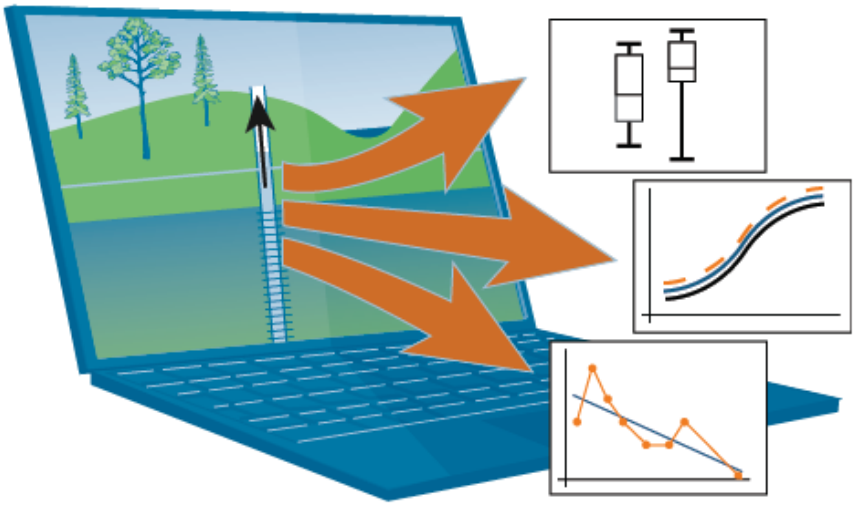
Guidance Document

Groundwater Statistics and Monitoring Compliance

Statistical Tools for the Project Life Cycle



December 2013

Prepared by
The Interstate Technology & Regulatory Council
Groundwater Statistics and Monitoring Compliance Team

ABOUT ITRC

The Interstate Technology and Regulatory Council (ITRC) is a public-private coalition working to reduce barriers to the use of innovative environmental technologies and approaches so that compliance costs are reduced and cleanup efficacy is maximized. ITRC produces documents and training that broaden and deepen technical knowledge and expedite quality regulatory decision making while protecting human health and the environment. With private and public sector members from all 50 states and the District of Columbia, ITRC truly provides a national perspective. More information on ITRC is available at www.itrcweb.org. ITRC is a program of the Environmental Research Institute of the States (ERIS), a 501(c)(3) organization incorporated in the District of Columbia and managed by the Environmental Council of the States (ECOS). ECOS is the national, nonprofit, nonpartisan association representing the state and territorial environmental commissioners. Its mission is to serve as a champion for states; to provide a clearinghouse of information for state environmental commissioners; to promote coordination in environmental management; and to articulate state positions on environmental issues to Congress, federal agencies, and the public.

DISCLAIMER

GSMC-1

Groundwater Statistics and Monitoring Compliance

Statistical Tools for the Project Life Cycle

December 2013

Prepared by
The Interstate Technology & Regulatory Council
Groundwater Statistics and Monitoring Compliance Team

GSMC-1 Revisions

Date	Revision
7/2014	Web-based document and PDF: Revisions include link updates, minor text edits, and formatting updates.
12/2014	Web-based document and PDF: Revisions include link updates, minor text edits, and formatting updates.

EXECUTIVE SUMMARY

The Interstate Technical and Regulatory Council (ITRC) Groundwater Statistics and Monitoring Compliance (GSMC) team has prepared this guidance document to help environmental practitioners better apply groundwater statistics in environmental projects. The purpose of this document is to help these practitioners to understand, interpret, and use statistical techniques to successfully manage groundwater compliance or cleanup projects. The guidance presented here is specifically for environmental project managers who must review or use statistical calculations for reports, who must make recommendations or decisions based on statistics, or who must demonstrate compliance for groundwater projects. These individuals typically have a technical background and experience in one or more disciplines related to site compliance or cleanup, but do not have specific expertise in statistics or access to in-house statistical expertise.

Groundwater statistical methods have applications throughout the life cycle of environmental projects. This document organizes the discussion of site management around five main stages in an environmental project life cycle: release detection, site characterization, remediation, monitoring, and closure. These tasks and their descriptions presented here correlate with the activities described in various regulatory programs. Although individual projects may vary in their progression through these stages, groundwater statistical tests can support decision making, regardless of how the project is defined.

Each of the project life cycle stages listed above progresses through development and refinement of the conceptual site model (CSM). This document explores some of the common problem statements that guide decision making throughout environmental projects and poses a list of typical study questions:

- Study Question 1: What are the background concentrations?
- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 3: Are concentrations above or below a criterion?
- Study Question 4: When will contaminant concentrations reach a criterion?
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?
- Study Question 8: How do contaminant concentrations change with distance from the source area?
- Study Question 9: Is the sampling frequency appropriate (temporal optimization)?
- Study Question 10: Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

These study questions relate one to another in the context of specific groundwater evaluation objectives that may be familiar to practitioners during various project life cycle stages. Study Questions 1 and 2 assess background concentrations. Study Questions 3 and

4 assess contaminant concentrations with respect to criteria (such as regulatory standards). Study Questions 5 and 6 evaluate temporal trends in data sets. Study Questions 7 and 8 assess temporal and spatial rates of change for contaminants. Study Questions 9 and 10 evaluate if the frequency of sampling and spatial coverage of wells are appropriate, leading to an optimal monitoring program.

These study questions serve as a bridge to connect life cycle activities with relevant statistical tests and methods. This document is not a tutorial on tests and methods, but rather illustrates how these tests and methods, along with general statistical approaches, can address the practical applications, challenges, and misapplications associated with groundwater statistics. In addition, available tools to conduct these tests and methods are presented in this document. This guidance document provides an overview of the United States Environmental Protection Agency's (USEPA's) March 2009 *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities* (known as the Unified Guidance) and other resources and shows how statistics are specifically applied to groundwater data for environmental projects.

ACKNOWLEDGMENTS

The members of the Interstate Technology & Regulatory Council (ITRC) Groundwater Statistics and Monitoring Compliance (GSMC) Team wish to acknowledge the individuals, organizations, and agencies that contributed to this web-based guidance document on groundwater statistics for monitoring and compliance during the environmental project life cycle.

As part of the broader ITRC effort, the GSMC Team effort is funded by the U.S. Department of Defense, the U.S. Department of Energy, the U.S. Environmental Protection Agency, and ITRC's Industry Affiliates Program. ITRC operates as a committee of the Environmental Research Institute of the States (ERIS), a Section 501(c)(3) public charity that supports the Environmental Council of the States (ECOS) through its educational and research activities aimed at improving the environment in the United States and providing a forum for state environmental policy makers.

The team recognizes the following state's support of team leadership and guidance preparation:

- California Department of Toxic Substances Control - Ning-Wu Chang, Team Leader
- Alaska Department of Environmental Conservation – Marlena Brewer
- Broward County Florida, Pollution Prevention, Remediation and Air Quality Division - Sat Mellacheruvu
- Delaware Department of Natural Resources & Environment Control – Stephen Johnson
- Florida Department of Environmental Protection – Elizabeth Bartlett
- Kentucky Division of Waste Management – Edward Winner and Todd Mullins
- Indiana Department of Environmental Management – Harold Templin
- Maine Department of Environmental Protection – Gail Lipfert
- New Jersey Department of Environmental Protection – Robert Soboleski
- Oklahoma Department of Environmental Quality – Orphius Mohammad
- South Carolina Department of Health and Environmental Control – James Bowman
- Utah Department of Environmental Quality – Helge Gabert
- Washington DC, Department of Environment - Nazmul Haque
- Washington [State] Department of Ecology – Dibakar Goswami

The team would also like to recognize the excellent participation from various federal agencies and other public institutions:

- U.S. Air Force – Philip Hunter
- U.S. Army Corps of Engineers – Thomas Georgian and Anna Butler
- U.S. Department of Energy – Beth Moore, Paul Beam, Allan Harris, and John Hathaway

- U.S. Environmental Protection Agency – Richard (Kirby) Biggs and Ray Ledbetter
- U.S. Navy – Palmer Anderson, William Major, and Kenda Neil

Stakeholder and academic perspectives are critical to the success of any ITRC document; the team would like to thank David Smit for his efforts as the stakeholder.

Finally, the team would also like to recognize the efforts and contribution of the following consultants and industry representatives:

- Barr Engineering Company – Jennifer Brekken
- CDM Smith – Ernest Ashley
- CH2M Hill – Devamita Chattopadhyay
- Chevron - Natalie Woodard
- Conestoga Rovers & Associates – Wesley Dyck
- ENVIRON International Corporation – Sarah Stoneking and Christopher Stubbs
- ERM – Carolyn Moore
- Exxon Mobil – Mark Malander
- Geosyntec Consultants Inc. – Arnab Charkrabarti, Keith Tolson, and Ehsan Rasa
- GSI Environmental Inc. – Thomas McHugh and Mindy Vanderford
- Kleinfelder - Lizanne Simmons
- Langan Engineering & Environmental Services - Jason Goff
- Lockheed Martin – Anita Singh (USEPA representative)
- MacStat Consulting – Kirk Cameron (Air Force and USEPA designated representative)
- Neptune and Company Inc. – Randall Ryti
- Papadopoulos & Associates – Prashanth Khambhammettu
- Plateau Geoscience Group, LLC – Mavis Kent
- Shell – Leroy (Buddy) Bealer
- Sterling Global Operations - Teresie Walker
- Tetra Tech - John (Brad) Peebles
- Trihydro Corporation – Jim Gleason

The GSMC Team would like to thank the team Program Advisor, Lesley Hay Wilson of Sage Risk Solutions LLC, for her efforts to keep the project focused and well supported to develop this web-based guidance document.

All parties who contributed to this document whether named or unnamed, team member, independent reviewer, or ITRC staff, are thanked by the GSMC Team for their efforts. Some made major contributions to the project while others made minor ones; all are appreciated for their time and effort.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

This guidance document explains statistical techniques to evaluate and optimize groundwater monitoring for environmental projects. The primary audience for this guidance is environmental practitioners who have technical and project management experience, but who may not have specific expertise in statistics. Public and tribal stakeholders reviewing environmental reports will also find this guidance helpful. Other good sources of information about statistics include the United States Environmental Protection Agency's (USEPA's) March 2009 *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities* (known as the Unified Guidance), *Statistical Methods for Groundwater Monitoring* (Gibbons, Bhaumik, and Aryal 2009), and several ASTM International (ASTM) publications on statistical methods for environmental monitoring (ASTM 2010a, 2010b, and 2012), as well as a wide variety of tools and software packages for performing statistical calculations and evaluations. Even with these resources, however, practitioners may still feel challenged when reviewing or implementing statistics. This guidance document provides an overview of the Unified Guidance and other resources and shows how to apply statistics specifically to analytical results from groundwater sampling in order to make better decisions in environmental projects.

1.1 Purpose

The purposes of this document and associated Internet-based training include the following:

- provide greater clarity to the planning, implementation, and communication of groundwater statistical methods and results
- provide information about available statistical tools and software in a useful format
- help practitioners to review and use statistical methods to improve the quality of their decisions
- help practitioners to identify the specific tasks within a project life cycle that benefit from statistical approaches
- provide better understanding of the statistical concepts that may be used for systematic planning such as USEPA's seven-step Data Quality Objective (DQO) Process (USEPA 2006a)

1.2 Scope and Limitations

This document offers practical tips for using information that is already published in other resources such as USEPA's Unified Guidance. This document is not a stand-alone tutorial on statistics, but instead addresses the practical applications, challenges, and misapplications associated with the use of groundwater statistics. The study questions, methods, and software packages are not all-inclusive and others may be appropriate. Concepts presented in this document apply to groundwater projects in many regulatory programs such as those under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), the Resource Conservation and Recovery Act (RCRA), voluntary cleanup, and underground storage tank (UST) programs. The

document is organized around life cycle stages of typical environmental projects including release detection, site characterization, remediation, monitoring, and closure. The guidance presented here will help users to ask the right questions about data and identify appropriate statistical methods to answer these questions.

This document does not address issues associated with hydrogeology, sampling methodology, Conceptual Site Model development, and analytical data quality. For example, it does not provide guidance on well installation, sampling, or hydrogeological interpretation. The default assumption, unless otherwise stated, is that data are representative, valid, and usable. In practice, statistical approaches such as tests for outliers and exploratory data analysis (EDA) described in Section 3.5: Testing Assumptions can be used with available data of uncertain quality. However, a more complete understanding of what makes data "ideal" may put the existing data into context so that it can be used with the appropriate level of confidence (see Section 3.3.1: Data Quality).

Additionally, this document does not provide guidance on geostatistical analyses or software, with the exception of some tools included in Appendix D that offer spatial analysis for optimizing well placement, redundancy, and sampling frequency.

In order to simplify discussions in this document, certain terms have been used to standardize various concepts. Some of terms may have different meanings for different regulatory agencies; their use here is not intended to undermine or change a particular regulatory meaning. Examples include "chemical" or "contaminant," which are used depending on the context, and "criterion," which is used instead of terms such as "Groundwater Protection Standard (GWPS)," "cleanup standard," or "cleanup level."

1.3 Background

The guidance presented here condenses and simplifies a selection of important methods from the Unified Guidance. The Unified Guidance was developed for the statistical analysis of groundwater data at RCRA facilities, but the statistical tests and graphical methods it describes are broadly applicable for a variety of other environmental programs.

The Unified Guidance is consistent with strategies for systematic planning and conceptual site model (CSM) concepts and was developed to help evaluate groundwater monitoring data for regulatory compliance. The statistical analyses described in the Unified Guidance and this document require acceptable data. A series of USEPA quality management guidance documents that provide information on data quality are also available:

- *Guidance on Systematic Planning Using the Data Quality Objectives Process*, QA/G-4 (USEPA 2006a)
- *Guidance on Choosing a Sampling Design for Environmental Data Collection*, QA/G-5S (USEPA 2002a)
- *Data Quality Assessment: A Reviewer's Guide*, QA/G-9R (USEPA 2006b)
- *Data Quality Assessment: Statistical Methods for Practitioners*, QA/G-9S (USEPA 2006c)

- *Guidance on Environmental Data Verification and Data Validation*, QA/G-8 (USEPA 2008b)

Additionally, the information in this ITRC guidance document and others may assist with evaluation of soil contamination, remedial system optimization, and meeting goals of green and sustainable remediation. The ITRC guidance document *Incremental Sampling Methodology* (ITRC 2012) addresses soil sampling approaches designed to ensure representative, reproducible, and defensible data. ITRC's *Improving Environmental Site Remediation Through Performance-Based Environmental Management* (ITRC 2007b) and related documents address how to use remedial process optimization to systematically evaluate and manage uncertainty associated with the remediation process. Recommendations in these documents include the use of statistical tools such as the Monitoring and Remediation Optimization System (MAROS) software and the Geostatistical Temporal/Spatial (GTS) optimization algorithm to optimize monitoring networks. These software tools and others are described in Appendix D of this document. In addition, ITRC's *Green and Sustainable Remediation: A Practical Framework* (2011a) and related documents provide a framework to achieve green and sustainable goals with better site management decisions. Finally, additional information can be found in the Department of Energy's *Scientific Opportunities for Monitoring at Environmental Remediation Sites: Integrated Systems Based Approaches to Monitoring* (Bunn et al. 2012).

1.4 Document Audience

Based on the groundwater statistics survey (Appendix E), the target audience for this guidance is a project manager (in industry, government, or consulting) who must review or use statistical calculations to generate a report or demonstrate compliance for a groundwater-monitoring project. This individual typically has technical experience in one or more disciplines related to site compliance or cleanup but does not have specific expertise in statistics. This individual may not have access to in-house statistical expertise, but still must make recommendations or decisions based on statistics. This document guides the project manager in using appropriate statistical methods to address common project objectives and recognizing common misapplications of statistics. A typical application might include evaluating whether a groundwater remedy is functioning effectively or whether there is a downward trend which supports a natural attenuation remedy selection.

1.5 Project Life Cycle Stages

Groundwater statistical methods can apply throughout the life cycle of environmental cleanup projects, including monitoring of active remediation systems. The terminology and regulatory framework for the stages of the project within its life cycle, however, often vary under different regulatory programs. For clarity, this document organizes the discussion of site management around five main technical tasks:

- Section 4.2: Release Detection
- Section 4.3: Site Characterization

- Section 4.4: Remediation
- Section 4.5: Monitoring
- Section 4.6: Closure

These tasks correlate with the activities described in various regulatory programs (such as RCRA, CERCLA, State Voluntary Cleanup, and UST Site Cleanup). Although individual projects may vary in their progression through these stages, groundwater statistical tests can support decision making regardless of how the project is defined. Figure 1-1 summarizes the correlations between the terms used in this guidance document and the terms used in several regulatory programs.

Project Lifecycle Stage	Release Detection	Site Characterization	Remediation and Monitoring				Closure
			Remediation		Monitoring	Remediation	
CERCLA/ Superfund	Preliminary Assessment/Site Inspection	Remedial Investigation	Feasibility Study or EE/CA	ROD	Remedial Action	Response Complete	LTM/LTMgt/ LTMO
RCRA	Release Detection	RCRA Facility Investigation	Corrective Measures Study	RCRA Permit	Corrective Measures Implementation/ Compliance Monitoring	Certification of Remedy Completion or Construction Complete	Post Closure Care
UST/ LUST	Varies by Regulatory Authority						
State	Varies by State						

Figure 1-1. Correlation of regulatory terms.

Source: Adapted from ITRC RRM IBT slide 2011.

1.6 Study Questions

Each of the project life cycle stages listed above progress through development and refinement of the CSM. This guidance explores some of the commonly identified problem statements that guide decision making throughout environmental projects and poses a list of typical study questions that are intended to connect life-cycle-based issues of concern with relevant statistical methods.

Ten common study questions were selected:

- Study Question 1: What are the background concentrations?
- Study Question 2: Are concentrations greater than background concentrations?

- Study Question 3: Are concentrations above or below a criterion?
- Study Question 4: When will contaminant concentrations reach a criterion?
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?
- Study Question 8: How do contaminant concentrations change with distance from the source area?
- Study Question 9: Is the sampling frequency appropriate (temporal optimization)?
- Study Question 10: Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

Study questions are discussed in more detail in Appendix C. Life cycle stages and related study questions are discussed in Section 4.0. Statistical methods are presented in Section 5.0.

1.7 Document Organization

The information presented in this document is organized into the following sections:

- Section 2.0: Regulatory Framework and Challenges for Groundwater Statistics
- Section 3.0: General Statistical Approach
- Section 4.0: Statistical Analysis for Project Life Cycle Stages
- Section 5.0: Statistical Tests and Methods
- Section 6.0: Data Management Considerations
- Section 7.0: Public and Tribal Stakeholders Perspective
- Section 8.0: Summary and Conclusions
- Section 9.0: References

Additional appendices provide in-depth supporting information:

- Appendix A: Case Studies
- Appendix B: Common Misapplication of Statistics
- Appendix C: Study Questions
- Appendix D: Software Packages
- Appendix E: Survey Results
- Appendix F: Methods to Verify Underlying Assumptions for Tests
- Appendix G: GSMC Team Contacts
- Appendix H: Acronyms
- Appendix I: Glossary

2.0 REGULATORY FRAMEWORK AND CHALLENGES FOR GROUNDWATER STATISTICS

This section identifies various issues and concerns to consider as you plan use of statistics at a specific site. Presented here is an overview of state and federal regulatory requirements and example applications of statistical methods as they are used in the private sector. The section also introduces some general challenges in the use and application of statistics.

2.1 Regulatory Issues and Barriers

Inconsistent federal, state, and local regulations often complicate the evaluation of groundwater data using statistical methods. Existing regulations differ widely and may not reflect the "state of the science" or recognize the advantages and limitations of current statistical practices. The following summary of regulatory challenges is not an authoritative guide to the regulations, but rather an overview of the widely varying requirements.

2.1.1 State Regulatory Perspective

Based on the ITRC groundwater statistics survey (Appendix E), state guidance on the use of statistics ranged from "not well defined" to "overly prescriptive." Differences in using statistics exist among various programs within a state. The survey also revealed that some states are in the process of developing statistical guidance. In a few cases, specific programs reject statistical analysis outright.

While some states have adopted and been authorized to implement statistical practices recommended in the Unified Guidance for sites regulated under the Resource Conservation and Recovery Act (RCRA), these practices are not consistently carried through to other cleanup programs. State environmental cleanup programs may have requirements or guidance for statistical methods that pre-date the Unified Guidance and do not reflect contemporary practices. Because tools and approaches presented in this guidance document may differ from state and local guidance, regulators and the regulated community should closely coordinate planning and implementation of statistical evaluations.

The following are examples of questions which may be discussed between regulators and the regulated community:

- What are the state-specific requirements for the test or procedure that is planned to be used?
- Are there multiple programs within the state with differing requirements?
- Are there requirements that have been implemented on a local (county or municipal) basis?
- Are there any common procedures for which the state has no guidance or preference?

Program Guidance

The following topic areas of environmental project management were identified where existing program guidance is likely:

- *background evaluations*
- *treatment of nondetects*
- *comparison to criteria*
- *trend analyses*

In instances where statistical evaluations are discouraged or prohibited, or guidance is not consistent with current best practices, provide results from statistical methods presented in this document to introduce alternatives that may be considered.

A snapshot of the variability in the statistical requirements of state programs is provided in the groundwater statistics survey (Appendix E). Contact the appropriate state regulatory agency to discuss the requirements for your particular project.

2.1.2 Federal Regulatory Perspective

USEPA's Unified Guidance provides recommendations for statistical methods and strategies that can be used to demonstrate compliance with federal RCRA regulations for regulated units such as landfills and surface impoundments. Statistical evaluations under CERCLA can also follow USEPA's Unified Guidance. The data quality objectives using USEPA's seven-step data quality objective (DQO) process (2006a), and characteristics of the data sets (sample sizes and frequency of nondetects), usually drive the test selection. However, for the RCRA cleanup process and other remediation programs, statistical approaches from the earlier USEPA documents may be inconsistent with either the USEPA Unified Guidance or generally accepted current best practices for statistical evaluations. In addition, USEPA has a publication for monitored natural attenuation (MNA) that includes information about using statistics in groundwater evaluations for MNA sites (USEPA 2011).

USEPA's optimization strategy (USEPA 2007 and 2012) institutes changes to remedial programs to take advantage of newer tools that promote more effective and efficient cleanup, and to achieve verified protective cleanup faster, cleaner, greener, and cheaper. The strategy encourages use of techniques throughout the life cycle of site cleanup, acknowledging that optimization techniques and their use throughout the cleanup life cycle have become numerous and are growing rapidly. In the early investigation and design stages of the cleanup process, statistical and geostatistical methods can be used to optimize sampling design (as in incremental sampling), at the design stage (using value engineering techniques), and during the remedial action stage.

The Remedial System Evaluation can be used to examine a site holistically to determine whether the remedy is on track for cleanup, offering alternate approaches if the remedy is stalled, and to develop a completion strategy for the final disposition of the site. Specifically, quantitative analyses using statistical and geostatistical approaches are often used for optimization of groundwater monitoring programs. Recent green remediation advances have also been incorporated into each USEPA optimization technical support event. The benefits of optimization strategies include "more cost effective expenditures, lower energy use, reduced carbon footprint, improved remedy protectiveness, improved project and site decision making, and acceleration of project and site completion" (USEPA 2013f).

2.2 Private Sector Perspective

Groundwater statistics are typically used in the private sector to support environmental project management, risk assessment, and decision making. Some of the uses typically include the selection of groundwater sampling frequency, comparison of results of different methods of sample chemical analysis, demonstrating results of differing sample techniques are comparable, and identification of background concentrations. Once statistical analyses are performed, private sector entities must secure regulatory and stakeholder concurrence on acceptability of results.

2.2.1 Examples of Statistics Usage in the Private Sector

Some examples of how statistics have been used for various types or stages of projects in the private sector and potential outcomes and concerns are presented below. Please note that these are for illustrative purposes only to demonstrate concerns that Private Sector users have addressed or need to address, and are by no means exhaustive with respect to either site type or method used. The methods described in these examples are further described in Section 5.0.

Superfund site. Mann-Kendall trend analysis is routinely used to determine sampling frequency. For example, to select groundwater sampling frequency for a Superfund site, Mann-Kendall trend analysis was applied to an existing data set, which included seven years of semi-annual sampling results. The sampling frequency was back-tested by reducing the data frequency to one-half and then one-fourth of the original frequency. The results showed that the trend results were similar in all three data sets, leading to a significantly reduced sampling frequency for the long-term monitoring program. Less than 20% of the wells are sampled annually or biennially; most are sampled only once every five years.

Purge/no-purge sampling. Regression analysis is routinely used to compare the results of different methods of sample collection or chemical analysis. For example, nonlinear regression analysis has been used to demonstrate that the analytical results of different sampling techniques, such as no-purge and low-flow groundwater sampling, are comparable. Results are generally comparable, and the two methods yield the same conclusion regarding the attainment of a certain performance criterion for compliance. A related application of regression analysis is the calibration of one method to another (such as field and laboratory analytical methods) so that the results obtained with one technique may relate to the other using their empirical regression relationship.

Refinery site. Crude and refined petroleum products are complex mixtures that frequently contain hundreds of individual compounds. Multivariate statistical analyses may be used at some sites to either identify the general types of source petroleum products present at a site, or to differentiate and quantify the relative contribution of various sources in a co-mingled plume. It can be a difficult analysis and not all data sets will reveal clear contributions of the various sources. For light end mixtures, such as gasoline, it is possible to apply a combination of trend analysis and ratio analysis to determine if a new release has occurred where a historical spill was previously documented. The ratio analysis is a comparison of the relative concentration ratios among a group of chemicals measured in groundwater at the site and how those ratios might vary depending on location (for example, the relative concentration ratios may differ between a new release and a historical spill area).

Natural attenuation sites. Nonparametric trend analyses (such as the Mann-Kendall trend test, the seasonal Mann-Kendall test, and the Theil-Sen trend line) and parametric trend analyses (such as linear regression) are routinely used to evaluate trends in groundwater concentrations of contaminants over time for natural attenuation remedies. These trend analysis techniques are used to assign direction of trends (increasing, decreasing, or no trend) and the statistical significance of the trends. For decreasing contaminant concentration trends, the natural logarithms of the concentrations are plotted versus time and the linear regression analysis is conducted. The result is used to predict the time required for groundwater contaminant concentrations to meet remedial objectives (See also Appendix A, Example A.6). This information can be used to design appropriate MNA remedies, or to request closure based on demonstration of limited risk and an expected short time to meet remedial objectives.

Solid waste units. Nine closed RCRA and active National Pollutant Discharge Elimination (NPDES) solid Waste permitted units are being monitored on a semi-annual basis at an oil refinery site. Various site-related chemicals are being evaluated for statistically-significant exceedances using a combination of traditional parametric and nonparametric tests (based on data distribution analyses). Two recently updated Solid Waste permits include a requirement for statistical planning and evaluations. The most recent permit includes intrawell, rather than interwell, comparison to address different background and compliance wells.

2.3 Challenges for Project Managers

Statistical analysis of groundwater data, and other forms of environmental monitoring data, can present challenges during different project activities (for example, planning, implementation, data interpretation, decision making, and communication). The following key challenges are addressed in this document and the Unified Guidance.

2.3.1 Planning Challenges

These challenges are related to the conceptual understanding of statistical methods, selecting and applying methods to answer study questions, and satisfying the project's objectives:

- When is it advantageous to use statistical tests?
- Why do I need statistics for my "small" site?
- Do I have enough data to perform a statistical analysis?
- How will the results of statistical tests help users make decisions?
- What are the purposes and limitations of statistical tests?
    - How can I estimate the value of using statistical tests during the various stages of a project's life cycle?
    - How many samples do I need to collect?
    - How do I balance the statistical certainty I need with the cost of data acquisition?
    - What are the best ways to explain the statistical approach to all involved – regulators, consultants, owner representatives and community members?
- How should historical data be optimally processed?
    - How to plan sampling to achieve data quality objectives?
    - What is the statistical parameter (for instance, mean or median) of interest?
    - What is the variability of the statistical parameter or trait that will be measured?
    - What are the acceptable false positive and false negative rates?
    - What statistical tests are appropriate?
    - What hypotheses need to be tested?
    - What is my alternate hypothesis?
    - Is it important to identify a directional or non-direction change (such as plume movement)?

2.3.2 Interpretation and Communication Challenges

These challenges are related to interpreting and communicating the results of statistical tests and evaluations to multiple audiences:

- How do I judge data adequacy and convey confidence in the data?
- How do I use graphics effectively and fairly? How do I spot misleading graphics?
- How can I explain the statistical tests and results to stakeholders?

2.3.3 Common Misapplications

Just as the application of statistics to groundwater and environmental monitoring presents many challenges, it also affords many opportunities for misuse, misapplication, and misinterpretation. The following list summarizes common misapplications. A more thorough discussion is provided in Appendix B, along with suggestions on how to best address each issue.

- ***Misapplication:*** *The site maximum sample is less than the risk-based decision criterion, therefore I can conservatively assume the site is low risk.*

  Response: If the number of samples is small, this assumption may be wrong and the population mean may actually be above the decision criterion.

- ***Misapplication:*** *The site maximum sample is greater than a "background" maximum or mean, therefore the site is contaminated.*

  Response: The data sets may be very different in size and in what they represent, so that comparisons may not be meaningful or appropriate. Statistical approaches such as the use of two-sample tests or background upper prediction limit (UPL) would be better alternatives.

- ***Misapplication:*** *The concentration this round is less than the last round and the round before that, therefore there is a decreasing trend.*

  Response: Assess trends over multiple rounds and consider the variability of the data. A few monitoring events are typically not adequate to characterize the variability and assess trends.

- ***Misapplication:*** *Three rounds of data below the decision criterion means a site is in compliance.*

  Response: An arbitrary rule or number of rounds may not adequately address the variability of the data being assessed.

- ***Misapplication:*** *If I have good correlation between my field screening data and laboratory analyzed splits, then I should be able to use the field data to make comparisons to the decision criteria.*

  Response: You must assess the variability of the responses (field technique to valid laboratory data) and quantify the potential error and confidence in the field results.

- ***Misapplication:*** *If the statistics indicate a result is an outlier, then it is acceptable to disregard that point.*

  Response: You must always have a well-documented, weight-of-evidence reason to eliminate data. Also evaluate the effect that eliminating data points has on the conclusions drawn from the remaining data. Often, outliers may be representative of unforeseen, yet important, site characteristics.

- ***Misapplication:*** *There is a statistically significant difference in the data, therefore it must be important.*

  Response: When data sets are large (such as for wells that have been monitored for long periods of time), statistical tests can be used to identify small changes in data trends. However, you must also note the magnitude of the change relative to the decision criteria.

- ***Misapplication:*** *The statistics "prove" the research hypothesis was correct.*

  Response: Statistics is the science of probabilities and the imprecise, therefore it is essential to state the degree of confidence the statistics provide for the conclusion.

Please refer to Appendix B for more common misapplications as well as suggestions for addressing these important issues.

2.3.4 Implementation Challenges

These challenges are related to selecting and using software, enhancing confidence in calculations, and gaining consensus on results:

- How do I select an appropriate statistical package to independently verify results and conclusions?
    - Is the selected software package accessible and user friendly?
    - What is its cost?
    - Are there upload/download security issues?
    - Who maintains the software and provides support?
- How do I overcome data entry difficulties resulting from software program structure?
- Does the software produce output with adequate explanations of intermediate results?
- Does the software produce presentation quality graphics with minimal effort?
- How does the user check the results and what should users check?
- What types of data need to be processed?
    - Transformed data (such as logarithms of the original results)
    - Censored data (such as nondetects)
- How, when, and why do I modify a data set before analysis?

3.0 GENERAL STATISTICAL APPROACH

This guidance takes a broad view of groundwater monitoring and compliance. Not every site undergoes the same project life cycle stages (see Section 1.5) or is governed by the same regulations, but groundwater monitoring at every site provides data for statistical analysis that can help support decision making. Some of the statistical approaches in this document might only be applied at larger sites with extensive data sets. Others can be used at even the smallest of sites, assuming that a reasonable minimum number of measurements are collected.

Throughout the project life cycle, systematic planning should form the basis for collection and analysis of groundwater data. One of the first steps for any site is to establish a working conceptual site model (CSM). The CSM is updated during the project as new information is gathered. In addition, the project planning team defines the data quality objectives (DQOs) and then determines the appropriate type and quality of data needed to answer questions of interest. From a statistical standpoint, exploratory data analysis (EDA) should generally be used to review data quality and select appropriate statistical methods.

Systematic planning results in clear data collection plans and objectives. Since appropriate and usable data are necessary for statistical analysis, this document generally assumes that groundwater data have been collected using a systematic planning process. The USEPA DQO process and the U.S. Army Corps of Engineers (USACE) technical project planning (TPP) process (USACE 1998) are two examples of systematic planning that can readily be used to help plan groundwater data collection. Additional information on systematic planning can be obtained from the following ITRC documents:

- RPO-7: Improving Environmental Site Remediation through Performance-Based Environmental Management (ITRC 2007b)
- SCM-1: Technical and Regulatory Guidance for the Triad Approach: A New Paradigm for Environmental Project Management (ITRC 2003)
- SCM-3: Triad Implementation Guide (ITRC 2007a)

This section describes a general approach for groundwater statistical evaluations, with an emphasis on CSM development and refinement, EDA techniques, statistical design, and the key assumptions common to groundwater statistics. This section also outlines steps used to assist in choosing an appropriate statistical method, along with options for data that have been not been collected systematically.

While statistics provides a quantitative basis for decision making, do not rely on statistics to the exclusion of other lines of evidence or to compensate for a poorly designed monitoring program. While this document focuses on the statistical analysis of groundwater measurement data, other critical lines of evidence may include related soils data, site history, soil gas measurements, ground-

water flow dynamics, lithology information, and well logs. A scientifically defensible and correct decision will often require multiple lines of evidence in addition to statistics.

3.1 Introduction to Conceptual Site Models

Developing a statistical approach based on a CSM is an initial investment that can save significant time and money and prevent poor decisions. The site CSM should be developed before deciding on the statistical methods to be used. A CSM is developed using site information such as information about sources, geology, hydrogeology, land use, and soil and groundwater data generated at sampling points from different locations on a site. Groundwater sampling points (or sampling locations) are most often monitoring wells, but could be other types of sampling such as direct push sampling, temporary probes, or field sensors. Locations may be upgradient of a release and plume, downgradient, side gradient, or within a groundwater plume. For some sites it is important to determine whether intrawell or interwell statistical testing will be useful. An example is presented below to illustrate comparing intrawell and interwell statistical testing.

An example in which knowledge of the working CSM can help is the choice between intrawell and interwell statistical testing. Traditional interwell tests compare upgradient background data with downgradient compliance well measurements. Groundwater compliance is then assessed by whether the downgradient values exceed background. At many sites, however, one or more of the monitored parameters occurs naturally in groundwater and varies substantially across the site due to natural geochemical factors (thus exhibiting natural spatial variability). At these sites, parameter concentrations larger than upgradient background might be attributed to contamination when the differences are actually natural and due to the locally-varying distribution of groundwater chemicals.

A statistical approach that first checks for statistically measurable spatial variability and, if present and natural in origin, then uses intrawell testing at each compliance well instead of interwell comparisons will likely avoid misleading conclusions. Intrawell testing compares earlier versus more recent data at the same sampling point. Because the comparison is made at a single sampling point, concentration differences between wells due to natural spatial factors do not affect intrawell tests. Only changes over time (indicating a trend or shift in concentration level) cause an intrawell test to be statistically significant and to show a change in groundwater quality.

Note above the importance of the qualifier ‘natural in origin’ when characterizing spatial variability. Compliance wells situated in the middle of a plume vary spatially from upgradient background wells, but in that case the variation is anthropogenic and indicative of contamination. Intrawell testing in those circumstances might not be helpful and could even obscure evidence of the plume. Note also that while natural spatial variation is a likely characteristic of most sites, it may not always be easy to identify. The variation may be real but low-level, in which case intrawell testing may be unnecessary. Or, the variation may be stronger but difficult to observe due to a small sample of measurements or a small number of sampling points.

Example: Natural Spatial Variability

Figure 3-1 below describes a landfill located just north of a coastal river. Measurements of specific conductance at the site consistently showed much higher (and statistically significant) readings at all three downgradient wells compared to upgradient background. Further investigation found, however, that the higher compliance point values of specific conductance were due to natural infiltration of salt water into the downgradient wells, due to tidal fluctuations where the river met the sea, and not groundwater contamination by the landfill. Only a refinement of the initial CSM captured this important feature of the groundwater system and allowed the cause of the spatial variation to be identified.

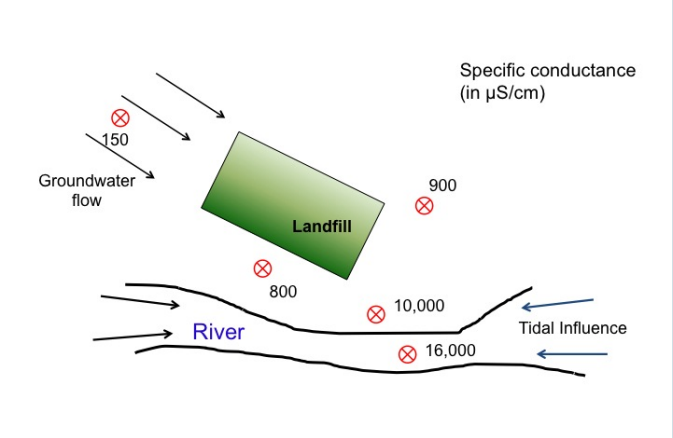


Figure 3-1. Landfill CSM.

Source: Adapted from USEPA 2009.

*The example site shown in Figure 3-1 illustrates the importance of developing a CSM **prior to** statistical evaluations and of refining the CSM on a periodic basis. The site-specific conditions and hydrogeology must be understood well enough to select a proper statistical approach, part of which will involve answering the following questions:*

- *Has background been selected from the right locations and is it statistically representative of local background conditions?*
- *Are the parameters of concern likely to be normally distributed?*
- *Will specific approaches or methods be needed to account for frequent nondetects?*
- *Are historical data representative of current groundwater quality, or have local conditions changed over time?*
- *Are enough data either available or planned for collection to enable accurate and statistically powerful testing and decision-making?*

3.2 Developing a Conceptual Site Model

An initial CSM is essential to formulating a statistical approach as well as to deciding which data and analyses are appropriate for the current stage of the project.

3.2.1 Target Population

For statistical purposes, knowledge about a site's hydrogeology and its CSM is critical for determining the nature and stability of the target population of groundwater measurements. In a highly stable, homogeneous, sandy geologic environment, groundwater concentrations may be fairly consistent over time. In a highly fractured or karst environment, significant discontinuities may exist in concentrations, even at nearby wells. For all sites, changes over time in regional conditions (such as a multi-year drought) may cause groundwater concentrations to change so much that past data may not be similar to more recent measurements. In that case, more than one target population may exist, with the newer population no longer being statistically the same as the older population, even though collected from the same site.

Understanding Target Populations

Perhaps the best way to make sense of sampling data is to understand the target population of measurements from which those data are drawn. For instance, an aquifer system may be conceptualized as a complex, dynamic four-dimensional object, with three dimensions representing groundwater subsurface volume over a prescribed boundary and depth, and one representing time. Physical groundwater samples are collected at specific locations and depths within the three-dimensional volume, but also at certain points in time. The target population associated with a given set of measurements could represent the entire history of the aquifer, but more commonly the goal is to say something about a specific time period or a specific hydrostratigraphic unit or layer, for instance, shallow zone circa 2013 or the local aquifer surrounding well A-1 over the past two years. Any statistical conclusion (or inference) drawn from the data only applies to the target population, so defining or understanding that target is of prime importance.

Since a CSM treats the subsurface as a dynamic, four-dimensional object, the best statistics will result from a clear understanding of the target population(s) of measured values. The target population refers to the entirety of the parameter of interest (such as the pH of an aquifer or the range of trichloroethylene concentrations within a plume). In statistics, 'population' is the total amount of a property and is generally defined by both spatial and temporal boundaries. Project managers should try to assess whether there is one homogenous population, many distinct localized populations, different populations by subsurface depth, populations that change over time, or some combination of these conditions. The relevant populations of measured values and groundwater flow at the site will affect sample point placement, sampling frequency, background data definition, and the

number of sample values and sampling points needed for reliable characterization.

Environmental populations generally cannot be fully characterized (that is, by analyzing every portion of soil at a waste site or all possible volumes of groundwater in an aquifer), so a statistical sample is drawn to represent a population. For example, it is possible to measure the concentration of a contaminant at a finite series of sampling events (for example, quarterly) over years, but it may not be practical to collect and analyze a continuous stream of samples over the same time frame. A representative sample is one with key statistical characteristics that parallel the characteristics of the target population.

Any decision made regarding the target population should be based on sample data collected from that target. Likewise, any statistical inference applies only to the target population and not necessarily to other populations. For instance, arsenic concentrations that exceed a regulatory limit during one year at one specific sampling point do not imply that arsenic concentrations will necessarily exceed the limit in successive years, or that arsenic concentrations exceed the limit at other well locations.

Defining the target groundwater population involves at least two tasks. The first task is to delineate the area of concern and note any clearly defined hydraulic boundaries or concentration change points that make for logical bounds. Consider the size of this area: too large an area may result in 'watering down' the parameter of interest (for example, the site-wide mean concentration), while too small an area may miss areas of potential contamination.

The second task is to define the temporal extent. What period of time is of interest? Is this a one-time evaluation or will the assessment involve multiple sampling events over time? If the groundwater system under study is highly dynamic, consider the age of existing data, since it may no longer represent current conditions. More frequent sampling may be required in areas where concentrations of contaminants change more rapidly or fluctuate on a seasonal basis, unless the variability of sample results is low relative to criteria.

Two important, often related questions in defining target populations are (1) is there a single target population or perhaps several? and (2) what is the decision support required of the existing or planned sampling data? Even when areal and temporal extents have been determined, a dynamic four-dimensional subsurface (time plus volume) may more appropriately be regarded as a series of distinct populations, including perhaps:

- separate aquifer units
- distinct hydrostratigraphic layers
- highly localized geochemistry, leading to substantial spatial variability among well locations and a separate target population per well

In this setting, consider what kinds of decisions or inferences the available data allow. In a dynamic environment with substantial natural spatial variation, one sample at a single well or sampling point could never spatially characterize the site as a whole nor the local groundwater population in the

well's zone of influence, since that one measurement would only provide a 'snapshot' of groundwater quality and give no information about temporal changes or trends. Even a larger number of measurements may not provide adequate information if coverage of the spatial and temporal extents of the target population are not taken into account.

If the existing data provide limited coverage (either spatial or temporal) then additional sampling is likely required. Spatial coverage is limited if not all hydrological zones have been characterized. Contaminant concentrations that vary substantially between locations signify high spatial variability and may require greater sample density. Temporal coverage may be insufficient if very few measurements are collected during the time interval of interest at a given sampling point, especially if possible trends cannot be captured or estimated accurately within the data record.

The target population may be defined in part by regulation. In some cases, the critical concern may be at the point of compliance, perhaps along the downgradient boundary of the site. Concentrations at other portions of the site may or may not matter from a regulatory viewpoint, as long as measurements collected from the point of compliance do not exceed applicable criteria. Even in such cases, however, it is usually difficult to estimate or predict changes along the compliance boundary unless measurements are also collected from other portions of the site. These measurements may be collected if for no other purpose than to establish and check the CSM. The effective target population is rarely confined strictly to the point of compliance.

3.2.2 Background Concentrations

For compliance purposes, project managers must determine what portion of the subsurface (regarded in four dimensions) adequately represents background concentrations in order to answer the following questions:

- Can the background population be unambiguously defined and sampled?
- Does background change over time?
- Is the local background population intermixed with potentially contaminated groundwater?
- Are groundwater gradients and flow paths consistent enough to ensure reliable monitoring of background conditions into the future?
- Are there multiple sampling points/wells dedicated to establishing and monitoring background levels?

Answers to these questions will facilitate good decisions as the project progresses and can ultimately reduce costs and avoid delays.

Given the dynamic nature of the subsurface, measuring background at a single point in space and time is generally inadequate. Background measured at a single time gives no indication of whether background conditions might change in the future. A single background sampling point confounds spatial variability and actual contamination. Multiple background sampling points allow for (1) assessment of the presence of significant spatial variation; (2) faster accumulation of adequate background data for statistical purposes; and (3) a better understanding of the uncontaminated subsurface.

See Appendix A Example, A.1 Comparing Two Data Sets Using Two-sample Testing Methods

See Appendix A Example, A.3 Calculating Prediction Limits

3.2.3 Multiple Source Areas

A good CSM is critical to statistical evaluations of overlapping plumes or multiple contaminant source areas. Questions relevant to these situations include:

- Must each plume be separately and independently assessed for compliance?
- Are the contaminants common to overlapping plumes?
- How will the sampling data be used to measure the relative impact and extent of each plume?
- Where should sampling points be placed or adjusted to optimize plume characterization?
- Are the contaminants from each source distinct enough (may apply only at some sites) to allow for a multivariate statistical 'fingerprinting' of each plume?
- How much sampling data are needed to adequately characterize the different sources?
- Are there differing aquifer zones and concentration profiles that vary with depth and must be analyzed separately for statistical purposes?
- Are the plumes in differing states of recharge or discharge? How will this impact what data must be gathered?

Again, a statistical approach based on the CSM and the answers it provides helps to ensure that the collected data are useful for making compliance decisions.

3.2.4 Monitored Natural Attenuation

A sound statistical approach can also help support the common remedy of monitored natural attenuation (MNA). Groundwater monitoring data at a sampling point can be initially tested for a statistically significant trend to determine whether a MNA remedy is or may be effective. However, consistent groundwater flow paths are essential, as are monitoring wells that accurately capture those paths. Later the groundwater monitoring data at a sampling point may be tested for a stabilizing trend. The amount of groundwater data needed will depend on the level of statistical

See Appendix A Example, A.2 Testing a Data Set for Trends Over Time

confidence required for detecting temporal trends and for deciding whether concentrations are projected to remain below criteria. Additionally, if monitoring is scheduled to continue indefinitely, the sampling frequency can be optimized statistically, but this will again require input from the CSM as well as the regulatory drivers governing the remedy.

3.3 Understanding the Data

Before conducting formal statistical evaluations, review the data. This review should include (1) reviewing data quality, (2) assessing the extent and usefulness of any historical data, and (3) exploring the data for general patterns and characteristics. One general way to aid in this understanding is through a collection of numerical and graphical statistical techniques known as exploratory data analysis (EDA, see Section 3.3.3). EDA can help to identify any data quality problems (such as anomalies or inconsistencies) as well as basic attributes of the data, such as its shape, spread (for example, standard deviation), and central tendency (for example, mean, median).

3.3.1 Data Quality

Site data must be of sufficient quality to be statistically usable. Among the questions that must be answered in order to assess data quality include:

- Are quantitation limits low enough to determine whether criteria have been exceeded?
- Are there outliers (that is values unrepresentative of the overall population of groundwater measurements) that might falsely imply detection of a release?
- Are quantitation limits consistent over time, or does measurement precision vary, perhaps associated with changes in analytical methods or sample interferences or dilutions?
- Are nondetect data reported to quantitation limits or detection limits, and is there an understanding of the difference in measurement uncertainty depending on which reporting method was used?
- Are measurements collected frequently enough to accurately characterize groundwater elevations and whether those elevations change over time? See USGS guidance (USGS 2013) on components of water-level monitoring programs.

One broad-based approach for acquiring and assessing environmental data is USEPA's seven-step DQO process (USEPA 2006a). The DQO process provides basic guidance on systematic planning, develops performance or acceptance criteria, and identifies resources and references for this process. It can be helpful to review the DQO process at each stage of a groundwater investigation or remedial effort. The DQO steps shown in Figure 2 of the USEPA guidance include the following:

1. State the problem.
2. Identify the goals of the study.
3. Identify information inputs.
4. Define the boundaries of the study.
5. Develop the analytic approach.

6. Specify performance or acceptance criteria.
7. Develop the plan for obtaining data.

USEPA's DQO process is general enough to potentially incorporate different lines and types of data-based evidence. Statistics is one useful tool in this framework, as highlighted by USEPA's Data Quality Assessment (DQA) (USEPA 2006b). The DQA process evaluates whether the level of data quality will enable the DQOs to be achieved. This latter, more specifically statistical, process consists of a series of complementary steps. Table 3-1 includes these steps (see Figure 11 of USEPA's [2006a] document) and illustrates example tasks for the steps.

Table 3-1. DQA steps

DQA Steps	Example Tasks
1. Review DQOs and sampling design.	Goal: estimate plume contaminant mass within 10% relative error.
2. Revisit DQOs if necessary.	Check sample design to see if spatial grid of locations is feasible
3. Conduct preliminary data review.	• Review quality assurance reports: Does available information support conversion of concentration data to mass estimates? • Calculate statistical quantities: Compute weighted mean and variance estimates of total contaminant mass. • Display the data graphically: Plot map of concentration estimates; has extent of plume been delineated by existing wells?
4. Select the statistical test.	95% confidence interval for total contaminant mass
5. Verify the assumptions.	Check normality of sample data; use non-parametric test if data cannot be normalized.
6. Draw conclusions from the data.	Estimate total contaminant mass with 95% statistical confidence.

Based on these steps, application of statistics should incorporate an iterative approach, including:

- up-front exploratory data analysis Section 3.3.3 to better understand the data set, its usability, and its representativeness
- a clear formulation of the study questions and the statistical inferences that need to be made
- selection of the appropriate target population (see Section 3.2.1) from which data will be drawn
- data quality assurance and quality control (QA/QC)—do the data meet required or appropriate QA/QC requirements?
- application of appropriate statistical methods, checks on the assumptions of those methods, and an assessment that reasonable answers have been obtained

Even with systematic planning, uncertainty is inherent in all scientific measurement. The level of uncertainty in a data set, however, must be low enough to answer the study questions with

sufficient statistical confidence. In some cases uncertainties can be addressed by collecting additional data or using more sensitive analytical methods. In other cases uncertainty reflects a basic lack of knowledge about how the natural system functions. Identifying and managing uncertainty (ITRC 2011b) supports informed decisions in all stages of the project life cycle.

While not the focus of this document, standard practices can help achieve and maintain appropriate data quality. These practices include collecting field duplicates, maintaining the chain of custody, and implementing good analytical practices such as laboratory replicates, spiked samples, and standard solutions. Even with these practices, limits to the precision of laboratory instruments will exist because of low signal-to-noise ratios at very low concentrations.

For many sampling methods, simple modifications to the current sample collection procedures can serve to reduce monitoring variability. These modifications can reduce variability by directly addressing some sources of variation such as in-well stratification of contaminant concentrations and by mitigating the impact of other sources of variability by minimizing differences in sample collection procedures between sampling events. For no-purge sampling methods, it can be important to consider seasonal changes in vertical temperature gradients when comparing samples from different times (McHugh et al. 2011). For low-flow or no-purge sampling methods, variability can be reduced by collecting the samples from exactly the same depth within the well (high precision sampler placement). For sampling methods that require transfer of the sample from the collection device to the sample container, specific bottom-fill transfer procedures will reduce variability associated with volatile loss (Parker and Britt 2012).

Example: Variability and Trend Analysis

Groundwater monitoring data are often affected by high levels of variability unrelated to the long-term temporal trend (McHugh et al. 2011). For example, poor data quality or precision may result in multiple nondetect or tied values that can lessen the ability of statistical tests to correctly identify trends (what is known as statistical power; see Section 3.6.1.2). A large number of ties or nondetects may obscure the distribution of the data and limit the selection of statistical methods that can be used. These data may also prevent estimates of temporal autocorrelation (see Section 3.4.4).

Selection of the time period to include in a trend analysis is also a trade-off between statistical power and interpretation of the results. A long time period can be evaluated as a whole or as multiple smaller data sets covering shorter time intervals. Use of shorter time intervals may be necessary to evaluate changes in attenuation rates (for example before versus after installation of an active remediation system). A single, larger data set (one that covers a longer time period) will have greater statistical power and is more likely to identify an actual trend with less variability associated with the estimated attenuation rate.

3.3.2 Historical Data

Even if not collected using systematic planning, historical data may be useful for statistical and

compliance purposes, depending on data quality and comparability with more recent measurements. Pre-existing data can be examined for general trends over time and to assess whether background concentrations are relatively stable or whether they are inconsistent with past data. Such exploratory comparisons may shed light on hydrogeologic changes, data anomalies, or other patterns, and often provide a longer-term perspective of the site.

A large amount of data is not the same as a large amount of *statistically usable* data. To be statistically usable, all the data points meant to represent a particular target population must have been drawn from that population using a similar, if not identical, collection and measuring process. Historical data collection and analysis may not be consistent with current methods. Sampling and analysis in the past may have been different enough from the present to bias the older values in one direction or another, or to introduce unacceptable levels of uncertainty. Local groundwater conditions may also have changed to such an extent that the data are no longer physically representative of current conditions of interest.

Not all historical data are useful for formal statistical analysis or even EDA. Quantitation or detection limits change with different laboratories, with different methods, and with improved laboratory techniques, potentially making comparison of data collected over time difficult. A common complication is the lowering of quantitation or detection limits as technology improves, resulting in poor understanding of low concentration levels early in the data record. This progressive lowering of quantitation or detection limits can mistakenly appear to be a decreasing concentration trend in a time series plot if the nondetects have been replaced with some fraction of the quantitation or detection limit (for example, historically one-half the detection limit has been used). Be familiar with changes in laboratory and data collection methods over time when using historical data.

Historical data that were not collected as part of the current systematic planning process may be valuable during the exploratory phase and for informing or checking the preliminary CSM. However, these data may not have sufficient or comparable quality to be used in a formal analysis or to assess regulatory compliance. EDA (see Section 3.3.3) can be helpful in comparing newer data against older data and in establishing which time period of data collection best represents relevant groundwater conditions and offers sufficient data quality.

3.3.3 Exploratory Data Analysis

EDA refers to a collection of mostly informal, descriptive and graphical statistical tools used to explore and understand a data set. Generally, EDA includes numerical summary statistics such as measures of centrality (for example, mean, median), measures of spread (for example, standard deviation, variance, interquartile range), and measures of shape (for example, skewness and kurtosis), as well as graphical displays such as histograms, box plots, scatter plots, time series plots, and probability plots. Section 3.5 includes information on how to use EDA to test statistical assumptions.

EDA methods allow you to check data quality and select appropriate statistical methods. EDA methods can also confirm whether or not the underlying assumptions of statistical methods are met.

For example, all parametric statistical tests assume that the data are drawn from a particular probability distribution, whether the normal, lognormal, gamma, or some other known statistical model (see Section 5.6). An initial assessment using EDA can help determine whether or not the measurements approximate such a theoretical population. On the other hand, EDA is not designed to confirm groundwater contamination or to measure remedial success. EDA can test and check assumptions and tentatively identify important changes or patterns, but confirmation of those changes or patterns is best done with formal inferential tests.

The typical objectives for EDA are listed in Table 3-2.

Table 3-2. Typical objectives for EDA

Objective	Example Tasks
Provide insight into a data set and check data quality.	Check for high fractions of nondetects or large field replicate variation, or both.
Uncover underlying structures.	Identify increasing trends on time series plots.
Extract important variables.	Calculate which chemicals frequently exceed criteria.
Detect outliers and anomalies.	Flag possible misreported values using box plots.
Test underlying assumptions.	Check normality with probability plots.
Qualitatively identify trends, relationships.	Assess correlation between chemicals using scatter plots.

For more information see Chapter 9, Unified Guidance, the NIST Engineering Statistics Handbook (2012), or Tukey 1977.

Graphical methods provide a critical overview of a data set. Histograms and probability plots are visualizations of the data shape that can help identify the best-fitting probability distribution, such as the normal or lognormal. Box plots graphically identify data characteristics such as the median, interquartile range (the measurement difference between the 25th and 75th percentiles; the latter percentiles are also known respectively as the lower and upper quartiles), range, and possible outliers. Scatter plots and time series plots can identify temporal trends and correlations.

EDA can also provide qualitative spatial analysis by plotting data on maps and observing spatial patterns. Such patterns are often enhanced by contouring or color-coding points on a map. Accurate spatial analysis generally requires a large number of sampling points, spread out to give good spatial coverage of the site. Although many software packages perform contouring, these packages may perform poorly if the data set is sparse (which is typically the case for corrective action sites; see, for instance, Siegel 2008). If a software package is used for contouring, you should carefully review the results for interpolation and extrapolation errors.

Changes in groundwater quality or remedy effectiveness can be qualitatively evaluated by plotting temporal trends on a map (also known as a trend map) and identifying any apparent spatial patterns. Typically, a symbol or color or both are used at each sampling point to represent the nature

and strength of the trend at that location (for instance, significantly decreasing). An alternative is to create a series of maps, each representing a particular time period, to evaluate changes in spatial patterns over time.

Project managers who lack expertise in statistics sometimes avoid an initial EDA to save time or money, especially when data exploration may not appear to be linked to specific compliance-related decisions. This practice is a false economy. Not only is EDA critical to properly navigating a systematic planning process and obtaining sufficient quantity and quality of data, it also helps avoid unnecessary or inappropriate statistical tests. As discussed in this document, all statistical procedures make assumptions about the nature of the data and the population from which those data have been collected. EDA helps to check these assumptions and select appropriate tests. For instance, a simple t-test to compare two groups assumes that the two populations are normally distributed. If this assumption is not checked (for example, by using probability plots or normality tests), an incorrect decision may be made, especially if the data sets are highly skewed or contain many nondetects.

EDA is also critical for examining data quality and checking for data anomalies and comparability. For example, visual examination of a parallel time series plot may suggest that all the measurements from a given sampling date (across wells and contaminants) are outliers. Such anomalous patterns can indicate laboratory or field sample collection problems that might arise from instrument miscalibration or perhaps sample mislabeling or mishandling. These outlier values should usually be deleted from statistical analysis since (1) they do not represent true water quality and (2) the cause of the aberration is known.

More generally, EDA can identify data quality issues; it can be used to determine whether site data require special statistical adjustments or if data quality is inadequate to make reliable decisions. Data sets with frequent nondetects often fall into the first category, while incomplete data (data sets with missing measurements) or data with elevated reporting limits (such as that arising from high dilution factors during chemical analysis) may fall into the second. Data sets that are characterized as 'completely usable' after laboratory QA/QC and data validation checks often contain significant anomalies and inconsistencies that are only identified after EDA. These cases show that EDA is an investment of project resources that can yield significant dividends.

3.4 Common Statistical Assumptions

Many assumptions are made during a groundwater investigation or in the course of long-term monitoring and compliance. This document focuses only on assumptions relevant to groundwater statistics and also assumes that the general principles of a systematic planning process have been followed during data collection and analysis, and the data are generally appropriate for the intended use (except perhaps for historical data).

Since parts of the systematic planning processes are statistical and iterative, exploratory statistical methods may be needed to ensure adequate data quality and quantity (see Section 3.3.3). Furthermore, statistical design considerations also inform systematic planning, so efforts to engage the

systematic planning processes are integrated with the statistics discussed here. Nevertheless, the primary discussion assumes that data have already been deemed usable for statistical purposes. See also Section 2.1 for more planning considerations.

For sites with no existing data, follow a systematic planning process to ensure that planned measurements have sufficient analytical precision, that the questions of interest are clearly defined, and that sufficient observations will be collected from a well-defined target population. For sites with historical data, as discussed earlier, those data should be examined prior to formal testing to determine whether they are usable (see Section 3.3.2).

Perhaps the most important assumption is that sufficient data exist to conduct a valid statistical analysis. All statistical tests assume measurements are drawn from a larger (often unseen or unobservable) target population of potentially measurable values. The conclusion from a formal statistical test reflects an inference *from* the sample values *to* the larger population and makes a statement about that population as a whole. To make such an inference (for instance, to estimate a characteristic of the population like the overall mean concentration) within a specified level of accuracy, a minimum number of measurements, termed the sample size, is needed. Sufficient sample size varies by statistical method and depends also on the level of desired statistical certainty or accuracy. Information regarding sample size is presented for the methods in Section 5.0.

3.4.1 Nonrandom Sampling Points and Sampling Times

An independent or random sample can be representative of the target population and its variance, and is useful for formal statistical inference. For groundwater, however, the subsurface target population may or may not be well mixed. Though dynamic and four-dimensional (time plus three-dimensional volume), the degree of natural 'mixing' will depend on multiple complex factors, including but not limited to flow rates, soil or rock composition, porosity, aquitards and hydraulic barriers, recharge rates, and the types and nature of the contaminants being monitored.

Combined with a population that may not be well mixed, groundwater sampling of the subsurface is generally nonrandom. Usually, it is not possible—either logistically, physically, or conceptually—to sample the subsurface at random locations and at random times. Sampling points (for example, groundwater wells) are at fixed locations and sampling teams must go out to the field at preset and logistically convenient times. The CSM may also dictate general rules for sampling point locations, usually based on professional judgment.

If there is a high degree of natural mixing and homogeneity within the subsurface, over a given time period, it should not be necessary to randomize the sampling points or times of sampling. Similar statistical results should be obtained at any sampling point and the combined data should approximate an independent sample from the target population. More often, groundwater plumes have a distinctive spatial and temporal footprint, meaning that concentrations vary substantially by location and time of sampling. In these cases, the subsurface population is not naturally well mixed (randomized), and—since the sampling process itself is nonrandom—it may not be possible to treat the data as if it represents an independent subset of the target population. That is, it may not be

valid to simply pool values across different sampling points and ignore the possibility of spatial correlation between different wells, or to ignore the possibility of temporal correlation among a series of samples from a single sampling point.

Examples where correlation can be problematic include (1) a well with regular seasonal fluctuations that is only sampled during the 'peak-concentration' summer months; or (2) sampling two spatially-correlated wells near the source of a plume and then assuming they accurately reflect the magnitude of the remaining plume area.Special geostatistical techniques such as kriging (see Section 5.14.2) have been developed to perform spatial analyses in the presence of significant spatial correlation. Unfortunately, an accurate kriging analysis generally requires a larger number of sampling points, often laid out on a systematic sampling grid, so this may be difficult at some sites.

It also may be possible to account or adjust for temporal correlation between adjacent sampling events at a single sampling point, especially if the lag time between measurements at the same location is small (for example, monthly or less), and the data set consists of a longer series of values. It is also important that enough different times of the year are sampled so that seasonal patterns are not missed. Again, a larger amount of data is usually required to both identify the presence of significant temporal or spatial correlation, and then to adjust the data so as to minimize any adverse statistical impact of that correlation.

In general, an independent sample (or one that has been adjusted for the presence of correlation) is important because:

- All standard statistical tests assume that the input data (if otherwise unadjusted) have been independently drawn from an underlying groundwater population of possible measurements (the target population).
- Since only a small fraction of any subsurface population can be observed, an independent sample ensures that all of the population has a chance to be selected and measured. By contrast, nonrandomized samples from a poorly mixed population tend to be biased and unrepresentative of the underlying target, possibly 'missing' important features of the measurement distribution.

3.4.2 Nondetects and Uncertain Measurements

Many chemical contaminants occur in very small concentrations or can be difficult to measure. This situations leads to many nondetects or to measurements with high degrees of unknown analytical uncertainty (for instance "J-flagged values" that are less than laboratory quantitation limits). Nondetects or "less than" values are technically known as "left-censored" values. Data censoring complicates statistical evaluations, especially when a large portion of a data set is nondetect. Uncertain measurements lead to data sets with varying analytical precision, also complicating statistical analysis. Although it may require more mathematically involved adjustments, the impact of nondetects (see Section 5.7) and uncertain measurements should be considered in statistical tests. Failing to utilize such observations can severely bias statistical estimates; adjusting for nondetects in the wrong way can also negatively impact the analysis.

3.4.3 Normality

Standard parametric statistical tests assume that the sample data are either normally distributed or follow another known statistical model (such as a lognormal or Weibull distribution). Many environmental and groundwater data sets are either nonnormal (skewed, lognormal, gamma) or contain too many nondetects to accurately check the normality assumption. In some cases, sample data can be normalized by mathematical transformation, for instance, by taking logarithms or square roots of the original results. Nonparametric statistical tests may be used when data do not seem to fit any known distribution. Either way, selection of an appropriate statistical method usually requires an initial check for normality (see Section 5.6).

3.4.4 Temporal Independence

As noted in Section 3.4.1, each measurement drawn from a population of groundwater measurements is assumed to be statistically independent of every other measurement. What this means statistically is that the occurrence of an event (sample value) makes it neither more nor less likely that a second event (sample value) occurs. Practically, this means that each sample value should provide an independent 'snapshot' of groundwater concentrations, not influenced by or correlated with other measurements. Otherwise, statistical results and summaries are likely to be biased and to underestimate the true variance.

In particular, to approximate independence over time at a fixed sampling point, sample measurements should not be collected too quickly after one another. Instead, a lag time should be allowed between sampling events, ideally governed by the degree of temporal correlation (that is, numerical similarity between consecutive or closely-timed sampling events) in the time series. Temporal correlation can be induced by a variety of physical factors, including among others the rate of groundwater flow, composition of the soil matrix, and the measuring process itself.

How long you should wait between sampling events will depend on site-specific conditions. A common rule of thumb is to sample no more frequently than quarterly, though this 'rule' is not based on formal studies. The Unified Guidance suggests that sites conduct a pilot study to estimate the correlation over a year's time at two or three representative wells. These correlation estimates can be utilized to establish a site-specific sampling frequency. The degree of temporal correlation can be checked using standard tools for time series analysis, like the sample autocorrelation function (see Section 5.8.3).

One common implication of the need for independence is that laboratory replicates and field duplicates should not be treated as independent measurements, since by design they should be highly correlated. To avoid such correlations, replicates and duplicates should either be averaged prior to statistical analysis or one duplicate or replicate from each set should be randomly selected to be included in the analysis data set. Deterministic rules such as always selecting the highest-valued replicate are discouraged, since they may bias the overall mean estimate but perhaps more importantly may cause the variance to be underestimated.

Another implication is that physical independence is not a guarantee of statistical independence. Even using Darcy's equation or similar method to ensure that physically distinct volumes of groundwater are sampled on different events does not necessarily ensure those measurements are statistically independent. Independence can be affected by other factors besides groundwater flow rates, including physical factors such as soil sorption and turbidity or the analytical measurement process itself (for example, periodic instrument miscalibration that biases some measurement batches but not others).

Seasonality is a special form of temporal dependence that can bias test outcomes if, as is typical, groundwater is not sampled at truly random times. If seasonality is present, a longer series of measurements is generally necessary to both characterize the seasonal pattern and to de-seasonalize the data (see Chapter 14.3.3, Unified Guidance) in order to remove the extraneous trend.

Example: Seasonality

In an extreme case, suppose contaminant concentrations always peak above a compliance criterion in the summer, but always drop below the criterion during the winter, with no long-term trend. Then, routine annual sampling in the summer (but not in the winter) will tend to identify the well as more contaminated than it really is, or perhaps falsely indicate that the long-term average exceeds regulatory limits.

Correlated data will both underestimate the true variance and represent the equivalent of a much smaller set of independent measurements. In fact, if the first order or 'lag-1' temporal correlation is equal to α, the number of equivalent independent values will be approximately $n(1-\alpha)/(1+\alpha)$ (Chatfield 2004). This means that a series of 20 measurements with a first order correlation of 0.3 will be roughly equivalent to only 11 independent observations.

3.4.5 Outliers, Identically Distributed Measurements

Any sufficiently-sized set of measurements drawn from a given population is assumed to have an identical distribution to that of the parent population. In traditional upgradient-to-downgradient comparison tests, the groundwater measurements at both upgradient and downgradient sampling points are assumed to be identically distributed unless the downgradient wells become contaminated. However, if there is significant natural spatial variability, the local distributions from well to well may differ even if the site is 'clean.' In another example, outliers are measurements that are either errors of some sort or do not come from the same statistical population as the rest of the data. Including one or more outliers in a background data set can dramatically affect statistical evaluations and often greatly decreases the statistical power of such tests. Check for both outliers and spatial variability in any groundwater data evaluation (see Section 5.10: Identification of Outliers and Section 5.5: Trend Tests). Outliers should generally be kept as part of the data set unless there is reasonable evidence that they are the result of an error.

3.4.6 Temporal Stability

Several groundwater statistical tests assume the input data are stable over time. This means the measurements should not exhibit obvious trends, but instead should be stable around a fixed mean. This assumption applies to t-tests, analysis of variance (ANOVA), confidence intervals around the mean, prediction limits, and control chart limits calculated using background data. Lack of temporal stability can substantially bias test outcomes, in large part because the estimated variance will be too high (and much higher than the nominal variance assumed by the test). When the sample data do not appear to be temporally stable, consider formal trend tests as an alternative or explicitly adjust for the apparent trend when designing the statistical method.

More generally, temporal stability and temporal stationarity are the same thing. Stationarity as a concept is more general in that it can also refer to spatial stationarity, referring to a local mean and variance that are stable across the site.

3.5 Testing Assumptions

EDA is described in Section 3.3.3 and is typically the first step in understanding data at a site and in helping to check the assumptions listed in Section 3.4. This section provides some guidance on how to implement EDA for testing statistical assumptions. Appendix F includes further information about checking the underlying assumptions of statistical tests. Effective EDA requires a decision logic or statistical process to sort through the decisions leading to a particular statistical design (see Section 3.6). The EDA process for each site will be different, but a general outline might include the following:

- **Testing normality.** Normality of the data distribution can be checked with formal tests such as the Shapiro-Wilk or with more subjective methods like probability plots. It may also be possible to normalize the data using a mathematical transform (for example, log, natural log) but note that (1) any subsequent parametric test must be run on the transformed data values and (2) back-transforming the results of the test may induce unacceptable bias. If the data cannot be sufficiently normalized, other distributions can be checked (such as Weibull or gamma); some parametric tests exist for these distributions. Unless the sample size is reasonably large, accuracy or statistical power may be lost when using a nonparametric test procedure. If possible, use a parametric test whenever the original data pass a normality test, or select an alternative nonparametric test method (see Section 5.6).
- **Testing for outliers.** Formal outlier tests, such as Dixon's or Rosner's test, usually assume normality. Therefore, check data for normality prior to running an outlier test. For data that cannot be normalized, use a nonparametric test method to minimize the effects of possible outliers. Including outliers in a statistical analysis, especially if they are part of a background data set, can lead to substantial loss of statistical power for detecting real changes. Sometimes, a nonparametric alternative will lessen the impact of one or more outliers, even if they are not removed prior to analysis. Examples include using a Wilcoxon rank sum test instead

of a t-test, a Kruskal-Wallis test instead of a parametric analysis of variance (ANOVA), or a Mann-Kendall or Theil-Sen line test instead of a linear regression.

- **Testing for background stability.** Use formal trend tests such as Mann-Kendall or the Theil-Sen line test to identify statistically significant downward or upward trends over time at compliance points without reference to background concentrations. Comparison tests against background, such as prediction limits, control charts, and tolerance limits also all assume that the background data are stable over the time frame being assessed. In these cases, apply the trend test to the background data as a diagnostic procedure to check the assumption. If a trend is found in background during this diagnostic check, a switch from, for instance, a prediction limit to a formal trend test at the compliance point may be required.
- **Testing for spatial variation.** Substantial spatial variation among sampling points can negate the use of traditional interwell (upgradient-to-downgradient groundwater) tests in favor of intrawell testing. Check for the presence or absence of spatial variation. If there are multiple background sampling points, these points can be compared formally with an analysis of variance (ANOVA) or informally using side-by-side box plots. If only one or no background sampling points exist, a similar check can be run on sampling events from compliance points that are known to be uncontaminated on the basis of other lines of evidence.
- **Testing for temporal independence.** Data that do not approximate independence through time can substantially bias test outcomes. Checking for independence requires a series of measurements, either from a single sampling point or from multiple sampling points all sampled on the same event. First check for trends or systematic patterns on a time series plot. If identifying a trend is of prime interest, independence should not be tested on the original data series, but rather on the residuals from the estimated trend if a linear regression is used. Alternatively, a nonparametric trend test like the Mann-Kendall or Theil-Sen line test might be used.
- **Accounting for nondetects.** A substantial fraction of nondetects often makes it impossible to normalize a data set or to accurately check for normality. Sometimes, a switch to an alternative nonparametric test method is warranted. As noted earlier, however, if the sample size is small a substantial loss in statistical accuracy or power may occur. Section 5.7 discusses managing nondetects.

3.6 Statistical Design Considerations

Statistics play a crucial role in properly evaluating groundwater throughout the project life cycle. Therefore statistical design, which is the intentional planning for statistical analysis and data collection, should always occur at the beginning of the project rather than the end. Ideally, statistical design should occur as part of a systematic planning process in the context of the project's DQOs and DQA process. To link this process more specifically to groundwater analysis, consider the following questions.

3.6.1 How good are my decisions?

Every statistical decision includes uncertainty. Upfront statistical design often allows the analyst to

anticipate the level of uncertainty attached to later statistical test results and to adjust the design if that uncertainty is unacceptable to stakeholders. Well-designed evaluations attempt to specify and control not only the confidence level of the test procedure, but also the expected false positive rates and false negative rates or statistical power (see Section 3.6.2).

3.6.1.1 False Positives and False Negatives

A false positive, Type I error, or alpha refers to rejecting the null hypothesis or conclusion about a population when it is actually true. A false negative, Type II error, or beta refers to failing to reject the null hypothesis or conclusion when it is actually false.

Since both false positives and false negatives can have regulatory and financial consequences, you should attempt to minimize both to the degree practicable, and consider both in the statistical design.

> ***Example: False Positives and False Negatives***
>
> *A 'false positive' would occur if the assumption that a site's groundwater is 'clean' is wrongly rejected; that is, the statistical evaluation erroneously concludes the groundwater is 'dirty.' A 'false negative' occurs when the groundwater is actually 'dirty' but the hypothesis of clean groundwater is accepted.*

3.6.1.2 Statistical Power

Statistical power is the complement of the false negative rate. It represents the probability that the null hypothesis will be rejected when the alternative hypothesis is true—the probability of not committing a Type II error. Higher power is always desirable since it implies that a correct decision will likely be made. Similarly, high statistical confidence is desirable since the confidence level indicates how likely it is that the null hypothesis will be accepted when it is true—the probability of not committing a Type I error. In the previous example, a high confidence level translates to a high probability of correctly deciding that 'clean' groundwater is indeed 'clean.'

3.6.1.3 Statistical Significance

Statistical significance is also required to assess the certainty of results. A statistically significant test conclusion is one with a low probability of occurring by chance. Often, the significance level of a test is equated with the false positive rate. A low false positive rate then sets a high bar for reaching significance, because the lower the false positive rate, the larger the difference, change, or trend in a data set needs to be in order to register a statistically significant result. By corollary, apparent differences that are due simply to chance variation are less likely to be identified at the lower the false positive rate.

Another helpful way to think of statistical significance is as the strength of the evidence *against* the null hypothesis. The more inconsistent the statistical evidence compared to the null hypothesis, the

lower the probability that those specific results would have been observed, assuming the null hypothesis is true. Numerically, this probability is expressed by what is called a p-value. P-values are commonly reported in statistical software to express the statistical significance of a test result, and measure how unlikely an observed set of results is relative to the null hypothesis.

3.6.2 What are site-wide false positive rates and power curves?

As discussed above (Section 3.6.1), the false positive rate for a single statistical test is the probability that the test will falsely indicate a statistically significant result when none exists. Since groundwater monitoring generally involves testing of multiple chemicals at multiple sampling points, the probability that at least one of those tests will falsely indicate a significant result is much higher than the individual test false positive rate. This alternate probability of error is known as the site-wide false positive rate (SWFPR).

To control the SWFPR and keep the number of false positive decisions to a minimum, Chapter 6.2, Unified Guidance, recommends designing any detection monitoring program to have an annual, cumulative SWFPR of 10%, regardless of the number of individual statistical tests that are run each year. To achieve this target, you can calculate the per-test false positive error rates for a specified number of tests using the equations found in Chapter 6.2.2, Unified Guidance. 'Per test' here refers to each sampling point and chemical combination statistically evaluated for compliance (for example, 10 chemicals measured in each of 5 compliance wells semiannually results in 100 annual tests). Some statistical software packages will also perform this calculation.

Ensuring an SWFPR of 10% at many sites entails assigning a very low false positive error rate to each individual test. This in turn tends to reduce the statistical power of those tests, since power always depends on factors such as sample size, significance level (Type I error rate, alpha), and the size of the difference or change in concentrations you want to detect (also known as the effect size). The effect in groundwater is often expressed as an increase over background in units of standard deviations (calculated from the background data). A large change in groundwater quality is easier to identify than a small change, so if the targeted effect size is large, the test will have higher power; conversely, smaller targeted effect sizes are associated with lower power.

Power curves may be used to estimate and visualize the statistical power of a test, or equivalently, a test's ability to correctly identify a 'significant increase' in chemical concentrations above background. A typical power curve graphs the statistical power of a test against a range of possible effect sizes (in terms of standard deviations above background); the effect size can be translated from standard deviations into concentration units. USEPA's Unified Guidance recommends regular use and reporting of power curves when designing statistical monitoring programs (see Chapter 6.2.3, Unified Guidance). This document also provides benchmark USEPA reference power curves (ERPC) with which to assess the adequacy of site-specific power curves. Generally, statistical software is needed to prepare power curves.

Because an inherent relationship and tradeoff exists between statistical power and the false positive rate of a test (that is, the lower the significance level, the lower the power), there is also a tradeoff

between the site-wide false positive rate (SWFPR) and power. Maintaining a balance between an acceptable SWFPR and sufficient statistical power is important to ensure that groundwater monitoring programs are protective of human health and the environment, while not placing undue burdens or unreasonable risk of false detections on responsible parties. To achieve this balance, the Unified Guidance recommends that the annual, cumulative SWFPR target be set at 10% and that the USEPA reference power curves be used to demonstrate that each single statistical test can detect a three or four standard deviation increase in groundwater contamination above background with reasonable power.

In practice, meeting these targets simultaneously at most sites—given the limited amount of groundwater data usually available or feasible—requires a three-point design strategy.

1. Examine the monitoring parameter list to see if any of the chemicals or indicators might be unrelated or perhaps distantly related to what would be present in groundwater should contamination actually occur. Eliminating chemicals from monitoring equates to fewer overall statistical tests, a smaller SWFPR, and greater power for the remaining tests.
2. The Unified Guidance recommends the Double Quantification Rule (DQR), instead of formal statistical testing, for any monitoring parameter that has never been detected in background. The DQR simplifies the assessment of such parameters, since you only need to observe quantified detections on two consecutive sampling events to identify a significant change in groundwater quality. Additionally, removing those parameters from formal testing again lowers the SWFPR and improves the statistical power of the remaining formal tests.
3. Institute a formal re-testing strategy (see Section 3.6.6) any time tests such as prediction limits, control charts, or tolerance limits are being used for release detection.

3.6.3 How much usable data do I have or need?

The statistical power, accuracy, and statistical confidence (see Section 3.6.1) of an analysis depend on the number of statistically usable measurements (also termed the sample size). As a general rule for parametric tests and many nonparametric ones as well, the larger the sample size the greater the power and the smaller the decision error risk (false positives and false negatives). Unfortunately, the relationship between sample size and statistical performance may be complicated and is somewhat different for each test. To minimize decision error risk, sample sizes should be determined ahead of time, if possible, as part of the design process.

Data usability depends on how closely the data set approximates an independent, identically-distributed sample and on how well those data represent the target population. Although it is difficult to verify these assumptions in a groundwater analysis, the statistical design and CSM should guide when and where to sample so as to best match the target population, minimize correlations between sampling events, and enable the collection of data related to the study questions. How, where, and when data are sampled all impact data usability and should be considered during statistical planning.

As a caution on sample size, while many statistical tests can be computed with just a handful of measurements (sometimes as few as three), such tests tend to have unacceptably low statistical power or high false positive rates. With parametric methods, critical points in published statistical tables or in software account for the desired false positive rate (or significance level) but give no indication of statistical power. Conversely, nonparametric statistical intervals (for example, prediction limits) computed with small sample sizes have high power but also very high false positive rates; nor can the false positive rate be specified in advance by the user without increasing the sample size.

Example: Background Sample Size

As an example of the impact of insufficient data, consider the achievable false positive rate using a nonparametric prediction limit based on the maximum observed background value to test for contamination at a compliance well when collecting a new measurement. Figure 3-2 plots background (BG, used in the figure) sample size versus the expected false positive rate of the test. The horizontal limits on the graph indicate standard 5% and 1% significance levels. The background sample size must be at least 19 to achieve 5% significance, and at least 99 to achieve 1% significance. Only 5 background values nets a 17% chance of falsely detecting a release, while 10 would result in a 9% false positive risk. The prediction limit itself can be computed with just a single background measurement, despite the high decision error risk (50%). It is also the case (not shown) that a much lower sufficient sample size can be used if formal retesting is pre-specified and incorporated into the procedure.

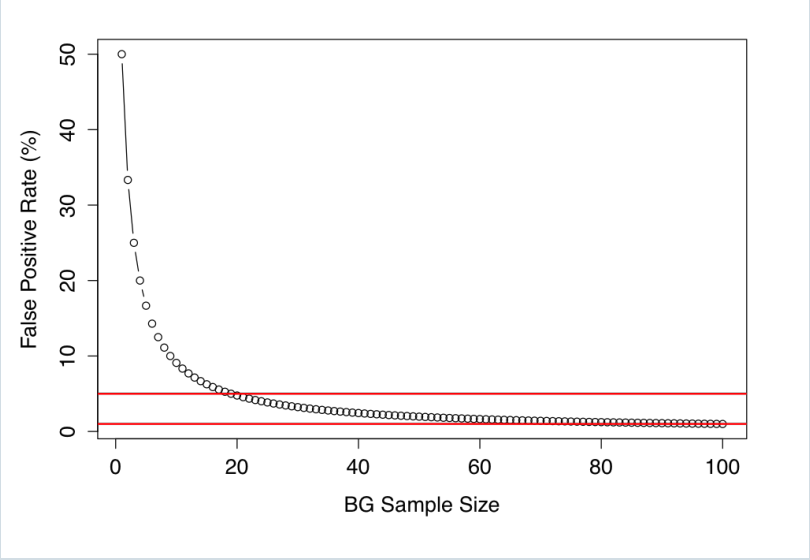


Figure 3-2. Background sample size versus the expected false positive rate of the test.

It is also difficult to judge whether a parametric statistical model fits the available data when the sample size is small. One method is to use 'bootstrapping' to augment the information contained in a small sample via computer-simulated resampling. Unfortunately, while standard bootstrapping is a valuable tool in moderate to large data sets for estimating sampling variability, it gives unreliable results for very small sample sizes and should be avoided. There is no substitute for real data. If the sample size is too small, additional measurements must be obtained prior to statistical analysis.

Most guidelines on sample size for groundwater tests recommend at least 8 to 10 background measurements when constructing prediction limits, control charts, or tolerance limits, and roughly the same number of compliance point measurements when calculating trend tests or confidence intervals. Some exceptions to these guidelines exist, but importantly, none of the recommendations directly indicate what statistical power will be achieved during parametric testing, or how much false positive rate control is likely in nonparametric tests. To figure both power and the expected false positive rate, further pre-planning is necessary. For some parametric tests without re-testing, you can use a minimum sample size formula to calculate the required sample size, as long as (1) a rough estimate of the variance is either known or can be bounded, and (2) you can specify the minimum difference of importance that should be detected.

