistical evaluations are not required during a PA, the following case study illustrates the value of a thorough qualitative evaluation of PA information.

2.2. Case Study 1—Examining Historical Data Sets. In the preliminary assessment of a landfill located on a manufacturing facility in Pennsylvania, some historical analytical data were available to the project team. The question raised, however, was whether or not those data would be usable in the PA. If the data were found to be usable and applicable, the landfill might be removed from further consideration in the CERCLA process. However, if the data were not found to be usable, then a Site Inspection (see Section II) would be needed. Moreover, if the data were used, prior to further validity testing (thus, explicitly assuming the data were reliable), and found later in the assessment to be erroneous, inaccurate and misleading conclusions would have been drawn.

2.2.1. Several different assessments of the data were required: i) Were the precision, accuracy, and representativeness of the data sufficient for the purpose? ii) Was the sampling design for the historical data sufficient for the purpose? and iii) Were the data comparable from historical event to historical event and could they be combined with new data, if necessary, to draw conclusions about the site?

2.2.2. The existing data were included in monitoring reports to the state. The reports consisted of little more than sample identification, date, and analytical results. Only positive detections were reported. Based on that information alone, the project team could not assess the quality of the data and concluded that unless additional information was obtained, the data could not be used as part of the PA. The site owners began to investigate the origins of the data.

2.2.3. In the interim, the project team assigned a geologist to examine the sampling design for the work. The facility had identified a single monitoring well, MW-02, as an upgradient location for comparison to a set of three downgradient wells, MW-03, MW-06, and MW-08. Through a review of well construction diagrams, as well as available topographic and hydrogeologic information, the geologist found that the well identified as upgradient was located within 3 feet of the landfill footprint, in a swale that received run-off from the landfill.

2.2.4. Thus, it was likely that the upgradient well was directly impacted by landfill operations and would not constitute an acceptable upgradient location. Further, MW-06 and MW-08 were found to be generally cross-gradient to MW-02 rather than directly downgradient, and that MW-03 had been screened in a perched aquifer, hydrologically isolated from the aquifer monitored by the other three wells.

2.2.5. Upon receipt of laboratory data packages for the historical data, the project team observed that a variety of different analytical methods and laboratories had been employed in the course of the work, resulting in mixed reporting limits and inconsistent detection of analytes. As a result of these assessments, the historical data were judged not to be usable for the PA.

2.2.6. In summary, prior monitoring appeared to indicate the presence of contamination (e.g., which would have triggered an RI), but additional evaluation data indicated that the data were not usable; therefore, an SI was initiated.

Section II
Site Inspection

2.3. Introduction. The Site Inspection (SI) is the next step in the CERCLA process. Statistical evaluations are often appropriate for an SI. Typically, the major objective of these evaluations is to establish the presence or absence of site contamination with respect to predefined decision limits. An SI is performed if the PA indicates the potential for hazardous materials to be present, if human or ecological receptors, or both, exist, and if there are potential complete exposure

pathways for the receptors. The SI generally focuses on establishing, through sampling and analysis, whether hazardous materials are present at concentrations that exceed some "screening criteria." The project planning team must establish decision limits or screening criteria prior to sampling and analyses. Generally, decision limits fall into the following categories:

2.3.1. Naturally occurring or known background levels (site-specific background information is typically unavailable at the SI stage).

2.3.2. Ecological benchmarks, which are dependent on analytes and media (typically developed with regulatory input).

2.3.3. Risk-based screening criteria for human health such as EPA Region IX Preliminary Remediation Goals (PRGs) or EPA Region III Risk-based Concentrations (RBCs) are available at the following Web sites.

http://www.epa.gov/region09/waste/sfund/prg/index.html

http://www.epa.gov/reg3hwmd/risk/index.htm

2.3.4. Applicable or relevant and appropriate requirements (ARARs). For example, Maximum Contaminant Levels (MCLs) for drinking water may be ARARs for some CERCLA sites.

2.3.5. During the DQO process, stakeholders identify the study questions, such as the presence or absence of contamination with respect to a set of decision limits, the nature and quantity of the data required to support the decision-making process, and the acceptable tolerances for decision errors. Selecting the screening criteria is *critical* for establishing both data quality objectives (DQO) and measurement quality objectives (MQOs). MQOs are established after DQO development. MQOs for analytical sensitivity must be adequate to report *quantitative* contaminant concentrations at levels less than the project decision limits. (Refer to Appendix G for a discussion of detection limits and quantitation limits.)

2.3.6. Team members must establish the DQOs for the project at the outset of the SI. In an SI, stakeholders must identify the problem at the site and how it will be evaluated, identify the decisions to be made using the data, and specify limits on that decision error. These will lead the project team to an optimal sampling design at a site. Appendix G discusses detection limits, quantitation limits, and censored data. Understanding the concepts in the context of ARARs guides part of the project planning.

2.4. Sampling Design. In general, statistical sampling designs are required to support statistical evaluations. Professional judgment, site-specific information, and DQOs must be used to select

the type of the statistical sampling design (e.g., *random*[*] as opposed to systematic sampling) and the required number of samples. The sampling design depends on factors such as the nature and distribution of the contamination in the study area, sampling cost, tolerances for decision error, and perceived level of decision uncertainty. For example, a small number of samples during the SI stage may be beneficial for short term cost considerations, but may not be adequate to achieve the desired tolerances for decision uncertainty and error and may, therefore, not be a cost-effective strategy by project closeout (as multiple sampling events rather than a single sampling event would typically be required to support decision-making).

2.4.1. Decision uncertainty refers to statistical variability, subjective judgment, randomness in the process, disagreement, and even imprecise wording inherent in the decision-making process (Moser 2000). Decision uncertainty is a function of the variability of the contaminant of concern in a study area and depends on the number of samples collected. For example, if the sample mean, $\bar{x}$, is an appropriate measure of site-wide contamination and the standard deviation of the sample mean, $s_{\bar{x}}$, measures the variability around $\bar{x}$, then the variability (and uncertainty) decreases as the number of samples n increases, because $s_{\bar{x}} = s/\sqrt{n}$. (Increasing the physical size of each sample would also decrease the variability.) It should also be noted that, in addition to decreasing the variability, $\bar{x}$ becomes a more accurate estimate of the population mean, μ, as n increases.

2.4.2. Site-specific information must be taken into account when selecting the sampling design. In particular, the team members need to identify potential source areas and any stratification they may represent. For example, suppose there are two sources of lead at a bomb reconditioning facility—stack emissions affecting surface soil and old buried waste piles affecting subsurface soil. This information can be used to design a sampling scheme for the "surface soil stratum" and a separate scheme for the "subsurface soil stratum." Likewise, there may be different study objectives for each stratum. Surface lead may be of concern for exposure of site workers and subsurface lead may be of concern for protection of groundwater. Stakeholders would need to identify these issues during project planning to develop an optimal site-wide sampling design.

[*] Appendices C and D.

2.4.3. Several different types of sampling designs are listed below. Appendix C presents a detailed explanation of these designs.

- Judgmental sampling.
- Random sampling.
  - Simple random sampling.
  - Stratified random sampling.
  - Systematic and grid sampling.
- Ranked set sampling.
- Adaptive cluster sampling.
- Composite sampling.

2.4.4. The TPP and DQO processes are used to develop an appropriate sampling design for the SI phase. Two case studies are presented below to illustrate sampling designs commonly used for SI.

2.5. Case Study 2—Judgmental Sampling, Oil/Water Separator. Project planners found an oil/water separator buried underground at a pipe mill. There was evidence of leakage to the surface soils around the tank and a release to groundwater was suspected. The objective was to determine if there was a measurable presence of oil floating on the water table.

2.5.1. Historical information and local knowledge allowed a hydrogeologist to determine the direction of groundwater flow. The hydrogeologist also knew of two monitoring wells in the area. One well was located upgradient to the separator; the second was cross-gradient.

2.5.2. The project planners decided to place a new monitoring well downgradient of the separator. Because they were looking for an oil product, the soil boring for the monitoring well was logged by a geologist who could then identify the water table depth. The well was installed so that the screen intersected the water table, where floating oil would most likely be visually detected.

2.5.3. Judgmental sampling was predominantly used in this example because the planners possessed significant existing site information. They knew the physical properties of the oil, they knew the hydrogeology of the site, and they were answering a nonquantitative question.

2.5.1. Case Study 4 predominantly illustrates the application of *composite sampling*[*] and *stratification*[†] for a SI, and the iterative nature of the DQO process when optimizing a sampling design.

[*] Appendices C and D.
[†] Appendix D.

2.6. Case Study 3—Arsenic Contamination in Soil. At an active manufacturing site, arsenic contamination was widespread in surface soils. Preliminary screening analyses and risk assessments identified worker exposure as the most likely concern. The site was initially divided (stratified) into 90 subunits related to work areas for a more in-depth evaluation of risk. Based on financial constraints, the project team was allocated a budget of $50,000 for SI sampling and analytical testing.

2.6.1. The aggregate initial cost of a field grab sample was $175, with $100 attributed to field collection and $75 attributed to laboratory analysis. The expected percent relative standard deviation (%RSD) for the analytical (laboratory) measurements was 5%. The estimated standard deviation, *s*, for the analytical method, at the decision limit of 600 ppm, was computed as 5% of 600 ppm or 30 ppm.

2.6.2. The planning team estimated the field component of the variability to be 10 times greater than the laboratory component of the variability. Thus, the %RSD for the field component of the variability was calculated by multiplying the %RSD for the analytical measurements by 10 (yielding a field component %RSD of 50%). This estimate was then multiplied by 600 ppm to yield a value of *s* equal to 300 ppm for the field component of variability (i.e., 50% of 600 ppm). The estimates for field and analytical variability (i.e., variance or s^2) were then combined and the standard deviation was calculated (s = 330 ppm). The maximum observed arsenic concentration was 720 ppm. The analytical method was deemed appropriate by the planning team. If historical sampling data were available, the data would be used to estimate the field variance and to test for normality.

2.6.3. The planning team principally considered two sampling design alternatives—simple random sampling and composite sampling (see Appendix C for a review of each sampling method). A *t*-test was used to calculate the sample size for simple random sampling (Appendix F). Given a decision error limit of $\alpha = 0.01$, more than 200 samples per work area would have been required (refer to Appendix L for a review of methods involved in setting and testing hypotheses). The total cost of this sampling effort would have exceeded $3 million.

2.6.4. Using similar methods, the team explored composite sampling, which would have required 30 samples to be collected per work area for a cost of over $1 million. Given the considerable cost burdens for both proposed sampling designs, the team decided to return to Step 6 of the DQO process and modify the decision error limits. The team found that by increasing α to 0.05, the composite sampling design would require the collection of 13 samples for each of the 90 work areas. This revised design had a total cost of $204,750, approximately one-fifth of the original estimate.

2.6.5. The team realized that they would have to find other means of generating an appropriate design while remaining within budget. To do this, the project team redefined the boundaries of the study (by revisiting Step 4 of the DQO process). The team recognized that one of the

drivers of the cost was the large number of separate study units (previously, the calculated sample size was applied to each of the study units). The planning team used exposure information for the contaminant to map out the potential or expected pathways in the surface soils through which the contaminant could spread. The potential pathways were categorized into four distinct spatial units.

2.6.6. Rather than collect data and make decisions for each of the 90 individual work areas, the team decided to sample and make decisions for each of the four risk areas. Recognizing that these larger areas carried greater decision error consequences, the team revisited Step 6 of the DQO process and established new limits for decision errors applicable to the four risk areas. The team established different decision confidence limits for each and recalculated the number of samples required. The cost of implementing this design was $38,850, which fell within the $50,000 budget for the sampling and analysis.

2.7. General Review of Sample Size Determination[*]. For typical statistical sampling designs, there are well-defined relationships between the number of required samples (i.e., sample size), tolerance for decision errors, and inherent variability of the analytical measurements and the target environmental population. One such relationship states that the sample size increases as the tolerance for decision error decreases or the variability increases. The sample size must be equal to or greater than the sample size required to achieve predetermined tolerances for decision errors. When confidence limits for the mean are of interest, an appropriate sample size is required to generate a sufficiently precise estimate of the true mean concentration of a chemical contaminant (refer to Paragraph 3.11 and Appendix K for additional discussion of confidence limits). For the example presented above, the sample size must be adequate to demonstrate that the upper limit of the CI for μ is less than the applicable regulatory threshold, RT. The required sample size must increase as s^2 increases and as the difference Δ (RT $-$ $\bar{x}$) decreases. In a well-conceived sampling plan for a solid waste, every effort should be made to estimate the values of $\bar{x}$ and s^2 before sampling starts. Case Study 3 illustrated that decision confidence affects sample size. Case Study 4 illustrates this concept in a different setting.

2.8. Case Study 4—Effect of Decision Confidence on Sample Number. Upon promulgation of the Toxicity Characteristic Leaching Procedure (TCLP) rule, a steel mill in Maryland contracted with a consultant to collect samples from various waste streams within the facility for TCLP analysis of metals (this case study considers only the cadmium data). One such waste stream was from a wastewater treatment system and consisted of collected sludges. Although no previous analysis of sludges had been done, cadmium had been monitored in the wastewater stream before treatment. The project manager believed that the wastewater data would be sufficient for establishing routine variability of cadmium in the sludge, assuming there were no great differences in the treatment process over time and a 10 times concentration factor from wastewater to sludge.

[*] Appendix C.

2.8.1. The project manager decided to use the past year's wastewater data to make preliminary estimates of the number of samples required to meet the statistical confidence requirements of the TCLP rule (i.e., $\alpha = 0.2$). Four results (in milligrams per liter [mg/L]) were available from the previous year as follows: 14.2, 9.6, 21.7, and 19.3.

2.8.2. The mean and variance of the results (as adjusted for concentration to sludge) were the following: $\bar{x} = 1.6$ mg/L and $s^2 = 2.2$ mg/L, respectively. The proposed water regulatory threshold value (RT) was 1 mg/L. Using the formula for simple random sampling, the project manager calculated the number of samples required as follows:

$$n = (t^2 \times s^2) \div (\mathrm{RT} - \bar{x})^2$$

where: n = number of samples required
t = Student's value for n–1 degrees of freedom and 0.8 confidence
s^2 = sample variance
$\bar{x}$ = sample mean
RT = regulatory threshold.

2.8.3. Thus, $n = [(0.9785)^2 \times 2.2]/(1 - 1.6)^2 = 6$ samples. Samples are an integer value, and should be reported without decimal fractions. (The value of t may be obtained from Table B-23, where $df = 3$ and $p = 0.8$.) Assuming a sampling cost of $50 per sample and an analytical cost of $25 per sample, this testing would cost $450.

2.8.4. The client's attorneys asked what the effect would be should they wish to establish a safety margin by increasing the decision confidence to $\alpha = 0.05$. The revised plan would require

$n = [(2.353)^2 \times 2.2]/(1 - 1.6)^2 = 34$ samples, or a sampling and analysis cost of $2,550.

2.9. Summary of Case Studies. Case studies 2 through 4 illustrate the multitude of related factors that must be considered when evaluating which sampling design to apply in a particular SI. When evaluating alternative sampling plans, planners may anticipate the concentration patterns likely to be present in the target population. Advanced information about these patterns can be used to design a plan that will estimate population parameters with greater accuracy and less cost than can otherwise be achieved.

2.10. Comparing On-site Data to Fixed Screening Criteria. In the data analysis phase of the SI, environmental scientists compare site data to screening values using either qualitative or quantitative statistical evaluations. The following provides a discussion of qualitative and quantitative evaluations.

2.10.1. *Qualitative Statistical Evaluations.* The EPA has developed risk-based screening criteria in the form of PRGs and RBCs. These criteria are frequently applied at the SI stage to

identify whether the site as a whole may need further attention in an RI/FS. Many screening criteria exist at both the Federal and state government level. Thus, comparisons are frequently made against the lowest of several screening criteria that can be applied to a given data set from a given location. The technical team must ensure that the criteria are being applied properly (i.e., not all screening criteria are applicable to every site), and that the implications are clear in the conclusions of the SI. For example, if site data exceed a standard developed to protect groundwater from soil leaching of contamination, but do not exceed an applicable human health standard, the team should report the results with the implications of these differences noted in the conclusions.

2.10.2. One typical qualitative method of comparing data decision limits entails the use of a spreadsheet or database. The decision limits and individual sample results are presented in a tabular format and each detected analyte concentration is compared to the corresponding screening values for that analyte. (It may be necessary to compare a single contaminant of concern to only the lowest decision limit or several different decision limits.) Table 2-1 is an example of such a spreadsheet.

2.10.3. The primary pitfall of this qualitative strategy is that the uncertainty associated with the reported results is not considered when the results are compared to the decision limits. Thus, the reported results may actually be equal to or exceed decision limits when uncertainty is taken into consideration. If this is the case, especially in the event the decision limit is exceeded, the wrong conclusion would be drawn. The ramification of an erroneous conclusion will vary, depending on the nature of the problem under investigation; nevertheless, this is an outcome that should be avoided or at least minimized.

2.10.4. Historically, environmental researchers have tended to screen analytical results into two categories—greater than the standard or less than the standard. Through advances in research and technology, three categories now exist against which analytical results can be compared: i) the reported value clearly exceeds the standard (when bias and variability are taken into account); ii) the reported value clearly does not exceed the standard; and iii) the result is inconclusive. This last conclusion is reached when the uncertainty is too large for reliable decision-making.

