Statistical Guidance for Determining Background Ground Water Quality and Degradation

State of Idaho
Department of Environmental Quality

March 2014

Printed on recycled paper, DEQ, March 2014, PID 9010, CA 82017. Costs associated with this publication are available from the State of Idaho Department of Environmental Quality in accordance with Section 60-202, Idaho Code.

Statistical Guidance for Determining Background Ground Water Quality and Degradation

March 2014

Prepared by
Idaho Department of Environmental Quality
Water Quality Division
1410 N. Hilton
Boise, ID 83706

Acknowledgments

The May 2009 guidance was prepared by Dr. Xin Dai, Idaho Department of Environmental Quality. Technical review of the May 2009 document and updates for the March 2014 guidance were completed by Dr. John Welhan, Idaho Geological Survey. Revisions for the March 2014 guidance were provided and reviewed by Don Carpenter, Brady Johnson, Scott Miller, and Edward Hagan, Idaho Department of Environmental Quality. Technical Editing was completed by Jill White, Idaho Department of Environmental Quality.

Table of Contents

1 Introduction 1
2 Authorities and Definitions 3
3 Statistical Characterization of Ground Water Quality 4
3.1 Elements of an Analysis 4
Analysis Tools and Documentation 5
3.2 Evaluation of Hydrogeologic Data 7
3.3 Defining Constituents of Concern 7
3.4 Adequate Sample Size 8
Limited Annual Sampling 9
3.5 Data Below Detection Limits 9
3.6 Evaluation of Background Ground Water Quality Data 10
4 Statistical Determination of Water Quality Degradation 12
4.1 Alternative Concentration Limit 13
4.2 New Versus Existing Activity 13
4.3 Interwell Versus Intrawell Analysis 13
4.4 Decision Thresholds and Confidence Levels 14
4.4.1 Interwell and Intrawell Tolerance Limits 15
4.4.2 Interwell Prediction Limits 15
4.4.3 Interwell Simultaneous Limits 15
4.5 Verification Resampling 16
4.6 Trending Data 16
5 Summary of Process 17
5.1 Determination of Background Ground Water Quality 17
5.2 Determination of Degradation 17
5.2.1 Intrawell Comparisons 17
5.2.2 Interwell Comparisons 18
5.2.3 Treatment of Verification Versus Confirmation Resampling Data 18
5.2.4 Interim Methodology for Trending Data 18
6 References 19
Appendix A. Alternative Concentration Limits 21
Appendix B. Exploratory Data Analysis/Descriptive Statistics 23
Appendix C. Data Independence 27
Appendix D. Determination of Normality and Choice of Distribution 33
Appendix E. Seasonal Trends 43
Appendix F. Secular Trends 49
Appendix G. Data Pooling 53

Appendix H. Parametric Upper Tolerance Limits 57
Appendix I. Nonparametric Upper Tolerance Limits 61
Appendix J. Parametric Upper Prediction Limits 63
Appendix K. Nonparametric Upper Prediction Limits 67
Appendix L. Interim Decision Thresholds in the Presence of a Secular Trend 73
Appendix M. Example Scenario for an Existing Wastewater Reuse Facility With No Chemical Impact 81
Appendix N. Applying Intrawell Analysis at Existing Facilities When Interwell Methods are Inadvisable 83
Appendix O. Statistical Concepts 91
Appendix P. Summary of Revisions 103

List of Tables

Table 1. Considerations for wastewater reuse sites. 13
Table B1. Data and resulting descriptive statistics for example scenario. 25
Table D1. Example of Shapiro-Wilk test for normality on TDS data from well B1. 36
Table D2. Partial list of coefficients a_i for the Shapiro-Wilk test of normality. 36
Table D3. Lower 1% and 5% critical values for Shapiro-Wilk test statistic W. 37
Table D4. Goodness-of-fit summary statistics for the Q-Q plots in Figure D3. 39
Table D5. Goodness-of-fit summary statistics for the Q-Q plots shown in Figure D5. 41
Table E1. A portion of the quantiles of the chi-square distribution with k-1 degrees of freedom. 44
Table E2. Testing for seasonality using the Kruskal-Wallis test. 46
Table F1. Mann-Kendall test set up. 50
Table H1. Partial table of factors (K) for constructing one-sided normal upper tolerance limits at 95% confidence and 95% coverage. 58
Table I1. Sample sizes for nonparametric upper tolerance limits.[a] 61
Table J1. K Factors at $\alpha = 0.05$ for a verification protocol where one or both verification resamples must confirm the initial exceedance. 64
Table K1. Confidence levels for a nonparametric prediction limit where exceedance is verified when one or both verification resamples also exceed the limit. 67
Table K2. ProUCL output showing calculated decision limits for a normal distribution. 69
Table K3. Decision thresholds calculated from the same background data for different assumed population distributions. 70
Table K4. Comparison of decision thresholds calculated from log-transformed background data that should have been modeled with a gamma distribution. 71
Table L1. Fabricated well B1 data. 74
Table L2. Two additional years of fabricated well B1 data. 78
Table N1. Background total dissolved solids measurements. 85
Table N2. Calculated test statistics for current year's monitoring data. 86
Table N3. Background total dissolved solids measurement with fabricated outliers. 89
Table O1. Summary of statistical notation used. 91

List of Figures

Figure 1. Implementation of Idaho's "Ground Water Quality Rule" IDAPA 58.01.11.400. 2
Figure 2. Process for determining background ground water quality. 6
Figure 3. Recommended process for handling censored data. 10
Figure 4. Process for evaluating background ground water quality and data adequacy. 11
Figure B1. Box plots created by ProUCL 5.0 24
Figure B2. Time-series graphs of concentration data in Table B1 created with ProUCL. 25
Figure C1. Example ground water TDS measurements used for evaluating the statistical independence of a time-series data set. 29
Figure C2. Box-Jenkins autocovariance plot created using the quarterly-averaged data in Figure C1. 29
Figure C3. An example of a semivariogram (Isaaks and Srivastava 1989) computed for all of the TDS data in Figure C1 29
Figure C4. A hypothetical example of a semivariogram based on a sufficient number of monitoring points to construct a well-defined semivariogram and identify the minimum interwell spacing necessary to maintain spatial data independence. 30
Figure C5. Semivariogram based on insufficient data to define spatial independence. 30
Figure D1. Decision tree for determining a population distribution. 34
Figure D2. Histogram of pooled TDS data from Table B1, superimposed on a normal distribution for visual comparison purposes. The mean (black line) and median (orange line) are shown for reference. 38
Figure D3. Goodness-of-fit results, in the form of Q-Q plots of the data represented in Figure D2 relative to (A) a normal distribution, (B) a lognormal distribution, and (C) a gamma distribution. 38
Figure D4. Data used to illustrate the decision process in Figure D1 when the sample is highly right-skewed. 40
Figure D5. Graphical summaries of goodness-of-fit to (A) a normal distribution, (B) a lognormal distribution, and (C) a gamma distribution. 41
Figure F1. Results of applying a Mann-Kendall test to the raw TDS data from well B1 in Table B1. 51
Figure F2. Results of applying the Mann-Kendall test to well B1's deseasonalized TDS data (Appendix E, Table E3). 52
Figure K1. Histogram of seasonally adjusted and pooled TDS data for wells B1 and B2. 68
Figure L1. Concentration versus time plot for fabricated well B1 data. 75
Figure L2. An example of applying a biannual slope recalculation procedure to identify changes in slope as future data is collected. 78
Figure L3. Recommended procedure for estimating interim decision thresholds in the presence of a trend (based on an approach outlined by Gibbons (1994) and the control chart methodology described in Appendix N). 80
Figure N1. Historical (background) total dissolved solids concentrations versus time. 86
Figure N2. Comparison of latest monitoring results to historical data and specified control limits. 86
Figure O1. Example histogram. 93
Figure O2. Example box plot. 94
Figure O3. Example time-series plot. 94

Figure O4. Example scatter plot. 95
Figure O5. Mode, median, mean for various distributions 96
Figure O6. Example of three distributions with various degrees of kurtosis (peakedness)..... 97

1 Introduction

This guidance describes a process for the Idaho Department of Environmental Quality (DEQ) to use when determining if ground water quality is degraded. The term *degradation* is defined in the Idaho "Ground Water Quality Rule" (IDAPA 58.01.11) as "the lowering of ground water quality as measured in a statistically significant and reproducible manner." This guidance provides a process and statistical tools, which can be used to determine statistically significant degradation. Other processes and statistical tools may be used; this guidance describes process and tools that are understood to achieve the intended goals.

The two principal goals of the guidance are as follows:

1. Describe a statistically based process for establishing background ground water quality.
2. Identify methods and criteria for identifying when ground water quality degradation is statistically significant.

An understanding of these two concepts is fundamental to addressing ground water quality issues. Knowledge of the background ground water quality is necessary before ground water quality degradation can be identified. Once background ground water quality is established, ground water quality degradation, if any, can be determined.

To achieve the two principal goals, the guidance is structured around the following four objectives:

1. Provide a standardized framework or process to objectively evaluate ground water quality data.
2. Provide flexibility to address site-specific conditions.
3. Provide a decision tree showing elements required to complete suggested parts of the process.
4. Suggest certain statistical tools but allow for alternatives.

Some suggested tools are presented in Appendices A to N (see specific topics in Table of Contents). Appendix O provides definitions of terminology used throughout the document. A summary of the revisions contained in this version of the guidance are identified in Appendix P.

The determination of what constitutes degraded ground water is essential for implementing DEQ programs that rely on the IDAPA 58.01.11 to protect the health of Idahoans and the environment. DEQ and the regulated community can use the methods contained in this document to estimate background ground water quality conditions and identify any degradation. This guidance document is intended to help interpret and apply IDAPA 58.01.11 at sites not addressed with existing state and federal program guidance. It may also complement existing guidance by addressing situations not covered with other guidance. This document does not impose legally binding requirements on DEQ or the regulated community. The document identifies an approach for defining ground water degradation, but DEQ retains the discretion to allow different approaches on a case-by-case basis that differ from this information. Interested parties are free to raise questions about the appropriateness of the application of the information in this document to a particular situation, and DEQ will consider whether or not the technical approaches are appropriate in that situation.

Once a constituent is detected in ground water, IDAPA 58.01.11 provides a process for DEQ to follow (Figure 1). Figure 1 illustrates the conceptual approach to implementing IDAPA 58.01.11.400.

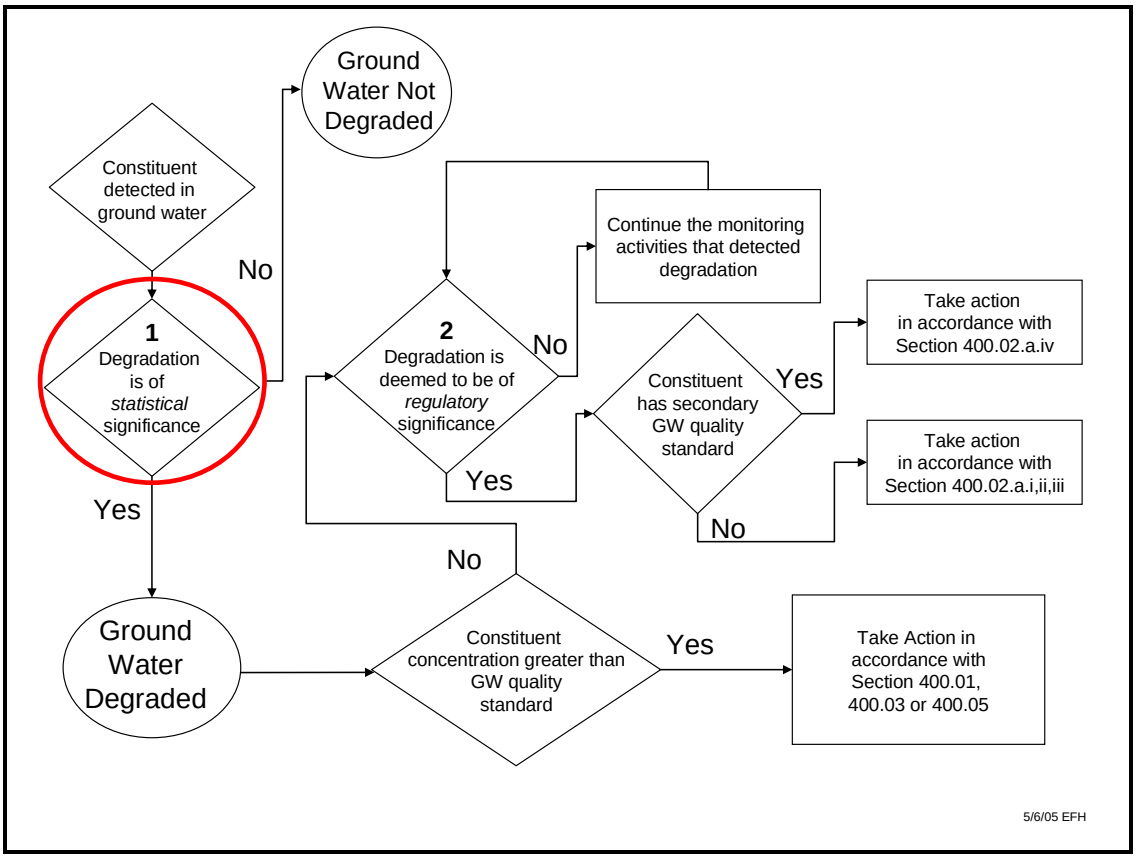


Figure 1. Implementation of Idaho's "Ground Water Quality Rule" IDAPA 58.01.11.400.

The first step, following a detection, is to determine if the constituent concentration is greater than background. (This step is shown in diamond 1 and circled in red). This guidance document addresses this step by providing a process that can be used to determine if a detection represents statistically significant degradation. **For ground water to be degraded, the concentration must be above background and the degradation must be of *statistical* significance.** If ground water is not degraded, then no further action is necessary. If ground water is degraded, but the constituent is present in a concentration below a ground water quality standard, DEQ must make a determination of whether the degradation is of regulatory significance; this step is shown in diamond 2.

This statistical guidance document does not address whether degradation is of *regulatory* significance. The criteria for determining regulatory significance are contained in IDAPA 58.01.11.400.02.b. and include the following:

- Site-specific hydrogeologic conditions
- Water quality, including seasonal variations
- Existing and projected future beneficial uses
- Related public health issues

- Whether degradation involves a primary or secondary constituent in IDAPA 58.01.11.200.

Additional guidance, using the criteria listed above, is being developed by DEQ to determine whether degradation that is of statistical significance is also of *regulatory* significance.

The following example illustrates the purpose of this guidance.

> Using the approach described in this guidance, background ground water nitrate concentration at a site is determined statistically to be 5.0 milligrams per liter (mg/L). Downgradient of the site, the nitrate concentration in ground water is found to increase from 5.0 mg/L to 5.5 mg/L. Verification sampling confirms the downgradient nitrate concentration exceeds the background level. These results indicate the degradation is of statistical significance and the ground water is degraded for nitrate (diamond 1 in Figure 1). Guidance regarding actions to address the degradation is beyond the scope of this guidance.
>
> Because the ground water is degraded, but the concentration is below a ground water quality standard (in this case, 10.0 mg/L), a determination of whether the degradation is of regulatory significance must be made (diamond 2 in Figure 1).

This statistical guidance does not provide a method for quantifying the magnitude of degradation. Statistics can be used to define the statistical uncertainty between sample values collected in different wells. But statistics cannot be expected to define the magnitude of the difference. Statistics can never prove that a difference between sample values is real, only the probability that one may exist, given the available data. Whether degradation is of regulatory significance is dealt with in another guidance document currently being developed by DEQ.

2 Authorities and Definitions

The legislation and rules addressing ground water quality issues in Idaho include the Idaho Ground Water Quality Protection Act of 1989 (Act) (Idaho Code §39-120 to §39-127) and the Idaho "Ground Water Quality Rule" (IDAPA 58.01.11). The Act created the Ground Water Quality Council and directed the council to develop the Ground Water Quality Plan (plan). The plan provides the overall direction and policies of the state with respect to ground water quality concerns. IDAPA 58.01.11 implements a portion of the plan.

Background ground water quality can be established using samples collected from monitoring wells that sample the ambient ground water quality in the same aquifer that is likely to be impacted by development. IDAPA 58.01.11 identifies two types of background ground water quality: natural and site.

Natural background level is defined by IDAPA 58.01.11 "as the level of any constituent in the ground water within a specified area as determined by representative measurements of the ground water quality unaffected by human activities." In areas where the natural background level of a constituent exceeds the standard, the natural background level shall be used as the standard.

Site background level is defined as the ground water quality at the hydraulically upgradient site boundary. In areas where the ground water quality is unaffected by human activities, the site background level is equivalent to natural background.

3 Statistical Characterization of Ground Water Quality

Before any data evaluation begins, it is useful to have a clear understanding of the issues that need to be addressed, including the constituents of concern (COCs). Once the main issues are defined, the data can be collected and then reviewed within the appropriate context. Existing data must be compiled and evaluated to determine if the information is sufficient to adequately characterize the ground water quality. In most cases, the goal of the statistical analysis will be to characterize background ground water quality in a manner such that decisions regarding ground water quality degradation are defensible.

The guidance provides flexibility by allowing options to determine background ground water quality, depending on the adequacy of the data for statistical analysis. If sufficient data are available to statistically characterize background ground water quality, then appropriate statistical methods may be employed. The determination of data adequacy is a site-specific decision that depends on many physical factors as well as the objectives of the project. Suggested methods to determine if the available ground water quality data are adequate to conduct valid statistical analyses are described in the appendices.

If data are not adequate to conduct valid statistical analyses, then a sampling plan to collect adequate data may be developed or background ground water quality may be estimated using an alternative concentration limit (ACL) in accordance with a DEQ-defined method. The ACL is designed to be protective of ground water quality by using the lowest value provided from three options as described in Appendix A. If a sampling plan is implemented, the ACL will be used for decision-making purposes until adequate data are collected to support valid statistical analyses. However, an ACL also may be selected even when appropriate data are available for valid statistical analyses if the interested party does not want to conduct statistical analyses and DEQ concurs with the decision.

3.1 Elements of an Analysis

The elements for characterizing background ground water quality are site-specific and dependent on the complexity of the area. If the process described in this guidance is used, the steps to be completed for each site include the following:

- State the objectives of the analysis.
- Delineate the study area and hydrogeologic features relevant to monitoring.
- Identify COCs and provide rationale for considering them.
- Evaluate and define data adequacy in the context of the analysis objectives.
- Identify appropriate statistical tools to address the issues.
- If the data are inadequate for the analysis, determine an appropriate temporal scale for the data collection program and provide a rationale for why it is appropriate.
- If selected data are used in (or excluded from) the analysis, provide a rationale.

The elements must be addressed within the context of the hydrogeologic framework. Individual aquifers must be defined at the appropriate scale. For each aquifer, the ground water flow direction and ground water gradient should be described and uncertainties in both should be estimated. Data on the ground water chemistry of each aquifer should be compiled and ground water quality trends should be identified, if data are sufficient. The sampling locations and sampling frequency should be evaluated to ascertain whether results can be used to represent the ground water quality within the area of concern.

The general process for defining background ground water quality is illustrated in the flow diagram in Figure 2.

Analysis Tools and Documentation

One of the principal goals of this guidance document is to help users navigate the array of statistical procedures, choices, and theoretical options that are available during the analysis of ground water quality data. By making the process more transparent, DEQ hopes to minimize analysis problems and decision errors and thereby streamline the review of users' data analysis results. One of the keys to success in this endeavor is that users submit sufficient relevant documentation with their analysis results to allow DEQ's technical review to proceed as smoothly as possible.