Unfortunately, analytical calculations of statistical power quickly become complicated depending on the type of test and application. Adding retesting to standard prediction limits or control charts dramatically changes the way minimum sample sizes must be computed, as does the less-than-mathematically-tractable nature of combined Shewhart cumulative sum control chart (Shewhart-CUSUM) control charts (see Section 5.13). In many cases, power and minimum sample size characteristics of a method must be simulated on a computer, so that professional statistical assistance may be required. One help in this regard is that the Unified Guidance provides pre-computed sample size and power values for many scenarios involving prediction limits with retesting (though not for Shewhart-CUSUM control charts).

3.6.4 What are the critical contaminants?

The selection of which contaminants to monitor might seem at first to be strictly a decision for regulators and responsible parties. Often the list of monitored contaminants is set as part of the regulatory record, based on the composition of the contaminant sources, interaction of those contaminants with the local hydrogeology (including mobility, persistence, toxicity), and at times political sensitivities. Nevertheless, consider the impact of the monitoring list during the statistical design phase because it can affect statistical results. For instance, the more chemical parameters subject to formal statistical testing, the larger the resulting risk of making false positive decisions, and the greater the difficulty in managing that risk.

In some cases, chemicals on the monitoring list may be only distantly, or not at all, related to source composition, meaning that detection of those parameters is not indicative of site contamination. Removing these chemicals from formal monitoring can help optimize analytical resources, improve compliance decisions, and allow the use of more powerful statistical tests for the remaining contaminants. Also, the list of contaminants should be screened for statistical usability. Parameters that

are heavily nondetect, that have low analytical precision, or that are only sampled in a small number of locations or across a small fraction of the site may be difficult to statistically analyze, especially within desired bounds of statistical error or confidence.

3.6.5 Should I use interwell or intrawell sampling?

Comparisons of compliance wells against local background data usually take one of two forms: interwell tests of upgradient versus downgradient wells, or intrawell tests of earlier versus more recent measurements at each compliance well (see Section 3.1). The type of comparison that is most appropriate depends on the specific contaminants and the nature of local hydrogeologic conditions. Substantial natural spatial variability may necessitate intrawell methods, a condition which can be identified in part by using graphical EDA tools (see Section 3.3.3) and diagnostic statistical tests. At some sites, both intrawell and interwell tests may be appropriate, because the degree of spatial variation may vary by contaminant (for example, levels of naturally-occurring chemicals may differ spatially from anthropogenic contaminants).

Intrawell methods are also the method of choice when attempting to identify trends at individual sampling points or when evaluating post-remedial progress in contaminant reduction. In these cases, local background may not be relevant to the comparison or may not be stable enough over time to allow the use of interwell tests.

3.6.6 Should I retest and how?

Long-term monitoring at many sites requires years of routine sampling at fixed intervals. If a new round of sampling at a given sampling point is inconsistent with background levels, evidence exists of possible contamination or a change in conditions. To confirm whether the apparent change in concentrations is real or simply due to chance sampling variation or the impact of an outlier, retesting may be needed. Retesting—collecting and testing one or more new, independent groundwater sample at that sampling point after the initial test—tends to confirm actual changes and simultaneously eliminate most false indications.

Because the amount of new data collected at a given sampling point during any sampling event is small (often one new measurement), retesting is generally imperative for a successful detection monitoring program. Retests must be explicitly built into the statistical procedure and incorporated into the statistical design that is established during project planning. All facets of retesting—including how many resamples to collect, how much independent background data are available, and the decision rules used to confirm or refute an initial indication—affect the statistical power and accuracy of groundwater tests. Consult Chapter 19, Unified Guidance, for additional information.

3.6.7 Does my monitoring network need to be optimized?

Depending on the study question, a monitoring network might include more sampling points or be sampled more frequently than is actually necessary. Conversely, there may be inadequate spatial or temporal coverage of the site suggesting that either additional sampling points should be added to

better characterize complex spatial trends or more frequent sampling at existing wells should occur. Both types of questions are part of monitoring network optimization.

Good statistical design should document networks that exhibit statistical redundancy in the numbers and placement of sampling points, or that are sampled more often than needed, and estimate the degree of redundancy present. In general, a high degree of spatial correlation between sampling points or a large temporal correlation between consecutive sampling events indicates statistical redundancy. Methods to identify redundancy and optimize your monitoring network are described in Section 4.5.3 and Section 5.14. Ideally, any network should be designed to most efficiently answer the study questions, while not wasting sampling and analytical resources.

It is also important to identify networks that provide too little statistical information to adequately characterize the site. In these cases, the spatial sampling design can be optimized by adding new sampling points at the most effective locations, or by increasing sampling efforts at existing wells to ensure compliance decisions can be made with adequate statistical confidence. Methods to optimize 'by addition rather than subtraction' are described in Section 4.5.3 and Section 5.14, generally relying on (1) locating specific site areas with the largest statistical uncertainty and fewest wells, and (2) determining whether trends over time are highly uncertain and linking this uncertainty to the current sampling frequency.

3.6.8 Is geostatistical or spatial analysis of groundwater necessary?

Geostatistical methods can be powerful tools for mapping and characterizing concentration or mass-flux patterns across a site. Geostatistical analyses, for instance, produce isopleth maps such as isoconcentration contour maps. Generating such maps and estimates usually requires more sampling points (sometimes arranged in a systematic pattern or along linear transects) than are available or feasible at many sites. As such, the statistical design should critically examine whether geostatistical analysis is necessary prior to well placement. Furthermore, the best sampling design for spatial analysis (for example, systematic grid) may not coincide with the requirements of the CSM, where wells may be sited along presumed preferential pathways or grouped near source areas.

Common sampling designs for spatial analysis include random sampling, systematic sampling, or multi-stage sampling. Random sampling of an aquifer (that is, by randomly locating sampling points) helps to ensure a low level of statistical bias in the data set, but often cannot be done because of physical obstacles (such as buildings) or logistical difficulties in locating and drilling at truly random coordinates. It also may be less efficient than systematic designs in estimating the variance of a spatially-correlated or spatially-stratified population. Systematic sampling is useful when uniform spatial coverage is desired or when attempting to identify localized contaminant plumes (such as hotspots), but may be cost-prohibitive for many groundwater sites unless the sampling points consist of temporary wells or perhaps field sensors.

Multi-stage sampling is a design option that can more efficiently use sampling resources and yet allow for spatial analysis. This design entails initially obtaining a higher density of screening or semi-quantitative data that are then used to focus the collection of samples to be analyzed using

more costly fixed-based laboratory analysis. Inexpensive field screening or direct push technologies may be a viable option for increasing the spatial resolution, as an alternative to drilling a smaller number of permanent wells only in locations targeted by professional judgment. Multi-stage sampling strategies must be developed during the statistical design phase.

Sampling points may also need to be stratified or apportioned into spatial groupings to represent distinct statistical populations or subpopulations. Such groupings may include upgradient and downgradient zones or reflect different aquifers, hydrostratigraphic units, or multiple screening depths.

3.6.9 Can I use field screening or the Triad approach?

On-site remediation of groundwater requires accurate characterization and understanding of any subsurface plumes. Often, the expense involved with drilling permanent wells and then using sophisticated analytical methods to measure each physical sample (particularly for organic chemicals) precludes locating more than a few sampling points relative to the areal extent of the site. Regulatory requirements, for instance under the Resource Conservation and Recovery Act (RCRA), have generally considered these realities so that a typical RCRA waste site is required to have a minimum of only one upgradient well and three downgradient wells.

A parallel concern is that less expensive field screening techniques were in the past typically associated with substantially greater levels of analytical uncertainty, making it difficult to accurately measure chemicals at low concentrations. USEPA has recognized the significant progress in sampling and measurement technologies, and that lower analytical uncertainty in individual laboratory analyses is often more than balanced by the large risk of decision error introduced from having too few sampling points. This understanding supports USEPA's Triad approach (see *Triad Implementation Guide*, ITRC 2007b).

Triad applies three key concepts to the statistical and sampling design at a cleanup site:

1. Systematic planning, including development of a CSM, upfront assessment and management of the risks of decision error and uncertainty, and allowance for the CSM to evolve as new information comes to light from sampling activities
2. Dynamic work strategies, in which pre-approved decision logic is used to flexibly adapt the sampling design and subsequent sampling activities to the information generated
3. 'Real-time' measurement technologies, including either rapid turn-around analyses from a traditional laboratory, or field-based screening and measurement methods, and direct push technologies.

The Triad approach consciously trades the lower analytical uncertainty associated with traditional laboratory and sampling procedures for the benefits gained from less expensive individual measurements, quicker turn-around time, and the statistical reality that more sampling points are needed to accurately characterize contaminated groundwater plumes than are typically available using traditional sampling plans. The gain in decision certainty from a larger number of, say, field screening measurements can often outweigh the better analytical precision of a small number of laboratory-based analyses.

Whether the Triad approach is useful at a particular site depends on whether more inexpensive methods of measurement exist for the contaminants, or for surrogates of those chemicals, and a proper weighing of the potential gains in project costs, flexibility, and decision certainty. All of these factors must be addressed in the systematic planning phase.

4.0 STATISTICAL ANALYSIS FOR PROJECT LIFE CYCLE STAGES

This section explains the role of groundwater statistics in the activities of typical project life cycle stages. These project life cycle stages include release detection, site characterization, remediation, monitoring, and closure. The environmental projects may be cleanup projects, or compliance monitoring (for example Resource Conservation and Recovery Act (RCRA) facilities) projects. Study questions serve as a bridge connecting project life cycle stages with relevant statistical methods and were selected based on common project objectives that require statistical analyses. Ten common study questions and their associations with each of the life cycle stages are presented in Table 4-1. A more detailed discussion of the study questions is presented in Appendix C.

Table 4-1. Statistical Study Questions for life cycle stages

	Project Life Cycle Stages				
Study Questions	Release Detection	Site Characterization	Remediation	Monitoring	Closure
1. What are the background concentrations?	X	X			X
2. Are concentrations greater than background concentrations?	X	X		X	X
3. Are concentrations above or below a criterion?	X	X	X	X	X
4. When will contaminant concentrations reach a criterion?			X	X	X
5. Is there a trend in contaminant concentrations?			X	X	X
6. Is there seasonality in the concentrations?			X	X	X
7. What are the contaminant attenuation rates in wells?			X	X	X
8. How do contaminant concentrations change with distance from the source area?			X	X	X
9. Is the sampling frequency appropriate (temporal optimization)?	X	X	X	X	X
10. Is the spatial coverage of the monitoring network appropriate (spatial optimization)?	X	X	X	X	X

Statistical analyses are conducted to answer the study questions listed in Table 4-1. The questions start simply and move to more complex analyses. You may need to reconsider the questions, however, as the project progresses through the life cycle stages. For example, when initially assessing a release or characterizing a site, you will likely determine background concentrations for comparison to site concentrations. Later, you may need to revisit background concentrations when determining compliance with criteria. Some study questions are relevant through all stages of a

project life cycle.

Study questions also have a relationship to one another in the context of specific groundwater evaluation objectives such as background or attenuation, which may be important during various project life cycle stages. Study Questions 1 and 2 assess background concentrations that may be of interest during release detection, site characterization, monitoring, and closure stages. Study Questions 3 and 4 assess contaminant concentrations with respect to criteria that may be important in all of the life cycle stages. Study Questions 5 and 6 evaluate temporal trends in data sets that may be of concern during remediation, monitoring, and closure. Study Questions 7 and 8 assess temporal and spatial rates of change for contaminants and are also important concerns during remediation, monitoring, and closure. Study Questions 9 and 10 evaluate if the frequency of sampling and spatial coverage of wells are appropriate, leading to a more optimal monitoring program, which is an important consideration for all life cycle stages. These final questions are used to determine whether a monitoring well network may need to be expanded for sufficient compliance point or plume migration coverage. These questions may also be used to determine whether more or less frequent sampling is necessary to characterize or evaluate changes in contaminant concentrations.

The five project life cycle stages do not cover all possible project situations, but are provided to link statistical methods and tools to the typical waste management facility or contaminated site investigation, monitoring, or cleanup actions. Many of the project life cycle stages share common groundwater evaluation objectives, where groundwater statistical methods might be used.

The discussion in this section provides the following guidance for each life cycle stage in relation to the study questions:

- selecting and characterizing a data set relevant to the study question
- appropriate statistical methods and tools for the study question
- interpretation of the results and the associated uncertainty

The discussion in the following sections does not provide guidance on the specific assessment or remediation tasks associated with each life cycle stage (such as groundwater remediation methods) but rather provides guidance on the use of statistics to support the groundwater evaluation objectives. Additional information about common mistakes in applying statistical methods is presented in Appendix B.

4.1 Considerations for Statistical Analysis

Exploratory data analysis (EDA) is a common step for all project life cycle stages or study questions. As discussed in Section 3.3.3, EDA is valuable for inspecting the quality and character of groundwater sample results. Data should be evaluated for frequency of detection, multiple detection limits, and outliers (either high or low values). Check the distributions of the sample data for normality or lognormality to select appropriate statistical methods. In general, do not make unverified assumptions regarding the statistical distribution of data. When the impact of nondetects on the data set is combined with the impact of a small data set or data sets containing extreme values,

distributional tests may fail to identify a known distribution and nonparametric methods must be used.

The number of samples that must be collected in order to apply the statistical methods varies; however, in general, more samples better characterize groundwater concentrations. USEPA recommends a minimum of 8 to 10 independent observations (Chapter 5.2.1, Unified Guidance) for most of the statistical tests. States may require a specific number of minimum observations by rule. Data sets of 20 or more observations may be possible, and methods to expand the data set, discussed in Chapter 5.2.6, Unified Guidance, can include additional sampling or pooling data from more than one well if the data characteristics allow (for example, use an analysis of variance (ANOVA) test to show concentrations do not differ) in order to statistically increase the number of samples. However, when pooling data, the statistical significance of a single contaminated well may be reduced by pooling with uncontaminated wells.

In addition, collecting samples that are not separated by sufficient time intervals can lead to redundant measurements. Such data are not statistically independent. Collecting samples separated by too long an interval may miss important aspects of the data record. See Study Question 9 for more information on temporal optimization methods.

Ideally, the project planning process documents the data analysis procedures, including how nondetects are managed. If not established prior to sampling then a decision should be made during EDA as to how nondetect data will be handled in the data set. Parametric statistical tests typically utilize the mean and standard deviation of a data set, both of which may be significantly skewed by using the detection limit or other substitution methods in these calculations. Nonparametric tests are not as impacted by a small number of nondetects, but the results of almost all statistical tests can be confounded by a large number of nondetects especially if associated with varying detection limits. See Section 5.7 for guidance on how to handle nondetects in statistical analyses.

Prior to analyzing the groundwater data and after the EDA, review the assumptions of the statistical tests. For example, some of the tests require that the data be derived from a normal statistical distribution. The assumptions for each of the statistical methods are discussed in detail in Section 5.0. Section 3.4 provides more information on the general statistical assumptions, and Appendix F includes information about checking the underlying assumptions of statistical tests.

4.2 Release Detection

At cleanup sites or at waste management facilities, groundwater monitoring may be used to determine if a release has occurred. A release may be detected by comparison of compliance well data to a criterion or by detection of a trend in the compliance well data. Groundwater concentrations may be compared to a criterion to determine facility compliance. A release may also be detected when the concentration of a chemical in groundwater exceeds background. Accordingly, an important aspect of release detection monitoring is determining the background concentrations (either natural or anthropogenic) for chemicals. Natural background would be representative of pristine or pre-

industrial conditions. Anthropogenic background refers to concentrations in the surrounding groundwater that may be impacted by human activity, but not by the site.

The study questions that are most applicable for release detection are shown here.

Release Detection Study Questions
1. What are the background concentrations?
2. Are concentrations greater than background concentrations?
3. Are concentrations above or below a criterion?
9. Is the sampling frequency appropriate (temporal optimization)?
10. Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

4.2.1 Background Conditions

When discussing the concept of background, it should be clarified whether background represents natural or anthropogenic conditions as well as whether background conditions are defined by location or by a time period that is uninfluenced by the site. Few sources exist to determine the analytical qualities of pristine groundwater, so groundwater monitoring results are typically compared to the conditions in wells unaffected by the contaminants associated with the activities being monitored. For example, monitoring wells upgradient of a landfill are generally compared to monitoring wells downgradient of the same landfill. Significantly greater contaminant concentrations in the downgradient wells than the upgradient wells would support the conclusion that contaminants were released from the landfill. Most guidance documents and research papers on determining natural or anthropogenic background concentrations primarily employ statistical techniques. However, geochemical evaluations may be used to identify potentially contaminated samples and naturally occurring metals (Thorbjornsen and Myers 2007).

Background data sets can be compared to site data using a variety of statistical tests depending on the geology, well development, and chemicals being examined as well as the release mechanism identified in the conceptual site model. The number of planned background wells or groundwater sampling points should reflect the site being investigated in terms of site size, the variability of the concentrations for a chemical, and analytical detection limits. Background may also be represented by a single value based on regulatory requirements or a value derived from literature. Not all background values reported in the literature are appropriate for comparison to site concentrations. Background values obtained from literature should be reviewed to determine whether the regional geology and land use for the values are similar to conditions at the site.

4.2.2 Location or Selection of Background Wells

In locating or selecting the background wells or data set, consider the common requirements for

statistical analyses. The monitoring wells should represent the same hydrogeologic unit and not be too close together or immediately up- or downgradient of one another. In determining proximity of wells, some factors to consider include formation transmissivity, range of contaminant concentrations, and size of area being investigated. Proper well placement assures that the water samples drawn from different wells are independent such that the groundwater from the same location is not being sampled twice. A sufficient number of samples must be collected to demonstrate that the analytical results do not correlate with either the time of collection or nearby wells. There also should be a sufficient number of samples to determine whether seasonality is important. In general, if the concentration of a chemical increases or decreases over time or is significantly higher in some background wells as compared to others then the background data drawn from those wells may not be suitable for use as background.

To help avoid misrepresentation of chemicals in background groundwater, avoid placing background wells near typical, known sources of contamination such as burial areas and underground storage tanks. Collect basic geologic data to determine which sources of contamination should be of concern. The hydrogeologic characteristics such as groundwater movement, direction, volume, and stability all are important and are discussed further in Section 4.3.

The appropriate number of samples will be related to site conditions as characterized in the conceptual site model (CSM); see Section 3.6 for further discussion of statistical design and number of samples.

4.2.3 Monitoring for a Release

When sample data are available from both background and potentially impacted wells, plots can be used to graphically assess whether the two data sets are derived from the same or different statistical populations. Probability plots and box plots can be useful for such qualitative evaluations.

An upper criterion can be calculated from background data as either an upper prediction limit or upper tolerance limit. Control charts can be used as an alternate graphical tool to assess whether concentrations within a well (intrawell) have increased above a control limit based on data from an earlier (background) time period. Sample results from the potentially impacted wells are then compared to these limits to determine if a release has occurred. Prediction or tolerance limits are simple to implement and communicate results. Care must also be taken to account for high site-wide false positive rates (SWFPR) when multiple contaminants and multiple wells are being compared. Multiple comparisons with a fixed level of statistical significance indicate that each repeat of the test has the same chance of incorrectly rejecting the null hypothesis and those repeats accumulate or sum the error (increasing the risk of at least one mistaken decision).

Two-sample tests are an alternative to the prediction limit or tolerance limit. These tests compare the mean or mean rank of the potentially impacted wells with the same statistic for the background data. The parametric comparison tests are Welch's t-test and pooled variance t-test and the nonparametric equivalents are the Wilcoxon rank sum test and Tarone-Ware test. The t-test is sensitive to outliers; the nonparametric tests are not sensitive to outliers.

4.2.4 Statistical Methods for Release Detection Objectives

When evaluating whether or not a release has occurred, the following methods are most applicable:

- Prepare box plots or probability plots to qualitatively evaluate whether the background and potentially impacted data sets appear to be drawn from the same population.
- Determine if the background samples appear to be from a single statistical population or if multiple aquifer characteristics exist (see one-way ANOVA or t-test, Wilcoxon rank sum test, Tarone-Ware), and assess whether background concentrations are stable (see parametric trend tests, Mann-Kendall trend tests).
- Calculate an interval that represents a background distribution of concentrations of chemicals from one or a set of wells that are unaffected by a contaminant source or a contaminated site with a stated coverage and certainty. See prediction limits, control charts, and tolerance limits.
- In some circumstances you may want to consider the following approach: use two sample tests to determine if there is a difference between potentially impacted wells and background. See t-test, ANOVA, Wilcoxon rank sum test, Tarone-Ware.

4.3 Site Characterization

Site characterization is typically the first phase of a cleanup project. Site characterization describes the physical conditions of the site such as soils, geology, hydrology, the presence of existing contamination, the potential for contamination to be released, and the actual and potential pathways and mechanisms for contamination transport. This stage of the project life cycle considers the chemical characteristics of the contaminants and their potential to be mobile in the environment. All of these aspects of site characterization are needed to develop an appropriate groundwater monitoring program, understand groundwater contaminant concentrations, and select and interpret groundwater statistical analyses.

Information collected during the characterization phase may support, refute, or provide additional details for the initial assumptions regarding the site. While some information may not be initially known, information must be collected to support the CSM. Development of the CSM starts at the beginning of the cleanup project and continues as additional information becomes available (see Section 3.2).

To ensure that the data collected will support the project goals, use a structured process such as the Triad approach or USEPA Data Quality Objectives (DQO) process to guide data acquisition. Developing, implementing, and optimizing groundwater monitoring all depend upon data of known quality and sufficient quantity. Section 3.3.1 includes more information about data quality.

The study questions that are most applicable for site characterization are shown here.

Site Characterization Study Questions
1. What are the background concentrations?
2. Are concentrations greater than background concentrations?
3. Are concentrations above or below a criterion?
9. Is the sampling frequency appropriate (temporal optimization)?
10. Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

4.3.1 Physical Site Conditions

The geology, geography, and climatic conditions associated with a site strongly influence groundwater hydrology. Site hydrology describes not only the presence or absence of water, but also its flow direction, its velocity, and its volume. The location of the aquifer and its character must be well understood in order to assure that the groundwater samples with which comparisons are made are drawn from the same sample population (same aquifer).

Surrounding bodies of water and local drainage patterns also influence hydrology. For example, it is possible for a local stream to be gaining during one season (groundwater moves into the stream) and losing during another season (stream water moves into groundwater) so that samples drawn from groundwater are influenced by differing sources of water in different seasons.

The natural (background) geochemical composition of the groundwater is directly related to the aquifer mineralogy and the rate of dissolution of the aquifer minerals into the groundwater. Most of these reactions are slow and only subtle changes in inorganic concentrations occur over time within a hydrogeologic unit. Faster reactions and changes can occur locally where there are changes in aquifer mineralogy such as a sandy stream channel within a limestone unit or an influx of groundwater with a different composition than the pore water of the aquifer such as from a losing stream.

Variables such as temperature, exposed surface area of minerals, pH, oxidation-reduction potential, and presence in solution of other ions will affect not only the rate of dissolution of minerals into the groundwater but also influence the precipitation or co-precipitation of minerals from the groundwater, potentially changing the groundwater chemistry. These factors will also affect the rate of absorption (uptake into the physical structure) of ions to the aquifer material or particulate matter in the groundwater itself.

Aquifer characteristics such as the type of aquifer solids, pore structure, and fracture systems may alter the transport of dissolved chemicals by physical mechanisms such as absorption onto the aquifer matrix. Depending on the persistence of the contaminant, the same mechanisms that hinder movement may also allow the slow release of contaminants over time. Soil type (such as a clay mineral) can also affect the mass of contamination that may be released into groundwater over time by absorbing the contamination onto the minerals within the soils. These contaminants are then slowly

released over time into the groundwater or infiltrating rainwater. The characteristics of pores and fractures in rock aquifers are also important controllers in the interaction of contaminants with aquifer material.

While not often emphasized, the microbiology of the subsurface can change contaminant half-lives, support biodegradation, cause changes in water chemistry, and affect contaminant movement. For example, changes in biota can alter key characteristics such as pH, free metals, and dissolved oxygen concentrations. You should understand and monitor biological parameters if microbial activity affects the goals for the site. See the ITRC *Environmental Molecular Diagnostics* (ITRC 2013) web-based document for information about a group of advanced and emerging techniques (referred to as EMDs) that analyze biological and chemical characteristics of environmental samples.

4.3.2 Existing Contamination and its Sources

The nature and extent of contamination should be understood sufficiently to support appropriate groundwater monitoring. Consider the following questions when evaluating existing contamination and its sources:

- How many potential sources exist?
- Are all present and potential sources releasing into the groundwater being monitored?
- What are the contaminants?
- Is contaminant co-mingling occurring from multiple potential source areas or does it have the potential to occur?
- Is a contaminant being examined against a background concentration that may include a naturally occurring element like chromium? Or is monitoring focused on a contaminant that is not present in natural background but may be present in anthropogenic background such as polychlorinated biphenyls (PCBs)? See Section 4.2.1 for a detailed discussion on background.
- With the above concerns in mind, what is the extent or mass of contamination?

4.3.3 Pathways and Mechanisms of Transport

The pathways for transport of contamination determine whether data are being appropriately evaluated and also help to anticipate how sample results may change over time.

Different geologic units within the same aquifer can have differing water quality. For example, for a saturated clay layer underlain by sand, the contaminant concentrations in the two geologic units can be significantly different due to different transmissivity. Some contaminants, such as depleted uranium, are generally immobile but can be mobilized under certain geochemical conditions. Transport mechanisms then become crucial to understanding results for these contaminants.

4.3.3.1 Contaminant Chemical and Physical Nature

Comparison of Two Aquifers

A site being monitored has both an upper and lower aquifer and the lower aquifer is not in direct contact with the area of contamination. As a result, the upper aquifer may be contaminated while the lower aquifer may be clean. Statistical comparisons of the concentrations in these aquifers could support and help to refine the evaluation of groundwater transport pathways.

The physical and chemical characteristics of the contaminant must also be considered in groundwater monitoring. A contaminant such as benzene, a light nonaqueous phase liquid (LNAPL), will both float on the surface of groundwater and dissolve into groundwater. If LNAPL is present, groundwater samples which are drawn from the water table can differ from deeper samples that only represent the dissolved phase. Benzene is also highly mobile and may travel further in groundwater than other contaminants, even those originating from the same source area. Trichloroethene, a dense nonaqueous phase liquid (DNAPL), is heavier than water and also dissolves in water, so it may exist below downgradient monitoring wells screened at the same depths as upgradient wells. An inorganic element such as arsenic can bind tightly to clay, but can be released by changes in parameters such as pH.

In order to properly characterize a site, all of the above contaminant characteristics must be considered. While some information may not be initially known, it is important that enough information be collected to support the CSM (see Section 3.0: General Statistical Approach).

4.3.3.2 Groundwater Monitoring Networks

Groundwater monitoring networks are built around either identified potential sources of contamination (such as a landfill, spill, or release) or are designed in an effort to identify and delineate unknown sources of contamination. In either case, these monitoring well networks should ideally be developed by examining the site history and geology and considering the following questions:

- What was the material released and what was its chemical composition?
- Where were materials stored, handled, processed or disposed?
- Can the volume of material released be estimated?
- Do available records shed light on the potential pathways by which contamination may have entered the soil and groundwater, such as drainage swales, underground tanks or piping, storm water or sanitary sewer systems?

More information on considerations for designing groundwater monitoring networks can be found in Section 3.6.

Historical records and data can provide context for site characterization, but site geography, geology, and hydrogeology are critical as well. Section 3.3.2 has more information on the issues

associated with using historical data. Once the available site data have been collected and assessed, the information can be used to identify the possible pathways for contamination to migrate, the sources of contamination, the possible contaminants, and then to plan the initial site characterization investigations. When information that is needed to support a CSM is not available in the historical records, a plan must be developed to fill the data gaps. Soil and groundwater sampling are typically needed to delineate the contamination extent, to confirm pathways, and to differentiate sources. Once the initial investigation has been performed, graphical and statistical evaluations of the data can be used to identify the extent of contamination to the degree possible with the data collected. This information can be used to develop an initial CSM. In this iterative process, graphical data analysis and statistics help to direct data collection, which is used to refine and focus the CSM.

4.3.4 Statistical Methods for Site Characterization Objectives

When conducting site characterization the following statistical methods are most applicable:

- Determine whether or not the mean concentration of the contaminant is increasing or decreasing over time. If concentration is plotted versus time, either linear regression (a parametric test) or Mann-Kendall and Theil-Sen (nonparametric tests) can be used to verify this trend. When comparing the trends between wells or in different time frames, you should check to see if the slopes of the two regression lines are significantly different. See linear regression, Mann-Kendall trend test, and Theil-Sen trend line.
- Determine whether individual wells within a monitoring well network have a contaminant concentration greater than that expected for a certain percentile (for example, 95th or 99th percentile) of the wells in the network historically. Calculation of an upper quartile, upper tolerance limits or upper prediction limits may help identify areas of highest concentration that may warrant further characterization.
- Compare a data set to a criterion. Whether that criterion is a maximum contaminant level (MCL) or similar regulatory value, first determine what the value is intended to convey. Does the criterion represent a not-to-exceed value, a mean value, or a value intended to represent a percentage of the population? Then compute an appropriate confidence interval on the data set to determine if the criterion has been exceeded. See parametric confidence interval, nonparametric confidence interval, and confidence interval for upper percentile.

4.4 Remediation

Remediation of contaminated groundwater is a challenge. On average, attainment of regulatory closure at sites with contaminated groundwater takes significantly longer compared to sites that have contaminated soil but no groundwater impacts. At many sites, it is estimated that attainment of groundwater standards will take decades or more (See NAP 2012 for a discussion of the challenges of managing complex contaminated groundwater sites).

This section provides guidance on the use of statistics to support remedy selection and evaluation of remedy effectiveness. For remediation, statistical analyses are most useful for evaluating changes

in concentration over time (trend analyses). An objective and accurate evaluation of changes in contaminant concentrations over time can help to resolve groundwater remediation issues.

The study questions that are most applicable for remediation are shown here.

Remediation Study Questions
3. Are concentrations above or below a criterion?
4. When will contaminant concentrations reach a criterion?
5. Is there a trend in contaminant concentrations?
6. Is there seasonality in the concentrations?
7. What are the contaminant attenuation rates in wells?
8. How do contaminant concentrations change with distance from the source area?
9. Is the sampling frequency appropriate (temporal optimization)?
10. Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

4.4.1 Remedy Selection

Selection of technologies for remediation of groundwater is commonly based on evaluation criteria. The CERCLA remedy selection process evaluates potential remedial alternatives using nine criteria (USEPA 1988). These CERCLA criteria are used to select the best overall remedy for the site. State cleanup regulations may also contain similar evaluation criteria that are used for selection of remedial technologies. Although statistical analyses are not always directly relevant to remedy selection, statistics can, for example, support natural attenuation as a potential remedy.

When natural attenuation is being considered as a potential remedy, trend analyses for existing groundwater monitoring data can be used evaluate short-term and long-term effectiveness and to predict remediation time frames. The results of these analyses can support a comparison of natural attenuation to other remedies. Because these analyses use existing data, the evaluation methods are essentially the same as those used for the evaluation of remedy effectiveness for a selected and implemented technology.

4.4.2 Remedy Effectiveness

After a groundwater remedy has been implemented, statistical analysis of groundwater monitoring results can show the degree of remedy effectiveness. Analyses that can be used to evaluate remedy effectiveness include groundwater plume contouring, an examination of contaminant concentration versus time (temporal trends analysis) and an examination of contaminant concentration versus distance from the source (spatial trend analysis).

Contouring may be used to better understand the spatial pattern. Many software packages perform contouring; these packages often perform poorly with the sparse data sets typical of corrective action sites. As a result, hand contouring is often preferred (Siegel 2008). If a software package is used for contouring, carefully review the results for interpolation and extrapolation errors. Remedy effectiveness can be evaluated by 1) plotting the temporal trends on a map and evaluating the spatial pattern in the trends or 2) creating a series of maps and evaluating the change in spatial pattern over time before and during remedy operation (see Section 3.6.7: Does my monitoring network need to be optimized?).

Statistical evaluation of remedy effectiveness may employ a number of methods, and may address a variety of site parameters; some examples are listed below:

- Determine whether the change in concentration over time represents a statistically-significant long-term trend (temporal trend analysis, see regression analysis and Mann-Kendall trend test).
- Estimate the rate of concentration change over time (the attenuation rate, see Example A.5 and Example A.6). Use the confidence interval for this attenuation rate to evaluate the uncertainty in the estimate (see regression analysis and Theil-Sen test).
- Evaluate the areal extent of remedy effectiveness by identifying the wells with higher attenuation rates. The confidence intervals for the attenuation rates can be used to determine whether the observed differences in attenuation rates are statistically significant (see confidence interval bands on regression analysis and Theil-Sen test).
- Estimate future contaminant concentrations using the current concentration and the estimated attenuation rate. The confidence interval for the attenuation rate can be used to evaluate the uncertainty in the concentration estimate (see confidence interval bands on regression analysis and Theil-Sen test).

Comparative statistical tests can also be used to evaluate remedy effectiveness. Comparative tests are most commonly used to evaluate differences in performance parameter values between groups of spatially-associated wells (that is, wells identified as inside rather than outside a treatment area). Appropriate comparative tests, such as t-test and Wilcoxon rank-sum, are discussed in Section 5.11 of this guidance. Comparisons that may be made include contaminant attenuation rates, change in contaminant concentration before and after treatment, or change in concentration of treatment compound before and after treatment.

4.4.2.1 Statistical Methods for Remediation Objectives

When conducting remedy selection or remedy effectiveness activities the following methods are most applicable.

- Estimate the rate of concentration change over time (the attenuation rate). Use the confidence interval for this attenuation rate to evaluate the uncertainty in the estimate (see regression analysis and Theil-Sen test).

- Evaluate the areal extent of remedy effectiveness by identifying the wells with higher attenuation rates. The confidence intervals for the attenuation rates can be used to determine whether the observed differences in attenuation rates are statistically significant (see confidence interval bands on regression analysis and Theil-Sen test).
- Estimate future contaminant concentrations using the current concentration and the estimated attenuation rate. The confidence interval for the attenuation rate can be used as a line of evidence to evaluate the uncertainty in the concentration estimate. However, note that any extrapolation of the attenuation rate or its associated confidence interval beyond the available data range likely includes much greater uncertainty in the projected concentrations from the statistical estimates (see confidence interval bands on regression analysis, Theil-Sen test, and Example A.2).

4.5 Monitoring

Groundwater monitoring is conducted to observe and assess characteristics of interest at cleanup, RCRA facility, or waste disposal sites. Often, monitoring is conducted on a long-term basis, sometimes for decades. Monitoring may be required even after closure of a site during post-closure monitoring. As such, monitoring may be conducted to describe characteristics at a specific location or point in time or to show how these characteristics change over time or space.

Changes in groundwater quality may have either natural or human causes. Proper evaluation of groundwater data helps you understand whether the criteria or goals of the monitoring program are met or if significant, adverse changes in groundwater concentrations have occurred. Proper design of the monitoring network depends upon the type of the site, the contaminants present, and the regulatory program. Prior to implementing a monitoring program, review well placement, parameter selection, sampling frequency, and whether or not a release has been identified. Typical activities in the monitoring stage include observing changes in concentration levels over time and space, and comparing concentrations to numerical criteria. After sufficient data are collected, it may also be possible to optimize sampling locations and sampling frequencies to improve and streamline the monitoring program (see Section 4.5.3).

The study questions that are most applicable for monitoring are shown here.

Monitoring Study Questions
2. Are concentrations greater than background concentrations?
3. Are concentrations above or below a criterion?
4. When will contaminant concentrations reach a criterion?
5. Is there a trend in contaminant concentrations?
6. Is there seasonality in the concentrations?
7. What are the contaminant attenuation rates in wells?
8. How do contaminant concentrations change with distance from the source area?
9. Is the sampling frequency appropriate (temporal optimization)?
10. Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

4.5.1 Monitoring for Concentration Changes

As discussed in Section 3.6, before implementing a monitoring program, consider the statistical design of the program and the methods that will be used to statistically analyze the measurements. These choices impact the kinds of data that must be collected and the frequency of monitoring. For example, routine, periodic groundwater monitoring lends itself to the use of prediction limits with retesting (see Section 5.4) in order to assess whether concentration levels exceed background. To do this appropriately, however, (1) data representing background concentrations must be collected from either dedicated background wells (interwell testing) or from (earlier) uncontaminated sampling events at compliance wells (intrawell testing); (2) the background concentrations should be stationary (stable, nontrending); and (3) there should be enough background observations to give the prediction limit a reasonable chance of identifying a significant change in concentrations (adequate statistical power).

No matter what evaluation methods are selected, always first graph your data on time series plots. This simple, graphical procedure can be used both to help verify that background concentrations are stable/stationary or to reveal apparent trends over time. Time series plots can also reveal the presence of seasonal or cyclical patterns, which might necessitate special data adjustments (de-seasonalization) or test methods specifically adapted for seasonal data (such as seasonal Mann-Kendall).

In cases where trends are apparent on time series plots, formal trend tests (for example, linear regression, Mann-Kendall, Theil-Sen) can be used to verify whether or not a statistically significant trend exists. Statistical trend methods are also applicable if the purpose of monitoring is to identify the rate at which groundwater contaminants are diminishing, or if attenuation is occurring more quickly in one location over another. Similarly, trend analysis may be used to evaluate the natural attenuation of a contaminant in groundwater. Evaluating and identifying trends in concentrations is

an important line of evidence to support monitored natural attenuation (MNA) as part of a groundwater remedy.

4.5.2 Compliance with Criteria

A common compliance goal during long-term monitoring is to determine whether groundwater concentrations meet, exceed, or have dropped below a numerical criterion or decision criterion. Under some regulatory programs, an extensive monitoring program may be established when a release occurs. At this point, statistical tests are used to test whether or not the concentrations exceed a specified criterion. Additionally, monitoring may be used to demonstrate that remediation activities have lowered concentrations below a criterion for cleanup. In both settings, a type of confidence interval or limit is an appropriate statistical method. But in selecting a specific method, consider first what the numerical criterion is meant to represent as a statistical quantity, for example a long-term average or an upper percentile. Decision criteria can be established as MCLs, alternate compliance limits (ACLs), background limits, risk-based concentrations for protection of human health and the environment, or other bases. Most of these criteria are designed to be long-term averages based on chronic exposures; more rarely, a criterion may be based on acute or episodic exposures and thus more akin to a concentration upper percentile. The key statistical principle is to match the type of statistical interval to the type of criterion (for example, using a confidence interval around the mean when comparing against a long-term average-based MCL).

An important corollary to this discussion is the need for multiple, independent statistical measurements with which to decide whether or not groundwater concentrations meet or exceed any criterion with a high degree of statistical confidence. To be specific, one observation below a criterion does not prove that the maximum or mean concentration of the contaminant population is below the criterion. Neither does one concentration above a criterion indicate that the decision criterion has been violated.

4.5.3 Optimization of Long-term Monitoring Networks

Optimization of a groundwater program can occur at any stage of the life cycle, especially if it makes the program more accurate, efficient, and cost-effective (see, for instance, the resources and options compiled by the Navy's Facilities Command (NAVFAC) in its *Optimization Roadmap* (US Navy 2013a). Statistical optimization of monitoring networks is generally practical during long-term monitoring when a larger amount of data has been accumulated or the number of wells is more extensive, or both. In that setting, the optimization objective is to create efficient data collection—in which the right amount of data are collected in order to make accurate decisions in a cost-effective manner.

Statistical methods can be used to judge whether a monitoring program is optimized. At a very high level, this involves (1) estimating the degree of statistical correlation or redundancy between sampling events or sample locations, or both, and (2) estimating the statistical uncertainty associated with trends or spatial maps of concentrations. Similar sample results from neighboring wells or closely-timed events indicates a positive correlation among the observations and possibly

statistical redundancy. An optimized sampling and network design tends to show little redundancy while retaining sufficient data to enable accurate and defensible decisions. Importantly, statistical optimization can lead to either more or fewer monitoring wells, sampling events, or monitored chemicals, depending on what best design meets the project goals. Results of any optimization should also be compared to what is known or hypothesized about the site in the (CSM).

Typically four modes of optimization are either directly statistical or substantially affect statistical decision making:

- *Choice of monitoring parameters (chemicals)*. The more parameters that must be collected and statistically analyzed, the greater the cost of the monitoring program, the greater the risk of making false positive decision errors, and the greater the site-wide false positive rate (SWFPR; see Section 3.6). In an optimal program, parameters that are not directly or indirectly related to possible contaminant sources or waste composition, or which are primarily nondetect and therefore statistically non-informative, may not be useful for routine monitoring. Furthermore, as discussed in Section 3.6, if a regulatory program requires or recommends a fixed SWFPR, the fewer the parameters, the greater the statistical power associated with each of those tests for detecting real changes in groundwater quality.
- *Choice of data collection (see Section 3.6)*. For some parameters, it may be feasible to use less expensive field screening techniques or temporary sampling points (for example, Hydropunch) in order to collect a larger number measurements over a broader area, while simultaneously reducing costs relative to traditional sampling and laboratory analysis of dedicated wells. Often, such data may be less precise than traditional groundwater measurements, but the statistical advantage is a much larger sample size, leading to greater overall statistical power and decision accuracy.
- *Temporal optimization (see Section 3.6 and Section 5.8)*. This statistical approach aims to optimize sampling frequencies, using one of several methods. One method, known as cost effective sampling (CES), uses linear trend estimates and the statistical uncertainty of those trends to bin wells into less and more frequent sampling categories. Modifications of the CES approach have been incorporated into software tools like MAROS and the 3-Tiered Monitoring and Optimization Tool (3TMO). Another method is iterative thinning, based on constructing a trend and then determining how much of the sample data can be removed or 'thinned,' yet still allow the original trend to be accurately reconstructed. The greater the percentage of data removed, the greater the degree of redundancy, and the less sampling required for an optimal sampling frequency.
- *Spatial optimization (see also Section 3.6 and Section 5.14)*. This approach attempts to optimize the number and placement of wells in a monitoring network. Different approaches seek to measure either (1) statistical redundancy between sampling points to assess whether some of those locations can be dropped from routine monitoring, or (2) statistical uncertainty and how it varies across the site. Areas with high uncertainty and no or few wells are candidates for adding new sampling locations. In general, both of these tasks rely on geostatistical techniques involving a significant degree of complexity. Specialized software tools such as GTS,

Summit Tools, and VSP (in addition to those referenced above) have been developed to perform these statistical analyses.

4.5.4 Statistical Methods for Monitoring Objectives

When implementing groundwater monitoring programs, the following methods are most applicable.

- Calculate the monotonic trends of concentrations over time at a single location to identify statistically significant concentration trends (see linear regression, Mann-Kendall trend test, and Example A.2).
- Estimate attenuation rates (rates of change), and use of confidence intervals (uncertainty) for the attenuation rate (see linear regression, Theil-Sen trend line).
- Compare the estimated attenuation rates in two wells by comparing the slopes. This comparison does not, however, demonstrate the relationship between the wells (see confidence interval bands on linear regression, Theil-Sen trend line, and time series plots).
- Calculate a confidence interval for a monotonic trend around the criterion to estimate when compliance can be reached (see confidence interval bands on linear regression and Theil-Sen trend line).
- As appropriate, consider the adequacy of sampling (both events and wells) to meet project objectives. Relevant tools are iterative thinning, CES, and spatial analyses.

4.6 Closure

Closure is the final stage of the project life cycle and therefore is subject to extra scrutiny. At this stage in the process, data planning and collection should have been managed through a systematic planning process, and the CSM is assumed to be complete for the purposes of making a final determination on whether monitoring may be permanently discontinued and the site closed. This decision point may be reached at any time during the life cycle process (for example, during site characterization, remediation, or monitoring). Significant variation occurs across regulatory programs, but in general, when contaminants are no longer detected in any wells for several sampling events or over a specified period of time, the remedial goals are deemed complete and the groundwater is no longer considered contaminated. However, in instances where contaminants decrease but remain measurably present, or are present in natural or anthropogenic background, statistics can support a closure decision.

Given the importance of this decision, managers must have a high degree of confidence that the data fully support closure. Closure should verify that site contaminants are no longer present in the groundwater, or are not present at concentrations that pose an unacceptable risk to human health or the environment. In cases where concentrations of contaminants are allowed to remain (such as under institutional or engineering control scenarios), trend analysis results may be used to show that contaminants will not migrate or increase concentrations outside of the defined boundaries of the controlled area.

Following remediation, formal statistical testing will usually involve an upper confidence limit around the mean or an upper percentile compared against a criterion. The overriding concern in corrective action is that remediation efforts must have sufficient statistical proof to be declared successful. Since groundwater is now presumed to be contaminated, a facility should not exit corrective action until there is sufficient evidence that contamination has been abated (see Chapter 7.2, Unified Guidance).

By the time a site reaches the closure life cycle stage, the evaluation assumes that contamination exists. Therefore, the statistical approach may involve comparing an upper confidence limit of the data to a criterion. The upper confidence limit (UCL) should lie below the criterion to accept the hypothesis that concentration levels support closure. If the entire confidence interval (considering both the lower and upper confidence limits) lies below the criterion, there is statistically significant evidence that the true value of the parameter (for instance, the mean) is less than the criterion. When the confidence interval straddles the criterion, the correct decision is uncertain within the stated confidence level. The true value of the parameter might be less than or greater than the criterion and no clear decision with high statistical confidence is possible (see Chapter 5, Unified Guidance).

The study questions that are most applicable for closure are shown here.

Closure Study Questions
1. What are the background concentrations?
2. Are concentrations greater than background concentrations?
3. Are concentrations above or below a criterion?
4. When will contaminant concentrations reach a criterion?
5. Is there a trend in contaminant concentrations?
6. Is there seasonality in the concentrations?
7. What are the contaminant attenuation rates in wells?
8. How do contaminant concentrations change with distance from the source area?
9. Is the sampling frequency appropriate (temporal optimization)?
10. Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

4.6.1 Compliance with Criteria

Statistical tools for comparison of groundwater concentrations to a fixed criterion include confidence intervals or a one-sample hypothesis test. For this comparison to be valid, the sample population must be stable, with no increasing or decreasing trends. If groundwater concentrations are

above the criterion and are changing over time, a trend analysis should be conducted. A confidence band around the trend line can be estimated and compared to the criterion to determine when compliance can be reached. The site owner may also evaluate other closure options such as closure with institutional or engineering controls, or should take additional action (remedial or other) to address the remaining contamination.

The choice of confidence interval should be based on the type of fixed criterion to which the groundwater data will be compared. State or federal regulatory programs determine the appropriate statistical parameter for comparison to a criterion. If a mean- or median-based parameter is chosen, fairly straightforward confidence interval testing is implemented. If the maximum or not-to-exceed criterion is the regulatory goal, then the program must identify a specific upper proportion and confidence level that the criterion represents. If nonparametric upper proportion tests must be used for the maximum or not-to-exceed criterion, then it will be very difficult to document compliance (see Chapter 5, Unified Guidance) because of the large number of samples required. Care must be taken that the confidence interval for an upper percentile concentration (such as the upper 95th percentile for a maximum or not-to-exceed criterion) should not be compared to a confidence interval constructed around the arithmetic mean (under a mean or median-based criterion); see Chapter 5, Unified Guidance.

4.6.2 Trends Toward Compliance Criteria

To show compliance with a fixed criterion, groundwater concentrations must not be increasing with time. Check for trends and make corrections as needed to ensure validity of the statistical evaluation. If contaminant concentrations remain above designated fixed criteria, then demonstration of a stable or downward trend, combined with institutional controls, or engineering controls may be sufficient to justify closure for the site.

4.6.3 Statistical Methods for Closure Objectives

When evaluating a site for closure, the following methods are most applicable.

- Compare a data set to a numerical criterion. A criterion may be an MCL, risk-based or fixed background concentration. Comparisons to a criterion are generally one-sample tests based on confidence interval testing against a fixed criterion, well by well and chemical by chemical. Pooled data from multiple wells can be compared to a numerical criterion if the numerical criterion was developed based on a limit consistent with such comparisons.
- Evaluate the UCL against the criterion. Choose a confidence interval consistent with the basis of the criterion. If the criterion represents an average concentration, the UCL of the mean or median concentration of the monitoring wells should not exceed the criterion. If the criterion represents an upper percentile or maximum, no more than a small, specified fraction of the individual concentration measurements should exceed the criterion (see confidence intervals for more information).
- In some cases, background is used to demonstrate that a site is suitable for closure. Test to see if contaminant concentrations are not different from background. If a fixed background

value has been established, use single-sample confidence interval methods to compare concentrations at closure to background. Use two-sample methods for interwell comparison, comparison of compliance wells to background wells, and comparison of compliance wells to established site background. For more information, see Shewhart-CUSUM control charts (intrawell), tolerance limits, prediction limits, t-test or Wilcoxon rank sum (Mann-Whitney), one-way ANOVA, and Kruskal-Wallis test (interwell).

- To show compliance with a fixed criterion using standard confidence intervals, concentrations must not be increasing. If the data are not stable, a trend should be estimated along with a confidence band around the trend. Then compliance can be documented if a point in time occurs at which the confidence band drops below the criterion and remains so. Alternatively, if contaminant concentrations remain above designated fixed criteria, demonstrating a stable or downward trend, combined with institutional controls, engineering controls, or land use controls, may justify closure for the site (see linear regression and Theil-Sen trend line).
- Evaluate the areal extent of remedy effectiveness by identifying the wells with higher attenuation rates. Use the confidence intervals for the attenuation rates to determine whether the observed differences in attenuation rates are statistically significant (see linear regression and Theil-Sen trend line).
- Estimate future contaminant concentrations using the current concentration and the estimated attenuation rate. Use the confidence interval for the attenuation rate to evaluate the uncertainty in the concentration estimate (see linear regression and Theil-Sen trend line).

5.0 STATISTICAL TESTS AND METHODS

This section includes practical information for some of the common statistical tests and methods used to evaluate groundwater data for cleanup sites. When evaluating data and making decisions at sites with impacted groundwater, project managers must consider a variety of information such as site history, site and area geology and hydrogeology, and data for other media. Statistics can provide an additional "line of evidence" under a multiple lines of evidence approach to decision making.

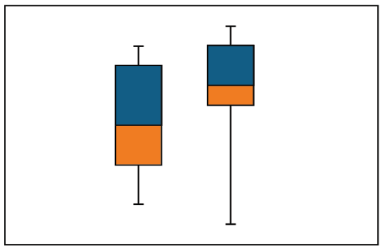
Statistical methods are grouped in this section based on their application. Information that applies to the entire group of methods is presented in the beginning of each subsection and method-specific information follows. Where USEPA's Unified Guidance describes a method, a reference is provided. USEPA guidance regarding the application of certain tests has been modified over time; information relating to these changes and modifications is provided in Appendix B, Unified Guidance.

Before applying statistical methods, you must confirm that the data collected is appropriate to the question being posed. For example, different data needs might be required for health and safety evaluations as compared to plume stability or mass flux across a property boundary. Site data should be reviewed for both temporal and spatial applicability to the question at hand. Different degrees of confidence in data may be acceptable, depending upon client or regulatory needs. The statistical power, accuracy, and statistical certainty or confidence (see Section 3.6.1) of an analysis depend on the number of statistically usable measurements (also termed the sample size). As a general rule for parametric tests and many nonparametric ones as well, the larger the sample size the greater the power and the smaller the decision error risk (false positives and false negatives). Although recommended minimum sample sizes are provided for many statistical tests discussed in this document, various project needs or regulatory frameworks may require a sample size different than that recommended.

Section 3 provides a discussion of the general statistical approach, including systematic planning processes and conceptual site models, defining a target population, data quality, exploratory data analysis, common statistical assumptions, and statistical design. Review Section 3 when applying the statistical tests and methods included in this section to a specific project.

Section 6.0 of this document includes information about data management for implementing the tests and methods described here using software tools. Appendix D includes descriptions for some commonly used software packages. Appendix F includes information about checking the underlying assumptions of statistical tests.

5.1 Graphical Methods



Graphs are powerful data evaluation tools. They provide quick, visual summaries of essential data characteristics. A few simple plots can replace complex statistical equations or tests to interpret environmental data. Box plots, histograms, and normal probability plots are examples of graphs that are commonly used to display environmental data. These graphs can provide information about concentration ranges, shapes of distributions, extreme values (outliers), relationships between different data sets, and trends (increasing, decreasing, and cyclic). Because graphical methods are qualitative, however, they may not be appropriate as a stand-alone technique to make inferences or support conclusions.

Graphical methods are typically used with quantitative statistical evaluations. Graphical methods provide information that may not be otherwise apparent from quantitative statistical evaluations, so it is a good practice to evaluate data using these methods prior to performing statistical evaluations. Graphical methods are also a key component of exploratory data analysis (EDA). In EDA, various graphical techniques are used initially to display data for qualitative assessments prior to selecting appropriate statistical tests. Brief descriptions of some useful statistical plots are presented in the subsections below.

5.1.1 Time Series Methods

Time series methods graph data of interest, such as concentration, on the y-axis versus time on the x-axis. When plotting multiple series, it may be helpful to standardize or normalize data prior to plotting. Time series plots include lag-plots, correlograms, and variograms.

Lag-plots. Lag plots display observations for a time series against a later set of observations, or against the difference between the two (for example, a plot of x(t) versus x(t-1). If the lag plot exhibits a linear pattern, it follows that data are nonrandom and that you may need to use an autoregressive model. If no patterns are discernible in the lag plot, data are likely random. Plotting data for a greater number of observational periods or lags can be helpful in evaluating data for seasonality. An example of a lag plot is provided in Figure 5-1.

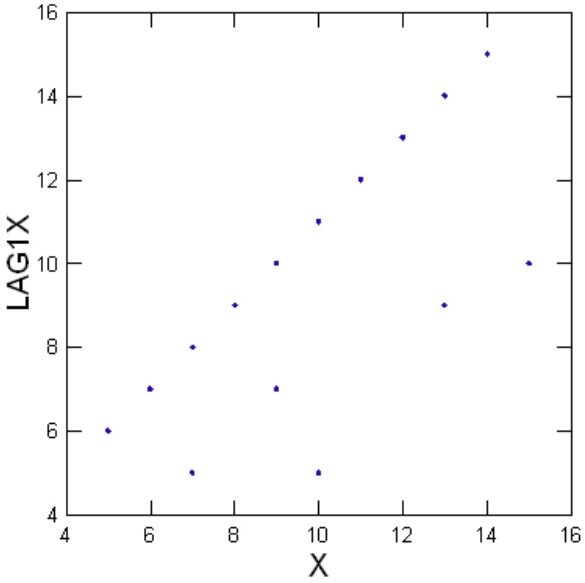


Figure 5-1. Lag plot example.

Correlograms. Correlograms are commonly used to evaluate the randomness in a data set. Correlograms (or, autocorrelation plots) display the correlation between two variables (for example, a plot of the autocorrelation function versus the lag) and provide a graphical evaluation of temporal dependence. Autocorrelations may be calculated for data values at varying time lags. If the data are random, the autocorrelation value should be near zero for all time lags (i.e., the autocorrelation plot at time x+1 should not be significantly different than the plot for time x+2, and so forth). A sample correlogram displaying nonrandom data are provided as Figure 5-2.

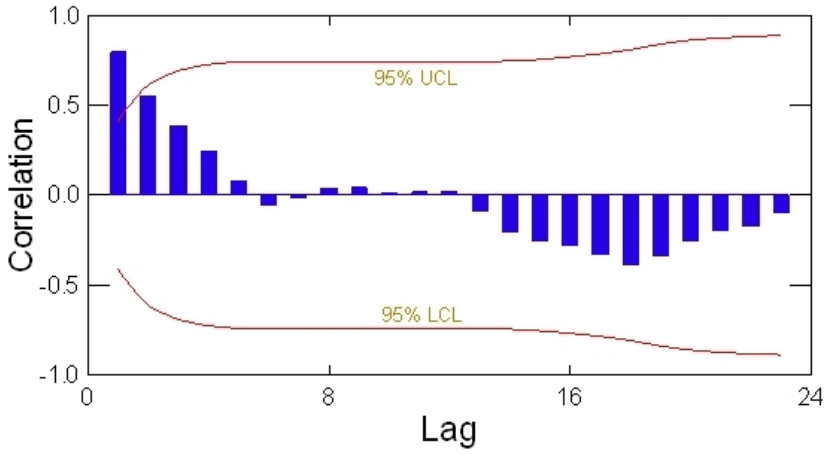


Figure 5-2. Correlogram example.

Variograms. Variograms (also known as a semi variogram) plot a variogram coefficient associated with a selected model of temporal or spatial correlation versus data from different lags and angles in an effort to fit the selected model to the data. The selected model is subsequently used in kriging for contouring of the data. An example of a variogram is provided in Figure 5-3.

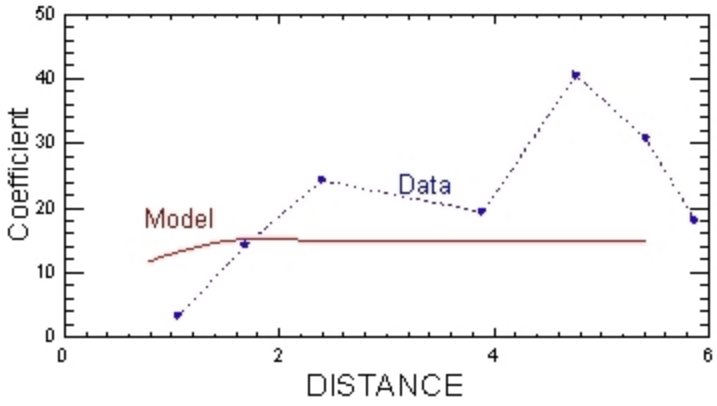


Figure 5-3. Variogram example.

Time series plots show the following:

- concentration trends over time
- lack of randomness
- changes in location (for example, of a plume or of the highest concentrations)
- degradation (when concentration vs. time plots are viewed for a contaminant and its degradation by-products)

Figure 5-4 illustrates a time series plot with data from two monitoring wells over seven years.

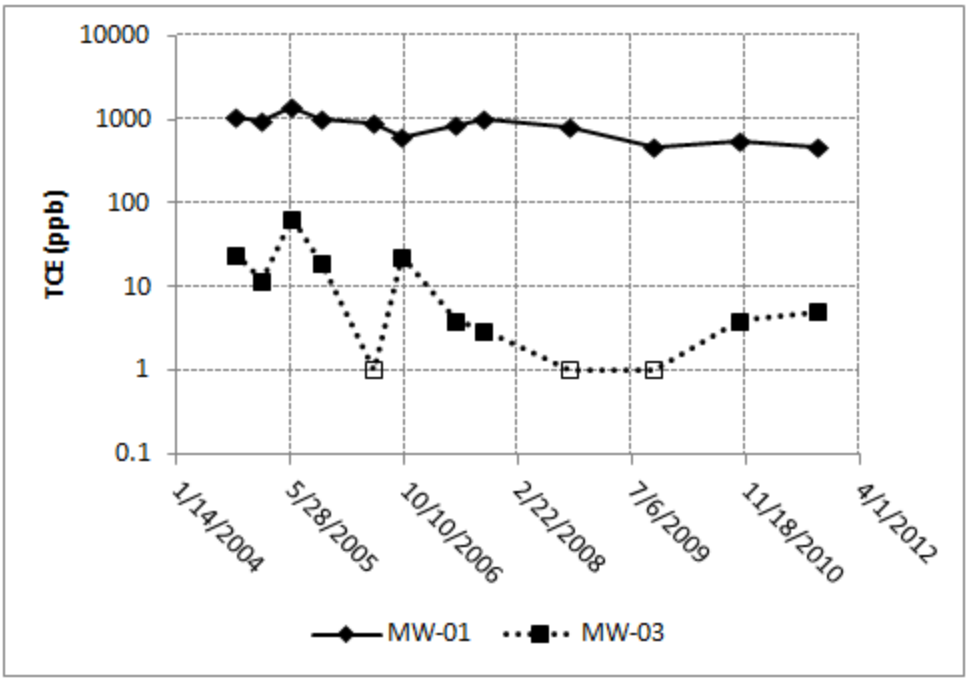


Figure 5-4. Time series plot example.

Applications and Relevant Study Questions

- Study Question 1: What are the background concentrations?
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?
- Study Question 8: How do contaminant concentrations change with distance from the source area?
- Time series methods may also be used to investigate stationarity, which is an underlying assumption for many statistical methods.

Assumptions

Data come from a consistent set of representative wells over a series of sampling events.

Requirements and Tips

- Assign a value to nondetects.
- Use different symbols to depict nondetects versus measured data values on the plot.
- Be sure that data are collected with sufficient frequency and at a sufficient number of points to answer the questions of interest. For instance, annual monitoring would be insufficient to evaluate seasonal variation, but may be sufficient to identify directional trends. A minimum of two measurements are needed, but a greater number of measurements increases the degree of confidence in detecting patterns. Also consider whether the series of monitoring events is sufficient to be representative of site conditions. For example, a series of four monitoring events conducted one month apart, or four annual monitoring events, may not be representative if the plume is affected by seasonal effects.
- Consider the scale of each axis of plots (cover the full range of data; highlight fluctuations by shrinking or spreading an axis as needed; consider use of a log scale).
- When comparing time series, use comparable scales. Standardizing or normalizing each variable might be necessary for plotting multiple chemicals on similar scales for subsequent comparison. Use of a log scale is recommended when data cover a large range of values (for instance, when graphing concentrations near a source area and at distal portions of a plume).
- If the wells selected for long term monitoring are not representative of the plume, the point of exposure, or other site characteristic, then statistical representations of data will also not be representative of the site conditions.

Strengths and Weaknesses

- These plots are quick and easy to construct using ordinary spreadsheet programs like Excel.
- These plots are not quantitative. They are typically used in conjunction with other quantitative information.
- These plots are useful for quickly and easily assessing patterns in data over time.

Further Information

A description of how to construct a time series plot is found in Chapter 9.1, Unified Guidance.

Chapter 14.2.1 provides an example of how to construct a time series plot for multiple series (parallel time series plot).

5.1.2 Box Plots

Box plots divide data into four groupings, each of which contain 25% of the data. The box most typically depicts the 25th (bottom of the box), 50th (horizontal line within the box) and 75th (top of box) percentile values while the whiskers can be selected to represent various extremes such as 1.5 times the interquartile range (Tukey 1977), or 0% and 100% values. Points falling outside of the range depicted by the whiskers are plotted as individual points; you can evaluate these points as potential outliers. The mean and the 95% upper confidence limit (UCL) and lower confidence limit (LCL) are often depicted on a box plot as well.

The extent of the box is the interquartile range, which is the range of values between the 25th and 75th percentiles. A common convention is for whiskers to extend to 1.5 times the interquartile range on either side of the box. In this case, values between 1.5 and 3 times the interquartile range outside the whiskers are typically considered "mild" outliers while values greater or less than 3 times the interquartile range are considered "extreme" outliers. Graphing two data sets on side-by-side box plots provides an easy method of data comparison.

Figure 5-5 illustrates a box plot.

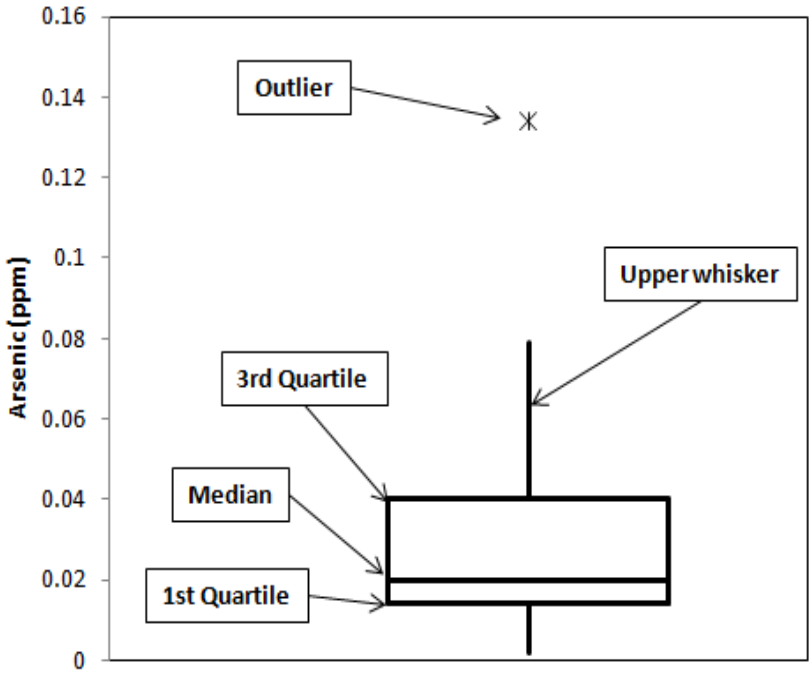


Figure 5-5. Box plot example

Applications and Relevant Study Questions

- Two subsets may be compared to evaluate spatial variability.

- Background data from different sources can be evaluated on side-by-side box plots to confirm that they represent a single data set.
- A comparison of a box plot representing background data to a box plot of data from individual wells may be used to evaluate whether concentrations from a particular well are above background concentrations.
- Box plots are useful for initial identification of potential outliers.
- Box plots are also useful for investigating and visualizing the mean value (centerline of the box), the variation or spread of the data (interquartile range or height of the box), the symmetry (sizes of box halves and whiskers), and the skewness of the data (the relative size of the box halves).
- Study Question 1: What are the background concentrations?
- Study Question 2: Are concentrations greater than background concentrations?

Assumptions: None

Requirements and Tips

- Assign a value to nondetects.
- This method is most useful with data sets containing eight or more values.

Strengths and Weaknesses

- Box plots are a quick, convenient way to view the distribution of a data set.
- These plots can be used for any type of data distribution.
- Box plots are a simple graphical method; results can be readily interpreted.
- This method is useful for comparing data sets side by side.
- The use of box plots for purposes such as identification of outliers is not quantitative.
- Generally, software is required to display box plots, although it is possible to construct them in spreadsheet programs with some effort.
- Box plots illustrate the characteristics of data for only a single variable.
- Depending upon the software used to construct the plot, a box plot may not show all individual data points.
- Identification of outliers depends on the extent of the tail, is fairly arbitrary, and not conclusive.

Further Information

Refer to Chapter 9.5 and Chapter 12.2, Unified Guidance. An example of an application of box plots may be found in Chapter 9.2, Unified Guidance.

5.1.3 Scatter Plots

Scatter plots display the relationship between two or three variables when comparing data sets consisting of multiple observations per sampling point. Linear relationships will manifest in points clustering about a straight line. Figure 5-6 illustrates a scatter plot.

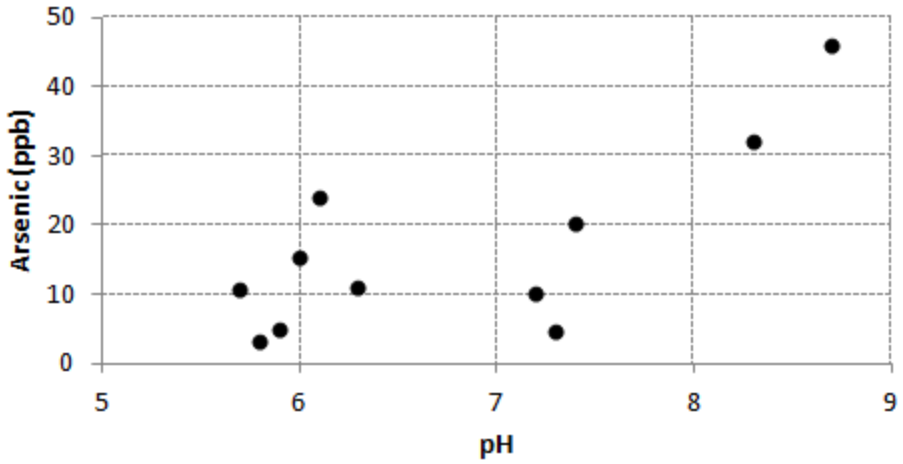


Figure 5-6. Scatter plot example.

Applications and Relevant Study Questions

- Evaluate the relationship of two or three variables to one another.
- Identify potential outliers.
- Identify clustering of data.
- Determine if the concentrations of contaminants are related in a definable way.
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?

Assumptions

- The data range is sufficiently large to be representative of the data set.
- X and Y values are not affected by outside factors. See Section 5.1.1: Time Series Methods .

Requirements and Tips

- Data sets should consist of multiple observations per sampling point and a sufficiently large data range.
- Assign values to nondetects.
- Assign different symbols to nondetect values.
- It is possible for variables with non-linear relationships to appear linear if the data range is small.

Strengths and weaknesses

- Scatter plots are a simple graphical method and results can be readily interpreted.
- This method is useful for comparing data sets side by side.
- The use of scatter plots for purposes such as identification of outliers or evaluation of trends is not quantitative.

- No special software is needed to create two dimensional plots; some software can plot three axes.
- Scatter plots only show relationships between two (or three) variables on a given plot.
- X and Y values may appear to have no clear relationship when influenced by an outside factor that was not taken into consideration.

Further Information

See Chapter 9.4, Unified Guidance for further information and a sample problem using scatter plots.

5.1.4 Histograms

Histograms present data in terms of bars of height (Y) in relation to a parameter (X), permitting a comparison of the shape and size of the plot, and of the placement of the plot along the x-axis.

Figures 5-7 illustrates a bimodal distribution of data in a histogram.

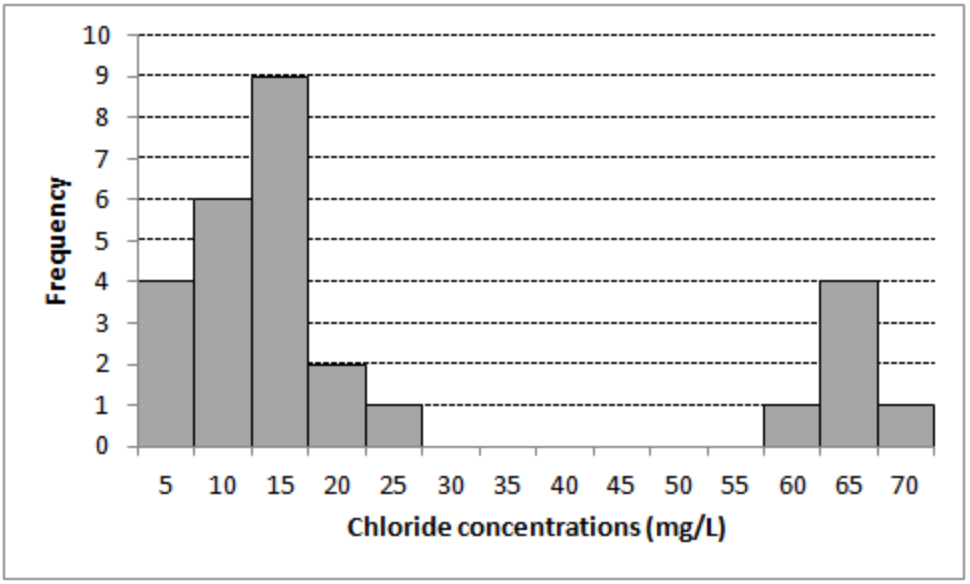


Figure 5-7. Histogram example (bimodal distribution).

Figure 5-8 illustrates a non-normal and skewed distribution of data in a histogram.

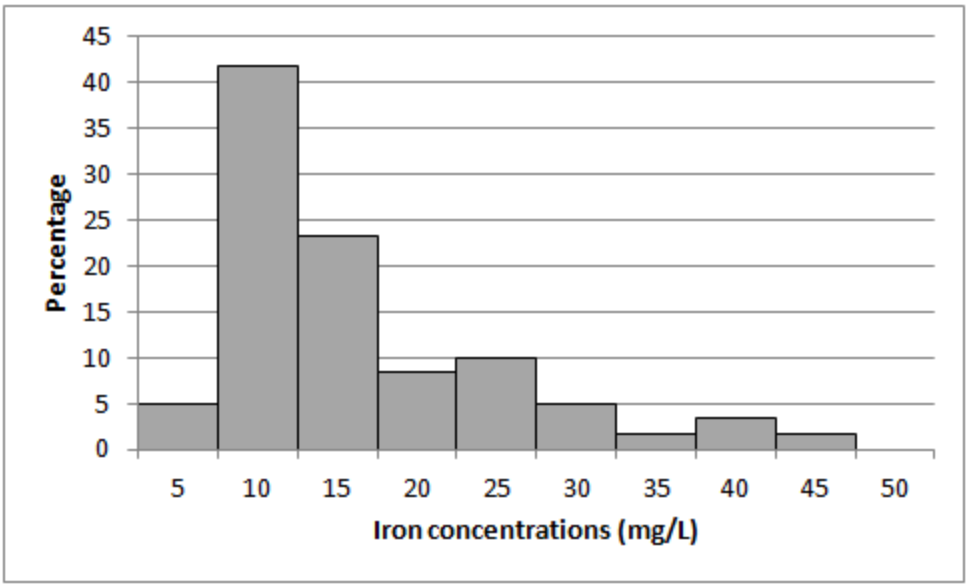


Figure 5-8. Histogram example (non-normal and skewed distribution).

Applications and Relevant Study Questions

- Histograms can be used to identify whether data are representative of a single population (one peak) or whether data may be representative of two separate populations (such as background data and data representing site contamination).
- This method is useful in EDA to evaluate the underlying data distribution.
- Study Question 1: What are the background concentrations?
- Study Question 2: Are concentrations greater than background concentrations?

Requirements and Tips

- This method is best applied to data representing a snapshot in time (as opposed to continuous measurements).
- Modifying the bin size can affect the shape of the plot.
- When comparing histograms for multiple data sets, consider placing the histograms one above another rather than side by side.

Assumptions: None

Strengths and weaknesses

- Construction of histograms does not require highly specialized software and is relatively quick and simple.
- Histograms provide a quick and easy method to investigate the skewness and symmetry of data.
- The accuracy of the visual data representation provided by histograms depends on the bin size selected for the plot (x-axis).
- This method does not provide a good representation of the center of the distribution.

- Y-axis data can be plotted as counts (for example, number detections) or as a percentage (for example, percent of detections).

Further Information

See Chapter 9.3, Unified Guidance for further information and an example problem.

5.1.5 Probability Plots

Probability plots help to evaluate how well data fit a theoretical distribution, such as a normal distribution, or gamma distribution. Probability plots express the theoretical distribution as a straight line and departures from the distribution appear as departures from the straight line. Data skewness or asymmetry, presence of outliers, and heavy tails of the data distribution (non-normal distribution) are obvious on probability plots. If the data do not fit the selected distribution, data can be transformed using a lognormal or other transformation in order to determine whether data fits an alternative distribution. A quantile-quantile plot may be used to compare two empirical distributions.

To generate probability plots, order the data, and calculate matching percentiles from the normal distribution. Plot the ordered data against the percentiles and examine the plot for a straight-line fit. The straightness of the plot indicates how closely the data fit a normal distribution. If all of the raw data closely follow a straight line, the suspected outliers are probably part of the same distribution and should not be considered outliers. Points that appear off of a linear pattern in the rest of the data may be outliers; however, be aware that other reasons, such as non-normal data, can also explain nonlinearity.

Figure 5-9 illustrates a data set as a probability plot. Figure 5-10 presents the same data in a histogram. Figure 5-11 presents the logarithms of the same data as a probability plot and Figure 5-12 presents the histogram of the log transformed data.

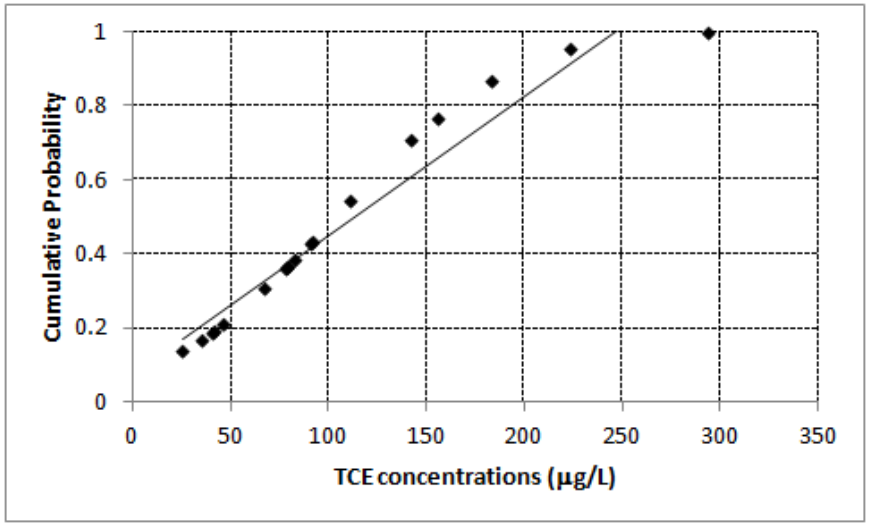


Figure 5-9. Data set as a probability plot.

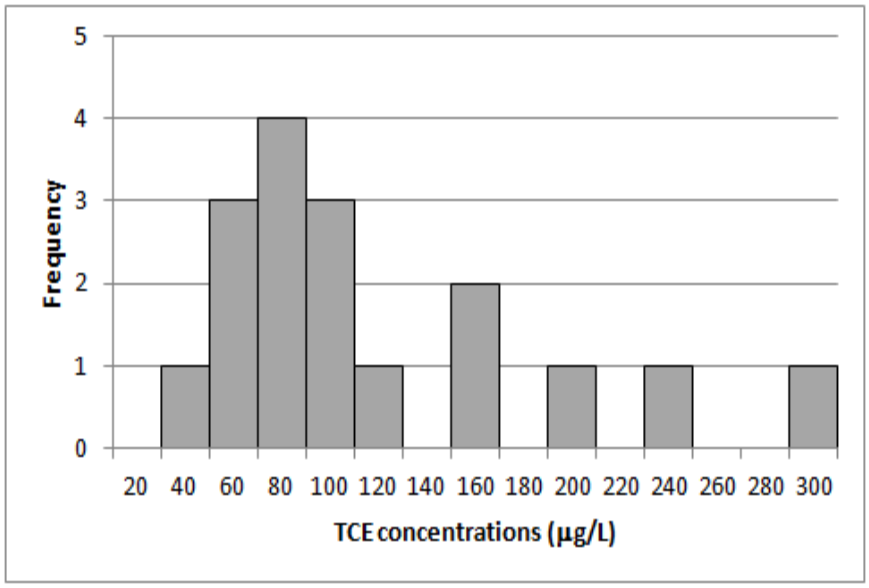


Figure 5-10. Data set as a histogram.

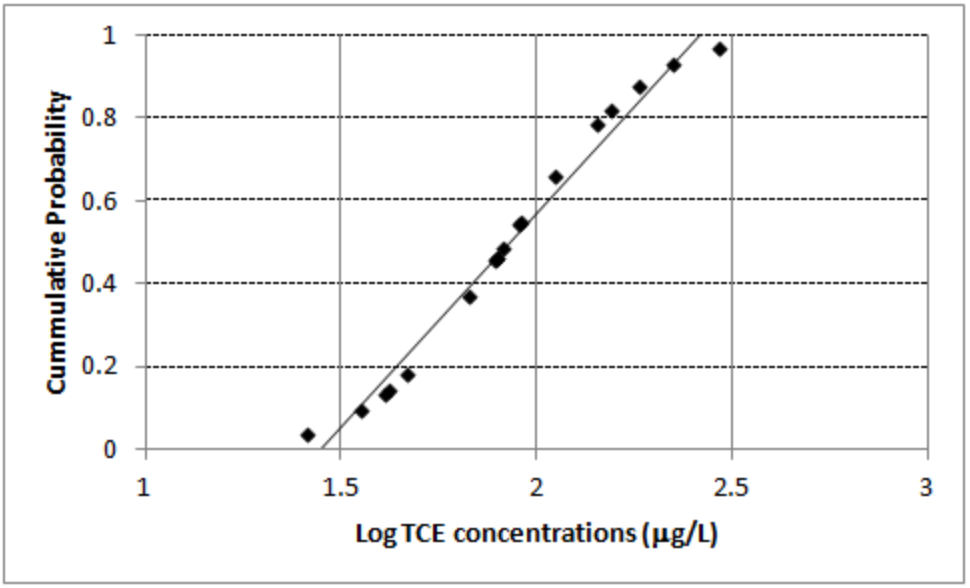


Figure 5-11. Logarithms of data set as a probability plot.

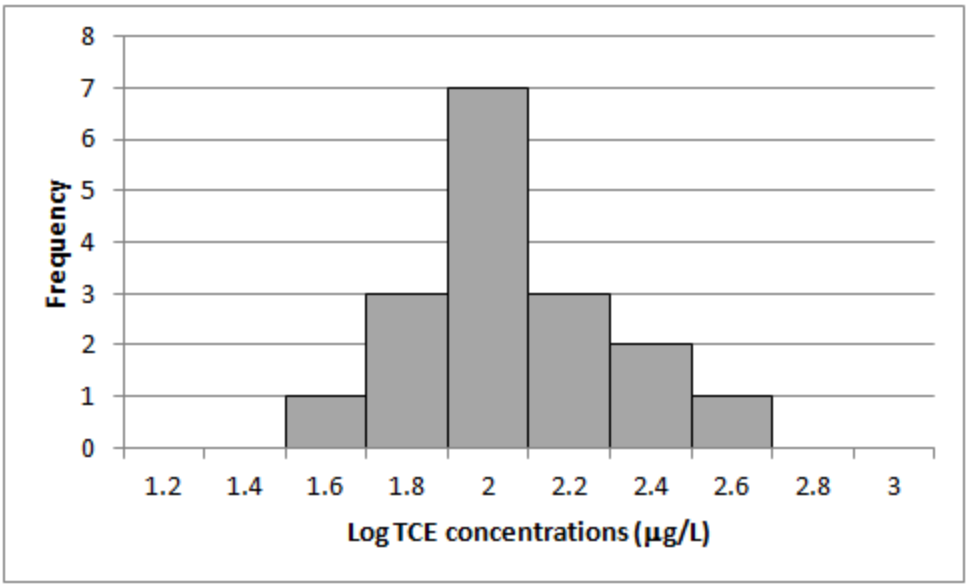


Figure 5-12. Histogram of log-transformed data.

Applications and Relevant Study Questions

- Probability plots can be used to identify whether data are representative of a single population or whether data may be representative of two separate populations (for example, background data and data representing site contamination).
- These plots can help to evaluate underlying data assumptions prior to application of other statistical tests.
- How well do data fit a theoretical distribution?
- What is the reason for departure of the data from the theoretical distribution?
- Study Question 1: What are the background concentrations?
- Study Question 2: Are concentrations greater than background concentrations?

Assumptions

Data follow a single distribution, typically the normal distribution; it is possible to use this test with data that can be normalized, such as lognormal data, or to evaluate other distributions, such as a gamma distribution.

Requirements and Tips

- Although examination of the probability plot will help assess whether the data are normal or not, you should confirm normality using another test. Further tests for normality are correlation coefficients and Shapiro-Wilk tests.
- If there are nondetect data, simple substitution will result in nonlinearity, so use an appropriate method for dealing with censored data, such as ROS, maximum likelihood estimation (MLE), or Kaplan-Meier.
- If the data do not appear to be linear, try normalizing the data by log-transforming the data and creating another probability plot.

- If the log-transformed data fits a straight line with no points off the line, the data are lognormal and there are probably no outliers.
- If neither of these plots fit a straight line and one or more data points appear to be off the line of the rest of the data, remove these points and re-plot the data.
- If removal of data results in a straight line then this is evidence that the removed data do not follow the same distribution as the rest of the data.