2.10.5. Table 2-1 illustrates how qualitative information may be used to support the decision making process when SI data are qualitatively, rather than statistically, compared to decision limits. In particular, information regarding the quality of the data, obtained in the data validation process, is used to determine whether contamination is present at concentrations greater or less than project decision limits. All applicable screening criteria are displayed in Table 2-1. For example, the “S” column reports the results of comparing each analyte concentration and the lowest screening limit. One of three codes is entered in this column for the three possible conditions identified in the preceding paragraph. An “X” is recorded if the reported values appear to be well above the decision limit, an “I” if the result is inconclusive, and a blank space if

the result appears to be well below the limit. Select results from Table 2-1 are discussed below to illustrate the nature of the screening evaluation.

2.10.5.1. Tetrachloroethene results in IRP-49 (1.2 ppb) and IRP-51 (17.08 ppb) both exceed the PRG (1.1 ppb). Although the value in IRP-49 is barely above the PRG, it reports the results as two significant figures, so we must accept its value as exceeding the PRG. However, accounting for analytical error, typically between 20 and 30% (as a conservative estimate), this result would be inconclusive. The researcher then must choose whether to conduct additional testing or accept the value of IRP-49 as an exceedance. The latter would be selected only if a conservative estimate was desired.

2.10.5.2. In IRP-49 (0.2 ppb) and IRP-51 (0.2 ppb), the reported concentration is not distinguishable from the PRG when compared on the basis of just one significant figure. Therefore, these results are inconclusive.

2.10.5.3. Several chloromethane results are marked inconclusive because of blank contamination. The only sample without blank contamination, IRP-39, was below the PRG (PRG = 1.5 ppb; IRP-39 = 0.2 ppb). The reported concentration was qualified with a J flag because it is less than the quantitation limit of 1 ppb. (The quantitation limits are not listed in Table 2-1, but were obtained from the laboratory's data package.)

2.10.5.4. For bromodichloromethane in sample IRP-48 (0.2 ppb), the reported concentration is biased low and is less than the quantitation limit of 1 ppb, so this exceedance of a PRG (0.18 ppb) is conclusive. In sample IRP-51 (0.1 ppb), the result is also biased low and is just below the PRG, so this result is also not conclusive.

2.10.5.5. For chloroform in sample IRP-39 (0.4 ppb), the reported concentration is qualified with a J flag because it is less than the quantitation limit of 1 ppb. As the reported result is quantitatively estimated, it does not reliably demonstrate that chloroform is present above the PRG.

2.10.5.6. Benzo(a)pyrene was reported in sample IRP-49 (0.278 ppb) above the PRG limit (0.0092 ppb). However, the detection limit (0.014 ppb) is above the PRG for the remaining samples. Only by achieving a lower detection limit is it possible to determine whether the non-detects are a problem. The results for benzo(a)pyrene are marked inconclusive. All of the arsenic non-detects are inconclusive based on a similar rationale.

Table 2-1.
Site Screening Data Table

		EPA MCL	Region IX PRG (1999)	IRP-39				IRP-48				IRP-49				IRP-51			
Analyte	**Units**	**Tap Water**			**L**	**V**	**S**		**L**	**V**	**S**		**L**	**V**	**S**		**L**	**V**	**S**
Organics																			
Bromodichloromethane	µg/L	—	0.18	0.1	U			0.2	J	L, s	I	0.1	U			0.1	L, s		I
Carbon Tetrachloride	µg/L	5	0.17	0.1	U			0.1	U			0.1				0.4	J	J	I
Chloroform	µg/L	—	0.16	0.4	J	J	I	0.1	U			0.1	U			0.1	U	U	
Chloromethane	µg/L	—	1.5	0.2	J	J		6.1		B	I	1.6		B	I	3.7		B	I
Methylene Chloride	µg/L	5	4.3	0.1	U			0.1	U			0.1	U			0.1	U		
Trichloroethene	µg/L	5	1.6	0.4	J	J		0.1	U			18.7			X	18.1			X
Tetrachloroethene	µg/L	5	1.1	0.1	U			0.1	U			1.2			X	17.1			X
Benzo(a)pyrene	µg/L	0.2	0.0092	0.014	U		I	0.014	U		I	0.278			X	0.014	U		I
Inorganics																			
Arsenic	mg/L	50	0.045	0.7	U		I	0.7	U		I	0.7	U			0.7	U		I
Chloride	mg/L	250	—	311			X	15.8				265			I	134.7			
Lead	mg/L	15	—	0.3	U	K		0.3	U	K		8				10			
Nickel	mg/L	—	730	590				29.0				214				198.0			
Sulfate	mg/L	250	—	44.0				5.98				41.6				21.45			
Thallium	mg/L	2	2.9	1.4				0.8	U			0.8	U			0.8	U		
Vanadium	mg/L	—	260	1.4				1.0	U			3.0				5.0			

Notes: L column contains the laboratory flags. V column contains the validation flags. S column contains screening results.

Flags: U – Not detected above reported detection limit.

B – Not detected substantially above a laboratory or field blank.

L – Biased low.

K– Biased high.

s – Surrogate failure.

J – Quantitatively estimated

Screening Codes:

X – sample concentration unequivocally exceeds the lowest screening standard.

I – sample concentration comparison to screening standard is inconclusive.

– A blank cell indicates that the sample concentration unequivocally does not exceed the lowest screening standard.

2.10.5.7. Though the reported concentration of chloride in sample IRP-49 (265 mg/L) is not qualified as estimated and exceeds the decision limit (250 mg/L), the result is marked inconclusive because the difference between the detected concentration and the decision limit is less than 5%, which is smaller than the analytical error for the test method (e.g., the error tolerance for the test method is typically 5 to 20%).

2.10.6. These results illustrate the critical importance of estimating and incorporating into decision-making knowledge of both the field and laboratory components of variance. One fundamental error is treating the reported results as conclusive when in fact they are not. The values represented in this table are measurements, and measurements contain bias and variability that must be accounted for in decision-making. (See EM 200-1-10 for additional guidance on the data review strategies that were primarily used to qualify the results in Table 2-1.)

2.11. Quantitative Statistical Evaluations. When the results of the qualitative statistical evaluations are inconclusive, further investigation is required. DQOs must be revised so that the parameter of interest is no longer a single datum per location. Instead, multiple samples are collected for those uncertain locations and the resulting distribution of values is compared to the decision limit using quantitative statistical tests. The results would typically be statistically compared to decision limits using *one-sample tests*[*] for central tendency, as discussed below.

2.11.1. All statistical tests require the user to make certain assumptions about the data to perform the statistical test. The user must demonstrate that the underlying assumptions for a particular statistical test are reasonable before doing the test. With respect to these underlying assumptions, statistical tests can be roughly categorized as either *parametric*[†] or *non-parametric*.[†] When non-parametric tests are conducted, data sets are required to satisfy fewer assumptions than for the corresponding *parametric tests*.[†] In particular, a parametric statistical test assumes a specific *distribution*[†] for the data (i.e., the entire population is described by some specific mathematical function), such as the bell-shaped curve for the *normal distribution*[‡]. Statistical plots of actual measured sample concentrations must be substantively consistent with the corresponding plots generated using the theoretical functional relationship. Tests that require normal or log normal distributions are most commonly used. (A data set is log normal if, when the log of each datum is calculated, the resulting set of values is normally distributed.) Common graphical methods (i.e., plots) are presented in Appendix J. In addition, an overview of the evaluation of distribution assumptions is presented in Section III of Chapter 3.

2.11.2. It should also be noted that parametric tests become problematic, and may not be possible to perform, when the data sets contain a significant number of *censored*[§] values (i.e., analyte concentrations reported as non-detects). However, as described in Appendix H, it may be

[*] Appendix L.
[†] Appendices H and I.
[‡] Appendices E, F, and J.
[§] Appendix H.

possible to use the *Poisson distribution*[*] for highly censored data. Parametric tests are also problematic when there are outliers. The possibility of *outliers*[†] should be considered in every analysis.

2.11.3. Non-parametric tests do not assume a specific functional relationship for the data distribution. These tests tend to be less sensitive to outliers and non-detects than parametric tests. Although non-parametric tests are more applicable relative to parametric tests, non-parametric tests tend to be less statistically *powerful*[‡] than parametric tests. In essence, this means that more samples must be collected for a non-parametric test relative to the corresponding parametric test to make decisions at the same level of confidence.

2.11.4. Background concentrations of naturally occurring and anthropogenically derived compounds are also possible screening criteria. However, there are few instances in which such background levels are available at the SI stage. Sometimes a "site-wide" statistical background study has been done. If such a study is available, *two-sample statistical tests*[§] would be used to compare the study area data set with the "site-wide" background data set. (As the name implies, a two-sample statistical test is predominantly a statistical evaluation to compare two separate sets of data.) Because an RI often includes specific sampling for background, the determination of background levels and their usefulness is described in Chapter 3. If the SI is the first sampling event for a site, there is a low probability that site-specific background sample data exist.

[*] Appendices E, G, and H.
[†] Appendix I.
[‡] Appendix O.
[§] Appendix M.

CHAPTER 3
Remedial Investigation/Feasibility Study (RI/FS)

3.1. Introduction. If, based on the PA/SI, a site warrants listing on the National Priorities List (NPL), an RI/FS is performed at the site.

3.1.1. The RI is the stage in the CERCLA process for collecting data to do the following.

3.1.1.1. Characterize site conditions (e.g., thickness of unsaturated soil [vadose zone], depth to groundwater, vegetative cover, background conditions).

3.1.1.2. Determine the types, conditions, and distribution of the waste contamination in affected media.

3.1.1.3. Assess risk to human health and the environment.

3.1.1.4. Conduct treatability tests to evaluate the potential performance and cost of the treatment technologies that are under consideration.

3.1.2. The FS is the stage for the development, screening, and detailed evaluation of remedial actions.

3.1.3. The RI and FS are intimately linked. Data from the RI influence the development of remedial alternatives in the FS, which in turn affect the data needs and scope of treatability studies and additional field investigations. This phased approach encourages the planning team to continually plan the site characterization effort, which minimizes the collection of unnecessary data and maximizes data quality.

3.1.4. As in the SI phase, the initial statistical elements in the RI process involve the development of DQOs. The statistical evaluations used for the RI typically include those performed for the SI. For example, as in the SI, site data are often statistically compared to some set of fixed decision limits and upper confidence limits are often established (as discussed in Chapter 2). In general, the statistical evaluations are more common for RIs than SIs, and the statistical analysis tends to be more comprehensive. In part, this is because typically data coverage is greater and the RI data quality objectives are more robust. For example, while the SI predominantly focuses on statistical evaluations to resolve the presence or absence of contamination, the RI reaches for a determination of the extent of contamination. Critical to the onset of an RI is the identification of Applicable or Relevant and Appropriate Requirements (ARARs), which, in turn, may influence the identification of areas requiring remediation. Both sampling strategy and extent of contamination are influenced by the selection of ARARs. ARARs help identify the best analytical procedures needed to reach decision limits. This aspect of DQOs is addressed in Appendix C.

Section I
Site Characterization

3.2. Introduction. The first two objectives of the RI (subparagraphs 3.1.1.1 to 3.1.1.4) are combined for discussion in this Paragraph. The process of site characterization is linked to the procedures described in Section II of Chapter 2, where sampling distribution design was discussed. In the RI stage, sample design is likely to be influenced by SI data. In turn, these SI results affect the statistical methods at the planner's disposal for collection of site data.

3.2.1. When scoping for the SI, project planners have expectations about the probable location and nature of contamination. By the time a site reaches the RI, some usable information is usually available. In particular, if a contaminant was identified in the SI, planners may have an idea of the mean and standard deviation of contaminant concentrations. These initial estimates assist in devising a statistical sampling design at the RI stage. Two examples of using site data to support sampling design are presented in this Paragraph. These are "hot spot" sampling and geostatistical sampling, the fundamentals of which are presented in Appendices C, J, and Q.

3.2.2. A "hot-spot" typically refers to a localized area of high concentration, but is often otherwise poorly defined (e.g., criteria for the size and concentration of hot spots are often arbitrary or not specified). Hot-spots are not uncommon at sites where waste was released in an isolated region, perhaps during a spill. In addition, hot-spots may occur within broader regions with low, but detectable, levels of contamination. One example of this may be when an area was used to process waste disposal over some time and, at times when a shop or operation was cleaning house, a high concentration of waste would be deposited. However, sample concentrations that exceed a regulatory threshold or other decision limit should not be considered to be hot-spots if these concentrations appear to be randomly distributed and will not necessarily be of concern if they represent a small portion of study area and contain a small contaminant mass.

3.2.3. Case study 1 presents an RI application of the hot-spot identification method discussed in Appendix C.

3.2.4. In this instance, professional judgment led to the determination of the size and shape of the hot-spot. The reader is urged to vary S and L to identify the sensitivity of hot-spot sampling grids to the assumptions.

3.2.5. As stated previously, there is typically some knowledge of contaminant distribution at a site by the time an RI begins. Geostatistics allow an investigator to extrapolate (and interpolate) what is known in one location to other nearby related locations. Its application relies on the fact that, given a known concentration at one location, an adjacent location is likely to have a similar concentration. The greater the distance from the known concentration, the greater uncertainty there is in predicting a concentration at an unsampled location. This situation can be de-

scribed as a spatial correlation, because correlations are related to how close samples are to one another. Geostatistical methods are described in detail in Appendices J and Q.

3.2.6. Case Study 2 illustrates the use of geostatistics for reducing uncertainty in a project. Although geostatistical techniques are more common for RIs than SIs, they may also be used for SIs if sufficient site data are available.

3.2.7. One of the major RI objectives is identifying the distribution of contamination at a site. As useful as geostatistics are in helping with sampling design, they may also be used in interpreting sample data. The geostatistical method known as kriging (Appendix J) is an effective method for interpolating site concentration data under conditions where spatial correlation exists. Kriging is a weighted-moving-average interpolation method. The USEPA developed a two-dimensional kriging package, which is useful in providing a fundamental introduction to the technique (Geo-EAS; EPA/600/4-88/033). Kriging as a method of contouring is described in several readily available texts, and typically requires the use of commercially available computer software with kriging options for contouring (e.g., Surfer, EVS).

3.3. Case Study 1—Hot-Spot Identification. The project team attempted to locate a hot-spot resulting from an uncontrolled water release within a larger storage area. The total storage area was approximately 150 by 200 feet. Because the suspected waste was spilled as a liquid, the hot-spot was assumed to be approximately circular. A best estimate of the diameter was approximately 20 feet. The method proceeded in steps as follows:

3.3.1. A circular hot-spot means S equals 1.

3.3.2. The radius of the target spot is 10 feet.

3.3.3. The team assigns a value of 0.1 to the acceptable risk of not finding the hot-spot.

3.3.4. Using S and β, refer to Table D-1 (or nomographs presented in Gilbert, 1987) to determine that L/G is 0.55 for a square grid and 0.50 for a triangular grid.

3.3.5. Using the relationship L/G and the assumed radius of 10 feet, we see that square grid spacing is 18 feet and triangular grid spacing is 20 feet (values are rounded to the nearest foot to reflect the significant figures).

3.3.6. One sample will be placed at each grid node in the storage area, so that a square grid requires 88 samples and a triangular grid requires 75 samples.

3.4. Case Study 2—Using Geostatistics in Project Planning to Reduce Uncertainty and Cost. At a site in the Midwest, project planners were asked to assess a site potentially contami-

nated with lead at levels exceeding risk-based limits. A SI was conducted using a grid system over areas that were suspected of being contaminated based on historical information.

3.4.1. The project team identified lead concentrations in soil exceeding threshold values in various areas of the site (red circles in Figure 3-1). They were required to move on to an RI/FS to more fully characterize the nature and extent of contamination and develop preliminary estimates of cost for a removal action. Initially, the team intended to collect numerous additional samples on a grid (green circles in Figure 3-1) to more fully delineate the extent of contamination. However, the project geologist suggested the use of geostatistics as a means of reducing the number of samples without increasing uncertainty.

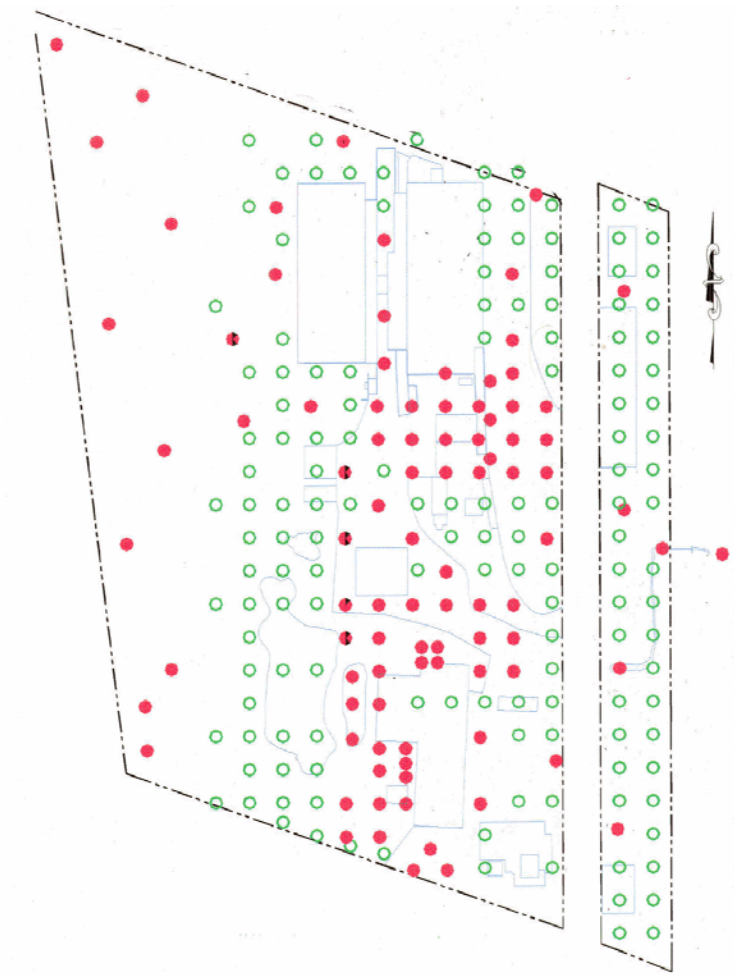


Figure 3-1. Initial sampling grid and proposed new samples.