Data Submission

All water quality data considered in the statistical analysis must be provided to DEQ in digital spreadsheet form together with any relevant documentation necessary to understand the data structure, data fields, and analytical details (e.g., concentration units and detection limits).

Analysis Methodology and Tools

To facilitate the evaluation of the user's statistical analysis, clear documentation is required of the type(s) of software used in the analysis, including version number(s) and relevant information on the software source and publisher. The use of nonstandard methodologies should be avoided to minimize interpretational problems or inappropriate conclusions. All software should be well documented and widely accepted in the ground water profession as to its utility in the kind of statistical analyses covered in this guidance document.

The United States Environmental Protection Agency's (EPA's) ProUCL v.5.0 (EPA 2006; 2013a) statistical software is a good example of acceptable software due to its ease of use, excellent documentation, wide acceptance, and free availability. The software is available for download at *http://www.epa.gov/osp/hstl/tsc/software.htm*; it is easy to install and includes many of the analysis tools described in this guidance document. Examples of its application are provided in many of the appendices herein.

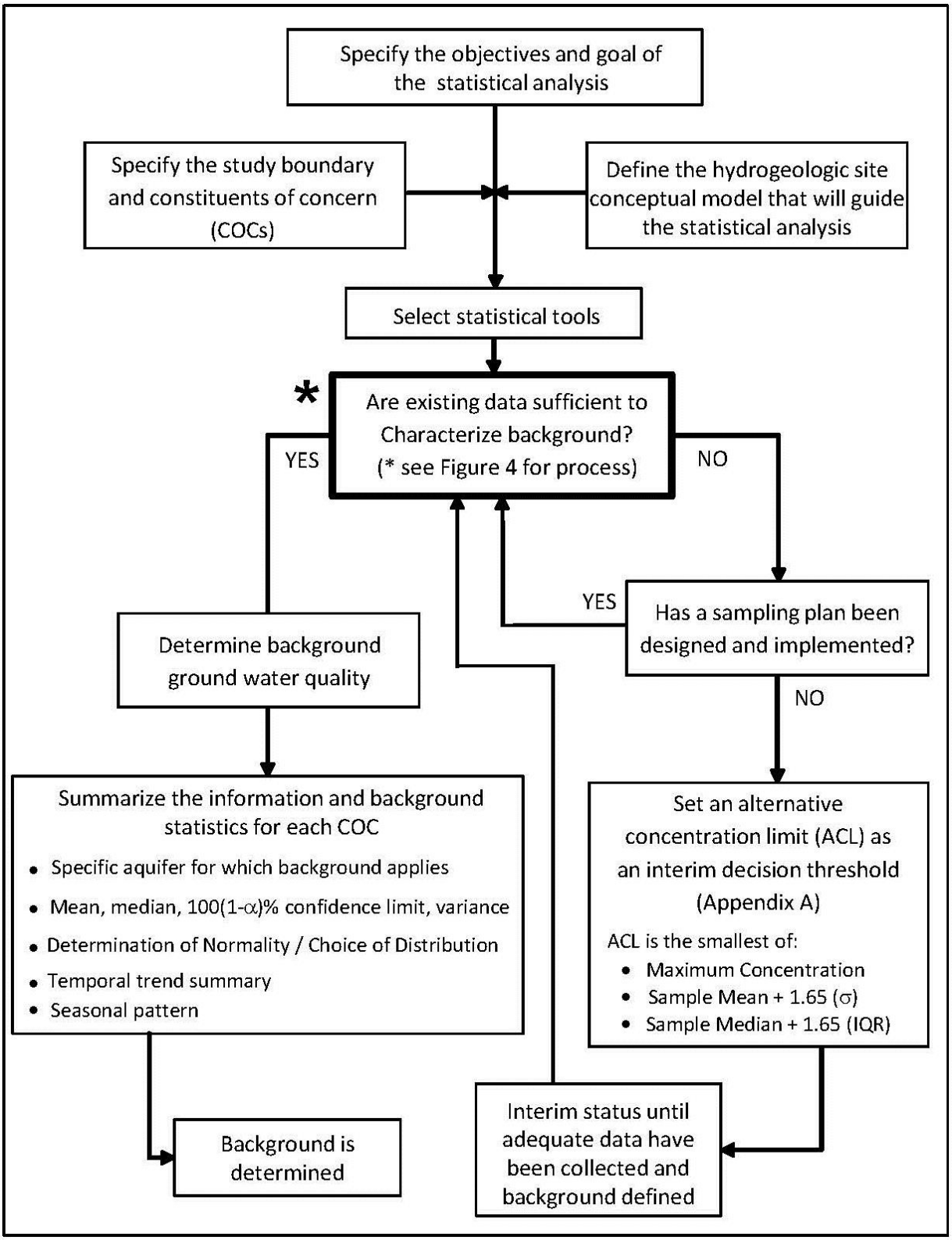


Figure 2. Process for determining background ground water quality.

3.2 Evaluation of Hydrogeologic Data

When defining background water quality or making comparisons against a compliance threshold, one of the first and most critical requirements is a clear and hydrogeologically defensible conceptual model of the site's subsurface architecture. Depending on the site, different statistical populations of ground water quality may occur at different aquifer depths and in different aquifer media. An adequate amount of water quality data is required for each subpopulation so it is statistically representative of the strata, sample depths, or other characteristics that may affect water quality differently. In such cases, background water quality may have to be defined separately for several subpopulations, and future comparisons with background may then have to be conducted with the same consideration in mind so that any statistical conclusions are hydrogeologically defensible in the context of the site conceptual model.

Ambient ground water quality typically varies spatially (between wells) and temporally (over time) due to natural conditions; anthropogenic impacts can contribute additional variability to water quality. A minimum of one upgradient well is needed to determine background ground water quality, but additional upgradient wells may be necessary to characterize site variability in complex hydrogeological situations (Fisher and Potter 1989; Cressie 1993; Gibbons 1994). At so-called *green fields* sites where human activities have not affected site water quality, both upgradient and downgradient wells can be used initially to determine background ground water quality and spatial variability.

The hydrogeologic characteristics of a site determine the number of ground water monitoring wells required and their locations. The depth to water, flow direction, net recharge rate, aquifer and soil characteristics, topography, thickness and lithology of the vadose zone, and hydraulic conductivity of the aquifer are all important in determining the vulnerability of an aquifer and the necessary spacing and depth of monitoring wells (Ogden 1987). The geology of a site should be characterized with data from well logs, geologic maps, and cross sections. Structural features, such as faults, fractures, fissures, impermeable boundaries or other features that can influence flow direction should be delineated. Additional hydrogeologic information relevant to assessing the adequacy of monitoring data should be summarized, including but not limited to ground water flow velocity, transmissivity, storage coefficient, porosity, and dispersivity.

3.3 Defining Constituents of Concern

A COC is a chemical that is disturbed, generated, used, or disposed at the site in sufficient quantity to pose a risk to beneficial uses of ground water or interconnected surface water. This includes degradation products or chemicals released during chemical reactions in the environment. COCs defined for each site will depend on site operations. When deciding what chemicals may be COCs, the following should be carefully considered:

- Industrial/commercial processes resulting in the generation of the chemical(s) that are permitted to be handled, stored, or reused (land-applied) on the site
- Physical and chemical properties of the chemical(s)
- Complexity and sensitivity of the hydrogeologic environment

Once a COC is determined the following should be considered:

- Methods of sample collection, handling, and transportation that are appropriate for the COC
- Laboratory analysis procedures used to measure chemical concentration that are appropriate for the COC

3.4 Adequate Sample Size

This section specifically addresses quantifiable measurements above the detection limit not affected by censoring. Procedures for dealing with censored data are discussed in section 3.5. The quality and quantity of available monitoring data are two of the most important factors in determining background ground water quality for a COC. Individual ground water samples are only representative of ground water quality at a particular time in a particular location. Ground water quality often varies seasonally or changes with time and/or location, so a single ground water sample may not be representative of ground water conditions throughout the site or over a period of time. The greater the number of independent samples collected over time, the more representative the characterization of the ground water quality. Larger sample populations also increase the statistical confidence in the evaluation of ground water quality. Valid statistical testing depends upon collection of adequate data. Statistical tests rely on using estimates of the true mean and true variance of a population. For example, the estimate of the true mean is the average of the data points collected. The estimate of the true standard deviation is the standard deviation of the data points collected.

The number of samples needed to conduct a statistical analysis meeting the objectives and goals of a project depends on the site-specific conditions, which in turn controls the data variability. Site-specific conditions may include physical factors such as land use, hydrogeologic environment, and seasonality and social considerations. The EPA's Unified Guidance document (EPA 2009) recommends that a minimum of 8 to 10 independent samples be available to estimate the standard deviation of a parametrically distributed statistical population (e.g., normal, gamma or lognormal distributions). **DEQ recommends collecting 12 independent samples for most statistical analysis methods discussed in this guidance document.** In stark contrast, a tolerance interval estimate for a nonparametric distribution requires a minimum of 59 independent data points to achieve 95% coverage[*] at 95% confidence (Conover 1999; EPA 2009; Gibbons 1994).

In other situations, such as the presence of a seasonal trend, the Seasonal Kendall Test requires a minimum of 3 years of monthly data, or 36 data points (Gilbert 1987). When quarterly data are sparse, the Kruskal-Wallis test can be used as long as there are at least 3 years of quarterly data collected in the same months (a minimum of 12 independent data points). To quantify serial correlation effects (temporal dependence), Harris et al. (1987) state that at least 10 years of quarterly data, or 40 data points, may be necessary.

As illustrated in the previous paragraphs, adequate sample size varies on a case-by-case basis and is a site-specific decision that must consider factors unique to each project and site. The goal of determining sample size in a statistical study is to find the number of samples that provides

[*] where 95% of future samples will fall within the interval

adequate yet practically feasible evidence with which meaningful conclusions can be made relative to the goals of the study. The final determination of what constitutes adequate sample size will be made by DEQ in cooperation with the regulated entity.

Limited Annual Sampling

In some cases, quarterly or more frequent sample data may not be available throughout the year. For example, at sites with limited or no access during winter and early spring, regular year-round sampling may not be physically possible. In such situations, DEQ may modify the quarterly sampling requirement to provide for adequate data coverage during the portion of the year when sampling is possible as well as some of the statistical analysis procedures pertaining to quarterly samples discussed in this document.

In the event that year-round sampling is not possible, then *quarterly* samples in this document shall be interpreted to comprise four or more evenly spaced sampling events during the annual sampling window. For example, if sampling is possible only during the 8-month period from April through November, then four samples should be collected, one every other month, at each well.

This strategy maintains sufficient temporal data coverage and minimum sample sizes for determination of background and seasonality. In most cases, depending on data independence considerations, it would allow for timely verification resampling of wells discovered to be out of compliance during approximately the first 5 months of the annual sampling window. Wells deemed to be out of compliance after September may not be testable again until the following spring. In such cases, particular care needs to be taken to ensure that appropriate seasonality adjustments can be applied so that comparisons between fall and springs samples are statistically defensible.

3.5 Data Below Detection Limits

Data sets that contain nondetect values make it more difficult to determine the type of statistical distribution that characterizes the population from which samples are drawn. These data sets are referred to as censored data throughout the remainder of this guidance. For most nonparametric methods, the presence of censored data is not an issue, but their effect in parametric analysis is very dependent on the statistical form of the data distribution (EPA 2009). The procedure to evaluate censored data is outlined in Figure 3 and conforms to recommendations in EPA's Unified Guidance (EPA 2009), as well as Gibbons (1994) and Helsel (1990; 2005).

The first step when evaluating censored data is to distinguish between detection-only applications (such as identifying the first arrival of a constituent) and ground water quality characterization (such as defining background). If the data are used to determine whether a constituent is present, then the results should be handled on a case-by-case basis independent of the process outlined in Figure 3. If the censored data will be used to estimate summary statistics, then the procedure outlined in Figure 3 is applicable.

A number of considerations, including sample size, percentage of censored data and the form of the data distribution determine the preferred method for handling censored data (EPA 2009). The use of substitutional methods (e.g., substituting half the method detection level for censored

values) is fraught with theoretical considerations and is to be avoided; in situations not covered in Figure 3, DEQ recommends following the recommendations of the Unified Guidance in this regard (EPA 2009, chapters 10 and 15) as well as those of Helsel (2005) and consulting with DEQ on a case-by-case basis.

In general, imputation of censored values should be avoided in small (<15–20) data sets and is unnecessary in very large data sets (>500). If censored measurements comprise less than 50% of the measurements of an analyte and the data set appears to be parametrically distributed (either normal, lognormal, or gamma), then the statistical parameters of the distribution are best inferred using distributional methods such as the maximum likelihood estimator (e.g., Helsel 1990; 2005; the utilities available in ProUCL 5.0 [EPA 2013a]) are recommended for such situations. If censored measurements comprise more than 50% of the data set, nonparametric analysis is generally preferred unless special circumstances apply (EPA 2009); in that case multiple methods for estimating the distribution's parameters should be evaluated, including a sensitivity analysis of the results, before deciding on the best outcome. In special cases, such as where the nondetect percentage is very high, DEQ may approve alternative methods for handling censored data on a case-by-case basis.

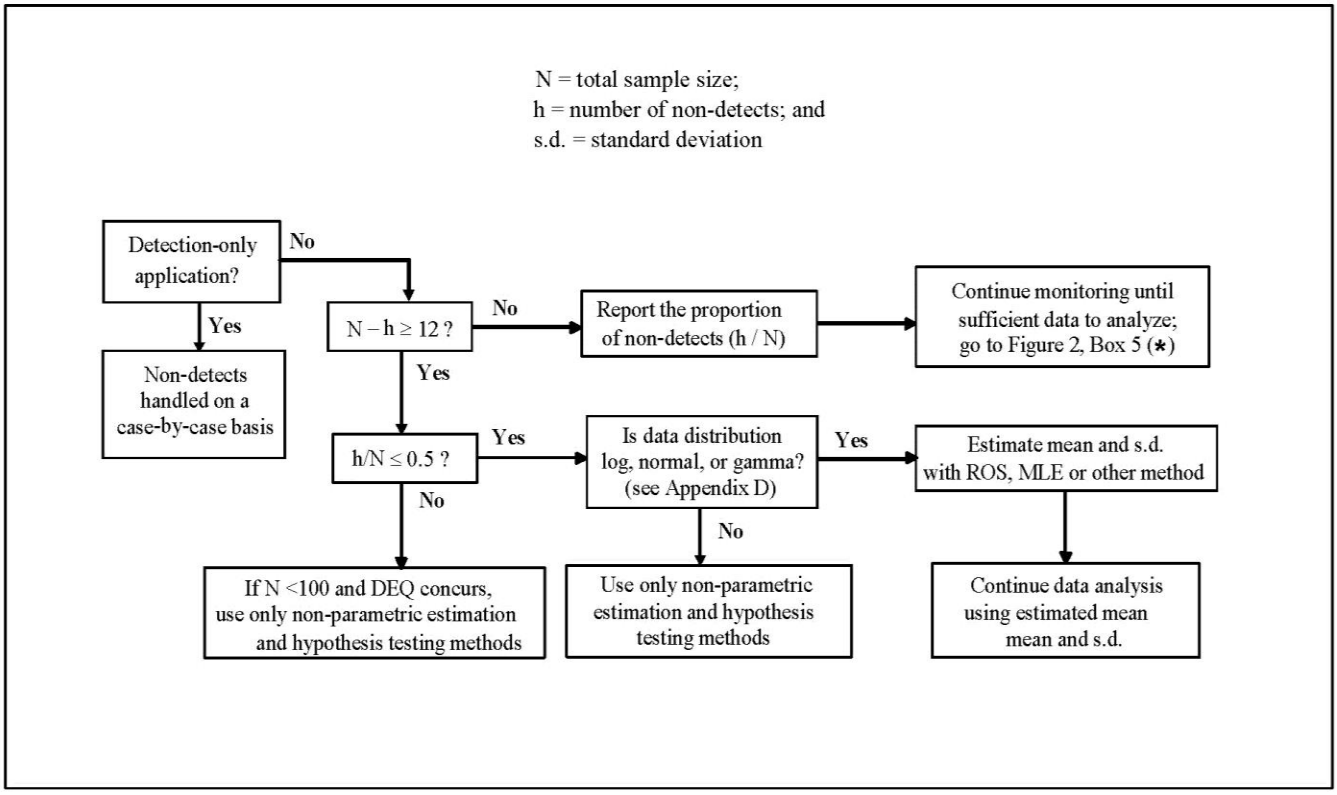


Figure 3. Recommended process for handling censored data.

3.6 Evaluation of Background Ground Water Quality Data

The procedure for evaluating data to determine its suitability for statistical analysis, along with the information and analysis required to substantiate a statistical characterization of background ground water quality, are outlined in Figure 4. The steps include data compilation; exploratory

analysis and descriptive statistics; evaluation of data independence; analysis of frequency distribution and parametric behavior; seasonal and secular trend analysis; justification for data pooling if used; and an assessment of the adequacy of the sample size of the available data to support a statistical characterization of background ground water quality (section 5 provides more details on these terms). It is necessary to accurately characterize background ground water quality based on a sufficient number of samples to determine average concentrations and variability at the site. Most importantly, the correct form of the population distribution must be determined so that subsequent hypothesis tests on the data will be as accurate and statistically powerful as possible. This is because all statistical testing assumes that the sample data are representative of a mostly unobservable population of the entire range of aquifer water quality. In practice, the most commonly inferred population distributions are the normal, lognormal, and gamma. If the sample data do not conform to one of these parametric distributions, then the underlying population distribution is assumed to be nonparametric.

Background ground water quality should be analyzed using the most current data available and, if the available data are deemed adequate to justify such an analysis, then the results of the statistical characterization should be summarized as part of the documentation submitted to DEQ for technical review.

Appendices A through G provide suggested methods for the analyses indicated in Figure 4.

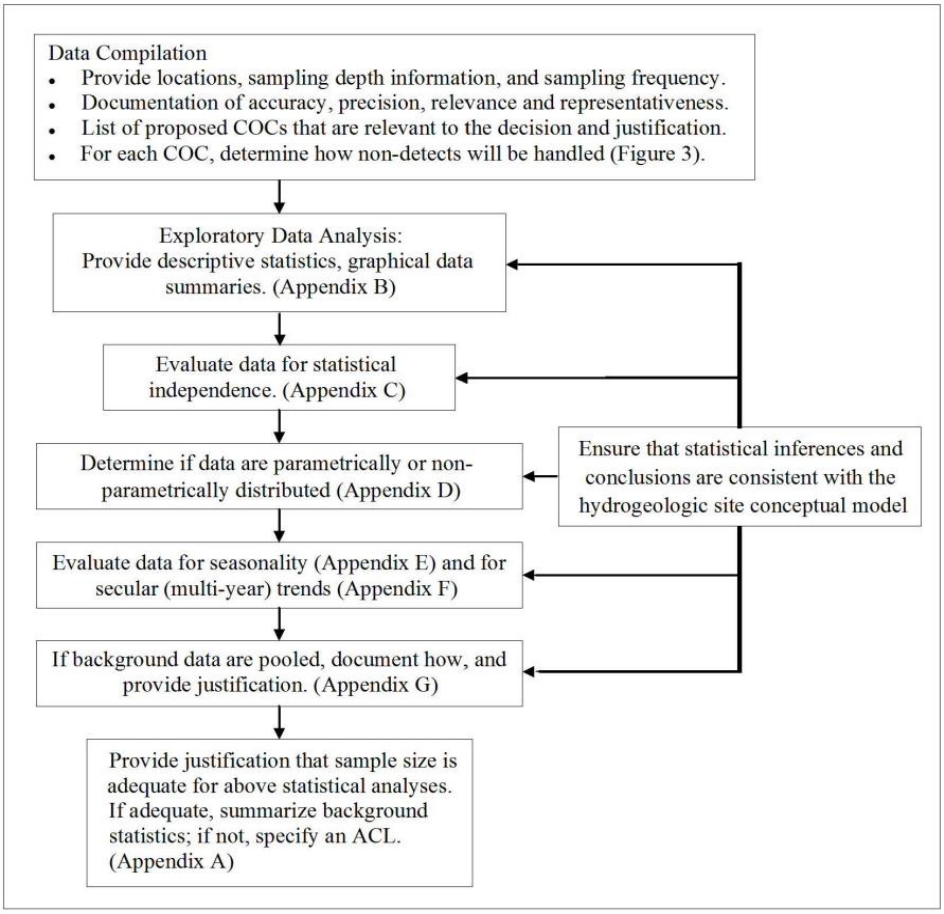


Figure 4. Process for evaluating background ground water quality and data adequacy.