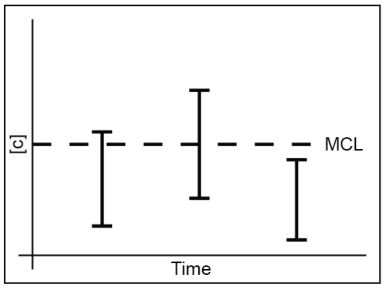
Strengths and weaknesses

- While departures from the theoretical distribution are easy to identify, you must evaluate the significance of the departure.
- Probability plots are useful in comparing characteristics of groups of data, such as skewness.
- Lognormal data can be transformed and the test conducted on the transformed data. If the data still do not follow a straight line, test whether the removal of some points results in a straight line. Probability plots offer an excellent graphical method for identifying normal data and data points that lie outside the normal distribution. Using probability plots for identifying outliers is only applicable for distributions that have been verified.

Further Information

A description of how to construct a probability plot is found in Chapter 8.3, Chapter 9.5, and Chapter 12.1, Unified Guidance.

5.2 Confidence Limits



One of the strengths of statistics is that they quantify uncertainty about data. Confidence limits (sometimes called "confidence intervals") clearly illustrate that uncertainty, thus, regulators often require them. For example, confidence limits may be used to compare groundwater monitoring data to a fixed threshold, such as a compliance criterion, or for placing an upper limit on background. Confidence limits are the maximum and minimum values bracketing the statistic of interest (usually the arithmetic or geometric mean) based on the distribution of the data (usually the normal or lognormal distribution) at a certain confidence level (usually 95%). In other words, confidence limits are the maximum or minimum values above or below which you are confident (at a selected confidence level) that the statistic will occur. Confidence limits can be parametric or nonparametric. For the calculation of parametric confidence limits, the underlying statistical distribution must be known in order to select the appropriate confidence limit. Certain more robust methods (e.g., calculation of robust confidence limits) may permit the calculation of confidence limits without removal of outliers within background data (USEPA 1999).

To illustrate confidence intervals, suppose you must compare the mean concentration of a contaminant in groundwater at a site to a protection standard. If you know the true mean of the entire population–all the groundwater at the site–then you can simply compare that value to the standard. The true mean, however, is never known. Usually, the mean is estimated from a measured sample (that is, the data set of groundwater concentration results), which is often a very small subset of the entire population. Estimating a parameter based on a sample always results in some uncertainty. A good way to account for this uncertainty is to estimate the upper and lower limits for the true mean based on the distribution of the data, the spread of the data, and a desired confidence level. For a given distribution, the confidence interval estimates a data interval within which, the actual statistic of the true population will fall, for a selected confidence level. For example, a 95% confidence interval of the mean chemical concentration in groundwater at a site means that if a group or network of wells was sampled 100 times, 95 of those times, the measured mean will fall within the calculated interval if the distribution model fits the data.Table F-1 includes information about checking assumptions for confidence limits. Confidence limits and tolerance limits (see Section 5.3) are distinct, even though in some cases the one-sided upper limits for both methods are equivalent.

Parametric and Nonparametric Confidence Limits

Confidence limits can be parametric or nonparametric. For the calculation of parametric confidence limits, the underlying statistical distribution must be known in order to select the appropriate confidence limit.

A confidence interval for a given data set may be calculated based on the sample statistic of interest, typically a mean or percentile, the sample standard deviation, the data distribution, and a selected level of confidence.

5.2.1 Determining Which Confidence Limits Are Needed

When using confidence limits, you must determine if one-sided or two-sided confidence limits are needed. This determination ensures that confidence limits are not over- or underestimated. If you are comparing data to a criterion and only need to know whether concentrations fall above or below a criterion, then only one of the confidence levels is of interest. The two-sided approach is appropriate when assessing the uncertainty of hydraulic parameters, such as the hydraulic conductivity estimates of a well.

Confidence intervals are often applied in the following scenarios:

- Compliance or assessment monitoring where it is assumed that concentrations do not exceed a criterion and you must determine if concentrations have exceeded the criterion. In this case, a calculated lower confidence level (LCL) exceeding the standard, indicates confidence that the measured concentrations are above the criterion.
- Corrective action sites where it is assumed that concentrations exceed a criterion and confirmation must be provided that the site media have been remediated to concentrations below

the criterion. In this case, a calculated upper confidence level (UCL) below the criterion, indicates that the criterion has been met.
- To determine the strength of evidence for an upward or downward trend in data, two sided confidence limits may be calculated for the estimated slope of the trend line. The calculation of two-sided confidence limits that do not include the value zero, are indicative of evidence of a trend at the selected confidence level (such as 95%).

Before calculating confidence limits, the data should be examined to evaluate what distribution fits the data, whether the underlying assumptions for constructing confidence limits are valid, and whether the selected confidence level is appropriate for the planned application (that is, the question you are trying to answer). Confidence limits may be constructed in several ways, depending on the distribution of the data and the question of interest, when assessing environmental data. Some common applications of confidence limits are listed below:

- Confidence interval around a normal mean. See Confidence Interval Around a Normal Mean (Chapter 21.1.1, Unified Guidance).
- Lognormal geometric mean. See Confidence Interval Around a Lognormal Geometric Mean (Chapter 21.1.2, Unified Guidance).
- Lognormal arithmetic mean. See Confidence Interval Around a Lognormal Arithmetic Mean (Chapter 21.1.3, Unified Guidance).
- Upper percentile. See Confidence Interval Around an Upper Percentile (Chapter 21.1.4, Unified Guidance).

5.2.2 Confidence Interval Around a Normal Mean

If the data are normally distributed, if the data pass normality tests (such as probability plots or the Shapiro-Wilk test), or are reasonably symmetric, choose the confidence interval around a normal mean. This method estimates the upper and lower confidence limits (UCL and LCL) around the arithmetic mean of a data set based on an underlying normal distribution model. Construct a one-sided test instead of a two-sided test if that is most appropriate. These confidence intervals are most appropriate when comparing concentration means to criteria.

Applications and Relevant Study Questions

- This method is used when comparing normally-distributed concentrations to a criterion that is based on a mean, as is common in risk assessment.
- Confidence limits may be used to evaluate whether a mean concentration is above a mean-based criterion using theLCL, or below a mean-based criterion using the UCL.
- Study Question 1: What are the background concentrations ?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- The data must belong to a normal distribution.

- Data are stationary; there are no trends in the data or data characteristics over time.
- The criterion to which data will be compared is based on the mean.

Requirements and Tips

- Check the data for normality and skewness before using. You can test for normality using a probability plot, correlation coefficient, or Shapiro-Wilk test.
- Use of a minimum of eight values is recommended, a larger data set may be required if data are skewed or contain nondetects
- If a temporal component to the data exists, check that there is no temporal correlation by using the autocorrelation function or the rank von Neumann ratio test.
- If you suspect a temporal trend, test for trends using a time-series plot, Mann-Kendall test, or linear regression.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, or Rosner's test to further evaluate the suspected outliers.
- See Section 5.7 for information regarding treatment of nondetects.
- Select a level of confidence, such as 95%. This level of confidence may be determined by federal or state regulatory requirements or guidance, or by project-specific needs.
- Determine whether a one-sided or two-sided limit is necessary.

Strengths and Weaknesses

- The confidence interval decreases with larger sample sizes, thus helping to distinguish the statistic of interest from a criterion.
- The converse is that for small sample sizes, the confidence interval may be so wide as to not allow for identification of a statistical difference.

Further Information

A description of how to construct and use confidence intervals is found in Chapter 8.3 and Chapter 21, Unified Guidance. A description of how to construct a confidence interval around a normal mean is given in Chapter 21.1.1, Unified Guidance.

5.2.3 Confidence Interval Around Lognormal Geometric Mean

Typical environmental data are not normally distributed but instead are heavily right-skewed. One way to handle these data is to transform them logarithmically. The transformed lognormal data may fit a normal distribution. The log-transformed data are no longer in the arithmetic domain, but the logarithmic domain.

Sometimes, it may seem easiest to simply log-transform the data, calculate the arithmetic mean of the log-transformed data, construct a confidence interval around this value, and then back-transform the confidence levels back to obtain the correct confidence interval. Unfortunately, this approach results in a confidence interval around the geometric mean, not the arithmetic mean, which usually results in an underestimate of the true mean. Be aware that a confidence interval calculated in this way may not meet regulations applicable to the site.

Applications and Relevant Study Questions

- This method is used when comparing lognormally-distributed concentrations to a criterion that is based on a mean, as is common in risk assessment.
- Confidence limits may be used to evaluate whether a mean concentration is above a mean-based criterion using theLCL, or below a mean-based criterion using the UCL.
- Study Question 1: What are the background concentrations ?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- After calculating the logs of the data, the resulting log-transformed data must belong to a normal distribution
- Data are stationary; there are no trends in the data or data characteristics over time.
- The criteria are based on the mean.

Requirements and tips

- Check the lognormal transformations of the data for normality and skewness before using them. You can test for normality using a probability plot, correlation coefficient, or Shapiro-Wilk test. If the lognormal data still do not fit a normal distribution, use a nonparametric confidence interval.
- Use of a minimum of eight values is recommended, a larger data set may be required if data are skewed or contain nondetects.
- If a temporal component to the data is present, check that no temporal correlation exists by using the autocorrelation function (ACF) or the rank von Neumann ratio test.
- If a temporal trend is suspected, test for trends using a time-series plot, Mann-Kendall test, or linear regression.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, or Rosner's test.
- See Section 5.7 for information regarding the treatment of nondetects.
- Select a level of confidence, such as 95%. This level of confidence may be determined by federal or state regulatory requirements, or guidance, or project-specific needs.
- Determine whether a one-sided or two-sided limit is necessary.

Strengths and Weaknesses

- The confidence interval decreases with larger sample sizes, thus helping to distinguish the statistic of interest from a criterion.
- The converse is that for small sample sizes, the confidence interval may be so wide as to not allow for identification of a statistical difference.
- This method may result in an underestimate of the mean.

Further Information

A description of how to construct and use confidence intervals is found in Chapter 8.3 and Chapter 21,Unified Guidance. A description of how to construct a confidence interval around a lognormal

geometric mean is given in Chapter 21.1.2.

5.2.4 Confidence Intervals Around Lognormal Arithmetic Mean

Confidence intervals about the arithmetic mean, the statistic commonly required by regulations, are useful for skewed, lognormal data. This method is appropriate when you need to compare your data to an arithmetic mean and the data fit a normal distribution when log-transformed. Be aware that the available procedures for constructing this type of confidence interval can produce unacceptable results. Land's procedure is commonly used, but if the lognormal data have a high coefficient of variation, consider a bootstrap confidence interval around the arithmetic mean.

Applications and Relevant Study Questions

- This method is used when comparing lognormally-distributed concentrations to a criterion that is based on a mean, as is common in risk assessment.
- Confidence limits may be used to evaluate whether a mean concentration is above a mean-based criterion using theLCL, or below a mean-based criterion using the UCL.
- Study Question 1: What are the background concentrations ?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- An important underlying assumption is that the after calculating the logs of the data, the resulting log-transformed data belong to a normal distribution
- Data are stationary; no trends exist in the data or data characteristics over time.
- The criteria are based on the mean.

Requirements and tips

- Check the lognormal transformations of the data for normality and skewness before using. You can test for normality using a probability plot, correlation coefficient, or Shapiro-Wilk test. If the lognormal data still do not fit a normal distribution, use a nonparametric confidence interval.
- It is recommended that a minimum of eight values be used, a larger data set may be required if data are skewed or contain nondetects. In addition, data that are poorly fit by a lognormal curve may produce upper confidence bounds that are unrealistic or inappropriate for comparison.
- If you suspect a temporal trend, test for trends using a time-series plot, Mann-Kendall test, or linear regression.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, or Rosner's test.
- See Section 5.7 for information regarding the treatment of nondetects.
- Select a level of confidence, such as 95%. This level of confidence may be determined by federal or state regulatory requirements, or guidance, or project-specific needs.
- Determine whether a one-sided or two-sided limit is necessary.

Strengths and Weaknesses

- The confidence interval decreases with larger sample sizes, thus helping to distinguish the statistic of interest from a criterion.
- The converse is that for small sample sizes, the confidence interval for a lognormal arithmetic mean can be remarkably wide and require larger sample sizes than the confidence interval for a lognormal geometric mean to allow for identification of a statistical difference.
- This method may yield unacceptable results.

Further Information

A description of how to construct and use confidence intervals is found in Chapter 8.3 and Chapter 21, Unified Guidance. A description of how to construct a confidence interval around a lognormal arithmetic mean is given in Chapter 21.1.3, Unified Guidance.

5.2.5 Confidence Interval Around Upper Percentile

Sometimes you must construct confidence intervals around a percentile. For example, if the criterion is a concentration that represents the 90th percentile, then a confidence interval around the upper 90th percentile should be calculated. If the standard is a fixed criterion, such as a "not to exceed" maximum, then it is appropriate to use a confidence interval around a high percentile, such as the upper 95th or 99th percentiles. Be cautious when selecting a percentile as it may be extremely difficult to demonstrate corrective action success if too high a percentile is selected.

Applications and Relevant Study Questions

- This method is used when comparing concentrations to a fixed criterion that is based on a percentile or maximum.
- An alternate background threshold value may be calculated based on the upper confidence level around an upper percentile.
- Confidence limits may be used to evaluate whether a mean concentration is above a mean-based criterion using theLCL, or below a mean-based criterion using the UCL.
- Study Question 1: What are the background concentrations ?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- An important underlying assumption is that the data belong to a normal distribution or can be normalized.
- Data are stationary; there are no trends in the data or data characteristics over time.
- The criteria are based on an upper percentile or fixed value, not a mean.

Requirements and tips

- Check the data for normality and skewness before using. You can test for normality using a probability plot, correlation coefficient, or Shapiro-Wilk test. If the data are not normal, check if the data can be normalized by a log or other transformation.

- Use of a minimum of eight values is recommended, a larger data set may be required if data are skewed or contain nondetects.
- If a temporal component to the data is present, check that no temporal correlation exists by using the sample autocorrelation function or the rank von Neumann ratio test.
- If you suspect a temporal trend, test for trends using a time-series plot, Mann-Kendall test, or linear regression.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, or Rosner's test.
- See Section 5.7 for information regarding the treatment of nondetects.
- Need to select a level of confidence, for example, 95%. This level of confidence may be determined by federal or state regulatory requirements or guidance, or project-specific needs.
- Users should take care to note whether a one-sided of two-sided limit is necessary.

Strengths and Weaknesses

- Confidence intervals around a percentile do not suffer inaccuracies due to back transformation of log data.
- The confidence interval decreases with larger sample sizes, thus helping to distinguish the statistic of interest from a criterion.
- The converse is that for small sample sizes, the confidence interval may be so wide as to not allow for identification of a statistical difference.
- When testing that concentrations do not exceed a maximum value, a very high confidence level can make it difficult to demonstrate corrective action success.

Further Information

A description of how to construct and use confidence intervals is found in Chapter 8.3 and Chapter 21, Unified Guidance. A description of how to construct a confidence interval around an upper percentile is given in Chapter 21.1.4, Unified Guidance.

5.2.6 Nonparametric Confidence Interval Around a Median or Percentile

If your data do not fit a normal, lognormal, or other distribution, or if there are too many nondetects, use of a nonparametric confidence interval is appropriate. Nonparametric methods do not assume a particular distribution. Unfortunately, this generally results in wider confidence intervals and the need for larger data sets for making confident decisions. This method is appropriate when comparing concentrations to a percentile, such as the median (50th percentile) or 90th percentile. If you need to compare concentrations to a maximum criterion, a large percentile, such as the 95th or 99th percentile may be applied.

Applications and Relevant Study Questions

- This method is used when comparing concentrations to a fixed criterion that is based on the percentile or median.

- Confidence limits may be used to evaluate whether a mean concentration is above a fixed criterion using theLCL, or below a fixed criterion using the UCL.
- Study Question 1: What are the background concentrations?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- The criteria are fixed.
- The sample size is sufficient to achieve the desired confidence level.
- Data are stationary; there are no trends in the data or data characteristics over time.

Requirements and tips

- The confidence level depends on the sample size. It may be necessary to increase the sample size in order to make decisions at a desired confidence level.
- Use of a minimum of eight values is recommended, a larger data set may be required if data are skewed or contain nondetects. In addition, it is likely that more than eight values will be needed for the required confidence level.
- Select a level of confidence, such as 95%. This level of confidence may be determined by federal or state regulatory requirements, or guidance, or project-specific needs. This confidence level may not be attainable if the sample size is too small.
- See Section 5.7 for information regarding the treatment of nondetects.
- Determine whether a one-sided of two-sided limit is necessary.

Strengths and Weaknesses

- No particular distribution is needed; this method will work on most data sets.
- The confidence interval decreases with larger sample sizes, thus helping to distinguish the statistic of interest from a criterion.
- The converse is that for small sample sizes, the confidence interval may be so wide as to not allow for identification of a statistical difference. For a small data set, the confidence level may be so low as to provide little value for making decisions.
- You may need a large data set to achieve the desired confidence level.

Further Information

A description of how to construct and use confidence intervals is found in Chapter 8.3 and Chapter 21, Unified Guidance. A description of nonparametric confidence intervals is given in Chapter 21.2, Unified Guidance.

5.2.7 Confidence Interval Band Around Linear Regression Lines

If a linear trend is present in your data, you can describe the uncertainty in these data by constructing a confidence band around the trend line over the range of the data set. The confidence band is constructed of the individual confidence intervals around the mean as a function of time,

not an upper percentile. This method is most appropriate for cases where the fixed criterion represents a mean concentration and not an explicit upper percentile or "not to exceed" value.

Applications and Relevant Study Questions

- This method estimates the confidence intervals around a trend line.
- Confidence limits may be used to characterize the uncertainty in the slope when estimating attenuation rates.
- Study Question 1: What are the background concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?

Assumptions

- The residuals from the regression are approximately normal or reasonably symmetric.
- The variation about the mean should not be increasing or decreasing (that is, it should be stationary).
- Enough data exist to not only estimate the trend, but also to compute the variance around the trend line.
- Few if any nondetects are present.
- A linear trend exists.

Requirements and tips

- After fitting the regression line, test that the residuals from the regression are approximately normal or reasonably symmetric using a probability plot, correlation coefficient, or Shapiro-Wilk test.
- Plot the residuals versus concentrations. Check that the resulting scatter cloud is essentially uniform in vertical thickness or width, that is there is no tendency of the cloud to increase in width with concentration, or that the scatter cloud exhibits any kind of regular pattern.
- Use of a minimum of 8 to 10 data points, with few if any nondetects, is recommended; a larger data set may be required if the data are skewed or contain nondetects.

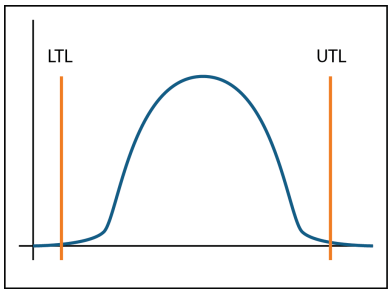
Strengths and Weaknesses

- The data are not required to be normal or lognormal, however the residuals are assumed to be normal.
- A large number of confidence intervals comprising the confidence band will not result in an increase in the false positive rate.
- If the variability changes along the trend, a wider confidence interval results.
- This method provides a graphical assessment of the uncertainty around a trend line.

Further Information

A description of how to construct and use confidence intervals is found in Chapter 8.3 and Chapter 21, Unified Guidance. A description of parametric confidence band around linear regression is given in Chapter 21.3.1, Unified Guidance.

5.3 Tolerance Limits



Tolerance intervals are statistical ranges typically constructed from on-site background data. Tolerance limits define the range of data that fall within a specified percentage with a specified level of confidence. The upper tolerance limit has been commonly used to establish a background threshold value, however, prediction limits are often favored for establishing a background threshold value in groundwater because they account for repeated measures. An upper tolerance limit (UTL) is designed to contain, but not exceed, a large fraction (that is, 95%, 99%) of the possible background concentrations, thus providing a reasonable upper limit on what is likely to be observed in background. Similarly, the lower tolerance limit (LTL) is designed to contain at most a certain percentage of the possible background concentrations, thus providing a reasonable lower limit on what is likely to be observed in background. The fraction to be contained or 'covered' by the limit is the coverage parameter, and must be specified along with a desired confidence level. Tolerance limits explicitly account for the degree of variation in the background population and the size of the sample of measurements used to construct the limit. Table F-2 includes information about checking assumptions for Tolerance limits. Tolerance limits and confidence limits (see Section 5.2) are distinct, even though in some cases the one-sided upper limits for both methods are equivalent.

Applications and Relevant Study Questions

- Tolerance limits can be used to represent the typical upper end of background concentrations (background upper tolerance limit).
- In compliance monitoring and corrective action (where allowed), a tolerance limit may serve as an alternate compliance limit (ACL) when no published compliance limit (for example, a maximum contaminant level (MCL) exists).
- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- Parametric tolerance limits assume the data follow a statistical distribution – typically normal (or can be normalized). If a transformation (for example, computing the logarithms of all of the data points is a transformation) is needed to normalize the measurements, the tolerance limit can be computed using the transformed values and then back-transforming the results to get the final limit.
- Nonparametric tolerance limits do not assume normality or any particular distributional form (but generally require larger samples sizes than parametric tolerance limits).
- Tolerance limits assume the population is stable (or stationary) over the period of time during which measurements are collected. No obvious trends or temporal patterns should exist in the background data.

- Since tolerance limits typically involve interwell comparisons, their use in detection monitoring tests assumes minimal spatial variability.
- Tolerance limits assume that the measurements are independent.
- Comparison of compliance data against an upper tolerance limit assumes that the two populations being compared have similar variances. This condition can be assessed using a homogeneity of variance test, but will be difficult to test directly unless you have at least four independent observations from each population (background and compliance).

Requirements and Tips

- Use of a minimum of 8 to 10 values is recommended, a larger data set may be required if data are skewed or contain nondetects.
- When using a parametric tolerance limit, test the normality of the original measurements or find a transformation that normalizes the data.
- When using a parametric tolerance limit, compute the sample mean and standard deviation.
- When using a nonparametric tolerance limit, rank the values.
- Check for temporal correlation and the presence of temporal trends in the background data.
- If you suspect outliers, examine the background data using a probability plot, Dixon's Test, Rosner's test, or another appropriate method.
- Check for spatial variation before using tolerance limits for interwell comparisons used in detection monitoring.
- When constructing a parametric tolerance limit, you must pre-specify both a confidence level and coverage level. A coverage level is usually pre-specified for a nonparametric tolerance limit, but the achieved confidence level is computed after the fact based on the available sample size.
- High false positive rates can occur when a large number of comparisons are done. However, the selected confidence level can be increased to lower the false positive rate.
- Lack of normality is not always a problem, since a nonparametric tolerance limit can be computed if the background data are not normal and cannot be normalized. However, you may need a larger sample size to achieve the desired coverage and confidence level targets.
- For parametric tolerance limits, see Section 5.7 for information regarding nondetects.
- If the population is not stationary over time, but instead the measurement levels are actively changing, the data used to construct a tolerance limit will be too variable and will tend to bias the limit on the high side (that is, it will be too large).
- If significant natural spatial variation exists, distinct well locations may exhibit substantially different levels independently of the presence of contaminants. In such cases, an interwell tolerance limit comparison may not answer the question: are there statistically significant differences from background levels attributable to groundwater contamination?
- If the measurements are not independent but instead positively correlated over time, the nominal degrees of freedom will be too large, leading to parametric tolerance limits that are biased low and nonparametric limits that achieve less confidence (for fixed coverage) than the nominally stated level.

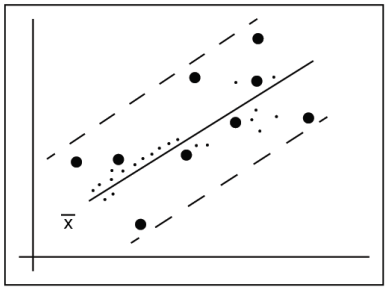
Strengths and Weaknesses

- A tolerance limit approximates an upper percentile of the background population and can be interpreted as such. For example, a 95% coverage upper tolerance limit approximates the population 95th percentile.
- Unlike prediction limits, tolerance limits can account for any number of future comparisons; however, a coverage level must be specified in place of the number of future comparisons.
- A tolerance limit on background can serve as an alternate criterion when background levels exceed published criteria.
- Tolerance limits are less flexible than prediction limits when incorporating formal retesting in detection monitoring.
- Nonparametric tolerance limits typically require a much larger sample size than parametric tolerance limits to achieve both a high coverage and high confidence level.
- Unlike prediction limits, a small percentage of true background values will tend to exceed a tolerance limit, at a rate equal to the complement of the coverage level (for example, 5% of the time for 95% coverage). As a result, greater uncertainty may exist in determining when an exceedance of a tolerance limit truly indicates a statistically significant difference from background levels.

Further Information

Chapter 17.2.1, Unified Guidance discusses parametric tolerance limits and provides a sample problem (Example 17-3). Chapter 17.2.2, Unified Guidance discusses nonparametric tolerance. See Example 17-4, Unified Guidance for application of nonparametric tolerance limits.

5.4 Prediction Limits



Prediction limits have several uses in groundwater monitoring, all of which involve predicting the upper limit of possible future values based on a background or baseline data set and comparing that limit to compliance point measurements or statistics. An upper prediction limit is constructed from upgradient or historical data and is designed to equal or exceed a specified number of future comparisons. If any of those values exceed the prediction limit, then the analysis suggests that groundwater concentrations have risen above the background levels. Prediction limits explicitly account for the degree of variation in the background population and the size of the sample of measurements used to construct the limit.

Prediction limits can be constructed with either a parametric or nonparametric statistical model. Parametric prediction limits are based on the mean and standard deviation of the background or baseline data set, whereas nonparametric prediction limits are based on ranking of the observations. Table F-2 includes information about checking assumptions for prediction limits.

Interwell prediction limits compare background and compliance data collected from distinct spatial locations (upgradient versus downgradient). Intrawell prediction limits compare historical data (labeled intrawell background) versus current data from a single location (see Section 3.6.5: Should I use interwell or intrawell sampling?).

Applications and Relevant Study Questions

Prediction limits are primarily used as formal detection monitoring tests of compliance data against background. Since they are very flexible statistical tools, prediction limits can be used successfully as interwell or intrawell tests and can be readily adapted to modern sampling schedules.

- For interwell testing, data from upgradient wells are used as background data to construct the upper prediction limit, to which the compliance data are compared.
- For intrawell testing, historical data from each compliance well are used as background data to construct the upper prediction limit, to which the current data are compared.
- It is possible to use prediction limits to compare means, medians, or other statistical measures of background and compliance data sets.
- Lower bound prediction limits are occasionally used to indicate that certain parameters (for example, pH, dissolved oxygen) are below a target interval
- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- Parametric prediction limits assume the data follow a known distribution or can be

transformed to a known distribution.
- Nonparametric prediction limits for testing future medians do not assume normality or any particular distributional form.
- All prediction limits assume the population is stable (or stationary) over the period of time during which measurements are collected. That is, no obvious trends or temporal patterns should exist in the background data.
- Prediction limits used for interwell testing assume minimal (nonsignificant) spatial variability. Intrawell prediction limits can be used when spatial variation is significant, but require the assumption that intrawell background is uncontaminated.
- Prediction limits for testing future means assumes a normal distribution or data transformed to a normal distribution. Perform all computations and comparisons on the transformed scale to avoid transformation bias.
- Comparison of compliance data against an upper prediction limit assumes that the two populations being compared have similar variances. This condition can be assessed using a homogeneity of variance test, but will be difficult to test directly unless you have at least four independent observations from each population (background and compliance).

Requirements and Tips

- A minimum of 8-10 values is recommended, a larger data set may be required if data are skewed or contain nondetects.
- The number of future comparisons (m) against the prediction limit must be pre-specified.
- Evaluate the distribution of the data (for example, test for normality) to determine whether a parametric or nonparametric model is appropriate. It may also be possible to transform data to fit a normal distribution.
- Check for temporal correlation and rule out the presence of temporal trends in the background data.
- If you suspect outliers, examine the background data using a probability plot, Dixon's test, Rosner's test, or another appropriate method. See Section 5.7 for information regarding the treatment of nondetects.
- Check for spatial variation before using as an interwell test in detection monitoring.
- A confidence level must be pre-specified when constructing a parametric prediction limit. For nonparametric prediction limits, the achieved confidence level is computed after the fact based on the available sample size.

Strengths and Weaknesses

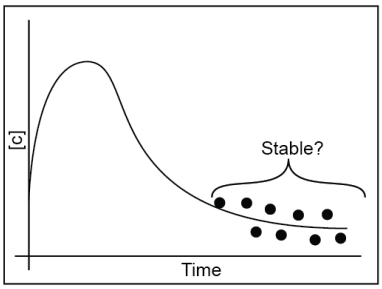
- A prediction limit estimates a firm 'cap' on the background population for a specified number of future sampling events (comparisons). It thus allows for clear interpretation of when background levels have been exceeded.
- Prediction limits have the advantage of incorporating a formal re-testing strategy into the calculation of the test statistic, making it possible to precisely control false positive rates and statistical power.

- Prediction limits for future means are a powerful but more flexible alternative to analysis of variance (ANOVA), more statistically powerful than prediction limits for future individual observations, and more adaptable to groundwater sampling than ANOVA.
- Nonparametric prediction limits typically require a much larger sample size than parametric prediction limits to achieve the desired confidence level. However, the formal retesting strategy can partially mitigate this requirement.
- Prediction limits are only 'valid' for a pre-specified number of future comparisons; when those comparisons have been exhausted, the prediction limit may need to be recomputed and a new test set up.

Further Information

Prediction limits are discussed in Chapter 18, Unified Guidance. Parametric prediction limits are discussed in Chapter 18.2, Unified Guidance. Nonparametric prediction limits are discussed in Chapter 18.3, Unified Guidance. See Example 18-1 and Example 18-2, Unified Guidance for applications of parametric prediction limits for future values and for a future mean, respectively. See also Example 18-3 and Example 18-4, Unified Guidance for the application of nonparametric prediction limits for future values and for a future mean, respectively.

5.5 Trend Tests



A trend refers to an association or correlation between concentration and time or spatial location, but can also refer to any population characteristic changing in some predictable manner with another variable. Trends take various forms, such as increasing, decreasing, or periodic (cyclic).

Detecting and assessing temporal and spatial trends is important for many environmental studies and monitoring programs. Trend tests are generally recommended as an intrawell alternative to prediction limits or control charts for use in detection monitoring. Trend evaluations are frequently used to determine whether it is reasonable to assume concentrations are temporally stationary (for example, to perform statistical evaluations that require stationary means) and to detect or model decreasing trends to support natural attenuation studies. Table F-1 includes information about checking assumptions for Trend Tests.

5.5.1 Linear Regression (Parametric Methods to Test and Model Trends)

Linear regression is used to test for linear temporal trends. Ordinary least squares regression is used to fit the "best" straight line. A linear trend is reported when the slope of the regression line is demonstrated to be statistically different from zero (using a t-test); a positive slope indicates an increasing trend and a negative slope a decreasing trend. The linear correlation coefficient Pearson's r (Equation 3.5 in Chapter 3.3, Unified Guidance), which is the correlation coefficient between the observed and calculated concentrations, provides information about the direction and "strength" of the linear trend. A positive value of r indicates an increasing linear trend and a

negative value a decreasing linear trend. The trend is "strong" if the absolute value of r (which ranges from – 1 to 1) is near one.

A regression line models concentrations for the period of time over which the concentrations were measured. However, this method is often used to predict concentrations at future times as well, under the assumption that the same linear relationship will be observed—an assumption that may not be valid for monitoring events in the distant future). For a decreasing trend, the regression line is often extrapolated to estimate the time at which a criterion will be met. However, even when it is assumed that the regression line is valid for future monitoring, this approach will not necessarily result in conservative estimates (underestimate the actual time required to achieve a criterion), because it does not take into account the uncertainty of the regression fit that arises from the variability of the data around the calculated regression line. A set of concentrations that fall nearly on the calculated regression line will result in estimates that are more reliable than concentrations that exhibit much larger scatter about the same regression line. A confidence interval is often calculated for the regression line (Chapter 5.2) to account for the uncertainty of the mean concentration as it varies linearly with time (for instance, to provide upper bound estimates of cleanup times or contaminant concentrations).

Many commercial statistical software packages offer multiple options for nonlinear regression fits. For example, many provide quadratic and cubic polynomials that can be used to model nonlinear trends. Many software packages also calculate confidence limits for nonlinear regression fits.

Applications and Relevant Study Questions:

- Use trend tests to determine if the mean of the population is stationary, which is a requirement for the use of many statistical tests.
- Study Question 4: When will contaminant concentrations reach a criterion?
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?

Assumptions

- Linear regression assumes the residuals (the differences between the measured and calculated concentrations) are independent and normally distributed with a constant variance (with respect to time and concentration).

Requirements and Tips

- Generate a time series plot initially to qualitative assess whether an apparent linear relationship exists.
- For ordinary least square regression fits, use scatter plots of the residuals (the differences between the measured and calculated concentrations) versus concentration and time to qualitatively evaluate whether the variance of the residuals is constant. For example, a "cloud" of points of relatively uniform width over the entire time or concentration range suggests the variance is constant.

- The residuals can be evaluated for normality using normal probability plots or statistical tests for normality.
- When regression residuals are not normally distributed, use mathematical transformations to normalize them. For example, taking logarithms of the concentrations and subsequently calculating a new regression line of the form $ln(y) = c + dt$, may normalize the residuals. This approach ultimately results in a nonlinear equation that models the trend. For example, when a log-transformation is done, the regression line is "back transformed" by exponentiation, resulting in a nonlinear equation of the form $y = c' + exp(dt)$.
- Use an autocorrelation test to verify that regression residuals are statistically independent.
- Linear regression is sensitive to outliers.

Strengths and Weaknesses

- Normality assumptions cannot be violated.
- Parametric methods are very sensitive to outliers.
- Nondetects cannot be readily addressed. The substitution of surrogate values for nondetects (for example, multiples of the reporting limit) can produce erroneous results.

Further Information

- Chapter 3.3, Unified Guidance, Common Statistical Measures, Sample correlation coefficient (Pearson's r)
- Chapter 17.3.1, Unified Guidance, Linear Regression
- Chapter 21.3.1, Unified Guidance, Parametric Confidence Band Around Linear Regression

5.5.2 Mann-Kendall Test (Nonparametric Method to Test and Model Trends)

The Mann-Kendall test is a nonparametric test for monotonic trends, such as concentrations that are either consistently increasing or decreasing over time. Therefore, the test is not appropriate when there are cyclic trends (where concentrations are alternatively increasing and then decreasing). The Mann-Kendall statistic provides an indication of whether a trend exists and whether the trend is positive or negative. Subsequent calculation of Kendall's Tau permits a comparison of the strength of correlation between two data series.

The Mann-Kendall test can be used to evaluate the following:

- Are contaminant concentrations increasing or decreasing in upgradient or downgradient wells?
- Does contaminant flux, as measured across a plume cross section, indicate an increasing or decreasing trend?
- Are concentrations within a well stable?

The Mann Kendall statistic (S) is calculated through pair-wise comparisons of each data point with all preceding data points, and determining the number of increases, decreases, and ties. Pairs of

nondetects below the reporting limit are "ties" that do not increase or decrease the value of S. A positive value for S implies an upward or increasing temporal trend, whereas a negative value implies a downward or decreasing trend. A value of S near zero suggests there is no significant upward or downward trend. The magnitude of S measures the "strength" of the trend. A statistically significant trend is reported if the absolute value of S is greater than the "critical value" of S (obtained from a table).

The nonparametric correlation coefficient Kendall's tau (τ) can be calculated to evaluate the nonparametric correlation between two data series. It is essentially a scaled measure of S; $\tau = S/[n(n-1)/2]$, where n denotes the number of concentration measurements. Therefore, a statistical trend is equivalently demonstrated when τ is significantly different from zero. However, it is more convenient to evaluate trends using Kendall's tau, because like the parametric linear correlation coefficient r, τ ranges from -1 to 1. A trend is "strong" if the absolute value of τ is near one.

Applications and Relevant Study Questions

- Trend tests may be used to determine if the mean of the population is stationary, which is a requirement for the use of many statistical tests.
- Study Question 4: When will contaminant concentrations reach a criterion?
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 7: What are contaminant attenuation rates in wells?

Assumptions

This test assumes independent concentration measurements.

Requirements and Tips

- Trend tests should be accompanied by time-series plots.
- The influence of nondetects should be evaluated. See Section 5.7 for more information regarding nondetect data.
- A minimum of 8 to 10 measurements is recommended; a larger data set may be required if data are skewed or contain nondetects.

Strengths and Weaknesses

- This test can be used when data sets contain nondetects.
- Results are not impacted by the magnitude of extreme values as with regression/correlation tests.
- This test is difficult to apply to data sets containing mixed detection limits and estimated values between the reporting limit and the detection limit.

Further Information

- Example A.2, Testing a Data Set for Trends over Time
- Chapter 17.3.2, Mann-Kendall Trend Test, Unified Guidance.

5.5.3 Theil-Sen Trend Lines (Nonparametric Method to Test and Model Trends)

When a monotonic trend is demonstrated using the Mann-Kendall test and the trend appears to be linear, you can use a Theil-Sen line to estimate the slope of the trend. The Theil-Sen line is a non-parametric alternative to the parametric ordinary least squares regression line. An ordinary least squares regression line models how the mean concentration changes linearly with time; a Theil-Sen line models how the median (50th percentile) concentration changes linearly with time. Therefore, the approach may not be appropriate when more than 50% of the concentration measurements are nondetects. The slope of the Theil-Sen line will be significantly different from zero when Kendall's tau is significantly different from zero (and vice versa). Like the parametric linear regression line, confidence intervals can be calculated for the nonparametric Theil-Sen line.

Bootstrapping to Obtain Theil-Sen Confidence Intervals

Chapter 21.3.2, Unified Guidance describes how to calculate confidence intervals for a Theil-Sen line using a computationally intensive procedure referred to as "bootstrapping." The Theil-Sen is initially calculated from a set of n sequential concentration measurements (y) over time (t): (t_1, y_1), (t_2, y_2) … (t_n, y_n). The bootstrapping method is subsequently used to calculate confidence limits. This method entails randomly selecting, with replacement, a set of n of these pairs. For example, if there are only two pairs (t_1, y_1) and (t_2, y_2), each selection or "iteration" results in one of four possible outcomes:

(t_1, y_1) and (t_1, y_1)

(t_1, y_1) and (t_2, y_2)

(t_2, y_2) and (t_1, y_1)

(t_2, y_2) and (t_2, y_2)

This selection is repeated a larger number of times, B (such as B = 1,000). A Theil-Sen line is calculated for each of the B iterations (sets of n pairs). At each time t_i (i = 1,…n), a concentration can be calculated using each of the B Theil-Sen lines, which results in a set of B calculated concentrations for each time.

Nonparametric confidence limits are obtained by calculating upper and lower percentiles of the B concentrations for each time. For example, if B = 1000, the 95% confidence interval of for concentration calculated from the original Theil-Sen line is obtained from the 25th largest concentration (2.5 percentile) and 975th largest concentration (97.5th percentile) of the B bootstrap concentrations.

Applications and Relevant Study Questions

- Trend tests may be used to determine if the mean of the population is stationary, which is a requirement for the use of many statistical tests.
- Study Question 4: When will contaminant concentrations reach a criterion?

- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?

Assumptions

This test assumes independent concentration measurements.

Requirements and Tips

- The influence of nondetects should be evaluated. See Section 5.7 for information regarding nondetect data.
- Ensure that statistical software used for the Mann-Kendall test treats nondetects as inequalities. It is recommended that you do not use software for which it is necessary to assign surrogate values to nondetects, since this can produce unreliable results.
- A minimum of 8 to 10 measurements is recommended, a larger data set may be required if data are skewed or contain nondetects.
- Consider verifying that trend residuals are statistically independent (for example, using an autocorrelation test).

Strengths and Weaknesses

- This test can be used when data sets contain nondetects, but may not provide useful information if a large portion of the data set is nondetect.
- Results are not impacted by the magnitude of extreme values as with regression or correlation tests.
- This test is difficult to apply to data sets containing mixed detection limits and estimated values between the reporting limit and the detection limit.

Further Information

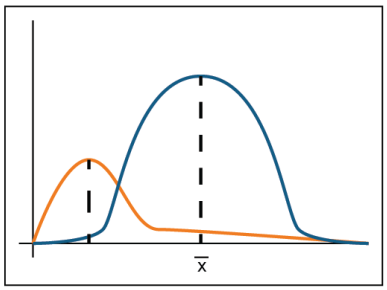
Chapter 21.3.2, Unified Guidance, Nonparametric Confidence Band Around Theil-Sen Line

5.5.4 Spearman's Rank Correlation Test

Spearman's rank correlation coefficient rho (ρ) is a nonparametric correlation coefficient that can be used to test for monotonic trends. The Spearman rank correlation test is discussed further in Section 5.12.2 of this document.

5.6 Distributional Tests

Distributional tests are commonly used to evaluate data distribution and to test data for normality. Many commonly applied statistical tests are parametric (i.e., they assume that the data follow a specific distribution, that they have a certain shape, and that the data can be described by a few parameters, such as the mean (a measure of centrality) and standard deviation (a measure of spread).



Of the many different types of distributions used in statistics, the most commonly used are the normal distribution, (also known as the bell curve) and distributions that can be transformed to a normal distribution (such as a lognormal distribution). In addition, the gamma and Weibull distributions are used. The normal distribution (bell curve) is well known because of its common use in scholastic grading. This curve plots the frequency of occurrence on the vertical axis and the ordered values of interest, in our case, concentration, on the horizontal axis. If the data follow a normal distribution, most of the data concentrations are near the mean, or average, value and the likelihood of obtaining values away from the mean in either direction tapers off the further the concentration is from the mean.

Appendix A includes several case examples that provide examples of evaluating groundwater data with distributions.

Normal data distribution, bell curves, and histograms

The mathematical model of the normal distribution produces a perfectly smooth, symmetrical, bell-shaped curve. The mean and standard deviation of the data determine the shape of the bell. The mean locates the bell peak on the horizontal axis, and the standard deviation determines the width of the bell. A large standard deviation means that the bell will be broad and flat. A small standard deviation means that the bell will be narrow and skinny (the concentrations in the data set do not deviate much from the mean).

A histogram presents a rough depiction of the data distribution that can be matched with the mathematical model of the normal curve. The histogram, which orders the values, counts the number (frequency) of values within a fixed range of values (a bin) and plots the frequency of values within each bin on the y-axis at the bin's central value on the x-axis.

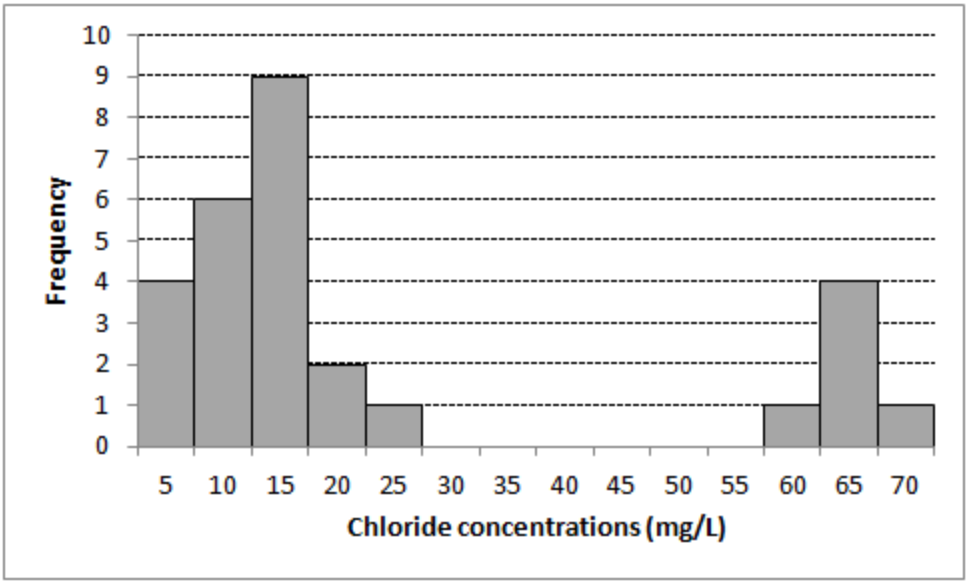


Figure 5-13. Histogram example.

Are the data normal?

The first task when using a parametric test is to test the underlying assumption of normality. If the

data do not produce a nicely shaped bell, for example, if the bell is lopsided or has several peaks, then the underlying mathematical model for the test will not match the data and may produce erroneous results. Other complications might cause the data to appear non-normal, such as outliers, the presence of nondetects, or changes over space or time (nonstationarity). Testing for normality should be conducted in conjunction with tests for outliers and nonstationarity.

Nondetects are left-censored data, meaning that, below a certain reporting limit the concentrations are not known. Most tests for normality depend on the values at the ends or tails of the ordered data. Too many nondetects in the data set (the Unified Guidance recommends having no more than 10-15% nondetects), can cause problems with the normality tests because the concentrations at the lower tail of the sample distribution are unknown, yet a value is needed for standard normality test to be run. Use caution in substituting values for nondetects, even at low percentages of nondetects. Apply nonparametric methods if there is doubt regarding the usability of the data due to the presence of nondetects. See Section 5.7: Managing Nondetects in Statistical Analyses for more information on nondetects.

Outliers are anomalous data found at the tails of data distributions, so their presence may cause problems in testing for normality. If outliers are suspected and a test for normality fails, try removing the suspected outliers and rerunning the test. See Section 5.10: Identification of Outliers for more information on outliers.

Nonstationarity can be an issue with data collected over space or time. The change of concentrations over time or the inconsistency of data over a large area may introduce data that are not in the same distribution. Distribution tests might fail when grouping data sets together even if the original data sets are independently normally distributed. Trend tests or analysis of variance (ANOVA) tests should be used if non-stationarity is suspected. See Section 3.4.6, Section 5.5, and Section 5.8 for more information on evaluating stationarity.

Many specific methods can test for normality of data distributions, including the goodness-of-fit tests, which compare a chosen distribution with the data set of interest. The following are commonly applied methods:

- Coefficient of Skewness and Variation
- Kolmogorov-Smirnov test
- Graphical assessment of normality (probability plot), probability plots
- Shapiro-Wilk test
- Shapiro-Francia normality test

5.6.1 Coefficients of Skewness and Variation

Because a normal, bell-shaped distribution is symmetric about the mean, normally distributed data will have zero skewness. Therefore, measuring the degree of skewness aids in evaluating data for normality and in evaluating the degree of non-normality. A coefficient of skewness greater than one indicates that the data are not normally distributed. Also, because of the symmetry of the

normal curve, the median value will be equal to the mean value. The coefficient of variation (the standard deviation divided by the mean) will also provide some measure of departure from normality. A coefficient of variation greater than one similarly indicates that the data are not normally distributed.

Application and Relevant Study Questions

Calculation of the coefficients of skewness and variation can aid in evaluating a data set for normality.

Assumptions: None

Requirements and Tips

- These methods are not appropriate for data that have been changed by a log transformation.
- Use of a minimum of 8 to 10 values is recommended, a larger data set may be required if data are skewed or contain nondetects
- See Section 5.7 for information on handling nondetects.

Strengths and Weaknesses

- These methods are useful for a quick and easy evaluation of data that will reveal a possible non-normal distribution.
- These methods do not confirm normality, but can provide evidence against normality. Therefore, these methods should be used in conjunction with other tests.

Further Information

Chapter 10.4, Unified Guidance includes discussion of the coefficient of variation and coefficient of skewness.

5.6.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (K-S test) is a common nonparametric goodness-of-fit test that compares the measured data distribution function with the normal distribution function (the mathematical model that generates the normal distribution). Thus, the K-S test compares the graphical curve (in this case, a cumulative fraction plot) of the measured data with that of the normal cumulative fraction plot. The method then calculates maximum distance between the two curves and estimates the p-value. A p-value greater than the selected confidence level indicates that the data likely fit a normal distribution. A p-value below the selected confidence level indicates that the data do not fit a normal distribution.

Application and Relevant Study Questions

- Goodness of fit tests are used to test the assumption of normality prior to applying other statistical tests.
- Study Question 9: Is the sampling frequency appropriate (temporal optimization)?

Assumptions

The K-S test only applies for continuous distributions, but these distributions are usually expected in environmental systems.

Requirements and Tips:

- If the K-S test fails (p-value is less than the selected significance level), try transforming the data and re-testing for normality.
- Use of a minimum of 8 to 10 values is recommended, a larger data set may be required if data are skewed or contain nondetects.

Strengths and Weaknesses

- The K-S test is a robust test that only considers the relative distribution of the data, therefore log-transformation of the data do not negatively affect this test.
- The test is more sensitive around the center of the curve than near the tails.
- The K-S test is not as powerful as the Shapiro-Wilk test.

Further Information

Chapter 10, Unified Guidance provides information regarding fitting of distributions to data sets

5.6.3 Shapiro-Wilk Test

The Shapiro-Wilk test calculates an SW value. The SW value indicates whether a random sample comes from a normal distribution. If a data set is normally distributed, then a correlation should exist between the ordered data and the normal distribution. Large values of SW indicate a strong correlation while small values of SW are evidence of departure from normally distributed data. This test has performed well in comparison studies with other goodness-of-fit tests.

Application and Relevant Study Questions

- Used to test for normality.
- If the SW value exceeds the critical value, the data set is probably normally distributed.
- If the SW is less than the critical value, the data set is not normally distributed. In this case, you may use a data transformation and re-test the transformed data for normality.

Assumptions: None

Requirements and Tips

Use caution when applying this method to data sets with a large number of nondetects; a larger number of detects will give a better result. For best results, chose a coefficient (α) = 0.10 for very small data sets ($n < 10$), $\alpha = 0.05$ for moderately sized data sets ($10 \leq n < 20$), and $\alpha = 0.01$ for large data sets ($n \geq 20$). This approach is not useful for very large data sets ($n > 50$).

Strengths and Weaknesses

- Because it involves null hypothesis significance testing, if you reject null hypothesis you may conclude that the population is not normally distributed. Rejecting the null hypothesis means that population is not normally distributed, but it does not indicate whether the reason for non-normality is because of a flat-tailed distribution, a skewed distribution, or something else.
- If the null hypothesis is not rejected, you may only conclude that the test failed to show that the population is not normally distributed. In other words, the test can substantiate that the population is not normally distributed, but it cannot prove that the data set is normally distributed.
- The tests are influenced by power. If you have a small sample (n is the number of values), then the test may not have enough power to detect normality in the population. If you have a very large sample, then the test will detect even a trivial deviation from normality.

Further Information

Chapter 10.5.1, Unified Guidance includes further information and an example for the Shapiro Wilk test.

5.6.4 Shapiro-Francia Normality Test

The Shapiro-Francia test is a simplified version of the Shapiro-Wilk test. The test is generally considered equivalent to Shapiro-Wilk test for large, independent samples. Like the Shapiro-Wilk test, the Shapiro-Francia test calculates an SF statistic to indicate whether a random sample comes from a normal distribution. If a data set is normally distributed, a correlation should exist between the ordered data and the z-scores taken from the normal distribution. Large values of SF indicate a strong correlation while small values of SF are evidence of departure from normally distributed data. The Shapiro-Francia test calculates an "SF" statistic. If the SF statistic exceeds the critical value, the test indicates that data likely fit a normal distribution. If the SF is less than the critical value, the test indicates that the data are not normally distributed. You may subsequently apply a data transformation, and retest for normality.

Applications and Relative Study Questions

The Shapiro-Francia method is used to test for normality.

Assumptions: None

Requirements and Tips

Use caution when applying this method to data sets with a large number of nondetects; a larger number of detected values will give a better result.

Strengths and Weaknesses

- Because it involves null hypothesis significance testing, if you reject null hypothesis you may conclude that the population is not normally distributed. Rejecting the null hypothesis

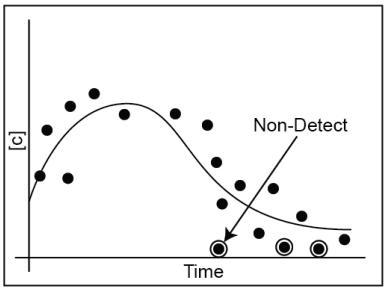
means that population is not normally distributed, but it does not indicate whether the reason for non-normality is because of a flat-tailed distribution, a skewed distribution, or something else.

- If the null hypothesis is not rejected, you may only conclude that the test failed to show that the population is not normally distributed. In other words, the test can substantiate that the population is not normally distributed, but it cannot prove that the data set is normally distributed.
- The tests are influenced by power. If you have a small sample (n is the number of values), then the test may not have enough power to detect normality in the population. If you have a very large sample, then the test will detect even a trivial deviation from normality.

Further Information

Chapter 10.5.2, Unified Guidance includes information about the Shapiro-Francia test.

5.7 Managing Nondetects in Statistical Analyses



Environmental statistics is constrained by a practical reality of laboratory analysis: it is technically impossible for a laboratory analysis to confirm the complete absence of a chemical or compound of interest. Instead, a chemical may be present at some unknown concentration below the low end of the concentration range that the analysis is able to detect. Since the true level is unknown, laboratories report the nonzero value representing the lowest concentration that can be reliably detected for the given analytical method. This alternate value is often used in environmental statistical applications, even though the true value can only be narrowed to a range of possible concentrations (for example, from zero up to the reporting limit).

5.7.1 Definition of Detection Limits

In environmental testing, a detection limit is the concentration that is statistically greater than the concentration of a method blank with a high level of confidence (typically, 99%), or the lowest level of a given chemical that can be positively identified when using a particular analytical method. Signal intensity below the detection limit cannot be reliably distinguished from a method blank or "baseline noise." Therefore, an analyte is confidently reported as present in an environmental sample only when the measured concentration is greater than the detection limit.

Statisticians refer to any threshold at which a nondetect is reported as a "censoring limit." Nondetects are sometimes referred to as censored values. Censoring limits affect how one should manage data. For instance, reporting nondetects to larger censoring limits (higher detection limits) than needed tends to adversely impact data quality and increase data uncertainty. Unfortunately, different environmental testing laboratories use different types of censoring limits and reporting conventions for nondetects. No standard industry practice exists for establishing censoring limits.

Often the method detection limit (MDL) described in 40 CFR Part 136, Appendix B is used as one of the censoring limits. The MDL is designed to minimize false positives (that is, reporting a compound as present when it is really not). Another common censoring limit is the reporting limit which typically refers to the smallest concentration at which analytical results will likely achieve specified or acceptable tolerances for precision and bias. The reporting limit is generally larger than the MDL and is also referred to as a "quantitation limit" or "limit of quantitation." Detected results less than the reporting limit (but larger than the MDL) may be reported with "J" flags (or "qualifiers") to denote their lack of quantitative reliability.

In general, a measured concentration (detect) greater than the MDL but less than the reporting limit only reliably demonstrates the chemical is present in the sample at some concentration significantly greater than that of a method blank. Nevertheless, it is generally preferable to utilize qualified

detections at their measured values in statistical evaluations — despite their greater analytical uncertainty — rather than treating them as censored values reported to a reporting limit (that is, less than values). Too much statistical information is lost by converting such detections to higher censored values.

Reporting conventions differ from laboratory to laboratory, and often nondetects are reported to the reporting limit (usually noted as "< RL") instead of the lowest detectable measurement. Therefore, prior to processing data containing nondetects, consult the laboratory (or an environmental chemist) to evaluate the thresholds to which nondetects are reported.

Good statistical evaluations attempt to minimize data censoring both in terms of the proportions of censored values (nondetects) and the magnitudes of the censoring limits. This practice frequently avoids potential problems and simplifies statistical calculations. No statistical technique can fully compensate for the information loss due to data censoring. The larger the proportion of censored data and the larger the censoring limits, the greater the information loss and uncertainty. Some analytical methods (such as metals analyzed by inductively coupled plasma spectroscopy) — at times referred to as 'uncensored methods' — are capable of reporting numerical values for method blanks. With these methods, negative values may sometimes be obtained for method blanks, even though negative concentrations are not physically meaningful as individual values. However, this practice is acceptable and expected when random measurement variability is present and the 'true' mean is equal (or nearly equal) to zero (a similar pattern may be observed in measurements of radionuclides). Remember that statistical decisions and results are based on the aggregate information contained in a data set, and not on any single estimated value. Because detection and reporting limits often change over time with improved analytical methods, or because differing levels of turbidity or interference may necessitate sample-specific reporting limits, many data sets contain multiple reporting limits. Observations in the same data set that are censored at differing levels present additional statistical complexity. However, special methods for handling such data have been developed (see Section 5.7.6, Section 5.7.7, and Section 5.7.8).

5.7.2 Managing Nondetects

Despite considerable research in recent years on handling nondetects, regulatory agencies have published no comprehensive guidance on the recommended approach to use in a particular situation. As a result, approaches to handling nondetects in groundwater projects vary widely.

The following are the general strategies for handling nondetects:

1. Use statistical approaches specifically designed to accommodate nondetects, such as the Tarone-Ware two-sample alternative to the t-test.
2. Use a rank-based, nonparametric test, such as the Mann-Kendall trend test.
3. Use a censored estimation technique to estimate sample statistics, such as the Kaplan-Meier method for calculating an upper confidence limit on the mean.
4. Impute an estimated value for each nondetect prior to further statistical analysis.

The most commonly used methods are described in the sections below.

5.7.3 Use of Nonparametric Methods

Nonparametric methods that treat nondetects as inequalities are probably the most versatile and effective approach for handling censored data sets. Some of these methods, such as the Tarone-Ware test, are specifically designed to accommodate censored data. Others entail ordering the measurements (from smallest to largest) and replacing the values with their corresponding ranks. Nondetects are treated as 'ties' and are assigned the same rank (without substituting any imputed or surrogate values). The advantage of nonparametric methods over parametric methods is that a specific (parametric) distribution is not assumed. The Wilcoxon rank-sum (Chapter 16.2, Unified Guidance) and Kruskal-Wallis tests (Chapter 17.2.2, Unified Guidance)—which test whether the medians of two or more environmental populations differ significantly—are examples of rank-based nonparametric tests that can be used for data sets which contain nondetects.

Tips regarding nonparametric methods include the following:

- A larger number of data points are generally required for nonparametric methods to achieve the same level of confidence and false positive rate control as parametric methods.
- Although nonparametric methods can tolerate a relatively large proportion of censored values, they generally lose significant statistical power if most of the data are censored.
- There is a distinction between nonparametric methods based on ranks (such as Wilcoxon rank-sum) and those based on counting values below a threshold (for instance, Tarone-Ware or test of proportions). Ranking methods assume that the data can be fully sorted and ranked (apart from ties). If a large proportion of the data is tied due to nondetects, an alternative strategy may be needed.
- Generalizations can be difficult to make because there are many types of nonparametric tests and their 'robustness' to data censoring depends on the nature of test. For example, it is problematic to compare medians (50th percentiles) when more than 50% of the results are censored, but comparisons of larger (e.g., 95th) percentiles may be possible.

A commonly encountered problem is how to estimate a linear trend when some of the data are censored. In some cases, a trend may be suspected from a time series plot (see Figure 5-14) when in reality most or all of the data are nondetect and reporting limits have decreased due to improved analytical techniques. In others, correct identification of a trend and its apparent slope may depend on appropriate handling of the censored values. Assigning each nondetect to, say, half the reporting limit will not properly account for the analytical uncertainty of the data set, especially if some of the reporting limits are elevated due to dilution factors or poorer historical precision.

Two alternatives include Turnbull's method and the Akritas-Theil-Sen technique (both discussed in Helsel 2012). Turnbull's method extends the Mann-Kendall trend test to censored data by computing a slope and significance test for the slope that is equivalent to testing whether Kendall's tau is different from zero. Akritas-Theil-Sen similarly extends the nonparametric Theil-Sen trend method to properly account for censored data by computing the slope that makes a Theil-Sen trend line estimate applied to the trend residuals (what is left over after subtracting out the trend) equal to

zero. Application of Turnbull's method to the example benzene data shown in Figure 5-14 gives a nonsignificant Kendall's tau value of 0.01, demonstrating that the apparent trend is illusory, a function of decreasing reporting limits but not a real change in concentration levels over time.

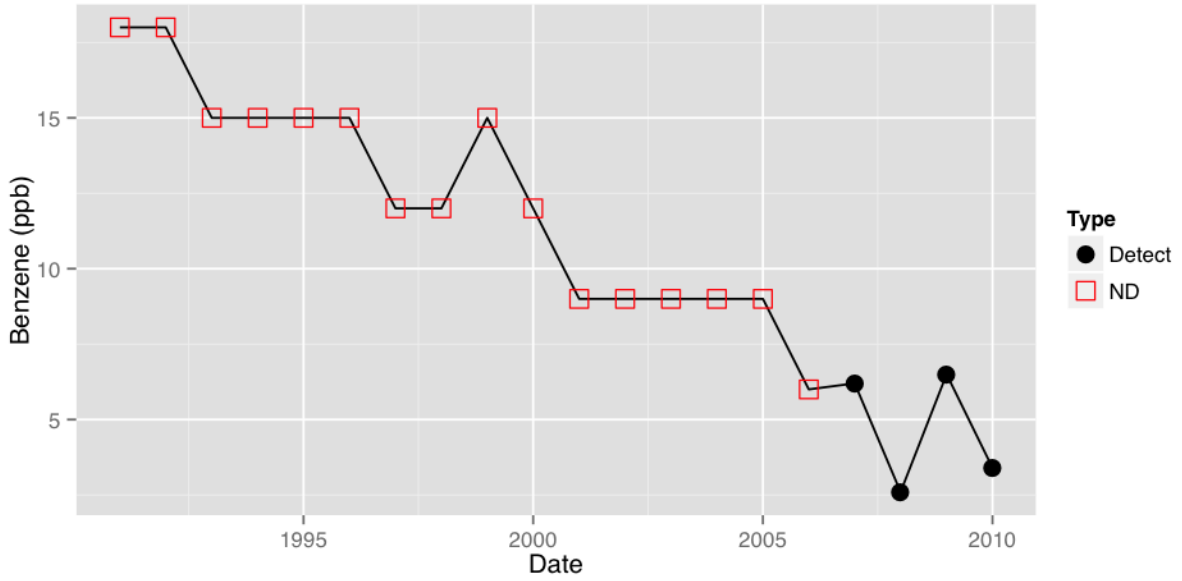


Figure 5-14. Example time series plot of benzene data with nondetects.

5.7.4 Omission of Nondetects

Omitting nondetects from a statistical analysis can bias outcomes and prevent the statistical tests from detecting real differences (thus decreasing the statistical power of the method). However, there are select circumstances in which it may be permissible to omit nondetects. For example, suppose a large number of measurements are available, only a small percentage of the data is nondetect, and the censoring limit is much smaller than the site's risk-based decision criterion. Under those conditions, if a statistical evaluation using only the detections indicates that contamination is present at levels significantly below the decision criterion, the omission of nondetects is unlikely to affect the outcome.

As a more general rule, nondetects should not be omitted but rather utilized and properly accounted for. The presence of nondetects provides valuable information about an environmental population. Eliminating nondetects often results in inaccurate test outcomes and can lead to greatly elevated mean or median concentration estimates and, importantly, underestimated variances. In addition, a large proportion of nondetects all well below a risk-based decision criterion constitutes strong evidence for the absence of significant contamination (regardless of whether a statistical evaluation can be done).

5.7.5 Simple Substitution Method

In the simple substitution method, proxy or surrogate numerical values are assigned to each of the

nondetects. The surrogate value for each nondetect is typically some fraction of the censoring limit (such as one half the detection limit or reporting limit). The impact of using simple substitution for nondetects depends greatly on what kind of statistical evaluation is to be performed. If the goal is to compute summary statistics such as the mean or variance (or quantities that depend on these statistics) — particularly for larger data sets or those with more than a low percentage of nondetects — it may be inappropriate to impute the censoring limit (or some fraction of this limit) to nondetects in statistical formulas because the substitution method distorts the data. In these cases, simple substitution can produce erroneous conclusions, particularly for data sets with very low concentrations or a large number of nondetects. For example, substituting the censoring limit for nondetects could result in a sample mean that is biased high, and substituting zero could result in a mean that is biased low, while simultaneously causing the variance to be either overestimated or underestimated.

In such applications, you should treat each nondetect as an inequality rather than a fixed numerical value. Although substituting one-half the censoring limit(s) for nondetects may not bias the mean, it can adversely affect estimates of the variance and statistics such as the upper confidence limit on the mean (which depends on the variance). Furthermore, the magnitudes of the substitutions — and perhaps the statistical evaluation itself — will depend on how the laboratory reports nondetects, including the size of the censoring limits, rather than on the actual but unknown concentrations in those samples. In general, the larger the fraction of nondetects and the more elevated the reporting limits, the greater the distortion of the data.

EPA's Unified Guidance suggests that the substitution method can be acceptable when only a small portion of the data set (10-15 percent) consists of nondetects. When the nondetect proportion is quite low, statistical results based on using simple substitution are not likely to vary substantially from other methods. However, it also recommends more sophisticated methods be used to handle nondetects for any larger data set when summary statistics are needed. For descriptive and graphical purposes, simple substitution may also be used to present preliminary summary statistics such as the sample mean and variance (for instance, when quantitative statistical evaluations are not planned), or when creating graphs like time series plots.

One additional consideration is that simple substitution may work fairly well in cases where parametric prediction limits are used with retesting to compare two populations (for instance, in release detection tests against background). Simulation studies have shown that simple substitution worked better than or as well as more complicated methods in that particular setting (McNichols and Davis 1988; Gibbons 1994; Gibbons and Coleman 2001), as long as the censoring proportion was not too high (more than 50%) and retesting was utilized as part of the test.

5.7.6 Kaplan-Meier Method

The Kaplan-Meier method is a nonparametric technique for calculating the (cumulative) probability distribution and for estimating means, sums, and variances with censored data. Originally, the Kaplan-Meier approach was developed for right-censored survival data. More recently, the method was reformulated for left-censored environmental measurements (e.g., nondetects). USEPA's

Unified Guidance also recommends the Kaplan-Meier method for use as an intermediate step in calculating parametric prediction limits, control charts, and confidence limits for censored data sets. In this latter application, the Kaplan-Meier estimate of the mean and variance is substituted for the sample mean and variance in the appropriate parametric formula.

Kaplan-Meier is one of a class of 'counting' techniques useful for accommodating censored data. It counts the number of data points below each detected concentration, and uses that information to generate an estimate of the probability distribution function. Kaplan-Meier accounts for the fact that data sets with nondetects can only be partially ranked (e.g., while a value of <5 is presumably less than a detected concentration of 10, it is not possible to determine whether or not it exceeds a detected concentration of 2). To get around this difficulty, the method only determines how many data values cannot exceed any given detected level. Once the (cumulative) probability distribution is estimated, statistics of interest like the mean or variance can be computed via areas under the distributional curve.

Applications and Relevant Study Questions

- Kaplan-Meier is most commonly used to calculate summary statistics like means and variances. It can be used in conjunction with bootstrapping and other methods to calculate upper confidence limits (UCLs) on the mean.
- The Kaplan-Meier method can also be used to sum data that include both censored and non-censored values. This approach is often used in environmental data when calculating toxicity equivalency (TEQ) for dioxins and benzo(a)pyrene equivalents.
- The Kaplan-Meier approach can also be used to improve parametric estimates of quantities like prediction and control chart limits that require means and standard deviations and an estimate of the cumulative distribution function (CDF) properly adjusted for the presence of nondetects. This strategy is discussed in USEPA's Unified Guidance.

Assumptions

- Kaplan-Meier is nonparametric, so it does not assume the data follow a known distribution.
- When applied as an intermediate step to calculate parametric statistics, Kaplan-Meier assumes that all data values come from a single underlying (non-negative) statistical population. In particular, contaminants are assumed to be present in nondetects at some low level not readily quantified by the analytical method.

Requirements and Tips

- To calculate Kaplan-Meier, you must have at least three detected concentrations, more than one reporting limit, and a detected value larger than all censored data. At least 8-10 measurements with no more than 50-70% nondetects are recommended. Note: if there is only one reporting limit, Kaplan-Meier is equivalent to simple substitution at the reporting limit, a strategy known to bias estimates of the mean and variance.
- This method can only include censored data less than the highest detected value, so be sure to consider information about censored data with high detection limits. The Kaplan-Meier

method cannot rank censored data points with reporting limits above the highest detected concentration. Thus Kaplan-Meier may not give accurate estimates in data sets with elevated reporting limits (perhaps due to high dilution factors during chemical analysis). One possible solution is to count the highest censored value as a detected concentration. This approach will tend to bias the mean upward but still allow computation of the Kaplan-Meier probability distribution.

- Another potential problem occurs when the lowest value is a censored value. Because estimates of the Kaplan-Meier (cumulative) probability distribution are only reported at the levels of detected concentrations, the distribution of censored data below the lowest detected value is unknown and not estimated. You can use Efron's bias correction to reduce the bias that occurs in this case. For this correction, simply convert the lowest censored data point to a detected value. Use care when performing this bias correction to ensure that the modified data point is ranked below other censored data points at the same reporting limit.

Strengths and Weaknesses

- Kaplan-Meier is well-suited for many environmental data sets because it is nonparametric, so that no underlying distribution need be assumed.
- Kaplan-Meier can accommodate multiple reporting limits and is routinely used with data sets having a lower than 50% detection frequency.
- One weakness of Kaplan-Meier is that it cannot rank censored data points with reporting limits above the highest detected concentration.
- Another weakness is that if only one reporting limit is present, Kaplan-Meier is equivalent to simple substitution at the reporting limit.

Further Information

Additional information on how to implement the Kaplan-Meier method and tools for calculating Kaplan-Meier distributions and sample means can be found in the Unified Guidance, the ProUCL documentation, and Helsel (2012). Helsel (2012) includes details on calculation of the mean and standard deviation using Kaplan-Meier methods. Beal (2009) presents an SAS macro for implementing the Kaplan-Meier method.

Kaplan-Meier Example

Consider the following data set:

20, 20, 10, 1, <0.5, <0.5, <25, <0.5, <2, <2, 1, 100, 30, 3, <3, 2, <25, 1, 3, 5

Order this data set in decreasing order: 100, 30, <25, 20, 20, 10, <10, 5, 3, 3, <3, 2, <2, <2, 1, 1, 1, <0.5, <0.5, <0.5

Apply Efron's bias correction:

100, 30, <25, 20, 20, 10, <10, 5, 3, 3, <3, 2, <2, <2, 1, 1, 1, <0.5, <0.5, 0.5 (note that the lowest censored data point, <.5, was converted to 0.5)

The highest value is a detected concentration at 100 and the data set includes 20 total data points. Since 19 data points are below 100, the probability of getting a data point less than 100 is 19/20 = 0.95.

The next highest data point is 30, which is also a detected concentration. Once the value of 100 is removed from the data set, the probability of getting a value below 30 is 18/19 = 0.947. The location of 30 on the probability distribution function is obtained by multiplying its probability by the probability of the next highest detected value: 0.95 * 18/19 = 0.90.

The next highest detected value is 20. Two data points are detected with concentrations at 20. Once all the detected and censored data points above 20 are removed, the probability of getting a value below 20 is 15/17 = 0.882. The location of 20 on the probability distribution function is 0.90 * (15/17) = 0.794.

This exercise can be continued to generate this probability distribution table:

Table 5-1. Kaplan-Meier Example Data

Value	Probability
100	0.95
30	0.9
20	0.794118
10	0.741176
5	0.684163
3	0.570136
2	0.506787
1	0.253394
0.5	0

When graphed, it looks like this:

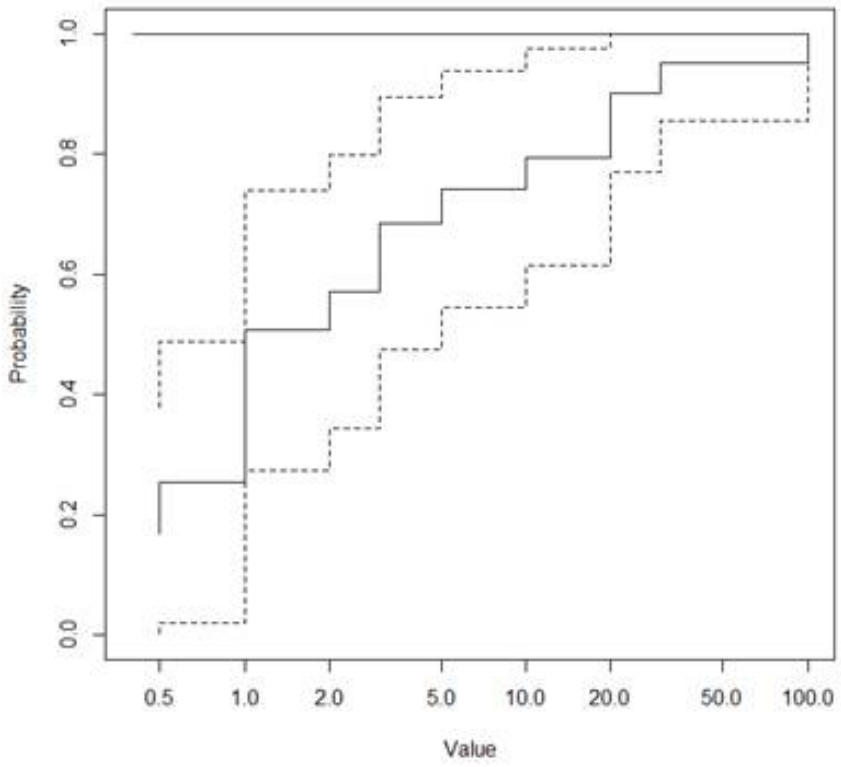


Figure 5-15. Kaplan-Meier method example data plot.

Integrate the area under this curve to determine a sample mean of 10.281.

5.7.7 Robust Regression on Order Statistics

Robust regression on order statistics (ROS) is a semi-parametric method that can be used to estimate means and other statistics with censored data. Unlike Kaplan-Meier, ROS internally assumes that the underlying population is approximately normal or lognormal. However, the assumption is directly applied to only the censored measurements and not to the full data set (hence the term 'semi-parametric'). In particular, ROS plots the detected values on a probability plot (with a regular or log-transformed axis) and calculates a linear regression line in order to approximate the parameters of the underlying (assumed) distribution. This fitted distribution is then utilized to generate imputed estimates for each of the censored measurements, which are then combined with the known (detected) values to summary statistics of interest (e.g., mean, variance). The method is labeled 'robust' because the detected measurements are used 'as is' to make estimates, rather than simply using the fitted distributional parameters from the probability plot.

Applications and Relevant Study Questions

- Robust ROS is most commonly used to estimate summary statistics like means and variances. It can be used in conjunction with bootstrapping and other methods to calculate upper confidence limits (UCLs) on the mean.
- ROS can also be used to improve parametric estimates of quantities like prediction and control chart limits that require means and standard deviations, properly adjusted for the presence of nondetects. This strategy is discussed in the Unified Guidance.

Assumptions

- To compute ROS, at a minimum, there must be at least three detected values and a detection frequency greater than 50%. More realistically, you should have at least 8-10 measurements.
- Robust ROS is semi-parametric. It assumes that the detected data can be fit to a known distribution on a probability plot, from which imputations are made for the nondetects. The estimated summary statistics are computed from a combination of the known and imputed measurements, rather than from the parameters of the fitted model.
- ROS assumes that all data values come from a single underlying (non-negative) statistical population. In particular, contaminants are assumed to be present in nondetects at some low level not readily quantified by the analytical method.

Requirements and Tips

Robust ROS will impute a value for each censored data point. However, these estimated values should not be used for any additional calculations other than estimating summary statistics for the data set as a whole.

Strengths and Weaknesses

- Robust ROS is widely applicable to many environmental data sets. However, as a semi-parametric method, you must be able to fit a known distributional model to the detected measurements on a probability plot.
- ROS can accommodate multiple reporting limits as well as (unlike Kaplan-Meier) a single reporting limit.

Further Information

Additional information on how to implement the Robust ROS method and tools for calculating statistics using the Robust ROS method can be found in Chapter 15.4, Unified Guidance, in the ProUCL documentation, and Helsel (2012).

Robust ROS Example

Consider the following data set, then list in descending order:

Table 5-2. Robust ROS data

Data set	Descending order
<1	3.2
<1	2.8
1.7	<2
<1	<2
<1	<2
<2	<2
3.2	<2
<2	<2
<2	<2
2.8	<2
<2	1.7
<2	1.5
<2	<1
<2	<1
<2	<1
0.7	<1
0.9	<0.9
0.5	0.9
0.5	0.7
<0.9	0.7
0.5	0.6
0.7	0.5
0.6	0.5
1.5	0.5

Starting at the highest detection limit, determine the probability of exceeding that percentile. For example, in the data set above with 24 values, the probability of exceeding the detection limit of 2 is 1 – 22/24 = 0.083333.

The remaining probability (0.9166667) is divided up evenly between the 8 values with a censoring limit of 2, resulting in probabilities of 0.8148, 0.7130, 0.6111, 0.5093, 0.4074, 0.3056, 0.2037, and 0.1019 assigned to these 8 samples.

Once this process has been completed for all detection limits, a regression line is fit to the detected portion of the data set. Then values are picked off the regression line and used as estimated concentrations for each censored data point.

This process results in a data set as shown in the following table:

Table 5-3. Robust ROS final data

Input Concentration	Probability	Estimated Concentration
3.2	0.972222	3.2
2.8	0.944444	2.8

Input Concentration	Probability	Estimated Concentration
<2	0.814815	1.41819
<2	0.712963	1.142868
<2	0.611111	0.953652
<2	0.509259	0.806562
<2	0.407407	0.682854
<2	0.305556	0.571858
<2	0.203704	0.464954
<2	0.101852	0.349209
1.7	0.873016	1.7
1.5	0.829365	1.5
<1	0.628571	0.982359
<1	0.471429	0.758549
<1	0.314286	0.581086
<1	0.157143	0.414444
<0.9	0.34375	0.612533
0.9	0.736607	0.9
0.7	0.589286	0.7
0.7	0.491071	0.7
0.6	0.392857	0.6
0.5	0.294643	0.5
0.5	0.196429	0.5
0.5	0.098214	0.5

The mean of the estimated data set (rightmost column in the table above) is 0.9725.

This data set is shown graphically below, where closed data points show detected concentrations and open data points show estimated concentrations for censored data points.

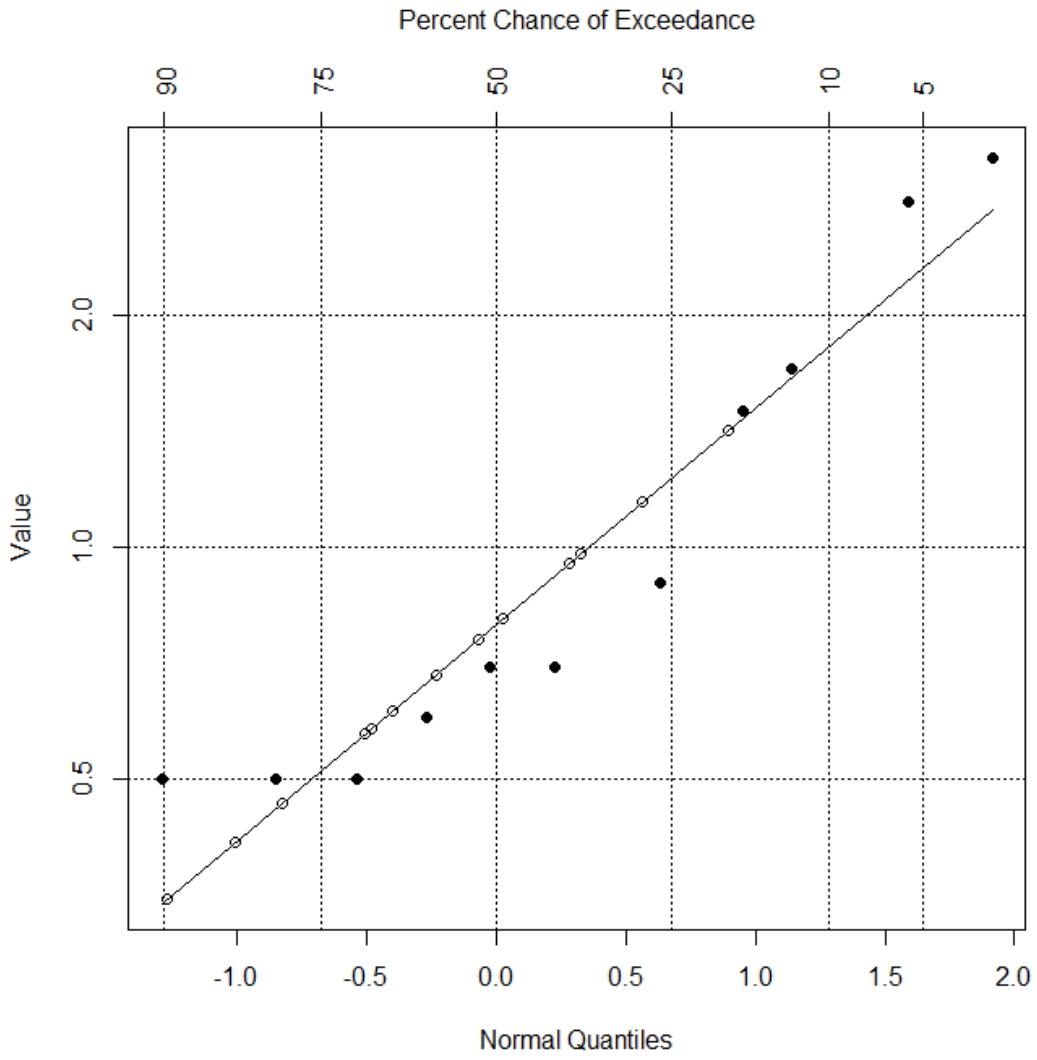


Figure 5-16. Robust ROS example data plot.

5.7.8 Maximum Likelihood Estimation (Including Cohen's Method)

Maximum likelihood estimation (MLE) is a parametric, model-based method that can be used to estimate means and other summary statistics with censored data. In this approach, you must know or assess what distribution (such as normal or lognormal) will best model the data set. The model parameters for that distribution (mean and variance) are then estimated by maximizing the likelihood of the observed values, while simultaneously treating each nondetect as an inequality. Once the model parameters are determined, other statistics can be estimated from the model.

Cohen's method (Chapter 15.5.1, Unified Guidance) is a simplified application of the MLE approach, where the underlying model is assumed to be normal (or transformed to normality) and the data contain but a single reporting limit, with all detected values larger than the nondetects.

Applications and Relevant Study Questions

The MLE approach (including Cohen's method) is most commonly used to estimate means and variances in larger data sets with known or assumed distributions. It can be used in conjunction with bootstrapping and other methods to calculate upper confidence limits (UCLs) around the mean. The mean and variance estimates (adjusted for censoring) can also be used in parametric formulas for prediction limits and control charts.

Assumptions

- To use MLE, the sample size must be large enough to assess the best-fitting underlying distribution. With multiple reporting limits, you might need at least 50 data points and a detection frequency greater than 50%. Cohen's method can be used with somewhat smaller data sets due to its more stringent assumptions.
- Data analyzed using MLE (including Cohen's method) are assumed to follow a known distribution, since the calculations depend explicitly on the assumed model. Distributional fitting using MLE works best on data sets with no obvious outliers and — if a normal model is assumed — that are not significantly skewed.
- Application of MLE assumes that nondetects are distributed in a manner similar to the detected values. Accurate estimates can only be anticipated when a common distributional model is valid for the both the detects and nondetects.