3.4.2. Geostatistics can predict both the concentration and the uncertainty for an unsampled portion of the study area. In essence, spatial correlations for contaminant concentrations es-

tablished from the existing data set are used to "extrapolate" sample concentrations and uncertainty for other portions of the study area. Consequently, the team was able to use a geostatistical evaluation to assess the value of collecting additional samples at any given location in the grid. Simply put, the team recognized that in any sampling and analysis system there will be bias and variability, and that estimates of that bias and variability could be made using the existing data. Thus, at any location where the estimate of uncertainty from the geostatistical prediction was less than the uncertainty from sampling and analysis, the team reasoned that there was no value in collecting additional samples.

3.4.3. The final sampling plan required the addition of only seven new sampling points (shown as black circles in Figure 3-2) with associated cost savings of over $12,000.

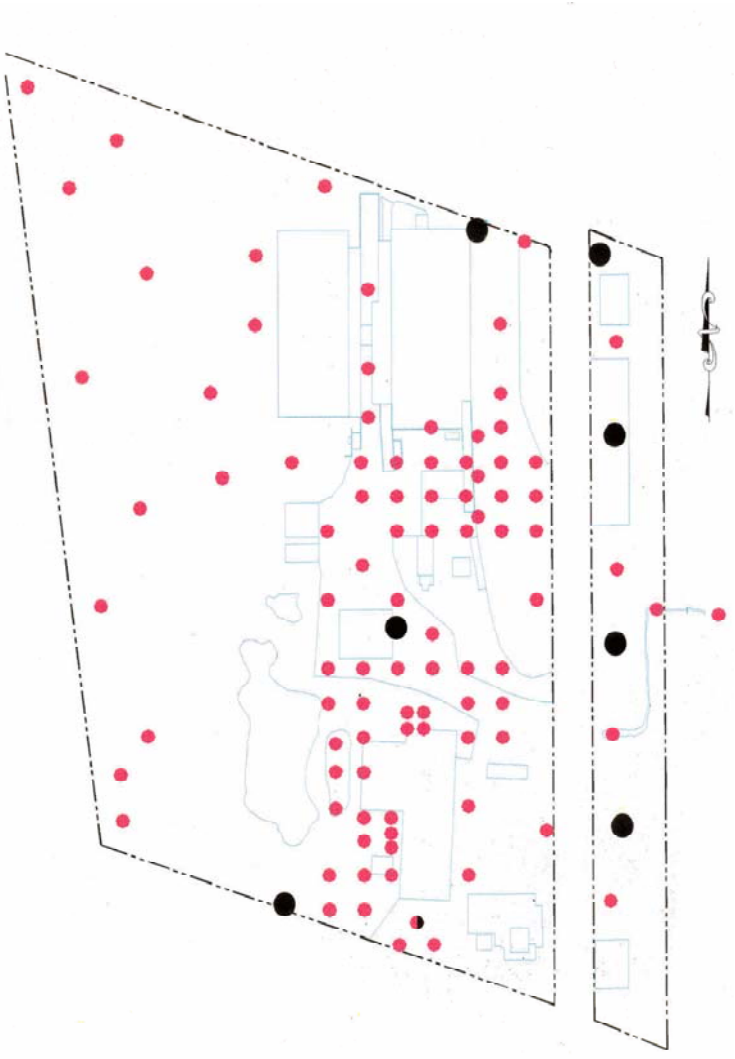


Figure 3-2. Samples required after geostatistical analysis.

Section II
Background Comparisons

3.5. Introduction. Not all chemicals detected at hazardous waste sites originate from site-related activities; for example, metals in soil and groundwater are often present because of natural geological conditions. Similarly, anthropogenic activities unrelated to a site frequently contribute certain organic chemicals (e.g., polycyclic aromatic hydrocarbons [PAHs] or pesticides derived from urban or agricultural sources; EPA SOW No. 788). If site sample concentrations for a specific compound are similar to or lower than background concentrations[*], there may be no need to consider potential remedial actions with respect to that compound. This determination can be quantitatively defended by use of statistical comparison methods.

3.5.1. The project team should determine the background sampling locations and parameters during the planning stages of the RI. Separating and identifying background sample locations from portions of the study area that have been potentially affected by waste handling activities is an example of stratification. The critical factor distinguishing a background sample from the site lies in understanding where contaminated areas end and natural conditions begin. Such samples may be located upwind, upstream, or upgradient from the waste site. Background data should be drawn from media that physically represent the study area; they should be from the same soil type or geological deposit, same type of surface water system (for example, freshwater versus saltwater; wet season versus dry season), or from the same aquifer as the site data. It is also critical to collect the background samples in substantively the same manner that the site samples are collected (same analytical method, volume of sample, etc). The sampling design and analytical methodology for the background and the site study areas must be similar. For example, erroneous conclusions can result if judgmental sampling is done for the site study area but random sampling is done for the background study area.

3.5.2. Background locations should be in a nearby portion of the region unaffected by site activities. As a caveat, site planners should be skeptical if regulators prefer to limit background sampling to only pristine areas; doing so will potentially result in erroneously concluding that the study area has been adversely impacted by site-related waste handling activities.

3.6. Does Background Soil Differ From Site Soil? The USEPA has developed guidance for addressing whether site soil characteristics differ from background (EPA/540-R-01-003 and EPA/540/S-96/500). The guidance EPA/540-R-01-003 emphasizes the formulation of DQOs in devising background sampling design and subsequent site to background testing. The focus of the cited guidance is only to determine whether site and background soil chemistry differ. It does not establish comparison standards, or levels of background that may replace unnaturally low risk-based clean-up goals.

[*] Background does not mean pristine or unaffected by human activity, especially at sites in heavily industrialized areas.

3.6.1. Fundamentally, the USEPA guidance (EPA/540-R-01-003) identifies two forms of background testing:

3.6.1.1. *Background Test Form 1*. Tests the null hypothesis that the mean contaminant concentration in samples from the site waste handling area is less than or equal to the mean concentration in background areas.

3.6.1.2. *Background Test Form 2*. Tests the null hypothesis that the mean contaminant concentration in samples from the site waste handling area exceeds the mean concentration in background areas by more than a specified margin (e.g., by 50 ppm).

3.6.2. Before continuing with this approach, investigators need to be certain that these tests are applied to random sample data sets collected from both the site and background locations. Typically, site sampling may have a component of judgmental sampling, meaning samples were biased to expected contaminated areas of a site. In such cases, the background testing cannot be applied.

3.6.3. The project planning team should establish which form of background testing will be applied at the onset of the RI planning process. In addition, the planning team needs to establish the levels of acceptable levels of error in the decision-making. This will differ from site to site, and will depend on the desires of the project planning team members.

3.6.4. The USEPA guidance also provides examples for the application of test methods that may be applied to the background test forms (EPA/540-R-01-003; Table 3-1). These are:

3.6.4.1. *Descriptive Summary Statistics*. These (e.g., mean, median, standard deviation, variance, percentiles—see Appendix D) may be used as a preliminary screening tool for comparison with site history and land use activities in the establishment of background. EPA considers these "simple and straightforward [but having low] statistical rigor."

3.6.4.2. *Simple Comparisons*. These (i.e., greater than maximum) may be used with very small data sets. This approach is not recommended.

3.6.4.3. *Parametric Tests*. These (e.g., Student *t*-test–see Appendix F) may be used if a larger number of data points is available ($n > 25$). EPA states that parametric tests require approximate normality of the estimated means and recommends that, for smaller data sets, investigators examine data for normality or lognormality in distribution. EPA considers this application statistically robust enough to be used frequently in parametric data analysis.

3.6.4.4. *Nonparametric Tests*. These (e.g., Wilcoxon Rank Sum Test—see Appendix M) may be used when data are not normally distributed, as rank-ordered tests make no assumption

on distribution. Again, EPA considers this approach statistically robust and to be used frequently in background estimation.

3.6.5. The list of methods is not complete, but, by reviewing the appropriate Appendix, users of this Manual may identify the most appropriate statistical method for site application. USEPA guidance leans heavily toward parametric and nonparametric tests, which in turn rely on establishing whether data are normal or lognormal (see Appendix F).

3.6.6. The U.S. Department of the Navy (DON) also developed statistical guidance for evaluating background in soils (UG-2049-ENV). Like the USEPA method, the guidance suggests comparative methods for testing whether site data differ from background. However, DON guidance is unique, in part, because it also relies on geochemical relationships. UG-2049-ENV provides guidance for evaluating the geology of the site and the geochemical characteristics of site soils as they relate to background analyses. The procedures outlined in UG-2049-ENV can be quite useful for USACE projects and are recommended as a resource for additional reading.

3.6.7. This "geochemical method" is often used when reference area data are not available. The method may be used to extract background concentration ranges by evaluating correlated background chemicals using on-site data only (i.e., no background area need be sampled). The key concept is that if the site has not been affected by a release, then only one population exists at a site; if a release has affected the site, then overlapping of different population characteristics would be evident in the data.

3.7. Simple Background Comparison. Investigators are more likely to rely on regional background at the SI stage than the RI. As the text below states, site-specific background is more desirable, but SI project budgets rarely allow for a full background study and such regional comparisons are still useful. Background concentrations are typically not known prior to RI activities, and sampling for background should be scoped in the planning stages of the RI. In some instances, background criteria are available as regulatory limits, as Case Study 3 illustrates. (Although the case study could also apply in an SI [Chapter 2], it is presented here to illustrate the concepts that arise for background comparisons all in one section of this document.)

3.8. Case Study 3—Comparison to Regional Background. Site-specific background concentrations are typically not known prior to RI activities, and sampling for background should be scoped in the planning stages of the RI. In some instances, regional background values may be compared to site data.

3.8.1. Texas has established soil background levels that can be used in the screening process if site-specific background levels are not available. Soil data from one site proposed for redevelopment were compared to Texas background levels. Texas regulation states that if the maximum concentration of the chemical under investigation does not exceed the Texas soil background level, then that chemical is not of concern. The site analytical data were reviewed for

quality and applicability. Based on the review, the project team was satisfied that the site analytical data were of sufficient quality for use in evaluating the site. The soil analytical data (in mg/kg) for chromium were:

6.17	4.31	4.38	6.07	5.68
2.86	5.08	4.98	2.22	15.30
4.75	3.56	4.48	3.46	2.63

3.8.2. The maximum concentration for chromium at the site is 15.30 mg/kg. The Texas soil background level for soil is 30 mg/kg. Therefore, chromium would not be a chemical of concern at the site.

3.8.3. As indicated in the USEPA guidance, such a comparison lacks statistical rigor, but is useful for guiding the project planners in the next phase of investigation.

3.8.4. At this stage, the comparison to regional background is merely sufficient to proceed to additional phases of site chromium evaluation.

3.9. Parametric and Nonparametric Tests. In the preceding case study, the regulatory community established background concentrations. It is far more desirable for local background levels to be assessed and applied. Differences related to sample medium, sampling method, or analytical method are less likely to arise in site-specific background data than regional background data. However, the project must be budgeted for a sufficient number of samples to characterize site-specific background conditions; a large number of samples may be required to characterize heterogeneous background media. If the regional background data (e.g., the background data from a very limited site-specific background study) are shown to be statistically different from a waste site, it may also be attributable to differences in water quality or soil types between the site and the location where the regional background data were collected, and not necessarily related to a waste release. Therefore, a thorough evaluation of local background conditions is preferred to the use of regional background levels.

3.9.1. Instructions and guidance for selecting analytical procedures as part of DQOs should be applied to the background data set with the eventual uses of background data in mind. For statistical comparison, background measurements need to be random. In addition, the power of statistical comparison may be greater if the background results are normally or lognormally distributed. Although the distribution of background measurements cannot be guaranteed, either random or systematic sampling of background should be a component of the sampling plan. (Note that given spatial correlation, systematic samples spaced closer than the geostatistical range may not be independent. Sampling methods are addressed in Appendix C.) Once a set of background samples have been collected, comparison methods are applied using the statistical procedures addressed in Appendix M or N.

3.9.2. A *random sampling*[*] design is typically used to characterize the background study area. *Two-sample statistical tests*[*] are then typically used to compare the site data set to the background data set. Two-sample tests, described in Appendix M, are summarized in Table 3-1.

3.9.3. An example of determining COPCs using background population tests is presented in case study 4.

Table 3-1.
Background Population Comparison

Percent Detections in Site Data	Percent Detections in Background Data	Test
0–100	0	No comparison
> 0–100	< 10	Poisson UTL
10–50	10- 50	Test proportions
> 50	> 50	Mann-Whitney test,
85–100	85–100	Student's *t* test* or Mann-Whitney test
*Student's *t* test should be used if the distributions in the site and background data sets are the same; otherwise, the Mann-Whitney test should be used.		

3.10. Case Study 4—Establishing and Comparing Background Concentrations to On-site Data. At a military installation in Utah, samples were collected for metals in soil—seven on site and four at background locations. This case study focuses on chromium. The chromium results were as follows (mg/kg):

SS01	SS02	SS03	SS04	SS05	SS06	SS07	BKG1	BKG2	BKG3	BKG4
4.3	2.7	2.2	3.2	<1	3.6	2.4	1.6	1.8	2.6	1.6

3.10.1. Because the site data had an 85% detection rate, one-half the reporting limit was substituted for each non-detect for the statistical calculations.

3.10.2. Both background and site data were determined to be normally distributed at a 90% confidence level. An *F-test* was used to compare the variance of the background data set to the variance of the site data set. The result of the *F*-test indicated that the variances are equal.

3.10.3. Thus, a *two-sample t-test* (with equal variances) was used to compare the background and on-site data sets. At the 95% confidence level, the calculated $p = 0.172$. Based on this evidence, a statistical difference between background and on-site data could not be demonstrated at the 95% level of confidence; thus, no further action with respect to chromium was required. Note that, for this simple example, the conclusion of "no further action" is drawn because a statistical difference was not obtained. The power of the test is normally calculated when the

[*] Appendix C.

null hypothesis is not rejected. Additional investigation would be required if the power was not adequate.

3.11. Upper Tolerance Limits. *Upper tolerance limits*[*] (UTLs) are sometimes used to determine whether site concentrations are elevated relative to background concentrations. The UTL defines a threshold value for the background data set. (More accurately, it is an upper confidence limit for some percentile of the background data.) Individual site contaminant concentrations are compared to this value. Study area detections that are greater than the background UTL are considered to be indicative of contamination from site-related waste handling activities. Tolerance limits are used in this manner in the USEPA guidance for the statistical treatment of groundwater monitoring data (EPA 530-SW-89-026, EPA 9285.7-09A). However, this approach must be used with caution. In particular, it is often erroneously concluded that site-related contamination exists if a single detection exceeds the UTL. For example, the "95% UTL" is typically used to evaluate site contamination relative to background. If the background and site concentrations are not different from one another, we will be 95% confident that *at least* 95% of all site measurements will fall below the *95% UTL with coverage of 95*%. (For brevity, this is often referred to simply as the "95% UTL.") Therefore, we would expect a small percentage of site measurements to exceed the UTL, even when overall site contamination is not elevated relative to background. When a large number of samples are taken, we should not definitively conclude that a small number of detections greater than the UTL necessarily indicate site-related contamination.

3.11.1. Furthermore, regulators have criticized the use of UTLs to compare site to background contamination because UTLs do not minimize false negatives but, rather, minimize false positives. In other words, if many detected study area concentrations were greater than the background UTL, this would constitute strong evidence of site-related contamination. This scenario would be unlikely if the site and background concentrations were similar. Alternatively stated, the probability of a false positive—erroneously concluding that the site is contaminated relative to background—would be low. However, if detected site concentrations were less than the UTL, strictly speaking; *no conclusion would be possible*. This would not be sufficient to demonstrate the absence of site contamination relative to background. If we were to conclude the absence of site-related contamination using the UTL, false negatives could result (i.e., erroneously concluding that site concentrations are not elevated relative to background concentrations).

3.11.2. Because of the problems with tolerance intervals discussed above, two-sample statistical tests are usually preferred (and are typically more appropriate) to compare site and background data sets. It is recommended that UTLs be used only when two-sample tests are not practical (or when the primarily objectives is to demonstrate that site contamination is elevated relative to background contamination). For example, a two-sample statistical test cannot be performed when the site data set is extremely small (when only one or two samples are available for the study area). If a large data set was available for the background study area (e.g., because a

[*] Appendices G and K.

"site wide" background study had been done for a prior investigation), then the study area results could be compared to the background UTL.

3.11.3. The UTL background comparison methods are discussed Appendix K. These methods are summarized in Table 3-2

3.11.4. There are parametric UTLs and non-parametric UTLs. The parametric UTL require the data to follow a specified distribution such as a *normal* or *lognormal* distribution. (Distribution tests are addressed in Appendices F and J.) As shown in the table above, the proportion of non-detects must be taken into account when selecting an appropriate UTL. (UTLs that rely upon the normality assumption cannot be calculated when a large portion of the data are reported as non-detect.) The nonparametric UTL represents a high-end value in the distribution. The following case study illustrates an example of calculating background UTLs for metals.