4 Statistical Determination of Water Quality Degradation

The term *degradation* is defined by IDAPA 58.01.11 as "the lowering of ground water quality as measured in a statistically significant and reproducible manner." To be statistically significant and reproducible suggests that multiple measurements over time are required to determine whether degradation has occurred. The number of measurements and the length of time will likely be site-specific and dependent on the complexity of the situation.

Once background ground water quality is established, the next step is to determine the concentration at which a change in ground water quality would be statistically significant and constitute degradation. Whatever statistical method is used, a statistical decision threshold and a confidence level are necessary. Future downgradient measurements will be compared to this threshold to determine if degradation has occurred. The process for determining degradation is outlined below in Figure 5.

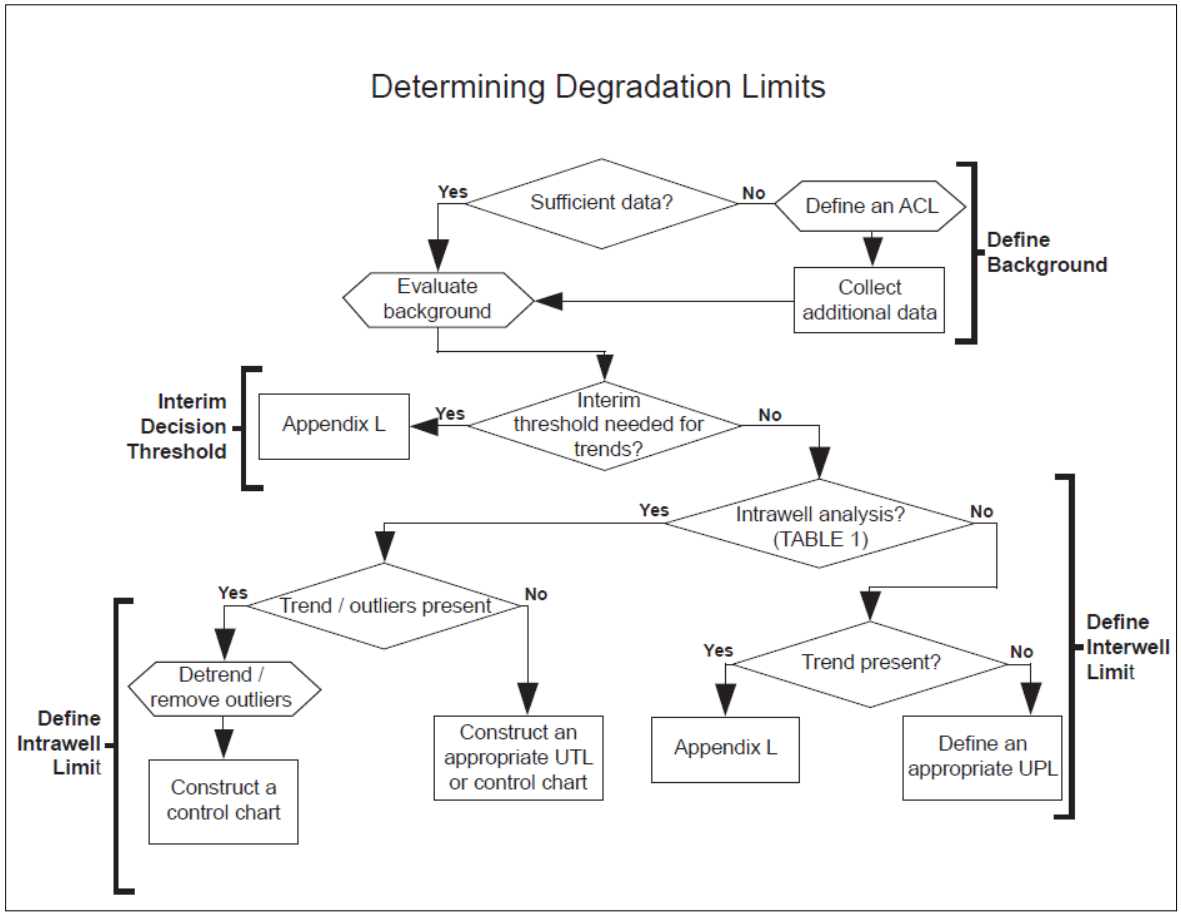


Figure 5. Process for determining degradation limits.

Issues to address when selecting an appropriate statistical decision threshold include the following:

- Are the data adequate to justify a statistically-based decision threshold (or would an ACL be more appropriate [Appendix A])

- Is the activity being evaluated new to the site? (This determines how future downgradient water quality will be evaluated to identify site impacts.)
- Should interwell or intrawell comparison methods be used? (This depends on whether upgradient and downgradient comparisons of wells are possible and defensible.)
- Is a tolerance interval or a prediction interval appropriate and justified for the problem at hand?

4.1 Alternative Concentration Limit

ACLs for COCs are to be estimated when data are insufficient to meet the statistical assumptions for a more detailed statistical analysis. An ACL is to be used as an interim upper limit of background ground water quality. The ACL is anticipated to be used primarily in situations where sufficient data are lacking to adequately define background ground water quality and/or an appropriate statistically defensible upper threshold based on background is not available.

However, an ACL also may be selected even when appropriate data are available for valid statistical analyses if the interested party does not want to conduct statistical analyses and DEQ concurs with the decision or when a rigorous statistical evaluation is not desired, practical, or necessary. ACLs are to be established on a case-by-case basis in consultation with DEQ. The ACL estimation process is described in Appendix A.

4.2 New Versus Existing Activity

An appropriate statistical decision threshold should be chosen in consideration of whether an activity is new or existing. For example, the considerations recommended for wastewater land application sites are shown in Table 1.

Table 1. Considerations for wastewater reuse sites.

Facility with No Previous Site Impact (new or existing)	Facility with Existing Site Impacts
Downgradient wells can also be used to define background ground water quality	Only wells unaffected by the facility's operation can be used to define the background ground water quality
Decisions are made via intrawell comparison	Decisions are made via interwell comparison (or intrawell, if warranted)
Upper tolerance limit (UTL) or Shewhart-CUSUM control chart limits are used set decision threshold	Upper prediction limit (UPL) is used as a decision threshold
Multiple downgradient wells compared to background UTL or individual well compared to its control chart limits	Multiple downgradient wells compared to upgradient UPL; verify exceedance with two independent verification samples.

4.3 Interwell Versus Intrawell Analysis

The objective of degradation analysis is to identify an appropriate background data set against which concentrations in wells potentially affected by a facility can be compared, so as to monitor the facility's impact on local water quality. Generally, interwell comparisons are appropriate

where water quality is spatially homogeneous, sample locations provide statistically independent data, and appropriate upgradient-downgradient comparisons can be identified and defended in the context of the hydrogeologic site conceptual model. Intrawell comparisons can be applied in wells where water quality has not been impacted by site activities and therefore represents background ground water quality at that location. Intrawell comparisons may be preferable in situations where strong spatial variability exists or where a single upgradient well makes it impossible to assess site variability (Appendix N).

4.4 Decision Thresholds and Confidence Levels

Decision thresholds that are commonly used in ground water monitoring are the prediction limit, tolerance limit, simultaneous limit, and confidence limit. An upper prediction limit (UPL) specifies the maximum allowed concentration that 100% of k future measurements must fall below in order to remain in compliance at a designated level of confidence (e.g., 95%); an upper tolerance limit (UTL) specifies the upper limit that a designated percentage (e.g., 95%) of all future measurements at a designated level of confidence (e.g. 95%) must fall below; an upper simultaneous limit (USL) represents the maximum concentration below which 100% of all future comparisons at a designated confidence level (e.g., 95%) must fall; and a confidence interval brackets the range of a specified population parameter (e.g., the mean) at a designated level of confidence (e.g., 95%) (EPA 2009). A discussion of simultaneous intervals involving multiple COCs is beyond the scope of this guidance; DEQ recommends that users consult the Unified Guidance (EPA 2009) for more on USLs. This guidance document assumes that future measurements will be compared to background data on a constituent-by-constituent basis, regardless of the decision threshold.

Prediction and tolerance limits may be applied for compliance sampling in detection, assessment, and monitoring programs since only one initial sample per well is required during the compliance period. These limits also may be used for establishing background-based ground water concentrations. Confidence intervals are most often used when comparing water quality measurements to a ground water standard that is based on a mean or median value (Virginia DEQ 2003). Before calculating these limits, it should be confirmed that the background data are statistically stationary, independent, and parametrically distributed (Appendix O provides more details on these terms).

Concepts to keep in mind when considering confidence intervals are as follows:

- Wider statistical intervals are associated with higher confidence levels (1-α)/lower significance levels (α). However, too high a confidence level decreases the power of the test (the probability of detecting an exceedance) so the confidence level should not be set higher than necessary. Conversely, too low of a confidence level may result in an excessive number of exceedances.
- The conservative choice when testing for a trend or a difference is to use a narrower interval or a lower confidence level (90% or 85%). This lower confidence level would reduce the probability that a difference or exceedance may be missed. In most cases, a 95% confidence level ($\alpha = 0.05$) provides the best compromise between power and confidence.

- For nonparametric methods in which the confidence level depends on sample size, select the highest confidence level dictated by the available sample size. Larger sample size may be needed to achieve a desired confidence level.

4.4.1 Interwell and Intrawell Tolerance Limits

At sites where ground water has not been previously affected by site activities, future water quality measurements can be compared to background ground water quality via an intrawell UTL or an interwell UTL. An intrawell UTL sets the background water quality for each COC in a given monitoring well, and compliance decisions are based on future samples from the same well; it is described in more detail in Appendix H. An interwell UTL is based on background water quality for each COC in one or more upgradient wells, and compliance decisions compare future samples from downgradient wells with the UTL. Application of either method requires that the data have been corrected for seasonal effects and do not display secular trends. For data that meet the above requirements but are not parametrically distributed, a nonparametric UTL can be calculated. Appendix I contains information on the sample sizes needed for nonparametric UTLs. DEQ may accept other decision thresholds for intrawell determination of degradation.

4.4.2 Interwell Prediction Limits

In cases where site conditions indicate that the ground water quality in downgradient wells differs from background conditions (because of existing site practices), data from up to six[†] downgradient wells can be compared to upgradient wells (an interwell analysis) via a parametric or nonparametric UPL calculated from upgradient background data.

Application of the method requires that the data are parametrically distributed, deseasonalized, and that at least eight background measurements are available (EPA 2009). The use of fewer background samples can result in an unacceptably large decision threshold and limited ability to detect exceedances. For data that meet the above requirements but are not parametrically distributed or that have a large proportion of nondetects, a nonparametric UPL can be calculated. However, larger background sample sizes are generally required (Appendix K). DEQ may accept other decision thresholds for interwell comparisons to determine degradation.

4.4.3 Interwell Simultaneous Limits

Typically, when only a small number ($n < 6$; EPA 2013b) of downgradient values are compared with background, an UPL is a suitable decision threshold. When many downgradient measurements need to be compared, an interwell UTL or USL should be used. A USL is similar to an interwell UTL in that it is an upper interval estimate below which the majority of future measurements are expected to fall, and both account for the expectation that multiple comparisons against a decision threshold increase the probability that at least one will exceed the threshold as the number of comparisons increases. The USL overcomes this limitation by appropriately raising the decision threshold to control the false positive rate while maintaining the statistical power of the test to detect an actual exceedance. A USL requires that all

[†] Comparison of more downgradient wells leads to questions of hydrogeologic homogeneity and interwell comparability (EPA 2009)

measurements in the background data set are below that value; if a downgradient value exceeds this threshold, it is by definition not part of the background population and signifies the presence of contamination. Typically, a USL results in fewer false positives than a UTL, especially for larger background sample sizes (n > 15; EPA 2013b).

EPA strongly *recommends* that a USL only be used for downgradient data sets of moderate size (<30 wells) and when the background data set represents a single, well-defined statistical population without outliers (anomalous observations, unrepresentative of background; EPA 2013b).

4.5 Verification Resampling

Resampling is not performed when a UTL is used as the decision threshold. However, when parametric or nonparametric UPLs (Appendices J and K) are used to detect degradation, EPA's Unified Guidance recommends that a verification resampling strategy be used to minimize false positives and maximize the statistical power of the decision threshold. For each compliance well where a prediction limit is exceeded, DEQ recommends collecting up to two additional samples (allowing sufficient time between resampling to ensure statistical independence). This protocol conforms to EPA's *1-of-3* retesting option: only if the first resample is in compliance is a second resample necessary; otherwise the first resample, by also being out-of-bounds, confirms the initial exceedance (EPA 2009, section 19.1).

4.6 Trending Data

In cases where background water quality data describe a trend, the trend must be evaluated to determine its cause. Water quality may show a trend in response to (1) natural circumstances or (2) anthropogenic activities such as land use changes. A trend can be seasonal, monotonically positive or negative and can disappear or grow over time. If monotonic, particular care should be exercised that a future change in trend can be detected (Appendix L provides an example). Only if trend removal can be rigorously justified both statistically and hydrogeologically—and quantified (e.g., via deseasonalization or ordinary least squares)—can the trend be removed from a data set prior to background analysis. In that case, the trend is also removed from all future samples and verification resamples prior to compliance testing with a UTL, UPL, or other exceedance-detection threshold. An example is provided in Appendix F, and Appendix L describes methods for establishing interim decision thresholds when a monotonic trend may be approaching a new stationary background level, either because of natural flow system change or modified site or land use practices.

Appendix M and Appendix N provide examples of how the guidance may be applied to new and existing wastewater reuse sites.

5 Summary of Process

5.1 Determination of Background Ground Water Quality

New sites have the advantage that all monitoring wells, regardless of whether they are upgradient or downgradient, can be used as background monitoring wells. For example, at wastewater reuse facilities where land is being converted from another land use (such as irrigated agriculture) to treated wastewater application, it is possible that some or all wells will not have attained a steady state condition or that downgradient wells will have reached a different steady state condition than upgradient wells.

For a new site or new unused acreage at an existing site, such as a wastewater reuse permit facility that has yet to have any treated wastewater applied, the first step is to conduct descriptive statistics on the COCs in all wells (Appendix B). Following the initial descriptive statistical tests, each of the monitoring wells should be evaluated for data independence (Appendix C). The form of the data distribution (parametric or nonparametric) should be determined next (Appendix D).

Statistically significant seasonal trends for each of the COCs (Appendix E) are then evaluated and such trends removed to produce a seasonally stationary data set. As the regulated entity is required to evaluate at least 3 years of quarterly data (where each quarter's data represents the same month from year to year), some of the background ground water quality variation may be due to changing land use practices (e.g., nearby agricultural activities and river and canal flows) or climatic changes (e.g., precipitation patterns, and evapotranspiration). The preferred method for determining seasonal stationarity is the nonparametric Kruskal-Wallis test (Appendix E).

The resulting data set should then be checked for the presence of secular (long-term) temporal trends (Appendix F). If a trend exists, then setting degradation thresholds may not be statistically valid and can lead to erroneous conclusions. The recommended method for testing for temporal stationarity is the nonparametric Mann-Kendall test for trend.

If the Mann-Kendall test shows that there is a statistically significant secular trend (either positive or negative), then an alternative method needs to be followed to set the standard(s) that the regulated entity will need to follow (section 5.2.1). If the Mann-Kendall test reveals no secular trends, the regulated entity can proceed to determine whether the data from multiple background wells can be pooled (Appendix G).

5.2 Determination of Degradation

At this point, background ground water quality has been rigorously evaluated and its statistical characteristics identified. The next step is to define appropriate thresholds against which future measurements can be compared to identify potential water quality degradation.

5.2.1 Intrawell Comparisons

Parametric tolerance levels for intrawell comparisons can be set using the methodology provided in Appendix H. An intrawell analysis allows future constituent levels in a well to be compared to the limit established by that well's own background ground water quality. To use this method,

one must have a data set that (1) is stationary (free of secular trends and has no statistically significant seasonal effects or has been corrected for seasonality), (2) is parametrically distributed, and (3) represents a site where the ground water has not been impacted by previous site activities. Appendix I provides a methodology for determining nonparametric tolerance limits (where the same assumptions apply). Future water quality for each COC in each well is to be compared to the UTL in each well. If the rate of exceedances observed in future sampling is greater than that used to establish the tolerance limit (e.g., 5% of all future measurements), then the ground water is deemed to be degraded. As an alternative to the use of tolerance limits, Appendix N describes the Shewhart-CUSUM control chart method, which monitors gradual and rapid contaminant impacts within a well in a time-series context.

5.2.2 Interwell Comparisons

The methodology for setting parametric prediction levels for interwell analyses is provided in Appendix J. In this case, an UPL is defined on the basis of upgradient water quality data. To use this method, one must have a data set that (1) is stationary (free of secular temporal trends and has no statistically significant seasonal effects or has been corrected for seasonality), and (2) meets the parametric distribution assumptions. Site conditions must be such that the downgradient well water quality can be compared to upgradient water quality (an interwell analysis). Appendix K assumes the same conditions as Appendix J except the distribution is nonparametric. A specified number of future water quality measurements in downgradient wells are compared to the UPL established in upgradient wells; any exceedance should be verified by the resampling procedure discussed in Appendix J.

Within the context of this guidance, a verification sample is a sample that exhibits data independence from the result it is attempting to verify. It is distinct from a measurement confirmation sample, which is a sampling event used to confirm the initial result.

5.2.3 Treatment of Verification Versus Confirmation Resampling Data

Verification resampling measurements collected in response to a potential exceedance can be retained and treated as part of the overall background data set if (1) the samples were collected so as to be temporally independent (section 4.5 and Appendix J, section J.1), and (2) an exceedance was not confirmed. In contrast, confirmation (i.e., duplicate) sampling results should be averaged to provide a single measurement value for use in subsequent statistical analyses.

5.2.4 Interim Methodology for Trending Data

Appendix L outlines a suggested procedure for setting an interim UPL for situations that violate the stationarity assumption (i.e., where the data show a secular trend).

6 References

Conover, W.L. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York, NY: John Wiley & Sons.

Cressie, N.A.C. 1993. *Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics*. New York, NY: John Wiley & Sons, Inc.

EPA (United States Environmental Protection Agency). 2006. *Data Quality Assessment: Statistical Methods for Practitioners.* Washington, DC: Office of Environmental Information. EPA QA/G-9S. EPA/240/B-06/003. *http://www.epa.gov/quality/qs-docs/g9s-final.pdf.*

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

EPA (United States Environmental Protection Agency). 2013a. *ProUCL 5.0 Software and User Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

EPA (United States Environmental Protection Agency). 2013b. *ProUCL 5.0 Technical Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

Fisher, S.R. and K.W. Potter. 1989. *Methods for Determining Compliance with Groundwater Quality Regulations at Waste Disposal Facilities*. Submitted to the Wisconsin Department of Natural Resources.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York, NY: Van Nostrand Reinhold.