Requirements and Tips

MLE will perform poorly if a well-fitted or closely matching distribution cannot be found to model the underlying population. Censored probability plots and other goodness-of-fit techniques should be utilized to help assess this critical assumption.

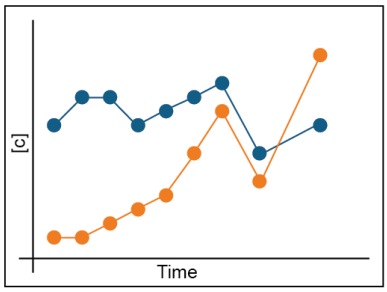
Strengths and Weaknesses

- The general MLE approach can handle multiple reporting limits and can be a rigorous way to estimate summary statistics of data sets when the sample size is sufficiently large. Cohen's method requires there to be only a single reporting limit.
- If the underlying distribution is known, MLE will explicitly account for distribution type in calculating estimates. If the assumed model is incorrect, MLE may lead to misleading results.
- MLE is most generally applicable to larger data sets ($n > 50$) with high detection frequencies. Under the stricter assumptions of Cohen's method, model-based mean and variance estimates can be computed with relatively high censoring rates (up to 50-80%).

Further Information

Additional information on how to implement Cohen's method can be found in Chapter 15.5, Unified Guidance.

5.8 Temporal Analysis



Temporal statistical analysis enables you to examine and model the behavior of a variable in a data set over time (e.g., to determine whether and how concentrations are changing over time. The behavior of a variable in a data set over time can be modeled as a function of previous data points of the same series, with or without extraneous, random influences (such as an earthquake or a new release). Common temporal analyses discussed below include time series plots, one-way ANOVA, sample autocorrelation, the rank von Neumann test, seasonality correlations, or the seasonal Mann-Kendall test. Table F-4 includes information about checking assumptions for multi-sample tests.

5.8.1 Time series plots

The time series plot provides a graphical view of the raw data. Time is plotted on the x-axis, and the data series observation or observations (for multiple series) are plotted on the y-axis. See Section 5.1.1: Time Series Methods of this document for a complete overview of time series plots.

Example

Figures 9-1, 14-1, and 14-2, Unified Guidance.

5.8.2 One-way ANOVA

ANOVA is a general purpose statistical approach used to compare data from three or more populations (with the data divided into one group/subset per population). Because of its flexibility and generality, ANOVA has utility for spatial analyses (for example, measuring contaminant level differences across multiple wells/sampling points), temporal analyses (for example, evaluating seasonality or temporal correlations across sampling events), as well as diagnostic testing (for example, testing for equal variances or identifying significant spatial variation).

For temporal analysis, the statistical populations to be compared by ANOVA represent distinct time periods, rather than distinct sampling points as in a spatial analysis. For instance, in cases of apparent seasonality at an individual well, each season (for example, spring or fall) is treated as a distinct population. In order to test for seasonality, each data subset must include representative observations from each distinct season — with a minimum of one sampling event per season collected over a period of at least three years.

When evaluating data sets for temporal patterns due to factors other than seasonality (but which impact a set of wells in common), each sampling event is treated as a separate population. The data are pooled across sampling points and then grouped/divided by sampling event. The ANOVA then compares the average levels per sampling event to look for differences between events that signify temporal patterns common to the set of wells.

In all parametric ANOVA analyses — regardless of how the data are grouped into subsets — the test (parametric F-test) returns an F-ratio statistic and an associated p-value. A large F-ratio (and small p-value) indicates that the observed differences between the subsets of data are more than expected based on chance alone, whereas an F-ratio close to one (large p-value) suggests that the differences may be due to random variation.

The Kruskal-Wallis test is a nonparametric counterpart to ANOVA that does not require normality of the ANOVA residuals. In this version, ranks of the data are used instead of the observed measurements, and an H-statistic is produced instead of an F-ratio, but the basic thrust of the test is the same. Average ranks are computed for each group being compared. If the differences in rank averages are larger than expected by random variation, the H-statistic will be large (with correspondingly small p-value), indicating a probable difference in the populations.

For diagnostic testing, one-way ANOVA can aid decisions about whether to conduct interwell or intrawell tests by identifying the presence of significant spatial variability among a group of sampling points. If the spatial variation is a natural phenomenon, the ANOVA results can help justify use of intrawell groundwater tests. Conversely, the lack of significant spatial variation can point to the use of interwell upgradient-downgradient testing.

Another variation of ANOVA, Levene's test, can also diagnose whether or not multiple populations have similar variances (see Chapter 11.2, Unified Guidance). In Levene's test, the absolute values of the residuals from a set of wells are treated as the 'data' in a standard one-way ANOVA. This tests whether the typical deviations from the mean of each well differ significantly among the wells, thus signifying differing levels of variance.

Applications and Relevant Study Questions

- This method can be used to evaluate stationarity (lack of a shift of the means over time).
- Use this method to check for the absence of spatial variability when evaluating temporal variations.
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?

Assumptions

- Residuals of the data must follow a normal distribution.
- Observations are statistically independent over time.
- Data must have constant variance.

Requirements and Tips

- Measurements collected at each well are performed on dates common to all wells.
- Data may need to be transformed (for example, using Box-Cox power transformation) if the assumptions of normality and equal variances are violated, and subsequently tested to evaluate the validity of the assumptions on the transformed scale.

- A level of confidence, such as 95% must be selected; this level of confidence may be determined by federal or state regulatory requirements or guidance, or project specific needs.
- Small sample sizes make it difficult to test the assumptions and may not allow for sufficient power. In order to test for seasonality, a minimum of three sampling events per distinct season, with events spanning at least three years is recommended.
- A minimum of 8 to 10 measurements is recommended when evaluating temporal variation due to factors other than seasonality.
- This test may be sensitive to outliers. Data should be checked for outliers before applying this test; see Section 5.10.
- If the purpose of the one-way ANOVA is to design an interwell prediction limit which accounts for temporal dependence, spatial variability must not be present.
- See Section 5.7 for information regarding handling of nondetects.
- If the data cannot be normalized, a similar test for a temporal or seasonal effect can be performed using the nonparametric Kruskal-Wallis test.

Strengths and Weaknesses

This method can be applied without specialized statistical software.

Further Information

Use of one-way ANOVA for spatial variability and an example problem are discussed in Chapter 13.2.2, Unified Guidance. Use of ANOVA to improve parametric intrawell tests is described in Chapter 13.3, Unified Guidance. Chapter 14.2.2, Unified Guidance discusses application of ANOVA for temporal effects and also provides a sample problem. A more generalized discussion of ANOVA is provided in Chapter 17.1, Unified Guidance.

5.8.3 Sample Autocorrelation Function

Autocorrelation is a correlation of a variable, such as a contaminant concentration, with itself over a series of time steps. Autocorrelation may be used to evaluate the frequency of sampling (for example, if subsequent sampling events are correlated, a reduction in sampling frequency may be supported). By computing the first few sample autocorrelation coefficients (ACFs), a plot of ACFs versus the time lags can be prepared; this graph is known as a correlogram (Figure 5-2). The shape of the ACF plot provides information regarding the variability of a given value over time.

A stationary but nonrandom series will often exhibit a large first-order autocorrelation coefficient, followed by one or two other significant coefficients, with the remaining coefficients tending towards zero. A seasonal series will exhibit a sinusoidal ACF. If the first order autocorrelation coefficient is significant and negative, the series tends to alternate between high and low values. If the series contains a trend, the ACF coefficients will not drop to zero with increasing lag.

Applications and Relevant Study Questions

- A comparison of the ACF coefficients that shows correlated consecutive time steps may support a reduction in sampling frequency.

- Study Question 6: Is there seasonality in the concentrations?
- Study Question 9: Is the sampling frequency appropriate (temporal optimization)?

Assumptions

Data must follow a normal distribution or be reasonably symmetric.

Requirements and Tips

- Check the data for normality.
- Use of at least 8 to 10 measurements is recommended, although a greater number of measurements may be necessary to obtain the desired confidence level or power.
- If outliers are suspected, examine the data with a probability plot, Dixon's test, or Rosner's test. Remove outliers from the data set.
- Select a level of confidence, such as 95%; this level of confidence may be determined by federal or state regulatory requirements or guidance.
- If you suspect autocorrelation of a series, change the sampling frequency. The smallest lag between sampling events with no serial discernible correlation indicates the minimum sampling frequency needed for statistical independence. If you suspect seasonal autocorrelation (at the appropriate lag), see Section 5.8.5: Seasonality Correlations. An ACF plot exhibiting a sinusoidal shape indicates seasonality.

Strengths and Weaknesses

Requires a higher level of analysis than other methods to interpret results.

Further Information

Further information on the sample autocorrelation function and an example problem are provided in Chapter 14.2.3, Unified Guidance. Partial autocorrelation coefficients can also be computed (see Chatfield 1994). Significant autocorrelation and partial autocorrelation coefficients can be combined for the construction of a Box-Jenkins autoregressive integrated moving average (ARIMA) time series model (Box and Jenkins 1976). Such a model can be used for prediction purposes. For two series, a cross-correlation function (CCF) can be constructed (Box and Jenkins 1976).

Example

See Example 14-3, Unified Guidance (which includes Figure 14-5, a sample autocorrelation function), and case example A.3.

5.8.4 Rank von Neumann Ratio Test

The rank von Neumann ratio is used to evaluate seasonality in a data set and is constructed from the sum of differences between the ranks of lag-1 data pairs (for example, data pairs generated by comparison of data collected in a monitoring event to data generated in the previous monitoring event). When these differences are small, the pattern of observations of the data series will be somewhat predictable, and the data series is likely to be autocorrelated. Large differences indicate no autocorrelation. The test is formally conducted by comparing the Rank von Neumann ratio to the

tabulated critical points (at a given sample size and desired significance level; see Table 14-1 of Appendix D, Unified Guidance). The Rank von Neumann Ratio test is a nonparametric method.

Applications and Relevant Study Questions

Study Question 6: Is there seasonality in the concentrations?

Assumptions

No distributional assumptions are required; however, frequent nondetects in the data may lead to a poor estimation of the Rank von Neumann ratio and critical points.

Requirements and Tips

- Check the data for nondetects; replace each tied value by its mid-rank.
- Use of a minimum of 10-12 observations from a single well is recommended.
- Check that the data are not autocorrelated.
- Select a level of confidence, such as 95% (or 99%); this level of confidence may be determined by federal or state regulatory requirements, or guidance, or by project specific needs.

Strengths and Weaknesses

- This method is easily applied to nonparametric tests.
- This method identifies simple temporal correlations.
- You must apply this method to a single series of data at a single data point, not to multiple series of data.
- Use this method only on data sets with few nondetects.
- Compared to other tests of statistical independence, the Rank von Neumann ratio is more powerful than certain other nonparametric methods. The Rank von Neumann ratio correctly detects dependent data for a variety of underlying data distributions.

Further Information

The Rank von Neumann ratio test is discussed further in Chapter 14.2.4, Unified Guidance, which also gives an example problem. The literature on time series analysis is extensive for other potential tests as well (such as Runs test, Durbin-Watson test, and Kendall's tau).

Example

See Example 14-4, Unified Guidance.

5.8.5 Seasonality Correlations

If the seasonal pattern in a data series is highly regular, then you can model the data with a sinusoidal function. Moving averages and lag-based differencing (for example, lag-4 for quarterly data, or lag-12 for monthly data) can be used to evaluate the data; see Chapter 14.3.3.1, Unified Guidance. When a significant temporal dependence is identified across a group of wells (for instance, by one-way ANOVA), the adjustment process (moving averages) can be conducted simultaneously for several sets of wells as described in Chapter 14.3.3.2, Unified Guidance.

Applications and Relevant Study Questions

- Use this method to de-seasonalize data.
- Study Question 6: Is there seasonality in the concentrations?

Assumptions

Seasonal correction is only appropriate for wells where a cyclical pattern is clearly present.

Requirements and Tips

- Use at least a two-year period of data.
- A minimum of three measurements per season is recommended for the application of seasonal corrections.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, Rosner's test, or another appropriate method.
- See Section 5.7 for information regarding the handling of nondetects.
- For interwell comparisons, the same seasonality effect must be present in all wells.
- For interwell comparisons (such as simultaneously collected background and downgradient data), a seasonal correction may not be necessary if the background and downgradient values are on the same cycle. To dispense with seasonal correction in this case, average groundwater velocities must be high enough for groundwater to migrate through both background and downgradient wells in the same season.

Further Information

A description of how to de-seasonalize a data set (or multiple data sets) is given in Chapter 14.3, Unified Guidance. De-seasonalizing data can also be conducted by differencing (see Box and Jenkins 1976). A two-way ANOVA may be conducted to test for both spatial variation and temporal autocorrelation (Davis 1994). Example problems are provided in Chapter 14.3.3, Unified Guidance.

5.8.6 Seasonal Mann-Kendall Test

The seasonal Mann-Kendall test is a simple modification to the Mann-Kendall test for trend that accounts for seasonal fluctuations. The data series is divided into subsets, with each subset representing the measurements collected during a common sampling event. The standard Mann-Kendall test is performed separately on each subset, with a test statistic performed for each individual subset. The separate, seasonal statistics are subsequently summed to arrive at the overall Mann-Kendall statistic, which is then compared to the critical points of the standard normal distribution.

Applications and Relevant Study Questions

- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?

Assumptions

The long-term mean of the data series should be stationary.

Requirements and Tips

- The sample series should span at least three seasons, with an observable seasonal pattern.
- Each season should include at least three measurements in order to compute the Mann-Kendall statistic.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, Rosner's test, or another appropriate method.
- See Section 5.7 for information regarding the handling of nondetect data.
- A normal approximation to the overall Mann-Kendall test statistic plot must hold.
- Select a level of confidence, such as 95%; this level of confidence may be determined by federal or state regulatory requirements, guidance, or project specific needs.
- You can also choose to remove the seasonal autocorrelation first (see Section 5.8.5: Seasonality Correlations) and subsequently conduct a formal trend test on the entire series.

Strengths and Weaknesses

This is a nonparametric test.

Further Information

See Chapter 14.3.4, Unified Guidance for additional information and a sample problem. See also Gilbert (1987) for a description of the seasonal Mann-Kendall trend test and slope estimator.

5.8.7 Temporal Optimization (Cost-Effective Sampling and Iterative Thinning)

Temporal optimization is best represented by the cost-effective sampling method (CES; Ridley et al. 1995; Ridley and McQueen 2005) and later modifications to this approach. In CES, a linear trend is estimated for each chemical-well pair and then classified according to the slope of the apparent trend as well as how much variation exists around the trend. Trends with relatively 'flat' slopes (small rates of change) and low variation are recommended for less frequent sampling, while trends with higher slopes or higher degrees of variation are targeted for more frequent sampling. The overriding principle is to (1) sample more frequently at locations where the apparent changes are more dynamic and associated with the greatest statistical uncertainty, and (2) sample less frequently when the trend is changing little and is statistically more certain (that is, less variable).

The second approach is the iterative thinning method (Cameron 2004). Iterative thinning examines whether sampling frequencies can be reduced due to temporal redundancy in the sampling events. This approach identifies redundancy by first estimating a baseline trend using the full data set, after which the trend is repeatedly re-estimated using subsets of the full data to identify the average number of data points needed to accurately reconstruct the baseline. The computations in iterative thinning create a series of 'what if' scenarios estimating the nature of the trend that would have been identified if only some of the existing data had been sampled. The overriding principle in iterative thinning is that if a trend can be accurately reconstructed using fewer sampling events, the optimal sampling frequency should be based on this smaller number.

Applications and Relevant Study Questions

Study Question 9: Is the sampling frequency appropriate (temporal optimization)?

Assumptions

The basic assumptions underlying temporal optimization methods are similar to those for most trend tests. CES and its modifications assume the trend is linear. Also, if linear regression is used to measure the trend, the trend residuals must be normal and homoscedastic. Iterative thinning can be performed on linear or non-linear trends, but typically requires at least 8 observations from which to form the (non-linear) baseline trend.

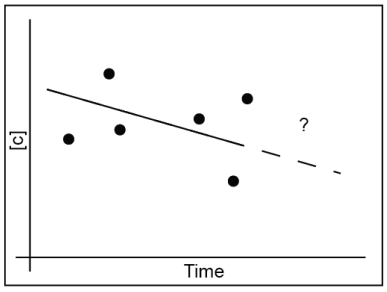
Requirements and Tips

- CES and its modifications generally require at least 4 observations per well; Iterative thinning usually requires at least 8 observations per well in order to estimate non-linear trends.
- CES is found in specialized optimization software packages such as MAROS and 3TMO. Iterative thinning is deployed in the geostatistical temporal-spatial (GTS) software and VSP.

Strengths and Weaknesses

- CES and its modifications can employ either parametric (linear regression) or nonparametric (Mann-Kendall) trend methods.
- Iterative thinning, as deployed in GTS software, can be applied to either linear or non-linear trends.

5.9 Time Series Forecasting



Time series forecasting allows univariate or multivariate forecasting of future values of an observed time series or multiple time series over a specified forecasting horizon (time frame). For example, what might the anticipated concentration of a chemical be in a given compliance well in two years? Forecasts are based on a model fitted to present and past observations. Either an automated model or a user specified model may be used. Time series forecasting follows on the discussion of sample autocorrelation function (Section 5.8.3); review Section 5.8.3 if you are not familiar with time series forecasting and autocorrelation functions.

Automated and User-specified Approaches

5.9.1 Automated models (such as Holt, Holt-Winters Forecasting)

For automated models such as Holt or Holt-Winters forecasting, a program automatically analyzes the data, selects forecasting techniques, and generates a forecast. Using an automated approach, exponential smoothing procedures rely on simple, recursive updating equations (geometrically weighted sums of past observations, with more emphasis placed on recent observations and less emphasis on more distant observations). These procedures can also account for trends and seasonal variations (see Chatfield 1994). The smoothing parameter is generally subjectively chosen to be

between 0.1 and 0.3 (its exact value is typically not critical), but can be numerically estimated as well.

5.9.2 User-specified models (such as ARIMA).

Autoregressive integrated moving average (ARIMA) procedures rely on the analyst's subjective judgment or knowledge to select an appropriate model from a broad class of available models for a given data series. Interpretation of correlograms produced by the autocorrelation function (ACF), as well as the partial autocorrelation function (plotted against the lag in time), suggests which model might be appropriate. If the groundwater data series contains a trend, differencing the data (by calculating new data points based on the calculated lag) usually produces a stationary series. The residuals of the fitted model must be analyzed to verify the appropriateness of the model (such as, by the Portmanteau lack-of-fit test or by the Durbin-Watson statistic). Forecasts for a given lead time can then be readily computed by the difference equations (Box and Jenkins 1976).

Applications and Relevant Study Questions

- Study Question 4: When will contaminant concentrations reach a criterion?

Assumptions

- Time series consists of at least eight observations, recorded at equally spaced (or nearly equally spaced) intervals in time. If seasonal variation is present, the time series should encompass at least two full cycles (for example, quarterly data requires at least eight observations, data collected monthly requires at least 24 observations). A minimum of three full cycles is recommended if the seasonal variation is not clearly defined.
- Concentrations (or other variables) change in a set way with time (cyclical variation, steadily increasing/decreasing rate).
- Prediction intervals typically assume that the forecasts are unbiased, and that the forecast errors are normally distributed.

Requirements and Tips

- If you suspect outliers, examine the data (with a probability plot, Dixon's test, or Rosner's test) and consider removing verified outliers.
- Select a level of confidence for the forecasts, such as 50% and 95%; this level of confidence may be determined by federal or state regulatory requirements or guidance.
- Forecast future values of an observed time series in conjunction with the calculation of a prediction interval at a given confidence level, because typically, with increasing lead time, the uncertainties of point forecasts increase rapidly.
- The forecast should be limited to one quarter of the length of the observed time series.
- Regularly update the time series, as soon as new observations become available, to decrease the forecast error for a given lag (forecast horizon).
- If data appear to be stationary, consider fitting a simple model, such as exponential smoothing, or an autoregressive moving average (ARIMA) model with few parameters.

- If data appear to be nonstationary, consider either exponential smoothing with a trend term, or data transformation such as log or differencing before fitting a general integrated ARIMA model.
- If data exhibit seasonal fluctuations, consider either seasonal exponential smoothing, or differencing the data with a seasonal lag before fitting a general, additive or multiplicative ARIMA(or seasonal ARIMA) model.
- If the variance of the data changes with time, consider the use of a model that can consider more than one distribution for the data and preserve that information, such as autoregressive conditional heteroscedasticity (ARCH) and generalized autoregressive conditional heteroscedasticity (GARCH).
- Consider the use of external predictor or explanatory variables, if additional information is available and relevant.

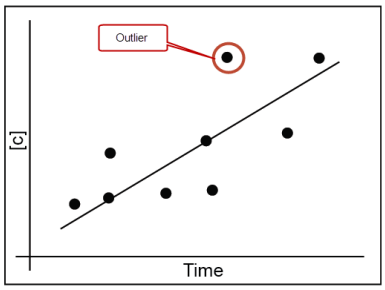
Strengths and Weaknesses

- The automatic (Holt-Winters) approach to time series forecasting is typically easier to implement and should be adequate for most groundwater data series where univariate forecasting is sufficient.
- The ARIMA (Box-Jenkins) approach requires a more thorough understanding of the underlying stochastic process and order of autoregressive and moving average terms, but is useful for more sophisticated forecasting analyses, especially if correlations with other time series are also being considered.

Further Information

If you suspect that the observed time series is oscillatory, and revolves around a constant mean, methods from the theory of linear prediction can also be used (see Yaglom 1962).

5.10 Identification of Outliers



Outliers are data that appear anomalous or outside the range of expected values. Outliers may indicate errors, may indicate data unrelated to the rest of the data set, or may be perfectly valid data that indicates contamination or unusual hydrogeological conditions. In assessing chemical analyses of groundwater, it is often difficult to determine the reason for outliers. Possible reasons for outliers are recording errors, unusual sampling and laboratory procedures or conditions, or inconsistent sample turbidity. The outlier may represent an unusual hydrological condition, sampling of unrelated groundwater, or the presence of locally controlled conditions. An outlier may also be an indication of contamination. It is crucial, therefore, to carefully evaluate the possible causes for outliers.

This section presents useful tests to identify outliers; unfortunately, identifying outliers in environmental contexts is not an exact science and there is no list of clear rules to follow in identifying outliers. The goal of outlier identification is to properly analyze the data to determine which outliers

are representative of valid data points (and should be kept), and which outliers likely represent errors, and should be removed from the data set. Data should not be excluded simply because they are identified as outliers. Once you have been identified outliers should be further evaluated to determine the reason for their existence. Outliers should generally be kept as part of the data set unless there is reasonable evidence that they are the result of an error. Many statistical tests require that outliers resulting from error be removed; some statistical tests may also require removal of valid, but extreme outliers that are not representative of the general population. The presence of outliers may preclude the use of some statistical methods altogether, requiring for example, a non-parametric alternative.

Box plots and probability plots are good tools for screening the data to identify possible outliers. Dixon's test may be used to evaluate a single suspected outlier. If multiple outliers are suspected, each outlier should be tested individually, beginning with the least extreme and progressing to each of the next extreme values until an outlier is confirmed. At that point, all values that are more extreme are also confirmed as outliers. Data sets with more than 20 values can be tested for multiple outliers using Rosner's test. Dixon's and Rosner's tests are more formal outlier tests involving the computation of a statistic that is compared to tabulated critical values.

As with all statistical procedures, data sets with many nondetects require care in applying outlier tests and in evaluating the practical implications of performing the tests on detects only or on data which include nondetects. In all cases, if nondetects are present in a data set, the results of outlier testing should be carefully examined to ensure validity from both a practical standpoint and a numerical basis.

5.10.1 Box plots or box and whisker plots

Box plots can be used as an initial screening tool for outliers as they provide a graphical depiction of data distribution and extreme values. Some software can also be programmed to display as outliers data values that exceed a specified distance from the measure of central tendency (mean or median). See Section 5.1.2: Box Plots of this document for more detail. Chapter 12.2, Unified Guidance provides a discussion on screening for outliers using box plots.

5.10.2 Probability plots

Probability plots are used for graphically displaying a data set's conformance to a normal distribution, and can be used as a screening tool for the initial identification of outliers. See Section 5.1.5: Probability Plots for more detail. A brief description of screening for outliers using probability plots is provided in Chapter 12.1, Unified Guidance.

5.10.3 Dixon's test

Dixon's test for single high or low outliers is relatively easy to perform and is offered in many statistical software packages. Generally, the data are ordered from lowest to highest, and the test computes a ratio between two values: the difference between the suspected outlier and a population value "near" the potential outlier, compared to the range of sample values in the population. The

value of this ratio, (the test statistic), is then compared to a tabulated critical value that is based on the sample size and desired confidence level; if the test statistic is greater than the critical value, the suspected outlier is confirmed as a statistical outlier.

Dixon's test evaluates a single suspected outlier. If you suspect more than one outlier, test these outliers individually, beginning with the least extreme and progressing to each of the next extreme values until an outlier is confirmed; at that point all values that are more extreme are also confirmed as outliers. Data sets with more than 20 values can be tested for multiple outliers using Rosner's test.

Applications and Relevant Study Questions

This test is a formal statistical test to identify outliers; it is most useful for small ($n \leq 25$) data sets with a single suspected outlier.

Assumptions

Only one outlier is present.

Requirements and Tips

- Data are normally-distributed (when suspected outlier removed).
- Number of sample values is < 25.

Strengths and Weaknesses

- The test is simple to implement and you can perform calculations by hand.
- This test can be used with small data sets.
- This test is widely available in statistical software packages.

Further Information

Additional information regarding the application of Dixon's test is provided in Chapter 8.3 and Chapter 12.3, Unified Guidance. Chapter 12.3, Unified Guidance includes a sample problem.

5.10.4 Rosner's test

Rosner's test helps to identify multiple outliers in a data set with at least 20 normally-distributed values. To use this test, you must first determine the number of potential outliers or extreme values present. This number of extreme values is then tested as a group and if the test is significant, all are determined to be outliers. If the test is not significant, the least extreme value of the group is removed and the reduced group is evaluated again. The process continues until either a significant result is obtained or the entire group is eliminated from outlier status.

Since Rosner's method tests a group of suspected outliers, a significant result can be returned even though not all of the data values in the group are individual outliers. Thus it is particularly important to pre-screen the data (visually or graphically) to make the selection of potential outliers as accurate as possible prior to running Rosner's test. However, this test may identify multiple outliers

in a single pass, unlike Dixon's test, which must be performed iteratively to screen for multiple outliers.

Applications and Relevant Study Questions

A formal statistical test to define outliers; useful for multiple suspected outliers in moderate to large data sets.

Requirements and Tips

- Data are normally-distributed (when suspected outliers are removed).
- Required sample size is at least 20.

Assumptions

The data follow a normal distribution (or can be normalized) and the outliers come from a different distribution.

Strengths and Weaknesses

- You must determine which points are potential outliers before conducting the test.
- This test is good for larger data sets.
- This test is not as simple to perform as Dixon's test for a single outlier.
- This test is widely available in statistical software packages.

Further Information

A description of how to conduct Rosner's test is found in Chapter 8.3 and Chapter 12.4, Unified Guidance. A sample problem is also provided in Chapter 12.4, Unified Guidance.

5.11 One Sample and Two Sample Tests

One-sample tests are used to compare the data set to a fixed criterion (for example, population mean, population percentile). Examples of one-sample tests have already been implicitly presented in Section 5.3 (Tolerance Limits) and Section 5.4 (Prediction Limits), as well as Section 5.6 (Distributional Tests). Other examples are goodness-of-fit tests, where, for example, you would like to know if the data support predictions regarding the value of the population mean. The null hypothesis would be:

$$t = \frac{\overline{x}_C - \overline{x}_{BG}}{\sqrt{\frac{S_{BG}^2}{n_{BG}} + \frac{S_C^2}{n_C}}}$$

H_0: $\mu = \mu_o$ where μ = actual true population mean

μ_0 = hypothesized population mean (under H_0)

and the alternative hypothesis H_A: $\mu <> \mu_0$.

A one sample t-test can be applied in this case, if the following assumptions hold:

- the data are normally distributed
- the sample drawn from the population is random
- the cases of the samples are independent
- the population mean is known

However, many groundwater monitoring scenarios require the comparison of two populations, such as a population of compliance (potentially impacted) data to a population of spatial or temporal background (unimpacted) data. The statistical tests used for these comparisons are referred to as two-sample tests and are used to determine if the two populations are statistically different at a specified level of significance. Examples of parametric two-sample tests include Welch's t-test and the pooled variance t-test. Nonparametric tests include the Wilcoxon rank sum test, the signed rank test, and the Tarone-Ware two sample test for censored data. These two-sample tests and their applications are described briefly below. Table F-3 includes information about checking assumptions for two sample tests.

5.11.1 Welch's T-test

Welch's t-test assumes that each population is normally distributed and requires that no temporal trends exist in the data, no spatial variability is present, and samples are statistically independent. One advantage of Welch's t-test is that it does not require you to assume that population variances are equal. Another advantage is that while Welch's t-test provides statistical power comparable to other two-sample tests, it is much simpler to use than other similar tests. The only calculations required are computing the mean, standard deviation, variance, t-statistic, and degrees of freedom. Many statistical software packages offer Welch's t-test, but most do not determine if the requirements and assumptions are met.

When applying Welch's t-test, the calculated t-value is compared to a critical t-value which is based on the selected significance level of the test and on the number of degrees of freedom. If the calculated t-value is less than or equal to the critical value, then no evidence exists for a statistically significant difference between the two population means at the selected confidence level. The equations for the necessary calculations, including the critical t-values for common significance levels, can be found in most statistical texts and in the Unified Guidance.

Applications and Relevant Study Questions

- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 5: Is there a trend in contaminant concentrations?

Assumptions

Data are normally distributed. This test will still provide relatively reliable results if data are not heavily skewed (coefficient of variation is less than or equal to 1.5).

Requirements and Tips

- No naturally-occurring spatial variability can be present.

- Samples must be spatially and temporally independent.
- No temporal trends in the data can be present.
- Use of 8 to 10 measurements is recommended, a larger data set may be required if the data are skewed or contain nondetects.

Strengths and Weaknesses

- This test does not require equal population variances.
- This test should not be used on lognormal data which are transformed from normal data.
- This test is simpler to use than other two-sample tests with comparable statistical power.

Further Information

Additional information on Welch's t-test including examples of how to perform the test can be found in Chapter 16.1.2 and Chapter 16.1.3, Unified Guidance.

5.11.2 Pooled Variance T-test

The pooled variance t-test shares the same underlying assumptions and requirements of Welch's t-test but, provides greater statistical power and therefore is helpful in identifying smaller differences. However, the pooled variance t-test has the added requirement that the variances of the two populations be equal; this requirement can be evaluated using box plots, or more robust methods such as Levene's test for equal variances (see Section 11.2, Unified Guidance). If these assumptions are met, the t-statistic can be calculated. Many statistical software packages offer versions of the pooled variance t-test, but most do not determine if the requirements and assumptions are met.

As with Welch's t-test, the calculated t-value is compared to a critical t-value, which is based on the selected significance level of the test and on the number of degrees of freedom. If the calculated t-value is less than or equal to the critical value, then no evidence exists of a statistically significant difference between the two population means at the specified confidence level. The equations for the necessary calculations, including the critical t-values for common significance levels, can be found in most statistical texts and in the Unified Guidance.

Applications and Relevant Study Questions

- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 5: Is there a trend in contaminant concentrations?

Assumptions

This test assumes that data are normally distributed. If this assumption cannot be met, consider using other parametric or nonparametric two-sample tests such as those discussed in this section.

Requirements and Tips

- No naturally-occurring spatial variability can be present.
- This method requires spatially and temporally independent samples.
- No temporal trends can be present in the data.
- Population variances must be equal.

- If you suspect outliers, examine the data using a probability plot, Dixon's test, Rosner's test, or another appropriate method.
- See Section 5.7 for information regarding the handling of nondetects.
- Use of 8 to 10 measurements is recommended, a larger data set may be required if the data are skewed or contain nondetects.

Strengths and Weaknesses

- This method is relatively simple to implement and interpret (when assumptions are met).
- Use on lognormal data which are transformed is not recommended.

Further Information

Additional information on the Pooled Variance t-test, including examples of how to perform the test can be found in Chapter 16.1.1, Unified Guidance.

5.11.3 Wilcoxon Rank-sum Test

The Wilcoxon rank-sum test is a nonparametric two-sample test that may be used to compare two populations when the groundwater data are not normally-distributed and cannot be normalized by transformation. The Wilcoxon rank-sum test is equivalent to the Mann-Whitney U-test. Requirements for the Wilcoxon rank-sum test include the assumption of equal variances, the assumption of a common (unknown) distribution, a lack of spatial variability, and temporal stability. The Wilcoxon rank-sum test can handle data sets with a limited number of nondetects (10-15%) with uniform reporting limits.

As the name implies, the Wilcoxon rank-sum test is performed by ordering the combined data from smallest to largest and ranking the values from 1 to N. Tied values receive a midrank which is the average of the ranks they would receive were they not tied. The resulting numerical ranks of the background samples are denoted as B_i and the compliance samples are C_i. The Wilcoxon statistic (W) is computed as the sum of the compliance ranks and the result is standardized to compute a Z-score for comparison to a tabulated critical statistic. Calculations for W, the expected value E(W), standard deviation SD(W), and the test statistic Z, for data with no ties are available in most statistical references and the Unified Guidance.

A computed Z is greater than the tabulated critical Z at the selected significance level, indicates that the compliance well concentrations are statistically different from the background at the significance level.

The Wilcoxon rank-sum test is available in most statistical software packages as a default selection for nonparametrically-distributed data; however, most packages do not automatically evaluate for compliance with the necessary underlying requirements or assumptions.

Applications and Relevant Study Questions

- Study Question 2: Are concentrations greater than background concentrations?

- Study Question 5: Is there a trend in contaminant concentrations?

Assumptions

Although there is no assumption of normality, violations of the requirements listed below may invalidate the results of the test. Always verify that the data comply with the requirements.

Requirements and Tips

- Equal population variances
- Common (shared) distribution between populations
- Absence of naturally-occurring spatial variability
- Samples are spatially and temporally independent
- Temporal stability
- The number of nondetects should be minimal (typically, less than 10 to 15%) and should be treated as tied data.
- Use of 8 to 10 measurements is recommended, a larger data set may be required if the data are skewed or contain nondetects.

Strengths and Weaknesses

- no requirement for normality
- can accommodate nondetects, but a large number of nondetects may decrease the usefulness of the result.

Further Information

Additional information on the Wilcoxon Rank-Sum test including examples of how to perform the test can be found in Chapter 16.2, Unified Guidance.

5.11.4 Sign or Signed Rank Test

The signed rank test is used to evaluate differences between groups of "paired" data such as analytical results from a group of wells before and after remediation efforts. The signed rank test evaluates whether a statistically significant difference exists between the medians of two groups by evaluating the difference between each pair of observations. The pairs are ranked in ascending order of the absolute value of their difference, and each rank is multiplied by the sign of the paired difference. The sum of those products is the test statistic W, which is compared to a tabulated critical value that is based on the selected statistical significance of the test and the number of sample pairs (differences). A computed test statistic W greater than the tabulated critical W at the selected significance level, indicates that the two groups of data are statistically different at the selected significance level. The signed rank test is available in some statistical software packages and is relatively straightforward to implement in spreadsheet software.

Applications and Relevant Study Questions

Study Question 5: Is there a trend in contaminant concentrations?

Requirements and Tips

- Use of 8 to 10 measurements is recommended, a larger data set may be required if the data are skewed or contain nondetects
- See Section 5.7 for information on handling of nondetect data.

Assumptions

- Data are paired and come from a common population.
- Each pair is independent of the other pairs.

Strengths and Weaknesses

- This test has no requirement for normality.
- This test is not designed to accommodate nondetects.

Further Information

Additional information on the signed-rank test, including examples, can be found in a variety of statistical texts and guidance documents (Gilbert 1987).

5.11.5 Tarone-Ware Two-sample Test for Censored Data

The Tarone-Ware two-sample test provides the added versatility of dealing with nondetect data. Like other nonparametric tests, Tarone-Ware assumes identical distribution of background and compliance populations, and requires equal variances. Also, as with the other tests, the Tarone-Ware two-sample test also requires temporal stability and lack of spatial variability. To perform this test, the two data sets (for example, background and compliance data) are combined and the distinct (unique) detect values ordered from lowest to highest. The number of values (including nondetects) less than or equal to each ordered value is computed for compliance, background, and combined data. The Tarone-Ware statistic is then calculated using equations found in some statistical references, including the Unified Guidance. Variations of this test (such as Gehan's (1965) generalized Wilcoxon test) are also found in some statistical software packages, although compliance with the underlying assumptions and requirements is generally not automatically evaluated.

A computed Tarone-Ware statistic (TW) greater than the tabulated critical value at the selected significance level, indicates, given the example above of comparing background and compliance data, that the compliance well concentrations are statistically different (greater) than the background at that significance level.

Applications and Relevant Study Questions

- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 5: Is there a trend in contaminant concentrations?

Assumptions

- Equal population variances
- Samples are spatially and temporally independent
- Temporal stability

Requirements and Tips

- Use of 8 to 10 measurements is recommended; a larger data set may be required if the data are skewed or contain nondetects.
- Although the method does not require normality, significant deviations from the requirements listed can invalidate results.
- Although this test is robust with respect to the presence of nondetects, general equality of variance should be visually checked using box plots.

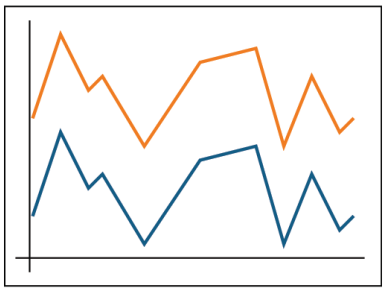
Strengths and Weaknesses

- This test does not require normality in the data set.
- This test addresses nondetect related limitations found in other nonparametric methods.
- This test is not as available in standard environmental statistics software packages as other nonparametric methods.

Further Information

Additional information on the Tarone-Ware two-sample test, including examples of how to perform the test, can be found in Chapter 16.3, Unified Guidance.

5.12 Correlation Tests



Correlation tests can be used to assess whether two groundwater variables have a linear relationship with each other. Correlation tests may be used to evaluate both positive (when one variable increases, the other variable increases) and negative (when one variable increases, the other variable decreases) correlations. An example of a positive correlation would be an observation that chemical concentrations in a well increase when water levels in the well increase. An example of a negative correlation would be an observed decrease in concentrations when the pumping rate for a groundwater extraction system is increased. These tests may also be used to test for monotonic trends or to compare trends.

5.12.1 Pearson Correlation Test

The parametric Pearson correlation test provides a measure of the linear association between two continuous variables. To conduct the test, correlation coefficients are calculated for each (x,y) pair, and the values of x and y are subsequently replaced with their ranks. Application of the test results in a correlation coefficient that ranges from -1 to 1. The sign of the coefficient indicates the

direction of the relationship (that is, negative values imply an inverse relationship or a decreasing trend), and its absolute value indicates its strength, with larger (absolute) values indicating stronger linear relationships.

Applications and Relevant Study Questions

The Pearson correlation coefficient is a common numerical measure of the degree of linear association between two continuous variables.

Study Question 5: Is there a trend in contaminant concentrations?

Assumptions

- Linear relationship between variables should hold.
- Variables should be identically distributed (but not necessarily independently).
- Assumes parametric distribution

Requirements and Tips

- A minimum of two variables with at least three observations for each variable are needed in order for the test to be meaningful.
- Use 8 to 10 paired observations, although a larger data set may be needed if the data sets are skewed or contain nondetects.
- The degree of confidence in order to detect patterns in the data increases with larger sample sizes.
- See Section 5.7 for information regarding the handling of nondetects.
- You may need to standardize each variable for plotting purposes in order to preserve the scales.

Strengths and Weaknesses

This test does not recognize nonlinear relationships between variables.

Further Information

A description of how to construct scatter plots is found in Chapter 9.4, Unified Guidance. Formula [3.5] of Chapter 3.3, Unified Guidance shows how to construct the Pearson correlation coefficient.

5.12.2 Spearman Rank Correlation Coefficient

The Spearman rank correlation test is essentially the nonparametric version of the Pearson correlation coefficient test, and provides a measure of the linear association between two variables. Spearman's rank correlation coefficient rho (ρ) is a nonparametric correlation coefficient that can be used to test for monotonic trends. To calculate the correlation coefficient ρ for any pair of variables x and y, each value of x is replaced with its rank R(x) and each corresponding value of y is replaced with its rank R(y). For concentrations sequentially measured over time (such as those, from a monitoring well), the x variable denotes time and R(x) is the sampling event order (R(x) = 1

for the first sampling event). The rank of the smallest concentration measurement is 1 (when it is not tied with other values).

Spearman's ρ is similar to Pearson's r that is calculated for the paired ranked results (1, $R(y_1)$), (2, $R(y_2)$), … (n, $R(y_n)$) (for instance using Equation 3.5 in Chapter 3.5, Unified Guidance). Like the Pearson's r, Spearman's ρ ranges from -1 to 1 and can be tested to determine whether it is significantly different from zero; a positive value indicates an increasing trend and a negative value indicates a decreasing trend. The absolute value of the coefficient indicates its strength, with larger (absolute) values indicating stronger linear relationships.

When the sample size n is large (n > 20), the test statistic $t = \rho (n- 2)^{½}/(1 - \rho^2)^{½}$ approximately follows the Student's t distribution with n – 2 degree of freedom. To test whether there is a significant trend, the statistic t is compared with upper and lower percentiles of the Student's t distribution. A large value of t (for example, greater than the 95th percentile of the Student's t distribution with n-2 degree of freedom) suggests a significant increasing trend; a negative value (less than the 5th percentile) suggests a decreasing trend. For small sample sizes statistical tables can be used to determine whether ρ is significantly different from zero.

Applications and Relevant Study Questions

- The Spearman correlation coefficient is a common numerical measure of the degree of linear association between two variables.
- Use this test to evaluate stationarity of the mean (the absence of a trend) for parametric data sets, which is a requirement for many statistical methods. A slope differing from zero may indicate the presence of a trend.
- Study Question 5: Is there a trend in contaminant concentrations?

Assumptions

- This test assumes a monotonic relationship between two variables (that is, as one variable increases, the other variable either increases or decreases, but does not fluctuate).
- This test assumes no seasonal trends are present, which generally require more sophisticated evaluations.
- Variables should be identically distributed (but not necessarily independently).

Requirements and Tips

- A minimum of two variables with at least 8 to 10 observations for each variable is recommended. Although it is possible to apply the test with fewer observations, such applications may provide a less meaningful result. A greater number of measurements may be needed if data sets are skewed or contain nondetects.
- The degree of confidence in detecting patterns in the data increases with larger sample sizes.
- Each variable may need to be standardized, for plotting purposes, in order to preserve the scales.
- See Section 5.7 for information regarding the treatment of nondetect data.

- Data should be matched pairs.
- This test does not recognize nonlinear relationships between variables.

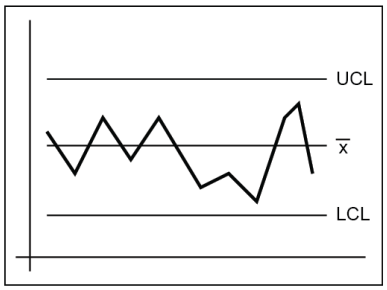
Strengths and Weaknesses

- This test does not require a particular data distribution.
- This test can be used with data sets that contain nondetects. Nondetects result in tied ranks when ρ is calculated.
- This test is not sensitive to outliers.
- This test can be used to detect nonlinear (monotonic) trends.
- Transformation of the data using logarithms (and other monotonic functions) does not alter the value of ρ.

Further Information

A description of how to construct scatter plots is found in Chapter 9.4, Unified Guidance. For all of the cases, the values of each of the variables are ranked from smallest to largest, and the Pearson correlation coefficient is computed on the ranks. Additional information is also available in *Statistical Methods in Water Resources* (Helsel and Hirsch 2002).

5.13 Control Charts



Control charts may be used as an alternative to parametric prediction limits for detection monitoring purposes and are commonly used to monitor the stability of groundwater data and to detect changes in data trends that may require further investigation. Control charts offer an advantage over prediction limits because they generate a graph of compliance data over time and allow for better identification of long-term trends. To generate control charts, a control limit is estimated from background data and subsequently compared to a set of compliance point measurements. A calculated comparison value that exceeds the control limit suggests that compliance point concentrations exceed background. Control charts may be constructed as interwell or intrawell comparisons; background data are collected from upgradient or other background wells for interwell comparisons, and from historical measurements at a targeted compliance well for intrawell comparisons. Baseline parameters (estimates of the mean and standard deviation) are obtained from the background data. As future compliance observations are collected, the baseline parameters are used to standardize the newly gathered data. A new observation is considered "out of control" if it exceeds the baseline control limits, thus indicating a spike or significant change in the trend of the data.

Examples of control chart tests include the Shewart control limit and the cumulative sum control chart (CUSUM). The Shewart control limit tests for and flags a sudden spike or change in trend of the data, which may indicate an event such as a new release at the site. The CUSUM control limit tests for and flags a gradual, but significant, increase or decrease over time, which may, for example, indicate plume migration.

Applications and Relevant Study Questions

- Control charts may be used for a graphical representation of upper and lower prediction limits.
- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 3: Are concentrations above or below a criterion?

Assumptions

- Data must follow a normal distribution or be reasonably symmetric, and also be independent.
- This method assumes a stationary mean for background data (meaning that no trends are present in the data set or subsets of the data).
- Comparison of compliance data against a control limit assumes that the two populations being compared have similar variances. This condition can be assessed using a homogeneity of variance test, but will be difficult to test directly unless you have at least four independent observations from each population (background and compliance).

Requirements and Tips

- Check the data for normality.
- Check the data for a stationary background mean using a test such as the Mann-Kendall trend test.
- Use of a minimum of 8 to 10 measurements to establish background is recommended.
- If you suspect outliers, examine the data using a probability plot, Dixon's test, or Rosner's test, or another appropriate method. Remove outliers from the data set, if appropriate.
- Consider conducting a test for autocorrelation of the background data to ensure that the sampling interval affords uncorrelated measurements.
- Nondetects generally should not exceed 25% of all samples (otherwise, the variance is not adequately defined).
- The Shewhart-CUSUMcontrol chart initially featured two control limits (h=5 and single control limit (SCL) =4.5, see Gibbons 1994). However, later research (Davis 1999) indicated a single control limit h=SCL=4.5 is sufficient and slightly more conservative. You should update the baseline statistics, including the preliminary data set mean and variance, every two years (Gibbons 1994).

Strengths and Weaknesses

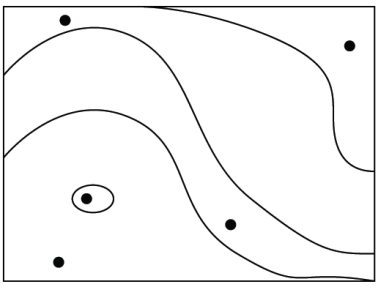
- Control charts can be constructed as either interwell or intrawell tests.
- Compared to trend tests, control charts provide a visual representation of compliance data over time and allow for better identification of gradual, long term trends.

Further Information

Control charts are discussed in Chapter 20, Unified Guidance. The basic procedure for constructing a Shewhart-CUSUMcontrol chart is presented in Chapter 20.2, Unified Guidance. Example 20-1 presents a data set, and Figure 20-2 displays the chart itself with the control limit

plotted as a horizontal line. Gibbons (1994) features combined Shewhart-CUSUM control charts in Chapter 8.4, "Intrawell Comparisons".

5.14 Spatial Statistics



Spatial statistics concerns the analysis of spatially-referenced data, for instance, well locations referenced to a coordinate grid. Originally developed for soil and crop surveys and mining applications, spatial techniques are now regularly used with groundwater data. The primary goal of a spatial statistical analysis is typically map making, which involves analyzing the spatial relationships between locations and utilizing those relationships to create accurate, defensible isocontours or other maps of concentration levels and groundwater elevations. Figure 5-17 illustrates groundwater elevations that are contoured.

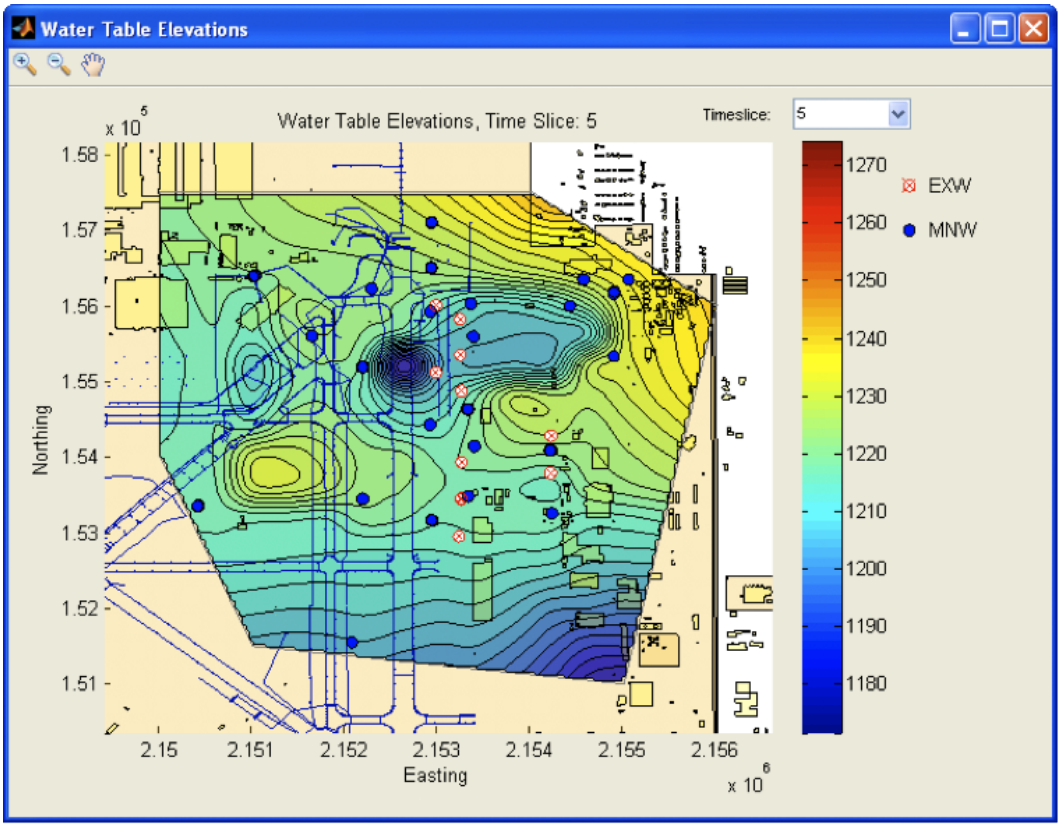


Figure 5-17. Map of contoured groundwater elevations developed in GTS using example data from the software.

Source: Example data courtesy AFCEC 2013.

For many, if not most, spatially-referenced populations — including groundwater aquifers — the possible sample values that might be collected are not evenly 'mixed' throughout the target area.

Instead, there is likely to be a (complex) spatial stratification of the measurement levels, perhaps due to the site topography or hydrogeology, or the existence of localized contaminant sources and groundwater plumes. In addition, wells are almost never randomly located within the target area (Section 3.2.1), but are placed nonrandomly according to the dictates of the Conceptual Site Model (CSM, Section 3.2) and professional judgment.

Because of these realities, spatially-referenced observations are generally not statistically independent. Instead, there is spatial dependency or correlation between different locations, generally meaning that chemical concentrations and/or groundwater elevations at a particular location are more similar to those at nearby locations than they are to more distant sampling points. Spatial statistical techniques are specially adapted to account for this correlation when drawing maps.

5.14.1 Spatial Interpolation and Smoothing

A common task in groundwater spatial analysis is to create a map of either the water table or the concentration isocontours for a particular contaminant. For example, one of the first steps in addressing groundwater contamination is to find the contaminant source(s). One way this can be accomplished is to map the contaminant concentrations in sufficient detail so that zones of lower and higher contamination become evident. The investigator can then 'follow' increasing levels of contamination back to the source(s).

Two general methods exist for statistical mapmaking: interpolation and smoothing. Both methods estimate the value of a variable at any specific location by calculating a weighted average of the (known) values at nearby observed locations (for example, monitoring wells). Both methods are typically used to take the known but scattered sampling locations and make estimates along a regular grid from which contour maps can be generated. The key difference between interpolation and smoothing involves what estimates are made at known locations. Interpolation 'honors' the known sampling points, treating the measured values as fixed and without error (analytical or otherwise). Interpolated maps always equal the observed value at any known sampling point.

Smoothing on the other hand is more akin to a spatial form of regression. When fitting a line to a scatter plot, the regression line may or may not coincide with any single one of the measured values. Instead, the line is chosen to best reflect the overall trend of the points, with the acknowledgment that some of the values may include measurement error or other random variation. The same assumption holds for spatial smoothing, where any single sample value may or may not be 'honored' in the resulting map, depending on what model represents the best overall spatial trend for the site.

Basic interpolation methods include inverse distance weighting, triangular irregular networks (TINs), splines, and nearest neighbor. These methods are simple to use and readily available in software, in part because they do not require estimation of a statistical model of spatial correlation before performing the interpolation. Smoothing methods include smoothing splines and locally-weighted regression, among others. These methods also do not require one to develop a spatial correlation model, though other diagnostics may be needed.

Software packages that are commonly used for spatial visualization of groundwater data using interpolation methods include Surfer (Golden Software 2013), ArcGIS (ESRI 2013), Spatial Analysis and Decision Assistance (SADA, University of Tennessee 2013), Geographic Resources Analysis Support System (GRASS 2013), and Visual Sampling Plan (VSP). Spatial smoothing is less widely available but can be found, for instance, in the Locfit package for R and in the Geostatistical Temporal-Spatial optimization software (GTS).

5.14.2 Kriging

A common but more complex method of spatial interpolation is known as kriging, which was developed in the mining industry. This is the method most popularly associated with the field of geostatistics. In contrast to simpler methods, the weighting of neighboring points in kriging to form estimates at unknown locations is based on a site-specific model of spatial correlation (usually depicted on a variogram) that must be estimated from the data.

Standard mapping software that uses kriging for interpolation may automatically estimate the spatial correlation model from the data without any user intervention. In many cases, the user may not be aware of what model has been selected or how well the selected model fits the data. For simple, approximate mapping purposes, this automatic approach can be acceptable as long as the resulting contour map is checked for consistency with the CSM and the principles of groundwater fate and transport (such as mass conservation). Hand-contoured maps can also be constructed to serve as a check against geostatistical maps.

There is an extensive literature on kriging and many varieties of the technique exist, including, for instance, ordinary kriging (for mapping a single variable), indicator kriging (for mapping probabilities), and co-kriging (for mapping a variable based, in part, on the spatial information provided by another co-sampled variable). There are also many applications of kriging beyond simple mapping of concentrations, including analysis of spatial uncertainty, estimation and mapping of contaminant mass, and monitoring network design and optimization. For these applications, it is usually important to carefully estimate the spatial correlation model (or variogram) and match it against the data. Geostatistical model fitting often requires considerable statistical expertise and is not described in this guidance. See Chiles and Delfiner 1999; Cressie 1993; Gooverts 1997; and Isaaks and Srivastava 1989 for more information.

Although widely used for spatial interpolation, kriging may not always perform significantly better than other simpler techniques. Partly this is due to the fact that it is often difficult to find a good-fitting correlation model, especially if the site does not possess an extensive set of existing sampling locations (such as on a regular grid). In addition, as an interpolation method, kriging is forced to honor the observed data values, even if some of those have been measured with error and are not representative of the underlying trend (for example, think of interpolating between two outliers). The standard kriging algorithms require selecting a single measurement to represent the sampling location in order to compute the interpolation, if data have been gathered over a long period of time this may be difficult. In these situations, a spatial smoothing technique (see Section 5.14.3) such as

spatial regression, which does not require selection of only one value per location, can estimate the broad spatial trend in the underlying data.

Even if data are collected on an equally-spaced regular grid, kriging may not always give satisfactory results if the data are anisotropic. The term anisotropy represents the tendency for the pattern of spatial correlation to be stronger in certain directions than others. The presence of anisotropy impacts how the nearby data should optimally be weighted when making estimates, and can also impact how mapping estimates will turn out. Highly anisotropic data on a regular grid may cause an artificial bunching of the estimated isocontour lines around the grid nodes. Thus, it is important to check for anisotropy prior to spatial mapping. In some cases, simpler techniques can also be adapted to appropriately account for anisotropy, data clustering, and other important spatial features of the data being mapped (Deutch and Journel 1997).

The appeal of kriging is that if (1) the spatial correlation model is correct, and (2) the observed data are without error, it provides optimal estimates compared with other methods, as well as an estimated kriging error at each mapped location. However, the kriging error estimate (or kriging variance) is independent of the measured data values and should not be used as a measure of mapping accuracy. Estimation accuracy can only be assessed via more advanced geostatistical techniques such as cross-validation (see Chiles and Delfiner 1999; Gooverts 1997; and Isaaks and Srivastava 1989 for more information).

5.14.3 Monitoring Network Design and Spatial Optimization

A very powerful application of spatial statistics is in designing and optimizing groundwater monitoring networks. Network design refers to the spatial placement and arrangement of sampling points (e.g., monitoring wells). Spatial optimization involves determining how many sampling points should be used and ensuring they are optimally located. Spatial optimization can be used to answer Study Question 10. For new or relatively new sites, the Conceptual Site Model (CSM), engineering judgment, and the project data quality objectives (DQOs) typically dictate where and how many sampling points are located. For mature sites, two questions can be important: (1) are any of the existing wells statistically redundant and not needed for routine monitoring? and (2) is there a need to locate any additional sampling points, and if so, where?

Several software packages are now widely available for the evaluation and spatial optimization of groundwater monitoring networks. These include:

- 3-Tiered Monitoring and Optimization Tool (3TMO)
- Monitoring and Remediation Optimization Software (MAROS)
- Geostatistical Temporal-Spatial (GTS)
- Visual Sampling Plan (VSP)
- Summit Tools

While each tool includes unique features and often a different approach to network design and optimization, almost all of them make explicit use of spatial mapping. Spatial mapping techniques

— including interpolation and smoothing methods — can be utilized to help answer both of the optimization questions listed above. In the case of the second question, kriging and locally-weighted regression can be used not only to map the site but also to directly estimate the degree of statistical uncertainty associated with locations on the map. By searching for areas of high uncertainty (and perhaps where few wells are already located), if any, the number and locations of new candidate sampling points can be targeted.

In a related fashion, statistical redundancies (first question) in a monitoring network can be identified, leading to fewer sampling points and more cost-effective monitoring. In this case, multiple tools and approaches exist to answer the question. For instance, nearest neighbor estimation can be combined with leave-one-out cross-validation to generate the slope factors in MAROS. Automated kriging can be combined with a genetic algorithm to find the optimal subset of well locations as in Summit Tools. Kriging can also be performed in a stepwise fashion by eliminating at each stage the well with the lowest kriging uncertainty (redundant wells have little uncertainty) as in VSP. Or, locally-weighted regression can be substituted for kriging and an intelligent, quasi-genetic search added to find the optimal networks in GTS.

In general, the optimal design of spatial groundwater monitoring networks is a difficult problem, and has been an active area of research. For smaller monitoring networks, analysis of the spatial design may be possible without the aid of software tools. For larger networks, statistical software tools greatly assist in evaluating the adequacy or redundancy of a monitoring network. Still, none of the widely available statistical approaches attempt to explicitly model complex geology, multiple sources, or multiple groundwater plumes at an individual site. All of the approaches also assume — implicitly or explicitly — that there is a consistent pattern of positive spatial correlation between sampling points. Consequently, the results of any statistically-based network design and optimization should be checked to ensure that the results are appropriate for, and consistent with, the CSM. At sites with complex geology and sources, it may also be necessary to develop a groundwater flow and transport model to assist with network design and optimization.

6.0 DATA MANAGEMENT CONSIDERATIONS

Methods presented in Section 5.0: Statistical Tests and Methods vary in complexity from relatively straightforward graphical methods to complex matrix-based procedures, such as kriging. Correspondingly, the software packages listed in Appendix D range in capabilities from specialized spreadsheet-based groundwater calculators to comprehensive high-powered statistical software suites that are not industry specific. Most of these packages will accept input data from spreadsheets or text files, and many commercial packages are able to connect directly to user databases. Regardless of the system used, input data files should always be provided with statistical analysis deliverables (in electronic format) to allow for verification and cross-checking with different models, as appropriate.

Data management strategies will vary depending on the amount and type of data collected using a systematic planning process, as presented in Section 3.0. For example, a small dry cleaner site may conduct trend analysis on source or boundary wells or both to evaluate concentration changes over time for post-injection monitoring of an in situ bioremediation remedy. Tracking groundwater monitoring data using spreadsheet software may be sufficient for a project of this nature.

For large, complex, multi-source CERCLA sites where there are numerous contaminants and separate monitoring systems a more sophisticated statistical approach may be warranted. With a large data set, preparing groundwater data for statistical analysis can be more time consuming than performing the analysis itself. For these sites, it can be more cost-effect to invest in a more robust data management solution. Commercial environmental data management software is available for this purpose. Comprehensive enterprise-level products developed under direction of the Department of Defense include:

- "The Environmental Restoration Information System (ERIS) is a Web-based database system for the storage of Army environmental restoration and range field data. It serves as a central repository for the Army installation chemical, geological, and geographical data." (US Army 2013a)
- "Environmental Resources Program Information Management System (ERPIMS) is the Air Force system for validation and management of data from environmental projects at all Air Force bases. These data contain analytical chemistry samples, tests, and results, as well as, hydrogeological information, site/location descriptions, and monitoring well characteristics." (USAF 2013).
- "Navy Installation Restoration Information Solution (NIRIS) is a web-based system that manages the Navy's environmental data, documents and records related to cleanup of hazardous waste sites. NIRIS provides the Navy’s remedial project managers (RPMs) and other environmental professionals with tools to effectively analyze, visualize, and present analytical and spatial data." (US Navy 2013b).

Most labs now deliver analytical results electronically and several state and federal organizations have established specific electronic data deliverable (EDD) format requirements. USEPA has developed the staged electronic data deliverable (SEDD) format to support uniform delivery, review, storage, and retrieval of laboratory data. (USEPA 2011a)

However, site data management may be complicated by turnover in site project managers and regulators over the life of a remediation project. Historical data may only be available as hard copy tables, presentation-level crosstab spreadsheets, or in other formats. Cleanup and conversion of legacy data can be very time and labor-intensive, so users must balance level of effort needed to convert data to usable form with the value of data to the statistical approach. See Section 3.3.2: Historical Data for additional discussion on usefulness of historical data for statistical evaluation. If the data set is small, it may be fastest to hand-enter data needed for analysis. Information regarding methods for automated data conversion and cleanup, such as scanning and optical character recognition, are available online

Good Practices for Managing Groundwater Monitoring Data

The general "good practices" listed below will help streamline data analysis and provide a basic structure listing for well construction, analytical results, field data, and geographical coordinates. This structure can be expanded upon as additional data needs are identified. The information presented here is intended as a starting point, you should determine the database formats and information requirements for each project. For more comprehensive data systems, users should follow established data standards such as USEPA's SEDD referenced above.

1. Provide well construction data for each well/monitoring interval in a single row for each well/screen interval:

Well Number	Well Diameter	Total Depth	Top Of Screen	Length Of Screen	Top Of Casing Elevation	Reference Datum

   - Total depth measurements are typically entered as a positive depth value qualified as below ground surface or BGS.
   - For sites with complex geology, parameters such as depth to first water after drilling, depth of drilling fluid circulation loss or other relevant measurements may also be tracked.
2. Provide analytical results and groundwater elevations in a "flattened" format, in which each row of data contains data collected from a single well/screen interval for a single contaminant on a single date. Tabulated analytical results should also include lab analysis qualifiers (such as I, J, and U), practical quantitation limits, and method detection limits to allow flexibility in identifying and managing nondetect values and potential outliers.

Well Number	Sample Date	Contaminant A	Concentration A	Lab Qualifier A	Pql A	Mdl A	Preparation Method	Analytical Method

- Contaminant listings should include a field for Chemical Abstract Service Registry Number (CASRN) or other standardized designation since many chemicals may be identified under multiple names. For example, tetrachloroethene is also known as per-chloroethene, perchloroethylene, Perc, and PCE.
- All numeric results should be formatted to a predetermined precision (number of decimals). For most contaminants whole numbers are adequate, however there are a few where the nth decimal place is the difference between leave and remediate. If this is not set before data collection, columns could be incorrectly formatted and values set to "0" by accident.

3. Present analytical measurements in three columns:
   - One column is for the quantified value for that sample or a reporting limit if the sample is nondetect
   - Another column is for a (possibly numeric, such as 1 for detected and 0 for non-detect) flag signifying the status of that sample (such as detected, trace, nondetect). Standardized lab qualifiers also serve this purpose and can be stored in this column.
   - A third column is for the units of the measurement consistent risk assessment or criteria such as µg/L or mg/L. Use of "parts per million" is not acceptable for groundwater evaluations.

Result	Status	Units
5	0	mg/l

Use this format rather than, for example, "<5", "5J", or a similar notation in the result column because most software will not function properly when numeric values are combined with text or symbols in the same column.

4. Tabulate field sampling results by single well/screen interval and date to verify that samples come from the same target population.

Well Number	Sample Date	Static Depth to Water	pH	Temperature	Conductivity	Dissolved Oxygen	Turbidity	ORP

5. While geospatial analysis is beyond the scope of this guidance, consistently collect and manage geographical coordinates and well survey elevations to simplify groundwater data analysis.

Well Number	Sample Date	Latitude (Decimal Degree Or Degree, Minute, Second)	Longitude (Decimal Degree Or Degree, Minute, Second)	Collection Method	Datum	Verification Method

6. Provide source references (such as lab reports, field notes) for all data stored in the system to verify integrity.

7. Always back up your data.

7.0 PUBLIC AND TRIBAL STAKEHOLDERS PERSPECTIVE

Public stakeholders should include members of the public affected by the site, environmental advocacy group members and community advocacy group members. In addition, tribal stakeholders are defined as Native Americans, Alaska Natives, Native Hawaiians, or persons who are affiliated with or are employees of Native American tribes. These public and tribal stakeholders are the voices of the communities and tribes that are affected by environmental problems and by remediation efforts.

Stakeholder involvement begins with identifying all applicable stakeholders. Stakeholders can be identified by mapping a project's area of influence or impact. This map helps to determine what groups, areas, or activities could be affected by the planned work. In the identification of tribal stakeholders, note that many tribes by treaty have hunting, fishing, and access rights to land that may not be near the present-day reservation. Thus you must look beyond simple reservation boundaries in order to identify tribal stakeholders. Tribes also have sovereignty and must be approached with the proper protocol afforded to a governing body. Stakeholders with an interest in groundwater statistics could be a subset of all stakeholders.

After the applicable stakeholders are identified, several key questions must be answered: the role of each stakeholder group in the environmental site management project, the potential impact each stakeholder group will have on project decisions, when stakeholders will be engaged, how stakeholders will be engaged, and how information will be disseminated. It is best to engage stakeholders early and often. Stakeholder engagement is not unique to groundwater statistical analyses; many regulatory programs have clearly defined checkpoints and procedures, such as the Resource Conservation and Recovery Act (RCRA) permitting process. The RCRA stakeholder identification guidance is described on USEPA's web page (USEPA 1996).

In communicating groundwater statistics to stakeholders, begin by explaining that the overall objective of the cleanup action is to protect human health and the environment. Groundwater statistical analyses are used to better understand the site conditions and inform the cleanup decision making. Statistical methods can be daunting to public and tribal stakeholders and many people are skeptical about the use of statistics to prove any point. A lack of agreement among the experts about statistical methods has created the perception that you can find a statistical argument to support almost any action or conclusion. Straightforward and transparent communication is the best approach. Stakeholders, and some regulatory agencies, may not have access to statistical experts, so present results as clearly as possible and provide opportunities for dialogue to answer questions.

Two general presentations of basic statistical concepts that may be helpful to stakeholders are found on the following web pages:

- NASA's Goddard Institute for Space Studies public educational site
- Texas A&M University's web site Statistics - Introduction to Basic Concepts

This public and tribal stakeholder section serves two purposes: 1) to help the state regulator to understand and anticipate likely issues, needs, and concerns of the stakeholders; and 2) to help make the document more useful to the public and tribal stakeholders. Important general principles that stakeholders should understand are that statistics are only as good as the data set and the data set is only as good as the conceptual site model. Thus, stakeholders must be aware of matters such as sampling well placement and sample validity:

1. Stakeholders should receive explanations of how choices pertaining to seemingly neutral statistical analyses may influence remedial decision-making.
2. A valid conceptual site model is always necessary. If a site is not properly characterized, the statistical results cannot be trusted.
3. Statistical analyses must be performed on a data set of valid size. Consider the following simple example: suppose eight sets of samples are the necessary minimum for valid statistics. If monitoring only occurs once a year, then it will take eight years before it is known whether attenuation is occurring and whether the remediation is occurring. In this situation, the stakeholders might insist on quarterly sampling, in which case it would take two years to learn whether remediation is occurring as projected.
4. To achieve appropriate and accurate statistical analysis, stakeholders should understand that, for data integrity, the monitoring must follow proper quality assurance and quality control procedures during sampling and laboratory analysis. Examples of appropriate procedures may include sample duplicates and blanks.

8.0 SUMMARY AND CONCLUSIONS

This guidance document discusses the statistical techniques and related groundwater monitoring evaluation, optimization, and measurement methods applicable to project life cycle stages. These techniques and methods can help practitioners successfully manage groundwater cleanup, demonstrate resource protection, and fulfill ongoing compliance requirements. Practical statistical tests and methods address typical study questions in each phase of the environmental project life cycle.

This document presents a general statistical approach for various methods (rather than a detailed tutorial; see the Unified Guidance for detailed mathematical approaches). Available tools to conduct statistical tests and methods, examples of appropriate use those tools and interpretation of the results are also presented. Using the project life cycle and study questions, this user-friendly guidance helps practitioners to better understand the application of groundwater statistics. Challenges and misapplications associated with the groundwater statistics are also discussed in this document, and illustrate pitfalls to avoid and the proper approach to correct them if they cannot be avoided. With the appropriate use of groundwater statistics in an environmental project, practitioners and stakeholders can make better project management decisions to achieve the overall goals of protection of human health and the environment.

One guidance document cannot cover all of the information that might be useful for applying statistical analysis to groundwater monitoring. Additional areas for future guidance may include the following:

- Geostatistical analyses and tools for groundwater data evaluation have been introduced here. Specific guidance on how geostatistical analyses can be used in groundwater compliance monitoring and long-term stewardship was beyond the scope of this document.
- Temporal and spatial optimization of groundwater monitoring well networks is included here in brief. Additional detailed guidance would be needed for state regulators.
- Groundwater monitoring programs are important to green and sustainable remediation activities. Additional practical guidance would be needed.
- Reviews of currently available software tools applicable to environmental projects are provided in Appendix D; you should always check for newer versions of any tools to be used at a specific project. Future updates of the software programs and tools information may become available.

9.0 REFERENCES

A

Abernethy, R. 2010. *The New Weibull Handbook*. 5th ed. Humble, Texas: Barringer & Associates, Inc.

Ahmadi, S.H., and A. Sedhamiz. 2007. "Geostatistical Analysis of Spatial and Temporal Variations of Groundwater Level." *Environmental Monitoring and Assessment* no. 129 (1-3):277-294.

Air Force Center for Environmental Excellence (AFCEE). 1997. "AFCEE Long-Term Monitoring Optimization Guide Version 1.1." Brooks AFB, Brooks City, TX: Air Force Center for Environmental Excellence.

Air Force Civil Engineer Center (AFCEC). 2012. "Monitoring and Remediation Optimization System (MAROS) Software, User's Guide and Technical Manual." In: Air Force Center for Environmental Excellence. http://www.gsi-net.com/en/software/free-software/maros-30.html.

ASTM. 2010a. Standard Guide for Optimization of Ground Water Monitoring Constituents for Detection Monitoring Programs for RCRA Waste Disposal Facilities. D7045-04(2010). West Conshohocken, PA: ASTM International.

ASTM. 2010b. Standard Guide for Applying Statistical Methods for Assessment and Corrective Action Environmental Monitoring Programs. D7048-04(2010). West Conshohocken, PA: ASTM International.

ASTM. 2012. Standard Guide for Developing Appropriate Statistical Approaches for Groundwater Detection Monitoring Programs. D6312-98(2012)e1. West Conshohocken, PA: ASTM International.

Aziz, J.A., C.J. Newell, M. Ling, H.S. Rifai, and J.R. Gonzales. 2003. "MAROS: A Decision Support System for Optimizing Monitoring Plans." *Ground Water* no. 41 (3):355-367.

B

Beal. D. 2009. "A Macro for Calculating Summary Statistics on Left Censored Environmental Data using the Kaplan-Meier Method." SDA-09. http://analytics.ncsu.edu/sesug/2010/SDA09.Beal.pdf.

Bowman, A.W., and A. Azzalini. 2012. The "sm" package for R. Smoothing methods for non-parametric regression and density estimation. www.stats.gla.ac.uk/~adrian/sm.

Bowman, A.W., and A. Azzalini. 1997. Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations. Oxford, England: Oxford University Press.

Box, G., and G. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. Revised, 2nd ed. San Francisco: Holden-Day.

Bunn, A.L., D.M. Wellman, R.A. Deeb, E.L. Hawley, M.J. Truex, M Peterson, M.D. Freshley, E.M. Pierce, J. McCord, M.H. Young, T.J. Gilmore, R. Miller, A.L. Miracle, D. Kaback, C. Eddy-Dilek, J. Rossabi, M.H. Lee, R.P. Bush, P. Beam, G.M. Chamberlain, J. Marble,

L. Whitehurst, K.D. Gerdes, and Y. Collazo. 2012. "Scientific Opportunities for Monitoring at Environmental Remediation Sites (SOMERS): Integrated Systems-Based Approaches to Monitoring." DOE/PNNL-21379. Richland WA: Prepared for Office of Soil and Groundwater Remediation, Office of Environmental Management, U.S. Department of Energy, by Pacific Northwest National Laboratory. www.pnnl.gov-/main/publications/external/technical_reports/PNNL-21379.pdf.

Burns Statistics. 2012. Spreadsheet Addiction. http://www.burns-stat.-com/documents/tutorials/spreadsheet-addiction/.

Burt, J.E., G.M. Barber, and D.L. Rigby. 2009. *Elementary Statistics for Geographers*. 3rd ed. Guildord Press.

C

Cameron, K.P. 2004. "“Better optimization of LTM networks.” *Bioremediation Journal* no. 8 (03-04):89-108.

Cameron, K.P, and P. Hunter. 2004. "Optimizing LTM networks with GTS: three new case studies." In Conference on Accelerating Site Closeout, Improving Performance, & Reducing Costs Through Optimization. Dallas, TX.

Cameron, K.P, and P. Hunter. 2000. "Optimization of LTM networks: statistical approaches to spatial and temporal redundancy." In Spring Natl. Meeting of American Institute of Chemical Engineers. Atlanta, GA.

Cameron, K.P, and P. Hunter. 2002. "Using spatial models and kriging techniques to optimize long-term ground-water monitoring networks: a case study." *Environmetrics* no. 13:629-656.

Cameron, K.P, and P. Hunter. 2003. "Optimization of LTM networks at AF Plant 6 using GTS. In In Situ and On-Site Bioremediation – 2003." Proceedings of the Seventh International In Situ and On-Site Bioremediation Symposium, edited by V.S. Magar and M.E. Kelley. Orlando, FL: Battelle Press.

Cameron, K.P, P. Hunter, and R. Stewart. 2011. "Demonstration and validation of GTS long-term monitoring optimization software at military and government sites." ESTCP Project ER-200714. www.serdp.org.

Chatfield, C. 2004. *The Analysis of Time Series: An Introduction*. 6th ed. Boca Raton, FL: Chapman and Hall.

Chiles, J.P., and P. Delfiner. 1999. *Geostatistics, Modeling Spatial Uncertainty*. Wiley Series in Probability and Statistics: Wiley.

Cressie, N.A.C. 1993. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. Wiley-Interscience.

D

Davis, C.B. 1994. "Environmental Regulatory Statistics." In *Handbook of Statistics*, Volume 12: Environmental Statistics, edited by GP Patil and CR Rao. New York: Elsevier Science B.V.

Davis, C.B. 1999. EnviroStat Technical Report 99-1: Comparisons of Control Chart and Prediction Limit Procedures Recommended for Groundwater Detection Monitoring. Henderson, NV.

Deutch, C. V. and A. G. Journel. 1997. *Geostatistical Software Library and User's Guide.* Applied Geostatistics Series. New York: Oxford Univ. Press.

Discerning Systems, Inc. 2012. CARStat Software. www.DiscerningSystems.com/carstat.html.

Discerning Systems, Inc. 2012. DUMPStat software. http://www.discerningsystems.com/dumpstat.html.

DOD (United States Department of Defense). 2013. *Quality Systems Manual (QSM) for Environmental Laboratories*, Based on ISO/IEC 17025:2005(E) and The NELAC Institute (TNI) Standards, Volume 1, (September 2009). Version 5: Department of Defense Environmental Data Quality Workgroup (EDQW). http://www.denix.osd.mil/edqw/upload/QSM-Version-5-0-FINAL.pdf.

E

Engle, R. 2001. "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics." Journal of Economic Perspectives no. 15 (4):157-168. doi: 10.1257/jep.15.4.157.

ESRI. *ArcGIS Software* 2013. http://www.esri.com/software/arcgis.

F

Faraway, J. 2002. "Practical Regression and ANOVA using R." http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Feenstra, S., J.A. Cherry, and B.L. Parker. 1996. "Conceptual Models for the Behavior of DNAPLs in the Subsurface." In *Dense Chlorinated Solvents and Other DNAPLs in Groundwater*. Pankow, J.F. and J.A. Cherry, Eds. Portland OR: Waterloo Press.

Forster, M.R. 2013. Extrapolation Error. http://-philosophy.wisc.edu/forster/papers/extrapolation.htm.

G

Gehan, E.A. 1965. "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples." *Biometrika* no. 52:203-223.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York: John Wiley & Sons.

Gibbons, R.D., D.K. Bhaumik, and S. Aryal. 2009. *Statistical Methods for Groundwater Monitoring*. 2nd ed, Statistics in Practice. New York: John Wiley & Sons.

Gibbons, R.D., and D.E. Coleman. 2001. *Statistical Methods for Detection and Quantification of Environmental Contamination*. New York: John Wiley & Sons.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York: John Wiley & Sons, Inc.

Golden Software, Inc. 2013. Surfer 11 2013. http://www.goldensoftware.com/products/surfer.

Goldwater, A. 2007. "Using Excel for Statistical Data Analysis - Caveats." In: Biostatistics Consulting Center, University of Massachusetts School of Public Health.

Gooverts, P. 1997. *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

GRASS Team GIS. 2013. *GRASS GIS software*. http://grass.osgeo.org.

H

Heilberger, R.M., and E. Neuwirth. 2009. *R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis and Graphics*. London: Springer Dordrecht Heidelberg.

Helsel, D.R. 2005. *Nondetects and Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Helsel, D.R. 2012. *Statistics for Censored Environmental Data Using Minitab and R*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons.

Helsel, D.R., and R.M. Hirsch. 2002. "Statistical Methods in Water Resources." In Book 4, Hydrologic Analysis and Interpretation, 522. United States Geological Survey.

Hornik, K. The R FAQ 2013. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

I

IBM. 2013. SPSS Statistics. IBM. http://www-01.ibm.com/software/analytics/spss/products/statistics/.

Isaaks, E. H., and R. M. Srivastava. 1989. *An Introduction to Applied Geostatistics*. New York: Oxford University Press.

ITRC (Interstate Technology & Regulatory Council). 2003. *Technical and Regulatory Guidance for the Triad Approach*. SCM-1. Washington, D.C.: Interstate Technology & Regulatory Council, Sampling, Characterization, and Monitoring Team. http://www.itrcweb.org/Documents/SCM-1.pdf.

ITRC. 2007a. *Triad Implementation Guide*. SCM-3. Washington, D.C.: Interstate Technology & Regulatory Council, Sampling, Characterization, and Monitoring Team. http://itrcweb.org/Guidance/ListDocuments?TopicID=27&SubTopicID=42#.

ITRC. 2007b. *Improving Environmental Site Remediation Through Performance-Based Environmental Management*. RPO-7. Washington, D.C.: Interstate Technology & Regulatory Council, Remediation Process Optimization Team. http://www.itrcweb.org/Guidance/ViewTopic?topicID=22.

ITRC. 2010. *Use and Measurement of Mass Flux and Mass Discharge*. MASSFLUX-1. Washington, D.C.: Interstate Technology & Regulatory Council, Integrated DNAPL Site Strategy Team. http://www.itrcweb.org/Guidance/ListDocuments?topicID=14&subTopicID=11.

ITRC. 2011a. *Green and Sustainable Remediation: A Practical Framework*. GSR-2. Washington, D.C.: Interstate Technology & Regulatory Council, Green and Sustainable Remediation Team. http://www.itrcweb.org/Guidance/ListDocuments?TopicID=9&SubTopicID=15.

ITRC. 2011b. *Project Risk Management for Site Remediation*. RRM-1. Washington, D.C.: Interstate Technology & Regulatory Council, Remediation Risk Management Team. http://www.itrcweb.org/Guidance/ListDocuments?TopicID=23&SubTopicID=32.

ITRC. 2012. *Incremental Sampling Methodology*. ISM-1. Washington, D.C.: Interstate Technology & Regulatory Council, Incremental Sampling Methodology Team. http://www.itrcweb.org/Guidance/ViewTopic?topicID=11.

ITRC. 2013. *Environmental Molecular Diagnostics, New Tools for Better Decisions*. EMD-2. Washington, D.C.: Interstate Technology & Regulatory Council, Environmental Molecular Diagnostics Team. http://www.itrcweb.org/Guidance/ListDocuments?TopicID=33&SubTopicID=14.

J

Jones, W.R., and M. Spence. 2012. "GroundWater Spatio-Temporal Data Analysis Tool (GWSDAT Version 2.0) User Manual." UK: Shell Global Solutions.

K

Kleinbaum, D.G., L.L. Kleinbaum, A. Nizam, and K.E. Muller. 2007. *Applied Regression Analysis and Multivariable Method*s - 4th Edition: Duxbury Press.

Krishnan, T., and R. Karandikar. "SYSTAT Tutorial." In. Bangalore, India: Cranes Software International Limited.

L

Ling, M., H.S. Rifai, J.A. Aziz, C.J. Newell, J.R. Gonzales, and J.M. Santillan. 2004a. "Strategies and Decision-Support Tools for optimizing Long-Term Groundwater Monitoring Plans-MAROS 2.0." *Bioremediation Journal* no. 8 (3-4):109-128.

Ling, M., H.S. Rifai, and C.J. Newell. 2005. "Optimizing Long-Term Monitoring Networks Using Delaunay Triangulation Spatial Analysis Techniques." *Environmetrics* no. 16 (6):635-657.