Table 3-2.
Background Comparison to Evaluate the Extent of Contamination

Percent Detections in Background Data	Type of UTL Calculated
0	No UTL calculated
< 10	Poisson UTL
10–85	Nonparametric UTL
≥ 85 (normal or lognormal distribution)	Parametric UTL

3.12. Case Study 5—Calculating Background UTLs for Metals. At a site in Utah, 56 soil samples were collected across a very large area to determine background concentrations for metals.

3.12.1. Chromium was detected above the detection limit in every sample, so there was no need to substitute for censored values. Manganese was not detected in one sample, and the geochemist elected to substitute one-half the detection limit for the censored value in that sample.

3.12.2. The chromium data were normally distributed and the manganese data were *lognormally distributed.*[*] Refer to Appendices D, E, and I for a review of these concepts.

3.12.3. For chromium, the 95% UTL was calculated from the sample results using the formula:

$$95\%\ \text{UTL} = \bar{x} + ks \ .$$

3.12.4. For 56 samples, k equals 2.032. Chromium results for background had a mean ($\bar{x}$) of 12.7 mg/kg and standard deviation of 5.1 mg/kg, so the UTL was 23.0 mg/kg. For manganese,

[*] The *Shapiro-Wilk test* (Paragraph F-3) was used to test for normality at the 95% level of confidence.

the log of each sample result was taken prior to the calculation of the UTL. (The individual concentrations are not shown.) For the set of log-transformed results, the sample mean and standard deviation were 5.41 and 0.75, respectively. The log UTL for manganese was 6.93 (using the above equation). All comparisons for manganese should occur in "log space" (that is the logarithm of the site manganese maximum would be compared to 6.93). (Alternatively, a *minimum variance unbiased estimator* of the manganese background concentration could be calculated using the methods described in Appendix E).

3.13. Extended Background Example. This paragraph illustrates the concepts of distributional assumptions presented in Appendix J through a case study.

3.13.1. Suppose surface soil samples (from 0 to 5 feet below ground surface) have been collected at Site A and a background location to evaluate chromium concentrations on site. Table 3-3 presents the analytical results from samples collected at the site and background areas. All chromium concentrations were detected so no proxy concentrations are needed to evaluate the data.

3.13.2. Further, suppose the objectives of this data evaluation are to identify whether chromium surface soil concentrations on site:

3.13.2.1. Exceed regulatory threshold levels.

3.13.2.2. Exceed background concentrations, on the average.

3.13.3. Several statistical tests can be used to make such comparisons. A "one-sample" test can be used to compare the mean site chromium concentration to regulatory risk-based levels (Appendix L). A "two-sample" test can be used to compare the mean concentration of chromium at the site to the mean background concentration of chromium (Appendix M). A background value, such as a UTL, can be estimated for comparisons to individual site concentrations to identify if any one sample has a concentration higher than background. However, before any statistical tests can be done, distributional assumptions must be evaluated for each population (site and background) of data to determine which statistical test is most appropriate. The distributions are evaluated for normality (or log normality) using statistical tests and graphical plots.

3.13.4. Graphical displays are the first approach taken to evaluate the distribution of the data (Appendix J). Histograms, box-and-whiskers plots, and probability plots are all useful in identifying how data are distributed and answering questions such as—are the data symmetrical, what is the range of concentrations, are there any outliers that may unduly influence future distributional tests, do the data seem to follow a normal distribution, and so on. Histograms, box-and-whisker plots, and probability plots for the site and background data are provided in Figures 3-3 and 3-4, respectively.

Table 3-3.
Analytical Results for Chromium at Site A and Background Locations

Site A Sample Location	Top Depth of Sample	Bottom Depth of Sample	Chromium Concentration (mg/kg)	Background Sample Location	Top Depth of Sample	Bottom Depth of Sample	Chromium Concentration (mg/kg)
SB01	1	2	4.76	BG01	1	2	4.99
SB01	4	5	4.42	BG01	4	5	4.35
SB02	1	2	4.68	BG02	1	2	4.61
SB02	4	5	4.82	BG02	4	5	4.83
SB03	1	2	4.36	BG03	1	2	3.92
SB03	4	5	4.37	BG03	4	5	5.09
SB04	1	2	4.09	BG04	1	2	5.19
SB04	4	5	4.14	BG04	4	5	4.54
SB05	1	2	4.78	BG05	1	2	5.49
SB05	4	5	4.94	BG05	4	5	4.3
SB06	1	2	3.35	BG06	1	2	5.67
SB06	4	5	3.08	BG06	4	5	4.16
SB07	1	2	10.1	BG07	0.5	1	5.41
SB07	4	5	18.5	BG07	2	2.5	4.98
SB08	1	2	10.6	BG08	1	2	5.64
SB08	4	5	4.87	BG08	4	5	4.98
SB09	1	2	10.3				
SB09	4	5	5.51				
SB10	1	2	6.4				
SB10	4	5	4.13				
SB11	1	2	4.96				
SB11	4	5	4.96				
SB12	1	2	4.91				
SB12	4	5	4.89				

3.13.5. These plots have been developed on the basis of the original data and the natural-log transformed data, as it is common that environmental data follow either a normal or log-normal distribution. Other less common transformations, such as the square root or inverse sine transformation, are *not* applicable in this case study because:

3.13.5.1. Chromium concentrations are continuous (values can be any number within a range of concentrations).

3.13.5.2. Detected chromium concentrations are not rare events to warrant review of the Poisson distribution.

3.13.5.3. Chromium concentrations are not binomially distributed.

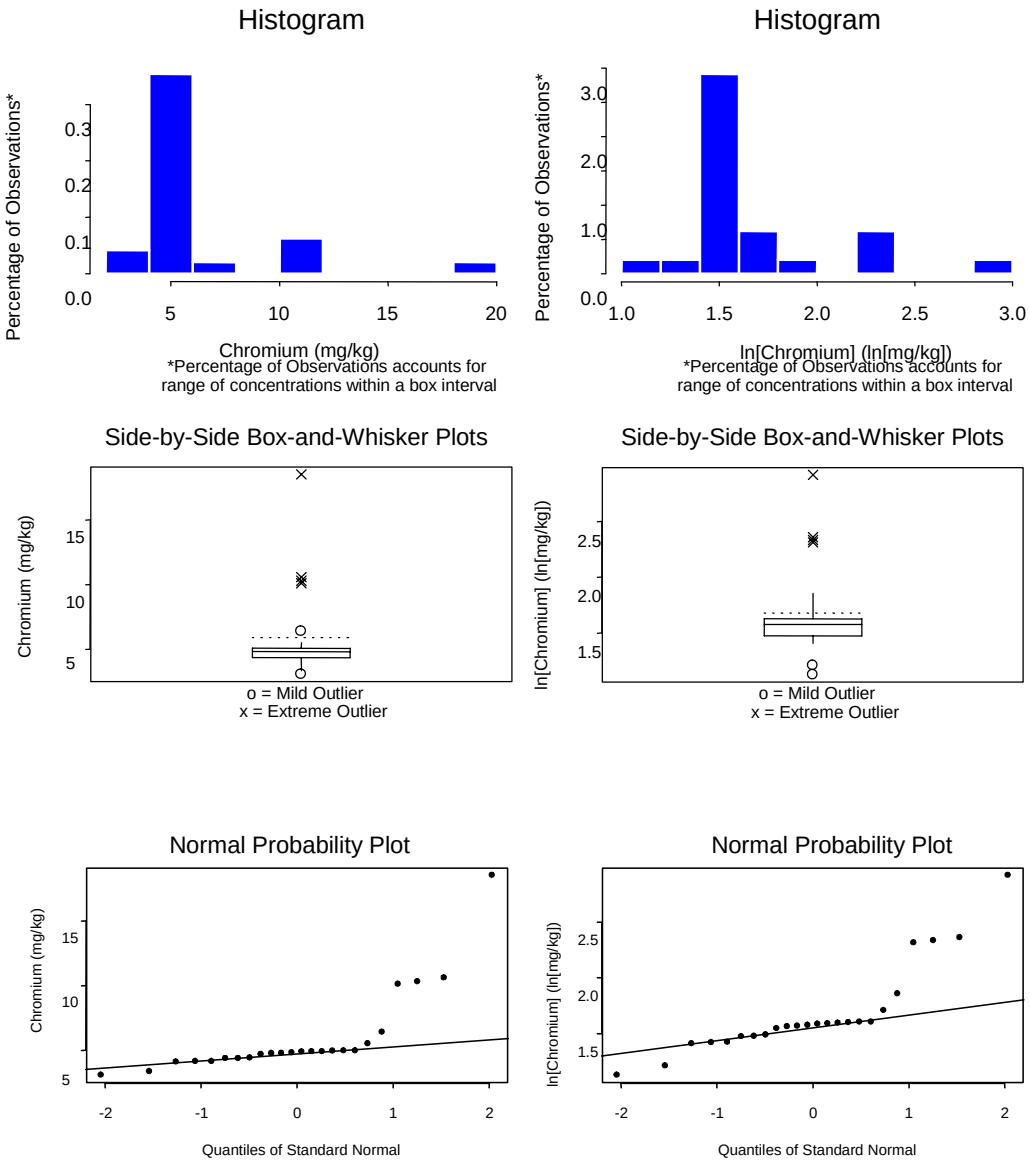


Figure 3-3. Chromium in Site A.

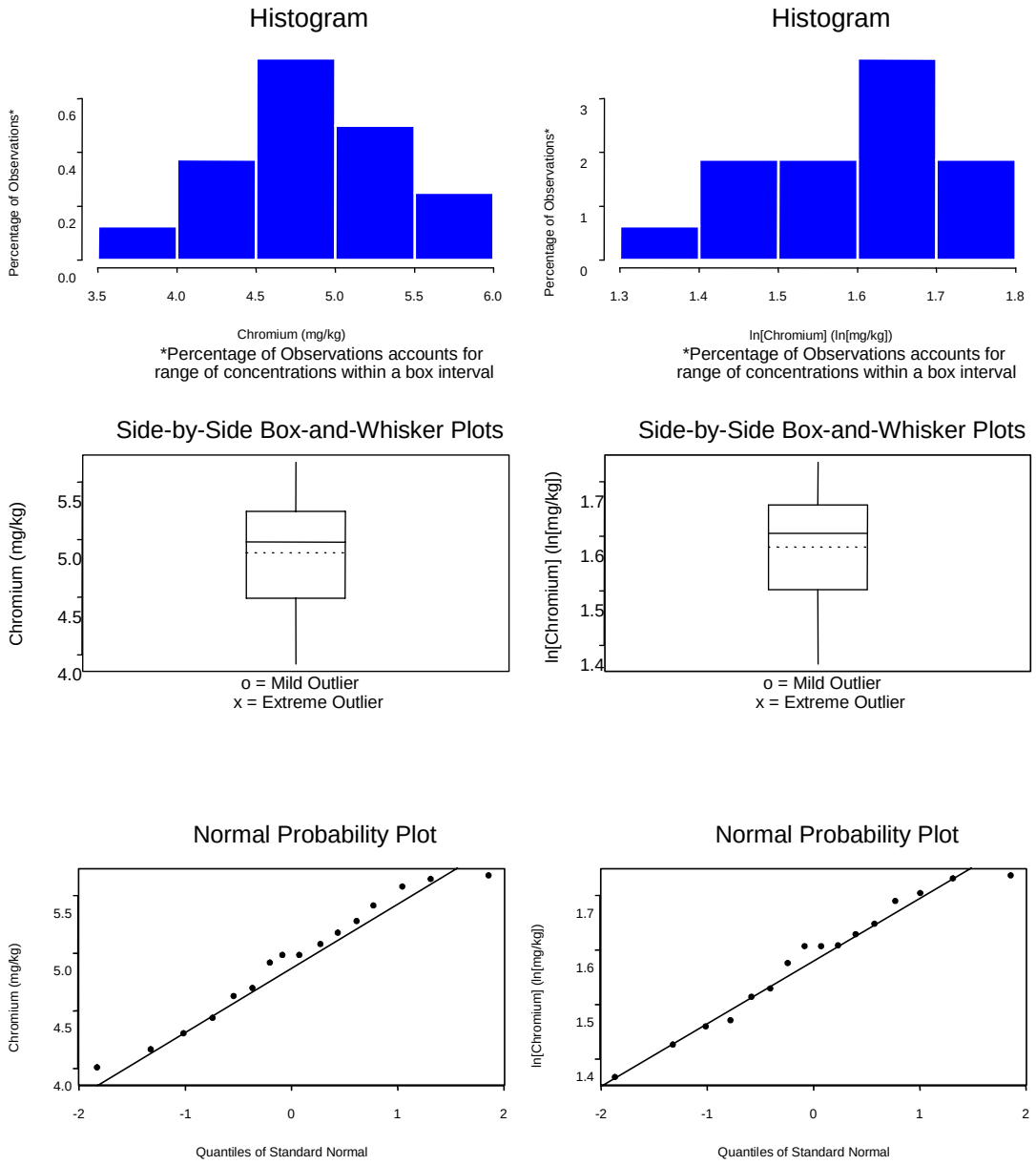


Figure 3-4. Chromium in background.

3.13.6. Based on just the plots in Figure 3-3, chromium at Site A does not appear to have a normal or lognormal distribution. The histograms for the original data and log-transformed data are not symmetrical, but are skewed. This is confirmed in the box-and-whiskers plots because the mean (the dotted line) is larger than the median (the solid line within the box) and the mean is even larger than the 75th percentile (the top part of the box). (If the data were normal, the mean

would be equal to the median.) As the mean is greater than the 75th percentile, this suggests that the mean is influenced by several considerably large concentrations. Outliers (each of point represented by an "X") predominantly occur only in the upper portion (the top) of the box plots. Lastly, as the normal probability plots for the original data and log-transformed data are not linear, this gives additional evidence that the data are not normal or lognormal.

3.13.7. The chromium data distributions possess heavier right tails relative to a normal distribution. Note the extreme deviation from linearity (Appendix F) at the right-hand side of each normal probability plot (appearing as a series of points above the straight line). The superimposed line on the normal probability plots illustrates the line that concentrations follow when data are normally or lognormally distributed. This line is related to Filliben's statistic in the sense that it provides a standard to compare the linearity of sample results. For these normal probability plots associated with Site A, it is apparent that the data do not follow a normal or lognormal distribution.

3.13.8. The plots in Figure 3-4 show evidence that chromium for the background data set appears to follow a normal or a lognormal distribution. The histogram for the original data seems to be symmetrical, though the histogram for the log-transformed data is not as symmetrical. However, histograms can be misleading if the boxes (i.e., concentration intervals) are too large or too small; therefore, another type of plot, preferably a normal probability plot, should be constructed to determine whether the data are normally (or lognormally) distributed.

3.13.9. One of the most powerful statistical methods for testing normality is the *Shapiro-Wilk*[*] test. Because the site data set has 24 sample results and the background data set has 16 sample results, this test would be appropriate for evaluating normality and lognormality for both the site and background data sets. The result of the Shapiro-Wilk test is presented in Table 3-4 for chromium at Site A and background based on the original data and log-transformed data. The Shapiro-Wilk test results in either a calculated value of the statistic W or the value p. There is acceptably strong evidence that the data set is not normal when either W or p is small relative to the corresponding acceptance limit for W or p.

3.13.10. For Site A, results of the Shapiro-Wilk test show evidence that the data *do not* follow a normal or lognormal distribution (i.e., since the calculated value of W is smaller than $W_{0.01}$, or equivalently, $p < 0.01$, there is less than a 1% chance that the data set is normal, or equivalently stated, there is at least a 99% confidence that the data are not normal). However, for background the results of the Shapiro-Wilk test suggest that the data seem to follow both a normal and lognormal distribution. It should be noted that there is more evidence that background data are normally distributed rather than lognormally distributed, because the value of W and the associated value of p are higher for the original data than for the log-transformed data.

[*] Appendix F.

3.13.11. The *coefficient of variation*[*] (*CV*) was estimated for each data set, and is provided in Table 3-4. A *CV* greater than 1 suggests a departure from normality. However, the evaluation of the *CV* is not as reliable as quantitative statistical tests for normality, such as the Shapiro-Wilk test. The coefficient of variation is useful only for identifying obvious departures from normality when *CV* is much greater than 1. Because the sample *CV*s for the site and background data sets based on the original data and the log-transformed data all are less than 1 (as discussed in Appendix F), one cannot conclude the data can be modeled by a normal distribution. Therefore, for these data sets, the *CV* does not provide any useful additional information.

3.13.12. Similarly, to illustrate the relative reliability of various distributional test methods, the *Studentized range test** was also performed on the data sets. The results of this test (Table 3-5) indicate that the Site A and background data sets follow normal and lognormal distributions. The range test failed to identify the lack of normality for Site A data. This happened because the data distribution for Site A is asymmetrical and this test does not perform well for asymmetrical distributions. However, according to the test, the background data follow a normal and lognormal distribution. Therefore, the Studentized range test for the background data set is consistent with the Shapiro-Wilk test, the coefficient of variation test, and the graphical plots (e.g., the normal probability and box plots).