Harris, J., J.C. Loftis, and R.H. Montgomery. 1987. "Statistical Methods for Characterizing Ground-Water Quality." *Ground Water*. 25(2): 185–193.

Helsel, D.R. 1990. "Less Than Obvious: Statistical Treatment of Data Below the Detection Limit." *Environmental Science & Technology*. 24(12).

Helsel, D.R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. New York, NY: John Wiley & Sons, Inc.

Ogden, A.E. 1987. *A Guide to Groundwater Monitoring and Sampling. Idaho Department of Health and Welfare*. Boise, ID: Division of Environment. Water Quality Report No. 69.

Virginia Department of Environmental Quality. 2003. *Data Analysis Guidelines for Solid Waste Facilities.* Richmond, VA: Virginia DEQ.

This page intentionally left blank for correct double-sided printing.

Appendix A. Alternative Concentration Limits

Alternative concentration limits (ACLs) for constituent(s) of concern are estimated when there are insufficient data to meet the statistical assumptions for a more detailed statistical analysis.

The following three measures of upper concentration limits are calculated from available data.

1. $ACL_1 =$ the largest of the 12 most recent data values collected
2. $ACL_2 = mean + 1.65s$
3. $ACL_3 = \text{median} + 1.65 * \text{IQR}$ (where IQR = the interquartile range)

The Idaho Department of Environmental Quality (DEQ) specifies that the lowest of these limits is then to be used as an interim upper limit of background ground water quality in order to be fully protective of human health and the environment in situations where sufficient data are lacking to adequately define background ground water quality and/or an appropriate statistically defensible upper threshold based on background is not available.

ACLs are to be established on a case-by-case basis in consultation with DEQ.

This page intentionally left blank for correct double-sided printing.

Appendix B. Exploratory Data Analysis/Descriptive Statistics

B.1. Descriptive Statistics

Once adequate data have been collected, the data are analyzed using descriptive statistics to characterize the overall population. For each constituent, at a minimum, the user should calculate the mean, standard deviation, skewness, median, minimum, and maximum for each constituent at each monitoring well, as well as summarize the sample size, proportion of censored data, and potential outliers and how they are treated in the subsequent analysis. In addition, visual representations of the distribution of each constituent should be provided in the form of box plots, histograms, and concentration-time plots for each constituent in each well.

As has been previously stated, the reason for collecting sample data is to understand the underlying statistical distribution of the ground water source from which the samples were drawn. The sample mean provides a measure of the central tendency of the population, whereas the sample standard deviation provides a measure of its spread, or dispersion. The measurements represent just one of many possible subsets of data that could have been collected from the entire population. Different samples will obviously lead to different values of the sample mean and sample standard deviation. These differences are the reason why statistical intervals are used to infer population parameters and set decision thresholds.

In any set of data, it is possible that there will be outliers (anomalous results). Outliers can have one of three causes: (1) a measurement or recording error, (2) an observation from a different population, or (3) a rare event from the tail of the population of interest. Outliers can be discarded from the data set with adequate justification. For example, a valid justification for removing an outlier would be the simultaneous occurrence of extreme values in four independent data sets on the same day. This type of event would strongly suggest either a field contamination issue or a lab error. The United States Environmental Protection Agency's (EPA's) Unified Guidance (EPA 2009) and ProUCL User's Guide (EPA 2013) provide additional guidance on how outliers should be handled.

Table B1 provides a simple example of calculating summary statistics manually. ProUCL v.5.0 also has useful tools for calculating summary statistics, with or without nondetect values, although the inclusion or exclusion of outliers is a decision that must be justified by the user.

Other descriptive statistics include the median, minimum, maximum, and quartiles. The median and quartiles are not affected by outliers unlike the sample mean, standard deviation, and skewness.

A graphical summary of the data, including the relevant constituents of concern (COCs), should provide box plots, showing at least the median, minimum, maximum, and quartiles for each COC and time-series plots. The latter provide a visual indication of whether there is a seasonal component to the data, whether there is a secular (long-term) trend, and/or whether the trend has changed or may be changing (approaching a new steady-state condition). An example of a box plot is provided in Figure B1.

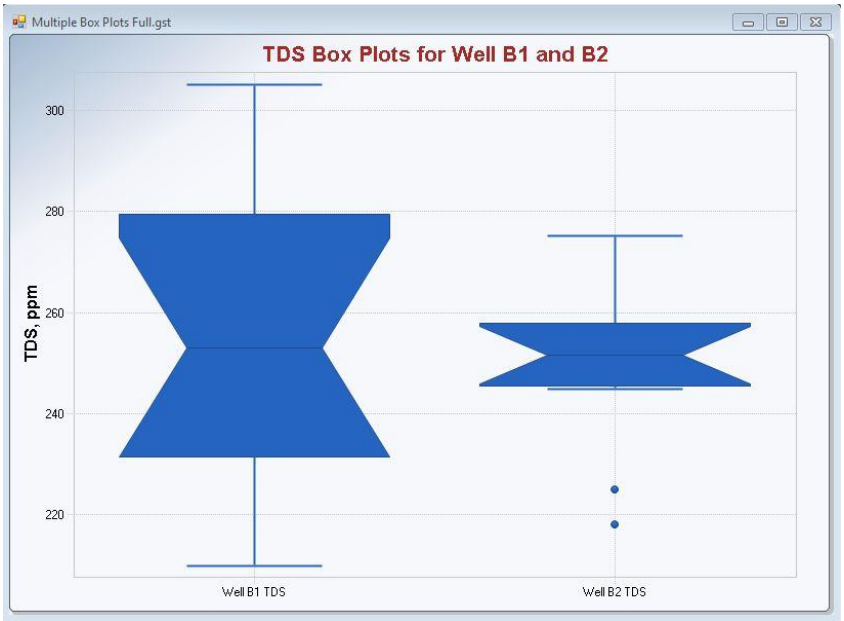


Figure B1. Box plots created by ProUCL 5.0 illustrating the centermost half of the data (the *hourglass* polygon) that straddles the median (the polygon's *waist*), the interquartile range (upper and lower limits of the polygon); the upper and lower adjunct limits (horizontal lines; Appendix O, section O.3); and extreme values (dots) that lie beyond the adjunct limits.

B.2. Example

Where appropriate, data from the following scenario will be used to illustrate selected applications in the various appendices of this guidance document. A wastewater reuse facility wants to determine a background ground level for total dissolved solids (TDS), above which there is a certain degree of statistical confidence that elevated values would indicate degradation of ground water quality. The facility has two background wells (B1 and B2). Well B1 is located near an irrigation canal and the canal may seasonally influence the water quality. Well B2 is located away from the irrigation canal. Three years of quarterly data have been collected at each monitoring well. Table B1 lists the TDS data in parts per million (ppm) and also provides summary statistics for the TDS data. Figure B2 summarizes these data graphically in the form of a time-series plot.

Table B1. Data and resulting descriptive statistics for example scenario.

Time Index	TDS Well B1 (ppm)	TDS Well B2 (ppm)
Year 1—1st quarter	305	252
Year 1—2nd quarter	228	251
Year 1—3rd quarter	258	245
Year 1—4th quarter	259	252
Year 2—1st quarter	285	260
Year 2—2nd quarter	210	248
Year 2—3rd quarter	274	275
Year 2—4th quarter	240	272
Year 3—1st quarter	290	256
Year 3—2nd quarter	216	246
Year 3—3rd quarter	248	218
Year 3—4th quarter	235	225
Descriptive statistics[a]		
Mean	254	250
Variance	904	268
Standard deviation	30.0	16.4
Skewness	0.20	-0.52
Minimum	210	218
Maximum	305	275
Median	253	252
1st quartile	231	246
3rd quartile	280	258

a. Excel software used for calculations.
Note: total dissolved solids (TDS); parts per million (ppm)

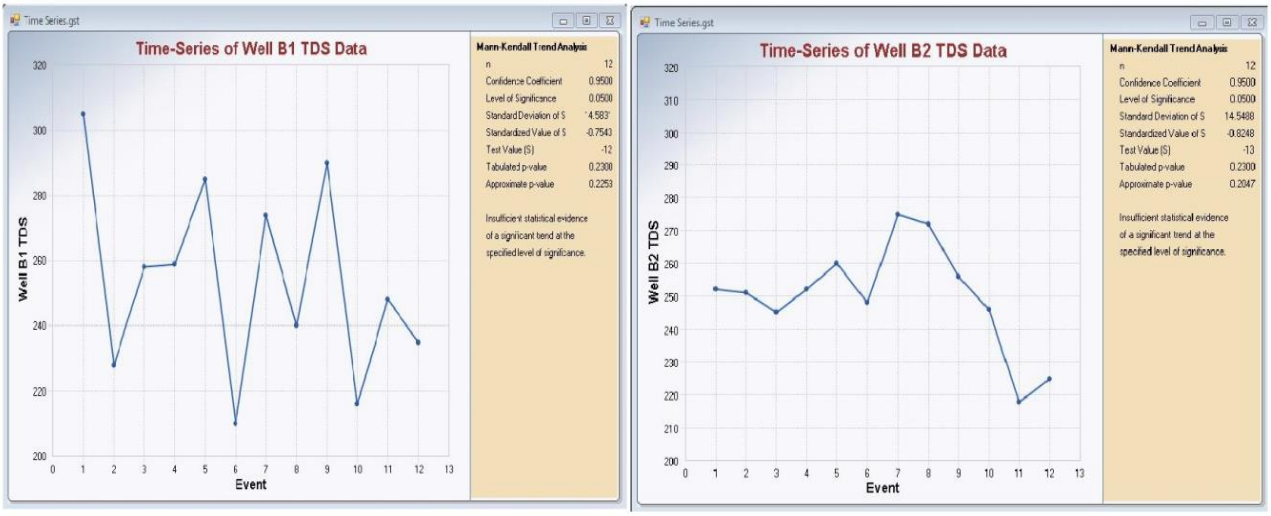


Figure B2. Time-series graphs of concentration data in Table B1 created with ProUCL.

References

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

EPA (United States Environmental Protection Agency). 2013. *ProUCL 5.0 Software and User Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

Appendix C. Data Independence

C.1. Introduction and Background

All of statistical theory and practice is based on three fundamental premises. First, a collection of measurements represents a random sample of the underlying population that is free of any bias imposed by the measurement process (e.g., the individual who conducted the sampling or the analytical method used to make the measurements).

Second, the statistics estimated for a population depend on the measurements used to make the estimate but bracket the true population statistics (i.e., the sample is representative), allowing us to infer the population statistics from any sample.

Third and perhaps most important from a practical standpoint, the data are assumed to be independent: that is, each measurement is randomly representative of the target population and its value is not influenced by any other measurement (i.e., each measurement is independent of every other). Dependent measurements exhibit less variability; for example, multiple measurements of dissolved nitrate collected from a well at 5-minute intervals are all very similar, which leads to an underestimation of the population variance that in turn affects the prediction limit and tolerance limit. In reality, every measurement of the physical world is to some degree dependent on (similar to, correlated with) previous or nearby measurements; such dependence is known as autocorrelation. For example, replicate measurements of stream chemistry are much more similar to each other than measurements collected a year apart. In another example, consider the analysis of dissolved nitrate in water from five wells (N = 5) where, for quality control, four aliquots of water are collected and analyzed from well 5. When calculating the average nitrate concentration in the five wells, the replicates cannot be treated as separate, random outcomes in a sample of N = 9 measurements because they constitute redundant information about the population. If nitrate in well 5 happens to be twice the average concentration of the other four wells and all nine measurements are averaged, then the apparent mean would be biased high by 50% and the apparent variance would be far lower than the actual population variance.

Every statistical procedure in this document assumes that the data being analyzed are independent. If the data are not independent, the effect is generally to decrease the power of hypothesis tests (e.g., reducing the ability to detect an exceedance), particularly tests that rely on an unbiased estimate of the variance.

Unfortunately, there is no general method for testing for data independence and little practical guidance is available in the literature. The onus is on the statistical analyst to evaluate the data on a case-by-case basis. The purpose of this appendix is to suggest possible approaches that can be used to evaluate data independence; where that is not possible, to provide some general guidelines to evaluating site-specific sampling conditions to minimize the risk of bias due to lack of data independence.

C.2. Example: Evaluating Data for Temporal Independence

Significantly more attention has been given to the issue of temporal autocorrelation than to spatial autocorrelation. For serial data (i.e., time-series, temporally sequenced, time-variable

data), two different approaches have been taken: (1) demonstrating physical independence between successive samples based on the minimum time required for ground water to move past the sampling point (EPA 2009), and (2) evaluating time-series data to characterize the time scale associated with statistical independence (Barcelona et al. 1989; Oswina et al. 1992; Johnson et al. 1996; Ridley and MacQueen 2005; EPA 2009).

The basic requirement is that sufficient time must elapse between sampling events to ensure independence. A commonly applied rule of thumb is that data to be used for statistical analysis and hypothesis testing should be collected no more frequently than quarterly (Gibbons 1994), but this guideline may not apply in aquifers with very slow ground water flow rates. The Idaho Department of Environmental Quality (DEQ) recommends that estimates of ground water flow velocity and travel times be used to confirm the validity of the quarterly rule of thumb using the procedure based on Darcy flow velocity outlined in the Unified Guidance (EPA 2009, section 14.3.2); For very small sample sizes ($N < 12$), a method proposed by Ridley and MacQueen (2005), based on decision-tree logic, could be adopted until sufficient historical data are available, but the method is cumbersome. Where sufficient data are available ($N > 20–30$), standard time-series analysis methods can be applied (Salas 1993). For example, a basic autoregression analysis (e.g., a Box-Jenkins autocorrelation plot available in statistical software packages such as R and MATLAB), can be applied. Prior to analysis, the data should be detrended but not deseasonalized. Alternatively, a one-dimensional semivariogram (Oswina et al. 1992) can be computed using standard geostatistical software (Deutsh and Journel 1998). However, there is little benefit to be gained from such an exercise if data have been collected no more frequently than quarterly (in which case, the quarterly rule of thumb can be applied).

As an example, the total dissolved solids (TDS) data in Figure C1 represent varying sampling intervals over a 3-year period, with an indication that monthly measurements tend to be more similar than quarterly measurements (i.e., autocorrelated). The raw (non-deseasonalized) data of Figure C1 were first reduced by calculating the quarterly averages of multiple measurements within each quarterly span (Washington State 2005). A plot of lag-dependent autocovariance (Box-Jenkins autocorrelation plot of Figure C2) or first-order autocorrelation coefficients (EPA 2009, section 14.2.3) can be created in a simple spreadsheet or in a statistical package like R or MATLAB. The autocovariance and autocorrelation coefficient are a function of lag (separation in time) and both decay from a value equivalent to the sample variance at zero lag to near zero at a lag of 180 days. Therefore, the TDS data used to create Figure C1 demonstrate that future samples will be statistically independent if collected at a frequency no greater than once every 180 days. However, this sampling interval may be longer than necessary because quarterly averages of the data were used in calculating the autocorrelation statistic rather than monthly data.

An alternative approach to using a time-series statistical package is to calculate a one-dimensional semivariogram of the detrended data using a geostatistical software package (Oswina et al. 1992; Johnson et al. 1996), where the spatial x,y coordinates are replaced by a time coordinate. Such a plot, using the monthly data in Figure C1, is shown in Figure C3. Unlike the autocorrelation statistics, the value of the semivariogram statistic in this plot rises rapidly to a sill value that is at or near the sample variance.

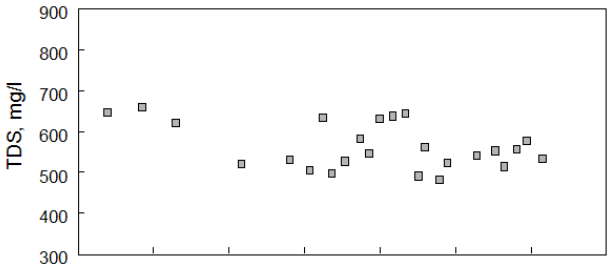


Figure C1. Example ground water TDS measurements used for evaluating the statistical independence of a time-series data set.

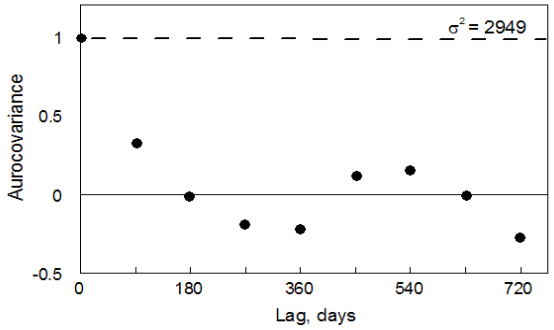


Figure C2. Box-Jenkins autocovariance plot created using the quarterly-averaged data in Figure C1. The value of the autocovariance statistic decays to zero at a lag of about 180 days, suggesting that measurements spaced at least semiannually are statistically independent.

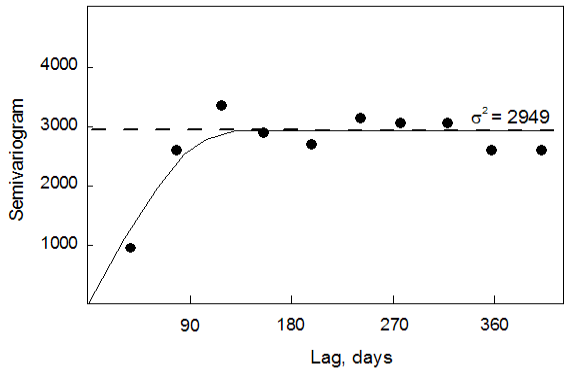


Figure C3. An example of a semivariogram (Isaaks and Srivastava 1989) computed for all of the TDS data in Figure C1, showing that time-series measurements at this location are statistically independent if made at least 90 to 100 days apart (i.e., about quarterly).

The lag at which the sill is achieved (approximately 90 to 100 days in this example) represents the minimum time interval over which measurements can be considered to be statistically independent. In this case, the minimum time interval proves to be shorter than that of Figure C2 because the monthly data were analyzed rather than the quarterly.

C.3. Evaluating Data for Spatial Independence

Guidance for evaluating spatial data independence is almost completely absent in the literature. Ideally, a geostatistical analysis could be conducted on spatial data in the same way that the one-dimensional semivariogram of Figure C3 was calculated for temporal data, the difference being that lags are defined in a two-dimensional spatial sense rather than a temporal sense (Isaaks and Srivastava 1989; Bertolino et al. 1983; Cameron and Hunter 2002). Because the number and spacing of monitoring locations (as well as the availability of data at each well) determine how useful such an analysis will be, the approach is rarely fruitful in small monitoring networks (<5 to 10 wells). Small monitoring networks typically have insufficient wells with which to calculate reliable semivariogram statistics, so that minimum interwell spacing to ensure data independence cannot be determined. Figures C4 and C5 illustrate the problem conceptually.