Ling, M., H.S. Rifai, C.J. Newell, J.A. Aziz, and J.R. Gonzales. 2003. "Groundwater Monitoring Plans at Small-Scale Sites: An Innovative Spatial and Temporal Methodology." *Journal of Environmental Monitoring* no. 5:126-134.

Henlopen Design, LLC. 2012. What is PAM? 20132012. http://www.henlopen.net/pam/aboutpam.htm.

M

Matzke, B.D., J.E. Wilson, L.L. Nuffer, S.T. Dowson, J.E. Hathaway, N.L. Hassig, L.H. Sego, C.J. Murray, B.A. Pulsipher, B. Roberts, and S. McKenna. 2010. "Visual Sample Plan Version 6.0 User's Guide." PNNL-19915. Richland, WA: Pacific Northwest National Laboratory.

McCullough, B.D, and D.A. Heiser. 2008. "On the Accuracy of Statistical Procedures in Microsoft Excel 2007." Computational Statistics and Data Analysis no. 52:4570-4578.

McHugh, T., L.M. Beckley, C.Y. Liu, and Newell C.J. 2011. "Factors Influencing Variability in Groundwater Monitoring Data Sets." *Ground Water Monitoring & Remediation* no. 31 (2):92-101.

McNichols, R.J., and C.B. Davis. 1988. "Statistical issues and problems in ground water detection monitoring at hazardous waste facilities." *Ground Water Monitoring Review* no. 8:135-150.

Microsoft, Inc. Excel Microsoft Store. 2012. http://www.microsoftstore.com/store/msusa/en_US/home.

Minitab, Inc. 2012. Minitab Software for Quality Improvement. http://www.minitab.com/en-US/default.aspx?langType=1033.

Myers, J.C. 1997. *Geostatistical Error Management*. New York: Van Nostrand Reinhold.

National Aeronautics and Space Administration (NASA). Goddard Institute for Space Studies. 2013. Public Education Site - Statistics. http://icp.giss.nasa.gov/education/statistics/.

N

National Academies Press (NAP). 2012. *Alternatives for Managing the Nation's Complex Contaminated Groundwater Sites*. National Research Council.

NCSS. 2012. NCSS Statistical Software 20132012. http://www.ncss.com/about/.

Newell, C.J., H.S. Rifai, J.T. Wilson, J.A. Connor, J.A. Aziz, and M.P. Suarez. 2002. "Groundwater Issue: Calculation and Use of First-Order Rate Constants for Monitored Natural Attenuation Studies " EPA/540/S-02/500. Washington, DC: United States Environmental Protection Agency. http://itrcweb.org/FileCabinet/GetFile?fileID=6932.

NIST/SEMATECH. 2012. "e-Handbook of Statistical Methods." In. http://www.itl.nist.gov/div898/handbook/ (accessed August 2013).

Nobel, C, and J.A. Anthony. 2004. "Three-Tiered Approach to Long-Term Monitoring Program Optimization." Bioremediation Journal no. 8 (3-4):147-165.

Nuffer, L.L., N.L. Hassig, L.H. Sego, B.A. Pulsipher, J.E. Wilson, and B.D. Matzke. 2009. "Validation of Statistical Sampling Algorithms in Visual Sample Plan (VSP): Summary Report." PNNL-18253. Richland, WA: Pacific Northwest National Laboratory.

P

Pacific Northwest National Laboratory. 2012. Visual Sample Plan. Pacific Northwest National Laboratory. http://vsp.pnnl.gov.

Parker, L., and S. Britt. 2012. "The Effect of Bottle Fill Rate and Pour Technique on the Recovery of Volatile Organics." *Groundwater Monitoring & Remediation* no. 32 (4):78-86.

Paul, H., C. Eilers, D.M. Rijnmond, and B.D. Marx. 1996. "Flexible smoothing with b-splines and penalties." *Statistical Science* no. 11:89-121.

Pennsylvania State University (PSU). 2013. STAT 510 - Applied Time Series Analysis http://onlinecourses.science.psu.edu/stat510.

Practical Stats, 2013. Is Microsoft Excel an Adequate Statistics Package? http://www.-practicalstats.com/xlsstats/excelstats.html.

R

Ridley, M.N., V.M. Johnson, and R.C. Tuckfield. 1995. Cost-Effective Sampling of Groundwater Monitoring Wells. Vol. UCRL-JC-118909. Livermore, CA: Lawrence Livermore National Laboratory.

Ridley, M.N., and D. MacQueen. 2005. A Cost-Effective Sampling of Groundwater Monitoring Wells: A Data Review and Well Frequency Evaluation. UCRL-CONF-209770. Livermore CA: Lawrence Livermore National Laboratory. http://www-erd.llnl.gov/library/CONF-209770.pdf.

Rong, Y. 2011. "Statistical Methods and Pitfalls in Environmental Data Analysis." In *Practical Environmental Statistics and Data Analysis*, edited by Y. Rong, 243-258. ILM Publications.

S

Sall, J. 2007. JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP. 4th ed: SAS Press.

Sanitas Technologies. 2012. Sanitas Statistical Software. http://www.sanitastech.com.

SAS Institute, Inc. 2012. JMP Software 20132012. www.jmp.com.

SAS Institute, Inc. 2012. Statistical Analysis with SAS/STAT Software. http://www.-sas.com/technologies/analytics/statistics/stat/index.html.

Science-dictionary.org. 2013. Science Dictionary. Babylon 9 2008. Accessed 9/13/13. http://mathematicsandstatistics.science-dictionary.org.

Siegel, D. 2008. "Reductionist Hydrogeology: Ten Fundamental Principles." *Hydrological Processes* no. 22:4967-4970.

Silva, E.L., and P. Lisboa. 2007. Analysis of the characteristic features of the density functions for gamma, Weibull and log-normal distributions through RBF network pruning with QLP. In Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases. Corfu Island, Greece.

Singh, A., R. Miachle, and S. Lee. 2006. "On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations." EPA/600/R-06/022: United States Environmental Protection Agency.

Singh, A., and J. Nocerino. 1997. "Robust Intervals for Some Environmental Applications." The Journal of Chemometrics and Intelligent Laboratory Systems no. 37:55-69.

Singh, A., A. K. Singh, and R.J. Iaci. 2002. "Estimation of the Exposure Point Concentration Term Using a Gamma Distribution." EPA/600/R-02/084: United States Environmental Protection Agency.

Singh, A. K., A. Singh, and M. Englehart. 1997. The Lognormal Distribution in Environmental Applications. Technology Support Center Issue Paper.

Stark, P.B. 2013. Glossary of Statistical Terms. University of California, Berkeley. http://www.stat.berkeley.edu/~stark/SticiGui/Text/gloss.htm.

Starpoint Software. 2012. ChemStat Software. www.pointstar.com.

StatSoft, Inc. 2013. Statistica. StatSoft. www.statsoft.com.

Systat Software. 2012. www.systat.com.

SurveyMonkey, Inc. 2011. www.surveymonkey.com.

T

Texas A&M University. *Statistics - Introduction to Basic Concepts*. http://-bobhall.tamu.edu/FiniteMath/Module8/Introduction.html.

The MathWorks, Inc. MATLAB, The Language of Technical Computing 2013. http://www.math-works.com/products/matlab/.

The R Project. 2013. The R Project for Statistical Computing http://www.r-project.org.

Thomas, L., and C. Drebs. 1997. "A Review of Statistical Power Analysis Software." *Bulletin of the Ecological Society of America* no. 78 (2):128-139.

Thyne, G., C. Guler, and E. Peoter. 2004. "Sequential Analysis of Hydrochemical Data for Watershed Characterization." *Ground Water* no. 42 (42):711-723.

Thorbjornsen, K., and J. Myers. 2007. "Identification of Metals Contamination in Firing-Range Soil Using Geochemical Correlation Evaluation." Soil & Sediment Contamination no. 16:337-349. doi: 10.1080/15320380701404391.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

U

United States Army Corp of Engineers (USACE). 1998. "Environmental Quality, Technical Project Planning (TPP) Process." EM 200-1-2. Washington, D.C.: Department of the Army. www.usace.army.mil/missions/environmental/technicalprojectplanning.aspx.

United States Air Force. (USAF). 2013. Environmental Resources Program Information Management System (ERPIMS). http://www.afcec.af.mil/resources/restoration/erpims/index.asp.

United States Army (US Army). 2013a. Environmental Restoration Information System (ERIS). http://aec.army.mil/Portals/3/reporting/eris00.html.

US Army 2013b. Environmental Statistics, EM 200-1-16: United States Army Corps of Engineers. http://www.publications.usace.army.mil/Portals/76/Publications/EngineerManuals/EM_200-1-16.pdf.

U.S. National Archives and Records Administration. Code of Federal Regulations. Title 40 Parts 258, 264 and 265.

United States Environmental Protection Agency. 1988. "Guidance for Conducting Remedial Investigations and Feasibility Studies Under CERCLA." EPA/540/G-89/004, OSWER Directive 9355.3-01. Washington, DC: United States Environmental Protection Agency, Office of Emergency and Remedial Response. http://rais.ornl.gov-/documents/GUIDANCE.PDF.

USEPA (United States Environmental Protection Agency). 1989. "Methods for Evaluating the Attainment of Cleanup Standards, Vol. 1 Soils and Solid Media." EPA 230/02-89-042 United States Environmental Protection Agency.

USEPA 1996. Wastes - Hazardous Waste - Treatment, Storage & Disposal (TSD), RCRA Public Participation Manual.http://www2.epa.gov/sites/production/files/2015-08/documents/rcra_pub_participtn_man.pdf.

USEPA. 1999. "Robust Statistical Intervals for Performance Evaluations." In. Las Vegas, NV: Office of Research and Development.

USEPA. 2002a. "Guidance on Choosing a Sampling Design for Environmental Data." EPA QA/G-5S. Washington D.C.: United States Environmental Protection Agency. http://www2.epa.gov/quality/guidance-choosing-sampling-design-environmental-data-collection-use-developing-quality.

USEPA. 2002b. Guidance for Quality Assurance Project Plans. EPA QA/G-5. EPA/240/R-02/009. Washington D.C.: United States Environmental Protection Agency. http://www2.epa.gov/quality/guidance-quality-assurance-project-plans-epa-qag-5-december-2002.

USEPA. 2002c. "Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites." EPA 540-R-01-003. Office of Emergency and Remedial Response. Washington, D.C. http://itrcweb.org/FileCabinet/GetFile?fileID=6931.

USEPA. 2005. "Uniform Federal Policy for Quality Assurance Project Plans. UFP-QAPP. EPA/505/B-04/900A.Washington D.C.: United States Environmental Protection Agency.

USEPA. 2006a. "Guidance on Systematic Planning Using the Data Quality Objectives Process." EPA QA/G-4. Washington D.C.: United States Environmental Protection Agency. http://www2.epa.gov/quality/guidance-systematic-planning-using-data-quality-objectives-process-epa-qag-4.

USEPA. 2006b. "Data Quality Assessment: A Reviewer's Guide." EPA QA/G-9R. Washington D.C.: United States Environmental Protection Agency.

USEPA. 2006c. "Data Quality Assessment: Statistical Methods for Practitioners." EPA/240/B-06/003 EPA QA/G-9S. Washington D.C.: United States Environmental Protection Agency. http://www2.epa.gov/quality/guidance-data-quality-assessment.

USEPA. 2007. "Optimization Strategies for Long-Term Ground Water Remedies (with Particular Emphasis on Pump and Treat Systems)." EPA 542-R-07-007. Washington, DC: United States Environmental Protection Agency. http://www2.epa.gov/remedytech/optimization-strategies-long-term-ground-water-remedies-particular-emphasis-pump-and

USEPA. 2008a. "USEPA Contract Laboratory Program National Functional Guidelines." OSWER 9240.1-48, USEPA-540-R-08-01.

USEPA. 2008b. "Guidance on Environmental Data Verification and Data Validation." EPA QA/G-8. Washington D.C.: United States Environmental Protection Agency.

USEPA. 2009. "Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities." Unified Guidance EPA 530/R-09-007. Washington DC: United States Environmental Protection Agency. http://www3.epa.gov-/epawaste/hazard/correctiveaction/resources/guidance/sitechar/gwstats/unified-guid.p-dfguid.pdf.

USEPA. 2010. "ProUCL Version 4.1.00 Technical Guide (Draft)." EPA/600/R-07/041. Washington, DC: United States Environmental Protection Agency. http://www2.epa.gov/land-research/proucl-version-4100-documentation-downloads.

USEPA 2011a. The Staged Electronic Data Deliverable (SEDD), November 10, 2011. Accessed 2013. http://www2.epa.gov/clp/staged-electronic-data-deliverable-sedd.

USEPA. 2011b. "An Approach for Evaluating the Progress of Natural Attenuation in Groundwater." EPA 600/R-11/204. Washington, DC. http://nepis.epa.gov-/Adobe/PDF/P100DPOE.pdf.

USEPA. 2012. "National Strategy to Expand Superfund Optimization Practices from Site Assessment to Site Completion." OSWER 9200.3-75. Washington, DC: United States Environmental Protection Agency. http://www2.epa.gov/superfund/cleanup-optimization-superfund-sites.

USEPA. 2013a. Abstract Scout, Scout 2008 Version 1.00.01. http://www2.epa.gov-/aboutepa/about-national-exposure-research-laboratory-nerl.

USEPA. 2013c. Pro-UCL Software. http://www.epa.gov/land-research/proucl-software.

USEPA. 2013d. ProUCL Software Documentation. http://www.epa.gov/land-research/proucl-software.

USEPA. 2013e. RCRA Public Participation Manual. http://www3.epa.gov-/epawaste/hazard/tsd/permit/pubpart/index.htm.

USEPA. 2013f. Remedy Optimization. Accessed October 30, 2013. http://www.epa.gov-/superfund/cleanup-optimization-superfund-sites.

USEPA. 2015. EPA Terminology Services. Search. Terms and Acronyms Web Page. http://i-aspub.epa.gov/sor_internet/registry/termreg/searchandretrieve/termsandacronyms/search.do.

USGS (United States Geological Survey). 2013. Essential Components of Water-Level Monitoring Programs. Circular 1217.http://pubs.usgs.gov/circ/circ1217/html/essential.html.

US Navy (United States Navy). 2013a. Optimization Roadmap. Navy Facilities Command (NAVFAC). Accessed July 2013. http://www.navfac.navy.mil/navfac_worldwide/specialty_centers/exwc/products_and_services/ev/erb/opt.html

US Navy. 2013b. Navy Installation Restoration Information Solution (NIRIS). Navy Facilities Command (NAVFAC)http://www.navfac.navy.mil/navfac_worldwide/specialty_centers/exwc/products_and_services/ev/erb/niris.html.

University of Tennessee, Spatial Analysis and Decision Assistance (SADA) 2013. http://www.sadaproject.net.

V

Vanderford, M. 2010. "A Comprehensive Approach to Plume Stability." *Remediation* no. Winter 2010:21-37.

Venables, W.N., D.M. Smith, and R Core Team. 2013. "An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics." Version 3.0.1. http://cran.r-project.org/doc/manuals/R-intro.pdf.

W

Wilkinson, L., G. Blank, and C.G. Gruber. 1996. *Desktop Data Analysis with SYSTAT*. Englewood Cliffs, NJ: Prentice-Hall.

Y

Yaglom, A.M. 1962. *An Introduction to the Theory of Stationary Random Function*. Englewood Cliffs, N.J: Prentice Hall. 235 pp.

Yalta, A.T. 2008. "The Accuracy of Statistical Distributions in Microsoft Excel 2007." *Computational Statistics and Data Analysis* no. 52:4579-4586.

APPENDIX A. CASE EXAMPLES

This appendix includes examples of statistical techniques applied to realistic groundwater data sets:

A.1 Comparing Two Data Sets Using Two-sample Testing Methods

You wish to compare copper concentrations at a compliance point well with background levels at an upgradient well at a Resource Conservation and Recovery Act (RCRA) site that has recently undergone remediation. The data sets are shown in Table A-1 below:

Table A-1 Two-Sample Test Data

Table A-1. Two-sample test data

Date	Compliance Well (µg/L)	Background Well (µg/L)
03/25/03	4.1	2
06/16/03	14	1.2
09/17/03	3.4	<1.0
12/09/03	3.8	<1.0
03/17/04	5.2	<1.0
06/15/04	11	1
09/14/04	15	1.2
11/09/04	10	<1.0
02/23/05	11	<1.0
05/24/05	9.4	0.6
08/31/05	9.7	0.6
11/29/05	12	0.6
03/01/06	18	0.65 J
05/24/06	21	0.73 U
08/15/06	12	1.1
10/12/06	15	0.9
01/24/07	12	1.4
04/18/07	9.7	1.4
07/26/07	7.5	0.92
10/25/07	8.8	0.84
01/23/08	6.2	<1.0

Table A-1. Two-sample test data (continued)

Date	Compliance Well (µg/L)	Background Well (µg/L)
04/17/08	7.9	1.3
07/17/08	10	0.63 J
10/16/08	7.4	0.77 U
01/15/09	2.5	<1.0
04/09/09	2.6	<1.0
07/17/09	1.1	1
10/15/09	1.1	1.2
02/12/10	1.5	<1.0
05/14/10	1.4	<1.0
08/13/10	1.1	1.7
12/09/10	1	1.4
02/03/11	3.4	<1.0
04/07/11	0.38	1.3
08/05/11	1.2	1.7
10/14/11	0.67	1.4
02/03/12	0.62	<1.0

A.1.1 Overview

Statistical tests for comparing two groups of data are known as two-sample tests (see Section 5.11). A two-sample test is a type of hypothesis test typically used to evaluate whether the mean or median concentrations of two sample populations are equal or one is greater or less than the other.

In this example, the site was contaminated in the past and now the statistical evidence is evaluated to determine if the corrective action was effective at cleaning up the site (for example, compliance well concentrations are consistent with background concentrations). Thus, the null hypothesis is that the mean or median concentration at the compliance point is less than or equal to the mean or median concentration at the background well. The alternative hypothesis is that the compliance concentration is greater than background. The hypothesis test will be used to determine the strength of the statistical evidence that the null hypothesis is incorrect. In other words, is there statistical evidence that the compliance well concentrations are not consistent with background concentrations? Depending on the purpose of the statistical testing, a different null hypothesis may be appropriate.

There are several types of two-sample tests that can be used to compare the mean, or another population parameter of interest, of two data sets. For sites in which a small number of wells and chemicals are compared, the Unified Guidance states that the following tests may be appropriate:

Parametric Student t-tests

- Pooled variance t-test (equal variance)
- Welch's t-test (unequal variance)

Nonparametric tests

- Wilcoxon rank sum test (also known as Mann-Whitney test)
- Tarone-Ware test (or Gehan's test)

The first two tests are classic parametric Student t-tests, while the latter two tests are nonparametric and are recommended when the assumptions of the parametric tests are violated. The Tarone-Ware Test and Gehan's Test are generalizations of the Wilcoxon rank sum test that are best for highly censored data sets. It would also be possible to use prediction limits to compare the two groups, but this is a more advanced approach which is not considered in this example.

A.1.2 Choose a statistical test

The most commonly used two-sample tests are Student t-tests. The pooled variance t-test is used when the two data sets have equal variance. The Welch's t-test is a modified version of the t-test used to compare data sets with unequal variances. You can visually compare the results of t-tests with side-by-side box plots. Since data sets will commonly have unequal variances, it may make sense to use the Welch's t-test as it has similar power as the pooled variance test without the requirement of equal variances.

In general, consider the use of a nonparametric two-sample test if any of the following key assumptions are violated:

- The population means are steady over time.
- Data are approximately normally distributed.
- Minimum sample size of six to eight measurements per group

The first key assumption is that the population mean is stable or steady over space and time, a property often referred to as stationarity. This assumption means that a two-sample t-test cannot be performed on a data set with a trend. The Unified Guidance recommends either performing a formal trend test at the compliance well or limiting the compliance point data set to values that are representative of current conditions. The plot of the data from Table A-1 (see Figure A-1) shows a downward trend in copper concentrations at the compliance well from 2006 to present.

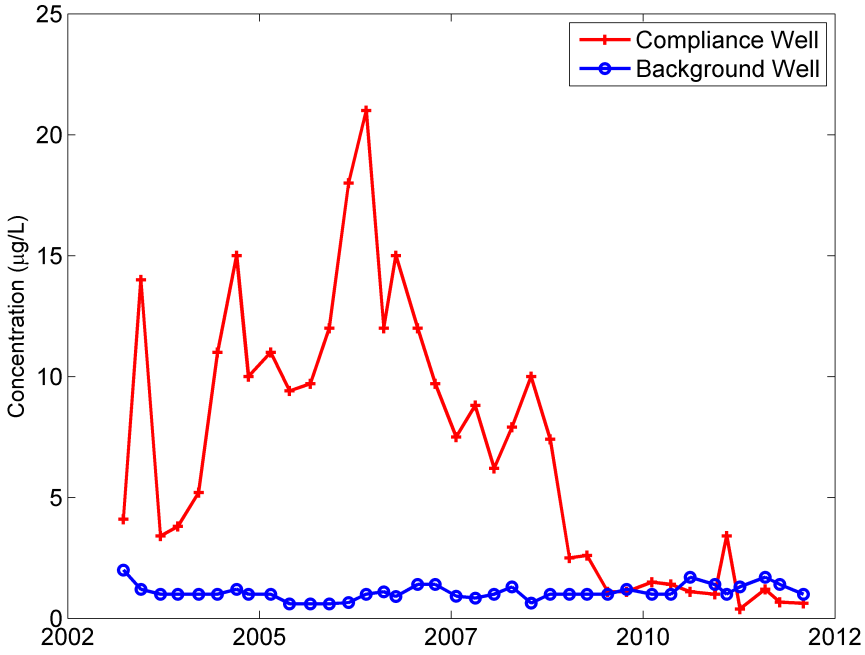


Figure A-1. Historical copper concentrations at Remediation Site A.

By visual inspection of Figure A-1, the most recent nine measurements do not have a significant trend and could be selected to represent post-remediation conditions. Since t-tests compare the mean concentrations of two data sets, using the entire compliance point data set in this case might erroneously yield the result that the compliance point concentration is greater than background.

Normality is also a key assumption of t-tests. The t-test is considered reasonably robust to this assumption, meaning that the results may still be accurate even if the normal assumption is not completely met. However, if the data set is highly skewed (coefficient of variation greater than 1.5), the test may give misleading results. For data sets that are lognormal, the data should be first log-transformed before the t-test is applied. When the variances are equal, normality can be checked by combining the residuals for both data sets on a normal probability plot. Otherwise, check normality using graphical methods or by using a multiple group test of normality like the Shapiro-Wilk test. Normality in this example is checked for both of the data sets with histograms and normal probability plots as shown in Figure A-2.

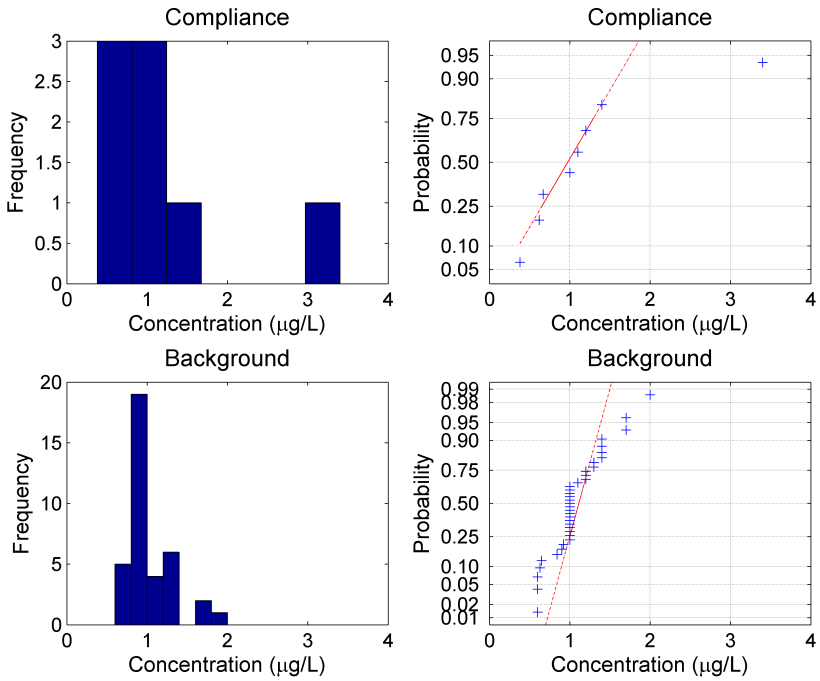


Figure A-2. Check assumption of normality.

Figure A-2 demonstrates that the data from the background well do not follow a normal distribution and both data sets are skewed. Thus, transform the data prior to analysis or use a nonparametric method. For highly censored data sets such as the data from the background well, it can be difficult to identify a particular statistical distribution. Given the significant number of nondetects in the background data set, a nonparametric method may be the best choice of statistical method for this example.

Although past guidance has recommended the use of the standard Wilcoxon rank sum test for censored data sets (assigning each group of data with the same reporting limit as a set of tied data), the current Unified Guidance cautions against this practice and recommends use of the Tarone-Ware test. The Tarone-Ware test is a generalized form of the Wilcoxon rank-sum test designed specifically for highly censored data sets. Although the test method is described in the Unified Guidance, the Tarone-Ware test is not implemented in many statistical software packages. A variant of the Tarone-Ware test that is more commonly included in statistical software is Gehan's test.

All of the nonparametric tests discussed have similar assumptions, which include the assumption of equal variances, the assumption of a common (unknown) distribution, and temporal stability. In many cases, these assumptions will not be met completely or may be difficult to verify. Figure A-3 presents a visual comparison of the variance of the background well with the last eight measurements of data at the compliance well. This visual method is recommended by the Unified Guidance for censored data sets.

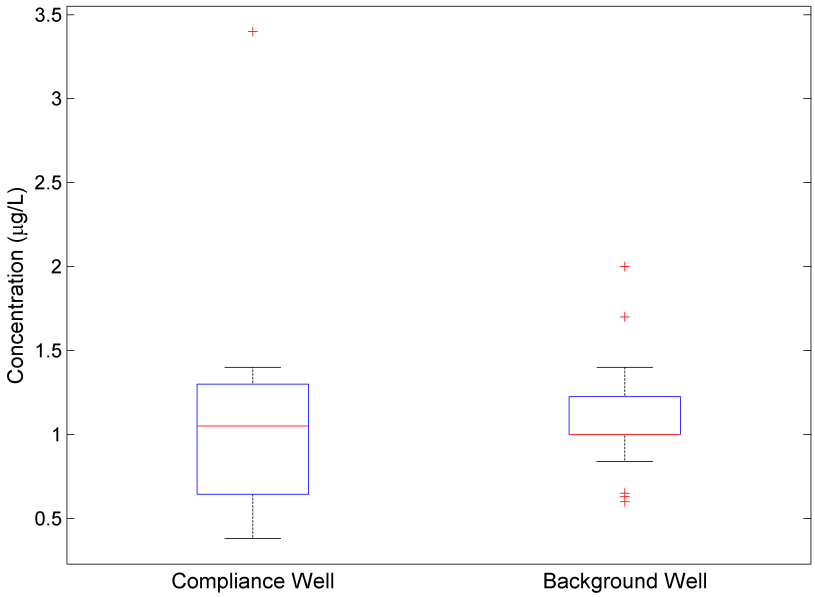


Figure A-3. Visual examination of variance.

The side-by-side box plots show that the data sets have similar buy not equal variances. It is fairly common that data from a compliance well will have a larger variance than data from a background well. From an examination of the probability plots in Figure A-2, it can be seen that the probability distributions are similar. From Figure A-1, neither data set has a temporal trend using only the final nine data points from the compliance well. Since there are no obvious violations of the test assumptions, perform the Wilcoxon rank sum and Gehan's tests. Because the variances of the groups are not equal, the power of the test to detect a difference between the groups when there actually is a difference may be decreased. However, if the tests find a difference between the groups, the difference is likely to be actually present.

A.1.3 Results

The USEPA software ProUCL was used to perform both the Wilcoxon rank sum test and Gehan's test, a variant of the generalized Tarone-Ware test. A significance level of 0.05 was selected. For this example, the null hypothesis is not rejected by either of the two-sample tests, indicating there is not a significant amount of evidence that the median concentration at the compliance well is greater than the median concentration at the background well. Based on this result, the remediation of metals at Site A appears to have been successful. Technically, the test results do not show that the site has been remediated, only that there is insufficient evidence (at a 5% significance level) that it hasn't. In other words, if more data were available, it's possible that the conclusion would change because our hypothesis tests would have more power to detect a difference between the two data sets.

To understand how confident you should be in this conclusion, it is useful to consider the p-values and power of the tests. Using the Wilcoxon rank-sum test the null hypothesis is accepted with a p-value of 0.17. Using Gehan's variant of the generalized Wilcoxon test yields a p-value of 0.15. The p-values give an indication of the strength of the evidence against the null hypothesis (site is less than background), with smaller p-values indicating stronger evidence. Since a significance level of 0.05 is selected, consider the evidence to be significant enough to conclude that the site is greater than background if the p-value falls below 0.05. However, a p-value of approximately 0.15 could be considered weak evidence against the null hypothesis, which should be considered along with other information in order to evaluate the total weight of evidence.

Since the null hypothesis was not rejected in the example, consider the power of the test. Power is the probability that the test would detect a difference between the two groups when there actually is a difference. ProUCL can be used to calculate the required sample size for the Wilcoxon rank sum test to achieve a certain level of power at a given significance level. Specify the size of difference in concentration between the groups that should be detected, which is referred to as the width of the gray region in ProUCL. In order to achieve a power of 75% in a single-sided test, ProUCL calculates a minimum sample size of 29, given a gray region width of 2 and the default standard deviation of the test statistic of 3. Since the sample size in the example is much less than 29, the power of the test is much less than 75%. The small power of the test means that there may actually be a small statistical difference between the site and background data sets. If a small difference would have an important impact on decisions at the site, then additional data or information should be collected.

A.2 Testing a Data Set for Trends Over Time

Chlorinated solvents have historically been detected in groundwater above regulatory limits at a former dry cleaning site. Suppose you want to test whether natural attenuation is taking place and if so, whether the compliance level will be reached within a reasonable time frame without active remediation. To test this hypothesis, the vinyl chloride data show in Table A-2 from a downgradient monitoring well are analyzed for a downward trend.

Table A-2. Time Series Data

Data Set 1 (µg/L)		Data Set 2 (µg/L)		Data Set 3 (µg/L)	
Date	Concentration	Date	Concentration	Date	Concentration
1/1/2000	14	1/1/2004	6	1/1/2008	10
4/1/2000	3	4/1/2004	3	4/1/2008	4
7/1/2000	10	7/1/2004	17	7/1/2008	6
10/1/2000	39	10/1/2004	30	10/1/2008	14
1/1/2001	11	1/1/2005	10	1/1/2009	8
4/1/2001	5	4/1/2005	4	4/1/2009	3
7/1/2001	19	7/1/2005	11	7/1/2009	5
10/1/2001	33	10/1/2005	31	10/1/2009	15
1/1/2002	13	1/1/2006	10	1/1/2010	9
4/1/2002	3	4/1/2006	2	4/1/2010	2
7/1/2002	10	7/1/2006	11	7/1/2010	11
10/1/2002	16	10/1/2006	12	10/1/2010	15
1/1/2003	12	1/1/2007	7	1/1/2011	6
4/1/2003	3	4/1/2007	4	4/1/2011	3
7/1/2003	11	7/1/2007	6	7/1/2011	9
10/1/2003	27	10/1/2007	26	10/1/2011	12

A.2.1 Overview

Various statistical tests can check the data set for a significant downward trend to demonstrate that natural attenuation is taking place. Linear regression is the most commonly used trend test, which tests the null hypothesis that the slope of the mean population is equal to zero. A one-tailed t-test is performed by comparing the test statistic to a critical point equal to the desired significance level. If the absolute value of a test statistic is larger than the critical point, then the alternate hypothesis is accepted that there is a significant downward trend at the specified significance. Alternately, a two-tailed test could be performed to test whether any trend exists. However, the use of linear regression is limited since it depends on a number of underlying assumptions that are not always satisfied in real-world environmental problems.

Alternatives to linear regression include the more robust, nonparametric Mann-Kendall and seasonal Mann-Kendall trend tests. These tests tend to perform better when analyzing real-world environmental data sets since the data do not have to conform to a particular distribution, and the tests can handle missing values and nondetects. These approaches also test the null hypothesis that there is no trend versus the alternative hypothesis that there is a trend. A Mann-Kendall statistic is formed by assigning a 1, -1, or 0 to each pair of data points depending on the sign (positive, negative, or equal) of the difference between the values. These values are then summed to form the Mann-Kendall statistic, which is compared to a critical value in a look-up table to determine if the null hypothesis is valid. When n is greater than 10, a normal approximation can be assumed based on the Central Limit Theorem. In this case, a Z statistic is formed and compared to a critical point from the standard normal distribution.

As opposed to linear regression, the Mann-Kendall and seasonal Mann-Kendall tests predict the median population at any given time. This is the reason nondetects and missing values do not necessarily affect the performance of the tests. According to the Unified Guidance, nondetects can typically be replaced by half the reporting limit as long as they make up less than 10-15% of the data. In this case, be sure that a downward linear trend is not merely an artifact of lower reporting limits due to improved analytical methods. For the Mann-Kendall and seasonal Mann-Kendall tests, nondetects can comprise up to 50% of the data; as long as these values occur in the lower half of the distribution, the median will always be based on a detected result.

Section 5.7 discusses managing nondetect data in statistical analyses. In addition, Helsel (2012) provides a summary of the pitfalls of simple substitution and cautions against the method recommended in the Unified Guidance. Helsel describes three alternative approaches to deal with censored data:

- nonparametric methods after censoring at the highest reporting limit
- maximum likelihood estimation
- nonparametric survival analysis procedures

The first method is intended for use when a simple analysis is warranted. The latter methods are more powerful, but require a more advanced understanding of statistics.

The seasonal Mann-Kendall test is also especially useful because seasonal fluctuations in real-world environmental data sets may not follow a predictable pattern. If this is the case, it may not be possible to de-trend a data set before analysis with the other tests.

In this example, linear regression and the Mann-Kendall and seasonal Mann-Kendall trend tests are performed on both the seasonally adjusted and unadjusted data to evaluate whether a statistically significant trend at the $\alpha = 0.01$ significance level exists. The results are then compared.

A.2.2 Visual Examination

The first step in trend analysis is to visually inspect the nature of the data set for apparent linear or cyclic trends. Cyclic trends such as seasonal patterns can mask linear trends and should be accounted for by either removing the trend from the data set prior to analysis or by using a method not affected by seasonality. More advanced statistical techniques may be required if complicated patterns are identified, such as abrupt changes in trends or impulses.

Evaluate statistical assumptions and limitations before formally testing a data set for a trend. For example, significant skewness, extreme data values, or non-normality can bias or invalidate linear regression results. Other assumptions inherent in linear regression depend upon the regression residuals and can only be checked after the test has been performed. Regression residuals must be normally distributed, show homoscedasticity, and be statistically independent. Homoscedasticity means that the variance is constant over time and mean concentration.

The following figures and analysis were developed using the Matlab software package. However, many other statistical packages could be used to create similar results. The four-plot in Figure A-4 is a common visual tool that contains a time-series plot, lag plot, histogram, and normal probability plot.

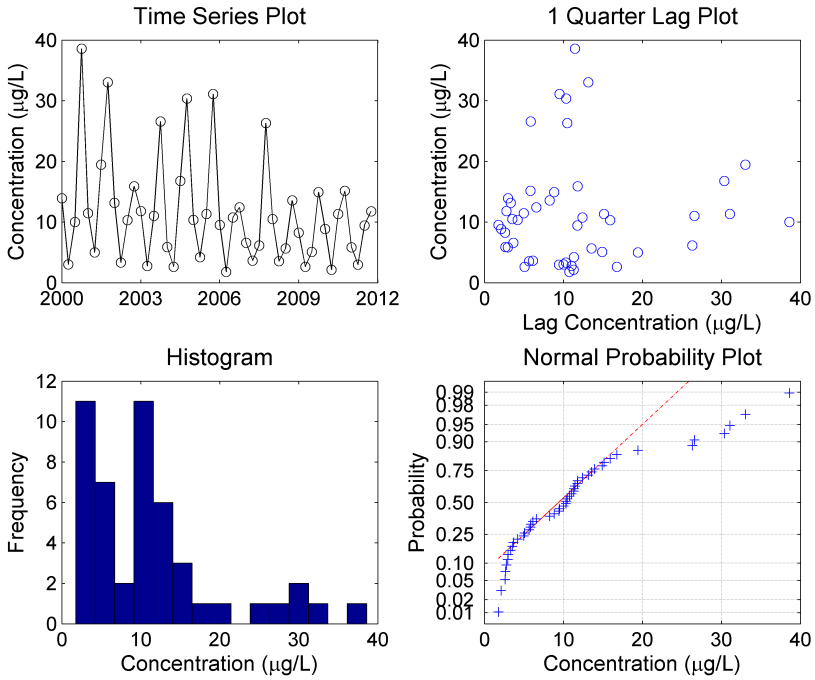


Figure A-4. Examination of original data.

The time series plot is used to check for apparent trends and outliers. Seasonal fluctuations are clearly evident in the time series plot. The lag plot is used to determine whether data exhibit auto-correlation. A structure-less lag plot would indicate lack of autocorrelation in the data set. The histogram and normal probability plot are used to check for normality and skewness. The data are from a normal distribution if the histogram is symmetric and bell-shaped and the normal probability plot is linear. The data in Figure A-4 appear to be log-normally distributed and left-skewed, as is often the case with environmental data.

In order to meet the requirements of linear regression, the data set is log-transformed and the revised four-plot is shown in Figure A-5.

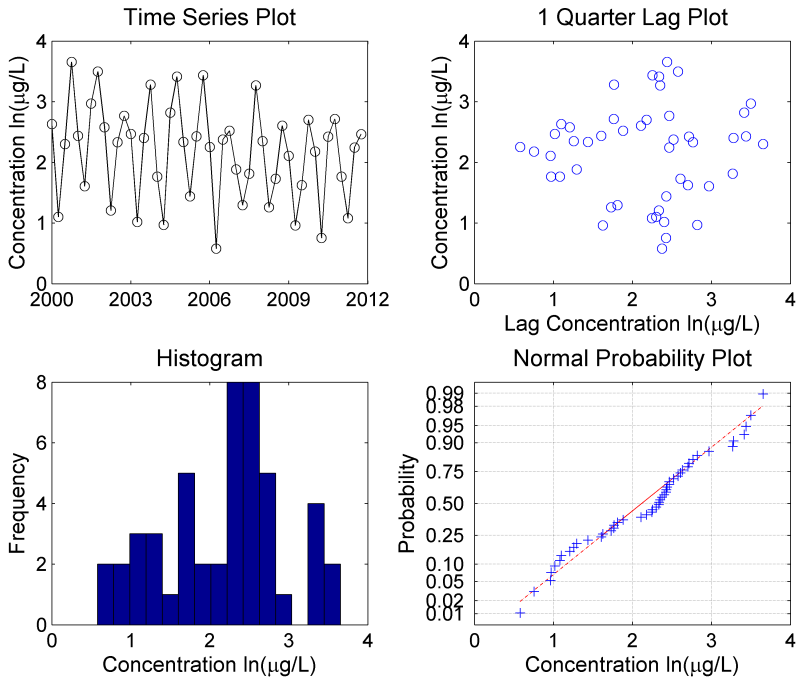


Figure A-5. Examination of log-transformed data.

The histogram is somewhat bell-shaped and the normal probability plot is close to linear, so the log-transformed data are approximately normally distributed. No skewness is now evident in the histogram and a linear trend is more evident in the time-series plot. The Lilliefors test was used to confirm that the transformed data are normal at the 5% significance level. Seasonality is also confirmed by the elliptical pattern in the lag plot.

Figure A-6 evaluates the assumptions of linear regression related to the regression residuals. The sample autocorrelation function confirms the presence of autocorrelation at lags that correspond to semi-annual and annual cycles, which violates the assumptions underlying linear regression. Sinusoidal patterns typically indicate a seasonal pattern is present. The nonparametric rank von Neumann ratio test could have also verified the assumption of statistical independence, if the data were not transformed to a normal distribution. Normality of the regression residuals is evaluated with a normal probability plot. The regression residuals appear to be approximately normal, slightly deviating from linearity. Homoscedasticity is verified by creating scatter plots of the regression residuals versus time and mean concentration. Since the scatter plots have uniform width and height and do not contain any discernible pattern it can be concluded that there is equal variance with respect to time and mean concentration.

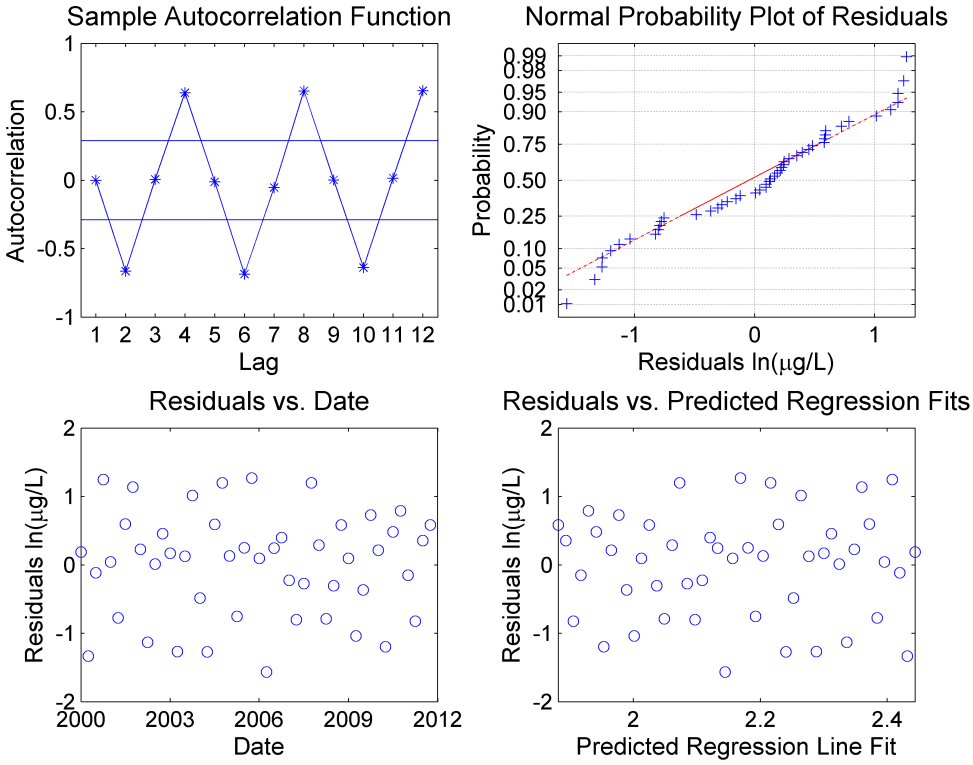


Figure A-6. Analysis of seasonally unadjusted data.

The seasonal pattern identified in the four-plot should be accounted for by either removing the trend from the data set prior to analysis or by using a method not affected by seasonality such as the Seasonal Mann-Kendall trend test. The results of the three trend tests on the seasonally unadjusted data are shown on Figure A-7 and in Table A-3. If the absolute value of the t- or Z statistics exceeds the target then the test demonstrates that there is a downward trend. Likewise, if the calculated p-value does not exceed the target p-value (significance level), then a significant trend exists. In this example, only the Seasonal Mann-Kendall trend test is capable of detecting a downward linear trend at the 0.01 significance level among seasonal fluctuations.

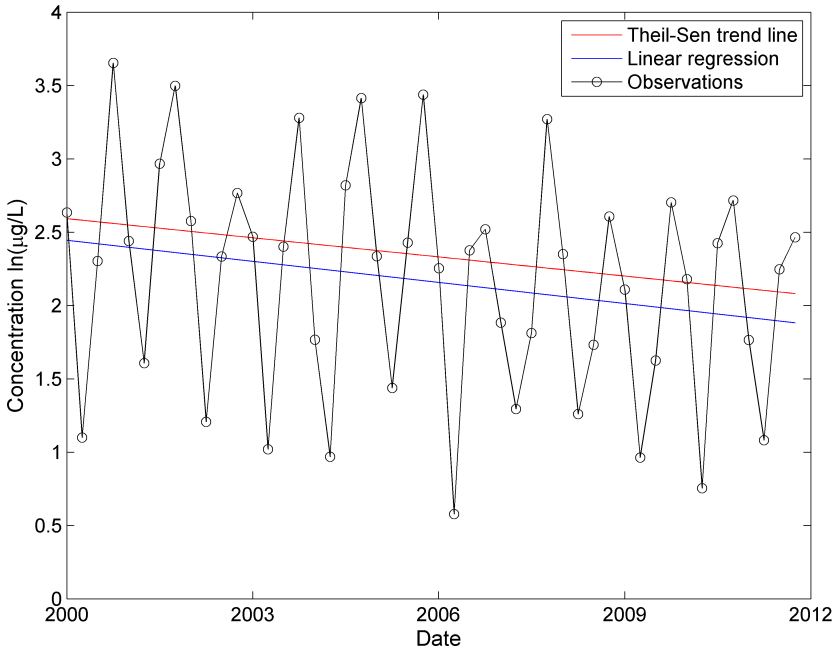


Figure A-7. Analysis of seasonally unadjusted data.

Table A-3. Trend test results on seasonally unadjusted data

Linear Regression Test				
t-statistic		p-value		Estimated Slope (µg/L/year)
Actual	Target	Actual	Target	
-1.63	2.41	0.055	0.01	-0.050
Mann-Kendall Test				
Z-statistic		p-value		Estimated Slope (µg/L/year)
Actual	Target	Actual	Target	
-1.70	2.33	0.045	0.01	-0.049
Seasonal Mann-Kendall Test				
Z-statistic		p-value		Estimated Slope (µg/L/year)
Actual	Target	Actual	Target	
-4.28	2.33	0.000009	0.01	--

Next, adjust the log-transformed data set to remove seasonal patterns according to the method described in Chapter 14.3.3, Unified Guidance. At least three complete, regular cycles of the seasonal pattern should be observed before adjusting the data in this manner. Figure A-8 compares the original and adjusted data sets.

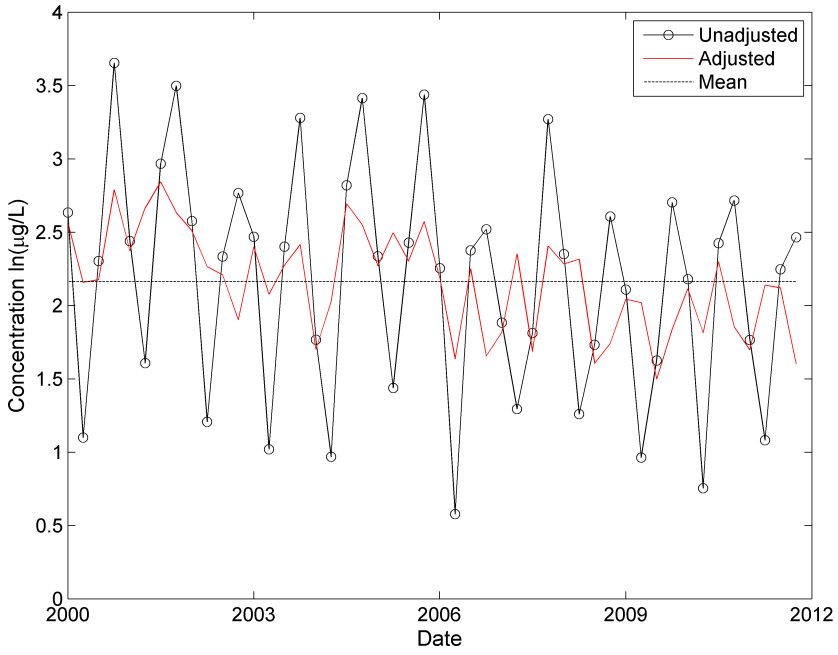


Figure A-8. Time series adjusted to account for seasonal variation.

Figure A-9 evaluates the assumptions of linear regression related to the regression residuals in the same manner as described above.

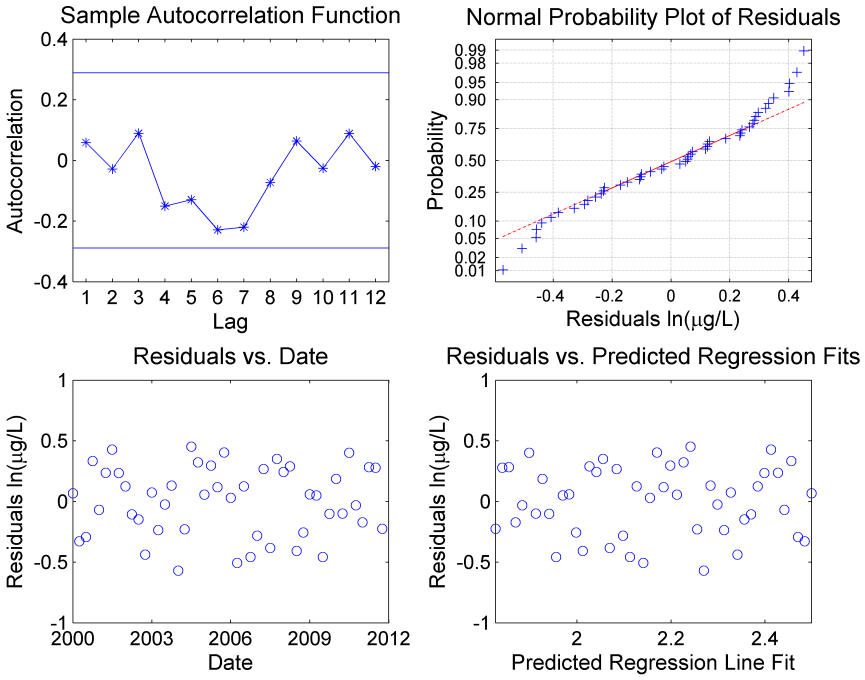


Figure A-9. Check assumptions of linear regression.

In this case, there is no significant autocorrelation identified and the normal probability plot is still only approximately linear.

Table A-4. Trend test results on seasonally adjusted data

Linear Regression Test				
t-statistic		p-value		Estimated Slope (µg/L/year)
Actual	Target	Actual	Target	
-5.14	2.41	0.000003	0.01	-0.059
Mann-Kendall Test				
Z-statistic		p-value		Estimated Slope (µg/L/year)
Actual	Target	Actual	Target	
-4.23	2.33	0.000012	0.01	-0.060
Seasonal Mann-Kendall Test				
Z-statistic		p-value		Estimated Slope (µg/L/year)
Actual	Target	Actual	Target	
-4.28	2.33	0.000009	0.01	--

All three trend tests are performed on the log-transformed and seasonally adjusted data set. The estimated trend lines and p-values are shown on Figure A-10 and in Table A-4. Once the seasonality is removed, both the regular Mann-Kendall test and linear regression also detect a significant downward trend in the data. This is evidenced by p-values lower than the target p-value (significance level).

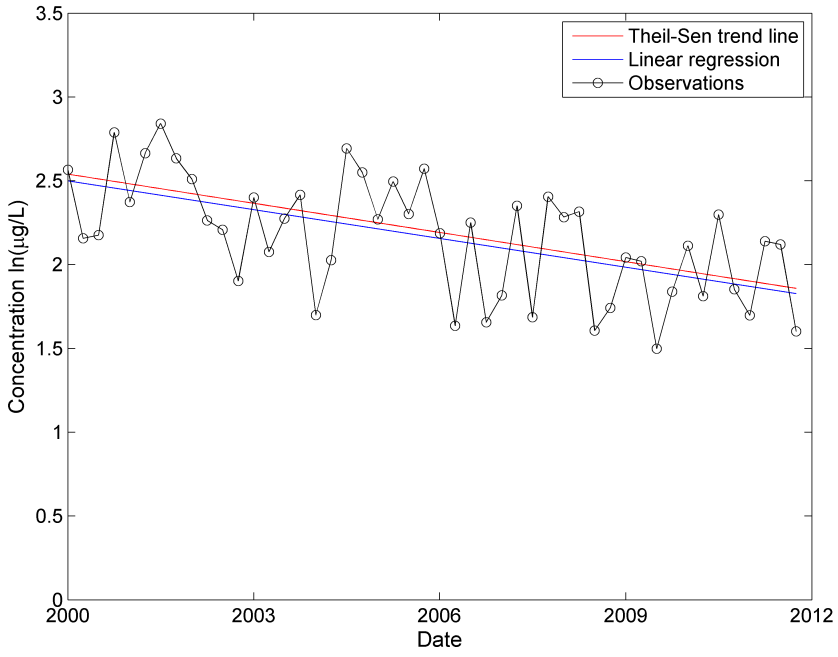


Figure A-10. Time series plot overlaid with linear regression and Theil-Sen trend line.

Once a significant decreasing trend has been identified, the estimated slope from the linear regression can be used to predict how long it will take to reach the compliance level (see Figure A-11). According to the Unified Guidance, the upper confidence limit on the mean is used to test for compliance in this situation. The confidence band is calculated according to the Unified Guidance for estimating confidence limits around a linear regression with an identified significant trend. However, the confidence band becomes quite large as the data set is extrapolated into the future so the estimated regression line is used predict that vinyl chloride concentrations will reach the federal maximum contaminant level (MCL) of 2 µg/L shortly after 2040. Site-specific information, along with the regulatory context, and input from stakeholders would be needed to determine whether this time frame is acceptable or if active remediation is warranted.

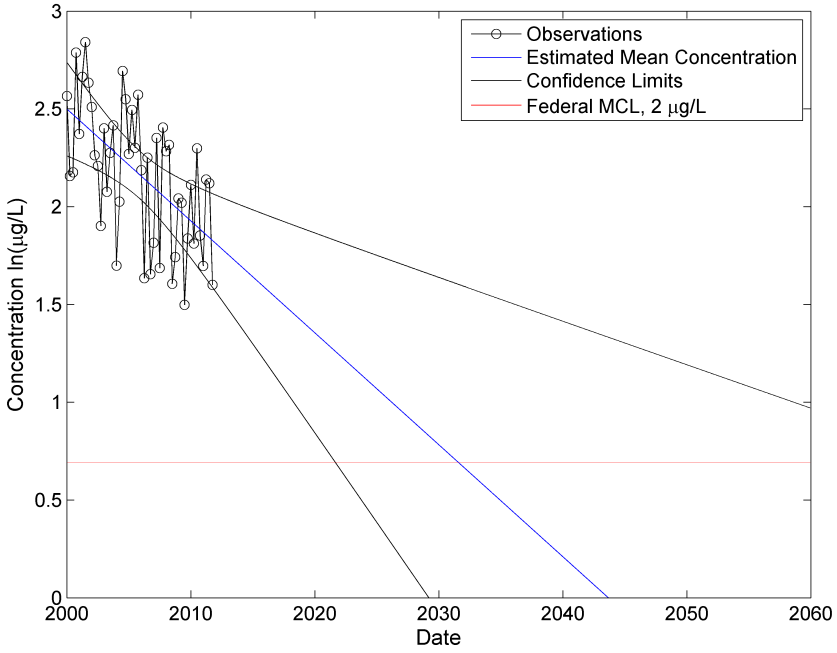


Figure A-11. How long until the compliance goal is met?

An alternative approach to assessing the progress of monitored natural attenuation was developed by the USEPA for use at CERCLA sites during the five-year review process (USEPA 2011b). This approach uses the calculation of an interim remedial goal at the end of each review cycle that indicates whether natural attenuation is proceeding adequately based on a first order decay rate law. This method could be applied to any situation by using a review period of any length.

A.3 Calculating Prediction Limits

Prediction limits apply to groundwater statistics in general, but are most commonly used to understand background concentrations. Comparisons to the prediction limit provide an answer to Study Question 2, Are concentrations greater than background concentrations?

There are four categories of prediction limits:

- prediction limits for m future values
- prediction limits for future means
- nonparametric prediction limits for m future values
- nonparametric prediction limits for a future median

This example considers the first category and was taken from Example 18-1, Unified Guidance. Table A-5 presents some data for an intrawell comparison.

Table A-5. Intrawell comparison data

Period	Date	Result (µg/L)
Background	1/1/2001	12.6*
	4/1/2001	30.8
	7/1/2001	52
	10/1/2001	28.1
	1/1/2002	33.3
	4/1/2002	44
	7/1/2002	3*
	10/1/2002	12.8
	1/1/2003	58.1
	4/1/2003	12.6*
	7/1/2003	17.6
	10/1/2003	25.3
Compliance	1/1/2004	48
	4/1/2004	30.3
	7/1/2004	42.5
	10/1/2004	15

*These values were also evaluated as nondetects in the supplemental analyses.

The 12 measurements taken during the background period are used to construct the upper prediction limit for the compliance period of four measurements. Construct and compare the upper prediction limit as follows:

1. Check the sample data for normality. For example, the Shapiro-Wilk test provides a test statistic equal to 0.947. Based on alpha=0.05 there is insufficient evidence to reject the assumption of normality. The diagnostic plots provided in Figure A-12 provide graphical information to support this conclusion.
2. Calculate the sample mean (27.52) and standard deviation (17.10).
3. Calculate the upper prediction limit based on the t-statistic. Suppose the overall confidence limit is 95% (1-alpha [0.05]). In this case there are four future measurements and the quantile for the t-statistic should be set to 0.9875 (1-0.05/4) based on assuming independence of these samples and making a simple Bonferroni adjustment. Note that Gibbons et al. 2009 warn that such simple Bonferroni adjustments do not account for the fact that the comparisons are correlated because all four compliance samples are compared to the same background.

Calculate the upper prediction limit (73.67) as follows:

$$PL = \bar{x} + t_{quantile,df} \times s \times \sqrt{1 + \frac{1}{n}}$$

where:

$\bar{x}$ = sample mean

s = standard deviation

n = number of values

$t_{quantile,df}$ = look-up value based on the t-distribution

df = degrees of freedom (df = n-1) = 11

4. Compare the four compliance samples to the upper prediction limit. Since none of the four measurements are greater than the prediction limit, there is no evidence for an increase in concentration for this well.

Supplemental Analyses. Nondetects are a common issue for groundwater sample results. Table A-5 includes three sample results from the background period as nondetect values. This change illustrates the impacts of nondetects on background statistics. The upper prediction limit (UPL) calculations were made using ProUCL for these supplemental analyses.

- For the 95% UPL for Next 4 Observations, with three nondetects and assuming a normal distribution, ProUCL returned a value of 77.71 using the maximum likelihood estimation (MLE) method.
- For comparison, ProUCL returned a nonparametric 95% UPL for Next 4 Observations, also based on three nondetects, of 69.68 using the Kaplan-Meier method
- For the 95% UPL for Next 4 Observations, using the original background data (all detects), and assuming a normal distribution, ProUCL returned a value of 73.67 (the same value as calculated above)
- For comparison ProUCL returned a nonparametric 95% UPL for Next 4 Observations, also based on the original background data (all detects), of 58.1, which was the maximum concentration reported

With the exception of the nonparametric UPL, the prediction limits calculated based on this example data set are fairly similar. This result is unusual, and for this example data set the gamma or lognormal based UPLs are several times larger than those calculated based on the normal distribution. Methods to evaluate nondetects in environmental data are an active area of statistical research and some of these tools are now readily available with statistical software like ProUCL. In contrast to the logic that USEPA has defined for calculating upper confidence limits (UCLs) based on varying levels of censoring and sample size there is no such guidance for UPLs or UTLs. However, it is good practice to select normal, gamma, lognormal, nonparametric UPL/UTL in that order assuming that these statistical distributions are not rejected. It is also recommended that you

plot the data and the resulting UPL/UTL. Lastly reviewing the UPL/UTL calculating with various methods can help you understand how sensitive the statistic is relative to distributional assumptions.

Figures A-12 and A-13 display the results of the supplemental analyses.

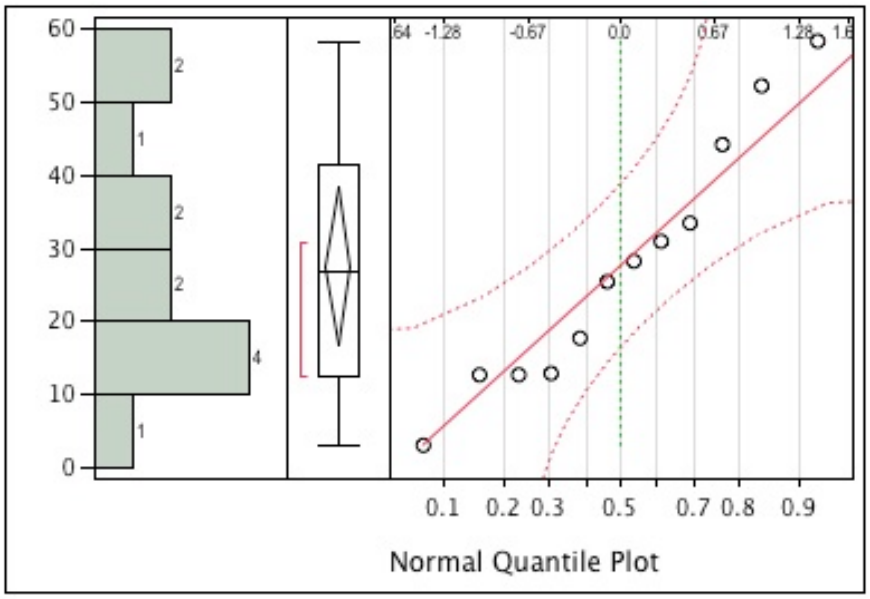


Figure A-12. Distribution diagnostic plots for background concentration data.

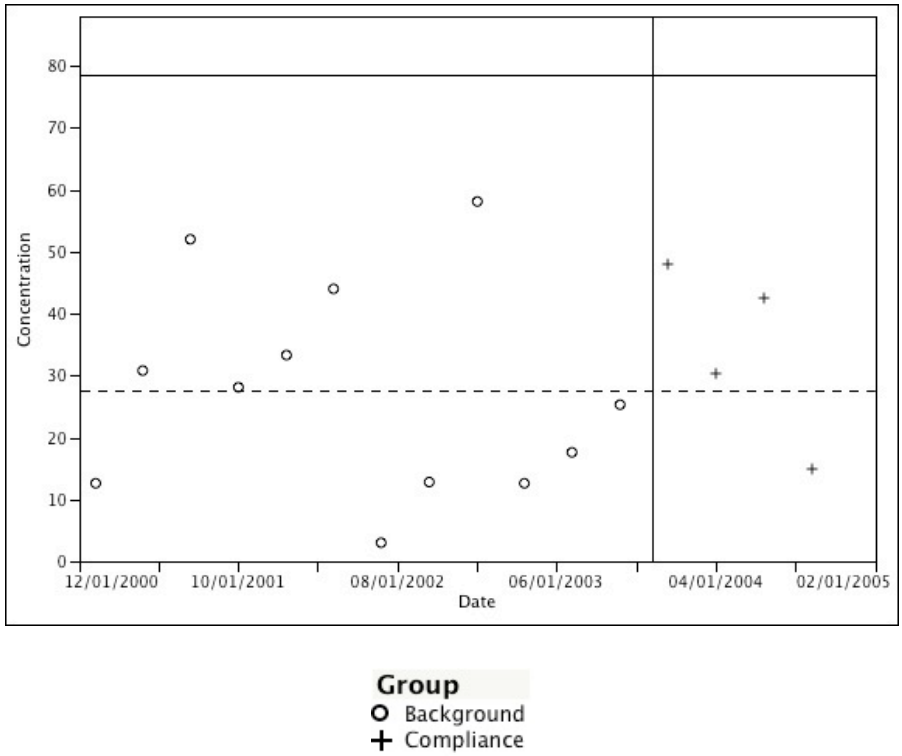


Figure A-13. Time series plot of background and compliance period data. Solid line is the upper prediction limit (73.67) and dashed line is the background period mean (27.52).

A.4 Using Temporal Optimization to Time Sampling Events

In this example, temporal optimization is used to examine the temporal spacing between sampling events. Groundwater sampling is often conducted on a default sampling schedule that may be monthly, quarterly, or semi-annually. In some cases, it may be possible to collect samples less frequently, but still be able to characterize contaminant concentrations over time. One statistical approach to this is iterative thinning.

Visual Sampling Plan (VSP) is an easy to use software package that can conduct this analysis. For this example, VSP was used to examine only two wells at a hypothetical site. More typically, there would be many wells at a site, and each of them would be analyzed using iterative thinning.

Figures A-14 and A-15 show the resulting plots from VSP. The original data are plotted using black dots, the smoothed data are shown with red dots, and the 90% confidence intervals are shown as blue lines. The important results is the overall trend in the data represented by the smoothed data. There is less variability in the concentrations from MW-1 (Figure A-14), so the confidence interval is much narrower for MW-1 as compared to MW-2 (Figure A-15). The iterative thinning algorithm identifies the frequency of sampling that would be required to reproduce the temporal trend.

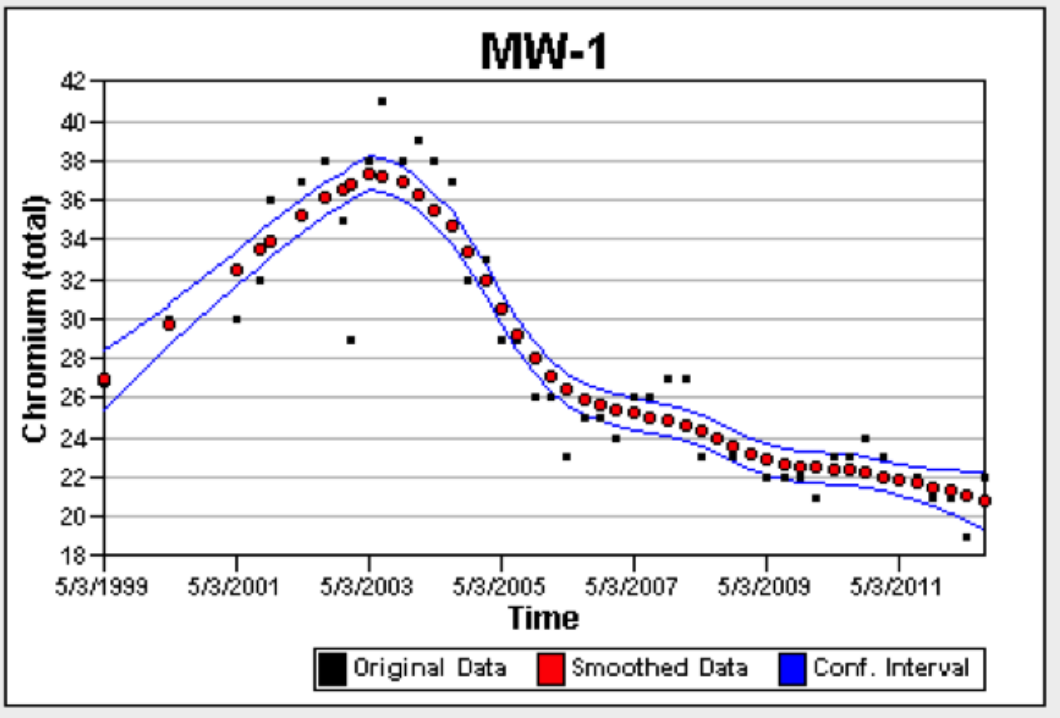


Figure A-14. VSP results for concentration (mg/l) from MW-1.

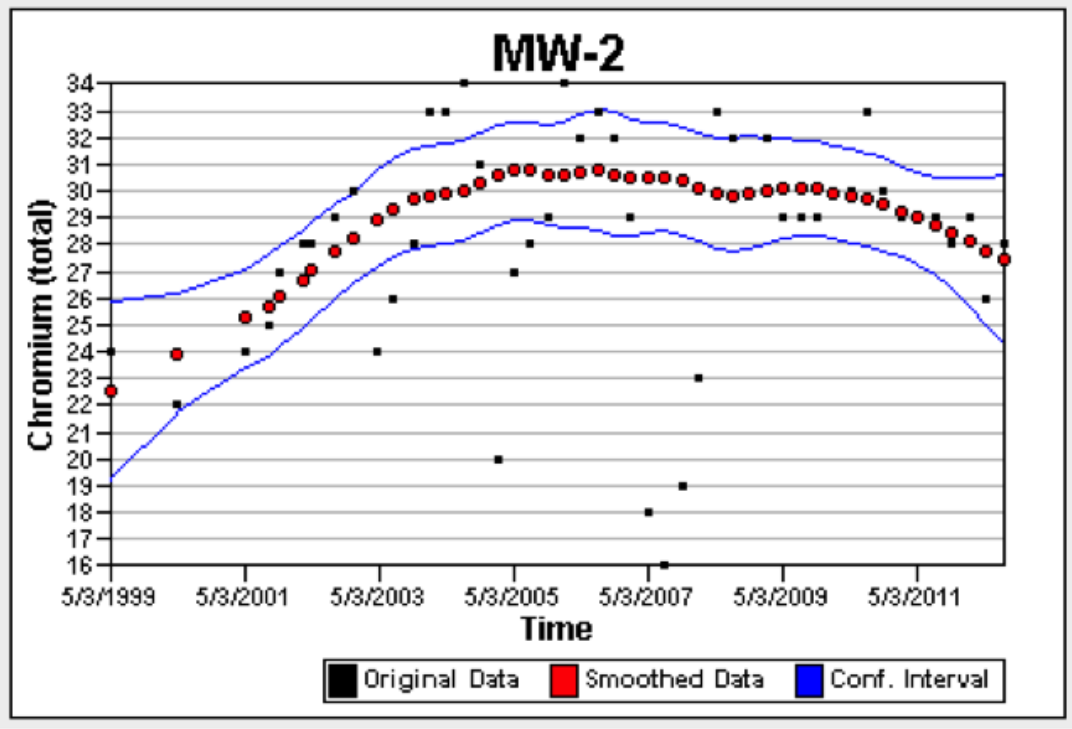


Figure A-15. VSP results for concentration (mg/l) from MW-1.

Both MW-1 and MW-2 were originally sampled quarterly. According to the VSP output, the optimal sampling frequency for MW-1 would be 202 days and for MW-2 would be 227 days. If a semi-annual frequency (180 days) were proposed for future sampling in both wells, there would be a 50% reduction in sampling costs with no significant difference in the ability to monitor trends in these wells.

A.5 Predicting Future Concentrations

This example presents the calculations for predicting future concentrations in a well based on an exponential (first-order) decay model.

For an exponential (first-order) attenuation rate, the predicted future concentration is:

Equation 1:

$$C_t = C_0 e^{-kt}$$

where:

C_t = predicted concentration at time t

C_0 = current concentration

k = attenuation rate

t = time between now and the date for the prediction

For Well A, the attenuation rate for vinyl chloride is 0.2 yr^{-1} with a 95% confidence interval of 0.1 yr^{-1} to 0.3 yr^{-1}. The current vinyl chloride concentration is 1000 µg/L. Based on this information, in 10 years the predicted vinyl chloride concentration would be:

$$C_t = C_0 e^{-kt} = 1000e^{-(0.2 \times 10)} = 135 \text{ µg/L}$$

A reasonable range for the prediction is:

$$1000e^{-(0.3 \times 10)} \text{ to } 1000e^{-(0.1 \times 10)} = 50 \text{ µg/L to } 368 \text{ µg/L}$$

The predicted time required to attain the MCL of 2 µg/L for vinyl chloride is 31 years.

$$t = Ln(C_0/C_t)/-k = Ln(2/1000)/-0.2) = 31 \text{ years}$$

A reasonable range for the prediction is 21 years to 62 years:

$$t = Ln(C_0/C_t)/-k = n(2/1000)/-0.3 \text{ to } Ln(2/1000)/-0.1) = 21 \text{ to } 62 \text{ years}$$

Equation 1 can be re-arranged to predict the time required to attain a specified criterion:

Equation 2:

$$t = Ln\ (C_0/C_t)/-k$$

where:
t = predicted time between now and attainment of the criterion
C_t = criterion
C_0 = current concentration
k = attenuation rate

For either equation, an estimate can be obtained by using attenuation rate determined in accordance with the procedures described in Example A.6. A reasonable range for the prediction can be evaluated using the confidence interval for the attenuation rate.

A.6 Calculating Attenuation Rates

A best estimate of the first-order attenuation rate can be obtained by fitting a first-order decay model ($C_t=C_0e^{-kt}$) to the concentration versus time data or by fitting a linear model for natural log concentration versus time data ($\ln(C_t) = \ln(C_0)-kt$). As illustrated in Figure A-16, when using the same data set, these two approaches yield identical results. In both cases, k is the first-order attenuation rate with units of $time^{-1}$. With the use of a bootstrapping method, many software packages also provide a 95% confidence band for the attenuation rate as described in Section 5.5 and Chapter 21.3.1, Unified Guidance. This confidence band is useful for evaluating the uncertainty associated with the estimated attenuation rate.

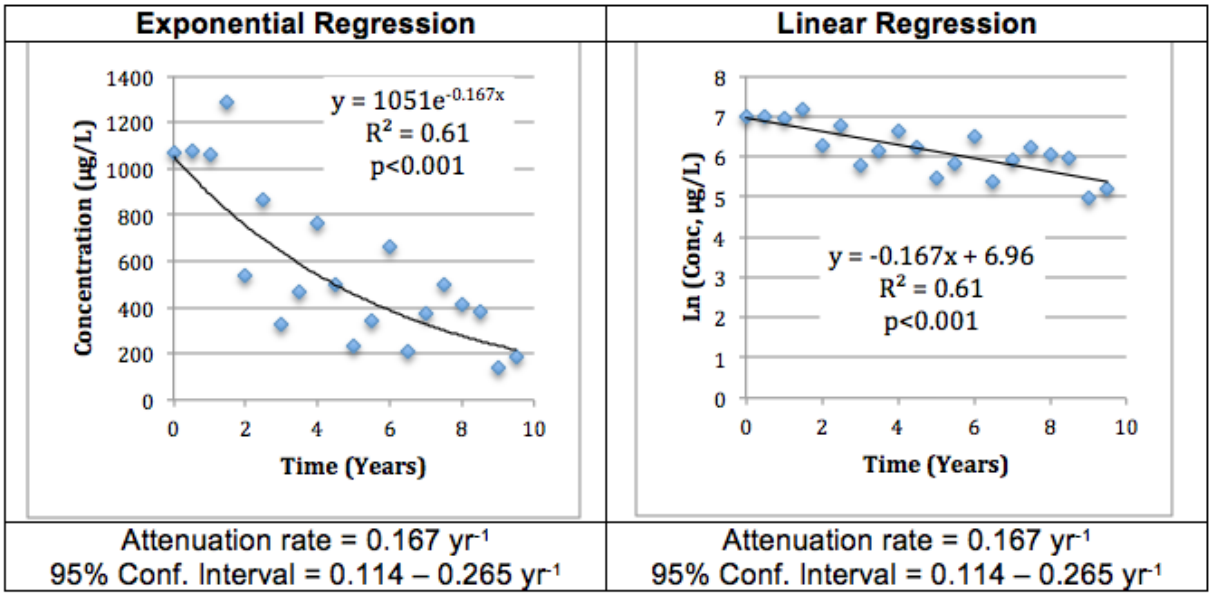


Figure A-16. Example of Regression Analysis for Temporal Trend Analysis.

A.7 Comparing Attenuation Rates

Two examples are presented here to illustrate the comparison of attenuation rates in two different wells at a site:

- Example 1, Different Attenuation Rates: The attenuation rate for Well A is 0.1 yr^{-1} with a 95% confidence band of 0.02 yr^{-1} to 0.18 yr^{-1}. The attenuation rate for Well B is 0.4 yr^{-1} with a 95% confidence band of 0.25 yr^{-1} to 0.55 yr^{-1}. It is reasonable to conclude that the attenuation rates in Wells A and B are different because the confidence bands do not overlap (that is the upper confidence limit for Well A (0.18 yr^{-1}) is smaller than the lower confidence limit for Well B (0.25 yr^{-1}).
- Example 2, Similar Attenuation Rates: The attenuation rate for Well A is 0.1 yr^{-1} with a 95% confidence band of 0.02 yr^{-1} to 0.18 yr^{-1}. The attenuation rate for Well B is 0.2 yr^{-1} with a 95% confidence band of 0.15 yr^{-1} to 0.25 yr^{-1}. It is not reasonable to conclude that the attenuation rates in Wells A and B are different because the confidence bands do overlap (that is the upper confidence limit for Well A (0.18 yr^{-1}) is larger than the lower confidence limit for Well B (0.15 yr^{-1}).

A.8 References

Gibbons, R.D., D.K. Bhaumik, and S. Aryal. 2009. *Statistical Methods for Groundwater Monitoring*. 2nd ed, Statistics in Practice. New York: John Wiley & Sons.

Helsel, D.R. 2012. *Statistics for Censored Environmental Data Using Minitab and R*. Second Edition. John Wiley & Sons, Inc. Hoboken, New Jersey, USA.

United States Environmental Protection Agency (USEPA). 2011. An Approach for Evaluating the Progress of Natural Attenuation in Groundwater. US EPA/600/R-11/204. December.

USEPA. 2009. Statistical Analysis of Groundwater Data at RCRA Facilities Unified Guidance. US EPA/530/R-09-007. March.

APPENDIX B. COMMON MISAPPLICATIONS OF STATISTICS

This appendix explains the mistakes that frequently appear in statistical reports. It also summarizes an appropriate alternative to the unacceptable practice. This information can help regulators, consultants, and stakeholders better understand how to apply statistics to groundwater data sets. The problems and errors below can occur during planning, implementation, or both (as noted for each).

B.1 Statistical Error and Resolution

1. *Problem/Error (Planning/Implementation): Concluding that, if the sample maximum is less than a decision criterion, this is likely conservative from a risk perspective.*

   It is not necessarily true that the study area contamination is less than the decision criterion because the sample maximum is less than the decision criterion. The study area population mean may be greater than the decision criterion when the sample maximum is less than the decision criterion, depending on the nature of the distribution and sample size.

   Recommendation: Use a one-sample hypothesis test.

2. *Problem/Error (Implementation): Concluding there is necessarily a problem because one grab sample result exceeds the decision criterion.*

   It is sometimes concluded that study area contaminant concentrations are elevated (and therefore additional remedial activities are needed) because one or more grab samples exceed the criterion. This conclusion is not necessarily valid. The study area mean may be less than the criterion, even when individual grab concentrations exceed the criterion (similar to Misapplication 1, above.)

   Recommendation: As part of a systematic planning process, determine whether numerical goals will be treated as "ceilings," simple averages, or other. Use a one-sample hypothesis to make inferences about the study area mean.

3. *Problem/Error (Planning/Implementation): Comparing the site sample maximum with a background threshold (for example, the background maximum or mean) without regard to potential decision errors.*

   The site sample maximum is being compared with the background sample maximum to determine if the site contamination is elevated relative to background. In general, do not compare maximums from the two data sets to make inferences about the means of the data sets as this does not control decision errors and can result in false positives.

   Recommendation: Use two-sample hypothesis tests or compare the study area results with background upper prediction limits (UPLs).

4. *Problem/Error (Implementation): Comparing the site sample maximum after a large number of site samples have been collected to a background 95% upper tolerance limit (UTL)*

and concluding that exceeding the UTL necessarily means the site concentration is elevated relative to the background concentration.

This approach does not control false positives. The probability of false positives approaches 100% as the numbers of study area results increase as when, for example, there are many samples for multiple analytes.

Recommendation: Use a two-sample hypothesis test.

5. *Problem/Error (Planning/Implementation): Using UTLs rather than upper prediction limits (UPLs) when false positives primarily need to be controlled.*

   Because only a small number of study area results were collected, but a large number of background results were available, hypothesis tests could not be performed to compare study area concentrations to background concentrations; instead the study area results were compared with UTLs. When background and site concentrations are not different from one another, all the study area results will be less than the background UPL with some specified level of confidence. For example, when the background threshold is a 95% UPL, there is a 95% chance 100% of the study area results will be less than the background threshold. However, only a proportion of the study area results will be less than the background UTL at the specified level of confidence. For example, when the background threshold is a "95% UTL with 95% coverage" (typically, referred to as a "95% UTL"), there is a 95% chance 95% (not 100%) of the study area results will be less than the background UTL.

   Recommendation: To best control false positives, use background UPLs rather than UTLs. Compare the k study area results (k is a variable representing the number of study area results) to the UPL for the next k future observations. The UPL depends on the number of study area results k that will be compared and increases as k increases.

6. *Error (Planning): Assuming that reliable inferences (decisions) can be made based on very small sample sizes (for example, n < 5).*

   There are many variations of this problem. For example, attempting to calculate an exposure point concentration as a 95% upper confidence limit (UCL) when the sample size is small (for example, n = 6), or, in the extreme case, comparing one measurement to the decision criterion, for example, the reporting requirement for polychlorinated biphenyls (PCBs) under the Toxic Substances Control Act (TSCA). Reliable inferences about the study area mean cannot generally be made when the sample sizes are small. The nature of the underlying measurement distribution cannot be reliably determined for example, to calculate a 95% UCL for exposure point concentrations for the risk assessment. The 95% UCL depends on the number of samples taken; it can be highly unstable in small data sets or in data sets with larger variation. The solution is to avoid undersampling by using a systematic planning process.

   Recommendation: Follow a systematic planning process, such as the seven-step DQO process, and establish tolerances for Type I (false positive) and Type II (false negative) decision

errors. Understand the requirements and limitations of the statistical methods that will be applied to the sample results (see Section 5.0 and Section 3.0) before deciding how many samples to collect. As a "rule of thumb," 8 to 10 measurements are often needed to do statistical evaluations, however, a larger sample size will likely be needed if the data set is very skewed or contains censored values (nondetects).The rule of thumb should not substitute for planning, determining what sample size is appropriate for the particular application, and documentation.

7. *Problem/Error (Implementation): Comparing sequential measurements in time with one another on a sample-by-sample basis to determine if there are increasing or decreasing trends.*

   This error may fail to distinguish random variability from long term changes in concentration. Statistical tests are needed to distinguish random variability from long term changes in concentration.

   Recommendation: Present time series plots with the results of statistical trend analyses (for example, p-values from Mann-Kendall tests).

8. *Problem/Error (Implementation): Substituting arbitrary multiples of the reporting limit for nondetects for statistical evaluations rather than treating the nondetects as inequalities.*

   This error distorts the data sets and can produce erroneous conclusions.

   Recommendation: Do not substitute arbitrary multiples of the reporting limits for nondetects for statistical evaluations (for example, ½ the detection limit) without considering the impact on the particular statistical evaluation to be performed (see Section 5.7.5). Treat nondetects as inequalities and use nonparametric statistical methods that do not rely on substitution of surrogates values for nondetects (for example, Kaplan-Meier methods). Section 5.7 includes discussion of managing nondetect data.

9. *Problem/Error (Implementation): Using an incorrect censoring limit; for example, reporting nondetects to the MDL in 40 CFR Part 136 Appendix B (which was designed to minimize Type I, false positive, error).*

   The reporting limits for nondetects need to minimize Type II (false negative) and not Type I error (given the null alternative hypothesis $\mu \leq 0$). This can underestimate contamination (resulting in false negatives) when comparing the nondetects to risk-based criteria.

   Recommendation: Report nondetects to the laboratory's quantitation limit or to a smaller reporting limit that was otherwise demonstrated to minimize false negatives (for example, the Limits of Detection (LOD) defined in the DOD Quality System Manual (DOD 2013).