3.13.13. Similarly, to illustrate the relative reliability of various distributional test methods, the *Studentized range test** was also performed on the data sets. The results of this test (Table 3-5) indicate that the Site A and background data sets follow normal and lognormal distributions. The range test failed to identify the lack of normality for Site A data. This occurred because the data distribution for Site A is asymmetrical and this test does not perform well for asymmetric distributions. However, according to the test, the background data follow a normal and lognormal distribution. Therefore, the Studentized range test for the background data set is consistent with the Shapiro-Wilk test, the coefficient of variation test, and the graphical plots (e.g., the normal probability and box plots).

3.13.14. To summarize, the background data appear to follow both a normal and lognormal distribution, but Site A data do not appear to follow either a normal or lognormal distribution. A dilemma exists regarding the distribution of the background data—is it normal or lognormal? As the log transformation did not appreciably improve the normality of the data set, it would be advisable not to perform the transformation.

[*] Appendix F.

Table 3-4.
Results of the Shapiro-Wilk Test of Normality and Lognormality for Chromium Surface Soil at Site A and Background

Area	Testing for Normality or Lognormality?	Number of Results	*CV*	Shapiro-Wilk Test Statistic, *W*	Critical Values (from Table B-20 of Appendix B)			*p* value for Shapiro-Wilk Test (from statistical software)	Conclusion: Is there evidence that the data are Normally or Lognormally Distributed? Yes/No
					Critical Value based on 0.05 level of significance, $W_{0.05}$	Critical Value based on 0.10 level of significance, $W_{0.10}$	Critical Value based on 0.50 level of significance, $W_{0.50}$		
Site A	Normality	24	0.5687	0.627	0.916	0.930	0.963	<0.0001	No
Site A	Lognormality	24	0.2426	0.791	0.916	0.930	0.963	0.0002	No
Background	Normality	16	0.1093	0.963	0.887	0.906	0.952	0.7177	Yes
Background	Lognormality	16	0.07041	0.958	0.887	0.906	0.952	0.6308	Yes

Table 3-5.
Results of the Studentized Range Test of Normality and Lognormality for Chromium Surface Soil at Site A and Background

Area	Number of Results	Test of Normality (based on original data)			Test of Lognormality (based on log-transformed data)		
		Ratio of Range of Results and Standard Deviation	Critical Values from Table B-21 of Appendix B, assuming a 0.05 level of significance	Conclusion: Is there evidence that the data are Normally Distributed? Yes/No	Ratio of Range of Results and Standard Deviation	Critical Values from Table B-21 of Appendix B, assuming a 0.05 level of significance	Conclusion: Is there evidence that the data are Lognormally Distributed? Yes/No
Site A	24	4.586	(3.308, 4.666)*	Yes	4.400	(3.308, 4.666)*	Yes
Background	16	3.278	(3.01, 4.24)	Yes	3.317	(3.01, 4.24)	Yes

*Critical Values for n = 24 are based linear interpolation of critical values from n = 20 and n = 25.

3.13.15. If a background value, such as a UTL, and other summary statistics are desired to characterize the background data set, then the assumed distribution should fit the data as much as possible. With respect to this objective, it would be more appropriate to define background as following a normal distribution because the Shapiro-Wilk test shows more evidence of normality than lognormality. Comparing the Shapiro-Wilk test's critical value or associated *p* value from the original data and from the log-transformed data is a reasonable approach for discerning which distribution is more appropriate and has more evidence of following a normal or log-normal distribution.

3.13.16. The first objective for this case study is to determine whether chromium contamination at Site A, on the average, exceeds a regulatory threshold value. As it cannot be assumed that the Site A data set is either normal or lognormal, a nonparametric test (e.g., the Wilcoxon signed rank test for the median as discussed in Appendices H and M) must be used to compare the Site A data to the regulatory threshold.

3.13.17. The second objective is to determine whether chromium exceeds background. Though the background data set could be reasonably assumed to be either normal or lognormal, this assumption could not be made for the Site A data set. As the Site A data set is neither normal nor lognormal, a *parametric two-sample test*[*] cannot be used to compare the Site A data set to the background data set (for example, to determine if the mean concentration at Site A exceeds the mean background concentration). Both data sets must follow the same distribution to use a parametric test. For example, both the background and site data sets must both be normally or lognormally distributed. As data from Site A does not follow a normal or lognormal distribution, only nonparametric tests such as the *Wilcoxon rank-sum test** can be used to compare the Site A and background data sets.

3.13.18. This case study illustrates the value of background data in project decision-making. The application of background data in identifying contaminants for inclusion in the risk assessment is presented in the following section. The data in the preceding discussion may be used as sample data to apply some of the nonparametric tests in Appendix M.

Section III
Risk Assessment

3.14. Introduction. Perhaps more than any other area in the CERCLA project life cycle, assessing site risk relies on statistics. Many of the techniques described in several of the appendices apply in quantifying and assessing risk at a hazardous waste site. The components of a risk assessment discussed in this report are:

[*] Appendices M and N.

- Identifying contaminants of potential concern (COPCs).
- Calculating exposure point concentrations (EPCs).

Statistics enter into risk assessment in one additional major area—the calculation of exposure levels. Specifically, a baseline human health risk assessment requires estimation of a reasonable maximum exposure (RME), and a central tendency exposure (CTE). The former relies on 95% *upper confidence level (UCL)* values for exposure parameters, and the latter on the mean of the exposure parameters. In either case, the exposure parameters are generally provided by EPA guidance, such as the *Exposure Factors Handbook* (USEPA, 1997). For all practical purposes, the environmental scientist will not need to statistically evaluate these parameters and, consequently, their derivation is not discussed here. However, understanding the concepts presented in Appendix E is very useful in deconstructing the data evaluations presented in the *Exposure Factors Handbook* (USEPA, 1997).

3.14.1. *Identification of Contaminants of Potential Concern for Risk Assessment.* Not all chemicals detected at a site are typically included in the quantification of risk. Those chemicals retained in the risk assessment are the COPCs. Note that the COPCs are media-specific; COPCs are evaluated for air, surface soil, subsurface soil, groundwater, sediment, surface water, and any other medium sampled in the RI at each site.

3.14.1.1. Chemicals are typically screened against background or other criteria (established by ARARs) and a subset is selected for inclusion in the risk calculations. Some of the screening criteria, other than background levels, include drinking water MCLs, or secondary MCLs, RBCs, and *Toxic Substance Control Act* (TSCA) values for PCBs (polychlorinated biphenyls) in soil. In addition, inorganics that are essential human nutrients (e.g., iron, potassium, magnesium, sodium, and calcium) may be excluded from the quantitative risk analysis in most cases. (ARARs are identified in the planning stage of the RI.)

3.14.1.2. Both qualitative and quantitative statistical evaluations are frequently performed to identify COPCs. A qualitative evaluation is initially conducted to determine whether select potential analytes of concern can be eliminated from future investigation; a statistical evaluation is subsequently done for a more in-depth look at of contaminants that were not eliminated during the qualitative assessment.

3.14.1.3. For example, for the qualitative evaluation of the data, if a chemical is detected infrequently in the sample data set, and is not considered to be associated with historical waste handling at a site, it may be screened out as a COPC. However, it is essential to use site-specific information before discarding such a chemical, as infrequently detected compounds may also represent hot-spots, depending on the sampling strategy used at the site. For every chemical detected at least once, the maximum detected concentration is compared to the chemical- and

medium-specific screening criterion. Chemicals with higher concentrations than their criteria are generally retained for quantitative evaluation in the risk assessment.

3.14.1.4. Contaminants that lack ARARs (usually because toxicity information does not exist) are retained as COPCs in the risk assessment and discussed in the uncertainty section of the report. One-sample tests for contaminants where the maximum exceeds the risk-based screening limit may be used to determine whether the mean is statistically less than the screening limit, even though a single value exceeds the screening limit. Anthropogenically derived contaminants (such as PAHs) that occur at concentrations below background levels are still retained in the risk assessment if they exceed ARARs. If the risk assessment indicates that such contaminants are a primary contributor to total risk at a site, then a quantitative statistical comparison with background (e.g., using appropriate two-sample statistical tests) would be done and the results would subsequently be discussed in the risk characterization at the end of the assessment.

3.14.2. *Calculating Exposure Point Concentrations*. For risk assessment, means and standard deviations are typically calculated as the basis for EPCs and as the basis for deriving UTLs for the background comparisons. However, the mean and standard deviation will frequently be inappropriate measures of central tendency and dispersion when the data are not normally distributed or a large portion of the data consists of non-detects. Under these circumstances, means and standard deviations should not be used to perform statistical evaluations. Before statistically valid means and standard deviations can be calculated, tests for normality should be conducted and non-detects must be appropriately addressed.

3.14.2.1. The EPC is used to calculate a COPC's carcinogenic risk and non-carcinogenic hazard index. It represents the concentration a receptor is likely to encounter. The USEPA requires the EPC to be a conservative estimator of central tendency—the 95% upper confidence limit (UCL) of the sample arithmetic mean concentration (OSWER 92-856-03, EPA 68-W0-0025). The 95% UCL is the concentration that, when calculated repeatedly for randomly drawn samples, equals or exceeds the true mean 95% of the time.

3.14.2.2. Calculating rigorous, statistically valid 95% UCLs requires that data be distribution tested and that non-detects be treated properly. Procedures for this are provided in Appendix H. Some of the older (pre-2000) RCRA and CERCLA guidance for calculating the UCL are outdated (and hence, are not recommended); modifications and updates are provided with the goal of improving scientific defensibility. Appendix G presents the most recent acceptable methods for estimating the 95% UCL at 95% confidence.

3.14.2.3. Calculating EPCs at a CERCLA site brings together many of the statistical procedures described in the attached Appendices. The correct steps are, in general, as follows

3.14.2.3.1. Identify the nature of the censoring limit and the proportion of censored values and substitute proxy values as directed in Appendix R.

3.14.2.3.2. Identify outliers as discussed in Appendices I and M.

3.14.2.3.3. Perform distribution testing as detailed in Appendix F.

3.14.2.3.4. Depending on the outcome of these steps, calculate the 95% UCL as directed in Appendix G.

3.14.2.4. Unfortunately, there are many pitfalls along the way, and this process does not always lead to a simple result. In part, this is attributable to the use of or adherence to older USEPA guidance. In particular, USEPA guidance for substituting for censored data is addressed in many separate risk assessment documents. In earlier documents, substituting one-half the detection limit is supported. Appendix E provides insight on the deficiency in this approach. In addition, even if the risk assessor has performed all of the statistical procedures, USEPA guidance for EPCs states that if a 95% UCL exceeds the maximum value of a compound detected at a site, the maximum should be substituted. This has the dissatisfying attribute of being completely *ad hoc*, giving rise to unquantifiable and unacceptable uncertainties for risk assessment decisions.

3.14.3. *Uncertainty Quantification*. A required element in a baseline human health risk assessment is to evaluate uncertainty for decisions. Statistical techniques alone will be unable to account for all sources of uncertainty in a risk assessment and a qualitative approach is normally taken. For example, there will be uncertainty in the risk assessment for analytes for which toxicity data do not exist, and the quantification of such uncertainty is not possible.

3.14.3.1. In risk assessment, uncertainty stems primarily from the following three sources.

3.14.3.1.1. Errors in the estimate of contaminant concentration.

3.14.3.1.2. Errors in the estimate of toxicity.

3.14.3.1.3. Errors introduced by large numbers of assumed values in the risk assessment formulations, which are by definition and intent very conservative.

3.14.3.2. In practical terms, there is little that can be done about the uncertainty in estimates of toxicity. The studies upon which toxicity data are based are taken “as is” simply because of the scarcity of available studies. Uncertainty in the assumptions employed in the risk assessment can sometimes be addressed, but only to a limited extent. An example for how the uncertainties listed in subparagraph 3.14.3.1.3 were taken into account is presented in Case Study 6.

3.14.3.3. Most statistical evaluations implicitly assume the absence of bias. The uncertainty predominantly depends on the distribution of field measurements. Even in the case of risk screening, as demonstrated in Chapter 2, we have seen that it is possible to qualitatively assess

the uncertainty of individual sample/analytical results before comparing those results to fixed threshold values using analytical QC information. For example, QC data can potentially be used to identify the direction of bias and to estimate the magnitude of the bias associated with a set of analytical results. This is illustrated in Case Study 6. It is also possible to make similar estimates of variability which may affect decision-making, as illustrated in Case Study 7.

3.14.3.4. The error introduced into the risk assessment by the uncertainty associated with each of the various assumptions and reference values is more likely multiplicative rather than additive, such that the calculated risk is conservative to an extraordinary degree. Consider, for instance, some components of a soil dermal absorption scenario. The risk assessor calculates an EPC, which represents the 95% UCL of the mean. Then, the skin area exposed to the contaminant is based on an upper 95% confidence level of all the U.S. adult population from EPA OSWER 92-856-03. These are combined with, say, the default average exposure duration and frequency values which, again, are upper estimates from some population. Combining all of these upper estimates results in a risk evaluation that has a far higher confidence than 95%. The Risk Assessor and Project Manager are encouraged to identify every opportunity to use site-specific values in place of assumptions in risk assessment to reduce uncertainty in the results and, thus, more appropriately apply the limited remediation resources available.

3.14.3.5. One method for estimating the true mean and distribution of risk estimates is to use the recommended RME and CTE values of exposure parameters. This methodology is recommended in *Risk Assessment Guidance for Superfund* (RAGS). The result of looking at each input parameter using the CTE is to provide an estimate of risk near the mean of the estimated exposure scenario. The RME is considered to represent an upper estimate of site risk. An alternative method of quantifying the range in risk estimates is to use Monte Carlo simulations.

3.15. Case Study 6—Refining Risk Assessment Assumptions.

3.15.1. A risk assessment was to be done as part of a RCRA Facility Investigation (RFI) at a steel mill in Pennsylvania. The project team approached the EPA Remedial Project Manager (RPM) regarding using site-specific assumptions for some of the exposure factors in the risk assessment calculations. This was possible because the facility maintained excellent records of employee longevity, promotion, and work assignments. For this case study, the focus is on site-specific estimates of exposure duration, which enters into quantification of risk.

3.15.2. Under the assumptions given by the EPA for the worker exposure scenario in OSWER 92-856-03, the risk assessor is to assume that a given worker will be exposed for a period of 25 years. However, by reference to detailed employee records for the facility, the project team was able to demonstrate concretely on a facility-specific, job-specific, and location-specific basis, the actual average lifetime exposure duration for the various site areas under study. Employing these actual values, which were approximately 3 to 5 years rather than 25 years, greatly reduced the exposure duration. More importantly, the site-specific value reduced the uncertainty

in the calculated lifetime risk. Using this lower value allowed the steel mill owner to limit the number of site areas proceeding to the Corrective Measures Study phase of the project.

3.16. Case Study 7—Direction and Magnitude of Bias. As part of a property transfer in Baltimore, Maryland, the project team was asked to estimate reserves that the seller would have to put in escrow against the potential need for site clean-up, before the seller would accept transfer of the property. For this case study, petroleum hydrocarbon contamination will be discussed.

3.16.1. The project team decided to divide the relatively small site into four quadrants and collect one composite sample from each to assess the potential need for remediation in each quadrant. The analytical results obtained from the laboratory were as follows:

Quadrant 1	1200 mg/kg	Quadrant 3	756 mg/kg
Quadrant 2	101 mg/kg	Quadrant 4	138 mg/kg

3.16.2. With the state's action level set at 100 mg/kg, it appeared that the seller would be required to reserve funds against a potential soil removal for the entire site. However, a review of the quality control data associated with the analytical results displayed significant potential bias.

3.16.3. A normal calibration curve was developed for the gas chromatograph used in the analysis that met method criteria for linearity. The laboratory then analyzed an Initial Calibration Verification (ICV) using a standard from an alternative source from that employed in the calibration. The ICV was essentially a blank spike set at the midpoint of the calibration curve. The result of this analysis was a percent recovery (%R) of 168%, which was within the acceptance limits provided with the standard by the manufacturer.

3.16.4. However, in its simplest form this QC result indicates that if the laboratory introduced the equivalent of 100 mg/kg of total petroleum hydrocarbons (TPH) into the analytical system, they would get a reported result of 168 mg/kg. This observation, applied to the results reported for the site, removed two of the four quadrants from further consideration, reducing the required reserves by half.

Section IV
Probabilistic Risk Assessments Monte Carlo Simulations

3.17. Introduction. The implementation of probabilistic risk assessment for environmental projects is beyond the scope of this document; however, a brief overview of the procedures is presented here. Monte Carlo simulation, the most common technique used for probabilistic assessments, is a statistical technique in which outcomes are produced using randomly selected values for input variables that possess a range of possible values. In some cases, a known probability distribution can be assigned to each input variable. By repeating the calculation many, many times, Monte Carlo simulations create a population of results representing (in theory) the

full range of possible outcomes and the likelihood of each. For example, when Monte Carlo simulation is used in risk assessment, risk is expressed as a distribution of possible values rather than a single point value.

3.17.1. There are two major practical limitations to the application of Monte Carlo simulations in general: i) it can be costly, and ii) few people are sufficiently qualified to do it. The EPA has also written a guidance document for probabilistic risk assessment titled *RAGS Volume 3 Part A: Process for Conducting Probabilistic Risk Assessment* (EPA 540-R-02-002) available at http://www.epa.gov/oswer/riskassessment/rags3a/index.htm. An EPA Region 3 publication (EPA 903-F-94-001) identified several technical limitations that preclude the Agency from relying on Monte Carlo simulations (http://www.epa.gov/reg3hwmd/risk/human/info/guide1.htm).