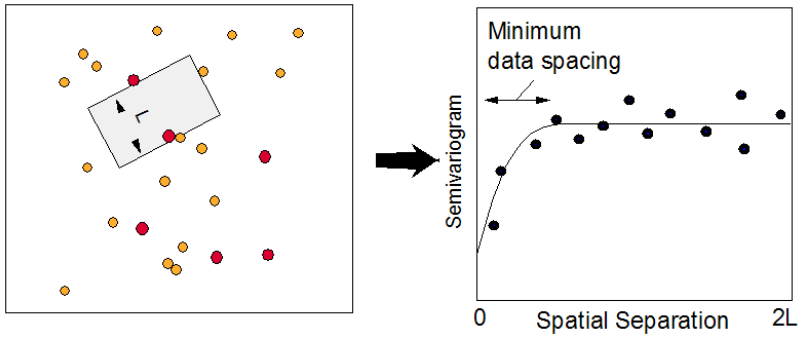


Figure C4. A hypothetical example of a semivariogram based on a sufficient number of monitoring points to construct a well-defined semivariogram and identify the minimum interwell spacing necessary to maintain spatial data independence. In this example, wells that are at least one-half L apart provide statistically independent information.

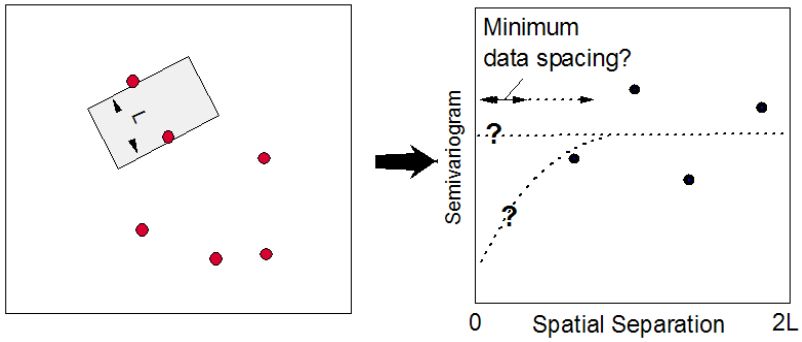


Figure C5. Semivariogram based on insufficient data to define spatial independence.

As Gibbons (1994) and others have pointed out, the spatial variability of water quality across a monitoring site is as important as interwell spacing considerations. If the aquifer is highly heterogeneous, then the assumption of spatial data independence may be violated for reasons other than well spacing: that is, contaminant concentrations in an active flow zone are biased high relative to concentrations in slow or inactive flow zones. Wells completed in hydraulically *tight* units will tend to reflect lower values of contaminant concentration than wells in more permeable zones, even if they are hydraulically downgradient of the contaminant source. It is for

such reasons that intrawell evaluation methods may be the only rational alternative when evaluating monitoring data from highly heterogeneous aquifers (Gibbons 1994).

Because of the above concerns, DEQ suggests that a qualitative assessment of spatial data independence should be performed, including but not limited to the following:

1. Estimates of ground water flow velocity and interwell travel times (EPA 2009); minimum well spacings in high flow-velocity zones will be greater than in low velocity zones.
2. The existence of a spatial trend in concentration across the facility tends to strengthen autocorrelation in that direction, so that minimum well spacing to ensure data independence in the direction of the trend will be less than across it.
3. If data from multiple upgradient wells cannot be pooled (Appendix G) because of hydrochemical variability across the site or if considerable hydraulic heterogeneity exists, then intrawell methods (Appendices H, I, N) should be adopted if at all possible.

In general, intrawell analysis methods should be explored wherever aquifer heterogeneity is significant and/or water quality is highly variable across a site. In such situations, DEQ may grant site-specific variances based on modifications of the methods contained in this document or other methods.

References

Barcelona, M.J., H.A. Wehrman, M.R. Schock, M.E. Sievers, and J.R. Karny. 1989. *Sampling Frequency for Ground-Water Quality Monitoring*. Washington, DC. EPA/600/4-89/032.

Bertolino, F., A. Luciano, and W. Racugno. 1983. "Some Aspects of Detection Networks Optimization with the Kriging Procedure." *Metron*. 41(3): 91–107.

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.

Deutsch, C.V. and A.G. Journel. 1998. *GSLIB: Geostatistical Software Library and User's Guide*. New York, NY: Oxford University Press.

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Isaaks, E.H. and R.M. Srivastava. 1989. *Applied Geostatistics*. New York, NY: Oxford University Press.

Johnson, V.M., R.C. Tuckfield, M.N. Ridley, and R.A. Anderson. 1996. "Reducing the Sampling Frequency of Ground-water Monitoring Wells." *Environmental Science & Technology*. 30(1): 355–358.

Oswina, A., U. Lall, T. Sangoyomi, and K. Bosworth. 1992. *Methods for Assessing the Space and Time Variability of Groundwater Data*. Washington, DC: United States Geological Survey. PB-94 116548.

Ridley, M. and D. MacQueen. 2005. *Cost-Effective Sampling of Groundwater Monitoring Wells: A Data Review & Well Frequency Evaluation*. Livermore, CA: Lawrence Livermore National Laboratories. UCRL-CONF-209770.

Salas, J.D. 1993. "Analysis and Modeling of Hydrologic Time Series." *Handbook of Hydrology*. D.R. Maidment, ed. New York, NY: McGraw-Hill.

Washington State Department of Ecology. 2005. *Implementation Guidance for the Ground Water Quality Standards*. Olympia, WA: Washington State Department of Ecology.

Appendix D. Determination of Normality and Choice of Distribution

D.1. Testing for Normality Using the Shapiro-Wilk Test

The importance of correctly determining the nature of the underlying population from which samples are drawn cannot be overemphasized. The primary reason to test whether data follow a normal or other theoretical (parametric) distribution is to determine whether or not parametric test procedures can be employed in subsequent statistical analysis. The ability to apply parametric statistical tests conveys higher statistical power, a lower false-positive error rate, and more confident conclusions overall.

Statistical hypothesis tests of a sample distribution are based on a null hypothesis, H_0, which states that the data set represents a specific type of parametric population (e.g., normal, lognormal, and gamma). An appropriate test statistic is calculated from the sample data (section D.2) and compared against a tabulated statistic to determine whether H_0 can be accepted or rejected. Failure to reject H_0 does not prove the data were drawn from a normal population (especially when the sample size is small), only that the hypothesis of normality cannot be rejected with the available evidence and at the stated level of significance (usually 0.05 or 5%). A significance level greater than 0.05 increases the power of the test (its ability to reject H_0) but at the expense of falsely detecting non-normality (a false positive result). The use of higher significance levels (e.g., 0.1 or 0.15) is particularly useful when testing very small sample sizes (Helsel and Hirsch 1995).

A critical aspect of the testing procedure concerns right-skewed data sets that contain a few high values. This is especially true when the sample size is small ($N < 20–30$). It has been a long-standing practice to logarithmically transform such data sets, test the log-transformed values for normality, and conclude that the distribution is lognormal. With growing awareness of the advantages and power of the gamma distribution (Appendix O) and of software to perform the statistical calculations, The United States Environmental Protection Agency's (EPA's) Unified Guidance now strongly recommends that skewed sample data sets should not be modeled as lognormal distributions but as gamma-distributed populations (EPA 2009, 2013a, 2013b). The principal reason is that a lognormal transformation disguises the effect of high values that may not represent background and exaggerates the apparent standard deviation of the modeled lognormal distribution. This, in turn, inflates decision thresholds that are based on the lognormal statistics, leading to incorrect and unrealistically high UPLs and UTLs that will be used to determine future compliance.

Therefore, the Idaho Department of Environmental Quality (DEQ) recommends that skewed sample data be modeled either as gamma-distributed or as nonparametric, particularly if the sample size is less than 20 and/or contains outliers. Because of the gamma function's flexibility in accommodating a wide range of symmetric and asymmetric (skewed) distributions, it is capable of representing a lognormally distributed data set without the risk of masking the effects of outliers.

Figure D1 illustrates the general approach to hypothesis testing and determining an appropriate population distribution model. The flow chart represents decisions that would be made when

manually calculating test statistics for various hypothesis tests. Many types of statistical software (e.g., ProUCL, Minitab, and S-plus) allow alternative hypothesis tests to be executed simultaneously, in which case Figure D1's logic applies to the decision process to be used in assessing the output of the software's hypothesis test results.

A Shapiro-Wilk test for normality is a widely used hypothesis test that would be applied in a series of tests like those outlined in Figure D1. This test is superior to the chi-square test (EPA 1988; EPA 2009; Fisher and Potter 1989) and is recommended because it is based on the normal probability plot (Helsel and Hirsch 1995). The Shapiro-Wilk test is designed for data with less than 10%–20% censoring, wherein all censored measurements up to this limit are withheld from the calculation. If censoring is greater than 20%, then either Royston's method (Royston 1993) or an appropriate adjustment to the sample standard deviation (Cohen 1991; Aitchison 1955) must be applied when using this test statistic. The test is based on the premise that the ranked sample values should be highly correlated with the corresponding quantiles taken from a normal distribution if the data set is normally distributed (Shapiro and Wilk 1965). An example of its application is given in section D.2.

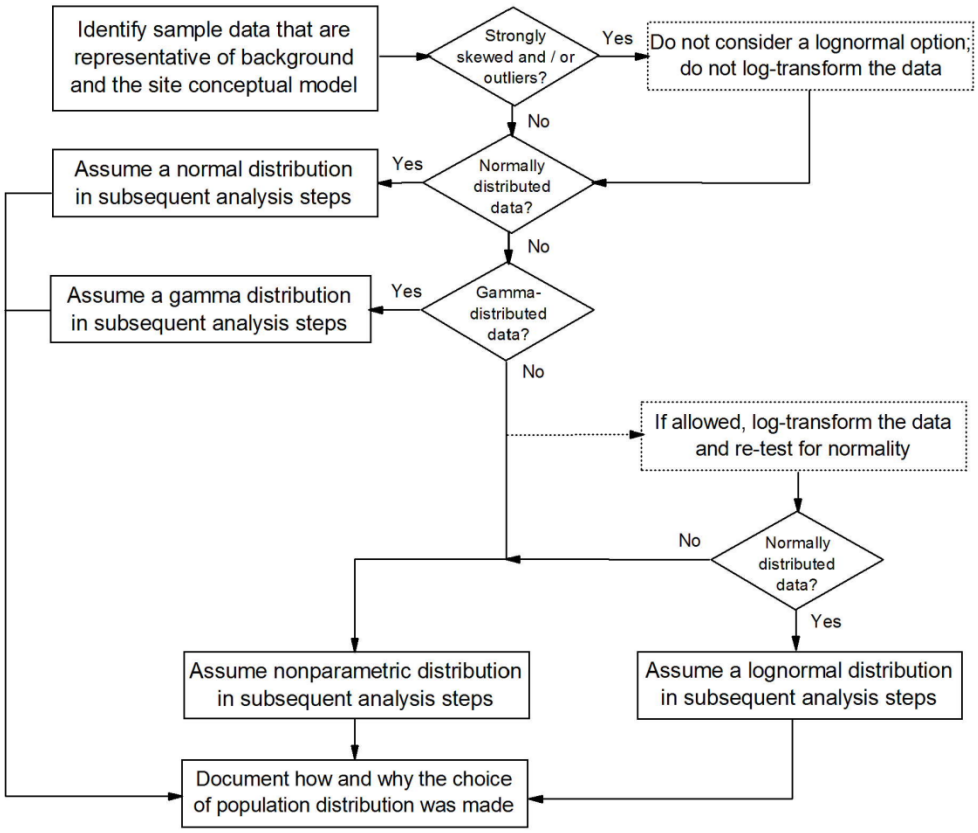


Figure D1. Decision tree for determining a population distribution.

D.2. Calculation Procedure

The Shapiro-Wilk statistic "W" is proportional to the ratio of the squared slope of the normal probability plot to the mean square estimate (Gibbons 1994):

$$W = \frac{\left(\sum_{i=1}^{n} a_{i,n} x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Consider the total dissolved solids (TDS) data for well B1 (Appendix B and Table D1). The null hypothesis, H_0, is that the sample data are normally distributed. The coefficients $a_{i,n}$ for the W statistic are given in Table D2 (Gibbons 1994 provides a more complete table). Recalling that *s* is the standard deviation, W can be re-expressed as

$$W = \left[\frac{b}{s\sqrt{n-1}}\right]^2$$

where

$$b = \sum_{i=1}^{k} a_{n-i+1}\left(x_{(n-i+1)} - x_i\right) = \sum_{i=1}^{k} b_i$$

The calculation procedure is as follows:

Step 1: Order the data from smallest to largest and list, as in Table D1. Also list the data in reverse order alongside the first column.

Step 2: Compute the differences $x_{(n-i+1)}$-x_i in column 3 of Table D1 by subtracting column 1 from column 2.

Step 3: Compute k as the greatest integer less than or equal to n/2. k = (n-1)/2 if n is odd and k = n/2 if n is even. Since n = 12, k = 6 in this example.

Step 4: Look up the coefficients a_{n-i+1} from Table D2 and list in column 4. Multiply the differences in column 3 by the coefficients in column 4 and add the first k products to get the quantity b.

Step 5: Compute the standard deviation of the sample (9.77) and calculate W (0.861).

Step 6: Compare the computed value of W to the 5% critical value, equivalent to an α value of 0.05 (Table D3) for a sample size of 12 (0.859).

Table D1. Example of Shapiro-Wilk test for normality on TDS data from well B1.

Ranked data value	x_i	$x_{(n-i+1)}$	$x_{(n-i+1)}-x_i$	a_{n-i+1}	b_l
1	242	268	26	0.5475	14.4175
2	244	268	24	0.3325	7.869167
3	246	266	20	0.2347	4.694
4	246	264	18	0.1586	2.8548
5	249	252	3	0.0922	0.245867
6	251	252	1	0.0303	0.0404
7	252	251	-1	—	b=30.12
8	252	249	-3	—	—
9	264	246	-18	Std_dev =	9.77
10	266	246	-20	—	—
11	268	244	-24	W =	0.8612
12	268	242	-26	—	—

The closer the value of W is to 1.0, the greater is the support for the normality assumption. The assumption of normality is rejected if the computed value of W is less than W's critical value in Table D3. In this case, the null hypothesis is accepted because W (0.861) is greater than the critical value (0.859). Therefore, the data can be assumed to be normally distributed.

The process for testing for lognormally distributed data is the same, except that the data are log-transformed prior to performing the Shapiro-Wilk hypothesis test.

Table D2. Partial list of coefficients a_i for the Shapiro-Wilk test of normality.

# of data	8	9	10	11	12	13	14	15	16
k									
1	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056
2	0.3031	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290
3	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521
4	0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939
5	—	0.0000	0.0399	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447
6	—	—	—	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005
7	—	—	—	—	—	0.0000	0.0240	0.0433	0.0593
8	—	—	—	—	—	—	—	0.0000	0.0196

Sources: Complete tables in Shapiro and Wilk (1965); (EPA 1992);and Gibbons (1994).

Table D3. Lower 1% and 5% critical values for Shapiro-Wilk test statistic W.

Sample Size	1% W Value	5% W Value	Sample Size	1% W Value	5% W Value
8	0.749	0.818	13	0.814	0.866
9	0.764	0.829	14	0.825	0.874
10	0.781	0.842	15	0.835	0.881
11	0.792	0.850	16	0.844	0.887
12	0.805	0.859	—	—	—

Sources: Complete tables in Shapiro and Wilk (1965); EPA (1992); and Gibbons (1994).

D.3. Calculation Example Using ProUCL 5.0: Weakly Skewed Data

For the purposes of this example, the raw TDS data in Table B1 for well B1 and B2 were pooled naively, that is, without considering whether they should be (Appendix G) and without testing and adjusting for seasonality or trend (Appendices E and F). Figure D2 shows a histogram of the data values with five histogram bins. The sample data appear to be only slightly skewed (skewness = -0.56); there are no apparent outliers; and the data distribution is close to symmetric (the mean and median are almost identical). A visual assessment cannot determine whether the data are normally distributed or not.

Figure D3 shows output from ProUCL's "Goodness-of-Fit Tests" option showing Q-Q plots and fitting statistics when the data are compared with three types of parametric distributions (normal, lognormal, and gamma). Table D4 is a screen capture from ProUCL's "G.O.F. Statistics" option that summarizes goodness-of-fit statistics for the three Q-Q plots. Note that the Shapiro-Wilks test described in section D.2 and the Lilliefors test are used to evaluate normal and lognormal fits, whereas two different statistical tests, the Anderson-Darling and Kolmogorov-Smirnov tests, are used to evaluate the fit to a gamma distribution.

In this case, the similarity of the regression fits to the various Q-Q plots (Figure D3) and the goodness-of-fit results (Table D3) indicate that there is no basis for selecting a *best* distribution. Therefore, using the logic of Figure D1, the normal distribution assumption should be used in all subsequent analysis steps. In other situations, the flow logic of Figure D1 should be used to determine which distribution is appropriate (e.g., a slightly skewed data set with no outliers that passes the gamma and lognormal distribution tests but not the normal test should be modeled with the gamma distribution). When in doubt, more than one distribution option can be carried forward to ensure that the choice does not have a major influence on subsequent conclusions and that decision thresholds computed on the basis of different possible distributions do not differ widely. Appendix K, section K.3, provides further discussion of the dependence of decision thresholds on the choice of a distribution.

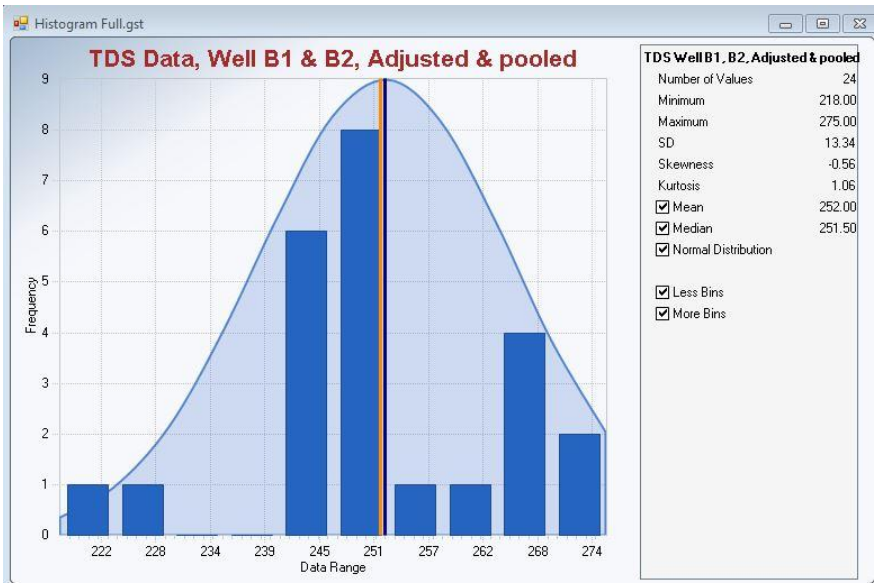


Figure D2. Histogram of pooled TDS data from Table B1, superimposed on a normal distribution for visual comparison purposes. The mean (black line) and median (orange line) are shown for reference.