10. *Error (Planning/Implementation): Using arbitrary decision rules; for example, concluding that groundwater is clean when three consecutive rounds of results are less than a criterion.*

    Arbitrary decision rules do not control decision errors.

Recommendation: Use a systematic planning process, such as the seven-step DQO process, to develop a statistical approach that controls decision errors to an acceptable degree. Then use one-sample hypothesis tests.

11. *Problem/Error (Implementation): Failing to use "like statistics" in comparisons.*
There are many variations of this problem. For example:

- Comparing the mean of Site A to some percentile of Site B
- Comparing the percentile of Site A to the same percentile of Site B and concluding on this basis that the mean of Site A is larger than the mean of Site B
- Comparing composite sample results from a study area (for example, which estimate the mean) to a background UPL or UTL determined from grab samples
- Comparing the 95% UCL of the mean from a site with few sample results to the 95% UCL of the mean from a site with many sample results

Even when the study area and background area possess the same underlying population, the percentiles of these distributions will differ. Replicate grabs will produce a distribution of measurements (x_i), but replicate composites, each prepared from n grabs, will produce the distribution of sample means of sample size n; that is, the distribution for the statistic $m_i = (x_{i1} + x_{i2} + \ldots x_{in})/n$ (versus the distribution for individual measurements x_i). Therefore, it is inappropriate to compare composite percentiles or individual composite results with grab percentiles.

For example, assuming both population distributions are normal, comparison of a study area composite to the 95th percentile of the distribution of background grabs is similar to comparing the 50th percentile (median/mean) of the study area with the 95th background percentile. Comparing unlike statistical parameters leads to unpredictable mistakes in decisions.

Recommendation: Understand the nature of the statistical parameters being compared. For example, as the number of samples increase, the 95% UCL will converge on the mean of the population but the 95% UTL converges to the 95th percentile. The 95% UCL and the UTL are therefore not comparable. In environmental applications, often the risk is calculated using the mean concentrations of the site but regulatory requirements may be based on a parameter in the upper tail of the distribution. A systematic planning process will identify the appropriate parameter.

12. *Problem/Error (Implementation): Failure to check assumptions required for statistical tests.*
There are many variations of this problem:

- Performing regression fits without testing the underlying assumptions required for these fits to be appropriate (for example, linearity, normality of the residuals, and constant standard deviation for the residuals).
- Assuming a distribution without testing the result to check whether it is reasonable.

For example, assuming measurements are normal or lognormal without testing for normality/lognormality.

- Assuming all temporal trends can necessarily be modeled by a simple equation of the form $y = at + b$ (thus ignoring periodic trends) and not investigating whether other models are more appropriate.

Recommendation: Use exploratory data analysis (EDA) techniques. Routinely graph data (box plots, scatter plots, and histograms) to qualitatively evaluate the distributions of the results. Use goodness-of-fit tests before doing other statistical tests. For time series data, investigate whether non-linear fits are more appropriate; for example, consider equations of the following form: $y = a\,t + b\,x + c\,\sin(2\,\pi\,t) + d\,\cos(2\,\pi\,t)$ where y denotes concentration, t the time, and a, b, c and d are constants.

Avoid software applications that do not emphasize using the correct test for the distribution.

13. *Problem/Error (Implementation): Using correlation as the sole criterion to evaluate the comparability of two different sampling or analytical methodologies. Believing that a correlation coefficient near one for two different methods means the two methods give comparable data.*
A correlation coefficient near one indicates the results are correlated, but does not mean that they are comparable.

    Recommendation: Use appropriate statistical tests for paired data, for example, refer to the U.S. Army Corps of Engineers guidance (USACE 1998).

14. *Error (Implementation): Failing to account for the uncertainty when determining a functional relationship between two methods/variables such as Y = F(X), so that when a measurement X is taken (for example, using a low cost method), Y can be predicted.*
This problem becomes more common as additional, less expensive analytical techniques are introduced. Here is an example: The study area was sampled with an inexpensive field method (giving results denoted by X). A limited number of split samples were also analyzed with a fixed-laboratory method (giving results denoted by Y). The field method was highly positively correlated with the lab method, giving a relationship $Y = F(X)$ (for example, $Y = a\,X + b$). The field method was then used to determine if portions of the study area are "clean" or "dirty." In effect, it was assumed that the field method produced reliable, definitive data for decision making using the relationship $Y = F(X)$. However, when X was measured to obtain Y, the uncertainty of the calculated value of Y was not reported or taken into account.

    Recommendation: In the example above, a prediction interval for the fit $Y = F(X)$ should have been calculated, as the field method was being used to determine extent of contamination by comparing individual measurements to the cleanup criterion.

15. *Problem/Error (Planning): Setting the null hypothesis as H_0: $\mu(site) > \mu(background)$ when comparing study area concentrations to background concentrations.*

This error may be committed out of consideration of the "precautionary principle." However, it requires the data user to show that site concentrations are less than background concentrations. As a practical matter, this condition would be nearly impossible even if the site were not contaminated.

Recommendation: Set the null hypothesis as "H_0: μ(site) < μ(background)." Note that in ProUCL, when using two sample hypothesis test for site and background data, the default null hypothesis H_0 is consistent with this recommendation.

16. *Problem/Error (Implementation): Discarding outliers solely on the basis of a statistical outlier test without presenting any physical justification, especially for a background data set (thereby biasing the results).*

    Obviously, discarding outliers will bias the results of statistical tests using the censored data. If a data point is a statistical outlier to the rest of the data set, however, it is not necessarily incorrect or unrepresentative. This principle applies whether the data set is from a potentially contaminated site or from a reference background site.

    Recommendation: Retain the outliers unless there is a strong weight of physical evidence to remove them. This decision is not a purely statistical consideration and should be made in the context of a systematic planning process. Document why these samples are likely not physically representative of anthropogenic or natural background conditions or why the quality of laboratory analytical results is substandard.

17. *Problem/Error (Implementation): Failing to distinguish between a statistically significant result and a result of no practical significance.*

    A test of normality will detect very small deviations from normality when the sample size (that is, the number of samples) n is large, but for the end data use the small deviation from normality may not be of any practical importance. Be sure to determine both whether a difference exists and the magnitude of the difference.

    Example: As reported, a test of normality detected a very small deviation from normality as the sample size n was very large. However, the normal probability plot and histogram indicated that the distribution is essentially normal. Also, the one-sample hypothesis test is being done to determine if the mean is greater than a decision criterion. Under the Central Limit Theorem, the large value of n and near normal distribution of the individual measurements, normality can be assumed to use a one-sample t-test.

    Recommendation: Use graphical methods to determine the reasonableness of tests for normality, lognormality, or other distribution. When statistically significant differences are detected, assess the magnitude of these differences in the context of project's objectives. Document decisions related to the appropriateness of statistical tests.

18. *Problem/Error (Implementation): Concluding that the failure to reject the null hypothesis "proves" the null hypothesis (when the power of the test is not also addressed).*

Typically, large random variability exists in environmental data. A trend or difference in concentration could be relatively small. Failure to reject the null hypothesis does not prove the null hypothesis if the statistical test is of insufficient power.

Recommendation: Use care when stating a conclusion based on failure to reject the null hypothesis (for example, "The null hypothesis could not be confidently rejected."). See Rong, Y. (2011).

B.2 References

DOD (Department of Defense). 2013. *Quality Systems Manual (QSM) for Environmental Laboratories Based on ISO/IEC 17025:2005(E) and The NELAC Institute (TNI) Standards*. Volume 1, (September 2009) DOD Quality System Manual Version 5.0.

Rong, Y. 2011. "Statistical Methods and Pitfalls in Environmental Data Analysis." *Practical Environmental Statistics and Data Analysis* 10:243-258.

United States Army Corp of Engineers (USACE). 1998. "Environmental Quality, Technical Project Planning (TPP) Process." EM 200-1-2. Washington, D.C.: Department of the Army.

APPENDIX C. STUDY QUESTIONS

- Study Question 1: What are the background concentrations?
- Study Question 2: Are concentrations greater than background concentrations?
- Study Question 3: Are concentrations above or below a criterion?
- Study Question 4: When will contaminant concentrations reach a criterion?
- Study Question 5: Is there a trend in contaminant concentrations?
- Study Question 6: Is there seasonality in the concentrations?
- Study Question 7: What are the contaminant attenuation rates in wells?
- Study Question 8: How do contaminant concentrations change with distance from the source area?
- Study Question 9: Is the sampling frequency appropriate (temporal optimization)?
- Study Question 10: Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

C.1 Study Question 1: What are the background concentrations?

Background is defined as groundwater which is not influenced by the releases from a site (Section 4.2.1). Specifically, a background groundwater data set may represent either a location or a time period that has not been influenced by a release from the site.

While it may be convenient to think of the background concentration as a single value, the concentration of a chemical will naturally vary both spatially and temporally. The analytical process introduces additional variability. Because of both natural and introduced variability, background can best be understood as a distribution, prompting the question: “Is the distribution of concentrations consistent with the distribution of concentrations in background?”

Background data may be collected by either of two methods. Data may be collected from a number of different wells (interwell data collection). Data may also be collected from the same well over time (intrawell data collection; see Section 3.6.4). If interwell comparisons are desired, a hydrogeologic assessment must be performed to evaluate whether the upgradient and downgradient wells are appropriately grouped. For example, are the wells screened in the same geologic formation (Section 4.2.2)? If representative upgradient wells are not available for use as background, or if spatial variability exists among background wells, intrawell comparisons may be better for evaluating background conditions. Intrawell evaluations assume that the background time period is uninfluenced by chemicals from the site.

Chemicals present at background concentrations in the groundwater may be either naturally occurring or anthropogenic. Naturally occurring substances present in the environment are those that are not a result of human activity. "Anthropogenic substances are natural and human-made substances present in the environment as a result of human activities not specifically related to the site in question" (USEPA 2002c).

Tests used to determine if a particular data set is consistent with background assume appropriate data collection and focus on data characterization.

This question is usually relevant in the release detection, site characterization, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Verify that the collected data set represents background. Using graphical methods (Section 3.3 and Section 5.1) and distributions, establish that a collected data set is consistent with background (natural or anthropogenic).

- Identify outliers using box plots, probability plots, Dixon's test, or Rosner's test.
- Plot data sets on maps and in three dimensions (vertical, horizontal, and time) and examine data sets for sources of contamination, important areas that have not been sampled, spatial correlations or trends in the data, and the location of suspected outliers. See Section 3.3.3: Exploratory Data Analysis.
- Check that mean and variance are stable over the data set time frame (stationarity) and seasonality in the data is accounted for and considered in the analysis.
- Analyze the data for significant trends.
- Note and appropriately address nondetects in the data set (see Section 5.7: Managing Nondetects in Statistical Analyses).
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools

Probability Plots

- These plots show the entire distribution of measured concentrations, ranging from the lowest value to the highest value, against percentile of the distribution of measured concentrations.
- Probability plots are useful for identifying data distribution, range, bunching and outliers.

Using this test and interpreting results

The goal of this test is to verify that the data set appears consistent with background. Examine the data set distribution for its range, its visual skew (or, inversely, its bunching), and for outliers. Do these elements support the conclusion that the data set appears to be drawn from a single population? You can use a scatter plot to examine a data set for the same parameters as outlined for the probability plot (Section 5.1.3).

Begin data analysis with a probability plot to identify potential weaknesses in the collected data set. Potential weaknesses include problems such as the data set not being distributed as expected. The data also may be bunched, or may have extreme outliers.

If the data set exhibits the above characteristics, you may need to investigate and address outliers, collect additional data, or select new sample points from which to collect potential background data.

Time Series Plots

- These plots show measured concentrations over time.
- Time series plots are useful for assessing trends, patterns, and inconsistencies in the data set.

Using this test and interpreting results

The goal of this test is to verify that the concentration of a chemical is in steady-state equilibrium over time, and that broad variations in chemical concentration are the result of identifiable events or seasonality.

Conduct this examination to identify potential trends over time. While background groundwater data may have seasonal variation in the concentration of many chemicals, seasonal variation should occur within a range and should be repetitive within the range over time. Trends in background data may occur due to changing hydrogeologic conditions or influences from upgradient sources. Some statistical tests require that background data remain stable, in which case background data should not demonstrate either an increasing or decreasing trend over time.

If concentrations exhibit either increasing or decreasing trends over time that are not attributable natural, cyclical events, then new sample points must be selected or monitoring may need to be extended until a stable trend is observed. Historical data that are no longer representative may be removed from the data set. Alternatively, trends in upgradient wells may indicate that intrawell tests are preferable. Additional information is presented in Chapter 5.2.5 and Chapter 5.3.4 of the Unified Guidance.

Note that the time series plot is a specific kind of scatter plot. If the goal is to examine the relationship between two variables, for example, the correlation between the concentration of chromium and the concentration of iron at a site, then refer to Section 5.1.3: Scatter Plots.

Outlier Identification

- This test examines the data set for extreme concentrations (outliers).
- This test is useful for ensuring that the data set is representative of background and does not include nonbackground samples.

Using this test and interpreting results

The goal of this test is to determine if any of the samples in the data set appear unrepresentative of the background data set. A statistical outlier in the background data set may indicate that one of the background samples was collected in a location that is not truly background.

If the concentration of a sample indicates that the sample is outside the background data set, then that data point may distort the statistical analysis. Statistical outliers, however, may represent real variations in the background data and should not be automatically removed unless there is a reason to suspect an error or data quality issue (see Chapter 5.2.3, Unified Guidance). Use professional judgment in evaluating whether or not a statistical outlier should be retained in a background data set (see Section 5.10).

Interpretation of Results and Associated Uncertainty

The natural variation, the anthropogenic variation, or both variations in concentrations must be understood before developing background values. The expected distribution, character of the probability plot, the potential concentration variation across seasons as well as over time, and the occurrence of apparent outliers are all a function of the chemical and its environmental setting. Examine published studies regarding the occurrence of the chemical to determine which analyses should be emphasized or more heavily weighted in decision making.

Based on the qualitative examination of the background data set, you may choose to analyze the data set and present its basic statistical characteristics. See discussions on characterizing the data set presented in Section 3.3.3, Section 5.1, and Section 5.6.

The background value is not determined only once. The individual wells or groups of wells which were used to support background determination or comparisons may develop trends. These trends could result from new contaminant sources influencing previously unimpacted wells. These trends could also result from changes in groundwater flow or chemistry. Trends that are not sustained could also result by chance.

Regardless of the reason for changes, background data must be updated. How often the data must be reconsidered for update depends on site-specific parameters such as groundwater flow velocity, nearness of other potential sources of contamination, and geochemistry. Frequency of updating the background data is also dependent on having sufficient new data to statistically identify a change; the Unified Guidance suggests four to eight new data points. The new data may be compared to the historical data by either the parametric t-test or the nonparametric Wilcoxon rank sum test depending on the distribution of the pooled intrawell data. If there is no significant difference between the new data and the historical data, then the new data can be considered background. Additionally, the absence of a trend in the data when historical and new data are combined, is indicative of background; see Section 5.5, Section 5.5.1, and Section 5.5.2.

Related Study Questions

Study Question 2: Are concentrations greater than background concentrations?

Key Words: Background, Compliance Monitoring, Concentration Comparisons, Release Detection, Site Characterization, Closure

C.2 Study Question 2: Are concentrations greater than background concentrations?

Determining whether a site's groundwater has been impacted usually requires either comparison of site data to a single criterion derived from concentrations measured in background samples, or a direct comparison of the site data to the background data set. To determine the statistical tools which will meet your specific needs, identify the type of comparison to be made (that is, comparison to a single criterion, or a two data set comparison). The distribution assumption for the data set, and whether interwell or intrawell tests will be used, also determine the selection of the proper statistical methods. Note that the site data and the background data to which they are compared must share the same hydrogeologic and geochemical parameters (see Section 4.3.1: Physical Site Conditions and Section 4.2.1: Background Conditions).

This question is usually relevant in the release detection, site characterization, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Determine that the data sets to be compared meet the assumptions of the test to be used. Verify that the background data set is representative (see Study Question 1). Refer to Section 3.4 for further discussion of how the following requirements may impact statistical analysis results.

- Check that no autocorrelation exists between successive sampling events associated with assumption of random samples (see Section 5.8.3).
- Confirm that no significant trends are present in the data set (see Section 5.8).
- Examine variance and the stability of the mean; similarly, ensure that seasonality is accounted for and considered in the analysis (see Section 5.8).
- Identify outliers. Use box plots, probability plots, Dixon's test, or Rosner's test to confirm outliers.
- Address nondetects in the data set appropriately (see Section 5.7).
- Determine the data distribution and use it to inform selection of the statistical methods (see Section 5.6).
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools

After checking that the data meet prerequisites common to most statistical tests, determine which tests will provide the information needed using the data you have or data that you will collect. Depending on the source of the background data, comparisons will either be interwell or intrawell. Background data set development is discussed in Study Question 1.

There are two general approaches for analyzing the site data and determining whether site chemical concentrations are above those measured in the background; individual compliance samples can be compared to pooled background results, or pooled compliance samples can be compared to pooled

background. Site-specific considerations and regulatory considerations usually determine which approach is used. In either case, parametric and nonparametric methods are available, so determining the distribution of the data is a typical initial data examination step (see Section 3.4.3). Prediction limits, tolerance limits, and control charts allow individual samples (an individual well sampled in time for control charts) to be compared to pooled background samples. T-tests and ANOVA-type tests only allow for comparing of pooled compliance samples to pooled background samples.

Tests That Support Examination of Individual Sample Points

Prediction Limits

Prediction limits (PLs) estimate an interval in which future observations will fall, with a defined probability, given the data which had been collected. The calculation of PLs takes into consideration the number of future results to be compared as well as the number of retests required to confirm a release. Once a background data set is established, prediction limits based on the data are used as the criteria for comparison of compliance samples.

- PLs are typically projected around means or medians (Section 5.4).
- An upper prediction limit represents a level that is predicted to equal or exceed future sample values based on past results.
- The number of future samples must be specified.
- The confidence level of a prediction limit represents the probability that a specified number of future samples drawn from the same population will be below the prediction limit.
- Prediction limits increase (or if viewed graphically "widen") as the number of testing events are increased into the future.
- Test can be constructed to examine a single compliance well using a single sample.

Using this test and interpreting results

Prediction limits depend on a PL factor, K. The value of K depends on the selected site-wide significance, the number of background measurements, the anticipated number of resamples, and the number of chemicals examined. As the number of chemicals examined increases or the number of resampling instances increases, or both, the upper prediction limit increases because the K value increases. An increased K value corresponds to a decrease in the power of the test, that is, the probability of missing a true exceedance in a well increases. To reduce this source of error, it is important to limit the number of chemicals examined.

Site data can be compared to interwell prediction limits to evaluate whether site data are above background, or upgradient, concentrations. Interwell prediction limits may be useful during site characterization or closure project stages. Intrawell prediction limits, calculated based on historical data collected from a single well, can be compared to current concentrations in that well to evaluate whether a statistically significant increase has occurred. Intrawell comparisons may be useful for release detection. Recommendations for use of prediction limits to calculate a fixed groundwater protection criterion for compliance monitoring are discussed in Section 0.1.

If the individual site well data or selected statistic are less than the prediction limit then you have evidence that the compliance data are consistent with background or at least not inconsistent with background. If the site data are above the prediction limit then you should conclude, that within the confidence level of the prediction limit, the data are inconsistent with background. Resampling to verify the result is appropriate.

Control Charts (individual wells over time)

- Control charts compare data collected sequentially in time to historical background data.
- Control charts can evaluate either intrawell or interwell data.
- Control charts are a parametric procedure.
- Individual samples collected over time must be sufficiently separated in time so as to support that the samples are independent, that is, you are not sampling essentially the same water multiple times.
- Collect a sufficient number of samples, 8 to 10 samples, to support a reliable estimation of the mean and standard deviation. A larger data set may be needed if the data are skewed or there are nondetects.

Using this test and interpreting results

Intrawell control charts are a useful tool for release detection at sites when historical data from the compliance well exists. Interwell control chart evaluate whether site data are above background, or upgradient concentrations. Control charts may be useful during site characterization or closure project stages. Recommendations for the use of control charts are discussed in Section 5.13.

Individual site well data are compared to the background data using a control limit, the calculation of which depends on the mean and standard deviation. If the data from an individual well are less than the control limit then you may conclude that the data set is consistent with background. If the data are above the control limit then resampling to verify the result is appropriate. The result could indicate that the mean concentration of a contaminant has increased or that the result was a chance occurrence.

Tolerance Limits

- Tolerance limits (TLs) are designed to contain a percentage (typically 90%, 95% or 99%) of the background data set with a specified level of confidence (typically 95%, see Section 5.3).
- Tolerance limits can be used in lieu of PLs or combined with PLs for re-testing to control false negatives.
- Tolerance limits can evaluate either intrawell or interwell data.
- Tolerance limits, are typically calculated around means or medians (Section 5.3).
- The confidence level of a tolerance limit represents the probability that a specified percentage of the population is captured.
- A test can be constructed to examine a single compliance well using a single sample.

Using this test and interpreting results

Individual site well data can be compared to tolerance limits developed using upgradient wells or within-well data for the background data set. An upper tolerance limit can serve as an alternate groundwater protection criterion. However, tolerance limits by definition do not cover the full range of the background data set. Therefore, use of tolerance limits for decision making should incorporate an acceptable failure rate. As a variation to PLs, exceedance of the tolerance limits (TLs) would probably require more retesting compared to using the PLs as the criterion. Recommendations for use of tolerance limits are provided in Section 5.3.

Individual site well data are compared to the background data using a tolerance limit. If the data from an individual well is greater than the tolerance limit, there is reason to suspect that the site is impacted and resampling to verify the result is appropriate.

Tests that Support Examination of Pooled Data

In many situations, such as site characterization, it is desirable or advantageous to compare pooled data sets. A key assumption when pooling data from multiple sampling points, is that variability between wells is minimal; however, in many natural systems this spatial variability is too great to be ignored and therefore, it should be tested before pooling data. This is particularly true when pooling data to represent background concentrations.

Sometimes, pooling background data is appropriate. For example, when building a background data set, it may be possible to combine data sets that are thought to be background, but which were collected at different times or were spatially separated from one another. Even though it would be exceptional, given the typical variability in groundwater chemical concentrations, monitoring networks that have low natural spatial variability like sand aquifers or artificial systems, may be examined by pooling data.

Parametric Two-Sample T-Tests

- These tests are used to compare two data sets for equality of means.
- The tests require normally-distributed data.
- Eight to ten samples are recommended.
- Nondetects must be assigned values; see Section 5.7.5 and Section 5.7.6
- Welch's t-test does not assume equality of variance; see Section 5.11.1.
- Pooled variance t-test assumes equality of variance; see Section 5.11.2.

Using these tests and interpreting results

Pool background data from one or more wells and compare to pooled site characterization data from a single well (or multiple wells) to determine if the means are significantly different. Each data set should contain at least 8 to 10 samples and sample sizes in each data set should be similar for the most robust test. The greater the inequality in data set sizes, the lower the accuracy of the estimated probability of erroneously concluding that background data are significantly different from site data.

A calculated t-statistic greater than the critical t-value indicates a statistically-significant difference between the means of the two data sets; this difference indicates that impact may have occurred.

Parametric ANOVA F-test

- This test compares two or more data sets for equality of means.
- The test requires a normal data distribution.
- Nondetect values should be assigned (see Section 5.7.5 and Section 5.7.6).
- The test assumes the two populations have equal variances.
- The test assumes samples are spatially and temporally independent.
- Eight to ten samples are recommended.

Using this test and interpreting results

Background data from one or more wells is combined then compared to site data by examining the variance between separated wells as compared to the variance between multiple samples taken from the same well. If the well to well variability is the same as the within-well variability then the means must be equal. If the means are not equal then well to well variability is greater than the within-well variability.

There are a number of reasons for variability in data taken from multiple wells and for variability in data collect from individual wells, for example seasonality and other temporal effects for within-well samples, and spatial variability from multiple wells. However, spatial variability is often large relative to temporal and analytical variability. Often ANOVA will conclude that the ratio of between-well variability to within-well variability is significant and the hypothesis of equal means will be rejected.

An F statistic greater than the tabulated critical value (based on degrees of freedom for between-well and within-well samples), indicates that the means are not equal. In that case, a follow-up test is needed to determine which mean is outside expectations.

Nonparametric Two-Sample Tests

Wilcoxon rank sum test (Mann-Whitney U-test)

- This test compares two populations using ranking methods when nondetects are present but have a common reporting limit.
- This test can accommodate a limited number of nondetects (typically 10% to 15%) in the data sets with a single reporting limit.
- The test assumes equal population variances.
- The test assumes the two data sets share a common, though unknown, distribution.
- A minimum of 8 to 10 samples are recommended.

Using this test and interpreting results

Pooled background data from one or more wells are compared to pooled site data by use of

ordered ranking to determine if the medians are equal.

A calculated W-statistic greater than the critical W-value indicates that the medians of the two data sets are not equal.

Tarone-Ware Two-Sample Test for Censored Data

- This test compares two data sets using ranking methods when nondetects and variable reporting limits are present in the data sets.
- Nondetect data with multiple reporting limits are acceptable.
- This is a nonparametric test.
- The test assumes the two populations have equal variances.
- This test assumes samples are spatially and temporally independent.
- A minimum of 8 to 10 samples are recommended.

Using this test and interpreting results

Pooled background data from one or more wells are compared to pooled site data by use of ordered ranking to determine if the difference in ranking between the two data sets is greater than that which would have occurred had the ordering occurred by chance.

A Tarone-Ware statistic (TW), greater than the tabulated critical value corresponding to the desired level of confidence, indicates that the test data are significantly different from the background data.

Nonparametric Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric counterpart to ANOVA that does not require normality of the ANOVA residuals (see Chapter 17.1.2, Unified Guidance and Section 5.8.2). In using this test, the interpretation is similar to the parametric F-test.

Related Study Question

Study Question 1: What are the background concentrations?

Key Words: Background, Compliance Monitoring, Interwell, Intrawell, Concentration Comparisons, Release Detection, Site Characterization, Monitoring, Closure

C.3 Study Question 3: Are concentrations above or below a criterion?

To ensure a valid test, it is important to understand how the criterion used for comparison was derived and based on that understanding, to have a well-defined null hypothesis. The criterion can be an MCL, a risk-based value or fixed background limit and may represent a single regulatory value, the mean of a population, or a percentile; therefore, when defining the null hypothesis (and ultimately the comparison method), it is important to take this into account to ensure selection of a test that reflects the intent of the criterion. For example, if the criterion is a not-to-exceed value, individual sample results can be compared to it in much the same manner as is done with prediction limits; alternatively, if the criterion is derived to represent an average concentration ceiling, an upper confidence limit around the mean of the compliance data is the appropriate test statistic.

This question is relevant during release detection, site characterization, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Examine the site data set to determine if you are going to compare either intrawell or interwell data to a criterion (see Section 3.6.5). Intrawell comparisons are most common. If interwell data are going to be used, ensure that the sample data share the same hydrogeologic and geochemical characteristics before combining these data, and test for significant spatial variability. In either case, examine the site data to determine what distributional assumption should inform selection of statistical tests (see Section 4.3.1: Physical Site Conditions and Section 4.2.1: Background Conditions. Refer to Section 3.4: Common Statistical Assumptions for further discussion concerning how the following requirements may impact statistical analysis results.

- Use box plots, probability plots, Dixon's test, or Rosner's test to check for outliers.
- Check that mean and variance are stable over the time frame (time series plot).
- No autocorrelation should exist between successive sampling events.
- Check that no significant trends exist (time series plot). If the data set exhibits significant trends, it may be appropriate to select a subset of the data to representing current concentrations.
- Determine distribution of the data (for example, normal, lognormal) (skewness coefficient, Shapiro-Wilk test, censored probability plots).
- Estimate the mean and standard deviation of left-censored sample using Kaplan-Meier when 50% or less of the data set is nondetect.
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools

There are two broad approaches for analyzing well data and answering the question as to whether chemical concentrations are above a criterion. These two approaches are comparison of pooled interwell compliance data to the criterion and comparison of intrawell compliance data to the criterion.

The statistical tests most commonly used are confidence intervals or limits, tolerance limits, prediction limits and one sample t-test. Confidence intervals are constructed around a statistic of interest (for example, mean, median, certain percentile) while prediction and tolerance limits are extreme values beyond which only represent a small portion of the data population. The one sample t-test compares a statistic of interest from the data to a criterion based on the same statistic of interest derived from the background. Site-specific considerations or regulatory requirements usually determine which parameters and tests are appropriate.

Limits are most often used to compare sampling data to a fixed criterion. There are two questions that can be asked. One question is whether the groundwater concentration of a specific chemical has exceeded a criterion, while the other question is whether the groundwater concentration of a particular chemical has fallen below a criterion. In determining if a criterion has been exceeded, the lower confidence limit is of primary interest. But the upper confidence limit, tolerance limit, or prediction limit are most important in determining if the concentration has fallen below a criterion.

As an example of limit selection, if the criterion being used is a health-based concentration, and the mean exposure should not exceed the criterion, then select a predetermined confidence that the upper confidence limit on the mean (UCL) is below the standard. Likewise, if you are examining groundwater data which has historically been above a criterion, you want the UCL to be below the standard. The scenario is different when assuming that the well being monitored is not contaminated. In this case, retain the assumption until the lower confidence limit is above the criterion.

If the fixed criterion is an average concentration, the appropriate statistical parameter to compare to is the mean or median concentration from site data by use of either a confidence interval or a one-sided t-test.

Parametric Confidence Intervals

Confidence intervals can be calculated for normal, lognormal or nonparametric distributions (see next section) using the methods below:

- confidence interval around a mean (see Section 5.2.2., Section 5.2.3, and Section 5.2.4).
- confidence interval around an upper percentile (see Section 5.2.5).
- robust confidence interval around a mean to modify the nonrobust calculations so that outlying observations in a data-set can be accommodated. (USEPA 1999).

Using this test and interpreting results

Data must be normal or capable of being transformed so that they are normal.

- Use lognormal methods when the underlying population is heavily right-skewed, meaning that a majority of lower concentration data are combined with fewer but much higher concentration data. When the data are transformed, the data should become reasonably symmetric about the mean or normally distributed (check with skewness coefficient, Shapiro-Wilk Test, probability plot).

- Some nondetects are acceptable. Use simple substitution for nondetect are approximately 10-15%. If nondetects are less than or equal to 50%, create a censored probability plot to check for normality.
- The parametric methods depend on t values from a student's t-table, small numbers of samples will correspond to large t values which in turn will make the intervals wide and the corresponding limits extreme; therefore, a minimum of 8 samples is recommended.

Nonparametric Confidence Interval

Nonparametric confidence intervals can be calculated for non-normal data and data which cannot reasonably be transformed so as to become normally distributed. They can also be used when the data set contains a high number of nondetects. Use of nonparametric confidence intervals in determining if a criterion has been exceeded is similar to the parametric confidence interval. As with parametric confidence intervals the assumption is that like parameters are being compared, for example, median to median. When data are ranked using nonparametric methods, it is relatively simple to estimate percentiles in which the data fall; but it is more difficult to estimate parameters such a mean and variance. Thus, nonparametric confidence intervals are built around medians or 50th percentile as opposed to means.

Using this test and interpreting results

For data sets that do not fit a normal or lognormal distribution.

- May be useful when the data includes a large number of nondetects
- Nonparametric confidence intervals are typically built around a median but can also be estimated at other percentiles such as a 90th or 95th.
- Confidence intervals for small data sets may too large to be useful, so the number of samples you need may be greater than typically needed using parametric methods.

Tolerance Limits

When a fixed criterion is an upper percentile or maximum, and no more than a small specified fraction of the individual concentration measurements should exceed the limit, a tolerance limit is a possible appropriate statistic. As with confidence limits, a tolerance limit is one side of a tolerance interval. Tolerance limits, as with confidence limits, may be calculated based on either parametric or nonparametric assumptions.

Using the tolerance limit for testing, you can state that, "I'm 95 percent confident that a particular tolerance interval brackets some percentage, say 99 percent, of the population." Similarly, for the upper tolerance limit (UTL) you could say that "I'm 95 percent confident that 99 percent of all data will be less than the UTL." Note that this statement is independent of the specific number of future samples, and this is what contrasts tolerance limits with predictions limits.

It may be useful to also note that there is no difference between a 95 percent confidence on the upper 95th percentile and an upper tolerance limit on the 95th percentile at 95% confidence.

Using this test and interpreting results

- Tolerance limits (TLs) are designed to contain a large fraction or coverage of the data set (typically 90%, 95% or 99%) with a specified level of confidence (typically 95%, see Section 5.3)
- TLs can be used in lieu of PLs or combined with PLs for re-testing to control false negatives.
- Individual compliance well samples (interwell) or samples collected from multiple wells (interwell) can be used to calculate tolerance limits
- Test can be constructed to examine a single compliance well using a single sample.
- Tolerance limits can be calculated based on either parametric or nonparametric assumptions.

When you must determine if a criterion has been exceeded, the lower tolerance limit (LTL) is compared to the criterion. If the LTL is greater than the criterion, then you can conclude that the data are higher than the criterion at the confidence level used to calculate the LTL. Similarly, if you are trying to determine whether data have fallen below a criterion, the UTL is compared to the criterion. If the UTL is below the criterion then you can conclude that the data are below the criterion. But note, tolerance limits by definition do not cover 100% of the data. Therefore, use of tolerance limits for decision making must incorporate an acceptable failure rate and a plan for retesting.

Prediction Limits

Prediction limits (PLs) estimate an interval in which future observations will fall, with a defined probability, given the collected data. The calculation of PLs takes into consideration the number of future data to be compared, as well as the number of retests required to confirm a release.

As the number of chemicals increase, and the number of resampling instances increases, the upper prediction limit also increases. A corresponding decrease will occur in the power of the test (the probability of missing a true exceedance). To reduce this source of error, limit the number of chemicals examined.

Using this test and interpreting results

Typically, background data are collected and PLs are developed for that data. A set number of future site samples are then compared to the PLs (see Study Question 2). When a criterion is based on an upper percentile or is a maximum, it is possible to develop PLs around site data and then ask if the upper prediction limit has exceeded the criterion.

- PLs are typically projected around means or median (Section 5.4).
- While TLs permit a specified percent of statistical failures, (false negatives); PLs are designed with the intent of no statistical failures.
- An upper prediction limit represents a level that is predicted to equal or exceed future sample values based on past results.
- The number of future samples must be specified.

- The confidence level of a prediction limit represents the probability that a specified number of future samples drawn from the same population will be below the prediction limit.
- Prediction limits increase (or if viewed graphically "widen") as the number of testing events are increased into the future.
- Test can be constructed to examine a single compliance well using a single sample.
- Interwell or intrawell compliance data can be used to construct PLs.

To determine if site data have fallen below a criterion, the upper prediction limit (UPL) may be used. If the UPL exceeds the criterion then you have an indication that the data set used to calculate the UPL may not be consistent with the criterion. Resampling to verify the result is appropriate.

One sample t-test

The one sample t-test compares a statistic of interest (generally the mean) from the data to a criterion representing the same statistic from the background population. The test can be used on either interwell data or intrawell data. It is a parametric test.

Using this method and interpreting results

All the assumptions which apply to the two sample t-test apply to the one sample t-test (see Section 5.12 and Study Question 2).

A calculated t-statistic greater than the critical t-value indicates a statistically-significant difference between the statistic of interest of the two data sets; this difference indicates that the statistic of interest of the site data is greater than the criterion. Typically the statistic of interest is the average (mean) of the site data and the criterion for the average concentrations. A significance level indicating the chance that the test will return an incorrect result and the size of the data set will be used to determine the critical t-value.

Interpretation of Results and Associated Uncertainty

In selecting the statistical method, understand what the groundwater criterion represents and the consequences of exceeding that criterion. The statistical methods selected and interpretation of their results may vary depending the null hypothesis selected (for example, site data are above the criterion or site data are below the criterion). When using a risk-based criterion or background, typically the UCL of the mean or median concentration is compared to the criterion.

Closure determination is supported only when the entire confidence interval (UCL) is below the criterion. Small sample size can result in a wide confidence interval, such that the interval is not useful in identifying a difference. In such cases, additional samples will need to be collected to increase sample size to narrow the interval. Chapter 21 and Chapter 22 of the Unified Guidance provide additional information regarding use of confidence intervals in monitoring for compliance and closure.

See also Section 4.2.4: Statistical Methods for Release Detection Objectives, Section 4.6.1: Compliance with Criteria, and Section 5.13: Control Charts.

Related Study Questions

Study Question 4: When will contaminant concentrations reach a criterion?

Study Question 5: Is there a trend in contaminant concentrations?

Key Words: Compliance, Comparison to Standards, Release Detection, Site Characterization, Monitoring, Closure, Target Levels

References

USEPA. 1999. "Robust Statistical Intervals for Performance Evaluations." In. Las Vegas, NV: Office of Research and Development.

C.4 Study Question 4: When will contaminant concentrations reach a criterion?

This question, associated with projecting future contaminant concentrations, is closely related to Study Question 5 and Study Question 7 regarding trends and attenuation rates. The attenuation rate determined for a chemical in a monitoring well (or for a data set representative of a group of monitoring wells) is useful for understanding how quickly concentrations are changing over time. The attenuation rate, estimated from existing monitoring data, can be used to predict concentrations in the future. The methods used to estimate how long it would take to reach a criterion could also be used to project concentrations at some future time.

This question is usually relevant in the remediation, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Verify that the data set can support trend analyses and modeling. Refer to Section 3.4: Common Statistical Assumptions for further discussion of how the following requirements may impact statistical analysis results.

- Check for outliers using box plots, probability plots, Dixon's test, or Rosner's test.
- Check for autocorrelation between successive sampling events.
- Verify that significant temporal trends using time series plots.
- Ability to detect trends can be impacted by aggregating data across wells.
- In general, you can obtain better detection of trends using longer records of data, but in many cases, attenuation rates will differ based in remedial methods.
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools for this Question

Estimating concentrations at a future time involves constructing a statistical model of chemical concentrations over time. Such models can reflect linear or nonlinear trends. These statistical models are closely related to attenuation rates and can be estimated by linear regression analysis (parametric) or a Theil-Sen trend line (nonparametric).

Linear Regression

- Linear regression assumes a normal distribution for the residuals (that is, the variability not associated with the long-term trend is normally distributed). When this assumption is not satisfied, the accuracy of the results is reduced.
- Regression is sensitive to outliers.
- Regression as a general tool provides flexible ways to develop models for your data. You may transform the data to be normally distributed using a log or other type of data transformation. In addition, regression can be used with a linear model, exponential model, or a

multivariate model that includes multiple factors such as water table elevation in addition to time.

Using this test and interpreting results

Regression is an easy procedure to apply and shows the relationship of pairs of data (time and concentration) to obtain a fit to a model (such as for linear regression, the slope and intercept of a line). A best estimate of the first-order attenuation rate (k) can be obtained by fitting a first-order decay model ($C_t=C_0e^{-kt}$) to the concentration versus time data or by fitting a linear model for natural log concentration versus time data ($\ln(C_t) = \ln(C_0) - kt$). Many software packages also provide a 95% confidence interval for the slope of the model or the attenuation rate in the form described above. This confidence interval is useful for evaluating the uncertainty associated with the estimated attenuation rate. An example of regression applied to groundwater data is included in Appendix A.6.

Theil-Sen Trend Line

- This method does not require a normal distribution for the residuals.
- Theil-Sen line analysis is less sensitive than regression analysis to outliers or extreme values
- This method can only be used to evaluate linear trends. However, a first-order attenuation rate can be estimated by analyzing natural log concentration versus time data.

Using this test and interpreting results

When the Theil-Sen trend line is used for a data set of natural log concentration versus time, the estimated slope is the negative of the estimated first-order attenuation rate with units of time^{-1}. In other words, if the slope is -0.25 and the time units for the data set is in years, then the estimated attenuation rate is 0.25 yr^{-1}. Many software packages also provide a 95% confidence interval for the attenuation rate. This confidence interval is useful for evaluating the uncertainty associated with the estimated attenuation rate.

Interpretation of Results and Associated Uncertainty

Any prediction of future concentrations that is made using an attenuation rate estimated from past data implicitly relies on several assumptions. The key assumptions include:

- Future site conditions will be the same as past conditions (same remedy, same groundwater flow conditions).
- The attenuation rate is determined using an appropriate model. For example, if a linear model was used to determine the attenuation rates, then the attenuation is assumed to be monotonically decreasing along a straight line.

For most sites, it is unlikely that these assumptions will be completely satisfied. For example, matrix diffusion effects may cause the attenuation to deviate from first order. In this case, the range of future concentrations or cleanup times calculated from the 95% confidence interval of the attenuation rate should not be considered a true 95% confidence interval for the prediction. Instead, the

calculated future concentrations and cleanup times are reasonable estimates based on the available data. The predictions should be interpreted in the context of the complete conceptual site model (CSM).

For any case where the 95% confidence interval for the attenuation rate includes zero (meaning the difference between the attenuation rate and zero is not statistically significant), any predictions made using the attenuation rate are highly uncertain.

See also Study Question 7,Section 4.5.1: Monitoring for Concentration Changes, Section 4.6.2: Trends Toward Compliance Criteriaand Section 5.9: Time Series Forecasting.

Related Study Questions

Study Question 3: Are concentrations above or below a criterion?

Study Question 5: Is there a trend in contaminant concentrations?

Study Question 7: What are the contaminant attenuation rates in wells?

Key Words: Cleanup Time, Concentration trends, Attenuation Rate, Remediation, Monitoring, Closure

C.5 Study Question 5. Is there a trend in contaminant concentrations?

Whether concentrations are increasing, decreasing, periodic, or stable over time is a question that generally requires analysis beyond simple graphical methods, especially when data fluctuate or exhibit high variability. The tests described below for general trend testing are closely related to the tests used for season or period trend analyses or for calculating attenuation rates (Study Question 6, Study Question 7). Statistical trend tests can be used as a diagnostic tool to determine if the mean of the population is stationary to qualify the use of the distribution for many other statistical tests. Trend tests can also be used to demonstrate decreases in contaminant concentrations over time. Temporal trend analysis of groundwater monitoring results often reveals differences in results between wells. Even at sites with overall decreasing chemical concentrations, the trend analysis often identifies some wells with statistically significant decreasing concentrations, some wells with decreasing concentrations that are not statistically significant, and some wells with increasing concentrations.

This question is usually relevant in the remediation, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Verify that the data set can support trend analyses and modeling. Refer to Section 3.4: Common Statistical Assumptions for further discussion of how the following requirements may impact statistical analysis results.

- Check for outliers using box plots, probability plots, Dixon's test, or Rosner's test.
- Check for autocorrelation between successive sampling events.
- Ability to detect trends can be impacted by pooling data across wells.
- In general, longer records of data are better at detecting trends but in many cases trends will differ based on remedial methods.
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools

To determine if there is a temporal change or pattern to the data, first use simple graphical techniques to observe significant trends. However, if cyclical effects complicate the pattern of the data consider other statistical methods to answer this study question. The statistical methods described below focus on the monotonic trends, as well as systematic variation in a temporal setting.

Time Series Plots

- These plots show concentration on the y-axis versus time on the x-axis.
- You must assign values to nondetects for these plots.

Using this test and interpreting results

Use this test to visualize temporal changes of concentrations of a single chemical. Data from multiple time periods can show existing patterns in the data. By combining multiple sample points, the temporal patterns usually show up as parallel traces. The spread of the lines would indicate the dispersion of the data over time. Qualitative comparison of the slopes can identify individual trends between wells.

ANOVA

- Data must follow a normal distribution and have a constant variance or spread of sample results.
- You must assign values to nondetects.

Using this test and interpreting results

This test identifies temporal differences among sample periods or nonstationarity (change in means over time). Evaluate the temporal effects of individual sampling events or cyclical event (season) by grouping concentrations across monitoring wells for each sampling date, or season.The one-way ANOVA for temporal effects can formally identify cyclical or nonmonotonic trends.

Significant cyclical variation usually tends to inflate the estimate of the current population variance. If the test identifies a significant temporal effect, the data set can be adjusted to account for seasonality or other cyclical patterns. Study Question 6 addresses how to identify and correct for seasonality or other periodic changes in concentrations.

Spearman's Test

- This nonparametric test does not require that data are derived from a particular statistical distribution.
- Spearman's test is not influenced by outliers or extreme values.
- This test is not appropriate for data sets with a large number nondetects.

Using this test and interpreting results

This test provides information on the direction of the trend (increasing or decreasing) and whether or not the trend is significant. This test is closely related to Pearson’s test discussed below. This test is not recommended if there are seasonal or periodic fluctuations in concentrations. Study Question 6 addresses how to identify and correct for seasonality or other periodic changes in concentrations. If there is autocorrelation among successive sampling events then the samples are not independent and the degrees of freedom for evaluating statistical significance are overstated.

Mann-Kendall Trend Test

- This nonparametric test does not require that data are derived from a particular statistical distribution.

- This test is not influenced by outliers or extreme values.
- This test is not appropriate for data sets with a large number of nondetects.

Using this test and interpreting results

This test identifies significant changes in the mean over time using at least eight sample points (not assuming a particular distribution). The slope of the concentration over time is monotonic. Trends are significant when the absolute value of the S statistic is greater than the critical value. This result indicates that the mean is not stationary over time at this sampling point. The total of the pair differences (S statistic) results in a large negative or positive value for decreasing or increasing trends, respectively. The size of the S statistic is not a measure of the magnitude of the slope. You can calculate the monotonic trends of concentrations over time at a single point to identify statistically significant concentration trends.

Theil-Sen Trend Line

- This test identifies the slope of the trend line.
- Confidence limits can be calculated around the slope of the trend line to ascertain whether the trend is statistically significant.
- Like the Mann-Kendall, this test does not require a normal distribution and can handle extreme values.
- This test is not appropriate for data sets with a large number of nondetects.

Using this test and interpreting results

This test estimates the slope of a trend line, which can be used to predict the mean concentration at some point in time. If the slope is positive and upper and lower confidence limits around the slope are also positive, the test indicates that there is a statistically significant increasing trend, and the opposite is true for a negative slope and confidence limits. If the upper confidence limit is positive and the lower is negative, there is insufficient evidence to indicate a statistically significant trend.

Pearson's Test

- This test requires a normal distribution.
- This test is sensitive to outliers or extreme values.
- This test requires a constant dispersion of the data over sampling events.

Using this test and interpreting results

This test provides information on the direction of the trend (increasing or decreasing) and whether or not this trend is significant. This test is closely related to the Spearman's test discussed above. This test is not recommended if there are seasonal or periodic fluctuations in concentrations. Study Question 6 addresses how to identify and correct for seasonality or other periodic changes in concentrations. If there is autocorrelation among successive sampling events then the samples are not independent and the degrees of freedom for evaluating statistical significance are overstated.

Linear Regression

- This test requires a normal distribution.
- This test is sensitive to outliers.
- This test requires constant dispersion of the data.

Using this test and interpreting results

Regression is an easy procedure to apply and shows the relationship of pairs of data (time and concentration) to obtain the slope and intercept of a line. To apply a linear regression appropriately, the relationship must be linear (monotonic increasing or decreasing). When the slope is not zero, either a positive or negative trend is present. The statistical significance of the trend is identified by the slope being significantly different from zero.

Interpretation of Results and Associated Uncertainty

Temporal trend analyses show whether and how chemical concentrations are changing over time. As discussed above, nonparametric tests, such as the Mann-Kendall test and Theil-Sen trend line, do not require assumptions regarding the data distribution. In contrast, linear regression analysis requires an assumption regarding the pattern of change over time (such as monotonically decreasing). Additionally, parametric regression analysis assumes that the variability not associated with the temporal trend is normally-distributed. If the required assumptions are not satisfied, then the accuracy of the regression analysis is reduced. However, if the assumptions are satisfied, regression analysis will be more accurate than the Mann-Kendall test because the regression analysis uses the information concerning data distribution as part of the test.

If the p-value is less than 0.05, then typically the change in concentration over time is statistically significant. A statistically-significant trend depends on a number of factors including the length of the monitoring record and the magnitude of variability not associated with the long-term trend relative to the magnitude of the long-term trend. In data sets with high variability, longer monitoring records are needed to identify statistically significant trends. See also Section 4.5.1: Monitoring for Concentration Changes and Section 4.6.2: Trends Toward Compliance Criteria.

Related Study Questions

Study Question 3: Are concentrations above or below a criterion?

Study Question 6: Is there seasonality in the concentrations?

Study Question 7: What are the contaminant attenuation rates in wells?

Key Words: Temporal Trends, Remediation, Monitoring, Closure

C.6 Study Question 6: Is there seasonality in the concentrations?

Most statistical tests assume statistical independence of the sample data. Temporally dependent groundwater data violate the assumption of independence. High levels of variability unrelated to the long-term temporal trend also make it difficult to identify statistically significant long-term trends and to estimate attenuation rates and remediation time frames (McHugh et al. 2011).

When temporal variability exists because of the distribution of the timing of the sample collection, the distribution exhibits time dependence or autocorrelation (for example, a cyclical pattern of data affected by the seasons). To verify statistical independence, demonstrate a low correlation between the concentration and the time of the sampling event. Evaluate the cyclical change (seasonal variation) to adequately understand the variance of the population. A cyclical pattern can bias the variance of the distribution or create a slope of the concentration that is not monotonic. You can evaluate the temporal evaluation and adjust the distribution to evaluate the trend accordingly.

This question is usually relevant in the remediation, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Refer to Section 3.4: Common Statistical Assumptions for further discussion of how the following requirements may affect statistical analysis results.

- Check for outliers using box plots, probability plots, Dixon's test, or Rosner's test.
- Check for autocorrelation between seasonal sampling events.
- The ability to detect trends can be impacted by aggregating data across wells.
- In general, you can obtain better detection of trends using longer records of data, but in many cases, attenuation rates will differ based on remedial methods.
- See also Section 4.1: Considerations for Statistical Analysis.

If the test does not assume a distribution, then no testing of the distribution is necessary. However, if a substantial number of nondetects are present, then the test cannot indicate autocorrelation. The samples should cover multiple years with an observable seasonal pattern each year. Each season should include at least three measurements.

When the objective is to determine if there is a temporal change or pattern to the data, simple graphical procedures can reveal significant trends. However, if cyclical effects complicate the pattern of the data then consider other statistical methods listed below to answer this study question.

Statistical Methods and Tools

Determine if there is a significant cyclical pattern in the data that creates autocorrelation between the samples; statistical independence of the data is a key assumption for many statistical tests. When the objective is to determine if there is a temporal change or pattern to single series data, use

the sample autocorrelation function or Rank von Neumann ratio test to identify correlated samples from specific seasons. When the objective is to determine if there is a temporal change or pattern to a group of wells, use time series plots or the Kruskal-Wallis test to identify correlated samples related to specific seasons.

Sample Autocorrelation Function

- Data must follow normal distribution for this test.
- A minimum of 8 to 10 measurements are recommended, although a greater number of measurements may be necessary to obtain the desired confidence level.
- This test is very sensitive to extreme values (outliers).

Using this test and interpreting results

Over several measurements, calculate an autocorrelation coefficient. If any coefficient exceeds the critical value, then assume samples are dependent.

Plot the autocorrelation coefficient over the number of lags overlaid (plus or minus the critical value) to identify the significance of the dependent nature of the samples. If the shape of the function is sinusoidal, then the data exhibit seasonal fluctuation. Adjust the values for seasonality. If seasonality occurs, change the frequency of sampling or adjust the data set.

Rank von Neumann Ratio Test

- This test requires no distribution assumptions.
- This test cannot handle a substantial number of tied values or nondetect values.
- A minimum of 10 to 12 observations from a single well is recommended.

Using this test and interpreting results

Since this test does not assume a distribution, no testing of the distribution is necessary. However, a substantial number of nondetects will cause the test to lose its validity for autocorrelation.

Seasonality is one of many reasons for temporal correlation. By evaluating von Neumann ratio and comparing it to a lower critical point, you can identify evidence for temporal correlation at a selected level of significance. Withne's test sufficient evidence of autocorrelation, you must adjust the data to evaluate the trend of the observations. If this occurs, the frequency of sampling should change or adjust the data set.

Time Series Plots

- Standardize the concentration on y-axis versus time on x-axis.
- Assign values to nondetects.
- Plot parallel lines for several wells.

Using this test and interpreting results

The standardized concentration is assigned by subtracting the mean concentration of each well

from the concentration and dividing by the standard deviation. You can visually identify temporal patterns by plotting the standardized concentrations of several wells over time. Seasonal fluctuations will show up in the time series as parallel traces. Since this is a qualitative method, adjust for seasonal variations in evaluating the trend.

Kruskal-Wallis Test

- As a nonparametric test, normality of the data is not required.
- At least three groups of data (see Section 5.8.2) must be present.
- If applied to a group of wells, little to no spatial variation should exist.

Using this test and interpreting results

This test identifies temporal differences among sample periods or nonstationarity (change in means over time). Evaluate the temporal effects of individual sampling events or cyclical events (seasons) by aggregating concentrations across monitoring wells for each sampling date, as described above. The Kruskal-Wallis test confirms whether median measurement levels differ by season, thus indicating the presence of seasonality.

If acceptable under the regulatory program, change the sample frequency as a simple remedy for temporal correlation. If the test identifies a significant temporal effect, adjust the data set to account for significant seasonal correlations (seasonality). Otherwise, use specific statistical tests that have a seasonal test method. For example, you can perform a seasonal Mann-Kendall test on each group (season), then combine the S statistics of the groups to calculate the overall S statistic. You can identify a significant trend over time at one location when the absolute value of S is greater than the critical point. This result indicates that the mean is not stationary at this sampling point, despite the seasonal fluctuations.

Interpretation of Results and Associated Uncertainty

The temporal trend analyses explain whether and how contaminant concentrations are changing over time. As discussed above, a nonparametric test such as the seasonal Mann-Kendall test does not require assumptions regarding the data distribution. However, significant seasonal fluctuations in the data will cause false negative errors in the statistical test methods. Temporally dependent and autocorrelated data generally contain both a truly random and nonrandom component. Only strong correlations are likely to affect the results of further statistical testing. See also Section 4.5.1 and Section 4.6.2.

Related Study Questions

Study Question 5: Is there a trend in contaminant concentrations?

Key Words: Temporal Trends, Remediation, Monitoring, Closure, Cyclic or Periodic Change

C.7 Study Question 7: What are the contaminant attenuation rates in wells?

When contaminant concentrations are identified as having a trend over time (see Study Question 5), this question follows to estimate the rate of change over time (attenuation rate). You can then use the attenuation rate to evaluate whether the rate of decrease in concentration is adequate to achieve the site goals. In addition, the attenuation rate can be used to estimate future concentrations including when concentrations will reach a cleanup criterion (see Study Question 4).

When using a data set from a single well or a set of wells in the source area, the attenuation rate determined from the concentration versus time data (sometimes called a "source attenuation rate") represents source depletion (Newell et al. 2002). In contrast, the decrease in concentration over distance downgradient of the source is called a "plume attenuation rate" (see Study Question 8).

This question is usually relevant in the remediation, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Verify that the data set can support trend analyses and modeling. Refer to Section 3.4: Common Statistical Assumptions for further discussion of how the following requirements may impact statistical analysis results.

- Check for outliers using box plots, probability plots, Dixon's test, or Rosner's test.
- Check for autocorrelation between successive sampling events.
- Ability to detect trends can be impacted by pooling data across wells.
- In general, you can obtain better detection of trends using longer records of data, but in many cases attenuation rates will differ based on remedial methods.
- See also Section 4.1: Considerations for Statistical Analysis.

When you evaluate multiple time intervals from a single monitoring record in order to identify changes in attenuation rates, be sure to evaluate the uncertainty in the attenuation rate estimates (the confidence bands) in order to determine whether an apparent difference in attenuation rates is most likely to be associated with true change in the source attenuation rate or an artifact of shorter-term random variability.

Determining whether two attenuation rates are different requires an evaluation of the uncertainty associated with each attenuation rate. The greater the difference in the attenuation rates relative to the uncertainty associated with each rate, the greater the confidence that the observed difference in the attenuation rates is real. See the case example in Appendix A.7.

Statistical Methods and Tools

Attenuation rates can be estimated using parametric or nonparametric methods. Both methods require assumptions about the concentration trend, for example zero order (linear trend over time)

or first order (exponential decay). For groundwater monitoring data, exponential decay is a commonly observed long-term trend (Newell et al. 2002). Therefore, the attenuation rate is usually best represented by a first-order decay rate.

Regression

- Regression assumes a normal distribution for the residuals (the variability not associated with the long-term trend is normally distributed). When this assumption is not satisfied, the accuracy of the results is reduced.
- Regression provides flexibility for the model fit to the data. Regression can be used with a linear model, exponential model, or a multivariate model that includes factors such as water table elevation in addition to time.
- Regression is sensitive to outliers.

Using this test and interpreting results

Regression is an easy procedure to apply and shows the relationship of pairs of data (time and concentration) to obtain a fit to a model (such as for linear regression, the slope and intercept of a line). A best estimate of the first-order attenuation rate (k) can be obtained by fitting a first-order decay model ($C_t=C_0\ e^{-kt}$) to the concentration versus time data or by fitting a linear model for natural log concentration versus time data ($\ln(C_t) = \ln(C_0) - kt$). Many software packages also provide a 95% confidence interval for the slope of the model or the attenuation rate in the form described above. This confidence interval is useful for evaluating the uncertainty associated with the estimated attenuation rate. An example of regression applied to groundwater data is included in Appendix A.6.

Theil-Sen Trend Line

- This test does not require a normal distribution for the residuals.
- This test is less sensitive than regression analysis to outliers or extreme values.
- This test can only be used to evaluate linear trends (however, a first-order attenuation rate can be estimated by analyzing natural log concentration versus time data).

Using this test and interpreting results

When the Theil-Sen trend line is used for a data set of natural log concentration versus time, the estimated slope is the negative of estimated the first-order attenuation rate with units of $time^{-1}$. In other words, if the slope is -0.25 and the time units for the data set is years, then the estimated attenuation rate is 0.25 yr^{-1}. With the use of a bootstrapping method, many software packages also provide a 95% confidence band for the attenuation rate as described in Section 5.2.7 and Chapter 21.3, Unified Guidance. This confidence band is useful for evaluating the uncertainty associated with the estimated attenuation rate.

When comparing the attenuation rates for two different wells (or two data sets that each represent one or more wells) if the confidence bands for the attenuation rates do not overlap, then you can conclude with reasonable certainty that the attenuation rates are different. If the confidence bands

do overlap, then you cannot conclude with confidence that the attenuation rates are different. Two examples of comparing attenuation rates are presented in Appendix A.7.

Interpretation of Results and Associated Uncertainty

When evaluating temporal trends in groundwater monitoring results, differences in results between wells are often present. Even at sites with overall decreasing contaminant concentrations, the trend analysis can identify some wells with statistically significant decreasing concentrations, some wells with apparently decreasing concentrations that are not statistically significant, and some wells with apparently increasing concentrations. In many cases, the apparent differences in concentration trends between wells can be attributed to random variability in the monitoring data rather than real differences in attenuation rates between wells.

A key challenge in the evaluation of concentration trends for multiple wells is determining whether these differences are due to random variability in monitoring results or due to true differences in attenuation between wells. This determination should be based on lines of evidence such as:

- Are the differences in attenuation rate statistically significant? If the 95% confidence bands for the attenuation rates overlap, then the difference is not significant.
- Does the variation in attenuation rates exhibit a spatial pattern? In other words, are wells with increasing concentrations or slower attenuation clustered together? Are wells with faster attenuation clustered together?
- Is there a potential mechanistic explanation for the observed differences? For example, are the wells with faster attenuation rates located closer to an active remediation system? Or are the wells with slower attenuation rates screened in lower permeability soils?

Unless a majority of the lines of evidence suggest true differences in attenuation rates between wells, then it is likely that the observed differences are due to random variation. For a group of wells without clear differences in attenuation rates, the best estimate of the overall plume attenuation rate can be obtained by either using a midpoint attenuation factor such as the average or median attenuation factor or evaluating the attenuation rate for groundwater concentrations that are representative of the group of wells (such as the average, median, or maximum concentration for each monitoring period).

When comparing attenuation rates, the evaluation of whether the attenuation rates are different is not an absolute yes or no. The results should influence the level of confidence in the conclusion. If the confidence bands for two attenuation rates almost overlap, then confidence in the conclusion that the attenuation rates are different should be lower than if the confidence bands are separated by a greater distance.

The 95% confidence band for the attenuation rate reflects the uncertainty associated with the estimate of the attenuation rate based on the variability in monitoring record that is not explained by the long-term trend. However, other sources of uncertainty may not be captured by the statistical analysis. For example, if a monitoring record was collected during an extended drought when the

water table was dropping over time, then you may have less confidence that the observed rate of attenuation would continue during a subsequent period of normal precipitation. For an evaluation of whether two attenuation rates are different, consider the statistical analyses discussed, the complete conceptual site model (CSM), and any other available information that may be relevant to the determination.

See also Study Question 7, Section 4.5.1: Monitoring for Concentration Changes, and Section 4.6.2: Trends Toward Compliance Criteria.

Related Study Questions.

Study Question 8: How do contaminant concentrations change with distance from the source area?

Key Words: Attenuation, Contaminant attenuation rate, changing contaminant concentrations, Remediation, Monitoring, Closure

C.8 Study Question 8: How do contaminant concentrations change with distance from the source area?

This question continues the progression of Study Question 7. Study Question 7 characterizes the temporal trends at different wells. Study Question 8 expands the comparison to the overall mass of a plume as you move away from the source area. As discussed in Study Question 7, comparing concentration trends in different wells is useful for evaluating the existence of spatial differences in attenuation rates and if concentrations are changing as you move further from the source area at a specific time. However, the question often involves the distance the plumes extends, and the overall behavior of the plume. If the slopes in concentration at each well are not statistically different, then the magnitude of those trend lines represents a consistent range of the plume attenuation rate. The rate of change of concentrations over distances is calculated by concentration difference over the distance. If there is not homogeneity of trends of wells along the centerline of the plume, then quantifying the decrease of contaminant concentrations as the plume moves downgradient of the source area becomes a complex spatial statistical problem. Comprehensive spatial analysis of change in concentration requires geostatistics, which is beyond the scope of this document.

The use of mass flux across an aquifer, however, is a better method to estimate the characteristics of the plume (Feenstra, Cherry, and Parker 1996). The uncertainty of the spatial variations of concentrations are handled by quantifying the mass discharge at transects of a plume as described in the ITRC document *Use and Measurement of Mass Flux and Mass Discharge* (ITRC 2010). This method appropriately gives manageable values of the mass discharge for an area of the plume. Then by evaluating the mass discharge between transects of a plume, you can understand the trend of the contaminant concentration over distance. The attenuation rate of the plume is the slope of the differences in the mass discharge between transects.

Comparing the trends in mass discharge across transects of a plume as you move further from the source area is similar to comparing the attenuation rates between wells (see Study Question 7). Determining whether two attenuation rates are different requires an evaluation of the uncertainty associated with each attenuation rate. If the slopes to the mass discharge trends of a plume are not statistically different as contamination moves away from the source area, then the slope of the trend lines will represent the range of the attenuation rate for the overall plume.

This question is usually relevant in the remediation, monitoring, and closure stages of the project life cycle.

Selecting and Characterizing the Data Set

Verify that the data set can support trend analyses and modeling. Refer to Section 3.4 for further discussion of how the following requirements may impact statistical analysis results.

- Check for outliers using box plots, probability plots, Dixon's Test, and Rosner's Test.
- Check for autocorrelation between successive sampling events.

- Ability to detect trends can be impacted by pooling data across wells.
- In general, you can obtain better detection of trends using longer records of data, but in many cases attenuation rates will differ based on remedial methods.
- See also Section 4.1: Considerations for Statistical Analysis.

When you evaluate multiple time intervals from a single monitoring record in order to identify changes in attenuation rates, be sure to evaluate the uncertainty in the attenuation rate estimates (the confidence bands) in order to determine whether an apparent difference in attenuation rates is most likely to be associated with true change in the source attenuation rate or an artifact of shorter-term random variability.

ITRC 2010 discusses the selection of points to define the transect and the calculation of the mass discharge. Each transect becomes a unique data point along the longitudinal axis of the plume, instead of looking at one location over time. This is a data intensive process, where at least four transects are needed to calculate the slope over the distance between transects.

Statistical Methods and Tools

The statistical methods of estimating the attenuation rates from monitoring wells along the center-line can be done by using parametric or nonparametric methods. Both methods require assumptions about the concentration trend for example zero order (linear trend over time) or first order (exponential decay). See Study Question 7 for details.

Linear Regression

- Regression is an easy procedure that shows the relationship between two variables (distance and mass discharge).
- The slope of the trend line is an estimate of the change in the mean of the mass discharges over distance between transects.
- To apply a linear regression appropriately, the relationship must be linear (monotonic and noncyclical trends to the mass discharge).

Using this test and interpreting results

Regression is an easy procedure to apply and shows the relationship of pairs of data (time and concentration) to obtain a fit to a model (such as for linear regression, the slope and intercept of a line). A best estimate of the first-order attenuation rate (k) can be obtained by fitting a first-order decay model ($C_t=C_0\ e^{-kt}$) to the concentration versus time data or by fitting a linear model for natural log concentration versus time data ($\ln(C_t) = \ln(C_0)-kt$). With the use of a bootstrapping method, many software packages also provide a 95% confidence band for the attenuation rate as described in Section 5.5.3 and Chapter 21.3.1, Unified Guidance. This confidence interval is useful for evaluating the uncertainty associated with the estimated attenuation rate.

Theil-Sen Trend Line

- The slope of the trend line is a measurement of the change in the median of the mass discharges over distance between transects.
- The trend line is constructed by combining the median pair-wise slope with the median mass discharge and distance between transects.

Using this test and interpreting results

When the Theil-Sen trend line is used for a data set of natural log concentration versus time, the estimated slope is the negative of estimated the first-order attenuation rate with units of time^{-1}. In other words, if the slope is -0.25 and the time units for the data set is years, then the estimated attenuation rate is 0.25 yr^{-1}. With the use of a bootstrapping method, many software packages also provide a 95% confidence band for the attenuation rate as described in Section 5.5.3 and Chapter 21.3.1, Unified Guidance.

Interpretation of Results and Associated Uncertainty

You can estimate the change in concentration over time (attenuation rate). Regression analysis (parametric) or Theil-Sen trend line (nonparametric) can be used to estimate attenuation rates. Then, by calculating a confidence interval around the median at a point in time, you can generate a nonparametric confidence band around the attenuation rate with the use of bootstrapping (see Section 5.5.3 and Chapter 21.3.1, Unified Guidance). This approach provides an estimate of the uncertainty you have with the magnitude of the slope.

If the confidence bands do not overlap, then the attenuation rates are statistically different (see Study Question 7 for details). By dividing the concentration difference by the distance between the wells you can estimate the concentration change over distance.

In order to compare the attenuation rates for two different transects, estimate the attenuation rates and provide confidence bands for the attenuation rates. If the confidence bands do not overlap, then you conclude that the attenuation rates are statistically different and you will need addition geostatistics methods to evaluate the overall plume.

The confidence band around trend lines reflects the uncertainty associated with the estimate of the attenuation rate. The uncertainty associated with the attenuation rate depends on a number of factors including the length of the monitoring record and the magnitude of variability not associated with the long-term trend (relative to the magnitude of the long-term trend). In data sets with higher variability and shorter monitoring, records will have more uncertainty (larger confidence bands) compared to data sets with lower variability and longer monitoring records.

When evaluating trends in groundwater monitoring results, differences in results between wells and different plume areas are often present. Even at sites with overall decreasing contaminant concentrations, the trend analysis can identify some wells or plume areas with statistically significant decreasing concentrations, some wells or plume areas with apparently decreasing concentrations

that are not statistically significant, and some wells or plume areas with apparently increasing concentrations. In many cases, the apparent differences in concentration trends between wells or plume areas can be attributed to random variability in the monitoring data rather than real differences in attenuation rates between wells.

A key challenge in the evaluation of concentration trends is determining whether these differences are due to random variability in monitoring results or due to true differences in attenuation. This determination should be based on lines of evidence such as:

- Are the differences in attenuation rate statistically significant? If the 95% confidence bands for the attenuation rates overlap, then the difference is not significant.
- Does the variation in attenuation rates exhibit a spatial pattern? In other words, are wells with increasing concentrations or slower attenuation clustered together? Are wells with faster attenuation clustered together?
- Is there a potential mechanistic explanation for the observed differences? For example, are the wells with faster attenuation rates located closer to an active remediation system? Or are the wells with slower attenuation rates screened in lower permeability soils?

Unless a majority of the lines of evidence suggest true differences in attenuation rates, then it is likely that the observed differences are due to random variation. For wells or plume areas without clear differences in attenuation rates, the best estimate of the overall plume attenuation rate can be obtained by either using a midpoint attenuation factor such as the average or median attenuation factor or evaluating the attenuation rate for groundwater concentrations that are representative of the group of wells (such as the average, median, or maximum concentration for each monitoring period).

An evaluation of whether two attenuation rates are different should include consideration of the statistical analyses discussed, the site conceptual site model (CSM) and any other available information that may be relevant to the determination.

Related Study Questions

Study Question 7: What are the contaminant attenuation rates in wells?

Key Words: Attenuation, Remediation, Monitoring, Closure

References

ITRC. 2010. "Use and Measurement of Mass Flux and Mass Discharge." In MASSFLUX-1. Washington, D.C.: Interstate Technology & Regulatory Council. http://www.itrcweb.org/Guidance/ListDocuments?topicID=14&subTopicID=11.

Feenstra, S., J.A. Cherry, and B.L. Parker. 1996. "Conceptual Models for the Behavior of DNAPLs in the Subsurface." In *Dense Chlorinated Solvents and Other DNAPLs in Groundwater*. Pankow, J.F. and J.A. Cherry, Eds. Portland OR: Waterloo Press.

C.9 Study Question 9: Is the sampling frequency appropriate (temporal optimization)?

Optimization and design of the monitoring program must assure sample independence while covering the site sufficiently and collecting adequate data over an appropriate time period for proposed statistical evaluations. If the monitoring program is in the early stages, statistical design options should be considered such that an adequate number of samples are collected (Section 3.6). For sites with existing long term monitoring data sets, sampling frequency can often be reduced while still providing adequate data for evaluation. The required frequency of sampling can be evaluated with statistical methods that assess whether there is redundancy of sample results for a particular well. You can also apply spatial statistics to evaluate sampling frequency among a set of wells. For an overview of spatial optimization methods, see Study Question 10. For effective optimization, you must establish the goal of the long-term monitoring program and identify an acceptable length of time to determine a change.

This question can be relevant in all stages of the project life cycle: release detection, site characterization, monitoring, remediation, and closure. Although it is more likely that there is enough information to conduct optimization at later stages in the project life cycle.

Selecting and Characterizing the Data Set

Verify that the data set can support optimization techniques. Refer to Section 3.4: Common Statistical Assumptions for further discussion of how the following requirements may impact statistical analysis results.

- Check for outliers using box plots, probability plots, Dixon's Test, and Rosner's Test.
- Check for autocorrelation between successive sampling events.
- Check significant temporal trends using time series plots.
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools

Using the results of the above plots and tests as a guide, you can use more sophisticated statistical methods to evaluate the redundancy of sample results for a particular well. These methods can also be applied to a network of wells. The two approaches highlighted for this question are an iterative thinning analysis or cost effective sampling (CES). There are also some other optimization methods including the modified CES method and genetic algorithms. Be aware that in some cases where the uncertainty is determined to be high, additional sampling may be recommended. See Appendix D for software packages.

Iterative Thinning

- This test analyzes a large data set and trims results.
- If trends or seasonality exist, then the performance metric is based on replicating these temporal trends with the subset of samples.
- If there are no trends, then trimmed data are compared to the stationary summary statistics.

Using this test and interpreting results

- Normality is not a requirement of the method, but you must use the appropriate underlying statistical method (for example, a parametric trend test for data that are derived from a normal statistical distribution).
- Some nondetects are allowed, but use caution in applying this method with frequent nondetects.
- This method can be used for concentrations that are trending with time (time series plots).
- The method works better with more data, so is best applied in later project stages.
- Specify the desired level of confidence for each well.

Cost Effective Sampling (CES)

- Base "the sampling frequency on the changes in concentration at a given well, rather than the well's location with respect to the plume" (Ridley and McQueen 2005; Ridley et al. 1995).
- "CES calculates quantitative measures of the trend and variability of important COCs [chemicals] at each monitoring location and interprets this information by means of decision trees to arrive at a recommended sampling frequency" (Ridley and McQueen 2005).
- "An essential aspect of the CES program has been to use simple statistics within a decision-logic framework to provide information that can be easily understood" (Ridley and McQueen 2005).

Using this test and interpreting results

- Normality is not a requirement of the method and nondetects are not explicitly an issue as long as the trend and variability in the chemical data can be assessed.
- This method can be used for concentrations that are trending with time (time series plots).
- The method works better with more data, so is best applied in later project stages.
- Specify bins of concentration trends and associated variability in the trend for each well and chemical – for example, small trends warrant annual sampling and large trends and variability merit quarterly sampling.

Interpretation of Results and Associated Uncertainty

Groundwater monitoring well network optimization often works best when the network is evaluated as a unit. Therefore, there is greater potential for project benefits when both spatial and tem-

poral information is considered. However, there are cases where a project could benefit by eliminating redundant sampling events or by adding sampling events to reduce uncertainties.

Related Study Questions

Study Question 5: Is there a trend in contaminant concentrations?

Study Question 6: Is there seasonality in the concentrations?

Study Question 10: Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

Key Words: Temporal Concentrations, Optimization, Release Detection, Site Characterization, Remediation, Monitoring, Closure

References

Ridley, M.N., V.M. Johnson, and R.C. Tuckfield. 1995. Cost-Effective Sampling of Groundwater Monitoring Wells. Vol. UCRL-JC-118909. Livermore, CA: Lawrence Livermore National Laboratory.

Ridley, M.N., and D. MacQueen. 2005. A Cost-Effective Sampling of Groundwater Monitoring Wells: A Data Review and Well Frequency Evaluation. UCRL-CONF-209770. Livermore CA:Lawrence Livermore National Laboratory.http://www-erd.llnl.gov/library/CONF-209770.pdf

C.10 Study Question 10: Is the spatial coverage of the monitoring network appropriate (spatial optimization)?

Optimization and design of the monitoring program must assure sample independence while providing adequate spatial coverage of the site. This question addresses how to use statistical methods to optimize the spatial coverage of the site. Optimization can lead to decreasing or increasing the number of wells. The concepts of sufficiency and redundancy are related but different tools are available to determine if existing wells are redundant (that is, wells can be removed from monitoring) if there are sufficient wells (you may either add or remove wells). If the monitoring program is in the early stages, statistical design considerations and site investigation data can be used to establish a well network (Section 3.6). Statistical spatial optimization methods are most applicable for a site with existing large data sets. For an overview of temporal optimization methods, see Study Question 9. For effective optimization, you must establish the goal of the long-term monitoring program and identify an acceptable set of wells to determine a change.

This question can be relevant in all stages of the project life cycle: release detection, site characterization, remediation, monitoring, and closure; it is more likely that enough information exists to conduct optimization at later stages in the project life cycle.

Selecting and Characterizing the Data Set

Verify that the data set can support optimization techniques. Refer to Section 3.4: Common Statistical Assumptions for further discussion of how the following requirements may impact statistical analysis results.

- Check for the presence of outliers using box plots, probability plots, Dixon's test, and Rosner's test.
- Check for autocorrelation between successive sampling events.
- Check for significant temporal trends.
- Verify that the mean and variance are stable over the data set (or subset) time.
- Verify that the data set exhibits normal distribution or normalize it using transformation, determine a suitable method for handling nondetects.
- See also Section 4.1: Considerations for Statistical Analysis.

Statistical Methods and Tools

Using the results of the above plots and tests as a guide, you can use more sophisticated statistical methods to evaluate the redundancy or sufficiency of sample results among wells. The two approaches highlighted for this assessment are the redundancy or spatial uncertainty analyses. Spatial optimization is a challenging objective and an active area of research. Generally these methods require a lot of data and broad spatial coverage of the plume. Optimization results should be checked versus what is known or hypothesized about contamination using the conceptual site model (CSM). Be aware that in some cases uncertainty can be high and additional sampling may be required. See Appendix D for software packages.

Redundancy Analysis

- This test analyzes a large data set and trims wells.
- If trends exist, then the performance metric is based on replicating these spatial trends (for example, maps of contaminant plumes) with the subset of wells.
- If there are no trends, then trimmed data are compared to the stationary summary statistics.

Using this test and interpreting results

- Normality is not required, but some of the methods may be sensitive to highly skewed distributions.
- Some nondetects are allowed, but use caution in applying these methods with frequent nondetects.
- You can use simple qualitative evaluations based on removing a single or multiple wells or more sophisticated analyses like slope factor analyses or genetic algorithms to search for optimal well reduction scenarios. These methods work better with more data, so are best applied in later project stages.

Spatial Uncertainty Analysis

- This test calculates spatial uncertainty for a large data set. Areas of higher uncertainty warrant more wells and areas with lower uncertainty need fewer wells.
- If trends or seasonality exist, then these trends should be eliminated for some of these methods.

Using this test and interpreting results

- These methods typically work best with normally distributed data or data that can be transformed to normality.
- Some nondetects are allowed, but use caution in applying this method with frequent nondetects.
- You cannot use this method when concentrations are trending with time (time series plots).
- These methods works better with more data, so it is best applied in later project stages.
- Spatial optimization methods generally employ geostatistics (for example, kriging). The current locations of wells and the error from the spatial model are used to identify where to place wells to improve estimates of contaminant concentrations. Detailed discussion of geostatistics is beyond the scope of this document.

Interpretation of Results and Associated Uncertainty

Optimizing groundwater monitoring well networks often works best when the network is evaluated as a unit. Therefore, greater potential for project benefits exists when both spatial and temporal information are considered. However, in some cases a project could benefit by eliminating

redundant wells or by adding wells to reduce uncertainties. It is important that optimization be conducted such with regard to and consistent with what is known or hypothesized using the CSM.

Related Study Questions

Study Question 5: Is there a trend in contaminant concentrations?

Study Question 6: Is there seasonality in the concentrations?

Study Question 9: Is the sampling frequency appropriate (temporal optimization)?

Key Words: Optimization, Efficiency, Spatial Coverage, Release Detection, Site Characterization, Remediation, Monitoring, Closure

APPENDIX D. STATISTICAL SOFTWARE TOOLS AND PACKAGES

Several widely available statistical software packages are summarized in this appendix. These packages are included based on responses to the groundwater statistics survey (Appendix E) and on the input of team members. The information provided here is intended as a summary introduction for project managers and not to replace a thorough review of the appropriateness of any software package. In addition, inclusion of software in this appendix is not an endorsement. Many other available statistical packages may also be useful for statistical analyses and evaluations. You must verify the applicability and accuracy of any selected statistical software package prior to use. Also, make sure that you understand the assumptions and input requirements for any statistical tests used in making decisions.

Note that not all of the packages are specifically designed for statistical analyses of environmental data, or specifically of groundwater data. Some are general statistical packages intended for use in different business or scientific application areas. Each package summary identifies the statistical functions that may be applied to groundwater problems.

Table D-1. Software packages included in Appendix D

Software Packages	
D.1 3TMO	D.13 PAM
D.2 CARStat	D.14 Pro-UCL
D.3 ChemStat	D.15 R for Statistics
D.4 DUMPStat	D.16 Sanitas
D.5 Excel	D.17 SAS
D.6 GTS	D.18 Scout
D.7 GWSDAT	D.19 SPSS
D.8 JMP	D.20 Statistica
D.9 MATLAB	D.21 Summit Tools
D.10 MINITAB	D.22 SYSTAT
D.11 MAROS	D.23 VSP
D.12 NCSS	

The included statistical software packages were described and evaluated using a number of categories. While most of the following software package information is self-explanatory, please note the following regarding the evaluation process for each package (these are the package descriptions included as Sections D.1 through D.23):

- **Current version:** Most recent available version at the time of publication. You should check for more up-to-date versions when considering a software package.
- **Ease of Use:** Ranked as Easy, Moderate, or Complex based on team reviews.

- **Statistical Functions:** Statistical software packages were evaluated on their capability to perform various statistical functions used in analysis of groundwater data as outlined below:
- **Capability Ratings:** The following capabilities ratings are included in the statistical functions tables for each software package:

  N/A = Not applicable or not available.

  ● = "Full capability" means that the ITRC team did not identify significant limitations on the identified tests using the software package.

  ◒ = "Some capability" means that some limitations exist for a given package. An example of limited capability might be the ability to map data using only certain geostatistical mapping methods or an inability to edit maps that are produced using the package. Not all users require all capabilities.

  In addition to the "as is" ratings, the ITRC team attempted to identify tests where "as is" capability may be limited or missing, but where some or full capability may exist with the use of add-ins, scripts, or programming. Because team members were not able to evaluate the use of each package for every type of analysis and for every possible situation, you should evaluate possible statistical packages for their particular intended use and, when possible, consult the package developers regarding specific capabilities or limitations.

- **Ease of Use and Data Import:** Packages were evaluated for the target audience for this document (average user). Experienced statisticians may find packages easier to use than indicated; users who are new to statistical software packages may find them slightly more difficult than indicated.
- **Primary Uses:** This section describes the primary identified uses of the statistical software package for groundwater data analysis. Be aware that certain packages were developed for a particular type or range of analyses and that other packages were developed to be used for a broader range of applications.
- **Benefits and Limitations:** The team evaluated benefits and limitations based on their experience with various statistical software packages and applications. You must understand the limitations of any software package or any particular test before applying statistics to groundwater data at your specific site.

The following tables (Tables D-1 through D-7) indicate available statistical functions for each software package described in Sections D.1 through D.23. Each individual software description provides the specific information about the capabilities for that package.

Table D-1. Handling of nondetects

Statistical Method	Simple Substitution	Kaplan-Meier	ROS	MLE/Cohen
	Section 5.7.5	Section 5.7.6	Section 5.7.7	Section 5.7.8
D.1 3TMO	✓			
D.2 CARStat	✓			
D.3 ChemStat	✓			✓
D.4 DUMPStat	✓	✓	✓	
D.5 Excel	✓	✓	✓	✓
D.6 GTS	✓	✓		
D.7 GWSDAT	✓			
D.8 JMP	✓	✓	✓	✓
D.9 MATLAB	✓	✓	✓	✓
D.10 MINITAB	✓	✓	✓	✓
D.11 MAROS	✓	✓		
D.12 NCSS	✓	✓	✓	✓
D.13 PAM	✓			✓
D.14 PRO UCL	✓	✓	✓	
D.15 R for Statistics	✓	✓	✓	✓
D.16 SANITAS for Groundwater	✓			✓
D.17 Statistical Analysis System (SAS)	✓	✓	✓	✓
D.18 Scout	✓	✓	✓	✓
D.19 SPSS	✓	✓		
D.20 STATISTICA	✓	✓		✓
D.21 Summit Tools	✓			
D.22 SYSTAT	✓	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software		✓		

✓ = Capability available for this method/test. See package description for specific information.