3.17.1.1. Software is unable to distinguish between measurement variability and lack of knowledge. Some input parameters are for well-described differences among individuals—these differences are variability. Other factors, such as frequency and duration of trespassing, are simply unknown, and assuming a distribution for them is ad hoc. But the simulated distribution of unknowns is presented in computer output as variability. The accuracy of the distributional assumptions limits the accuracy of the simulation.

3.17.1.2. Software is unable to account for sample dependency (e.g., spatial and temporal correlations for sample locations). However, this limitation also applies to all classical statistical methods (e.g., the methods predominantly discussed in this document and in EPA environmental statistical documents such as the QA-G4 and GA-G9 guidance documents). In classical statistics, the assumption of independence highly influences the applicability of a technique—the same limitation applies here.

3.17.2. In most statistical evaluations (excluding geostatistics), environmental scientists are resigned to the limitations of classical statistics for environmental data. The same is true for Monte Carlo simulations. Though Monte Carlo simulations require sample independence, the approach can be advantageous. The primary advantage is that it accounts for a range of input values and outputs a range of outcomes (such as risk values) with associated probabilities. Although a Monte Carlo approach is currently not recommended or required by the EPA, the approach may be beneficial for some projects. There are applications of such simulations. Moreover, future scientists may learn how to overcome some of the limitations and eventually develop reasonable and inexpensive computer applications.

3.17.3. Applications of Monte Carlo simulation are more prevalent in groundwater modeling than any other current environmental application. Case Study 8 shows how a Monte Carlo simulation of groundwater contamination was used to perfect a remedy.

3.18. Case Study 8—Monte Carlo Simulation in Remedial Alternative Selection

3.18.1. Monte Carlo analysis was coupled with decision tree analysis for a study site in Nebraska where the groundwater was contaminated with trinitrotoluene (TNT). The extent of TNT contamination was characterized during an RI. Three pump-and-treat alternative remedial actions were developed for the FS. The maximum concentration of TNT remaining in the saturated zone at the end of each alternative project lifetime was determined stochastically using a Monte Carlo model. The Monte Carlo model randomly generated values for site information for initial mass concentration, hydraulic conductivity, and retardation coefficient. Then these randomly generated fields were sampled and the output was combined into sets or ensembles. Probability functions were fitted to the output ensembles with the maximum simulated TNT concentrations. Because each of the treatment alternatives was associated with a different set of possible maximum concentrations, the Monte Carlo simulation made it possible to identify the optimal alternative quantitatively by analyzing the output ensembles for each alternative.

3.18.2. Applying Monte Carlo simulations requires the technical support of a specialist in this area; detailed methodologies are beyond the scope of this Manual. The technique does rely on the power of randomly generated data sets and the optimization of conditions based on the simulation.

CHAPTER 4
Remedial Design and Remedial Action

4.1. Introduction. During the RD/RA phase, engineers develop detailed designs for remedial actions, construct remediation systems, and operate and monitor sites with long-term remedies in place. The term remedial system is defined here in a broad sense; it includes removal actions and capping as well as more active treatment systems.

4.1.1. A number of statistical approaches that are applicable for prior stages of a project's life cycle are also applicable for the RD/RA. This Chapter will address environmental statistical applications for the RD/RA that have not been highlighted for the PA, SI, or RI/FS. In this Chapter, we consider adaptive sampling plans for removal actions and groundwater monitoring and trend analysis.

4.1.2. Although groundwater is most commonly subject to long-term monitoring, the same tools can be used to monitor and optimize remedial systems for other environmental media or demonstrate achievement of site closure criteria.

4.2. Comparisons to ACLs and MCLs. Confirmation sampling is often performed for the RD/RA and would typically entail one-sample statistical tests. These would be the same types of tests that would be conducted during the SI and RI, only the nature of the decision limits would differ (e.g., the decision limits for the RD/RA would be "cleanup goals" rather than the risk-based screening concentrations as in the SI).

4.2.1. As an example, consider data collected at a landfill. If a statistically significant difference is observed between upgradient and downgradient concentrations, a compliance monitoring program must be put into place. According to RCRA regulations, analysis of Appendix IX list constituents is required. Assuming that a release is confirmed, the facility must demonstrate that the release does not present a health or environmental risk. Generally, this entails comparing analytical results to fixed threshold values, called Alternate Concentration Limits (ACLs), which are often established in a jurisdiction-specific fashion. An alternative approach is to compare site data to MCLs. In the first case, tolerance or confidence intervals are recommended. In the second case, the tolerance limit is the preferred method.

4.2.2. An appropriate one-sample statistical test is to determine whether contamination exceeds the decision limit (e.g., an MCL). For example, if a set of measured contaminant concentrations is normal, a one-sample *t*-test could be used to compare the mean concentration to the decision limit. However, a reliable comparison using a one-sample test will not be possible if the data set is small (e.g., consists of only three points). If normality of the data set can be assumed, a conservative approach would consist of calculating an UTL and comparing it to the decision limit. If the UTL were less than the decision limit, there would be strong evidence that site contamination does not exceed the decision limit. *However, do not conclude that there is a contami-*

nation problem when the UTL exceeds the decision limit. To avoid false positives, when the UTL exceeds the decision limit, additional data should be collected to do an appropriate one-sample statistical test.

4.2.3. The confidence limit approach is used for comparisons to ACLs based on background data, whereas the tolerance limit approach is used when the comparison criteria are health-based and the comparisons are in relation to MCLs or health-based ACLs. The tolerance limit approach is more conservative than the confidence limit approach in that the UTL must be less than the MCL. However, Gibbons (1994) has pointed out the following.

4.2.4. Because at most four independent samples will be available during semiannual monitoring, the 95% confidence, 95% coverage tolerance limit is approximately five standard deviation units above the mean concentration. In light of this, even if all four semiannual measurements for a given compliance are well below the MCL, the tolerance limit will invariably exceed the MCL or health-based ACL and never-ending corrective action will be required.

4.2.5. Thus, special care must be taken in the design of compliance monitoring programs to ensure that the facility is not caught in the kind of regulatory trap described above.

4.2.6. In addition to one-sample statistical tests, multi-sample statistical tests can be appropriate for the RD/RA to perform comparisons with background values. Since long-term monitoring is commonly performed for groundwater during the RD/RA, Figures 4-1 through 4-5 summarize the types of one-sample and two-sample statistical tests that would be used for groundwater monitoring.

Section I
Groundwater Monitoring and Optimization Trend Analysis

4.3. Introduction. Monitoring remedial systems have significant, long-term costs. It is not difficult to anticipate that, over the course of 10 to 20 years, substantial economic resources available for environmental programs at military installations will be in long-term monitoring of sites actively under remediation or sites that require long-term monitoring. Project planners should ensure that these monitoring systems are optimized, and that they provide the necessary information at the least possible cost. Likewise, where active remediation is ongoing, optimization is important to minimize economic impacts to the facility. While optimization is desirable, compliance is mandatory, and at most installations, groundwater monitoring is required under various permits or consent agreements. This section reviews various methods of assessing groundwater systems over time with a view to both detection and compliance, and optimization.

4.4. Detection and Compliance Monitoring. Detection monitoring is a means of identifying whether a regulated hazardous waste site is releasing hazardous materials into the environment. Compliance monitoring entails the repetitive, periodic sampling and analysis of a select set of

monitoring locations for compliance with a fixed set of standards or requirements. The standards to which analytical results are compared are generally specified in regulations, permits, or consent agreements.

4.4.1. In detection monitoring, the results of sampling and analysis from a location that has recorded a release are compared to measurements from an unaffected or background location. In the case of groundwater monitoring, this generally entails selecting one or more monitoring wells upgradient of the site and selecting a representative set of downgradient monitoring wells. If the difference between the two sets of results is statistically significant, the owner is usually required to begin compliance monitoring to investigate how the release is occurring and to remedy the situation. These statistics fall into the category of hypothesis tests, specifically two- or multiple-population tests, and are addressed in Appendices M and N.

4.4.2. The selection of the statistical approach is generally open to discussion with regulators and the final determination will depend upon many factors. In general terms, the simplest approach (consistent with the requirements of local jurisdictions) is the best approach. For example, for detection monitoring, a two-sample *t*-test could potentially be used to compare upgradient (background) to downgradient (site) contaminant concentrations. Under the best of circumstances, a straightforward, parametric *t*-test would suffice; however, in practical terms, it is rare that environmental data meet all of the conditions that would make such a straightforward approach viable. And, in fact, by the time Figure 4-2 was published in EPA 530-SW-89-026, the use of the *t*-test had been largely discredited for this application because it failed to adequately control false positives when multiple site and background comparisons are required. Clearly, as of the time of its publication, the 1989 guidance recommended the use of ANOVA techniques (essentially a generalization of the two-sample *t*-test), and, to a lesser extent, alternatives such as tolerance intervals, prediction intervals, and control charting. By 1992, with the publication of *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities—Addendum to Interim Final Guidance* (EPA 68-W0-0025), a somewhat different statistical approach was highlighted. Preferences had shifted further with the use of intervals and resampling strategies receiving much greater attention. By 1994, when Gibbons published *Statistical Methods for Groundwater Monitoring*, ANOVA techniques had largely fallen out of use, replaced by prediction intervals with resampling strategies that have become, in some cases, very complex. This statistical approach currently represents what might be called the state-of-the-art for groundwater.

4.4.3. The alternative approach of using control charts has not gone altogether out of favor, however. A control chart is a type of plot (using data from a particular monitoring well) of some function of concentration (e.g., the mean concentration) versus time. The various statistical tests previously discussed are based on one of two possible approaches for detection monitoring. With the exception of the control chart approach, each new downgradient result is compared to the history (or historical data set) of upgradient results. These types of comparisons are called interwell (literally, "*between well*") comparisons. A potential flaw in this approach is that it as-

sumes the only variable that can make a difference between the upgradient and downgradient results is the intervening waste management unit. In reality, there are a number of other possible influences and, for this reason, intrawell (literally, "*within well*") comparisons are still considered quite useful in groundwater monitoring applications. The classic method of performing these *intrawell* comparisons is with control charting. The two types of control charts normally employed for these purposes are the Shewart and cumulative summation (CUSUM) control charts, which are often combined in normal use.

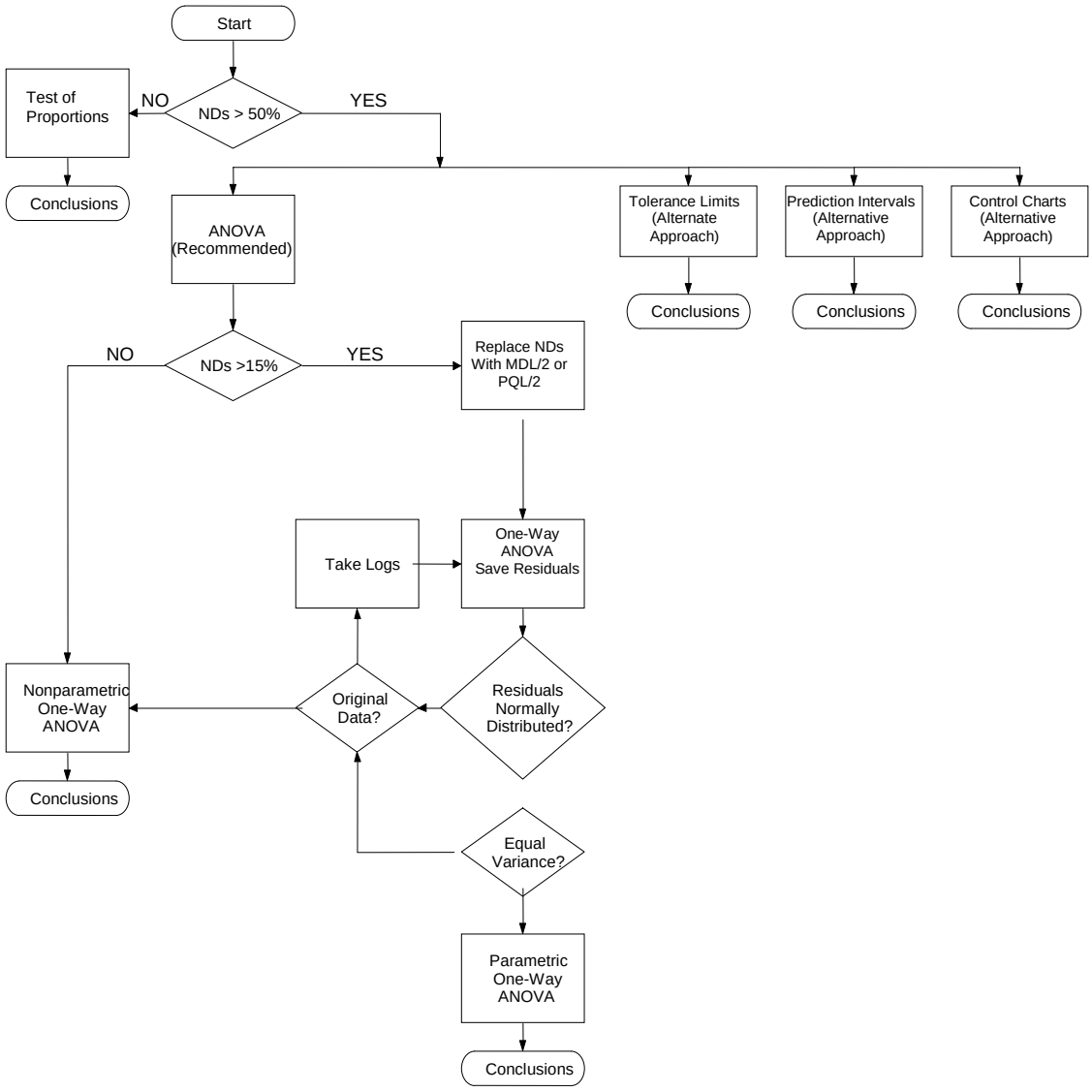


Figure 4-1. 1989 EPA decision tree for groundwater monitoring.

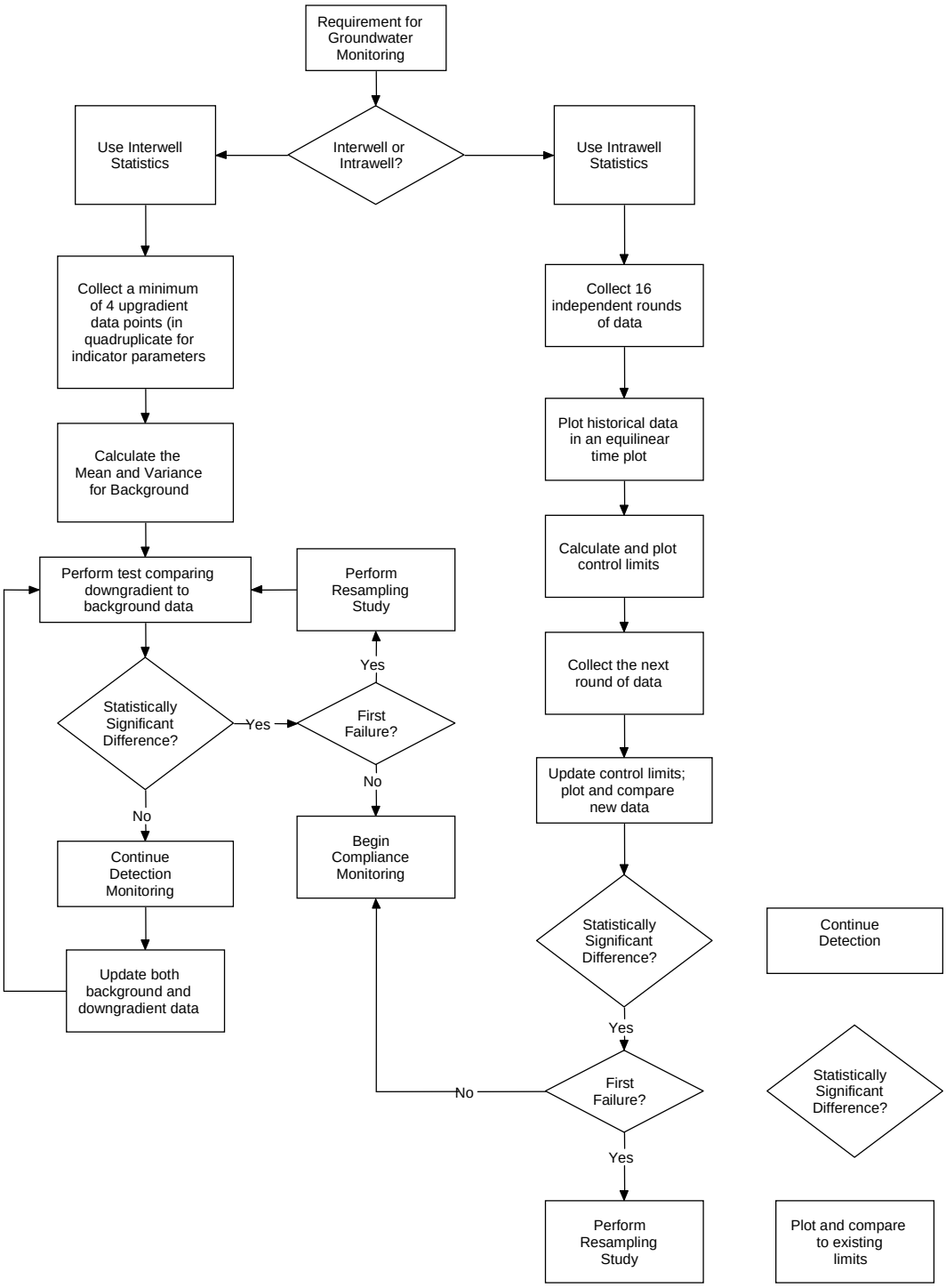


Figure 4-2. Statistical decision tree with options for groundwater monitoring—Part 1.