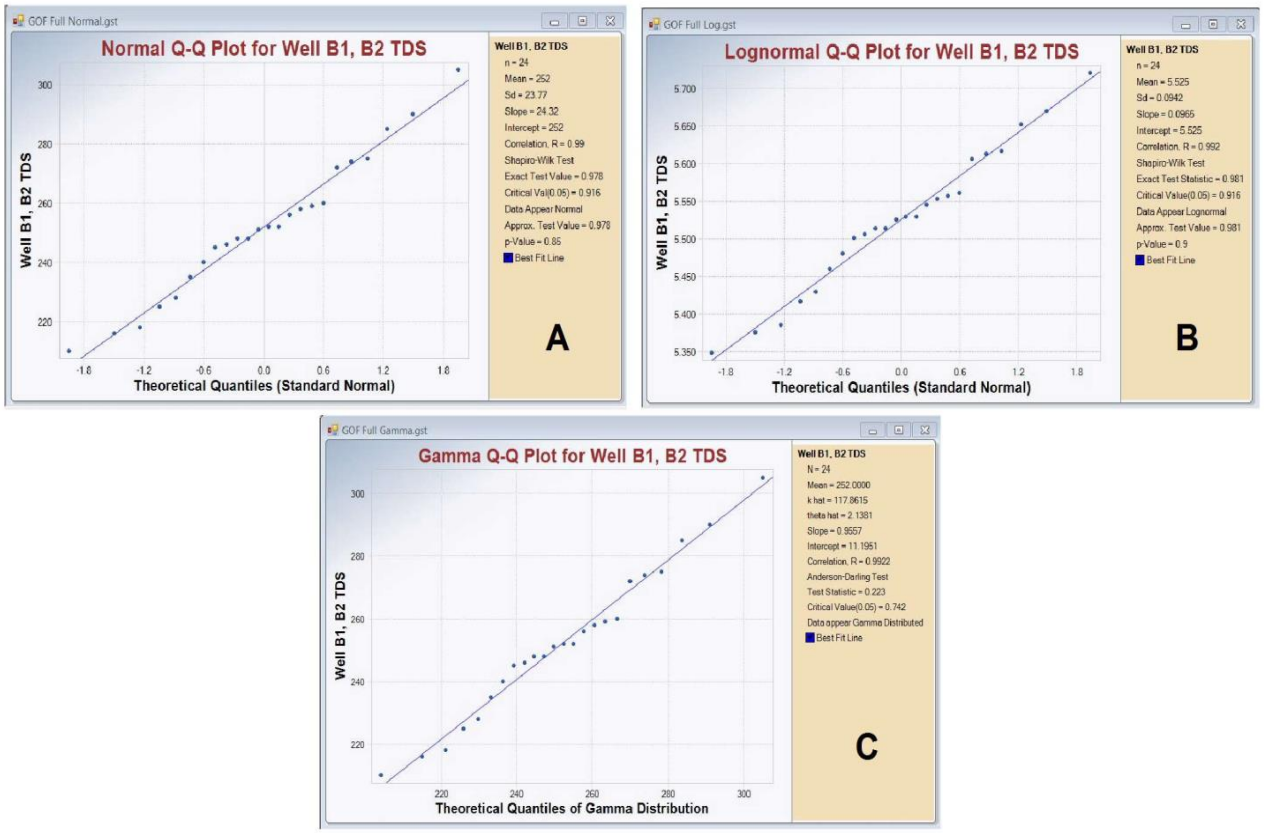


Figure D3. Goodness-of-fit results, in the form of Q-Q plots of the data represented in Figure D2 relative to (A) a normal distribution, (B) a lognormal distribution, and (C) a gamma distribution.

Table D4. Goodness-of-fit summary statistics for the Q-Q plots in Figure D3.

	A	B	C	D	E	F
24						
25				**Normal GOF Test Results**		
26						
27					Correlation Coefficient R	0.99
28					Shapiro Wilk Test Statistic	0.978
29					Shapiro Wilk Critical (0.05) Value	0.916
30					Approximate Shapiro Wilk P Value	0.85
31					Lilliefors Test Statistic	0.118
32					Lilliefors Critical (0.05) Value	0.181
33	**Data appear Normal at (0.05) Significance Level**					
34						
35				**Gamma GOF Test Results**		
36						
37					Correlation Coefficient R	0.992
38					A-D Test Statistic	0.223
39					A-D Critical (0.05) Value	0.742
40					K-S Test Statistic	0.105
41					K-S Critical(0.05) Value	0.177
42	**Data appear Gamma Distributed at (0.05) Significance Level**					
43						
44				**Lognormal GOF Test Results**		
45						
46					Correlation Coefficient R	0.992
47					Shapiro Wilk Test Statistic	0.981
48					Shapiro Wilk Critical (0.05) Value	0.916
49					Approximate Shapiro Wilk P Value	0.9
50					Lilliefors Test Statistic	0.108
51					Lilliefors Critical (0.05) Value	0.181
52	**Data appear Lognormal at (0.05) Significance Level**					

D.4. Calculation Example Using ProUCL 5.0: Strongly Skewed Data

A data set from ProUCL 5.0's library of examples[‡] is used to illustrate Figure D1's decision process with non-normal data. Figure D4 shows a histogram of the data values, indicating that the sample is strongly right-skewed and contains several extreme values well beyond the distribution's mode.

Figure D5 summarizes the Q-Q plots and goodness-of-fit results for three types of parametric distributions (normal, lognormal, and gamma) and Table D5 summarizes the goodness-of-fit statistics for all three parametric distribution options. As is visually apparent—and confirmed by hypothesis test results—these sample data cannot possibly represent a normal population at a confidence level of 95% (or even 90%).

[‡] The file "Ex-lognormal-Gamma.xlsx," is provided in ProUCL's installation directory under the \Data folder

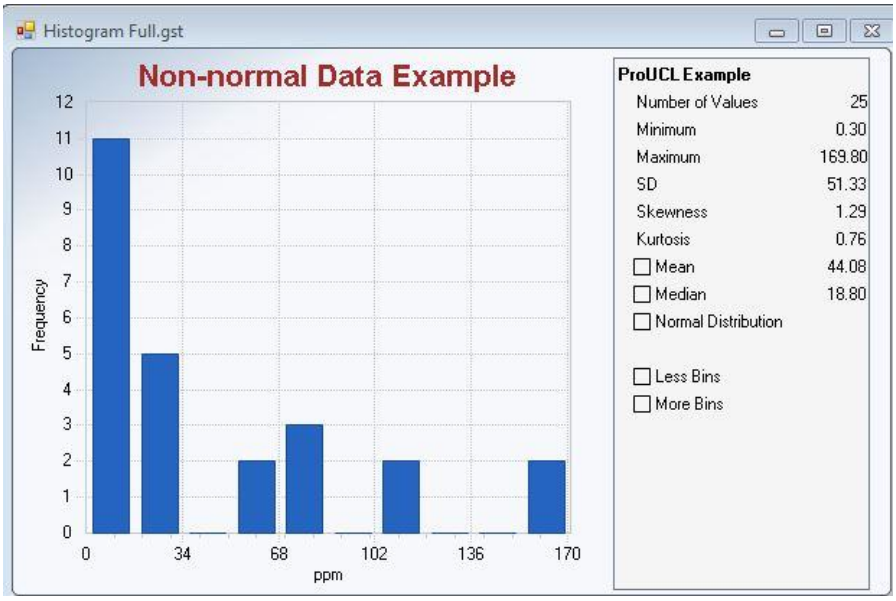


Figure D4. Data used to illustrate the decision process in Figure D1 when the sample is highly right-skewed.

Visual comparison of Figure D5(B) and D5(C) illustrates the effect that log-transformation has on the transformed data and the influence of extreme values. Relative to the gamma distribution's quantiles, the lognormal quantiles *bunch up* the sample quantiles along the regression line. The result is to weight extreme values much more heavily in a goodness-of-fit assessment of the lognormal distribution. This is, in turn, a direct consequence of the masking effect that log-transformation has on extreme values and its tendency to inflate the larger standard deviation of the transformed distribution thereby masking the presence of outliers and extreme values. For these reasons, DEQ follows EPA's lead in recommending that strongly skewed data be modeled with the gamma distribution wherever possible.

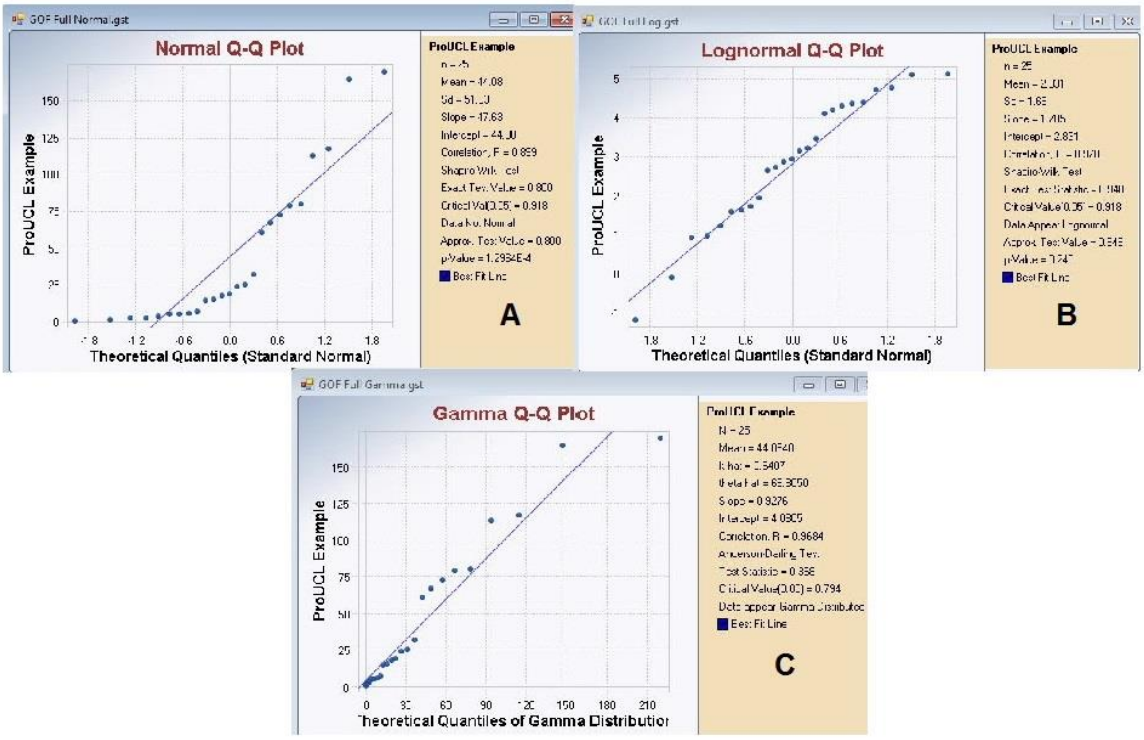


Figure D5. Graphical summaries of goodness-of-fit to (A) a normal distribution, (B) a lognormal distribution, and (C) a gamma distribution.

Table D5. Goodness-of-fit summary statistics for the Q-Q plots shown in Figure D5.

Normal GOF Test Results	
Correlation Coefficient R	0.899
Shapiro Wilk Test Statistic	0.8
Shapiro Wilk Critical (0.05) Value	0.918
Approximate Shapiro Wilk P Value	1.2964E-4
Lilliefors Test Statistic	0.245
Lilliefors Critical (0.05) Value	0.177
Data not Normal at (0.05) Significance Level	

Gamma GOF Test Results	
Correlation Coefficient R	0.968
A-D Test Statistic	0.368
A-D Critical (0.05) Value	0.794
K-S Test Statistic	0.113
K-S Critical(0.05) Value	0.183
Data appear Gamma Distributed at (0.05) Significance Level	

Lognormal GOF Test Results	
Correlation Coefficient R	0.978
Shapiro Wilk Test Statistic	0.948
Shapiro Wilk Critical (0.05) Value	0.918
Approximate Shapiro Wilk P Value	0.245
Lilliefors Test Statistic	0.135
Lilliefors Critical (0.05) Value	0.177
Data appear Lognormal at (0.05) Significance Level	

References

Aitchison, J. 1955. "On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin." *Journal of the American Statistical Association.* 50: 901–908.

Cohen, A.C. 1991. *Truncated and Censored Samples: Theory and Applications*. New York, NY: Marcel Dekker.

EPA (United States Environmental Protection Agency). 1988. *Statistical Methods for Evaluating the Attainment of Superfund Cleanup Standards, Volume 2: Groundwater*. Draft 2.0. Prepared by Westat Inc. under contract no. 68-01-7359.

EPA (United States Environmental Protection Agency). 1992. *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance*. Washington, DC: EPA. EPA/530-R-93-003.

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

EPA (United States Environmental Protection Agency). 2013a. *ProUCL 5.0 Software and User Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

EPA (United States Environmental Protection Agency). 2013b. *ProUCL 5.0 Technical Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

Fisher, S.R. and K.W. Potter. 1989. *Methods for Determining Compliance with Groundwater Quality Regulations at Waste Disposal Facilities*. Submitted to the Wisconsin Department of Natural Resources.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Helsel, D.R. and R.M. Hirsch. 1995. *Statistical Methods in Water Resources.* Studies in Environmental Science 49. New York, NY: Elsevier.

Royston J.P. 1993. "A Toolkit for Testing for Non-Normality in Complete and Censored Samples." *The Statistician*. 42(1): 37–43.

Shapiro, S.S. and M.B. Wilk. 1965. "An Analysis of Variance Test for Normality (complete samples)." *Biometrika*. 52: 591–611.

Appendix E. Seasonal Trends

E.1. Testing for Seasonality Using the Kruskal-Wallis Test

One of the important requirements for conducting statistical tests is temporal stationarity, specifically; do the data exhibit seasonal variations in concentration? Note that k, the number of seasons, is defined to be appropriate for the data being analyzed; for example, hourly stream temperature measurements might be grouped into two 12-hour *seasons* per day whereas monthly ground water measurements are usually grouped into four 3-month *seasons* For measurements collected quarterly over a multiyear period (each quarter tested in the same month), some of the variation in background ground water quality may be due to changing land uses (e.g., nearby agricultural activities and river and canal flows), which can obscure seasonal variations in water quality due to precipitation and evapotranspiration.

The Kruskal-Wallis test for seasonality is described below (Gilbert 1987; Helsel and Hirsch 1995). This test is considered a nonparametric test, which means that the underlying population distribution is not assumed. The Kruskal-Wallis test may be computed by an exact method used for small samples sizes (Lehmann 1998; Conover 1999), or by a large-sample or chi-square approximation (Helsel and Hirsch 1995) (Table E1). The null and alternative hypotheses are as follows:

H_0: All of the seasonally grouped subsets of data have identical distributions.

H_A: At least one group differs in its distribution.

In other words, do the measurements taken in one quarter of the year differ significantly from the measurements taken in any other quarter of the year?

To conduct the test, the data are ranked from smallest to largest, from 1 to N. If H_0 is true, the average rank for each of the k seasonal groups should be similar and also be close to the overall average of the N data. When H_A is true, the average rank for some of the groups will differ from others, reflecting the difference in magnitude of its observations. The test statistic, K, will equal 0 if all groups have identical average ranks and will be positive if group ranks are different. The distribution of K when H_0 is true is approximated by a chi-square distribution with k-1 degrees of freedom (df), where k is the number of seasons (Helsel and Hirsch 1995). For example, for quarterly data, k = 4 and df = 3.

All N observations are given a numerical rank from 1 to N, smallest to largest. When observations are tied, the average of their ranks is assigned to each (i.e., if observations 6 and 7 have the same value, assign 6.5 as the rank for both). These ranks, R_{ij}, are then used for computation of the test statistic. Within each group, the average group rank $\overline{R}_j$ is computed as follows:

$$\overline{R}_j = \frac{\sum_{i=1}^{n_j} R_{ij}}{n_j}$$

The average group rank, $\overline{R}_j$, is compared to the overall average rank, $\overline{R} = (N+1)/2$, squaring and weighting by sample size, to form the test statistic K:

$$K = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left[\overline{R}_j - \frac{N+1}{2} \right]^2$$

Reject H_0 if $K \geq \chi^2_{1-\alpha,(k-1)}$, the 1-α quantile of a chi-square distribution with k-1 degrees of freedom; otherwise do not reject H_0 (Gilbert 1987, Table A19).

For the minimum, the Idaho Department of Environmental Quality suggested data requirements (i.e., 3 years of quarterly data) N = 12, n_j = 3, and k = 4. Where, N is the number of samples, n_j is the size of the j^{th} group (years of seasonal data), and k = number of seasons. If the data cover only a partial year, the number of data points for one season may differ from those of another. Thus, n_j may differ depending on the season. The above equation reduces to the following:

$$K = \frac{1}{13} \sum_{j=1}^{4} 3 \left[\overline{R}_j - \frac{13}{2} \right]^2$$

Table E1. A portion of the quantiles of the chi-square distribution with k-1 degrees of freedom.

Degrees of Freedom (k-1)	Confidence Level	
	0.900	0.950
1	2.71	3.84
2	4.61	5.99
3	6.25	7.81
4	7.78	9.49
5	9.24	11.07
6	10.64	12.59
7	12.02	14.07
8	13.36	15.51
9	14.68	16.92
10	15.99	18.31
11	17.28	19.68

If the regulated entity discovers a seasonal trend to the collected water quality data, then this trend needs to be removed before continuing with further statistical testing. To remove the seasonal trend, apply the following calculations:

Step 1: Calculate the mean for all values from the same season in different years, $\overline{x}_k$.

Step 2: Calculate the universal mean for all values in the data set, $\overline{X}_N$.

Step 3: For each measurement, subtract the seasonal mean, $\overline{x}_k$ and add the universal mean, $\overline{X}_N$, to calculate the seasonally adjusted measurement. The seasonally adjusted values have lower overall variance due to removal of seasonal fluctuations.

E.2. Example

Table E2 summarizes the Kruskal-Wallis test calculations as applied to wells B1 and B2 in the example data set of Table B1. The total dissolved solids (TDS) values are ranked in ascending order from 1 to N = 12. The average quarterly ranks ($R_j(1)$, $R_j(2)$, $R_j(3)$, $R_j(4)$) and the test statistic, K, are calculated, and the resulting K statistic for each well is compared to the chi-square values from Table E1. In this case, the conclusion is that well B1 has statistically significant seasonal variability. The above three steps for removing the seasonal variability are applied to well B1's data and the resulting transformation is listed in the "Adjusted B1" column in Table E2. The result of the transformation is a data set with the same mean (254) but a significantly lower standard deviation (9.77 compared to 30.07).

Table E2. Testing for seasonality using the Kruskal-Wallis test.

	Well B1	Rank	Adjusted B1	Well B2	Rank
Year 1—1st quarter	305	12	266	252	7.5
Year 1—2nd quarter	228	3	264	251	6
Year 1—3rd quarter	258	7	252	245	3
Year 1—4th quarter	259	8	268	252	7.5
Year 2—1st quarter	285	10	246	260	10
Year 2—2nd quarter	210	1	246	248	5
Year 2—3rd quarter	274	9	268	275	12
Year 2—4th quarter	240	5	249	272	11
Year 3—1st quarter	290	11	251	256	9
Year 3—2nd quarter	216	2	252	246	4
Year 3—3rd quarter	248	6	242	218	1
Year 3—4th quarter	235	4	244	225	2
Determination of Seasonality					
Rj—1st quarter		11			8.8
Rj—2nd quarter		2			5
Rj—3rd quarter		7.3			5.3
Rj—4th quarter		5.7			6.8
K =		9.7			2.1
Critical statistic =		7.81			7.81
Seasonal variability?		Yes			No
Adjusting for Seasonality					
Mean_1	293.3				
Mean_2	218				
Mean_3	260				
Mean_4	244.7				
Mean_total	254		254		
Std_dev	30.07		9.77		

E.3. Comparing Seasonally Varying Populations

Where seasonality exists in background data, it is vital that the seasonal effect be removed from well data prior to estimating summary statistics, defining a decision threshold, or comparing downgradient data with background. For example, the standard deviation of well B1's raw TDS values is 30.0 compared to 9.8 after adjusting for seasonality (Table E2). Failure to account for seasonal variability can lead to statistical conclusions that are grossly inaccurate (Appendix F, section F.2 example).

References

Conover, W.L. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York, NY: John Wiley & Sons.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York, NY: Van Nostrand Reinhold.

Helsel, D.R. and R.M. Hirsch. 1995. *Statistical Methods in Water Resources.* Studies in Environmental Science 49. New York, NY: Elsevier.