Table D-2. Exploratory-diagnostic methods

Statistical Method	Summary Statistics	Distributional tests	Outlier tests	Data transformations
	Section 3.3.3	Section 5.6	Section 5.10	Appendix A
D.1 3TMO	✓			
D.2 CARStat	✓	✓	✓	✓
D.3 ChemStat	✓	✓	✓	
D.4 DUMPStat	✓	✓	✓	✓
D.5 Excel	✓	✓	✓	✓
D.6 GTS	✓		✓	
D.7 GWSDAT	✓		✓	✓
D.8 JMP	✓	✓	✓	✓
D.9 MATLAB	✓	✓	✓	✓

Statistical Method	Summary Statistics	Distributional tests	Outlier tests	Data transformations
D.10 MINITAB	✓	✓	✓	✓
D.11 MAROS	✓	✓	✓	✓
D.12 NCSS	✓	✓	✓	✓
D.13 PAM		✓		
D.14 PRO UCL	✓	✓	✓	
D.15 R for Statistics	✓	✓	✓	✓
D.16 SANITAS for Groundwater	✓	✓	✓	
D.17 Statistical Analysis System (SAS)	✓	✓	✓	✓
D.18 Scout	✓	✓	✓	✓
D.19 SPSS	✓	✓	✓	✓
D.20 STATISTICA	✓	✓	✓	✓
D.21 Summit Tools	✓			✓
D.22 SYSTAT	✓	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software	✓	✓	✓	
✓ = Capability available for this method/test. See package description for specific information.				

Table D-3. Statistical design

Statistical Method	Statistical Power	SWFPR	Contaminant ranking	Monitoring network optimization
	Section 3.6.1, Section 3.6.2	Section 3.6.2	Appendix C.9, C.10	Section 5.14.3
D.1 3TMO				✓
D.2 CARStat	✓	✓		
D.3 ChemStat	✓			
D.4 DUMPStat	✓	✓		
D.5 Excel	✓		✓	
D.6 GTS			✓	✓
D.7 GWSDAT				
D.8 JMP	✓	✓	✓	✓
D.9 MATLAB	✓	✓	✓	✓
D.10 MINITAB	✓			
D.11 MAROS	✓		✓	✓
D.12 NCSS	✓	✓	✓	✓
D.13 PAM				
D.14 PRO UCL	✓			
D.15 R for Statistics	✓	✓	✓	✓
D.16 SANITAS for Groundwater	✓	✓	✓	✓
D.17 Statistical Analysis	✓	✓	✓	✓

Table D-3. Statistical design

Statistical Method	Statistical Power	SWFPR	Contaminant ranking	Monitoring network optimization
System (SAS)				
D.18 Scout	✓	✓		
D.19 SPSS	✓			
D.20 STATISTICA	✓	✓	✓	
D.21 Summit Tools				✓
D.22 SYSTAT	✓	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software	✓			✓
✓ = Capability available for this method/test. See package description for specific information.				

Table D-4. Statistical limits

Statistical Method	Confidence Limits	Tolerance Limits	Prediction Limits	Testing Compliance Limits
	Section 5.2	Section 5.3	Section 5.4	Section 4.5.2
D.1 3TMO				
D.2 CARStat	✓		✓	✓
D.3 ChemStat	✓	✓	✓	✓
D.4 DUMPStat	✓		✓	✓
D.5 Excel	✓	✓	✓	✓
D.6 GTS	✓		✓	✓
D.7 GWSDAT	✓			✓
D.8 JMP	✓	✓	✓	✓
D.9 MATLAB	✓	✓	✓	✓
D.10 MINITAB	✓	✓	✓	✓
D.11 MAROS	✓			✓
D.12 NCSS	✓	✓	✓	✓
D.13 PAM	✓	✓	✓	✓
D.14 PRO UCL	✓	✓	✓	✓
D.15 R for Statistics	✓	✓	✓	✓
D.16 SANITAS for Groundwater	✓	✓	✓	✓
D.17 Statistical Analysis System (SAS)	✓	✓	✓	✓
D.18 Scout	✓	✓	✓	✓
D.19 SPSS	✓	✓	✓	✓
D.20 STATISTICA	✓	✓	✓	✓
D.21 Summit Tools			✓	
D.22 SYSTAT	✓	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software	✓	✓		
✓ = Capability available for this method/test. See package description for specific information.				

Table D-5. Graphics

Statistical Method	Plots/Charts	Batch plots	Tweaking of graphics
	Section 5.1		
D.1 3TMO	✓		✓
D.2 CARStat	✓	✓	✓
D.3 ChemStat	✓	✓	
D.4 DUMPStat	✓	✓	✓
D.5 Excel	✓	✓	✓
D.6 GTS	✓	✓	✓
D.7 GWSDAT	✓	✓	✓
D.8 JMP	✓	✓	✓
D.9 MATLAB	✓	✓	✓
D.10 MINITAB	✓	✓	✓
D.11 MAROS	✓		
D.12 NCSS	✓	✓	✓
D.13 PAM	✓	✓	✓
D.14 PRO UCL	✓		
D.15 R for Statistics	✓	✓	✓
D.16 SANITAS for Groundwater	✓	✓	✓
D.17 Statistical Analysis System (SAS)	✓	✓	✓
D.18 Scout	✓	✓	✓
D.19 SPSS	✓	✓	✓
D.20 STATISTICA	✓	✓	✓
D.21 Summit Tools	✓	✓	✓
D.22 SYSTAT	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software	✓	✓	✓
✓ = Capability available for this method/test. See package description for specific information.			

Table D-6. Comparison and analysis

Statistical Method	t-tests	ANOVA	Geostatistics/ Mapping	Kriging/ Interpolation	Spatial smoothing	Bootstrapping
	Section 5.11	Section 5.8.2	Section 5.14	Section 5.14.2	Section 5.14.1	Section 5.5.3
D.1 3TMO			✓			
D.2 CARStat						
D.3 ChemStat	✓	✓				
D.4 DUMPStat						
D.5 Excel	✓	✓	✓	✓	✓	✓
D.6 GTS			✓		✓	

Table D-6. Comparison and analysis

Statistical Method	t-tests	ANOVA	Geostatistics/ Mapping	Kriging/ Interpolation	Spatial smoothing	Bootstrapping
D.7 GWSDAT			✓	✓	✓	
D.8 JMP	✓	✓	✓	✓	✓	✓
D.9 MATLAB	✓	✓	✓	✓	✓	✓
D.10 MINITAB	✓	✓				✓
D.11 MAROS	✓					
D.12 NCSS	✓	✓				✓
D.13 PAM	✓	✓				
D.14 PRO UCL	✓	✓				✓
D.15 R for Statistics	✓	✓	✓	✓	✓	✓
D.16 SANITAS for Groundwater	✓	✓				
D.17 Statistical Analysis System (SAS)	✓	✓	✓	✓	✓	✓
D.18 Scout	✓	✓	✓			✓
D.19 SPSS	✓	✓				✓
D.20 STATISTICA	✓	✓				✓
D.21 Summit Tools			✓	✓	✓	
D.22 SYSTAT	✓	✓	✓	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software	✓	✓	✓	✓	✓	

✓ = Capability available for this method/test. See package description for specific information.

Table D-7. Regression and time series

Statistical Method	Trend Tests	Mann-Kendall	Linear regression	Nonlinear regression	Theil-Sen	Time-series Analysis
	Section 5.5	Section 5.5.2	Section 5.5.1	Section 5.5.1	Section 5.5.3	Section 5.1.1 or Section 5.8
D.1 3TMO	✓	✓				✓
D.2 CARStat	✓				✓	✓
D.3 ChemStat	✓	✓			✓	
D.4 DUMPStat	✓	✓			✓	✓
D.5 Excel	✓	✓	✓	✓	✓	✓
D.6GTS	✓			✓	✓	✓
D.7 GWSDAT	✓	✓	✓	✓		✓
D.8 JMP	✓	✓	✓	✓	✓	✓
D.9 MATLAB	✓	✓	✓	✓	✓	✓
D.10 MINITAB	✓	✓	✓	✓	✓	✓
D.11 MAROS	✓	✓	✓			
D.12 NCSS	✓	✓	✓	✓	✓	✓

Table D-7. Regression and time series

Statistical Method	Trend Tests	Mann-Kendall	Linear regression	Nonlinear regression	Theil-Sen	Time-series Analysis
D.13 PAM	✓	✓			✓	
D.14 PRO UCL		✓	✓		✓	
D.15 R for Statistics	✓	✓	✓	✓	✓	✓
D.16 SANITAS for Groundwater		✓			✓	
D.17 Statistical Analysis System (SAS)	✓	✓	✓	✓	✓	✓
D.18 Scout	✓	✓	✓		✓	✓
D.19 SPSS	✓	✓	✓	✓		✓
D.20 STATISTICA	✓		✓	✓		✓
D.21 Summit Tools						
D.22 SYSTAT	✓	✓	✓	✓	✓	✓
D.23 Visual Sampling Plan (VSP) Software	✓	✓	✓	✓		✓
✓ = Capability available for this method/test. See package description for specific information.						

D.1 3-TIERED MONITORING OPTIMIZATION TOOL (3TMO)

Approximate Cost: Free

Source: Request from Philip Hunter, AFCEC (formerly AFCEE), philip.hunter@us.af.mil, or John Hicks, Parsons, john.hicks@parsons.com

Current Version: v1.0

Operating System Needs: Windows XP Service Pack 3, Windows Vista, Windows 7

Input Structure: Microsoft Excel spreadsheet or comma-separated values (CSV) file

Overview

The 3-Tiered Monitoring Optimization Tool (3TMO) was developed by Parsons and ENVIRON International Corp. on behalf of Air Force Civil Engineer Center (AFCEC), known previously as Air Force Center for Engineering and the Environment (AFCEE), in 2011. This program is a comprehensive, public domain, user-friendly, long-term monitoring optimization (LTMO) decision support tool. Across the spectrum of monitoring approaches, this tool emphasizes the qualitative and intuitive aspects of optimization. The program offers more efficient performance and successful implementation of LTMO evaluations. In addition, 3TMO presents guidance for other aspects of monitoring programs, including use of passive or low-flow sampling methods, sample shipment, purge water disposal, field quality assurance/quality control samples, and data management. This program incorporates substantive qualitative considerations into the analysis, and output can be augmented by user-entered considerations and rationale.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	◒	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests		N/A
Outlier tests		N/A
Data transformations		N/A
Statistical Design		
Statistical Power		N/A
SWFPR		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Contaminant ranking		N/A
Monitoring network optimization	◒	N/A
Statistical Limits		
Confidence Limits		N/A
Tolerance Limits		N/A
Prediction Limits		N/A
Testing Compliance Limits		N/A
Graphics		
Plots/Charts	●	N/A
Batch plots		N/A
Tweaking of graphics	◒	N/A
Statistical Comparisons		
t-tests		N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping	◒	N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	◒	N/A
Mann-Kendall	●	N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line		N/A
Time Series analysis	◒	N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

3TMO applies user input for project information, well parameters, contaminants, and sample data. Enter tabular data directly into the data input grids or import information from correctly formatted Excel or CSV files. Enter other data through the functions accessible from the **Data Input** sub-menu on the **Framework Overview** home page or from the **LTMO Framework** menu links. Well parameters must be loaded before sample data; the sample data well IDs must be entered into the Well Parameters table before they can be successfully loaded. Example data input files are available from the **Help** menu, **Example Project** link, and can be used as templates for creating site-specific data upload files.

The 3-Tiered approach is sufficiently sophisticated, comprehensive, and rigorous to yield credible and useful results, yet at the same time it is intuitive and easy to understand. The spatial analysis that 3TMO offers is a relatively straightforward qualitative analysis that can be performed to evaluate the spatial importance of monitoring locations using the **Map Tool**.

Types of Distribution

The Mann-Kendall test for trends is included in 3TMO which is a nonparametric trend test. The software does not include tests to determine the distribution of the data (such as normal, lognormal).

Visualization

Results of statistical analyses can be exported to Excel for formatting to a user-defined specification.

Primary Uses for Groundwater Data Analysis

This public domain software tool supports LTMO evaluations at the groundwater contaminant plume or site level (not on an installation-wide basis). The tool is intended for use by environmental managers, contractors, and regulators at sites of varying size and complexity, and is scalable dependent on the available level of information. It provides a credible alternative for sites where LTMO practitioners must incorporate a substantial qualitative evaluation into the overall LTMO analysis. Although the tool was designed for optimization of groundwater monitoring networks, portions of it could conceivably also be used to assess surface water monitoring networks (such as temporal trend evaluation and spatial evaluation).

Benefits

- 3TMO can manage large data sets.
- The 3-Tiered approach to LTMO is unique when compared to existing LTMO statistical applications because it focuses on qualitative factors that are supported by quantitative statistical analysis.

- The presentation of results allows for transparency of decision-making with clearly defined and easily accessible decision rationale and backup data.
- The program applies a decision algorithm to assess the optimal frequency of monitoring, the optimal spatial distribution of the components of the monitoring network, and to develop final recommendations for monitoring program optimization.

Limitations and Data Requirements

- The spatial analysis included in 3TMO is a qualitative evaluation facilitated by the **Map Tool**; the spatial importance of each well is not quantitatively determined using geostatistics.
- 3TMO assigns nondetect results a value of zero, and does not allow the user other options such as assigning a value equal to the method detection limit (MDL) or half the MDL.
- The Mann-Kendall trend module does not provide the analysis statistics (such as S statistic or confidence in trend).
- Charts and maps are exported as image files that cannot be manipulated further outside of 3TMO.
- The map tool does not project layers with differing coordinate systems to a common coordinate system. Thus, to display properly on the map, the imported map layers need to be in the same coordinate system and map units as the well coordinates.

References

Nobel, C. and Anthony, J.A. 2004. "Three-Tiered Approach to Long-Term Monitoring Program Optimization." *Bioremediation Journal* 8 (3-4): 147-165.

D.2 CARStat

Approximate Cost: $4,000

Source: Discerning Systems Inc. (www.DiscerningSystems.com)

Current Version: 2.1.9

Operating System Needs: Windows XP (has been successfully installed and used on Windows 7 32-bit and 64-bit; 64-bit requires the Microsoft Virtual PC and XP mode)

Input Structure: ASCII (text) flat file; one data row per measurement; comma-separated values (CSV) file; delimited, or fixed column, flexible column order

Overview

CARStat is a statistical analysis system that automatically performs a complete analysis of all sampling locations, groups of locations, and contaminants for compliance, assessment, and remediation. This program was originally designed for the analysis of industrial plants, disposal facilities, brownfield sites and other installations requiring detailed investigation. CARStat statistically analyzes data for soil, groundwater, surface water, air and waste streams for assessment monitoring and corrective action programs and performs comparisons to background and to regulatory standards, as well as performing a natural attenuation analysis. This program extends DUMPStat's assessment monitoring capabilities with more sophisticated statistical techniques.

Monitoring locations can be combined and analyzed and graphed as Potential Areas of Concern. All statistical processing balances false negative and false positive rates for the entire facility. Results are presented in graphical and tabular formats and all intermediate calculations are shown in worksheets.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Outlier tests	●	N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR	◒	N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits		N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests		N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall		N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

CARStat merges structured ASCII (text) files into its environmental database. Spreadsheet and database files such as Excel and Access can export data into an appropriate format for import.

To use this program, select the monitoring network and contaminants for analysis, and specify statistical options. These selections persist as new data are added, so that you can perform subsequent analyses easily.

The program makes all decisions on the most appropriate methods to use within the parameters of the statistical analysis settings, which guide the processing automatically so that large quantities of environmental data can be analyzed. You can set up a series of zones so that multiple areas of a site (whether they are related physically or statistically) can be batch processed differently from other zones without re-entering the monitoring network and statistical options.

CARStat provides separate viewers for the graphical and tabular results and for worksheets, which show the individual statistical calculations in detail. Use a dedicated **Print Results** screen to print all output from one place with options for grouping and ordering of the results.

Types of Distributions

Normal, log normal, Poisson, and nonparametric distributions are available for data analysis.

Visualization

Specific options for data output include prediction limit graphs, confidence interval charts and power curves. The graphs can be cut and pasted to move into other applications.

Primary Uses for Groundwater Data Analysis

CARStat is used by engineers, environmental scientists, regulators, consultants, owners and operators to comply with state and federal regulations at industrial plants, waste disposal facilities and Brownfield sites. The program monitors the progress of corrective action, demonstrates clean closure, and helps parties transferring real estate. The statistical analyses are also consistent with USEPA Subtitle C and D and American Society for Testing and Materials (ASTM) Standard D7048-04 requirements for landfill detection monitoring.

Benefits

- CARStat imports, stores, and manages large amounts of environmental data. Data can be edited, aliased, reviewed, and printed. Units can be converted.
- Flexible data import works with many laboratory data file formats and qualifiers.
- Statistical options and user adjustable time windows provide a means to custom tailor analyses while maintaining batch processing for fast, automated results.
- The program maintains a database of regulatory standards organized by state, matrix, and land use, which can be shared among facilities.
- Correct application of statistical methods can minimize the cost of remediation.
- Centralized analysis and printing speeds report generation.
- Powerful display and filtering options for output highlight important results.

Limitations and Data Requirements

- Cost
- CARStat does not interface with other programs for graphical output, it is limited to cut and paste into other applications.

D.3 ChemStat

Approximate Cost: $990; ChemPoint, which is data management software that supports ChemStat is an additional $300.

Source: http://www.pointstar.com/

Current Version: v6.3

Operating System Needs: Windows

Input Structure: tab delimited text file; Sanitas, DUMPstat, GRIT, ChemPoint files; converter program to import Microsoft Excel, Microsoft Access, and DBF files.

Overview

ChemStat can be used as an assessment tool for analyzing groundwater contamination. The software can calculate upper confidence limits, upper prediction limits, and upper tolerance limits for comparison to background concentrations in groundwater or a designated value (such as MCLs). Other available tests that may apply to groundwater monitoring include one-way analysis of variance (ANOVA), trend evaluation including seasonal analysis, outlier, and goodness-of-fit tests.

ChemPoint 7.0 is a data management system that organizes environmental data collected during water, soil and air sampling. Although not required, ChemPoint is commonly purchased along with ChemStat and serves as the data repository. ChemStat also interfaces directly with files from ChemPoint, Sanitas, DUMPstat, and GRITS programs.

In conjunction with the data repository, ChemStat is used to plot time-series graphs of data, identify outliers, test for normal data, and calculate concentration limits of both parametric and non-parametric monitoring well data to be used for both intrawell and interwell comparisons in order to evaluate potential groundwater contamination. ChemStat generates reports and various plots including box plots, concentration versus time plots, and probability plots.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	◒	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE	◒	N/A
Exploratory/Diagnostic Tools		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Summary Statistics	◒	N/A
Distributional tests	◒	N/A
Outlier tests	◒	N/A
Data Transformation		N/A
Statistical Design		
Statistical Power	◒	N/A
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	◒	N/A
Tweaking of graphics		N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line	◒	N/A
Time Series analysis		N/A
Multivariate Analysis		
Multiple regression	◒	N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

To input data, format the data into columns and generate an ASCII text file. A conversion utility is included to convert tab delineated text, Microsoft Excel files, Microsoft Access files and DBF files. You can provide additional variables for regression analysis as part of concentration versus time plots; however, all other statistical methods are performed one parameter at a time. Select desired statistical tests from drop-down menus and relevant options from subsequent menus. ChemStat does support batch processing of statistical methods and graphs.

ChemStat is relatively straightforward to use; however, make sure that you understand the assumptions and input requirements for any statistical tests used to ensure validity and applicability in making decisions.

Types of Distribution

Data can be evaluated for fit to normal, lognormal, or gamma distributions. Statistical interval test results are available for all of these distributions in addition to several nonparametric options.

Visualization

Plots available in ChemStat include box plots, concentration versus time plots, probability plots, and Shewhart-CUSUM control charts for both intrawell and interwell comparisons. Limited options are offered for editing the output plots, and you can only export output to Microsoft Word or Acrobat PDF files.

Primary Uses for Groundwater Data Analyses

ChemStat 6.3 is a Windows-based program used for analysis of groundwater, surface water, soil, or air quality monitoring data at RCRA facilities. Analysis methods are consistent with the USEPA Unified Guidance document.

Benefits

- Relative simplicity of use
- Developed specifically for environmental applications

- Documentation is well-written and generally easy to understand

Limitations and Data Requirements

- Cost
- Primary use is calculation of upper statistical limits
- Cannot set percent false positive rates or false negatives to other values than those preset for statistical tests
- Limited options provided to modify plots and reports

D.4 DUMPStat

Approximate Cost: $2,500 for a landfill (or other facility) license; includes a copy of the software. Software is available separately for $500 per copy for use with already licensed sites. Upgrade pricing to be determined.

Source: Discerning Systems Inc. (www.DiscerningSystems.com)

Version: 2.3 (scheduled for early 2013)

Operating System Needs: Windows XP (has been successfully installed and used on Windows 7 32-bit and 64-bit; 64-bit requires the Microsoft Virtual PC and XP mode)

Input Structure: ASCII (text) flat file; one data row per measurement; comma-separated values (CSV) file, delimited, or fixed column, flexible column order

Overview

DUMPStat is a statistical package used to perform ongoing detection monitoring. Primarily designed for the analysis of waste disposal facilities, DUMPStat is also used at a wide variety of facilities including mining and industrial sites. It statistically analyzes data from groundwater, surface water, and air to comply with monitoring requirements.

DUMPStat is licensed on a per landfill (or other facility) basis and includes a copy of the software. Additional copies of the software can be purchased for use with licensed facilities. No ongoing fees are charged to monitor a facility.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	●	N/A
ROS	●	N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Statistical Design		
Statistical Power	●	N/A
SWFPR	●	N/A
Contaminant ranking		N/A
Monitoring network optim-ization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits		N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests		N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability
(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

DUMPStat merges structured ASCII (text) files into its environmental database. Once the data are merged, select the monitoring network and contaminants for analysis and specify statistical options so that each type of analysis can be custom tailored as required. These selections persist as new data are added, so that subsequent analyses can be performed easily.

An analysis screen provides a means to select one or several analyses to be performed and serves as a central location for accessing all results. The program makes all decisions on the most appropriate methods to use within the parameters of the statistical analysis settings, which guide the processing automatically so that large quantities of environmental data can be analyzed quickly and correctly. DUMPStat provides separate viewers for the graphical and tabular results and worksheets, which show the individual statistical calculations in detail. The viewers allow for the customization of display and printing with options for colors, scaling and page layout, and filtering options to highlight statistically interesting results such as exceedances or those containing trends. A new sample descriptor feature allows you to assign custom symbols to classes of points. This feature generates legend entries and all symbols can be scaled and line thickness adjusted. DUMPStat 2.3 also allows you to add annotations to graphs. You can edit, reposition, resize, and rescale these annotations. Annotations persist so that they do not need to be regenerated after every analysis. Use the dedicated **Print Results** screen to print all output from one place (includes options for grouping and ordering of the results).

You can set up a series of zones so that multiple areas of a site (whether they are related physically or statistically) can be batch processed in different ways from other zones without re-entering the monitoring network and statistical options.

DUMPStat includes a complete simulation module that can be used to compute site-wide false positive and false negative rates (statistical power) based on any combination of statistical methods available. Statistical power characteristics for different statistical monitoring plans can easily be compared and contrasted in terms of their Type I (false positive) and Type II (false negative) rates.

Types of Distributions

Normal, log normal, cube root, square root, square, cube, Poisson, gamma and nonparametric distributions are available for data analysis. Selection of the appropriate distribution is according to Helsel's Ladder of Powers: normal, lognormal, cube root, square root, square, and cube.

Visualization

Specific options for data output include control charts, prediction limit graphs, confidence interval charts and power curves. The graphs can be cut and pasted to move into other applications.

Primary Uses for Groundwater Data Analysis

DUMPStat is used by engineers, environmental scientists, regulators, consultants, owner, and operators to comply with state and federal regulations at waste disposal facilities, industrial plants, and mining sites. It is used primarily for ongoing detection monitoring. DUMPStat statistical analyses are consistent with USEPA Subtitle D, American Society for Testing and Materials (ASTM) Standard D6312-98 requirements for landfill detection monitoring, and methods described in the USEPA Unified Guidance.

Benefits

- DUMPStat imports, stores, and manages large amounts of environmental data.
- Data can be edited, aliased, reviewed and printed. Units can be converted.
- Flexible data import works with many laboratory data file formats and qualifiers.
- Statistical options and user adjustable time windows provide a means to customize analyses, while maintaining batch processing for fast, automated results.
- Correct application of statistical methods avoids costly assessments and remediation.
- Centralized analysis and printing speeds report generation.
- Powerful display and filtering options for output highlight important results.

Limitations and Data Requirements

- Cost
- Each site must be licensed independently.
- DUMPStat does not interface with other programs for graphical output, it is limited to cut and paste into other applications.

D.5 EXCEL

Approximate Cost: $140 (Most often bundled with Microsoft Office Suite, cost variable around $300)

Source: Microsoft Store (www.microsoftstore.com/store/msstore/home)

Current Version: 2010 for Microsoft Windows and 2011 for Mac OS X

Operating System Needs: Microsoft Windows or Mac OS X

Input Structure: Enter data into a grid of cells arranged in numbered rows and lettered or numbered columns. Data type varies and includes numeric, text, dates, time, and percentage.

Overview

Excel is a commercial, flexible spreadsheet-based program. This program comes preconfigured to perform several parametric statistical tests ranging from t-tests to two-way analysis of variance with replications. In the native capability configuration, Excel offers a narrow range of statistical functions that have limited usefulness when evaluating groundwater data. You must use add-ins to evaluate groundwater data.

Excel supports several types of expansion. Limited formula syntax allows you to develop statistical packages. In addition, numerous commercially-available statistical add-ins are available. Programing with Visual Basic for Applications (VBA) allows data manipulations that are difficult to accomplish with the packaged formula syntax.

Disclaimer: Statistical functions and capabilities presented for this software package have not been reviewed or verified by Microsoft.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution		●
Kaplan-Meier		●
ROS		●
Cohen/MLE		●
Exploratory/Diagnostic Tools		
Summary Statistics	●	●

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Distributional tests		●
Outlier tests		●
Data transformations	●	●
Statistical Design		
Statistical Power		●
SWFPR		N/A
Contaminant ranking		●
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	●
Tolerance Limits		●
Prediction Limits		●
Testing Compliance Limits		●
Graphics		
Plots/Charts	●	●
Batch plots	●	●
Tweaking of graphics	●	●
Statistical Comparisons		
t-tests	●	●
ANOVA	●	●
Spatial Analysis		
Geostatistics/Mapping		●
Kriging/Interpolation		●
Spatial smoothing		●
Regression/Time Series		
Trend Tests	●	●
Mann-Kendall		●
Linear regression	●	●
Nonlinear regression	●	●
Theil-Sen line		●
Time Series analysis		●

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Multivariate Analysis		
Multiple regression		●
Factor/Discriminant analysis		●
Bootstrapping		●

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

Available through Microsoft and other commercial entities. Some examples include Analysis ToolPak for Excel 2010 (www.microsoftstore.com), xlstatistician (www.xlstatistician.com), and XLStat (www.xlstat.com). Examine add-ins carefully to confirm that they contain the statistical procedures of interest.

Ease of Use and Data Import

You can enter data into Excel in a variety of ways ranging from simple keystroke to importing data contained in databases like Access. Standard rules of relational database development must be used if the data are going to be imported from an external database. For example, the fields can be delimited in a variety of ways, such as tab-delimited or comma-separated values (CSV), but must have specific field names. Each groundwater measurement must occupy one record of the input text file.

Types of Distributions

Excel accepts data of any distributional type. You can apply data transformations within Excel. Statistical procedures within the native capability of Excel are parametric.

Visualization

The native capability of Excel includes basic built-in graphics for data visualization (such as scatter plots, histograms, and line plots). You can alter the graphics formatting.

Primary Uses for Groundwater Data Analysis

Excel is a commercial spreadsheet application created as a general but flexible data analysis tool that can be applied across a broad range of disciplines (for example, in business, engineering, finance, and science). The native capability of Excel is not specifically tailored for statistical evaluation or optimization of groundwater monitoring networks. However, add-ins and custom code

creation (in VBA) allows you to create highly tailored statistical functions that can be used to evaluate or optimize groundwater monitoring networks. Excel is easy to use for simple statistical applications, ubiquitous and compatible with other Microsoft Office applications such as the Access database tools. Data and results can be easily exported from Excel into other applications such as ESRI geographic information system (GIS) tools.

Benefits

- Excel is a widely used spreadsheet application. The in-program help function is intuitive and the learning curve is short for native capability functions. Excel does not require advanced training or a specific statistical skill set.
- Many websites provide help for add-ins or for creating specific applications through the Excel-specific formula syntax.
- Excel is convenient for data entry and manipulating rows and columns of data.
- Excel can be used for rapid, preliminary analysis of data. Tables and graphs of results can be professionally formatted, rapidly and easily brought into presentation tools such as PowerPoint.
- Custom functions can be programmed using macros programmed in VBA.

Limitations and Data Requirements

- Native capability statistical tests do not include nonparametric methods or several other statistical tests (for example, prediction limits).
- No groundwater-tailored evaluation functions are offered in native Excel.
- Previous versions of Excel (for example, Excel 2007) produced erroneous results for some statistical procedures.
- Distributions are not computed with precision.
- Excel has limited accuracy with very large and very small numbers. The precision is confined to 15 significant figures, which may be further limited by rounding off and binary storage.
- Ranks of tied data are nonstandard.
- No record is made of the process to results.
- Historical compatibility issues with macros programmed in earlier versions of the software. Microsoft has a history of changing how custom buttons and objects are handled in applications built on the Excel platform.

References

Burns Statistics, http://www.burns-stat.com/, (click on tutorials, spreadsheet addiction).

Goldwater, A. 2007. *Using Excel for Statistical Data Analysis – Caveats*. Biostatistics Consulting Center, University of Massachusetts School of Public Health.

Heilberger, R. M., and E. Neuwirth. 2009. *R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis and Graphics*. London: Springer Dordrecht Heidelberg.

MCCullough, B. D., and D. A Heiser. 2008. "On the Accuracy of Statistical Procedures in Microsoft Excel 2007," Computational Statistics and Data Analysis 52: 4570-4578.

Practical Stats, http://www.practicalstats.com/xlsstats/excelstats.html.

Yalta, A. T. 2008. "The Accuracy of Statistical Distributions in Microsoft Excel 2007," *Computational Statistics and Data Analysis* 52: 4579-4586.

D.6 GEOSTATISTICAL TEMPORAL-SPATIAL OPTIMIZATION SOFTWARE (GTS)

Approximate Cost: Free

Source: http://www.itrcweb.org/team/GTS-Optimization-Software

Current Version: v1.0

Operating System Needs: Windows XP (has been successfully installed and used on Windows 7, though not formally supported)

Input Structure: ASCII (text) flat file with fixed column header names; one data row per measurement; tab-delimited preferred but not required

Overview

Geostatistical Temporal-Spatial software (GTS) is a statistical and geostatistical decision-logic groundwater monitoring optimization software that is publicly available as open-source freeware. GTS is a quantitative calculation tool that includes options to customize its use. It was developed for the Air Force Civil Engineer Center (AFCEC), known previously as Air Force Center for Engineering and the Environment (AFCEE). Given an existing long-term monitoring (LTM) network, GTS uses a combination of statistical techniques to answer two questions:

1. What is the optimum number and placement of wells in that network?
2. What is the optimal sampling frequency for wells in the network?

GTS has five modular components linked together in a user-friendly interface: Prepare, Explore, Baseline, Optimize, and Predict. The Prepare and Explore modules allow the user to import and manage analytical and water-level data, identify outliers, explore basic statistical features of the data (including simple trends), and also to rank contaminants in terms of optimization potential. The Baseline module creates nonlinear trends and trend maps, and constructs base maps to quantify and visualize plume extent. Baseline also allows you to create potentiometric surface maps. The Optimize component runs two distinct types of temporal optimization—iterative thinning and temporal variograms—as well as spatial optimization involving both a search for statistical redundancy and an assessment as to whether and where new wells should be added. The software is designed so that you may chose only to perform the temporal optimization as a standalone module. However, the spatial analysis depends on the temporal analysis being performed first in sequence to obtain the spatial results. Finally, the Predict module focuses on flagging newly imported data that are inconsistent with projected trends and maps.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	◒	N/A
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests		N/A
Outlier tests	◒	N/A
Data transformations		N/A
Statistical Design		
Statistical Power		N/A
SWFPR		N/A
Contaminant ranking	●	N/A
Monitoring network optimization	●	N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits		N/A
Prediction Limits	◒	N/A
Testing Compliance Limits	◒	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics/tables	◒	N/A
Statistical Comparisons		
t-tests		N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping	◒	N/A
Kriging/Interpolation		N/A
Spatial smoothing	●	N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall		N/A
Linear regression		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Nonlinear regression	●	N/A
Theil-Sen line	●	N/A
Time Series analysis	◒	N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

GTS is officially limited to the Windows XP platform, though some users have successfully installed and used it on Windows 7. The wizard interface offers a number of defaults and makes it easy to click through a basic analysis. Basic defaults can be configured and set with many preferences to allow a highly customized optimization. Process flow is logical from top to bottom and left to right when navigating the sequence of operations in the dialog window on each page. Interpreting the results properly requires an intermediate background or training in statistics and geostatistics. GTS includes simple plots, exploratory tools, and trend analyses, as well as sophisticated statistical techniques and optimization algorithms written in the open-source statistical computing environment R (www.r-project.org).

GTS requires input of a structured ASCII (text) flat file. The fields can be delimited in a variety of ways, such as tab-delimited or comma-separated values (CSV), but must have specific field names, generally corresponding to the format of AFCEC'sEnvironmental Resource Program Information Management System (ERPIMS) database. The order or sequencing of data fields is not critical. Shape files of facility boundaries, sites, roads, and other infrastructure can be imported. Each groundwater measurement must occupy one record of the input text file. Fields required for a GTS analysis are listed within the GTS Users Guide. Data files in Excel or spreadsheet format must be exported to text format prior to GTS input. A data filtering tool allows analysis of selected records.

Types of Distributions

GTS accepts data of any distributional type. Although you cannot apply data transformations within GTS, most of the procedures within GTS are quasi-nonparametric and do not require explicit fitting of parametric models or distributional testing.

Visualization

GTS includes sophisticated built-in graphics for data visualization, including contour mapping, complex nonlinear trends, post-plots, and shape file annotation. GTS provides automated batch processing of graphics in order to sequentially plot multiple wells, contaminants, aquifer zones, and time periods. Graphics are designed to be final pictures for reports, however, the program cannot batch print graphics. In addition, users cannot tweak or alter the graphics formatting. On the other hand, some interactive widgets are provided, for instance, zooming and scaling tools, and pointers for identifying specific locations on plan-view maps. Individual graphs are best exported using the Windows Snipping tool or an equivalent screen capturing application.

Primary Uses for Groundwater Data Analysis

GTS can be used at various stages in the life cycle of groundwater monitoring, but is best for optimizing long-term monitoring networks, once characterization has been completed and remedies are in place. Although the exploratory tools can be used during any stage of a facility's life cycle, GTS generally assumes that a given site has been adequately characterized, is undergoing long-term monitoring, and that enough well locations exist and sampling data collected so that statistical redundancy in locations and sampling events might exist.

Benefits

- Applicable to site-specific plumes or site-wide studies (for example, entire facilities or installations) involving multiple source areas, plumes, and monitoring conditions.
- Does not require plume-specific configuration data, fate-and-transport models, or other hydrogeologic modeling information
- Stand-alone spatial and temporal optimization modules that can be used independently
- Exploratory statistical tools for assessing data characteristics, ranking contaminants for optimization potential, and analyzing multiple aquifer horizons
- Fitting of nonlinear and seasonal time series data
- Semi-nonparametric surface map estimates made using quantile local regression, a smoothing technique not bound by the constraints of kriging
- Empirical, data-driven assessment of redundancy (reduced-network is optimal if it can accurately reproduce base maps).
- Automated redundancy searches, both during temporal and spatial optimization
- Use of multiple cost-accuracy tradeoff curves to gauge points of optimality
- No limitations on the number of monitoring wells or sampling events
- Spatial analysis uses quasi-genetic algorithm to determine essential and redundant wells
- Imports multiple shape files for boundaries and infrastructure

- Temporal analysis proposes optimal sampling intervals specific to the number of quarters
- Database filtering tool helps select records for "what if" analysis

Limitations and Data Requirements

- Preparing the data set can be challenging with potentially a large number of data fields
- Effective spatial optimization in GTS requires a minimum of 15-20 wells and at least two sampling events per well; temporal optimization requires at least one well and 6-8 distinct sampling events per location.
- Quantile local regression, the GTS spatial mapping engine, by design is a 'smoother' rather than an interpolator (thus may not replicate or 'honor' observed measurements when creating map estimates, unlike, for instance, kriging)
- Does not offer sophisticated handling of radiochemical data, particularly measurements recorded with non-positive values (zeros or negatives); must first convert these data to positive values, unless they represent nondetects with a known, positive detection or reporting limit
- Does not track changes in contaminant or plume mass or allow users to specify contaminant mass as an optimization criterion
- May not give valid spatial results in subsurface environments that are highly fractured and discontinuous with poor hydraulic connection.
- Note: Spatial mapping techniques in general (not just those in GTS) inherently assume that concentration patterns at known wells can be extended (interpolated, smoothed) to unsampled locations. This may be problematic at sites with large contrasts in hydraulic conductivity (preferential pathways).

References

Cameron, K., P. Hunter, and R. Stewart. 2011. Demonstration and validation of GTS long-term monitoring optimization software at military and government sites. ESTCP Project ER-200714. www.serdp.org.

Cameron, K. 2004. "Better optimization of LTM networks." *Bioremediation Journal*8 (03-04): 89-108.

Cameron, K., and P. Hunter. 2004. Optimizing LTM networks with GTS: three new case studies. Conference on Accelerating Site Closeout, Improving Performance, & Reducing Costs Through Optimization, Dallas.

Cameron, K. M., and P. Hunter. 2003. "Optimization of LTM networks at AF Plant 6 using GTS". In V.S. Magar & M.E. Kelley (Eds.), In Situ and On-Site Bioremediation – 2003. Proceedings of the Seventh International In Situ and On-Site Bioremediation Symposium (Orlando, FL; June 2003), Columbus, OH: Battelle Press.

Cameron, K., and P. Hunter. 2002. "Using spatial models and kriging techniques to optimize long-term ground-water monitoring networks: a case study".*Environmetrics*13: 629-656.

Cameron, K., and P. Hunter. 2000. "Optimization of LTM networks: statistical approaches to spatial and temporal redundancy." Spring Natl. Meeting of American Institute of Chemical Engineers, Atlanta.

D.7 Groundwater Spatio-Temporal Data Analysis Tool (GWSDAT)

Approximate Cost: Free

Source: http://www.api.org/GWSDAT

Current Version: v2.0

Operating System Needs: Windows XP, Vista or Windows 7, also requires Microsoft Office, XP, 2007 or 2010

Input Structure: Excel standardized data input template sheet

Overview

The Groundwater Spatiotemporal Data Analysis Tool (GWSDAT), was developed by Shell Global Solutions to help visualize trends in groundwater monitoring data. This program is designed to work with simple time-series data for contaminant concentration and ground water elevation, but can also plot non-aqueous phase liquid (NAPL) thickness if required. Spatial data are input in the form of well coordinates, and wells can be grouped to separate data from different aquifer units. The software also allows the import of a site base map in GIS shapefile format. Trend and contour plots generated using GWSDAT can be exported directly to Microsoft PowerPoint and Word to expedite reporting.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	◒	N/A
Distributional tests		N/A
Outlier tests	●	N/A
Data transformations	◒	N/A
Statistical Design		
Statistical Power		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits		N/A
Prediction Limits		N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests		N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping	●	N/A
Kriging/Interpolation	●	N/A
Spatial smoothing	●	N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression	●	N/A
Theil-Sen line		N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

GWSDAT has been designed to be as user-friendly as possible. The application is supported for Windows XP, Vista, Windows 7, and the corresponding version of Microsoft Office. The user entry point and data input platform to GWSDAT is a custom built Excel Add-in application. The statistical engine used to perform geo-statistical modeling and display graphical output is the open source statistical programming language R (www.r-project.org).

Enter groundwater monitoring data into GWSDAT by populating three tables in a standardized Excel input sheet. Include historical monitoring data, well coordinates, and GIS shape files. Two example data files are provided with the program for training and demonstration purposes. After entering the data, select "GWSDAT Analysis" from the Excel add-in menu to initiate a GWSDAT analysis.

The GWSDAT operates with a graphical, stand-alone point-and-click interface that allows you to analyze time-series, spatial, and (uniquely) spatiotemporal trends. By left-clicking on any of the user interface plots, an identical but expanded plot is generated in a separate window. You can save plots to a variety of different formats including JPEG, postscript, PDF, metafile, and Microsoft Power Point slide.

Types of Distributions

GWSDAT uses a wide variety of different nonparametric statistical methods for the analysis of trends in temporal, spatial and spatiotemporal components of the groundwater monitoring data set.

Visualization

GWSDAT includes sophisticated graphical visualization for trend detection:

- Spatial plot: This is for the analysis of spatial trends in solute concentrations, groundwater flow and, if present, nonaqueous phase liquid (NAPL) thickness. Overlaid on this plot are the predictions of the spatiotemporal solute concentration smoother which is a function that simultaneously estimates both the spatial and time series trend in site solute concentrations. User specified shape files can also be overlaid on this plot. The spatial plot can be automatically plotted in time series order to provide a movie depicting the changing trends in spatial solute concentrations.
- Well Trend plot: This is for the investigation of historical time-series trends in solute concentrations, groundwater elevation and, if present, NAPL thickness for individual wells. Users can overlay a nonparametric smoother which estimates the time-series trend in solute concentration. The advantage of this nonparametric method is that the trend estimate is not constrained to be monotonic, i.e. the trend can change direction.

- Trend and Threshold Indicator Matrix: This provides a summary of the level and time series trend in solute concentrations at a particular model output interval.

Benefits

- Early identification of increasing trends or off-site migration
- Evaluation of groundwater monitoring trends over time and space (holistic plume evaluation)
- Nonparametric statistical and uncertainty analyses to assess highly variable groundwater monitoring data
- Reduction in the number of sites in long-term monitoring or active remediation through simple, visual demonstrations of groundwater data and trends
- More efficient evaluation and reporting of groundwater monitoring trends via simple, standardized plots and tables.

Limitations and Data Requirements

- Spatiotemporal solute concentration predictions do not necessarily lie on observed data points because the program smoothes rather than interpolates.
- The quality of the spatiotemporal smoothing is directly influenced by the quality of the underlying data.
- The analysis may be skewed if data are input from monitoring wells with disparate construction or screened in different aquifers.

References

Ahmadi, S. H., and A. Sedghamiz. 2007. "Geostatistical Analysis of Spatial and Temporal Variations of Groundwater Level", *Environmental Monitoring and Assessment*, Vol. 129, No. 1-3: 277-294.

Bowman, A.W., and A. Azzalini. 2012. The "sm" package for R. Smoothing methods for nonparametric regression and density estimation. 2012. www.stats.gla.ac.uk/~adrian/sm.

Bowman, A.W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. Oxford University Press: Oxford, 1997.

Jones, W.R., and M. Spence. 2012. "GroundWater Spatio-Temporal Data Analysis Tool (GWSDAT Version 2.0) User Manual." Shell Global Solutions, UK.

Jones, W.R., M.J. Spence, A.W. Bowman, L. Evers, and D. A. Molinari. 2014. "A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data." *Environmental Modelling & Software*. Vol. 55: 242-249. http://www.sciencedirect.com/science/article/pii/S1364815214000309

Paul H. C. Eilers, Dcmr Milieudienst Rijnmond, and Brian D. Marx. 1996. "Flexible smoothing with b-splines and penalties." *Statistical Science*, 11:89-121.

R Development Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria, 2008. ISBN 3-900051-07-0, http://www.r-project.org.

Thyne, G., C. Guler, and E. Poeter. 2004. "Sequential Analysis of Hydrochemical Data for Watershed Characterization", *Ground Water*, V. 42, Issue 42: 711-723.

D.8 JMP

Approximate Cost: Individual annual license $1400; academic pricing and corporate pricing available

Source: http://www.jmp.com

Current Version: JMP 10

Operating System Needs: Windows or Mac OSX

Input Structure: Supports database connections, text or Microsoft Excel file formats, column header names; one data row per measurement; tab-delimited preferred but not required.

Overview

JMP is a general statistical analysis platform that uses a graphical user interface to display and analyze data. JMP is not specifically tailored for statistical analysis of groundwater monitoring data. JMP is software for interactive statistical graphics and includes:

- a spreadsheet for viewing, editing, entering, and manipulating data
- several graphical and statistical methods for data analysis
- a design of experiments module
- options to select and display subsets of the data
- data management tools for sorting and combining tables
- a calculator for each table column to compute values
- a facility for grouping data and computing summary statistics
- special plots, charts, and communication capability
- tools for moving analysis results between applications and for printing
- a scripting language for saving frequently used routines

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution		●
Kaplan-Meier		●
ROS		●
Cohen/MLE		●
Exploratory/Diagnostic Tools		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Summary Statistics	●	●
Distributional tests	●	●
Outlier tests	●	●
Data transformations	●	●
Statistical Design		
Statistical Power	◒	●
SWFPR		●
Contaminant ranking		●
Monitoring network optimization		●
Statistical Limits		
Confidence Limits	●	●
Tolerance Limits	●	●
Prediction Limits	●	●
Testing Compliance Limits	●	●
Graphics		
Plots/Charts	●	●
Batch plots	●	●
Tweaking of graphics	●	●
Statistical Comparisons		
t-tests	●	●
ANOVA	●	●
Spatial Analysis		
Geostatistics/Mapping	◒	◒
Kriging/Interpolation	◒	◒
Spatial smoothing	●	●
Regression/Time Series		
Trend Tests	●	●
Mann-Kendall	●	●
Linear regression	●	●
Nonlinear regression	●	●

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Theil-Sen line		●
Time Series analysis	●	●
Multivariate Analysis		
Multiple regression	●	●
Factor/Discriminant analysis	●	●
Bootstrapping	●	●

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

JMP includes JMP Scripting Language (JSL). JMP can also call R (the open-source statistical computing environment R; see www.r-project.org) or Statistical Analysis System (SAS) functions for additional capabilities.

Ease of Use and Data Import

JMP offers descriptive statistics and simple analyses for beginning users, as well as more complex model fitting for advanced users. The program organizes statistics into logical areas with appropriate graphs and tables, which help users find patterns in data, identify outlying points, or fit models. Appropriate analyses are defined and performed based on the types of variables input. JMP provides several statistical and graphical methods organized into a small number of interactive platforms.

Types of Distributions

JMP accepts data of any distributional type. You can evaluate the statistical distribution and make data transformations within JMP. JMP also offers nonparametric statistical tests.

Visualization

JMP includes sophisticated built-in graphics for data visualization, including contour mapping, complex nonlinear trends, post-plots, and shape file annotation. Through the JMP scripting language, the process of analyzing data and producing plots can be automated. Graphics can be modified by using the interface or by writing scripts. JMP has a journal option that allows you to export data summaries, statistical analyses, or plots into Microsoft Word or other formats.

Primary Uses for Groundwater Data Analysis

JMP is a general statistical analysis platform and is not tailored for statistical analysis of groundwater monitoring data. JMP can be used for various exploratory or more sophisticated analyses and can be specialized with JMP scripting language or call to R or SAS procedures. JMP can be used for reliability analysis.

Benefits

- visual tools to assess data, flag values and look for trends among contaminants
- can be used to evaluate multiple media
- exploratory statistical tools for assessing data characteristics
- fitting of nonlinear and seasonal time series data
- multivariate methods to look for patterns in data

Limitations and Data Requirements

- JMP is not specifically designed for groundwater monitoring data and therefore does not have all of the tests recommended for groundwater statistical analyses.
- Analyses of nondetects requires special data handling or writing scripts to handle these situations.

References

Sall, J. 2007. *JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP*, 4th ed. SAS Press.

Thomas, L., and C. Drebs. 1997. "A Review of Statistical Power Analysis Software," *Bulletin of the Ecological Society of America*. Vol. 78, No. 2: 128-139.

D.9 MATLAB + Statistics Toolbox

Approximate Cost: $2,150 for an individual license of MATLAB®; $1,000 for an individual license of Statistics Toolbox

Source: MathWorks (www.mathworks.com)

Current Version: R2013a

Operating System Needs: Windows 7, Windows Vista, Windows XP, Mac OS X

Input Structure: Text files or Excel files are easiest, but any data format can be imported using a script.

Overview

MATLAB® is a software application that consists of a high-level programming language, interactive environment, and execution environment for data analysis, visualization, and numerical computation. The basic MATLAB software allows you to fit regression lines, calculate summary statistics, and plot data. MATLAB is flexible and can perform additional analyses using scripts and add-ins.

MATLAB also has a wide variety of visualization options including line plots, bar plots, histograms, pie charts, topological maps, and images. Many of these plots are available for both 2-D and 3-D plots and can be animated to show changes over time. More advanced users can generate a script to import, analyze, and plot data. Scripting is helpful when the same series of plots are generated on a regular basis (such as plots for quarterly reports).

With the Statistics Toolbox and some basic scripting, MATLAB can also be used for many other types of data analysis and visualizations. Additional types of plots available in the statistics package include box plots, probability distributions, additional types of histograms (including 3-D histograms and scatter histograms), quantile-quantile plots, and multivariate analysis plots (including dendograms, biplots, and parallel coordinate charts). You can also perform hypothesis tests, analysis of variance (ANOVA), cluster analysis, more complicated regression and classification, bootstrapping, confidence interval calculations, and data transformations.

With this additional functionality, MATLAB can generate customized plots of data to analyze distributions, compare or display data, and visualize temporal changes. For example, it is possible to compare two data sets to determine if concentrations changed over time, to determine appropriate background levels, and to compare site data with background levels or cleanup goals. The Statistics Toolbox has a number of interactive applications for analysis of covariance, distribution fitting, density and distribution plots, contour plots, polynomial fitting, random number generation, regression diagnostics, robust regression, and response surface demonstration.

The capabilities designated in the table below for "Capability with Scripts/Add-ins" are based on available scripts.

Statistical Functions

Statistical Method	Capability As Is (using MATLAB + Statistics Toolbox)	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	●
Kaplan-Meier		●
ROS	●	●
Cohen/MLE	●	●
Exploratory/Diagnostic Tools		
Summary Statistics	●	●
Distributional tests	●	●
Outlier tests	●	●
Data transformations	●	●
Statistical Design		
Statistical Power	●	●
SWFPR		●
Contaminant ranking		●
Monitoring network optimization		●
Statistical Limits		
Confidence Limits	●	●
Tolerance Limits	●	●
Prediction Limits	●	●
Testing Compliance Limits	●	●
Graphics		
Plots/Charts	●	●
Batch plots	●	●
Tweaking of graphics	●	●
Statistical Comparisons		

Statistical Method	Capability As Is (using MATLAB + Statistics Toolbox)	Capability with Scripts/Add-Ins
t-tests	●	●
ANOVA	●	●
Spatial Analysis		
Geostatistics/Mapping	●	●
Kriging/Interpolation	●	●
Spatial smoothing	●	●
Regression/Time Series		
Trend Tests	●	●
Mann-Kendall		●
Linear regression	●	●
Nonlinear regression	●	●
Theil-Sen line		●
Time Series analysis	●	● (with Econometrics Toolbox)
Multivariate Analysis		
Multiple regression	●	●
Factor/Discriminant analysis	●	●
Bootstrapping	●	●

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

Add-ins relevant to groundwater statistics include Statistics Toolbox, Curve Fitting Toolbox (for fitting curves and surfaces to data as well as nonparametric modeling techniques, such as splines, interpolation, and smoothing), and Neural Network Toolbox (for data-fitting, pattern recognition, and clustering). There are also many other add-ins not directly relevant to groundwater statistics.

Ease of Use and Data Import

Versions of MATLAB are available for a variety of Windows, Macintosh, and Linux platforms. Most basic tasks in MATLAB require use of a script or understanding of basic commands, making it moderately difficult to use. Advanced users can generate scripts, which have the potential to perform almost any desired task. MathWorks offers both introductory and advanced classes.

MATLAB is matrix-based and has the capacity to handle and manipulate large quantities of data rapidly. Many data or file types can be directly imported into MATLAB, including spreadsheet, text, or image files. Advanced users can write scripts to import data from any format. The Database Toolbox add-in allows you to read data directly from databases. Once imported, MATLAB is able to easily transform or perform other calculations on large data sets. Plots and image files from MATLAB can easily be saved as most common image types, written to Excel, or printed (either to a printer or to a pdf).

Types of Distributions

MATLAB accepts data of any distributional type. You can apply data transformations using a script or the Statistics Toolbox. The Statistics Toolbox includes functions and graphical tools to work with both parametric and nonparametric distributions, both continuous and discrete distributions, and both univariate and multivariate distributions. You can fit distributions to data, evaluate goodness of fit, generate statistical plots, generate probability density functions and cumulative distribution functions, and generate random and quasi-random number streams from distributions.

Visualization

MATLAB can be used to generate or display basic plots (including line, bar, and pie plots), topographic plots, images, and 3-D plots (including objects, volumes, and lines). With the added statistics package, MATLAB can also generate box plots, histograms, probability distributions, and quantile-quantile plots. You can customize plot styles as well, including lighting or camera angle on 3-D graphs. Series of plots can be animated to show temporal or geographical changes. While advanced users may use scripts to generate plots, other users can use a graphical user interface to perform most customization options.

Primary Uses for Groundwater Data Analysis

MATLAB can perform a wide variety of tasks related to data analysis, visualization, and numerical computation. MATLAB is particularly well-suited for computations in large data sets (including data transformations, hypothesis testing, background evaluations and regression/trend analysis). MATLAB can also generate customizable plots and images and can be an excellent tool for generating figures that must be updated on a regular basis, such as quarterly monitoring reports.

Benefits

- flexible and provides customization options to generate plots and images
- multi-purpose software that can be used for other applications
- capable of handling computations in large data sets
- able to evaluate background data to determine appropriate background levels and test for consistency of data with background
- includes scripts for easy updating of figures for quarterly reports
- variety of trend tests available
- able to handle parametric, nonparametric, continuous, discrete, univariate, and multivariate distributions
- can generate and customize plots using the graphical user interface and have MATLAB create the script associated with generating the plot

Limitations and Data Requirements

- cost
- moderate to difficult to use and requires some experience with basic programming.
- advanced use (scripting) required for many features
- limited functionality without the addition of the statistics toolbox
- challenging to import data that are not in spreadsheet, text, or image files
- limited customization of plots using the graphical user interface

D.10 MINITAB

Approximate Cost: $1395 single-user license, $2940 for 5-user multi-user license

Source: www.minitab.com/en-US/default.aspx

Current Version: 16

Operating System Needs: XP, Vista, Windows 7, Windows 8

Input Structure: Similar to Excel spreadsheet (values for variables entered in columns).

Overview

Minitab was primarily designed for the evaluating manufacturing and service quality data rather than environmental data but is used across many industries.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	◒	N/A
Kaplan-Meier	◒	●
ROS		◒
Cohen/MLE	●	N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	◒	◒
Data transformations	●	N/A
Statistical Design		
Statistical Power	◒	N/A
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	◒	◒

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Prediction Limits	◒	◒
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	◒	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall		●
Linear regression	●	N/A
Nonlinear regression	●	N/A
Theil-Sen line		◒
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression	●	N/A
Factor/Discriminant analysis	●	N/A
Bootstrapping		◒

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

Many freeware macros are available (including some that are useful for censored data and environmental applications). Several are posted at the following websites:

- http://www.practicalstats.com
- http://www.minitab.com

Ease of Use and Data Import

Some operations (such as, generation of statistical plots and descriptive statistics) can be done with little or no training. This software contains an extensive **Help** menu, tutorials, sample data sets, and an **Assistant** menu to help select appropriate statistical procedures. It is relatively simple to import data from sources such as Excel, but environmental data with flags or nonnumerical characters can be interpreted as text and must be manually converted to numerical values (for instance, by separating or deleting the non-numeric characters). You can assign formulas to columns to automatically perform calculations. Data appear in spreadsheets that can be readily manipulated including merged, sorted, and segregated data. Graphs readily update when spreadsheet data are revised. Results are readily output to PowerPoint and MS Word.

Types of Distributions

Data can be fit to normal, lognormal, and Poisson distributions. The program generates random data and outputs probability densities, cumulative probabilities, and inverse cumulative probabilities for many distributions (such as normal, lognormal, chi square, gamma, beta, exponential, F-distribution, and Poisson).

Visualization

The program's powerful, simple, and versatile graphic capability includes scatter plots, box plots, pie charts, histograms, individual value plots, and various 3-D plots. The software is useful for exploratory data analysis (EDA) and generating reports.

Primary Uses for Groundwater Data Analysis

This package can be used for groundwater applications for graphing capabilities for EDA, one-sample and multi-sample parametric and nonparametric hypothesis tests, time series evaluations, and control charts.

Benefits

- Basic functions are simple to use.
- The program offers clear and extensive help system.
- The cost is moderate (about $1,400; periodic updates cost about $600).
- Programming language and free macros are available to expand the software's capability.

Limitations and Data Requirements

- Not designed for environmental applications
- Lack of as is capability for some common environmental statistical applications (such as prediction intervals, outlier tests, Theil-Sen lines, and user-friendly menus for left-censored data).

D.11 MONITORING AND REMEDIATION OPTIMIZATION SYSTEM SOFTWARE (MAROS)

Approximate Cost: Free

Source: GSI website (http://www.gsi-net.com/en/software/free-software/maros-30.html)

Current Version: version 3.0

Operating System Needs: Microsoft Office Access 2007 or 2010; Windows 7 (older versions of software are available compatible with Windows XP and earlier versions of MS Office)

Input Structure: Microsoft Excel spreadsheet or Access database

Overview

The Monitoring and Remediation Optimization System (MAROS) software was developed by GSI Environmental Inc. (GSI) on behalf of Air Force Civil Engineer Center (AFCEC), known previously as Air Force Center for Engineering and the Environment (AFCEE), in 1998. It is a public-domain, data management and evaluation tool specifically designed to improve long-term groundwater monitoring (LTM) programs using both qualitative and quantitative methods. MAROS provides both:

1. Optimization routines, to help determine the appropriate number of sample locations, sampling frequency, and laboratory analytes for the specified monitoring objective
2. Statistical analysis tools to evaluate the plume stability condition and the effectiveness of natural attenuation and active remediation efforts

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	●	N/A
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Data transformations	●	N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR		N/A
Contaminant ranking	●	N/A
Monitoring network optimization	●	N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits		N/A
Prediction Limits		N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots		N/A
Tweaking of graphics		N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression		N/A
Theil-Sen line		N/A
Time Series analysis		N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

The MAROS tool analyzes groundwater data collected from a network of permanent or semi-permanent groundwater monitoring locations. Modules within the MAROS tool evaluate affected groundwater sites where substantial site characterization has already occurred. MAROS is not specifically designed for detection monitoring programs. The intended use of this software is to review data and evaluate the efficacy of the network to address site monitoring and remediation objectives.

The MAROS software uses site-specific data from the full historical record for statistical analysis and review. Groundwater analytical data can be imported into the software from AFCEC's Environmental Resource Program Information Management System (ERPIMS) database or Microsoft Excel or Access files. Minimum data requirements for the MAROS tool include data from at least six monitoring locations with detected concentrations of contaminants and at least four sampling events. The software can evaluate data from as many as 200 wells (speed of calculations may be slow with more than 200 wells). Input files include data such as well name, X, Y coordinates, sample date, contaminant name, analytical result, detection limit, and data flags.

The software can currently import data for over 2,000 different chemicals that may be dissolved in groundwater. Internal MAROS databases contain regulatory screening levels from various programs and physical parameters for these chemicals. Data from these databases are used to prioritize contaminants for long term monitoring optimization (LTMO) analyses.

Types of Distributions

The software contains modules that prioritize contaminants, calculate summary statistics, find outliers, identify data distributions, determine temporal trends at individual wells (using both Mann-Kendall and linear regression techniques), and calculate plume stability metrics and their trends over time (such as total dissolved mass, center of mass, and spread of mass).

Visualization

MAROS has several data visualization options that allow you to review concentration versus time graphs and tables prior to more in-depth analyses. You can export results of analyses either as formatted reports, spreadsheet output, or database output at various locations in the software.

Primary Uses for Groundwater Data Analysis

The program uses individual well statistics to prioritize wells in the network and identify locations that have attained cleanup goals. Optimization analyses identify redundant locations using qualitative decision logic and a nearest neighbor spatial geometry approach, estimate plume

concentration uncertainty to recommend new well locations, and use a sampling frequency module to recommend optimal sampling intervals.

Benefits

- Uses simple statistics and decision frameworks to prioritize data collection efforts and link data to defensible site management decisions
- Can use results from this software to develop lines of evidence, which combined with professional judgment, can be used to inform site management decisions for safe and economical long-term monitoring of groundwater plumes.
- Can use MAROS to help design and calculate remediation performance metrics and as a tool to evaluate progress toward site remedial goals

Limitations

- MAROS provides a two-dimensional analysis of the plume, so analyze nested wells separately (wells with the same X, Y coordinates but different depths).
- Evaluate multiple saturated units separately. Some modules assume a single source location.
- MAROS does not evaluate the plume outside of the imported well locations.
- Some statistical modules have limitations in the number of sample events that can be analyzed. For example, in the current version, the Mann-Kendall trend tool has an upper limit of 40 samples while the outlier analysis (Dixon's method) is limited to 25 samples.

References

AFCEC (Air Force Civil Engineer Center). 2012. "Monitoring and Remediation Optimization System (MAROS) Software, User's Guide and Technical Manual." In: Air Force Center for Environmental Excellence.

AFCEE (Air Force Center for Engineering and the Environment).1997. AFCEE Long-Term Monitoring Optimization Guide Version 1.1. Brooks AFB, Brooks City, TX, Air Force Center for Environmental Excellence.

AFCEE. 2004. *Monitoring and Remediation Optimization Software User's Guide*. Air Force Center for Environmental Excellence. http://www.gsi-net.com/software/MAROS_V2_1Manual.pdf.

Aziz, J. A., C. J. Newell, M. Ling, H. S. Rifai and J. R. Gonzales. 2003. "MAROS: A Decision Support System for Optimizing Monitoring Plans." *Ground Water* 41(3): 355-367.

Ling, M., H.S. Rifai, C.J. Newell, J.J. Aziz, and J.R Gonzales. 2003. "Groundwater Monitoring Plans at Small-Scale Sites: An Innovative Spatial and Temporal Methodology", *Journal of Environmental Monitoring*, 5: 126-134.

Ling, M., H. S. Rifai, J. J. Aziz, C. J. Newell, J. R. Gonzales, and J. M. Santillan. 2004a. "Strategies and Decision-Support Tools for optimizing Long-Term Groundwater Monitoring Plans-MAROS 2.0", *Bioremediation Journal*, 8 (3-4): 109-128.

Ling, M., H. S. Rifai, and C. J. Newell. 2004b. "Optimizing Long-Term Monitoring Networks Using Delaunay Triangulation Spatial Analysis Techniques", *Environmentrics*, Accepted, August 2004.

Vanderford, M. 2010. "A Comprehensive Approach to Plume Stability." *Remediation* Winter 2010: 21-37.

D.12 NCSS 8

Approximate Cost: $249 (one seat, government); complete price lists are found on the NCSS web page.

Source: NCSS, LLC (http://www.ncss.com/about_ncss.html)

Current Version: v8

Operating System Needs: Windows 32 or 64 bit

Input Structure: Similar to ProUCL in that the result column is followed by a detect column to identify if the results are measured values or detection limits.

Overview

General statistical program with a range of procedures and tools.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	◒	N/A
Kaplan-Meier	●	N/A
ROS	◒	N/A
Cohen/MLE	◒	N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR	●	N/A
Contaminant ranking	●	N/A
Monitoring network optimization	●	N/A
Statistical Limits		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression	●	N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression	●	N/A
Factor/Discriminant analysis	●	N/A
Bootstrapping	◒	N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

Requires a Windows 32 or 64 bit operating system. This program imports 35 different file types including Excel (all versions) spreadsheets. Data import is straightforward and examples are provided. Data may be processed or input within NCSS. The program provides drop down menus and a toolbar offers short cuts to frequently used functions. The program also provides a macro recording function for repeated procedures. Default templates are provided for typical parameters required for report tables and graphs, or you can easily create and save new templates.

Types of Distributions

NCSS 8 accepts data of any distributional type. The program provides flexibility in performing transformations and in selecting appropriate tests. The Nondetects-Data Group Comparison procedure computes summary statistics, generates EDF plots, and computes hypothesis tests appropriate for two or more groups for data with nondetects (left-censored) values.

Visualization

All standard graphics are available plus some unusual ones such as empirical distribution function and violin plots.

Primary Uses for Groundwater Data Analysis

You can use NCSS 8 for any environmental statistics problem that does not require spatial analysis.

Benefits

- Many advanced tests and procedures useful to expert users
- Output easily inserted in reports and documents which can then be manipulated through various common Windows applications
- Well documented; statistical tests clearly explained (although at a moderate to advanced level) ; introductory topics for chapters (which give an overview and comparison of different approaches)
- Example problems and trial data sets provided for many procedures
- Extensive built-in data transformation functions, including IF-THEN functions, from within the program (unlike some statistical programs that require all data pre-processing to occur outside the program)
- Performs data simulations
- Easy to convert results into a database or to combine different databases in the program

Limitations and Data Requirements

- No geostatistical capabilities.
- Not dedicated to groundwater monitoring applications.
- Requires a moderate level of statistical expertise to use appropriately.

D.13 PAM

Approximate Cost: Free (freeware)

Version: 0.62 Beta

Source: www.henlopen.net/pam/index.htm

Operating System Needs: Windows (Macintosh OS X and Linux versions are available by request).

Usage Restrictions: PAM is freeware and can be downloaded and used free of cost. However, PAM is not open-source and cannot be distributed or resold or included in any other package. Full license details are provided with the downloadable installation program.

Input Structure: Data are provided to PAM in text files.

Overview

PAM is designed to support CERCLA (Superfund) analyses for the USEPA Region 5 GEOS Program and focuses on intrawell tests. The program can analyze multiple monitoring locations, contaminants, and events for a broad view of the changing groundwater conditions and contaminant concentrations. PAM creates outputs in standardized formats, including PDF documents, HTML pages, and PNG images. The program also evaluates trends, comparisons to standards, and comparison of recent data to baseline

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE	●	N/A
Exploratory/Diagnostic Tools		
Summary Statistics		N/A
Distributional tests	●	N/A
Outlier tests		N/A
Data transformations		N/A
Statistical Design		
Statistical Power		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	◒	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line	●	N/A
Time Series analysis		N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

According to the PAM website, the software can be customized to incorporate additional input data formats, new statistical tests, and revised summary report tables. Authorized customization is performed by Subterranean Research, Inc.

Ease of Use and Data Import

PAM is easy to run because it is designed to support CERCLA analyses, rather than a general purpose statistical package. Data input to PAM is provided in text files that can be easily generated from spreadsheets or databases.

Types of Distribution

PAM supports some basic distributional tests, but is not geared toward a comprehensive parametric, distributional analysis of data.

Visualization

The program displays results using charts and color-coded data tables. Charts show the original data, subsets of data used for each statistical test, and various confidence, prediction, and regulatory limits in a concise format.

Primary Uses for Groundwater Data Analysis

For each specified contaminant at a monitoring well, three types of tests can be performed:

- **Comparison to Standard Test:** The data are compared to a standard, such as a maximum contaminant level (MCL) or alternate compliance limit (ACL), to determine whether the data are in compliance with the standard.
- **Comparison to Baseline Test:** The most recent data value is compared to the prediction interval of the population.
- **Trend Hypothesis Test and Estimate:** The Mann-Kendall and Theil-Sen trend line test are used to determine whether a statistical trend is present in the time-series data set.

Benefits

- User-friendly environment for exploring data trends and analyzing whether the time series data are in compliance.
- Runs in batch mode to process large data sets
- Built atop MATLAB Component Runtime for fast computation
- Free

Limitations

PAM is not a general-purpose statistical software and can only be used to perform trend and compliance-type tests.

D.14 ProUCL 5.0.00

Approximate Cost: Free

Source: USEPA (http://www.epa.gov/land-research/proucl-software)

Current Version: v5.0.00

Operating System Needs: Windows 32 or 64 bit, XP, Vista, Windows 7

Software Needs: Microsoft.NET version 4.0 Framework

Input Structure: Cross-tabular Excel file (XLS or XLSX format) or comma-separated values (CSV) format

Overview

ProUCL has been developed by Lockheed Martin under a contract with USEPA. ProUCL has been developed to address statistical issues arising in the various Superfund and RCRA site projects, and is available from USEPA at no cost. The software is designed specifically for various environmental monitoring applications including background contaminant evaluations, risk analysis for understanding concentrations, trend identification, and provides both interwell and intrawell procedures for evaluating groundwater contamination, using several parametric and nonparametric (including bootstrap methods) approaches as well. Graphical analyses offered include normal, lognormal, and gamma quantile-quantile (Q-Q) plots, probability plots, histograms, box plots, and line/trend plots. In addition to parametric and nonparametric upper limits, ProUCL 5.0 provides Gehan and Tarone-Ware tests to compare two data sets with multiple detection limits. Results for statistical intervals are offered with several options and relevant cautions. ProUCL 5.0 has rigorous methods to compute statistical upper limits including confidence limits, prediction limits, tolerance limits, and simultaneous limits for data sets with and without nondetect observations covering a wide-range of skewness and sample sizes. A partial list of references used in the decision making process included in ProUCL is provided.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	●	N/A
ROS	●	N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations		
Statistical Design		
Statistical Power	◒	N/A
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	◒	N/A
Tolerance Limits	◒	N/A
Prediction Limits	◒	N/A
Testing Compliance Limits	◒	N/A
Graphics		
Plots/Charts	◒	N/A
Batch plots		N/A
Tweaking of graphics		N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	◒	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests		N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression		N/A
Theil-Sen line	●	N/A
Time series analysis		N/A
Multivariate Analysis		
Multiple regression	◒	N/A
Factor/Discriminant analysis		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Bootstrapping	◒	N/A

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

Use of ProUCL is straightforward. Although ProUCL requires no formal background in statistics, you should understand the assumptions and input requirements for any statistical tests used in making decisions. Input data sets are uncomplicated, requiring columns of detected values for contaminants and corresponding columns indicating whether each value is a detect or a nondetect at the quantitation limit. You can also add variables to provide grouping data, regression variables, or sample dates. ProUCL modules can handle missing data.

You can select desired statistical tests from drop-down menus, and relevant options from subsequent menus. You can also view results within the program and export them to an Excel spreadsheet.

Data can be evaluated for fit to normal, lognormal, or gamma distributions; statistical interval test results are available for all of these distributions in addition to several nonparametric options.

Types of Distribution

ProUCL provides goodness-of-fit tests for normal, lognormal, and gamma distributions for uncensored data sets (without nondetects) as well as left-censored data sets (with nondetect observations.)

Types of Upper Limits
ProUCL 5.0 computes parametric and nonparametric statistical upper limits including confidence limits, prediction limits for k future observations and mean of k observations, tolerance limits, and upper simultaneous limits for censored and uncensored data sets (Singh and Nocerino 1997). Statistical limits computation methods available in ProUCL 5.0 cover a wide range of skewness and sample size. All of these limits can be computed using GROS, LROS, and nonparametric Kaplan-Meier methods. ProUCL also provides bootstrap methods to compute confidence and tolerance limits.

Types of Two-Sample Hypothesis Tests for Data Sets with Nondetects

In addition to two-sample t-tests and the Wilcoxon rank sum test, ProUCL 5.0 provides two-sample hypothesis tests (Gehan generalized Wilcoxon test and Tarone-Ware) for left-censored data with nondetect observations.

Visualization

Plots available in ProUCL include box plots, histograms, quantile-quantile (Q-Q) plots for normal, lognormal, and gamma distributions, and normal probability plots. Multiple normal Q-Q plots by groups provide a point-by-point comparison of data from multiple groups (monitoring wells). These graphs can also be used on data sets with nondetect observations. Although the program offers some options for editing the output plots, the process is limited.

Primary Uses for Groundwater Data Analyses

As the name implies, ProUCL was initially developed to provide a package for computing statistical intervals. Iterations of the software have provided additional statistical tools as well as improvements to the original. Version 5.0 provides for upper confidence limits, upper prediction limits, upper tolerance limits, and upper simultaneous limits for data sets with and without nondetect observations covering a wide range of data skewness and sample sizes. Depending upon the sample size, data distribution, and number of detects present, ProUCL 5.0 provides suggestions and cautions on the output results. Other available tests that may apply to groundwater monitoring include analysis of variance (ANOVA), trend evaluation, outlier, and goodness-of-fit tests. The sample size module of ProUCL computes DQO-based sample sizes needed to address statistical requirements of environmental projects. The sample size module can also be used to perform power evaluations in retrospect.

Although ProUCL software does not perform more in-depth statistical evaluations or address issues of definition and investigation, it is a good fit for many sites.

Benefits

- free
- relative simplicity of use
- developed specifically for environmental applications
- results output with recommendations, cautions, and cited references
- documentation well written and generally easy to understand

Limitations

- primary use: calculation of upper statistical limits, background versus site comparisons, inter-well and intrawell comparisons, outlier identification, sample size determination and power evaluations, and trend evaluations in ground water data, and hypothesis testing
- limited opportunity for user intervention or modification of procedures

References

Singh, Anita and Nocerino, John. 1997. "Robust Intervals for Some Environmental Applications." *The Journal of Chemometrics and Intelligent Laboratory Systems*, Vol. 37: 55-69.

Singh, A.K., Singh, A., and Engelhardt, M. 1997. *The Lognormal Distribution in Environmental Applications*. Technology Support Center Issue Paper, 182CMB97. EPA/600/R-97/006, December.

Singh, A., Maichle, R., and Lee, S. 2006. *On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations*. EPA/600/R-06/022, March. http://www.epa.gov/land-research/proucl-software.

Singh, A., Singh, A.K., and Iaci, R.J. 2002. *Estimation of the Exposure Point Concentration Term Using a Gamma Distribution*. EPA/600/R-02/084, October. http://www.epa.gov-/osp/hstl/tsc/software.htm.

D.15 R FOR STATISTICS

Approximate Cost: Free

Source: http://www.r-project.org

Operating System Needs: Operates on Windows, Mac OS, and most versions of UNIX.

Input Structure: Scripts can be written in R to read and analyze data from a wide variety of data sources including, but not limited to text/binary files, spreadsheets, and databases.

Overview

According to the R FAQ (Hornik 2013), "R is a system for statistical computation and graphics consisting of a programming language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files." "R is an integrated suite of software facilities for data manipulation, calculation, and graphical display" according to the "An Introduction to R" document (Venables et al. 2013).

The statistical functions in R provide support for linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing, analysis of spatial data, and Bayesian analysis, among others. In addition to storing and manipulating data, a mature collection of functions help in the production of report-quality graphics. R can be downloaded free from the Comprehensive R archive network (CRAN; http://www.r-project.org/). It is distributed under a GNU-style copyleft (http://www.gnu.org/copyleft/copyleft.html) license and is part of the GNU project (http://www.gnu.org).

Functions and corresponding data sets are typically organized in units called 'packages'. The directory where packages are stored is called the library. R comes with a standard set of packages in the standard library. Other packages can be downloaded and installed as needed. Once installed, these packages must be loaded into the session to be used. The list of packages in the standard library and detailed descriptions and documentation for each of the packages can be found at http://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html. In addition to the standard packages, the user can install additional packages from the CRAN website or elsewhere. Additional contributed packages can be found at the CRAN website at http://CRAN.R-project.org/ and related sites such as Bioconductor (http://www.bioconductor.org/) and Omegahat (http://www.omegahat.org/). Advanced users can program their own packages for custom applications.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	◒	●
ROS	◒	●
Cohen/MLE	◒	●
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR	●	N/A
Contaminant ranking	●	N/A
Monitoring network optimization		◒
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Geostatistics/Mapping	◒	●
Kriging/Interpolation	◒	●
Spatial smoothing	◒	●
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression	●	N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression	●	N/A
Factor/Discriminant analysis	●	N/A
Bootstrapping	●	N/A

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

Several existing add-on packages extend the functionality of R. A partial list can be found at http://cran.r-project.org/doc/FAQ/R-FAQ.html●Add_002don-packages-from-CRAN.

Ease of Use and Data Import

The most common data structures in R are vectors and data frames. Higher order data structures such as lists and data frames are also available for advanced analysis. The R environment may challenge a new user; however, an interactive user interface and comprehensive help documentation are provided. In addition, active development is underway to generate graphical user interfaces that provide a method to access commonly used functions.

Types of Distributions

R can be used for calculating properties of probability distributions as well as to check whether a given data set fits a standard distribution. A number of distributions and distributional tests are supported in R, including: beta, binomial, Cauchy, chi-squared, exponential, F, gamma, geometric, hypergeometric, lognormal, logistic, negative binomial, normal, Poisson, Student's T, uniform, and Weibull.

Visualization

R has a mature graphics library and can produce presentation quality graphics for most of the commonly used plots, such as stem and leaf, box plots, scatter plots,histograms, and contours.

Primary Uses for Groundwater Data Analysis

R is commonly used to perform the following tasks:

- calculate summary statistics
- perform distributional tests
- get point estimates of population mean
- get interval estimates of population mean with known and unknown variance
- perform sampling size of population mean
- calculate point and interval estimates of population proportion
- test hypotheses
- perform linear and nonlinear regression
- perform analysis on time-series and spatial data
- snalyze nondetects in data using substitution-type methods and also more advanced maximum likelihood estimator methods
- develop custom applications

Benefits

- provides a flexible, interactive, and powerful environment for data analysis and visualization
- free
- built-in support for a variety of simple to the complex statistical analyses
- scripts for performing complex analysis
- easily produces presentation-quality graphics and automated reports
- active and knowledgeable online community for support issues.
- detailed online documentation

Limitations and Data Requirements

- The program provides the functions and libraries to read and process data from a variety of sources including, but not limited to ASCII Files, binary Files, spreadsheets, and databases.

- As long as the data format and structure is known, data can be imported into the R environment.
- The environment challenging to the first-time user, and presents a steep initial learning curve.

References

Faraway, J. 2002. Practical Regression and ANOVA Using R. http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Hornik, K. 2013. The R FAQ. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.r-project.org.

Venables W.N., D.M. Smith, and the R Core Team. 2013. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics*. Version 3.0.1.

D.16 SANITAS FOR GROUNDWATER

Approximate Cost: about $1,000 plus annual maintenance fee

Source: Sanitas Technologies (www.sanitastech.com)

Current Version: v9.2

Operating System Needs: Windows 7 and 8, XP, Vista

Software Needs: Microsoft.NET version 4.0 Framework

Input Structure: Flat file or cross-tabular

Overview

Sanitas is a commercially available software package designed specifically for evaluating groundwater monitoring data, particularly at RCRA Subtitle C and D facilities. Although earlier releases were less versatile, more recent versions allow for selection between USEPA (1992), California, Unified Guidance, or American Society for Testing and Materials (ASTM) standards, provide some user-defined treatment of nondetects, and relax earlier input file restrictions.

The software provides both interwell and intrawell procedures for evaluating groundwater contamination, and employs parametric and nonparametric approaches. You can select from several options for treatment of nondetects, outlier testing, data transforms and adjust for seasonality, low detection frequency, or trends in background. Graphical analyses offered include probability plots, piper and stiff diagrams, control charts (Shewhart-Cusum), and statistical interval (prediction and tolerance) test results.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE	◒	N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	◒	N/A
Outlier tests	●	N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Data Transformations		N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR	●	N/A
Contaminant ranking	◒	N/A
Monitoring network optimization	◒	N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	◒	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests		N/A
Mann-Kendall	●	N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line	●	N/A
Time Series Analysis		N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

The most recent (v9.0 and newer) versions of Sanitas software can run on any Windows version that supports the .NET framework; however, the older versions of Sanitas are not compatible with the newer Windows platforms. Older versions of Sanitas also require that the input data file be in cross-tabular format with some stringent formatting requirements; v9.0 and newer allow flat file input, but will also accept the earlier cross-tabular format.

The various statistical procedures available in Sanitas are all accessed from drop-down menus. You determine the general category of test, then choose more specific criteria from the menus. Depending on the number of sample points, collection dates, and contaminants monitored, each test may take several minutes to complete.

Although no statistical expertise is required to run Sanitas and the operational basics are relatively simple, as with any software, you should make sure that you understand the assumptions and input requirements for any statistical tests used in making decisions based on the output. Sanitas Technologies also offers training workshops at regular intervals for new users.

Type of Distributions

The Sanitas package includes distributions and distributional analyses such as parametric (normal, transformed normal, and Poisson-based distributions), nonparametric, Shapiro-Wilk normality, Shapiro-Francia normality, chi-squared, coefficient of variation normality, skewness and kurtosis, and Levene's test, equality of variance.

Visualization

Sanitas includes graphics for data visualization, including customized graphs, power curves, plots, and production of summary tables.

Primary Uses for Groundwater Data Analysis

Sanitas is a statistical application developed specifically for groundwater monitoring applications. Many of the statistical procedures needed for groundwater data analyses are included.

Benefits

- straightforward use
- unlimited number of users for each site license
- developed specifically for groundwater monitoring applications
- relatively inexpensive
- documentation is well-written and generally easy to understand

Limitations and Data Requirements

- Each site must be licensed independently (price reduction after several sites licensed).
- Evaluation of distribution is limited to normal and ladder-of-powers transforms.
- Configuration errors (for example, selection of nondetect treatment) may go undetected by the user.

D.17 STATISTICAL ANALYSIS SYSTEM (SAS)

Approximate Cost: about $3,500 and up, depending on version and add-ins

Source: SAS (www.sas.com/technologies/analytics/statistics/stat/index.html)

Current Version: v9.3

Operating System Needs: Windows, Unix/Linux

Data Input Structure: Varied, user-defined

Overview

SAS is a suite of statistical software packages that apply to a wide variety of sectors including financial services, medicine, insurance, manufacturing, and oil and gas. Although not designed specifically for environmental applications, SAS is practically unlimited in the types and levels of statistical analyses that it can support. Some background in statistics is required, however, to use all of the features in SAS and to ensure that approaches are valid and inferences are sound.

SAS Base contains many general statistical procedures, as well as graphical capabilities, and is useful for most standard groundwater monitoring applications. More complex matrix-based procedures (such as kriging) are provided in SAS IML, which is available for separate purchase. Similarly, process control applications (such as Shewhart-Cusum charts) are specifically addressed in SAS QC software, which can also be purchased separately.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier		●
ROS		●
Cohen/MLE		●
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Statistical Design		
Statistical Power	●	N/A
SWFPR	●	N/A
Contaminant ranking	●	N/A
Monitoring network optimization	●	N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping	◒	●
Kriging/Interpolation	◒	●
Spatial smoothing	◒	●
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression	●	N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression	●	N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Factor/Discriminant analysis	●	N/A
Bootstrapping	●	N/A

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

Add-in packages are available for a variety of applications and business sectors. SAS/IML and SAS/QC provide added functionality for groundwater monitoring statistics.

Ease of Use and Data Import

SAS version 9.3 (the most recent commercially available version) is available for use on most Windows platforms. SAS is versatile with regard to data file structure, and can read most file types (text files, CSV, and others), but you generally must define the structure and data format for the program.