4.4.4. Figures 4-1 through 4-5 present flow charts showing the options available and guidance on option selection. However, the decision regarding the type of statistical analysis program to employ should be made as part of the DQO development process for the monitoring effort. It is strongly recommended that the Project Manager involve a statistician in this process.

4.4.5. Case study 1 provides an example in which multiple techniques are used to assess groundwater monitoring data. Case study 2 provides an example of using a combined Shewart/CUSUM method to identify a release at a site.

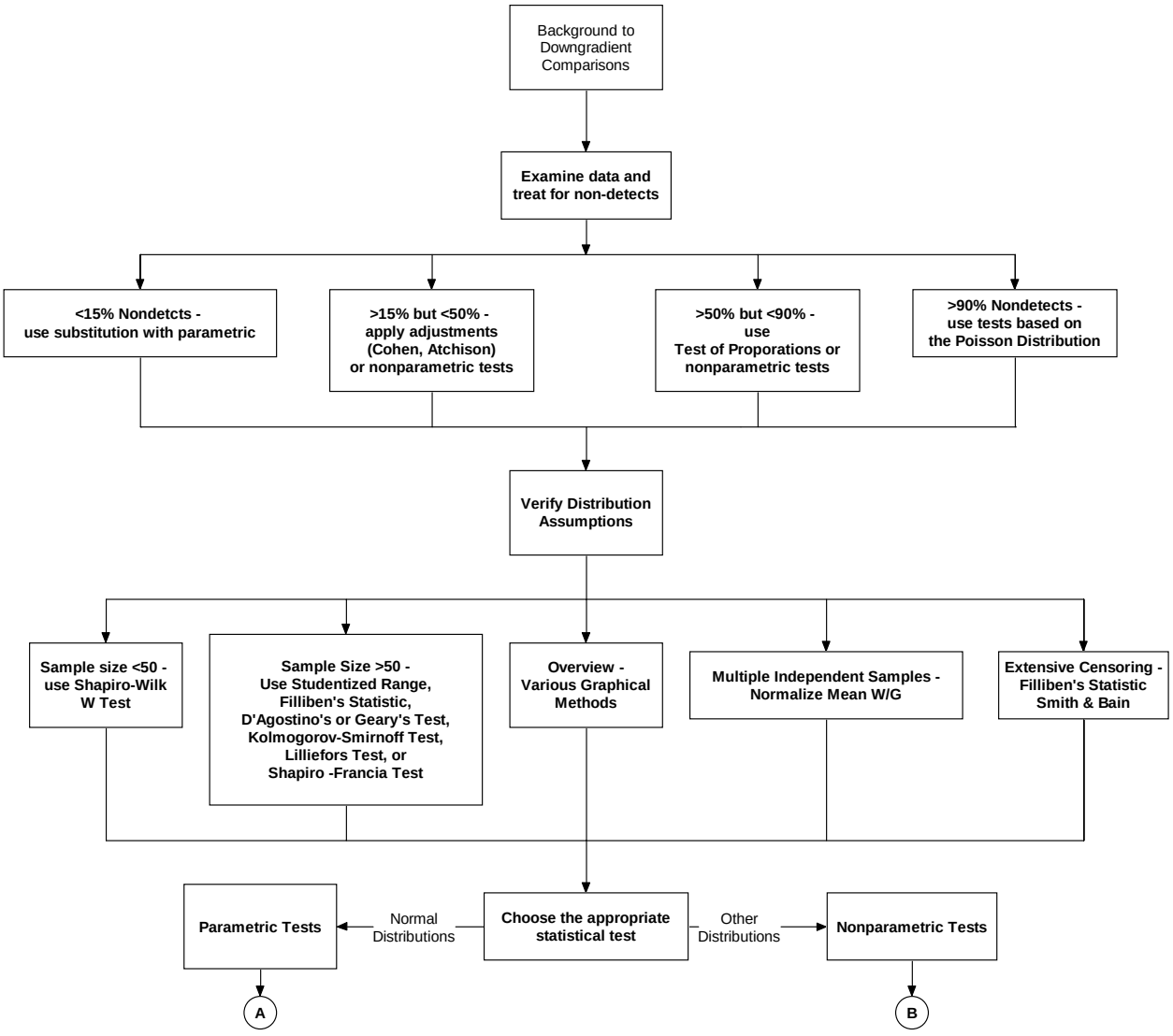


Figure 4-3. Statistical decision tree with options for groundwater monitoring—Part 2.

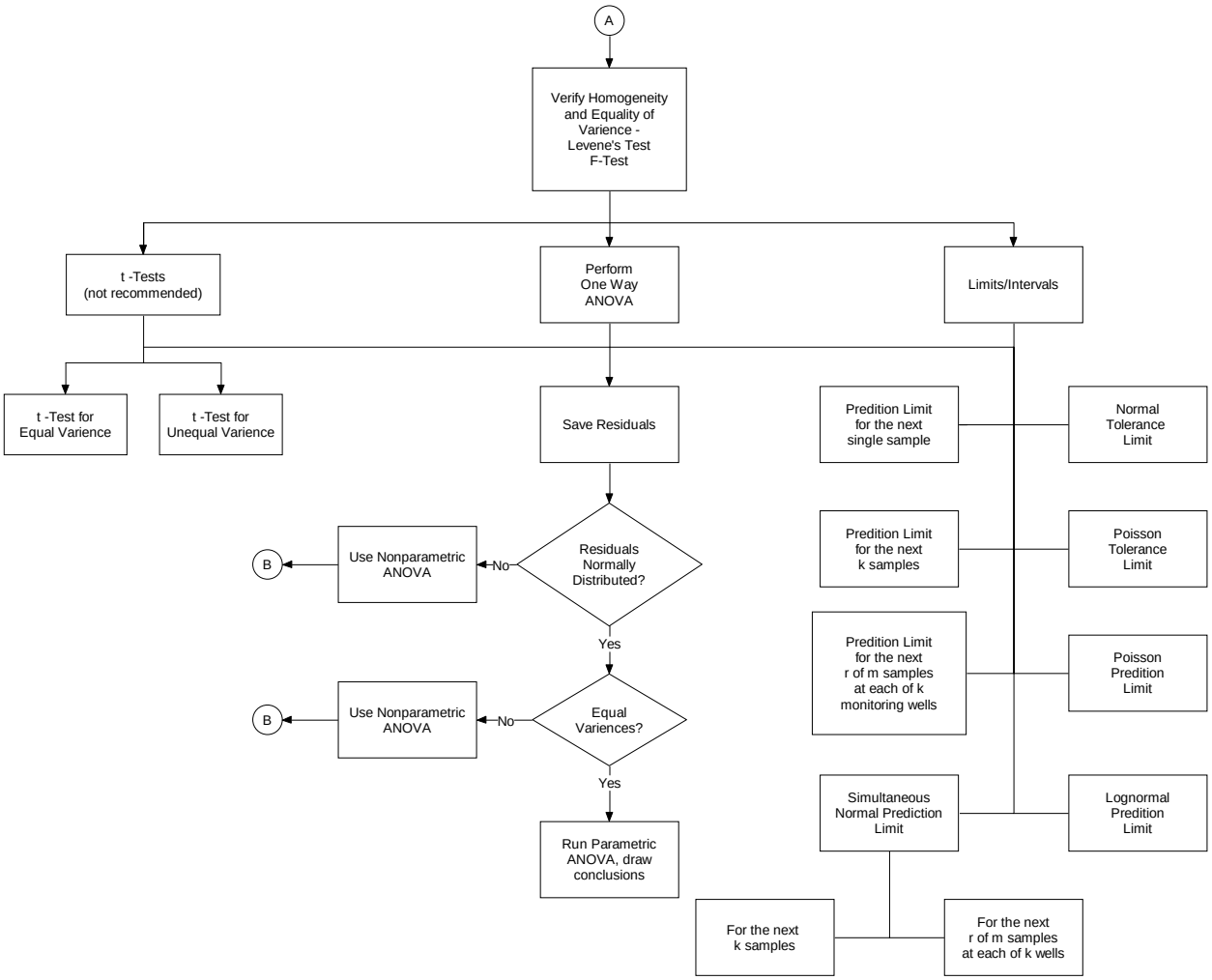


Figure 4-4. Statistical decision tree with options for groundwater monitoring—Part 3.

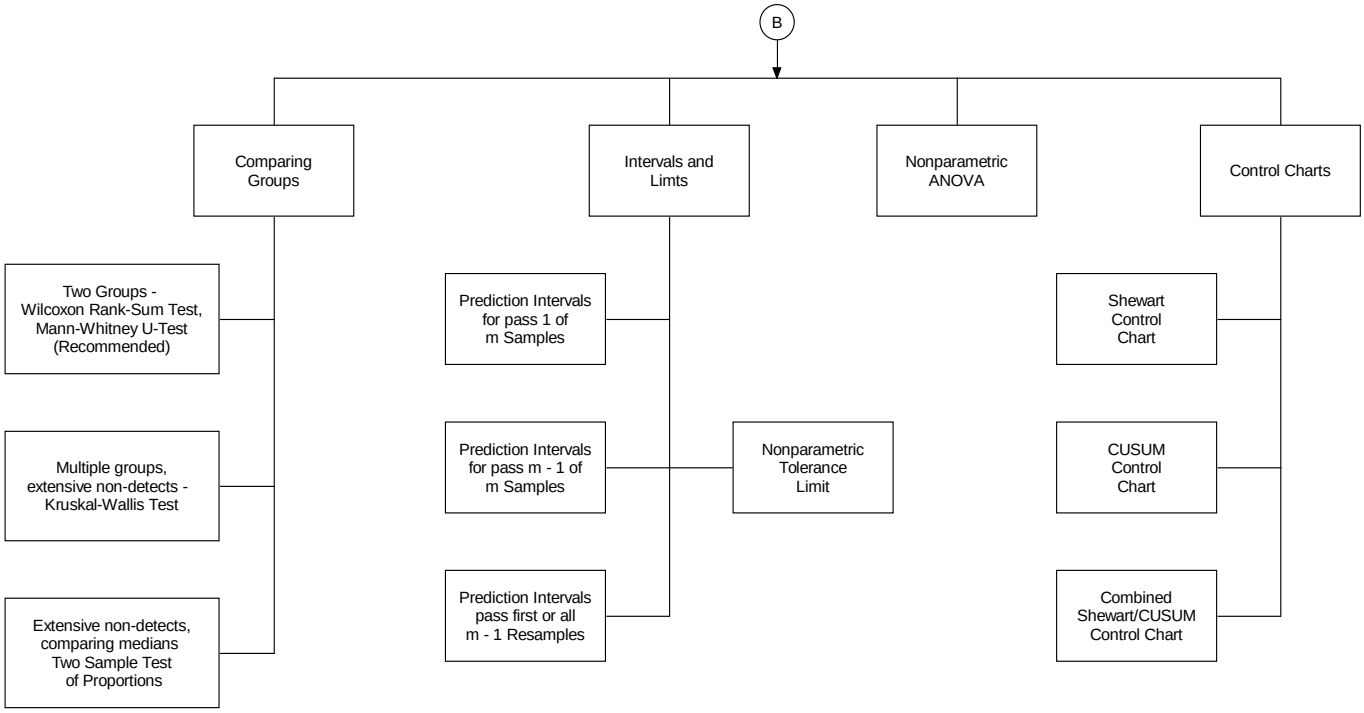


Figure 4-5. Statistical decision tree with options for groundwater monitoring—Part 4.

4.5. Case Study 1—Groundwater Monitoring. At a manufacturing facility in Virginia, a long-standing tetrachloroethene (PCE) plume is being hydrologically contained and treated with a combination of vapor extraction and groundwater pump-and-treat. The facility has been engaged in long-term monitoring for over 20 years and uses a variety of techniques to assess permit compliance. Sample statistics allow the facility to determine whether remediation at the site is causing reductions in PCE concentrations. Table 4-1 presents an example of summary statistics and testing results in a fashion that is easily understood for both compliance and detection monitoring.

4.5.1. For compliance monitoring at wells with known past contamination (MW1 to MW4), increasing or decreasing statistical trends were determined at the 90 and 95% level of confidence, respectively, as negotiated with state regulators at the site.

4.5.2. Trend analyses, control charts, and tolerance limits are being used for the four wells under the category “Comp” and for the three wells under the category “Trend.” Typically, differing DQOs would be set for compliance and detection wells and only one set of statistical tests would be performed. However, the regulatory negotiations at this site mandated identical tests for both types of wells. (This example demonstrates an opportunity for improving past negotiated monitoring with regulators.)

4.5.3. Additionally, the number of detections greater than the "tolerance limit" is specified for each well. The 95% UTL is constructed from a set of background wells, also as determined in the site permit at time of negotiation with regulators. Because there is background contamination the following case study provides an example of using a combined Shewart/CUSUM method to identify a release at a site.

Table 4-1.
Groundwater Monitoring Data for Case Study 1

Identification				Descriptive Statistics				Trend Significance		Excursions?	
Class	Well	*n*	Avg	Med	*s*	*W*	MK	95%	90%	Control Chart	Tolerance Limit
Comp.	MW1	46	5595.0	5610.0	982.0	Yes	No	Up	Up	None	3
	MW2	44	62.3	67.2	21.5	Yes	No	Down	Down	None	None
	MW3	40	1295.0	1198.0	367.8	No	No	Down	Down	None	None
	MW4	47	133.8	133.7	22.3	Yes	No	Down	Down	None	None
Detect.	MW5	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW6	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW7	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW8	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW9	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW10	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW11	16	0.369	0.4	0.307	Yes	No	None	None	None	None
	MW12	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW13	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW14	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW15	16	0.039	0.0	0.088	No	No	None	None	None	None

Notes:
Comp Compliance
n Number of samples
Avg Sample mean
Med Sample median
s Sample standard deviation
W Normal according to Shapiro-Wilk test at 95% confidence?
MK Seasonality according to Mann-Kendall test at 95% confidence?

4.6. Case Study 2—Shewart/CUSUM Monitoring. A groundwater plume at a site is currently being addressed via pumping and treating large amounts of groundwater. The system is very costly, and the site owner wishes to change the system configuration. Project regulators want to know whether changing the system (in this case, shutting off the treatment system) will increase measured trichloroethene (TCE) values near the leading edge of the plume. A special type of compliance monitoring was initiated to determine whether concentrations after system shutdown exceeded a "trigger" level. Table 4-2 lists the eight most recent TCE measurements at monitoring well B-37 prior to altering the system.

4.6.1. The sample mean for these data ($\bar{x}$) is 4.3 parts per billion (ppb) and the sample standard deviation (s) is 1.1 ppb. These values are used in statistical tests for normality, which did not indicate the data set is non-normal. (A hypothesis of normality cannot be rejected at the 90% significance level using any of the Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov, or D'Agostino tests [See Appendix F].)

Table 4-2.
Eight Most Recent TCE Measurements in B-37

Well ID	Sample Date	Measured TCE Concentration (µg/L)
B-37	7-Jun-99	3.0
B-37	29-Nov-99	3.2
B-37	26-Jun-00	4.5
B-37	3-Jan-01	5.8
B-37	16-May-01	5.9
B-37	4-Oct-01	3.2
B-37	27-Mar-02	4.6
B-37	10-Dec-02	4.3

4.6.2. Table 4-3 lists the measured TCE concentrations in this well over eight monitoring periods after system shutdown in mid-December 2002, and the associated Shewart/CUSUM statistical parameters (see Appendix K). The Shewart/CUSUM calculations shown in the table are plotted in the Figure 4-6.

Table 4-3.
TCE Measurements and Shewart/CUSUM Calculations

Hypothetical Sampling Event	Sampling Period, *i*	TCE Concentration (µg/L)	z_i	z_{i-1}	S_i
Winter 2002	1	4.9	0.6	–0.4	0
Spring 2003	2	5.7	1.2	0.2	0.2
Summer 2003	3	6.0	1.4	0.4	0.7
Fall 2003	4	3.9	–0.4	–1.4	0.0
Winter 2003	5	9.8	4.8	3.8	3.8
Spring 2004	6	8.1	3.3	2.3	6.1
Summer 2004	7	7.5	2.8	1.8	8.0
Fall 2004	8	10.6	5.5	4.5	12.5

z_i = standardized result (or normalized concentration)
S_i = cumulative sum

4.6.3. The quantities z_i and S_i (discussed in Appendix K) were calculated to determine whether changing the system configuration resulted in an unacceptable change (i.e., increase) in the TCE concentration in Well B-37.

4.6.4. The first out-of-control event occurred in winter 2003 when the z_i of 4.8 exceeded the Shewart threshold of 4.5. In addition, although the normalized concentration z_i decreases after the fifth sampling event following the start of shutdown, S_i continues to increase beyond and remains greater than the threshold of 5.0 for this quantity through fall 2004.