Lehmann, E.L. 1998. *Nonparametrics, Statistical Methods Based on Ranks*. Oakland, CA: Holden-Day.

This page intentionally left blank for correct double-sided printing.

Appendix F. Secular Trends

F.1. Testing for Secular Trends Using the Mann-Kendall Test

One of the most important requirements when determining background ground water quality levels for the constituent(s) of concern in upgradient monitoring wells is deciding whether temporal stationarity (steady state) exists. If the system is not in a steady state condition, then background is undefined and setting a level for comparison is not statistically valid and can lead to erroneous results. In situations where a secular trend exists, a detrending procedure can be applied (Appendix L, section L.4) but only if detrending is both statistically and hydrogeologically justified.

There are several methods for determining whether the collected data show an increasing or decreasing trend through time. The United States Environmental Protection Agency (EPA 1988) suggests two methods. One, linear regression analysis, is somewhat simple to apply with commercial software, wherein the slope is calculated and tested for statistical significance. Though this method may be easy to perform, The Idaho Department of Environmental Quality (DEQ) does not recommend it. Linear regression is heavily influenced by outliers (Kimsey 1996) and also makes stronger assumptions about the distribution of the data (normality of residuals, constant variance, and linearity of the relationship) (Helsel and Hirsch 1995). Instead, DEQ recommends that the regulated entity use the Mann-Kendall test for trend to determine if a steady state condition exists within the data.

The Mann-Kendall test is a nonparametric alternative to regression. A major advantage is that no assumption of normality is required. In addition, the procedure is useful if there are missing data values (e.g., a quarterly sample was missed). Data reported as less than the detection limit are assigned a common value smaller than the smallest measured value, typically one-half of the detection limit. The actual value does not matter because the test only uses the relative magnitudes of the data rather than the specific data values (Gilbert 1987). The procedure outlined below is for cases when the number of collected background ground water quality data points is 40 or less (Gilbert 1987; Gibbons 1994). For situations where more than 40 data points are available, the regulated entity is referred to the literature (Mann 1945; Kendall 1975; Gilbert 1987).

Refer to Table F1 for the general procedure in setting up the test. First, order the data as shown by sampling date: $x_1, x_2, ..., x_N$ where x_i is the measured value for sampling date i. Second, record whether the difference $x_{i'} - x_i$ is positive or negative (where event i' follows i) for all possible pairs as well as the number of total positive and total negative differences in the data set.

This procedure is equivalent to defining

$$\operatorname{sgn}(x_{i'} - x_i) = \begin{cases} 1 & \text{if } x_{i'} - x_i > 0 \\ 0 & \text{if } x_{i'} - x_i = 0 \\ -1 & \text{if } x_{i'} - x_i < 0 \end{cases}$$

and computing the Mann-Kendall statistic as

$$S = \sum_{i=1}^{n-1} \sum_{i'=k+1} \text{sgn}(x_{i'} - x_i),$$

S is equal to the number of positive differences minus the number of negative differences in the bottom two right-most entries of Table F1. Conceptually speaking, if S is a large positive number, then more measurements taken later in time tend to be larger than those taken earlier. Similarly, if S is a large negative number, then more measurements taken later in time tend to be smaller.

The Mann-Kendall probability tables (Kendall 1975) are used to test the null hypothesis of no secular trend (statistically insignificant slope) versus the corresponding alternate hypothesis of a significant upward or downward slope. The tabulated probability corresponding to the absolute value of S is compared to the test's specified significance level (α); H_0 is rejected if the tabulated probability is less than α.

If the Mann-Kendall test shows that there is no statistically significant temporal trend in the water quality data, then statistical analysis and calculation of decision thresholds (Appendices H–K) may proceed. If a trend exists, then a detrending procedure such as described in Appendix L should be used to estimate an interim decision threshold.

Table F1. Mann-Kendall test set up.

Measurement Ordered by Time							
x_1	x_2	x_3	…	x_{N-1}	x_N	**Number of + differences**	**Number of – differences**
	$x_2 - x_1$	$x_3 - x_1$		$x_{N-1} - x_1$	$x_N - x_1$		
		$x_3 - x_2$		$x_{N-1} - x_2$	$x_N - x_2$		
				$x_{N-1} - x_3$	$x_N - x_3$		
				…	…		
				$x_{M-1} - x_{N-2}$	$x_N - x_{N-2}$		
					$x_N - x_{N-1}$		
						Total number of + differences	Total number of - differences

As noted in Appendix E, section E.3, deseasonalized data should be used when testing for a trend, because the Mann-Kendall results will be biased by seasonal variability (Appendix F, section F.2 example). An alternative to the Mann-Kendall test is the seasonal Kendall test, which estimates the trend by adjusting for seasonal variation. The test performs well when the product of the number of seasons and number of years is at least 25 (Helsel and Hirsch 1995). For example, if 3 or more years of independent monthly data or 7 years of quarterly data are available, the seasonal Kendall test can be used to detect a seasonally adjusted trend. The seasonal Kendall test is not available in ProUCL, however, so other software packages must be used (*http://www.gns.cri.nz/Home/Our-Science/Energy-Resources/Groundwater*; or *http://www.xlstat.com/en/learning-center/tutorials/identifying-a-trend-using-mann-kendall-test-with-xlstat-time.html)*. Alternatively, the data can be deseasonalized manually following steps 1–3 in Appendix E, section E.1 prior to applying the Mann-Kendall trend test.

F.2. Example Using ProUCL 5.0

The total dissolved solids (TDS) measurements from well B1 (Table B1) are used to illustrate the importance of deseasonalizing data prior to testing for a trend. For this example, we use a significance level of $\alpha = 0.05$. As noted in Appendix E, section E.3, erroneous summary statistics, statistical tests, and compliance decisions are possible when evaluating data with statistically significant seasonal variability. Figure F1 shows the results of applying ProUCL's Mann-Kendall trend test to the raw TDS data. Considerable variability is apparent and second-quarter values appear to be much lower than at values measured at other times of the year. Despite the visual indication of a possible trend, the Mann-Kendall test fails to reject the null hypothesis (H_0 = no trend) because of the large range in raw data values (TDS plot scale = 208 to 308).

Figure F2 summarizes the results of the Mann-Kendall test on deseasonalized TDS data (Appendix E, Table E3). In contrast to the raw data, the scale range is 241 to 269 parts per million, less than 30% of the raw data's scale range, reflecting the marked decrease in the standard deviation before and after deseasonalization ($\sigma_{\text{raw data}} = 30.0$; $\sigma_{\text{deasonalized}} = 9.8$, Appendix E).

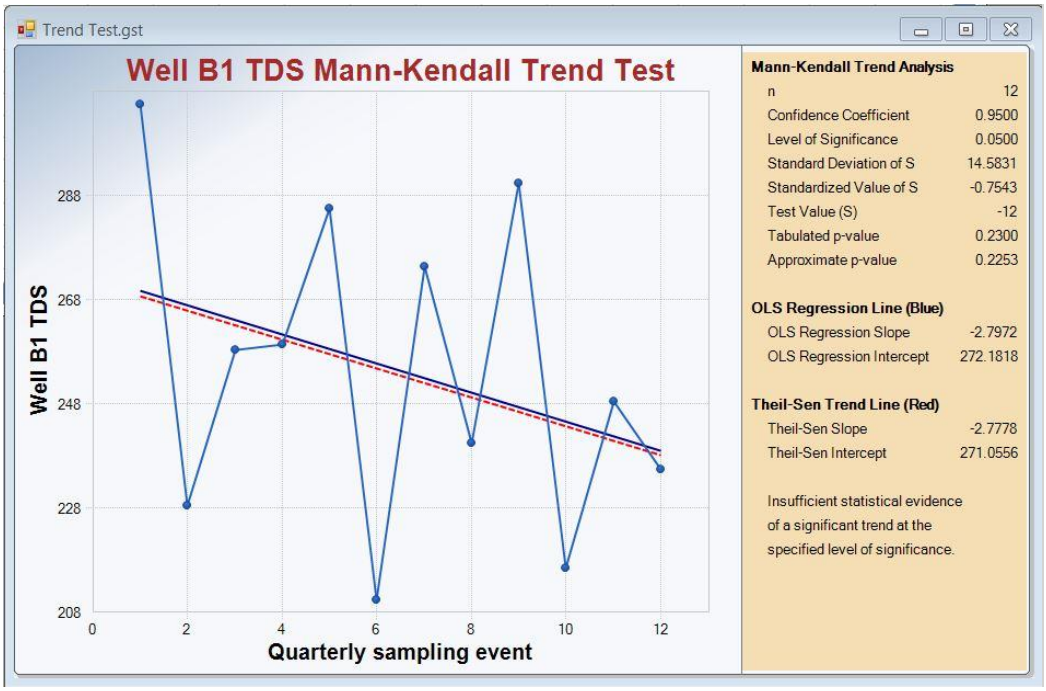


Figure F1. Results of applying a Mann-Kendall test to the raw TDS data from well B1 in Table B1. Sen's slope is shown in red and an ordinary least-squares linear regression slope, in black. H_0 is not rejected.

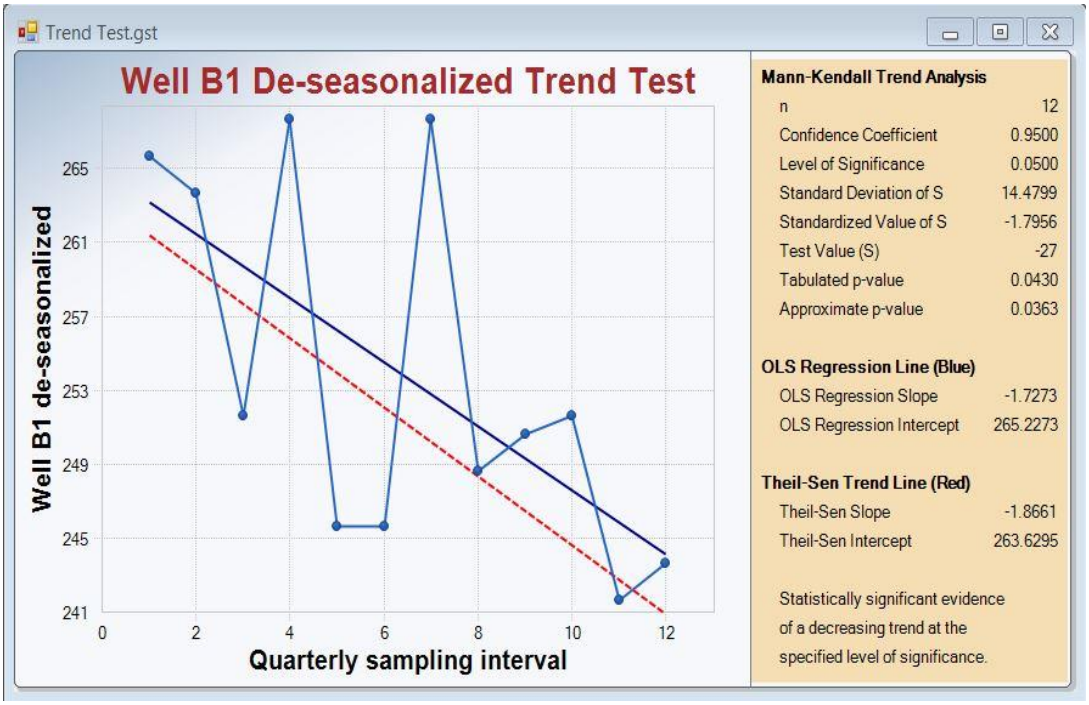


Figure F2. Results of applying the Mann-Kendall test to well B1's deseasonalized TDS data (Appendix E, Table E3). H_0 is rejected, and the correct conclusion that a trend does exist is achieved.

References

EPA (United States Environmental Protection Agency). 1988. *Statistical Methods for Evaluating the Attainment of Superfund Cleanup Standards, Volume 2: Groundwater*. Draft 2.0. Prepared by Westat Inc. under contract no. 68-01-7359.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York, NY: Van Nostrand Reinhold.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Helsel, D.R. and R.M. Hirsch. 1995. *Statistical Methods in Water Resources*. Studies in Environmental Science 49. New York, NY: Elsevier.

Kendall, M.G. 1975. *Rank Correlation Methods*. 4th ed. London: Charles Griffon.

Kimsey, M.B. 1996. *Implementation Guidance for the Ground Water Quality Standards*. Prepared for the Washington State Department of Ecology Water Quality Program, Olympia, WA.

Mann, H.B. 1945. "Nonparametric Tests Against Trend." *Econometrica*. 3: 245–259.

Appendix G. Data Pooling

G.1. Combining Data Sets for Normally Distributed Data

The advantage of combining background data from multiple upgradient monitoring wells is that, by increasing sample size, greater power can be realized for decision thresholds defined at a given confidence level. However, data sets can only be combined if it can be shown that they are statistically similar and the conceptual site model permits pooling of data from wells in different parts of the aquifer system as discussed in Section 3.2, "Evaluation of Hydrogeologic Data." Any differences in statistical subpopulations of ground water quality occurring at different depths and hydrogeologic units must be respected. Above all, water quality data representing different subpopulations must not be pooled. In such cases, background water quality must be defined separately for different subpopulations and comparisons of downgradient water quality with background must be conducted in a manner that is hydrogeologically defensible in the context of the site conceptual model.

The Idaho Department of Environmental Quality (DEQ) recommends that the data sets from two wells first be tested for similar variance using the F-test; then, if their variances are similar, they can be tested for similar means using the t-test (Larsen and Marx 1986). For two independent, normally distributed random samples (X and Y) having means of $\bar{x}$ and $\bar{y}$ and variances of s_x^2 and s_y^2, respectively, then the null hypothesis H_0 for the F-test is that $s_x^2 = s_y^2$. H_0 can be rejected at the α level of significance if

$$\frac{s_y^2}{s_x^2} \text{ is either } \begin{cases} \leq F_{\alpha/2,m-1,n-1} \; or \\ \geq F_{1-\alpha/2,m-1,n-1} \end{cases}$$

where m and n are the sample size for each data set. For the case of α = 95% and m and n, both equal to 12 data points (DEQ's minimum recommended data requirement), H_0 can be rejected if s_y^2/s_x^2 is either ≤ 0.29 or ≥ 3.48.

If H_0 is rejected, then the data sets cannot be combined. If H_0 cannot be rejected, then the data sets can be tested using the two-sample t-test. For two independent, normally distributed random samples (X and Y) having means of $\bar{x}$ and $\bar{y}$ and statistically equal variances of s_x^2 and s_y^2, then H_0 for the t-test is $\mu_x = \mu_y$. H_0 can be rejected at the α level of significance if

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \text{ is either } \begin{cases} \leq -t_{\alpha/2,n+m-2} \\ or \\ \geq +t_{\alpha/2,n+m-2} \end{cases}$$

where

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

For the case of α = 95% and m and n, both equal to 12, so H_0 can be rejected if t is either ≤ -2.07 or ≥ 2.07. If the null hypothesis is rejected, then the data sets cannot be combined.

G.2. Example

Appendix G uses deseasonalized well B1 data. Having adjusted background well B1 for seasonal affects and having found that both data sets (well B1 and well B2) are normally distributed (Appendix D), the question arises whether the two data sets can be combined into a single background data set. Looking at the descriptive statistics, the means are similar (254 parts per million [ppm] for well B1 and 250 ppm for well B2), and the standard deviations are similar (9.77 ppm for well B1 and 16.4 ppm for well B2), so the implication is that the data could be combined. To check if the variances are statistically similar, the F-test is conducted first.

$$\frac{s_2^2}{s_1^2} = \frac{16.4^2}{9.77^2} = 2.818$$

For the case of α = 95% and m and n, both equal to 12 data points, so s_y^2/s_x^2 must be ≤ 0.283 or ≥ 3.66 in order to reject H_0. In this case, H_0 cannot be rejected.

Next, the t-test is applied to the two data sets.

$$s_p^2 = \frac{(12-1)9.77^2 + (12-1)16.4^2}{12+12-2} = 182.2$$

$$t = \frac{254-250}{13.5\sqrt{\frac{1}{12}+\frac{1}{12}}} = 0.726$$

Since the test statistic (0.726) is neither ≤ -2.0739 nor ≥ 2.0739, H_0 cannot be rejected. Therefore, the total dissolved solids (TDS) data sets can be combined for the purposes of setting a background ground water quality level.

G.3. Combining Well Data Sets for Lognormally Distributed Data

The process for determining whether background data can be combined for lognormally distributed data is the same as that for normal data, except that the calculations of the mean and standard deviation differ. The following equations can be used to calculate the arithmetic mean and standard deviation before applying the F-test and t-test (Gilbert 1987).

$$\bar{x}_{ln} = \frac{1}{N}\sum_{i=1}^{N} \ln(x_i)$$

$$s_{ln}^2 = \frac{1}{N}\sum_{i=1}^{N} \left[\ln(x_i) - \bar{x}_{ln}\right]^2$$

where N is the size of the sample.

The decision to pool data is made using transformed data. Interpretation of the mean and standard deviation in original scale units can be obtained using the following equations:

$$\bar{x} = \exp\left(x_{ln} + s_{ln}^2 / 2 \right)$$

$$s = \sqrt{\exp\left(2x_{ln} + s_{ln}^2\right)\left[\exp\left(s_{ln}^2\right) - 1\right]}$$

G.4. Combining Well Data Sets for More Than Two Wells or Nonparametrically Distributed Data

Levene's test (Levene 1960) can be used to check the assumption of homogeneity of variance if there are more than two wells to be pooled. The formula is

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i \left(\overline{Z}_{i} - \overline{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i} \left(\overline{Z}_{ij} - \overline{Z}_{i.}\right)^2}$$

where $Z_{ij} = \left|Y_{ij} - mean_i\right|$. The group means of the Z_{ij} are $\overline{Z}_{i.}$ and the overall mean of Z_{ij} is $\overline{Z}_{..}$. N is the overall sample size; k is the number of subgroups (i.e., number of background wells to be pooled); and N_i is the sample size for the i^{th} subgroup. If the data are normally distributed, all calculations are based on the original scale units. If the data are lognormal, all calculations are based on logtransformed data.

As with normal and lognormal data, nonparametric data sets need to be checked for homogeneity of variance before a comparison of medians can be performed. Levene's test can be extended (Brown and Forsythe 1974) for working with the medians of the data sets. In the above formula, $Z_{ij} = \left|Y_{ij} - median_i\right|$ and other terms remain the same.

Levene's test rejects H_0 ($s_x^2 = s_y^2$) if $W > F_{(\alpha, k-1, N-k)}$ where $F_{(\alpha, k-1, N-k)}$ is the *upper critical value* of the *F distribution* with *k* - 1 and *N* - *k* degrees of freedom at a significance level of α.

If H_0 is rejected, then the data sets cannot be combined. If H_0 cannot be rejected, then the Kruskal-Wallis test can be applied to determine if the subgroup medians are statistically similar and the k data sets can be combined. The Kruskal-Wallis test is described in Appendix E.

The Shapiro-Wilk test for normality (Appendix D) can be used to determine if the combined data set is normally or lognormally[§] distributed; ProUCL 5.0 uses the Anderson-Darling and

[§] by transforming the data to their logarithmic equivalents

Kolmogorov-Smirnov tests to determine if the data are gamma distributed. If the data are parametrically distributed, then a parametric decision threshold can be calculated (Appendix H or J). If the data prove to be non-normal and nongamma, then the data should be logarithmically transformed and the Shapiro-Wilk test repeated to determine if a lognormal distribution is an option. If the data are found to be nonparametrically distributed, then a nonparametric decision threshold can be calculated (Appendix I or K).