Because of its power and versatility, you must understand statistical approaches and methodologies to ensure valid and defensible results from SAS. You can write "free-form" code, calling statistical procedures, and within one program regroup and revise data, perform multi-layered analyses, output new data sets, and provide presentation-grade graphical analyses. SAS can be used to perform reliability analysis.

Visualization

SAS can produce contours, plots, and graphs. For examples of SAS visuals, visit the website http://www.sas.com/technologies/analytics/statistics/stat/index.html.

Primary Uses for Groundwater Data Analysis

SAS performs statistics for a wide variety of applications, but specific methods and tests for groundwater monitoring include statistical intervals, hypothesis testing, regression, correlation, outlier tests, spatial weighting, trend analysis, and a variety of graphical methods. While these procedures are not necessarily available in a drop-down menu or as a specific application, all of them can be coded into an SAS program by a proficient user. User-coded methods include nonparametric or non-normal procedures as well as those noted above as available in additional purchased software.

Benefits

- online help and available support system

- many examples provided in the documentation

Limitations and Data Requirements

- cost
- requires at some knowledge of statistical methods
- may be challenging for inexperienced users who may find the initial learning curve steep

D.18 Scout 2008 v1.00.01

Approximate Cost: Free

Source USEPA

Current Version: 2008 v 1.00.01 (2009)

Operating System Needs: Windows 98 or newer

Software Needs: Microsoft .NET version 1.1 Framework

Input Structure: Microsoft Excel spreadsheet (XLS) or comma-separated values (CSV) file

Overview

Scout was developed by Lockheed-Martin under contract with the USEPA. This program is a comprehensive, public domain data analysis software package that performs statistical methods used for evaluating data sets for groundwater monitoring optimization, background contaminant evaluations, and risk analysis for quantifying cleanup criteria.

Scout was designed to allow practitioners and decision makers to learn and apply accepted statistical techniques. This tool allows you to incorporate qualitative considerations into the analysis and output may be adjusted based on user-specified considerations and rationales.

In addition, two stand-alone software packages, ProUCL4.0 and Parallax, are incorporated into Scout. ProUCL 4.0 (2009) has been updated to ProUCL 5.0 (2013) which is not part of Scout 2008 v 1.00.01.ProUCL 5.0 is a statistical software package useful for performing statistical evaluations including hypotheses testing and calculating upper limits of data sets with and without nondetect observations with multiple reporting limits. Parallax is a software package that offers graphical and classification tools to analyze multivariate data using parallax coordinates.

Four statistical modules are used with Scout, as described below.

Data Module

The data module generates univariate data sets from normal, lognormal, gamma, and uniform distributions, and multivariate data sets from normal distributions. The software can also perform transformation operations on univariate and multivariate data for data sets with and without nondetects. The software handles nondetects through substitution and ROS methods, as well as estimation of missing observations.

Graph Module

The graphs module generates plots for single or grouped data sets. Graphical capabilities are described further in the Visualization section.

Statistical Module

For univariate data sets, Scout can perform a variety of descriptive statistics, classical interval estimates, and goodness-of-fit tests. The software is capable of parametric and nonparametric methods including Kaplan-Meier, ROS, and bootstrap methods on left-censored data sets. The software also performs univariate and multivariate outlier evaluations, robust estimation, single-sample and two sample hypotheses tests (including quantile tests), and the Wilcoxon-Mann-Whitney test.

The goodness-of-fit tests may be used to theoretically test and verify the normality of a data set and also to identify outliers. The goodness-of-fit tests may also be incorporated into the quantile-quantile plots along with correlation coefficients and critical values for a specified significance level, α.

This module performs outlier identification and estimation for both univariate and multivariate data sets using both classical and robust methods. The univariate methods for uncensored and left-censored data sets include Dixon's test, Rosner's test, and Grubbs tests as well as Tukey's robust biweight method. Multivariate outlier identification and estimation methods include but are not limited to: Max MD and multivariate kurtosis sequential classical methods, iterative robust and resistant M-estimation methods based on Huber and PROP influence functions, MCD, and MVT method.

The module computes the various parametric (normal, lognormal, and gamma distribution based) and nonparametric (for example, bootstrap, central limit theorem) upper limits including confidence limit, prediction limit, tolerance limit, and simultaneous limit. Additionally, the software also computes parametric and nonparametric two-sided confidence intervals, prediction intervals, tolerance intervals, and simultaneous intervals.

This module can be used to perform one-way parametric and nonparametric ANOVA. The trend tests option can perform trend evaluations on time-series data sets using the Mann-Kendall test and Theil-Sen nonparametric trend line, supplemented with graphical displays.

Regression Module

The regression module can perform classical and robust regressions using several linear methods including least median of squared regression; least percentile of squared regression for classical methods; and M-estimation, Huber, biweight, and PROP influence for robust methods. In addition, the module also generates and displays prediction and confidence limits around fitted regression models for first order linear models. The graphical displays available within the module can be used to identify outliers, leverage points, and compare the performance of the various classical and robust regression methods.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Simple Substitution	●	N/A
Kaplan-Meier	●	N/A
ROS	●	N/A
Cohen/MLE	●	N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power	◒	N/A
SWFPR	◒	N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	●	N/A
Prediction Limits	●	N/A
Testing Compliance Limits	◒	N/A
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	◒	N/A
Spatial Analysis		
Geostatistics/Mapping	◒	N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression		N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression	●	N/A
Factor/Discriminant analysis		N/A
Bootstrapping	●	

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

Scout 2008 v 1.00.01 is a user friendly software with User Guide providing details with examples on how to implement the various methods available in Scout 2008. Scout 2008 can read Microsoft Excel spreadsheet (XLS) or CSV files. Output files generated by Scout can be saved in Excel spreadsheet or as Outlook offline storage table (OST) files.

Types of Distributions

Scout can perform statistical evaluations for normal, lognormal, and gamma distributed uncensored data sets and left-censored data sets consisting of nondetects with multiple reporting limits. Several robust, nonparametric, and bootstrap methods are also available in Scout 2008 v 1.00.01.

Visualization

Scout permits construction of plots for single and grouped data sets, including histograms, box plots, quantile-quantile plots, index plots, and 2D and 3D scatter plots. Scout can also generate univariate graphs for regression analyses including residual plots, observed versus predicted plots, and

bivariate regression line plots. Results from other statistical modules may also be displayed on some of the graphs. Graphical displays are interactive, enabling a limited amount of editing of the graphs, which may be performed within the software, and plots may be exported as an image file or copied into other image editing software.

Several of the other modules within Scout can generate graphs in order to determine data set distributions, compare grouped data, identify outliers, compare performances of various methods for outlier identification, compare leverage points, and provide univariate and multivariate classical and robust method quality assurance and control.

All modules of Scout generate graphical output displays (GST file), Excel spreadsheets, or both graphical displays and spreadsheets. Some of the graphics generated include side-by-side box plots, histograms, index plots, multiple quantile-quantile plots, interval graphs, control charts, and bivariate scatter plots of raw data.

Primary Uses for Groundwater Data Analysis

This software package performs statistical evaluations of data sets for groundwater or soil contaminants. Scout provides alternatives for sites at which practitioners must evaluate site contaminant cleanup criteria, perform trend analyses on groundwater contaminant monitoring data, or perform statistical evaluations on the data sets, including distribution assessments, outlier identification, and interval estimates to compute decision making statistics.

Benefits

- user friendly and offers easy presentation of results, which allows for interpretation and decision making
- capable of handling large data sets

Limitations and Data Requirements

Scout v1.00.01 (2009) is designed specifically for use in evaluation of soil and groundwater data and is not a general statistics package. There is no plan to upgrade the Scout software in the near future.

References

An overview and user guide is provided online and gives more detailed information on the capabilities of the software.

D.19 SPSS (IBM(R) SPSS Statistics 19 Base)

Approximate Cost: Depends on level of licensing and support: Standard Package, $2,300-$13,000 and Premium Package, $6,900-$39,000

Source: IBM (www.ibm.com/software/analytics/spss/products/statistics)

Current Version: v22 (2013)

Operating System Needs:

- IBM SPSS Statistics Base 22 for Windows: Microsoft Windows XP (Vista or Windows 7)
- IBM SPSS Statistics Base 22 for Mac: Apple Mac OS 10.5 (Leopard) or 10.6 (Snow Leopard),

Input Structure: Can accept data from multiple file formats

Overview

SPSS is a high-end, general purpose statistical package with a wide variety of capabilities. Originally developed for analyzing social science data, SPSS is now used in business analytics, medicine, academia, and some environmental settings. Like other general purpose packages, SPSS is not specifically tailored for groundwater analysis, yet can perform many of the tests typically conducted on groundwater data.

These tests include methods to compare groups such as t-tests and one-way analysis of variance (ANOVA), as well as trend analysis such as linear and nonlinear regression. SPSS also has multivariate methods that can be used to interpret patterns in groundwater data. Typically, multivariate analysis (such as principle component analysis, Q-mode factor analysis, and cluster analysis) examines correlation among variables in terms of a few weighted combinations of the component variables. Multivariate analysis can achieve great efficient compression of the original data, while gaining information to help interpret the environmental geochemical origin of contaminants.

Disclaimer: Statistical functions and capabilities presented for this software package have not been reviewed or verified by IBM.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	◒	N/A
ROS		N/A
Cohen/MLE		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization		N/A
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits		N/A
Prediction Limits	◒	
Testing Compliance Limits	●	N/A
Graphics		
Plots/Charts	●	●
Batch plots	●	●
Tweaking of graphics	◒	●
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping		N/A
Kriging/Interpolation		N/A
Spatial smoothing		N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression	●	●
Theil-Sen line		N/A
Time Series analysis	◒	●

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Multivariate Analysis		
Multiple regression	●	N/A
Factor/Discriminant analysis	●	N/A
Bootstrapping		●

Capability Ratings:
N/A = Not applicable or not available

● = Full capability

◒ = Some capability
(blank cell) = No capability

Add-Ins Available

Multiple add-ins are available for SPSS, including applications for bootstrapping, regression analyses, decision trees and others. SPSS also allows for integration of R to expand the range of available applications. A listing of available software packages is provided on the product website: www.ibm.com/software/analytics/spss/products/statistics.

Ease of Use and Data Import

SPSS Statistics 22 is a comprehensive system for analyzing data that can accept data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses. This program has simple menus and dialog box selections that make it possible to perform complex analyses without using command syntax.

SPSS has a data editor. This feature is user-friendly and resembles a spreadsheet. Using this feature, you can enter data directly into SPSS. In this editor, the columns represent the variables, and the rows represent the observations. You can also import data from a number of different sources, such as data stored in IBM SPSS Statistics data files; spreadsheet applications (such as Microsoft Excel); database applications (such as Microsoft Access); and text files.

Types of Distributions

SPSS is primarily a tool for data analysis rather than a tool to generate specific kinds of distributional data. The Simulation option, however, offers Monte Carlo simulation of a wide range of standard statistical distributions, including ones common to groundwater analyses like the normal, lognormal, gamma, exponential, Weibull, binomial, and Poisson distributions.

Visualization

This program generates commonly used charts such as scatter plots, histograms, and population pyramids. SPSS can create these charts more easily with Chart Builder. This chart creation interface allows you to create a chart by dragging variables and elements onto a chart creation canvas. The Graphics Production Language (GPL) can be used to customize charts.

Primary Uses for Groundwater Data Analyses

Since SPSS is not tailored for groundwater statistics, it is mostly limited in groundwater applications to standard statistical tests like t-tests, ANOVA, linear regression and their nonparametric counterparts. SPSS accommodates upper-tail censored data in survival analysis, but not lower-tail censored values such as nondetects.

Benefits

- easily available
- widely used, with active support community
- most standard statistical methods are available
- click and point interface as well as command interface

Limitations

- cost
- not tailored for groundwater statistical analyses
- not all typical groundwater statistical tests included in base package; must integrate with R and create customized functions
- difficult to produce customized analysis

D.20 STATISTICA

Approximate Cost: $1,195

Source: StatSoft (www.statsoft.com)

Current Version: STATISTICA 12

Operating System Needs: Windows 7 (recommended), Windows Vista, Windows XP

Input Structure: Can directly open spreadsheet, text, and database files

Overview

STATISTICA is user friendly, while still allowing for significant customization and functionality. The base package computes practically all common descriptive statistics and can produce a wide variety of customizable graphics. The base software includes graphics tools along with the following modules:

- descriptive statistics, breakdowns, and exploratory data analysis
- correlations
- basic statistics from results spreadsheets (tables)
- interactive probability calculator
- t-tests (and other tests of group differences)
- frequency tables, cross-tabulation tables, stub-and-banner tables, multiple response analysis
- multiple regression methods
- nonparametric statistics
- one-way analysis of variance (ANOVA)/multivariate ANOVA (MANOVA)
- distribution fitting

STATISTICA Query can be used to easily access data from databases using Microsoft's OLE DB conventions and allows easy statistical analysis of large and changing databases. STATISTICA interacts directly with the free software package R, allowing users to have additional features not present in the base software, while maintaining the customization and ease of use of STATISTICA.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	●
Kaplan-Meier		●

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
ROS		
Cohen/MLE		◒
Exploratory/Diagnostic Tools		
Summary Statistics	●	●
Distributional tests	●	●
Outlier tests	●	●
Data transformations	●	●
Statistical Design		
Statistical Power	●	●
SWFPR		●
Contaminant ranking	●	●
Monitoring network optimization		
Statistical Limits		
Confidence Limits	●	●
Tolerance Limits	●	●
Prediction Limits	●	●
Testing Compliance Limits	◒	●
Graphics		
Plots/Charts	●	●
Batch plots	●	●
Tweaking of graphics	●	●
Statistical Comparisons		
t-tests	●	●
ANOVA	●	●
Spatial Analysis		
Geostatistics/Mapping		
Kriging/Interpolation		
Spatial smoothing		
Regression/Time Series		
Trend Tests	◒	●
Mann-Kendall		
Linear regression	◒	●
Nonlinear regression	◒	●

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Theil-Sen line		
Time Series analysis	◒	●
Multivariate Analysis		
Multiple regression	◒	●
Factor/Discriminant analysis		●
Bootstrapping		●

Capability Ratings:
N/A = Not applicable or not available
● = Full capability
◒ = Some capability
(blank cell) = No capability

Add-Ins Available

STATISTICA also has a number of add-in packages and modules that can enhance the functionality of the base software package.

- The Data Mining add-in is a versatile tool that includes techniques for quick analysis of large data files.
- The Market Basket add-in uses sequence, association, and link analysis to build models and extract rules from large data sets.
- The multivariate statistical process control allows users to apply univariate and multivariate statistical methods for quality control, predictive modeling, and data reduction for complex processes; determine and optimize the most critical processes or factors; monitor process characteristics interactively; and build, evaluate, and use predictive models based on historical data.
- The quality control add-in includes a wide selection of quality control analysis techniques and additional quality control charts.
- Additional add-ins include advanced linear/nonlinear models, multivariate exploratory techniques, power analysis and interval estimation, and variance estimation and precision.

Ease of Use and Data Import

STATISTICA is designed as a user-friendly software package. For additional help with the program, you can watch a wide variety of training videos on the website or take a training seminar for STATISTICA basics or advanced topics.

STATISTICA can directly open many data types including databases, spreadsheets, and text files. Using STATISTICA Query and Visual Basic, you can easily query and import or export data

from databases for statistical analysis. Output data sheets and plots can be sent to workbooks, STATISTICA reports, or Microsoft Word. STATISTICA can also coordinate with the free software package R, allowing you to run R scripts and algorithms from STATISTICA and customize outputs and graphics directly in STATISTICA. You can also record a macro of process steps, allowing for easy reproduction and duplication of analysis.

Types of Distributions

STATISTICA contains a distribution fitting option which directly compares the distribution of data to a wide variety of distributions. Available distributions for fitting include normal, rectangular, exponential, gamma, lognormal, chi-squared, Weibull, Compertz, Binomial, Poisson, geometric, or Bernoulli distributions. Once a distribution has been fit, you can evaluate the fit using a variety of tests and plots.

Additional distribution fitting options are available in the STATISTICA Process Analysis add-in, including the option to calculate the maximum-likelihood parameter. The STATISTICA Advanced Linear/Non-Linear Model add-in allows you to fit data to complex, custom-defined functions.

Visualization

STATISTICA has a wide variety of plotting capabilities in both 2-D and 3-D. Plotting options in the base package include box plots, 2-D and 3-D histograms, bivariate distributions, 2-D and 3-D scatter plots, normal, half-normal, and detrended probability plots, quartile-quartile plots, probability-probability plots, contour plots, nonsmoothed surfaces, and icons. You can zoom in on portions of the graphs, which can be useful when visualizing larger data sets and when producing cross-section slices from 3-D graphics. STATISTICA also has the option of plotting multiple-subset scatter plots and categorized scatter plots. The program provides many options to customize and format figures and tables for reports and presentations.

Primary Uses for Groundwater Data Analysis

The STATISTICA base package and add-ins include a wide variety of customizable graphics, which are ideal for use in groundwater data analysis. These plots can be used to analyze distribution, illustrate general trends, and support conclusions derived from hypothesis testing and descriptive statistics. STATISTICA is also well known for its strong data mining add-in, which allows you to rapidly analyze large data files and data sets.

Benefits

- easy to use and a wide variety of basic and advanced training videos and seminars available
- able to import, store, and export data easily
- advanced user direct interaction with R for additional statistical capabilities
- STATISTICA Query to interact and query directly in databases
- ability to plot and customize a wide variety of graphics
- strong data mining capabilities with the STATISTICA data mining add-in module

- strong and easy-to-use multivariate analytic features not readily found in other statistical software
- macros provided to record steps for easy reproduction of data analysis

Limitations and Data Requirements

- cost
- some statistical tools missing or have limited functionality relative to other statistical software packages.
- requires purchase of add-ins and modules for complete functionality

D.21 SUMMIT TOOLS

Approximate Cost: Free (one seat, government, version 2.0)/$100 per sampling location per site (unlimited seats, version 3.0)

Source: Summit Envirosolutions, Inc. (www.summite.com/products.html)

Current Version: v3.0

Operating System Needs: Windows 32 or 64 bit

Input Structure: Simple cross-tab comma-separated values (CSV) file (date, location name, X coordinate, Y coordinate, parameter values)

Overview

SampleOptimizer and SampleTracker can be used for monitoring optimization to reduce sampling costs and perform statistical testing to find potential anomalies in sampling data.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier		N/A
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests		N/A
Outlier tests		N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power		N/A
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization	●	N/A
Statistical Limits		
Confidence Limits		N/A
Tolerance Limits		N/A
Prediction Limits	●	N/A
Testing Compliance Limits		N/A

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests		N/A
ANOVA		N/A
Spatial Analysis		
Geostatistics/Mapping	●	N/A
Kriging/Interpolation	●	N/A
Spatial smoothing	●	N/A
Regression/Time Series		
Trend Tests		N/A
Mann-Kendall		N/A
Linear regression		N/A
Nonlinear regression		N/A
Theil-Sen line		N/A
Time Series analysis		N/A
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

The product is compatible and integrated with Summit's EPIPHINY data management solution ($1000/seat) for import/export to and from multiple formats including spreadsheets, databases, and a variety of text formats.

Ease of Use and Data Import

The tools require Windows (32 or 64 bit, XP or later). You can input data using a CSV file, or in combination with EPIPHINY, to and from many other formats. Documentation and a brief tutorial

are provided. The software is designed around an easy-to-use interface that automatically configures the settings based on the input set and allows control of the entire optimization process.

Visualization

Graphics can be generated and modeling can be done with kriging or inverse distance weighting interpolation methods. Results can be exported to PNG file, to GeoTIFF, or to ESRI Arc Grid.

Primary Uses for Groundwater Data Analysis

SampleOptimizer and SampleTracker are tools for both spatial and spatio-temporal analysis for monitoring network optimization.

Benefits

- SampleOptimizer applies mathematical optimization to monitoring networks in an easy-to-use desktop software tool.
- SampleTracker reviews new monitoring data against historical data. The software identifies cases where current data deviates from expectations that are based on the historical data set.

Limitations and Data Requirements

- The software can perform spatio-temporal analysis but requires at least 4 rounds of past historical data. Sampling optimization is best achieved on sites with at least 15 sampling locations.
- Data import formats are limited to CSV files, unless EPIPHINY or another conversion tool is used (Excel can save as a CSV file, for example).

References

Summit Envirosolutions, Inc. 2013. Summit Envirosolutions, Our Products. http://www.summite.com/products.html.

D.22 SYSTAT

Approximate Cost: $999 (government/nonprofit Organization)

Source: www.systat.com

Operating System: Windows, UNIX, Macintosh

Input Structure: Data can be imported from a wide variety of data sources including but not limited to text/binary files, spreadsheets and databases. A BASIC-type language is available for extensive data management and manipulation within the program.

Overview

SYSTAT features 49 functions in the standard package library including bootstrapping and sampling, cluster analysis, linear regression, mixed regression, nonparametric tests, random sampling, smoothing, spatial analysis, and time series analysis.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution	●	N/A
Kaplan-Meier	●	N/A
ROS	◒	●
Cohen/MLE	●	N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	N/A
Distributional tests	●	N/A
Outlier tests	●	N/A
Data transformations	●	N/A
Statistical Design		
Statistical Power	●	N/A
SWFPR	◒	●
Contaminant ranking	●	N/A
Monitoring network optimization		◒
Statistical Limits		

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Confidence Limits	●	N/A
Tolerance Limits	◒	●
Prediction Limits	◒	●
Testing Compliance Limits	◒	●
Graphics		
Plots/Charts	●	N/A
Batch plots	●	N/A
Tweaking of graphics	●	N/A
Statistical Comparisons		
t-tests	●	N/A
ANOVA	●	N/A
Spatial Analysis		
Geostatistics/Mapping	●	N/A
Kriging/Interpolation	●	N/A
Spatial smoothing	●	N/A
Regression/Time Series		
Trend Tests	●	N/A
Mann-Kendall	●	N/A
Linear regression	●	N/A
Nonlinear regression	●	N/A
Theil-Sen line	●	N/A
Time Series analysis	●	N/A
Multivariate Analysis		
Multiple regression	●	N/A
Factor/Discriminant analysis	●	N/A
Bootstrapping	●	

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

SYSTAT Exact Tests extends the functionality of SYSTAT. This add-in uses an exact inference function to run statistical tests on sparse or imbalanced samples, where the usual asymptotical assumptions might produce misleading results.

A free student version called MYSTAT is also available. MYSTAT retains most of the full functionality of SYSTAT, except that only 100 variables are allowed, with no limits to the number of rows (cases).

Ease of Use and Data Import

The most common data structures in SYSTAT are rectangular (column/row) files, but triangular data files are also allowed. SYSTAT can import data import from ASCII files and from various popular formats such as Excel, dBase, SAS, and Minitab. Data files can be easily manipulated through the built-in BASIC-like language support. The user interface can be interactive (with all commonly used commands accessed through drop-down menus and dialog boxes), or command driven for more experienced users. Comprehensive online help documentation is available.

Types of Distributions

SYSTAT can be used for calculating properties of probability distributions, as well as to check whether a given data set fits a standard distribution. Some distributions and distributional tests supported in SYSTAT include beta, binomial, Cauchy, chi-square, exponential, F, gamma, Gompertz, Gumbel, lognormal, logistic, negative binomial, normal, Poisson, Student's T, uniform, and Weibull.

Visualization

SYSTAT has a comprehensive graphics library and can produce presentation-quality graphics for most of the commonly used plots. Graphs can be easily annotated and edited.

Primary Uses for Groundwater Data Analysis

SYSTAT is used for the following groundwater data analysis tasks:

- calculate summary statistics
- perform distributional tests
- calculate point estimates of population mean
- calculate interval estimates of population mean with known and unknown variance
- perform sampling size of population mean
- calculate point and interval estimates of population proportion
- test hypotheses

- perform linear and nonlinear regression
- perform analysis on time-series and spatial data
- analyze nondetects in data using substitution-type methods and also more advanced maximum likelihood estimator methods
- develop custom applications

Benefits

- flexible, interactive, and powerful environment for data analysis and visualization
- built-in support for a variety of statistical analyses ranging from simple to complex
- macros for performing complex analyses
- presentation-quality graphics and automated reports easily produced
- SYSTAT support staff and help through an extensive FAQ data base, as well as a knowledgeable online community for support issues

Limitations and Data Requirements

- Cost may be prohibitive.
- The program offers functions and libraries to read and process data from a variety of sources including but not limited to: ASCII files, binary files, spreadsheets, and databases.
- As long as the data format and structure is known, the data can be imported into SYSTAT. No data size limitations exist, other than hardware limitations of the computer running the software.

References

Krishnan, T. and R. Karandikar, Cranes Software International Ltd. An overview of the general-purpose SYSTAT software, with some examples. http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdf%20files/Manual%20I/13-SYSTAT%20TUTORIAL.pdf

Wilkinson, L, Blank, G. and Gruber, C.G.1996. *Desktop Data Analysis with SYSTAT*. Englewood Cliffs, NJ: Prentice-Hall.

D.23 Visual Sampling Plan (VSP) Software

Approximate Cost: Free

Source: http://vsp.pnnl.gov

Current Version: v6.5

Operating System Needs: Windows XP or later

Input Structure: Supports text, manual entry, ESRI file formats (SHP, ASCII grid), DBX formats, or copy and paste from other applications (such as spreadsheets)

Overview

The Visual Sampling Plan (VSP) software was developed by the Pacific Northwest National Laboratory. You can use VSP to develop a groundwater sampling plan that is based on statistical principles. You can use VSP to analyze groundwater sampling results. VSP Version 6.5 provides sample-size equations and algorithms for specific statistical tests that support environmental sampling objectives. VSP can also be used at sites with suspected contamination to support data quality assessment. The program is easy-to-use with graphical user interfaces (Matzke et al. 2010). Some of the statistical algorithms in VSP have undergone a rigorous validation assessment (Nuffer et al. 2009).

Designed for project managers, regulators and technical staff without expertise in statistics, VSP can be used for planning sampling of various media (groundwater, rooms and buildings, surface soil, piles, or water bodies) for studies of environmental quality. It can also be used to sample items such as drums, documents or equipment.

Statistical Functions

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Handling of NDs		
Simple Substitution		N/A
Kaplan-Meier	●	
ROS		N/A
Cohen/MLE		N/A
Exploratory/Diagnostic Tools		
Summary Statistics	●	
Distributional tests	◒	
Outlier tests	◒	

Statistical Method	Capability As Is	Capability with Scripts/Add-Ins
Data transformations		N/A
Statistical Design		
Statistical Power	●	
SWFPR		N/A
Contaminant ranking		N/A
Monitoring network optimization	●	
Statistical Limits		
Confidence Limits	●	N/A
Tolerance Limits	◒	N/A
Prediction Limits		N/A
Testing Compliance Limits		N/A
Graphics		
Plots/Charts	●	
Batch plots	◒	
Tweaking of graphics	◒	
Statistical Comparisons		
t-tests	●	
ANOVA	●	
Spatial Analysis		
Geostatistics/Mapping	●	
Kriging/Interpolation	●	
Spatial smoothing	●	
Regression/Time Series		
Trend Tests	●	
Mann-Kendall	●	
Linear regression	●	
Nonlinear regression	◒	
Theil-Sen line		N/A
Time Series analysis	●	
Multivariate Analysis		
Multiple regression		N/A
Factor/Discriminant analysis		N/A
Bootstrapping		N/A

Capability Ratings:

N/A = Not applicable or not available

● = Full capability

◒ = Some capability

(blank cell) = No capability

Add-Ins Available

None

Ease of Use and Data Import

VSP is menu driven software that performs statistical sample design and analysis. Map information can be imported from GIS and data for analyses can be imported from text files or pasted from the clipboard.

Types of Distributions

VSP offers statistical methods based on normally distributed data as well as nonparametric methods. In addition the user may apply log-transformations to their data.

Primary Uses for Groundwater Data Analysis

Users only interested in RCRA applications can select the RCRA Version, which focuses on designs for monitoring and trends. Trends can be evaluated either with or without seasonality based on the Mann-Kendall trend test. VSP also has a module to evaluate redundancy of wells. The well redundancy modules in VSP can identify redundant wells and identify a technically defensible temporal spacing of observations for wells. The redundant well module uses a geospatial analysis based on kriging. VSP also has a module to help with new well placement to reduce estimation uncertainty. The sampling frequency well evaluation is applied on a well-by-well basis and is based on iterative thinning methods.

Benefits

- free
- includes tools to develop statistically-based sampling design for a variety of environmental problems
- some tools developed specifically for groundwater monitoring sampling problems
- visualizes spatial mapping tools and facilitates export for use in other spatial tools like ArcGIS
- automatic reports produced for any work done in VSP that summarizes the statistical assumptions, analyses performed, and methods used.

Limitations and Data Requirements

The background tests developed for VSP are most applicable to solid media.

References

Nuffer, LL, NL Hassig, LH Sego, BA Pulsipher, JE Wilson, B Matzke. 2009. Validation of Statistical Sampling Algorithms in Visual Sample Plan (VSP): Summary Report. PNNL-18253. Pacific Northwest National Laboratory, Richland, Washington.

Matzke, B.D., J.E. Wilson, L.L. Nuffer S.T. Dowson, J.E. Hathaway, N.L. Hassig, L.H. Sego, C.J. Murray, B.A. Pulsipher, B. Roberts, S. McKenna. 2010. *Visual Sample Plan Version 6.0 User's Guide*. PNNL-19915. Pacific Northwest National Laboratory, Richland, Washington.

APPENDIX E. ITRC GROUNDWATER STATISTICS AND MONITORING COMPLIANCE TEAM SURVEY RESULTS

E.1 Survey Summary

The web-based Groundwater Statistics and Monitoring Compliance (GSMC) Survey was conducted using Survey Monkey during the summer of 2011.

- ITRC received 126 responses from 34 states.
- Most questions could be skipped. The only required questions were the state and role category questions.
- The highest participation was from the states as seen in Table E-1.

Table E-1. States represented in the survey responses

States Represented	
State	**Responses**
CA - California	15
IN - Indiana	12
FL - Florida	11
NY - New York	9
WA - Washington	8
AZ - Arizona	7
VA - Virginia	6

- Table E-2 includes the responses by role category.

Table E-2. Survey responses by role category

All Responses	
Category	**Responses**
State regulator	55
Federal employee, non-regulator	33
Consultant	21
Federal regulator	7
State employee, non-regulator	5
Industry representative	3
Public/tribal stakeholder	2

- Table E-3 includes the respondent answers for their technical background (individuals could choose more than one response).

Table E-3. Skill set of respondents

Skill Set	
Response	**Responses**
Hydrogeology	60
Geology	57
Environmental science	51
Project management	38
Environmental regulations and laws	36
Remediation technology	28
Chemistry	27
Engineering	26
Groundwater modeling	19
Risk assessment	19
Statistics	18
Biology	12
Toxicology	8
Geostatistics	6
Soil science	5
Physics, Health Physics	1
Community relations	1

E.2 Agency- or Program-Level Responses

The survey included questions that referred to the agency or program with which the respondent either worked (if a state or federal employee) or the agency or program with which the respondent mainly interacted. Results for several of these questions are presented.

- Respondents were asked if their program or agency has specific policies or guidance for the use of statistics for groundwater data. Figure E-1 summarizes the responses.

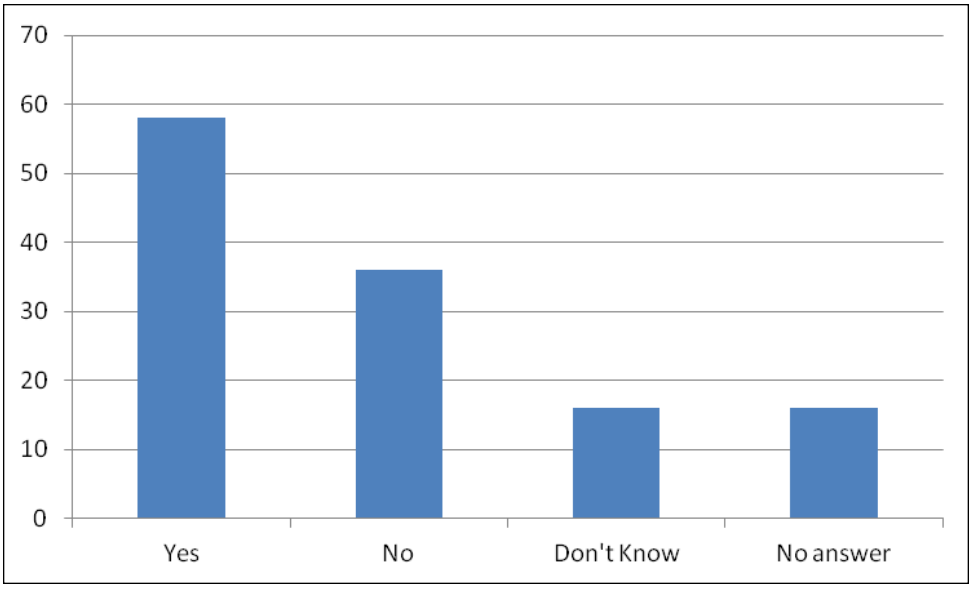


Figure E-1. Responses to the question: Does your program or agency have specific policies or guidance for the use of statistics for ground water data?

- Respondents were asked about who primarily reviews or performs statistical analyses within their program or agency. This question had “please check all that apply” and "other" write-in space. The four answers that were given are listed in Table E-4. The write-in answers were categorized approximately by one of the four responses, plus the “groundwater statistics are not used” answer. A few representative comments are included below the table.

Table E-4. Responses to question about who performs or reviews the statistical analyses

Who Performs or Reviews Statistics?	
Response	**Number**
Technical support group - in house	56
Each project manager	40
Technical support available through outside contractor	26
One specific expert who reviews or performs statistical analyses	21
Groundwater statistics not used	4

- Who performs or reviews statistical analyses?
    - Generally, each project manager reviews the project and consults, as needed, with the "specific expert."

- Statistics are assigned to staff geologists, regardless of their background with statistics.
- The technical support group is made up of chemists, engineers, and geologist. There is not a statistician in house.
- We do not use statistical analysis in the remedial programs.

- Respondents were asked: If groundwater statistics are not generally used, are any of the following factors? Table E-5 includes the responses to this question.

Table E-5. Factors as to why statistics are not used

Statistics Not Used	
Response	**Number**
Staff is not proficient in rigorous statistical analyses	35
Regulations do not require	24
No perceived need	21
Insufficient time and resources to review submittals	21
High cost of doing statistical evaluations	8
Regulations do not allow	5

- Respondents were asked about how challenging it is to use statistics for decision making. In addition, write-in answers describing the challenges were requested. Figure E-2 summarizes the responses for how challenging is it to use statistics.

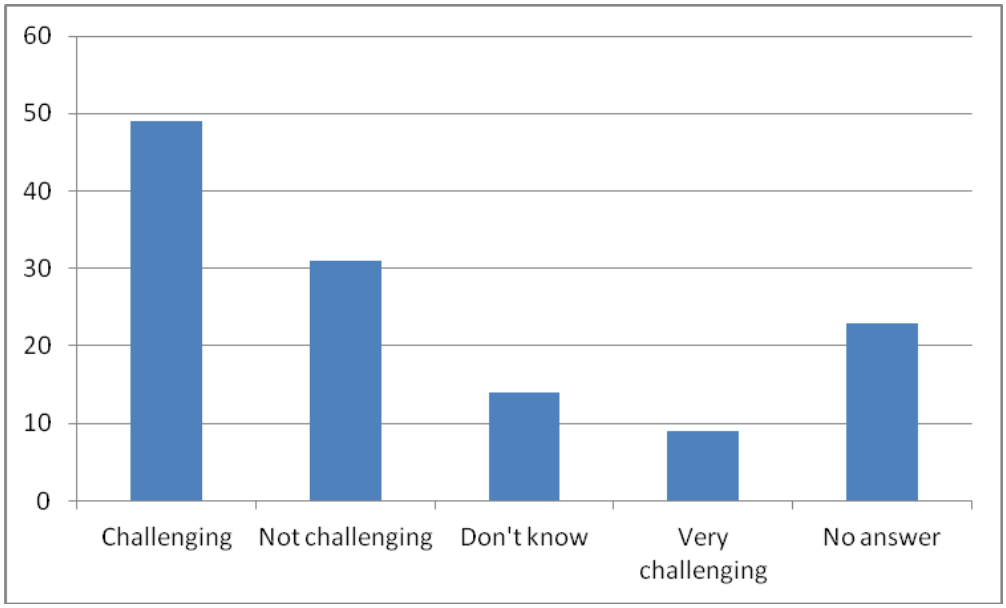


Figure E-2. Responses for how challenging is it to use statistics.

- The identified challenges included the following:
    - checking the accuracy of the data, applicability of the method to the data set
    - knowledge of reviewer or project manager and keeping current on statistical applications if they don't use every day; practical training; knowing when the presentation or results of statistical analyses are inaccurate or inappropriate
    - understanding of results by nonstatisticians
    - data that are suitable for statistical analyses (for example, poor planning, sparse data, not following DQO process)
    - misapplication of statistical analyses
    - knowledge of statistical software applications, formatting and exporting data for use in software
    - expense of software
    - expense of groundwater monitoring programs needed to generate the needed data set
    - not challenging, statistics not used in some or all programs

- Respondents were asked about the types of sites where policies or guidance are applicable. Figure E-3 summarizes the responses for this question.

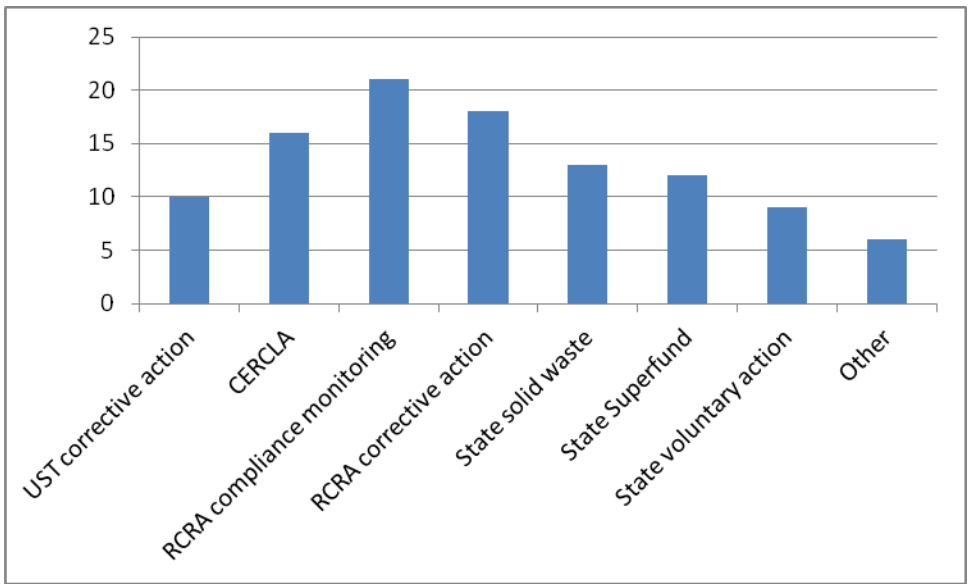


Figure E-3. Types of sites for which policies or guidance are applicable.

E.3 Individual-Level Responses

The survey included a group of questions about the individual's knowledge and experience with statistical analyses.

- Figure E-4 summarizes the responses about individual knowledge and experience.

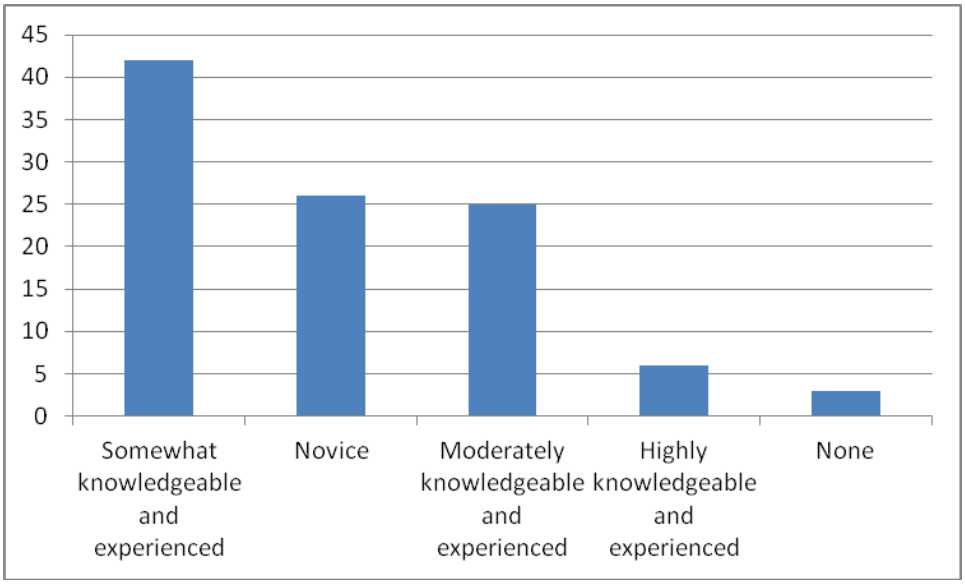


Figure E-4. Individual knowledge and experience with statistical analysis.

- Respondents were asked: How familiar are you with the USEPA's 2009 Unified Guidance? Figure E-5 includes the responses to this question.

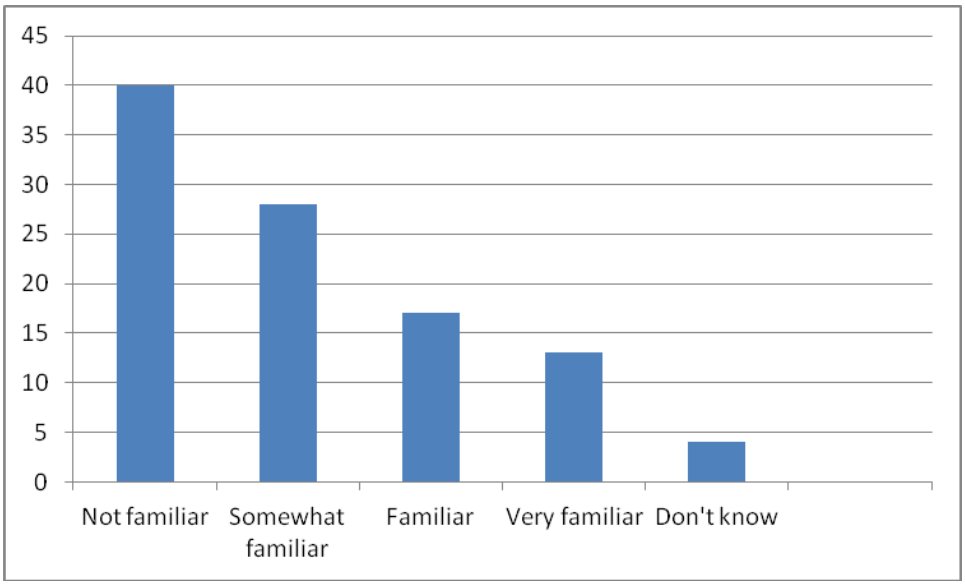


Figure E-5. Responses on familiarity with the USEPA's Unified Guidance.

- Respondents were asked: How do you use groundwater statistics? Figure E-6 summarizes the use of groundwater statistics by various tasks and project life cycle stages.

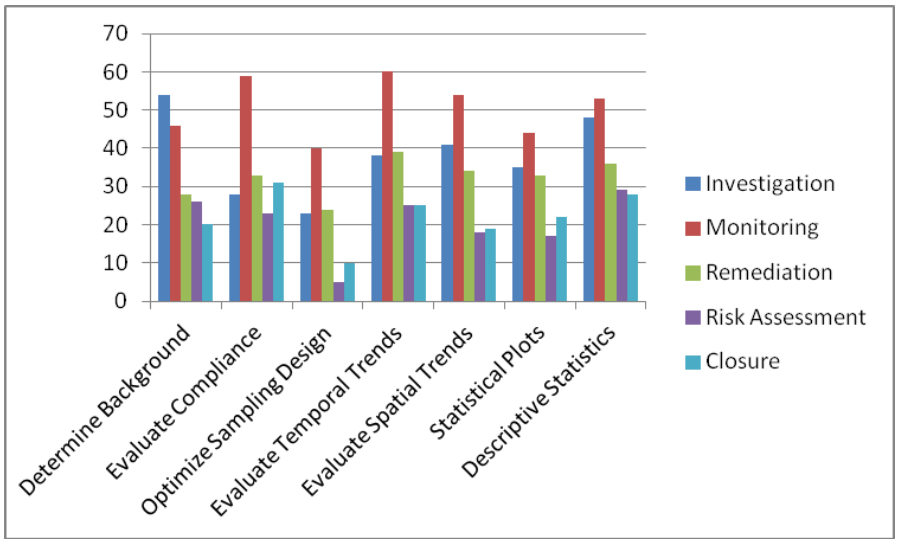


Figure E-6. Use of groundwater statistics by various tasks and project life cycle stages.

- Respondents identified the types of statistical tests that they use in reviewing groundwater data. A list of tests/methods identified by survey respondents is included in Table E-6 in order of most responses to fewest responses. In addition, several respondents indicated that they do not use or review any statistics or provided non-method-specific responses.

Table E-6. Statistical methods that are used or reviewed by survey respondents

Statistical Methods	Number
Simple descriptive statistics (such as sample mean, median, range)	67
Tests for simple trends (such as Mann-Kendall)	60
Confidence Intervals	50
Histograms	42
Box plots	37
Prediction limits	35
Tolerance limits	34
Normal probability or quantile plots	29
Two-sample hypothesis tests (such as t-tests)	29
Control charts	29
Multi-sample hypothesis tests (such as ANOVA)	24
One-sample hypothesis tests	18
Nonparametric Wilcoxon rank sum test	1
Outlier tests (such as Dixon's or Rosner's)	1
Principal component analysis - correspondence	1

Statistical Methods	Number
analysis (PCA-CA)	
Shewart-CUMSUM test	1
Time Series (ARIMA) discriminant analysis non-linear modeling of data	1

- Respondents were asked about the challenges or misuses of statistics. The answers included statements such as the following:
    - lack of knowledge
    - connection between conceptual site model (CSM), monitoring network, data quality, and data set collected
    - terminology issues and misunderstandings
    - drawing conclusions from insufficient data
    - ensuring that statistical assumptions are valid
    - using an overly simplified approach
    - for some software applications, a steep learning curve, expensive training, and expensive software
    - interpreting the uncertainties associated with statistical methods and tests
    - misrepresentation of the results, biased presentations
    - how to handle nondetect values
    - application of the wrong statistical test
    - issues with background
    - difficulties for the regulators in trying to recheck the calculations performed or using software to confirm that calculations were performed correctly
    - using statistics to avoid cleanup
    - software applications that are deceptively easy to use, but difficult for novice users to ensure that the calculations are appropriate
    - communicating the results to decision makers

- The respondents were asked to assign value to the topics that the GSMC team identified for the guidance document. Figure E-7 includes the responses.

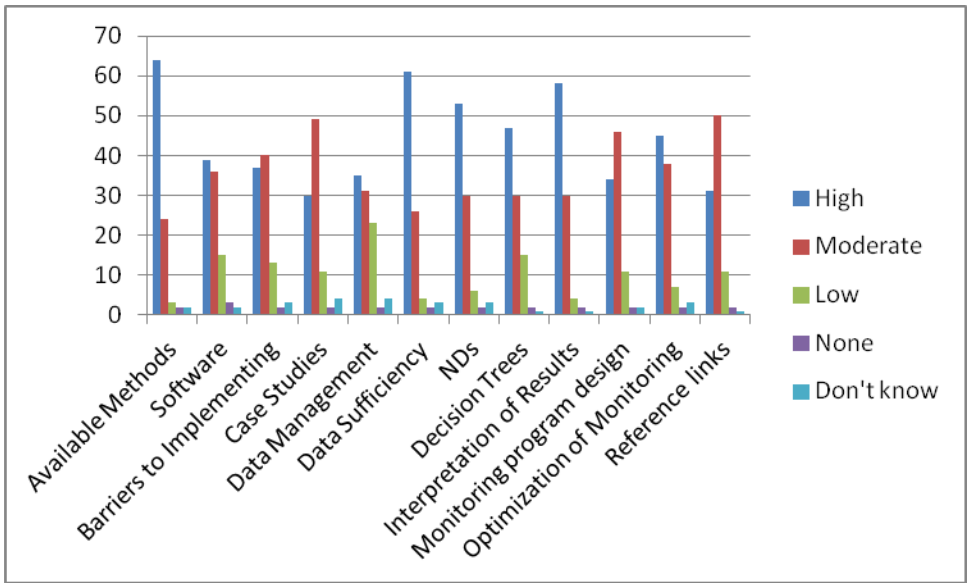


Figure E-7. Value of topics for the guidance document.

- The respondents were asked about the value of training for their agency or organization. These responses are included in Figure E-8.

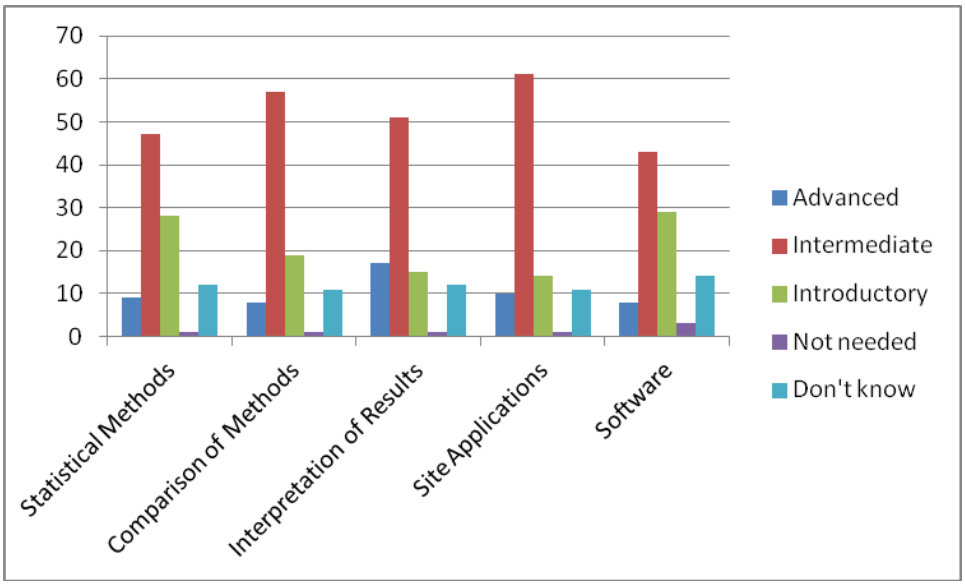


Figure E-8. Applicable level for training on identified topics for groundwater statistics.

APPENDIX F. METHODS TO VERIFY UNDERLYING ASSUMPTIONS FOR TESTS

Any formal inferential statistical test makes a series of assumptions about the underlying population from which the sample data are drawn and the manner in which those measurements are drawn. These assumptions might include 'the population is normal' or 'the populations being compared have the same variance' or 'the background population has a stable mean over the period in question.' In addition, certain kinds of data, such as nondetects, can be difficult to analyze unless special statistical adjustments are made.

The tables included in this section list the major assumptions that should be checked and adjustments that might be needed for each commonly-used groundwater statistical method. More information about the methods can be found in Section 5.0. In each case, the table offers diagnostic methods (where feasible) that can be used to check the assumption. Section 3.5 includes more information about the different assumptions used.

Table F-1. Checking the underlying method assumptions for confidence limits and trend tests

Tests	Confidence Intervals	Nonparametric Confidence Intervals	Confidence Bands	Linear Regression	Mann-Kendall	Theil-Sen
Normality	• probability plot • correlation coefficient • Shapiro-Wilk	N/A	• probability plot • correlation coefficient • Shapiro-Wilk (on residuals)	• probability plot • correlation coefficient • Shapiro-Wilk (on residuals)	N/A	N/A
Spatial variability	N/A	N/A	N/A	N/A	N/A	N/A
Homogeneity of variance	N/A	N/A	scatter plot (residuals vs. time; residuals vs. fitted trend)	scatter plot(residuals vs. time; residuals vs. fitted trend)	N/A	N/A
Temporal correlation (for example, seasonality)	• sample autocorrelation function (ACF) • rank von Neumann ratio test	• sample autocorrelation function (ACF) • rank von Neu-	• sample autocorrelation function (ACF) • rank von Neumann ratio test	• sample autocorrelation function (ACF) • rank von Neumann ratio test	Usually not tested (if seasonality, use seasonal	Usually not tested (if seasonality, use seasonal Mann-Kendall or first de-seasonalize data)

Table F-1. Checking the underlying method assumptions for confidence limits and trend tests (continued)

Tests	Confidence Intervals	Nonparametric Confidence Intervals	Confidence Bands	Linear Regression	Mann-Kendall	Theil-Sen
		mann ratio test (but difficult to test if large fraction of NDs)	(on residuals)	(on residuals)	Mann-Kendall or first de-seasonalize data)	
Temporal stability (presence of trends)	• time-series plots • Mann-Kendall • linear regression	• time-series plots • Mann-Kendall • linear regression	• time-series plots • Mann-Kendall • linear regression	N/A	N/A	N/A
Outliers	• probability plot • Dixon's test • Rosner's test	box plot screening; difficult to test formally (outlier tests assume normality)	• probability plot • Dixon's test • Rosner's test (residuals)	Difficult to test formally (must be able to compute normal residuals; use Mann-Kendall or Theil-Sen)	N/A (resistant to outliers)	N/A (resistant to outliers)
Nondetects	• simple substitution • Kaplan-Meier • regression on order statistics	N/A	simple substitution (<10-15% NDs if using linear regression); switch to Theil-Sen if more NDs	Use Mann-Kendall, Theil-Sen, or Akritas-Theil-Sen	Score pairs of NDs as ties (if small fraction of NDs). Use Kendall's tau-beta if many NDs (see Helsel 2005)	simple substitution (if <10-15% NDs; use Akritas-Theil-Sen if many NDs)

Table F-2. Checking the underlying method assumptions for tolerance limits, prediction limits, and control charts

Tests	Tolerance Limits	Nonparametric Tolerance Limits	Prediction Limits	Nonparametric Prediction Limits	Shewhart-CUSUM Control Chart
Normality	• probability plot • correlation coefficient • Shapiro-Wilk (on background data)	N/A	• probability plot • correlation coefficient • Shapiro-Wilk (on background data)	N/A	• probability plot • correlation coefficient • Shapiro-Wilk (on background data)
Spatial variability	• box plot • one-way ANOVA (for interwell limits)	• box plot • Kruskal-Wallis (for interwell limits)	• box plot • one-way ANOVA (for interwell limits)	• box plot • Kruskal-Wallis (for interwell limits)	• box plot • one-way ANOVA (for interwell charts)
Homogeneity of variance	Difficult to test except periodically (See Section 5.3, Assumptions)	Difficult to test except periodically (See Section 5.3, Assumptions)	Difficult to test except periodically (See Section 5.4, Assumptions)	Difficult to test except periodically (See Section 5.4, Assumptions)	Difficult to test except periodically (See Section 5.13, Assumptions)
Temporal correlation (for example, seasonality)	• sample autocorrelation function (ACF) • rank von Neumann ratio test	• sample autocorrelation function (ACF) • rank von Neumann ratio test (but difficult to test if large fraction of NDs)	• sample autocorrelation function (ACF) • rank von Neumann ratio test	• sample autocorrelation function (ACF) • rank von Neumann ratio test (but difficult to test if large fraction of NDs)	• sample autocorrelation function (ACF) • rank von Neumann ratio test
Temporal stability (presence of trends)	• time-series plots • Mann-Kendall • linear regression (on background data)	• time-series plots • Mann-Kendall (on background data)	• time-series plots • Mann-Kendall • linear regression (on background data)	• time-series plots • Mann-Kendall (on background data)	• time-series plots • Mann-Kendall • linear regression (on background data)
Outliers	• probability plot	box plot screening	• probability plot	box plot screening (on	• probability plot

Table F-2. Checking the underlying method assumptions for tolerance limits, prediction limits, and control charts (continued)

Tests	Tolerance Limits	Nonparametric Tolerance Limits	Prediction Limits	Nonparametric Prediction Limits	Shewhart-CUSUM Control Chart
	• Dixon's test • Rosner's test (on background data)	(on background data); difficult to test formally (outlier tests assume normality)	• Dixon's test • Rosner's test (on background data)	background data); difficult to test formally (outlier tests assume normality)	• Dixon's test • Rosner's test (on background data)
Nondetects	• simple substitution (if <10-15% NDs) • Kaplan-Meier • regression on order statistics	N/A	• simple substitution (if <10-15% NDs) • Kaplan-Meier • regression on order statistics	N/A	• simple substitution • Kaplan-Meier • regression on order statistics

Table F-3. Checking the underlying method assumptions for two-sample tests

Tests	Pooled Variance t-Test	Welch's t-Test	Wilcoxon Rank-Sum Test	Tarone-Ware
Normality	• probability plot • correlation coefficient • Shapiro-Wilk (on pooled residuals)	Shapiro-Wilk for Multiple Groups	N/A	N/A
Spatial variability	• box plot • one-way ANOVA (for interwell tests; on uncon-taminated/background data)	• box plot • one-way ANOVA (for interwell tests; on uncon-taminated/background data)	• box plot • Kruskal-Wallis (for interwell tests; on uncon-taminated/background data)	• box plot • Kruskal-Wallis (for interwell tests; on uncon-taminated/background data)
Homogeneity of vari-ance	• side-by-side box plots • Levene's test	N/A	side-by-side box plots(dif-ficult to test formally)	side-by-side box plots (dif-ficult to test formally)
Temporal cor-relation (for example, sea-sonality)	• sample autocorrelation function (ACF) • rank von Neumann ratio test (on each well)	• sample autocorrelation function (ACF) • rank von Neumann ratio test (on each well)	Difficult to test with many NDs	Difficult to test with many NDs
Temporal stability (presence of trends)	• time-series plots • linear regression • Mann-Kendall (on each well)	• time-series plots • linear regression • Mann-Kendall (on each well)	• time-series plots • Mann-Kendall (on each well)	• time-series plots • Mann-Kendall (on each well)
Outliers	• probability plot • Dixon's test • Rosner's test (on pooled residuals)	• probability plot • Dixon's test • Rosner's test (for each well)	box plot screening (on each well); difficult to test (outlier tests assume normality)	box plot screening (on each well); difficult to test (outlier tests assume normality)
Nondetects	simple substitution (if <10-15% NDs)	simple substitution (if <10-15% NDs)	Use midranks for tied NDs	N/A

Table F-4. Checking the underlying method assumptions for multi-sample tests

Tests	One-Way ANOVA	Kruskal-Wallis Test
Normality	• probability plot • correlation coefficient • Shapiro-Wilk	N/A
Spatial variability	• box plot • one-way ANOVA (on uncontaminated/background data)	• box plot • Kruskal-Wallis (on uncontaminated/background data)
Homogeneity of variance	• side-by-side box plots • Levene's test	side-by-side box plots(difficult to test formally)
Temporal correlation (for example, seasonality)	• sample autocorrelation function (ACF) • rank von Neumann ratio test (on each well)	difficult to test with many NDs
Temporal stability (presence of trends)	• time-series plots • linear regression • Mann-Kendall (on each well)	• time-series plots • Mann-Kendall (on each well)
Outliers	• probability plot • Dixon's test • Rosner's test	box plot screening (on each well); difficult to test (outlier tests assume normality)
Nondetects	simple substitution (if <10-15% NDs)	Use midranks for tied NDs

APPENDIX G. TEAM CONTACTS

Ning-Wu Chang, Team Leader
California Department of Toxic Substances Control
(714) 484-5485
Ning-Wu.Chang@dtsc.ca.gov

Lesley Hay Wilson, ITRC Program Advisor
Sage Risk Solutions LLC
(512) 327-0902
lhay_wilson@sageisk.com

Palmer Anderson
U.S. Navy
(805) 982-1488
palmer.anderson@navy.mil

Ernest Ashley
CDM Smith
(617) 452-6416
ashleyec@cdmsmith.com

Leroy (Buddy) Bealer
Shell
(484) 632-9755
Leroy.bealer@shell.com

Paul Beam
U.S. Department of Energy
(301) 903-8133
paul.beam@em.doe.gov

Richard (Kirby) Biggs
U.S. Environmental Protection Agency
(703) 823-3081
biggs.kirby@epa.gov

Jennifer Brekken
Barr Engineering
(952) 832-2700
jbrekken@barr.com

Marlena Brewer
Alaska Department of Environmental Conservation
(907) 269-1099
marlena.brewer@alaska.gov

Anna Butler
U.S. Army
(912) 652-5515
anna.h.butler@usace.army.mil

Kirk Cameron (Air Force and USEPA representative)
MacStat Consulting, Ltd.
(719) 532-0453
kcmacstat@qwest.net

Arnab Chakrabarti
Geosyntec Consultants
(510) 285-2755
achakrabarti@geosyntec.com

Devamita Chattopadhyay
CH2M Hill
(614) 209-4776
dchattop@ch2m.com

Wesley Dyck
CRA
(519) 884-0510
wdyck@CRAworld.com

Helge Gabert
Utah Department of Environmental Quality
(801) 536-0215
hgabert@utah.gov

Thomas Georgian
U.S. Army
(402) 697-2567
thomas.georgian@usace.army.mil

James Gleason
Trihydro Corporation
(307) 745-7474
jgleason@trihydro.com

Dibakar Goswami
Washington Department of Ecology
(509) 372-7902
dgos461@ecy.wa.gov

Nazmul Haque
Washington D.C. Department of Environment
(202) 535-1330
nazmul.haque@dc.gov

Allan Harris
U.S. Department of Energy
(513) 246-0542
Allan.Harris@emcbc.doe.gov

John Hathaway
U.S. Department of Energy
(509) 372-4970
john.hathaway@pnl.gov

Phillip Hunter
U.S. Air Force
(210) 395-8412
philip.hunter@us.af.mil

Stephen Johnson
Delaware Department of Natural Resources & Environmental Control
(302) 395-2600
stephen.johnson@state.de.us

Mavis Kent
Plateau Geosciences Group LLC
(360) 521-2592
maviskent@comcast.net

Prashanth Khambhammettu
S.S. Papadopulos & Associates, Inc.
(301) 718-8900
pk@sspa.com

Ray Ledbetter
U.S. Environmental Protection Agency
(702) 784-8008
ledbetter.ray@epa.gov

Gail Lipfert
Maine Department of Environmental Protection
(207) 287-7650
gail.e.lipfert@maine.gov

Mark Malander
ExxonMobil
(703) 846-6044
mark.w.malander@exxonmobil.com

Thomas McHugh
GSI Environmental Inc.
(713) 522-6300
temchugh@gsi-net.com

Orphius Mohammad
Oklahoma Department of Environmental Quality
(405) 702-5118
Orphius.Mohammad@deq.ok.gov

Beth Moore
U.S. Department of Energy
(202) 586-6334
beth.moore@em.doe.gov

Carolyn Moore
ERM
(925) 482-3710
carolyn.moore@erm.com

Kenda Neil
U.S. Navy
(805) 982-6060
kenda.neil@navy.mil

Ehsan Rasa
Geosyntec Consultants
(530) 574-8193
erasa@geosyntec.com

Randall Ryti
Neptune and Company
(505) 662-2121
rryti@neptuneinc.org

Lizanne Simmons
Kleinfelder
(858) 320-2267
lsimmons@kleinfelder.com

Anita Singh
Lockheed Martin (EPA representative)
(702) 563-9906
asingh428@gmail.com

David Smit
Mountain Area Land Trust
(303) 953-1924
smit9142@yahoo.com

Robert Soboleski
New Jersey Department of Environmental Protection
(609) 984-2990
bob.soboleski@dep.state.nj.us

Sarah Stoneking
ENVIRON
(703) 516-2407
sstoneking@environcorp.com

Chris Stubbs
ENVIRON
(510) 420-2552
cstubbs@environcorp.com

Harold Templin
Indiana Department of Environmental Management
(317) 232-8711
htemplin@idem.in.gov

Mindy Vanderford
GSI Environmental Inc.
(713) 522-6300
mvanderford@gsi-net.com

Teresie Walker
Sterling Global Operations, Inc.
(865) 988-6063
teresie.walker@sterlinggo.com

Edward Winner
Commonwealth of Kentucky
(502) 564-5981
edward.winner@ky.gov

APPENDIX H. ACRONYMS

3TMO	3-Tiered Monitoring and Optimization Tool
ACF	autocorrelation coefficient or function
ACL	alternate compliance limit
AFCEC	Air Force Civil Engineer Center (formerly AFCEE)
AFCEE	Air Force Center for Engineering and the Environment (changed to AFCEC)
ANOVA	analysis of variance
ARCH	autoregressive conditional heteroscedasticity
ARIMA	autoregressive integrated moving average
ASTM	American Society for Testing and Materials (now known just as ASTM International)
CASRN	Chemical Abstract Service Registry Number
CCF	cross-correlation function
CDF	cumulative distribution function
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CES	cost-effective sampling
CFR	Code of Federal Regulations
CSM	conceptual site model
CSV	comma separated values
CUSUM	cumulative sum control chart
DNAPL	dense non-aqueous phase liquid
DQA	data quality assessment
DQO	data quality objective
EDA	exploratory data analysis
EE/CA	engineering evaluation/cost analysis
ERIS	Environmental Restoration Information System
ERPIMS	Environmental Resources Program Information Management System
GARCH	generalized autoregressive conditional heteroscedasticity
GRASS	Geographic Resources Analysis Support System
GSMC	Groundwater Statistics and Monitoring Compliance
GTS	Geostatistical Temporal-Spatial optimization software
GWSDAT	Groundwater Spatiotemporal Data Analysis Tool

LCL	lower confidence level or limit
LTL	lower tolerance limit
LTM or LTMgt	long-term management
LTMO	long term monitoring optimization
LUST	leaking underground storage tank
MANOVA	multivariate analysis of variance
MAROS	Monitoring and Remediation Optimization Software
MCL	maximum contaminant level
MDL	method detection limit
MLE	maximum likelihood estimation
MNA	monitored natural attenuation
NAPL	non-aqueous phase liquid
ND	nondetect
NIRIS	Navy Installation Restoration Information Solution
NPDES	National Pollutant Discharge Elimination System
OST	offline storage table
PCB	polychlorinated biphenyl
PCE	perchloroethylene
PL	prediction limit
Q-Q	quantile-quantile
QA/QC	quality assurance/quality control
RCRA	Resource Conservation and Recovery Act
ROD	record of decision
ROS	regression on order statistics
SADA	Spatial Analysis and Decision Assistance
SAS	Statistical Analysis System
SCL	single control limit
SD	standard deviation
SEDD	staged electronic data deliverable
SWFPR	site-wide false positive rate
TEQ	toxic equivalent
TIN	triangular irregular network
TL	tolerance limit
TPP	technical project planning
TSCA	Toxic Substances Control Act
TW	Tarone-Ware statistic
UCL	upper confidence level or limit

UPL	upper prediction limit
USACE	U.S. Army Corps of Engineers
USEPA	United States Environmental Protection Agency
UST	underground storage tank
UTL	upper tolerance limit
VBA	Visual Basic for Applications
VSP	Visual Sampling Plan

APPENDIX I. GLOSSARY

A

alpha (α)
Decimal level of significance or false positive error rate of a statistical test (Unified Guidance).

analysis of variance (ANOVA)
A statistical method for identifying differences among several population means or medians.

ARCH/GARCH
The autoregressive conditional heteroscedasticity (ARCH) model is used to characterize time series data when variance may occur. The generalized ARCH (GARCH) model is used when the error variance is evaluated with an autoregressive moving average model. Heteroscedasticity is the inequality of the variances of error terms in a data set (Engle 2001).

ARIMA
Autoregressive integrated moving average (ARIMA) is a time series model consisting of autoregressive parameters (explaining the time series observation with past values) and moving average parameters (random shocks with an error structure that is usually Gaussian). The integrated portion of the model refers to the order of differencing (subtracting one observation from the previous one) in order to simulate stationarity in nonstationary data.

arithmetic mean
The sum of a list of numbers, divided by the number of values (Stark 2013).

autocorrelation
Correlation of values of a single variable data set over successive time intervals (Unified Guidance). The degree of statistical correlation either (1) between observations when considered as a series collected over time from a fixed sampling point (temporal autocorrelation) or (2) within a collection of sampling points when considered as a function of distance between distinct locations (spatial autocorrelation).

B

background
Natural or baseline groundwater quality at a site that can be characterized by upgradient, historical, or sometimes cross-gradient water quality (Unified Guidance).

bias

Systematic deviation between a measured (observed) or computed value and its true value. Bias is affected by faulty instrument calibration and other measurement errors, systematic errors during data collection, and sampling errors such as incomplete spatial randomization during the design of sampling programs (Unified Guidance).

bimodal distribution

A data distribution that has two peaks or two modes (science-dictionary.org 2013; NIST/SEMATECH 2012).

bootstrap

A computerized method for assigning measures of accuracy to sample estimates. This technique allows estimation of the sample distribution of almost any statistic using only very simple methods. Bootstrap methods are generally superior to ANOVA for small data sets or where sample distributions are nonnormal (USEPA 2010).

box plot

Graphic of selected descriptive statistics at a monitoring point such as mean, median, or upper and lower quartiles (Unified Guidance).

Box-Jenkins ARIMA

Forecasting or extrapolating future values from a stationary time-series data set using an autoregressive integrated moving average (ARIMA) model. The model requires a stationary data set (Box and Jenkins 1970).

C

censored data

Values that are reported as nondetect. Values known only to be below a threshold value such as the method detection limit or analytical reporting limit (Helsel 2005).

Central Limit Theorem

States that given a distribution with a mean, μ, and variance, σ^2, the sampling distribution of the mean approaches a normal distribution with a mean, μ, and a variance σ^2/N as N, the sample size, increases (USEPA 2010).

conceptual site model (CSM)

A living collection of information about a site which considers factors such as environmental and land use plans, site-specific chemical and geologic conditions, and the regulatory environment (ITRC 2007b).

confidence interval

Statistical interval designed to bound the true value of a population parameter such as the mean or an upper percentile (Unified Guidance).

confidence level
Degree of confidence associated with a statistical estimate or test, denoted as (1 – alpha) (Unified Guidance).

control charts
Graphical plots of compliance measurements over time; alternative to prediction limits (Unified Guidance).

correlation
An estimate of the degree to which two sets of variables vary together, with no distinction between dependent and independent variables (USEPA 2013b).

correlograms
A plot of the autocorrelation coefficients versus the time lags. This plot is also known as an autocorrelation plot.

criterion
General term used in this document to identify a groundwater concentration that is relevant to a project; used instead of designations such as Groundwater Protection Standard, clean-up standard, or clean-up level.

critical point (value)
A predetermined decision level for a test of statistical hypotheses (Unified Guidance).

D

Data Quality Assessment (DQA)
The scientific and statistical evaluation of data to determine if data obtained from environmental operations are of the right type, quality, and quantity to support their intended use (USEPA 2002b).

Data Quality Objective (DQO) Process
A systematic planning tool (based on the scientific method) that identifies and defines the type, quality, and quantity of data needed to satisfy a specified use. DQOs are the qualitative and quantitative outputs from the DQO process (USEPA 2002b).

data quality objectives
The qualitative and quantitative statements derived for the DQO process that clarifies the study's technical and quality objectives, defines the appropriate type of data, and specifies tolerable levels of potential decision errors that will be used as the basis for establishing the quality and quantity (USEPA 2002b).

degrees of freedom
The number of ways which members of a data set or data sets can be independently varied (Unified Guidance).

detection limit (DL)

The concentration that is statistically greater than the concentration of a method blank with a high level of confidence (typically, 99%), or the lowest level of a given chemical that can be positively identified when using a particular analytical method.

E

exploratory data analysis (EDA)

An approach for initial data evaluation using graphical methods to open-mindedly explore the underlying structure and model of a dataset to aid in selection of the best statistical methods. Typical techniques are box plots, time series plots, histograms, and scatter plots (Tukey 1977; NIST/SEMATECH 2012; Unified Guidance).

extrapolation errors

Two common errors in statistical inference are sample error and extrapolation error. An example of when extrapolation errors occur is in curve fitting for prediction outside of the data domain. Hypothesis testing does not account for extrapolation error (Forster 2002).

F

false negative

In hypothesis testing, if the alternative hypothesis (H_A) is true but is rejected in favor of the null hypothesis (H_0) which is not true, then a false negative (Type II, β) error has occurred (Unified Guidance).

false positive

In hypothesis testing, if the null hypothesis (H_0) is true but is rejected in favor of the alternate hypothesis (H_A) which is not true, then a false positive (Type I) error has occurred (Unified Guidance).

false positive rate

The frequency at which false positive or Type I error occurs. The false positive rate, or α (alpha), is the significance level of a hypothesis test. If a test is at an $\alpha = 0.01$ level of significance there would be a 1% chance that a Type I error would occur (Unified Guidance).

G

gamma

A gamma distribution or data set. A parametric unimodal distribution model commonly applied to groundwater data where the data set is left skewed and tied to zero. Very similar to Weibull and lognormal distributions; differences are in their tail behavior, and the gamma density has the second longest tail where its coefficient of variation is less than 1 (Unified Guidance; Gilbert 1987; Silva and Lisboa 2007).

geochemical factors
Geologic/chemical parameters such as oxidation/reduction potential, nitrate, and sulfate that may influence the distribution, concentration, or persistence of contaminants in the subsurface.

geometric mean
A summary statistic calculated by multiplying the data values and taking the Nth root, where N is the sample size (science-dictionary.org 2013).

geostatistical analyses
An analysis using a branch of statistics that focuses on the analysis of spatial or spatiotemporal data, such as groundwater data. One example of a geostatistical technique is kriging, which is an interpolation method that is based on a statistical model of spatiotemporal correlation (Gilbert 1987).

geostatistics
A branch of statistics that focuses on the analysis of spatial or spatiotemporal data, such as groundwater data (Gilbert 1987).

groundwater protection standard (GWPS)
Concentration limits set by the regulatory agency as a standard to be attained in groundwater monitoring. These limits may be fixed health- or risk-based limits (for example, MCLs) or a background level (Unified Guidance).

H

heteroscedasticity
The inequality of the variances of error terms in a data set (Engle 2001).

histograms
Graphical representation of frequency with data values grouped into specified numerical ranges (Unified Guidance).

homoscedasticity
The equality of variance among sets of data (Unified Guidance).

hypothesis test
A statistical test that determines whether one of two statements made about potential outcomes of a statistical test is true. The null and alternative hypothesis statements refer to the condition of a population parameter. The null hypothesis is favored, unless the statistical test demonstrates the greater likelihood of the alternative hypothesis (Unified Guidance).

I

interpolation errors
Error associated with interpolation or prediction of values within a data domain.

interquartile range
The middle range of an ordered set of sample values between the 25th and 75th sample percentiles (Unified Guidance).

interwell
Comparisons between two monitoring wells separated spatially (Unified Guidance).

interwell statistical testing
Statistical analyses of data collected from different monitoring wells (Unified Guidance).

intrawell
Comparison of measurements over time at one monitoring well (Unified Guidance).

intrawell statistical testing
Statistical analyses of data collected from one monitoring well over a period of time (Unified Guidance).

J

J flags
Laboratory qualifier or 'flag' indicating that an analyte is tentatively identified in the sample, but cannot be quantified with precision because the value is below the quantitation limit or lowest confirmed laboratory standard.

K

kriging
A weighted moving-average technique to interpolate the data distribution by calculating an area mean at nodes of a grid (Gilbert 1987).

kurtosis
A measure of whether the data are peaked or flat near the mean. High kurtosis would show a distinct peak near the mean and drop off rapidly to heavy tails (NIST/SEMATECH 2012).

L

lag plot
A plot that displays observations for a time series against a later set of observations, or against the difference between the two sets.

linear regression analysis
A parametric statistical method to measure the linear trend of a data set using data point regression residuals that are based on assumptions of normality, homoscedasticity, and independence (Unified Guidance).

lognormal
A dataset that is not normally distributed (symmetric bell-shaped curve) but that can be transformed using a natural logarithm so that the data set can be evaluated using a normal-theory test (Unified Guidance).

lower confidence limit (LCL)
The lower value on a range of values around the statistic (for example, mean) where the population statistic (for example, mean) is expected to be located with a given level of certainty (science-dictionary.org 2013).

M

maximum likelihood estimation (MLE)
A statistical method used to make inferences about parameters of the underlying probability distribution of a given data set (USEPA 2010).

mean
The arithmetic average of a sample set that estimates the middle of a statistical distribution (Unified Guidance).

median
The 50th percentile of an ordered set of samples (Unified Guidance).

midranks
When ranking a set of measurements by magnitude, assigning ranks to tied values by giving to each tie the average of the ranks those ties would have received were their ordering known. Example: for data values {1, 2.5, 4.2, 5, 5, 5, 9, 10, 10}, the set of midranks would be {1, 2, 3, 5, 5, 5, 7, 8.5, 8.5} since the 4th, 5th, and 6th values as ties would each get the average of ranks 4, 5, and 6, and the 8th and 9th values as a second group of ties would receive the average of ranks 8 and 9.

monotonic trend
The long-term movement in an ordered series, which regarded together with the oscillation and random component, generates observed values that are entirely increasing or decreasing. (EPA 2006c)

multivariate statistical analysis
Any statistical technique designed to simultaneously analyze observations consisting of multiple random variables, for example, a series of groundwater samples where measurements are made on each sample for several chemicals. An example of a multivariate method is cluster analysis, which can examine how groups of chemicals tend to form preferential patterns in groundwater samples.

MyTerm

N

nondetects
Laboratory analytical result known only to be below the method detection limit (MDL), or reporting limit (RL); see "censored data" (Unified Guidance).

nonparametric
Statistical test that does not depend on knowledge of the distribution of the sampled population (Unified Guidance).

normal distribution
Symmetric distribution of data (bell-shaped curve), the most common distribution assumption in statistical analysis (Unified Guidance).

null hypothesis
One of two mutually exclusive statements about the population from which a sample is taken, and is the initial and favored statement, H_0, in hypothesis testing (Unified Guidance).

O

outliers
Values unusually discrepant from the rest of a series of observations (Unified Guidance).

P

parametric
A statistical test that depends upon or assumes observations from a particular probability distribution or distributions (Unified Guidance).

pooled
Groundwater samples from more than one sampling point.

power
See "statistical power."

prediction limits
Intervals constructed to contain the next few sample values or statistics within a known probability (Unified Guidance).

probability plots
Graphical presentation of quantiles or z-scores plotted on the y-axis and, for example, concentration measurement in increasing magnitude plotted on the x-axis. A typical exploratory data analysis tool to identify departures from normality, outliers and skewness (Unified Guidance).

p-value
In hypothesis testing, the p-value gives an indication of the strength of the evidence against the null hypothesis, with smaller p-values indicating stronger evidence. If the p-value falls below the significance level of the test, the null hypothesis is rejected.

Q

quantile plot
A graph of the ranked data versus the fraction of data points it exceeds (USEPA 2006c).

R

range
The difference between the largest value and smallest value in a dataset (NIST/SEMATECH 2012).

regression analysis
A statistical tool for evaluating the relationship of one of more independent variables to a single continuous dependent variable (Kleinbaum et al. 2007).

S

sampling point
A specific spatial location from which groundwater is being sampled.

scatter plots
Graphical representation of multiple observations from a single point used to illustrate the relationship between two or more variables. An example would be concentrations of one

chemical on the x-axis and a second chemical on the y-axis. They are a typical exploratory data analysis tool to identify linear versus nonlinear relationships between variables (Unified Guidance).

seasonal autoregressive integrated moving average (SARIMA)
A model in which seasonal autoregression and moving average terms predict future values from, or extrapolate, time series data. The SARIMA model incorporates seasonal and non-seasonal factors. Statistical packages such as Minitab can be used (PSU 2012).

site-wide false positive rate (SWFPR)
The design probability of at least one statistically significant finding among a network of statistical test comparisons at a group of uncontaminated wells (Unified Guidance).

skewness
A measure of asymmetry of a dataset (Unified Guidance).

spatial variability
Spatial variability exists when the distribution or pattern of concentration measurements changes from well location to well location (most typically in the form of differing mean concentrations). Such variation may be natural or synthetic, depending on whether it is caused by natural or artificial factors (Unified Guidance).

standard deviation (SD)
An estimate of the degree of variability within a distribution indicating the degree to which the values vary from the average value or mean (Unified Guidance).

stationarity
Stationarity exists when the population being sampled has a constant mean and variance across time and space (Unified Guidance).

stationary
A distribution whose population characteristics do not change over time or space (Unified Guidance).

statistical bias
Quantitative term describing the difference between the average measurements made on the same object and its true value; see "bias" (NIST/SEMATECH 2012).

statistical confidence
Likelihood that a range of values will contain the population parameter of interest (NIST/SEMATECH 2012).

statistical inference
Conclusions drawn from observed data without seeing all of the possible data (Unified Guidance).

statistical power
Strength of a test to identify an actual release of contaminated groundwater or difference from a criterion (Unified Guidance).

statistical significance
Statistical difference exceeding a test limit large enough to account for data variability and chance (Unified Guidance). A fixed number equal to alpha (α), the false positive rate, indicating the probability of mistakenly rejecting the stated null hypothesis (H_0) in favor of the alternative hypothesis (H_A). Or, the p-value sufficiently low such that the analyst will reject the null hypothesis (H_0).

T

temporal autocorrelation
The correlation between observations on a single variable over successive intervals of time. This relationship is also called "serial correlation". Autocorrelation in temporal data is significant for time-series analysis (Unified Guidance; Burt et al. 2009).

test power
The probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true (Stark 2013).

time series plot
A graphic of data collected at regular time intervals, where measured values are indicated on one axis and time indicated on the other. This method is a typical exploratory data analysis technique to evaluate temporal, directional, or stationarity aspects of data (Unified Guidance).

tolerance limits
The upper or lower limit of a tolerance interval (Unified Guidance).

t-test
A t-test, or two-sample test, is a statistical comparison between two sets of data to determine if they are statistically different at a specified level of significance (Unified Guidance).

type I error rate
The frequency at which a false positive or Type I error occurs. The false positive rate, or α, is the significance level of a hypothesis test. If a test is at an $\alpha = 0.01$ level of significance there would be a 1% chance that a Type I error would occur (Unified Guidance).

U

upper confidence limit (UCL)

The upper value on a range of values around the statistic (for example, mean) where the population statistic (for example, mean) is expected to be located with a given level of certainty, such as 95% (science-dictionary.org 2013).

V

variance

The square of the standard deviation (EPA 1989); a measure of how far numbers are separated in a data set. A small variance indicates that numbers in the dataset are clustered close to the mean.

variogram

A plot of the variance (one-half the mean squared difference) of paired sample measurements as a function of the distance (and optionally the direction) between samples. Typically, all possible sample pairs are examined, distance and directions. Variograms provide a means of quantifying the commonly observed relationship that samples close together will tend to have more similar values than samples far apart (EPA 1989). A graphical tool used in geostatistical analysis.

W

Weibull distribution

A parametric unimodal distribution model commonly applied to groundwater data where the data set is right- or left-skewed. Also used for failure analysis. Very similar to gamma and lognormal distributions; differences are in their tail behavior, and the Weibull density has the smallest tail (Unified Guidance; Gilbert 1987; Silva and Lisboa 2007; Abernethy 2010).

Z

z score

The observed value of the z statistic (Stark 2013).

z statistic

A test statistic whose distribution under the null hypothesis has an expected value of zero and can be approximated by the normal distribution (Stark 2013).