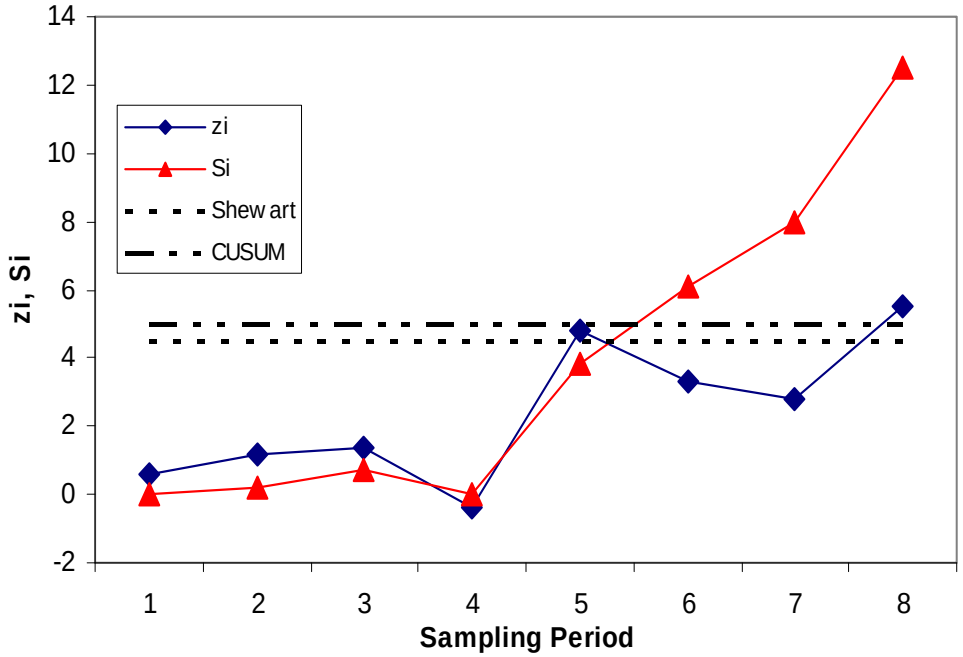


Figure 4-6. Shewart/CUSUM control chart, Well B-37.

4.6.5. The results of the testing showed that reconfiguring the system appeared to change the concentrations of TCE in this downgradient well at a statistically significant level. The reconfiguration was abandoned, and project planners began to reevaluate their understanding of groundwater movement at the site.

4.6.6. The Shewart/CUSUM method is commonly applied to landfills for detection monitoring, although it has obvious additional uses in other long-term monitoring applications. For instance, by looking for an insignificant change over time, a site stakeholder could suggest that monitoring at a natural attenuation site could be discontinued.

4.7. Optimization. The process of optimization is similar in many ways to the process of sensitivity analysis. In both cases, one makes planned adjustments to the system and looks for changes in the outcome. The process of optimization involves assessing whether or not a change made in the system results in a beneficial outcome—improving system performance, for example, by reducing cost, increasing efficiency, or shortening the time to completion. This can be ac-

complished by comparing data taken after the adjustments have been made to historical data for the process using a variety of hypothesis testing tools.

4.7.1. It is also possible to examine trends in the system after taking into account seasonal and other forms of cyclic correlation. For example, when a time plot is examined for trend after a system modification, one may find that the slope of the time plot line changes, indicating a change in system performance. A time series plot is a graph showing how a parameter (e.g., TCE concentration) changes over time. A trend is a statistically significant change upward or downward with a certain degree of confidence. Whether or not that change is significant and an assessment of the magnitude of its impact can be addressed using trend tests such as *Mann-Kendall* and *Sen's Slope Estimator.*[*]

4.7.2. Another example of system optimization is in addressing such issues as the monitored analyte list and the frequency of sampling, both of which have economic implications and can have regulatory implications as well. As a hypothetical extreme case for illustration, assume that a monitoring well network must be sampled four times each year; that there are 10 wells in the network; and that each well is monitored for 50 constituents, all of which must be non-detects.

4.7.3. The statistics underlying the determination of a detection limit (e.g., if normality is assumed and the detection limit is the "Type I detection limit" or "critical value" in Appendix C) are such that there is only a 1% probability of a false positive at the detection limit while, as the statistics employed are one-sided, there is a 50% probability of a false negative at the detection limit. Thus, in the course of a given year, based on probability alone, the facility could falsely report itself in violation an average of 20 times, while falsely reporting compliance 1000 times (on the average). In fact, it can be demonstrated that simply because of the inherent Type I error rate associated with any statistical test, where literally thousands of such comparisons may be required, whether at the detection limit or otherwise, the probability of a false conclusion of violation approaches unity. Thus, it is always in the best interest of the regulated facility to limit the number of analytes for which one tests to the smallest possible number. Every permit renewal period or 5-year review should be used as an opportunity to further limit the analyte list. Even hypothetically, one can see that this approach is inefficient (costly), and reaching the goal of all non-detect is an example of a poorly defined quality objective. Detection limits can differ across laboratories and over time, and, clearly, they are not related to risk management in any way.

4.7.4. Another approach currently under study is the use of statistics to establish predictable correlation between the analyte of interest and some parameter that is more readily or cost-effectively measured than the analyte of interest. This "harbinger" or "calibration" approach has its roots in the commonly accepted practice of monitoring for indicator parameters such as pH, conductivity, total organic carbon, and total organic halides in place of specific analytes. If a rig-

[*] Appendix P.

orous regression analysis of historical data suggests a quantitative linkage between the concentration of arsenic and magnesium at a given site, it should be possible to delete, or at least reduce the frequency of analysis, for one or the other analyte, particularly in the case where both analytes have historically displayed compliant behavior. It would also be useful in this type of situation if a functional relationship and the uncertainty associated with that relationship could be established.

4.7.5. To assess the viability of monitored natural attenuation as a remedial alternative, it is essential to demonstrate: i) degradation of VOCs from parent products through to mineralization; and ii) correlation between that degradation and appropriate geochemical conditions. An example of assessing the correlation of parameters at a site in Maryland is illustrated in Case Study 3. Correlation measures show how strongly variables (or parameters) are related, or change with each other.

4.8. Case Study 3—Trend Analysis and Correlation in Natural Attenuation Data.

4.8.1. The data used for a site in Maryland were organized along a single geographic line, from the suspected source to a groundwater discharge zone located along a creek bed. Location was displayed in feet from the center of the suspected source. The parent constituent was PCE. The primary geochemical indicators of interest (for purposes of this case study) were dissolved oxygen (DO) and oxidation-reduction potential (redox).

Table 4-4.
Attenuation Data

Distance from Source (feet)	PCE (µg/L)	DO (mg/L)	Redox (mV)
0	320	0	–210
50	1430	0	–220
100	960	0.2	–170
150	780	0.3	–140
200	570	0.6	–80
250	630	0.5	–30
300	580	0.8	10
350	340	1.1	40
400	430	1.4	70
450	130	1.7	90
500	12	3.5	120

4.8.2. The data for the three parameters of interest are presented in Table 4-4. The data were then plotted against distance from the origin (source) to identify trends over distance. A Mann-Kendall trend analysis showed that PCE concentration decreased over distance. Redox and DO are positively correlated to one another with a Pearson's *r* value of 0.84. Geochemical understanding of natural attenuation requires that redox and DO should be inversely correlated to

PCE concentration, and the Pearson's *r* values for DO and redox are –0.71 and –0.74, respectively. The results are displayed in the Figures 4-7 and 4-8. In summary, the results suggest that conditions for natural attenuation are present.

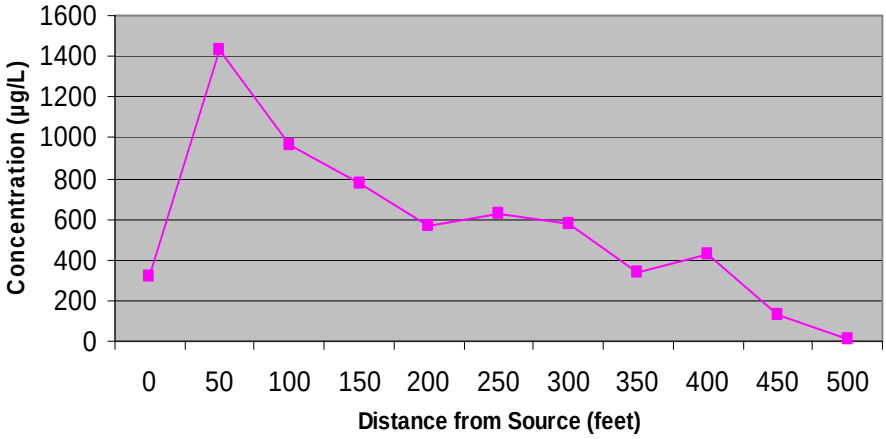


Figure 4-7. PCE concentration versus distance.

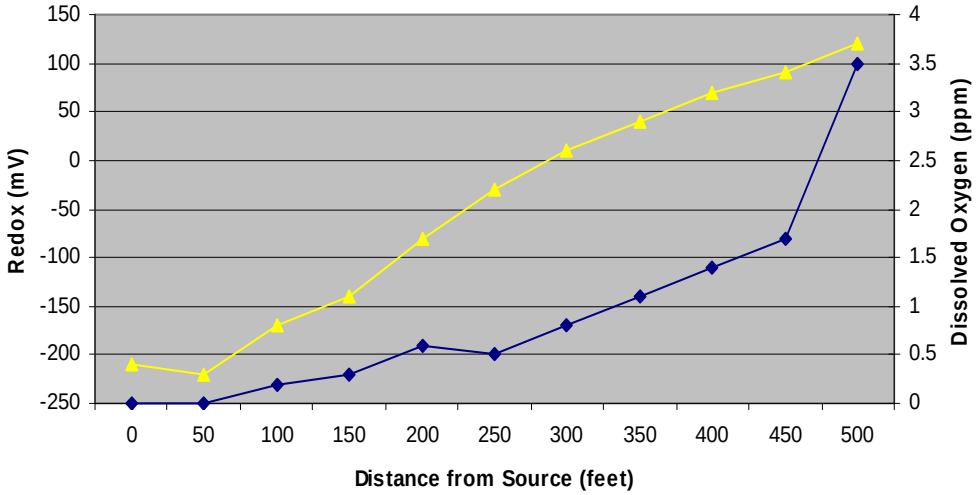


Figure 4-8. Geochemical parameters versus distance from source (yellow triangle—redox; blue diamond—dissolved oxygen).

Section II
Applying Cleanup Levels

4.9. Introduction. When derived in accordance with USEPA's risk assessment guidance, risk-based cleanup levels are intended to represent the average contaminant concentration within the exposure unit that can be left on the site following remediation (Schulz and Griffin, 2001). In contrast, a "not-to-exceed" cleanup level drives remediation solutions that involve treating or removing any and all media with contaminant concentrations that exceed the cleanup level. The result is that applying a not-to-exceed level may result in over-remediation.

4.9.1. Calculated using risk assessment principles, the cleanup goal concentration is usually defined as an exposure unit concentration that will meet the target risk level agreed to by the design team and regulatory authorities. Some sample concentrations exceeding the cleanup objective can remain in place as long as the overall exposure concentration, calculated to a predetermined level of certainty, meets the cleanup goal (and likewise the agreed upon risk level). Because of the uncertainty associated with estimating the true average concentration of a contaminant at a site, USEPA recommends use of the 95% one-sided, upper confidence limit of the arithmetic mean (95% UCL) of the sample data to represent the exposure unit concentration term in risk assessments (EPA 9285.7-09A and EPA OSWER 9285.6-10). Consequently, a risk-based cleanup level should generally be interpreted as the 95% UCL of the contaminant concentration within the exposure unit following remediation.

4.9.2. However, *draft* USEPA guidance suggests specific situations in which application of the cleanup level as an area average may not be appropriate (USEPA, 2002) These include the following.

4.9.2.1. Exposure within the exposure unit is not random.

4.9.2.2. The cleanup level is based on acute rather than chronic exposure.

4.9.2.3. The cleanup level is not risk-based (i.e., it considers factors other than risk).

4.9.2.4. The quality of site characterization data is not optimal but it is not worth investing in additional sampling.

4.9.2.5. Given the site conditions (complexity, size, characterization, contaminant distribution), it is not cost-effective to do the necessary sampling and statistical analysis.

4.9.2.6. The community will not accept leaving soil with contaminant concentrations that exceed the cleanup level on the site.

4.9.3. If applying cleanup levels as an area average is appropriate, there are two basic approaches: i) using non-spatial statistical methods to determine a not-to-exceed concentration, and ii) using spatial statistical methods to iteratively re-calculate the UCL until the optimal "design line" for the remedial action is determined.

4.10. Determining Not-to-Exceed Concentrations Using Non-Spatial Statistics. Draft USEPA guidance (USEPA, 2002) defines the remedial action level (RAL) as the maximum concentration that may be left in place within an exposure unit such that the average concentration (or 95% UCL) within the exposure unit is at or below the cleanup level. Non-spatial techniques may be appropriate for calculating the RAL when there is no spatial correlation between contaminant concentrations, such as at a dump site where small, randomly located spots of high contaminant concentrations are interspersed with areas of lower concentrations. Non-spatial techniques are based on the mean and standard deviation of the sample contaminant concentration data and on how those metrics change as soils with high contaminant concentrations are replaced with post-remediation concentrations during remediation. The draft guidance describes two non-spatial statistical methods for calculating remedial action levels that ensure that post-remediation area average contaminant concentrations achieve cleanup levels: i) iterative truncation method, and ii) confidence response goal method. These methods are also reviewed in Schulz and Griffin (2001). Both methods can be applied in a spreadsheet calculation or programming language.

4.10.1. *Iterative Truncation Method.*

4.10.1.1. The iterative truncation method is based on the identifying and removing (truncating) high values in the sample concentration measurements (hot spots), replacing them with the post-remediation concentration (e.g., concentration in clean fill), and calculating the hypothetical post-remediation average concentration (95% UCL) in the exposure unit. Starting with the highest concentration in the data set, the process is repeated iteratively until the post-remediation 95% UCL is less than or equal to the cleanup level. The highest sample concentration remaining in the data set is designated the RAL.

4.10.1.2. This method is sensitive to the completeness of site characterization and the range of resultant sample concentrations. According to the draft USEPA guidance, to use this method with confidence, good site characterization through extensive, unbiased sampling is required and the resulting data must adequately represent random, long-term exposure to receptors. This method is not reliable when samples are not independently and randomly located.

4.10.2. *Confidence Response Goal Method.* Bowers et al. (1996) developed a method for calculating a confidence response goal (CRG), which, like the RAL, is a not-to-exceed level. This method can be applied at sites where there is a non-spatial, lognormal distribution of contamination (USEPA, 2002).

4.10.2.1. As described in the draft USEPA guidance, the basic premise of the method is that the CRG can be expressed as a function of the geometric mean and the geometric standard deviation of contaminant concentrations, and the desired reduction in exposure, which is defined as the ratio of average post-remediation concentration to the average pre-remediation concentration. The guidance provides a summary of the method, documents the equation for calculating the CRG, and refers the reader to the original paper (Bowers et al., 1996) for details on the derivation of the function.

4.10.2.2. The Schulz and Griffin (2001) review of the two non-spatial methods concludes that the CRG method is less sensitive than the iterative truncation method to changes in the highest sample concentrations and recommends the use of the CRG method when the contaminant distribution is lognormal.

4.10.3. *Using Spatial Statistical Methods to Determine "Design Line" for Remediation.* The distribution of contaminant concentrations may be spatially correlated at many sites where there is an original source or release that is subject to environmental fate and transport mechanisms. Contaminant concentrations in and around the original source or release may be higher than those at greater distances, or they may be higher where there is a mechanism of accumulation or an environmental "sink." Biased sampling is frequently applied in such cases because a high number of samples is desired in areas with high variance and uncertainty (for example, near the source area), and a lower number of samples is often sufficient to characterize areas with expected low variance and uncertainty. The concept of taking "step out" samples in the vicinity of sample locations where high contaminant concentrations are detected also introduces bias into the sampling plan. Geostatistical techniques are statistical procedures designed to process spatially correlated data (see Appendix R on Geostatistics). Unlike the non-spatial techniques, geostatistical techniques are well suited for evaluation of biased data sets.

4.10.3.1. The draft USEPA guidance presents an example of the determination of RALs using geostatistical techniques. The example has two simplifying features that can be found on many (but not all) sites: i) contamination that is surface only, and ii) the importance of a residential scenario. For this example, the steps for determining RALs are as follows.

4.10.3.1.1. Create an iso-concentration map of the site by modeling the spatial correlation underlying measured values.

4.10.3.1.2. Superimpose a grid of exposure units over the site and compute average contaminant concentrations in each exposure unit.

4.10.3.1.3. Identify zones that must be remediated to reduce average concentrations in all exposure units to the appropriate cleanup level. This is an iterative process, where the higher contaminant concentrations are replaced with post-remediation concentrations and average con-

taminant concentrations in each exposure unit are re-calculated. The final cutoff concentration is the RAL.

4.10.3.1.4. Use the original iso-concentration map to define zones with concentrations in excess of the RAL. The contoured zone is the area that requires remediation.

4.10.3.2. The draft guidance cautions against using geostatistical techniques if contaminant concentrations show a random, non-spatial pattern, or if the anticipated benefits from geostatistical analysis do not justify the costs. For example, even in cases of conservatively biased data, spatial statistical methods may not be warranted when non-spatial methods are determined to result in cleanup objectives that are both sufficiently conservative from the risk perspective and acceptable from the cleanup cost perspective. Additionally, conservatively biased, non-spatial methods may be needed from a practical view when adequate technical or computational resources are not available. Proponents of geostatistical techniques counter that presentating the site contamination and remediation results as spatial is a highly intuitive and visually powerful approach, and therefore enhances communication among the parties during risk management discussions. Available computational tools make it possible to find the point of diminishing returns where an increase in remediation has little effect on reducing risk in a cost-effective manner.