References

Brown, M.B. and A.B. Forsythe. 1974. "Robust Tests for Equality of Variances." *Journal of the American Statistical Association*. 69: 364–367.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York, NY: Van Nostrand Reinhold.

Larsen, R.J. and M.L. Marx. 1986. *An Introduction to Mathematical Statistics and Its Applications*. 2nd ed. Upper Saddle River: NJ: Prentice-Hall.

Levene, H. 1960. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. I. Olkin et al. eds. Stanford, CA: Stanford University Press.

Appendix H. Parametric Upper Tolerance Limits

H.1. Parametric Upper Tolerance Limit as a Decision Threshold

This section of the guidance assumes that the regulated entity has a background data set that has been found to be in a steady state, to have no statistically significant seasonal effects or been corrected for seasonality, and to meet parametric distribution assumptions. In most cases, this procedure will be applied at a new site where future water quality measurements will be compared to background ground water quality measurements in the same wells (an intrawell analysis; Gibbons 1994). The intrawell upper tolerance limit (UTL) is used to set a background water quality threshold in a downgradient well for a constituent of concern using only data from that monitoring well.**

When monitoring ground water quality, the compliance point samples are assumed to come from the same population as the background values until significant evidence of contamination can be shown (EPA 1992). Once the UTL is set, each future compliance sample is compared to the UTL (Fisher and Potter 1989; Gibbons 1994; Kimsey 1996). To minimize the false negative rate and reduce the need for verification resampling, a specified exceedance rate (*coverage*) is tolerated, for example, no more than 1 exceedance in the next 20 comparisons (95% coverage). If this rate is exceeded, then significant evidence of contamination is indicated. In setting compliance limits, the Idaho Department of Environmental Quality (DEQ) suggests that UTLs be set so that 95% of the tested samples (coverage) are below the limit with 95% confidence. Therefore, in this discussion, the compliance standards will be calculated for 95% confidence and 95% coverage.

UTLs define a range within which some proportion of the population (the coverage; in this case, 95%) will fall some proportion (in this case 95%) of the time. The limit is calculated using the following formula:

$$\mathrm{UTL} = \bar{\mathrm{x}} + K\mathrm{s}$$

where UTL is the upper tolerance limit, $\bar{\mathrm{x}}$ is arithmetic mean of the background data, s is the arithmetic standard deviation of the data, and *K* is a constant that changes depending on the proportions used (Gibbons[1994] provides the complete mathematical formulation). Table H1 provides examples of *K* factors for 95% coverage and 95% confidence. The process for setting tolerance intervals for lognormally distributed data is the same as that for normal data, except that the UTL is set for and compared to the log-transformed data values. For gamma-distributed data, DEQ recommends using software that is capable of estimating decision thresholds for a gamma population distribution (e.g., ProUCL 5.0).

** A UTL can also be defined for interwell data comparisons by defining background water quality in an ensemble of upgradient wells and a UTL that is calculated from those data; the UTL is then applied to downgradient wells to determine compliance (EPA 2009).

Table H1. Partial table of factors (*K*) for constructing one-sided normal upper tolerance limits at 95% confidence and 95% coverage.

Sample Size	95% Coverage	Sample Size	95% Coverage
8	3.188	16	2.523
9	3.032	17	2.486
10	2.911	18	2.453
11	2.815	19	2.423
12	2.736	20	2.396
13	2.670	25	2.292
14	2.614	30	2.220
15	2.566	35	2.166

Sources: Complete tables in Gibbons (1994) and Guttman (1970).

H.2. Example

As was shown in Appendix G, the sample total dissolved solids (TDS) data for background well B1 and background well B2 could be combined into a single data set of 24 data points. The resulting data set is normally distributed and has a mean of 252 parts per million (ppm) and a standard deviation of 13.34 ppm. The appropriate K factor, therefore, as interpolated from Table H1, is 2.313 and the 95% UTL is

$$\text{UTL} = 252 + 2.313 * 13.34 = 282.8\,\text{ppm}$$

As long as no more than 5% of future TDS measurements exceed this threshold, the constituent of concern is deemed not to be affected by the facility's operation. Should a future exceedance of the UTL violate the 95% coverage criterion (e.g., six exceedances in 100 future sampling events), then water quality would be deemed to be degraded. As Gibbons (1994) points out, the UTL's specification of a coverage makes verification resampling unnecessary because a specified number of exceedances are expected and give the method its power without a verification requirement.

Appendix K provides an example illustrating how decision thresholds are applied and interpreted using ProUCL 5.0's output.

References

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

EPA (United States Environmental Protection Agency). 1992. *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance*. Washington, DC: EPA. EPA/530-R-93-003.

Fisher, S.R. and K.W. Potter. 1989. *Methods for Determining Compliance with Groundwater Quality Regulations at Waste Disposal Facilities*. Submitted to the Wisconsin Department of Natural Resources.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Guttman, I. 1970. *Statistical Tolerance Regions: Classical and Bayesian*. Darien, CT: Hafner Publishing.

Kimsey, M.B. 1996. *Implementation Guidance for the Ground Water Quality Standards*. Prepared for the Washington State Department of Ecology Water Quality Program, Olympia, WA.

This page intentionally left blank for correct double-sided printing.

Appendix I. Nonparametric Upper Tolerance Limits

I.1. Nonparametric Upper Tolerance Limit as a Decision Threshold

For background data sets that are not parametrically distributed but meet the steady state condition and have been corrected for seasonal effects, a nonparametric upper tolerance limit (UTL) can be used to set a decision threshold for the constituents of concern. In most cases, this method will be applied to a new site where future water quality measurements will be compared to background ground water quality measurements in the same well (an intrawell analysis; Gibbons 1994). Table I1 shows the background sample sizes (N) required to achieve the desired coverage at varying confidence levels. The UTL is set equal to the highest value in the background data set using N seasonally adjusted background sample values.

For example, to be 85% confident (column 1, row 5) that 90% of future comparisons (column 7) will fall below the UTL, then the highest background data value of the most recent 19 measurements is used as the limit. For 95% confidence and 95% coverage, background sample size must be at least 59.

Table I1. Sample sizes for nonparametric upper tolerance limits.[a]

1-α	q=0.500	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.980	0.990
0.700	2	4	5	6	8	12	24	48	60	120
0.750	2	4	5	7	9	14	28	55	69	138
0.800	3	5	6	8	10	16	32	64	80	161
0.850	3	6	7	9	12	19	37	75	94	189
0.900	4	7	9	11	15	22	45	91	144	230
0.950	5	9	11	14	19	29	59	119	149	299

a. q is the proportion of the population that is covered by the tolerance interval. The tabulated quantity, X, is the required sample size (N or greater) to ensure that the probability of at least q of all future samples is greater than or equal to 1 – α (Conover 1999, Table A5).

I.2. Example

Suppose that the total dissolved solids (TDS) data in the example data set had been nonparametrically distributed. In that case, with a pooled background sample size of 24, Table I1 would indicate that 95% coverage could be obtained at no better than a 70% confidence level (first row, eighth column). Where the sample size required to achieve a specified confidence level and coverage is less than the available background sample size, set the UTL equal to the highest value of the N most recent background values. To achieve 90% coverage at a 90% confidence, the UTL is equal to the highest of the most recent 22 seasonally corrected TDS values in Table B1 (268 parts per million). The new observation should be seasonally adjusted when comparing to the UTL.

Appendix K provides an example illustrating how decision thresholds are applied and interpreted using ProUCL 5.0's output.

References

Conover, W.L. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York, NY: John Wiley & Sons.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Appendix J. Parametric Upper Prediction Limits

J.1. Parametric Upper Prediction Limit as a Decision Threshold

This section of the guidance pertains to data sets that have been corrected for seasonality and show no statistically significant secular trend, and are parametrically distributed. Downgradient ground water quality will be compared to the limit to detect degradation. A parametric upper prediction limit (UPL) will be set on the basis of upgradient water quality data (from either pooled or individual wells). As in Appendix H, the arithmetic mean and standard deviation for lognormal distributions must be correctly calculated (Appendix G, section G.3) and the UPL defined for logarithmically transformed data.

In interwell comparison, future measurements in multiple downgradient wells are compared to an UPL based on background ground water quality data from upgradient wells (possibly pooled; Appendix G). To reduce the sitewide false positive rate, any exceedance encountered is verified through resampling (Gibbons 1994, sections 1.5 and 1.6). Within the context of this guidance, a verification sample is a sample collected to provide statistical verification of a possible exceedance (Section 4.5 "Verification Resampling"). It is <u>not</u> a measurement confirmation (i.e., duplicate) sample collected to document analytical precision. Therefore, verification samples must be temporally independent of the initial sample and of each other; in other words, sufficient time must elapse between the initial sampling and both resampling events to ensure statistical independence of all three measurements.

The method for verification resampling adopted by the Idaho Department of Environmental Quality (DEQ) for wastewater reuse permits and Resource Conservation and Recovery Act facilities is to allow the facility to take two subsequent verification samples. If one or both verification samples exceed the UPL, then the initial exceedance is confirmed. This verification resampling scheme is referred to as "one of three samples in bounds" (Gibbons 1994; EPA 2009) and has the lowest false negative rate and highest power of the three verification schemes discussed by Gibbons. Table J1 gives multiplication factors (*K*) for the UPL formula:

$$\mathrm{UPL} = \bar{x} + Ks$$

where $\bar{x}$ is the mean of N upgradient background measurements and s is their standard deviation. Note that the *K* factor depends only on the number of future comparisons. In general, the larger the number of future comparisons, *k*, the higher the *K* factor and the UPL; conversely, the larger the background sample size, N, the lower the *K* factor and UPL. The number of future comparisons, *k*, is defined as:

k = [number of measurements to be collected per well per year]

× [number of years] × [number of wells] × [number of COCs]

For example, for a downgradient well that will be sampled four times a year for 5 years (the permit reapplication period) and analyzed for two constituents of concern, $k = 1 \times 4 \times 5 \times 2 = 40$ comparisons.

If the upgradient, background well data cannot be pooled, then a separate UPL will be determined for each upgradient well and decisions made as to which downgradient wells will be

compared to which upgradient wells. These decisions will affect the value of N and *k* that are applicable for different downgradient wells because the upgradient background data sets will differ. Note that the *K* factors in Table J1 increase only slightly for $k > 50$; if *K* factors are required for higher *k*, they can be estimated by extrapolation.

Table J1. *K* Factors at $\alpha = 0.05$ for a verification protocol where one or both verification resamples must confirm the initial exceedance.

	***k* = Number of Future Comparisons**				
N	**10**	**20**	**30**	**40**	**50**
4	2.02	2.42	2.65	2.82	2.94
8	1.37	1.61	1.75	1.84	1.92
12	1.21	**1.42**	1.54	**1.62**	1.68
16	1.14	1.33	1.44	1.52	1.58
20	1.10	1.28	1.39	1.46	1.51
24	1.08	**1.25**	1.35	**1.42**	1.47
36	1.03	**1.20**	1.29	**1.36**	1.41
48	1.01	**1.17**	1.27	**1.33**	1.38

Source: Gibbons (1994), Table 1.6, as prepared by Charles Davis based on results in Davis and McNichols (1987).
Notes: Bold values indicate K factors that would apply to 5 years of quarterly future measurements (e.g., 20 comparisons of one constituent of concern [COC] for one well; 40 for two wells or for two COCs at one well) and for various background sample sizes, each representing 3 years of quarterly data collected at 1, 2, 3, or 4 background wells.

J.2. Example

The pooled total dissolved solids (TDS) background data set from wells B1 and B2 has an overall N, $\bar{x}$, and s of 24, 252, and 23.8, respectively. For a 5-year permit reapplication period, with quarterly samples to be collected at two downgradient wells and a single constituent of concern (TDS), *k* is 5*4*2=40 and the K-factor read from Table J1 is 1.42. The UPL is therefore

$$\text{UPL} = 252 \text{ ppm} + 1.42\ (13.34) = 271 \text{ ppm}.$$

Future measurements from the downgradient well will be compared to this UPL over the monitoring period. Every exceedance triggers a verification resampling event using the protocol specified in Table J1. At each permit reapplication, the UPL will be reevaluated using all available data, including the new upgradient data collected since last application as well as previous data and verification samples. If the data are seasonally adjusted, then every new observation should be adjusted before comparing to the UPL.

Appendix K provides an example illustrating how decision thresholds are applied and interpreted using ProUCL 5.0's output.

References

Davis, C.B. and R.J. McNichols. 1987. "One-sided Intervals for at Least *p* of *m* Observations from a Normal Population on Each of *r* Future Occasions." *Technometrics*. 29: 359–370.

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

This page intentionally left blank for correct double-sided printing.

Appendix K. Nonparametric Upper Prediction Limits

K.1. Nonparametric Upper Prediction Limit as a Decision Threshold

For data sets that are not parametrically distributed but are steady state, independent, and corrected for seasonal effects, a nonparametric prediction limit can be used to set background levels. Like nonparametric tolerance limits, nonparametric prediction limits require larger background sample size to provide high levels of confidence. The nonparametric upper prediction limit (UPL) is set equal to the maximum value out of N independent background samples required to achieve a specified confidence level for a specified number of future comparisons. Confidence level is a function of N, the resampling plan used, and the number of future comparisons *k*. For large *k* or small α, a large number of background samples is required (Gibbons 1994).

One assumption inherent in this procedure is that the downgradient monitoring well's water quality data represents the same population as the upgradient (background) well(s) to which it is compared. The interwell analysis method could be applied even at new facilities in situations where the data from downgradient wells are insufficient to justify an intrawell analysis.

As is the case with parametric UPLs, upgradient background data from multiple wells can only be pooled if the means and standard deviations of each upgradient well's data set are statistically indistinguishable. For nonparametric data, the statistical test to check for statistical differences of the variances and medians is Levene's test and the Kruskal-Wallis test, respectively (Appendix G).

Table K1 (Gibbons 1994, chapter 2) summarizes confidence levels for various background sample sizes and future comparisons and the same verification sampling scheme discussed in Appendix J (take two verification samples, if one or both also exceed the UPL then exceedance is verified). Table K1 shows how, as background sample size increases, so does the confidence level; conversely, confidence level decreases as the number of future comparisons (*k*) increases.

Table K1. Confidence levels for a nonparametric prediction limit where exceedance is verified when one or both verification resamples also exceed the limit.

	***k* = Number of Future Comparisons**							
N	**10**	**20**	**30**	**40**	**50**	**60**	**80**	**100**
4	.5585	.4393	.3759	.3347	.3050	.2822	.2491	.2257
8	.7616	.6522	.5836	.5348	.4976	.4678	.4225	.3890
12	.8538	.7676	.7072	.6613	.6246	.5942	.5463	.5095
16	.9023	.8356	.7852	.7449	.7115	.6831	.6368	.6001
20	.9305	.8785	.8369	.8024	.7729	.7473	.7044	.6695
25	.9516	.9126	.8798	.8516	.8268	.8047	.7668	.7350
35	.9729	.9492	.9279	.9087	.8912	.8751	.8463	.8211
50	.9858	.9727	.9604	.9488	.9379	.9275	.9083	.8908

Source: Gibbons (1994), Table 2.13, after Charles Davis based on results in Davis and McNichols (1993)

K.2. Example

The pooled sample size for background total dissolved solids (TDS) data from upgradient wells B1 and B2 is N = 24. For a single constituent of concern and a 5-year permit reapplication period with quarterly samples to be collected at two downgradient wells, k is 40. The UPL is set equal to the highest value of the N seasonally adjusted background measurements (275 parts per million [ppm], Table E2). Table K1 indicates that future comparisons to this UPL will be at a confidence level no higher than about an 85%. If a higher confidence level were desired, then fewer future comparisons would have to be specified; for example, to achieve a 90% confidence with N = 24, k would have to be limited to about 24 comparisons (interpolating the sixth row of Table K1). This would correspond to a 3-year comparison period in the above example.

Where the sample size required to achieve a specified confidence level and coverage is less than the available background, set the UPL equal to the highest value of the N most recent background values. To achieve a 90% confidence level with k = 20 future comparisons, as in the above example, interpolation within Table K1 indicates that N = 23. Therefore, set the UPL equal to the highest of the 23 most recent background sample values.

K.3. Example Using ProUCL 5.0: Weakly Skewed Data

In this example, ProUCL output is used to show how the various parametric and nonparametric decision thresholds calculated by the program are interpreted (and misinterpreted). The histogram of seasonally adjusted and pooled TDS data (Appendix G, section G.2), representing a statistically homogeneous background population, is shown in Figure K1.

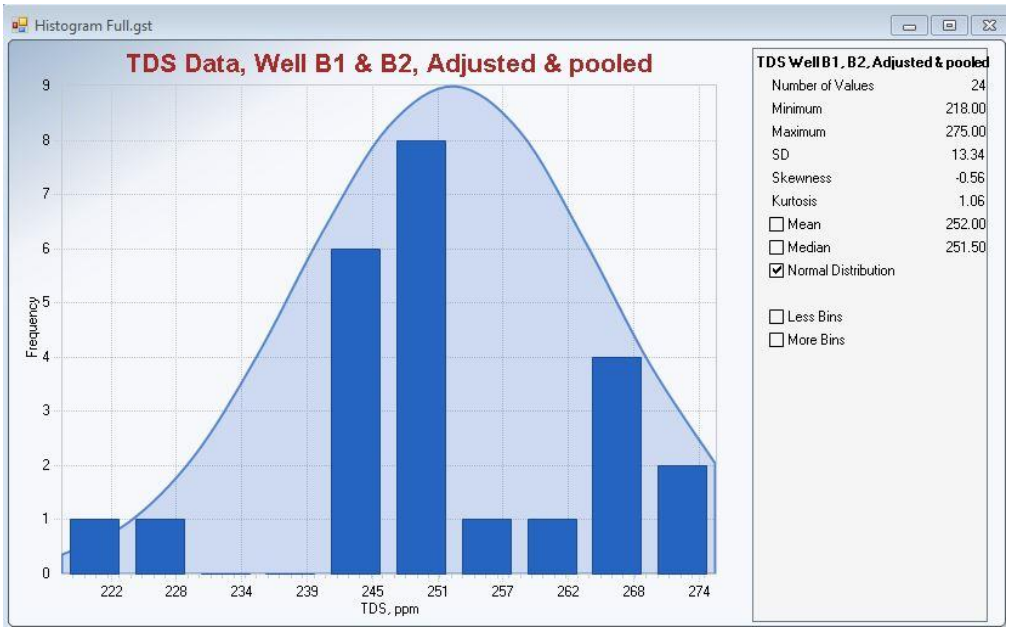


Figure K1. Histogram of seasonally adjusted and pooled TDS data for wells B1 and B2.

With the seasonal trend removed, the sample data set has a substantially smaller data range and standard deviation (s = 30) compared to the raw data (s = 13.34; Appendix D, Figure D2). Hypothesis tests to determine the form of the distribution, however, lead to the same conclusion: at 95% confidence the sample data can be fit to a normal, lognormal, or gamma distribution.

Following the decision logic of Figure D1, however, the data will be assumed to represent a normally distributed population.

The corresponding decision thresholds were calculated using the "Normal" option in ProUCL's "Upper Limits/BTVs" menu. In the Options dialog window, the number of "Future k Observations" was set to 5 and the confidence level and coverage were left at their default settings of 95%. A portion of the output file, shown in Table K2, displays a variety of information, but only the bottom (highlighted) portion is relevant to this discussion of decision thresholds.

Table K2. ProUCL output showing calculated decision limits for a normal distribution.

General Statistics			
Total Number of Observations	24	Number of Distinct Observations	17
Minimum	218	First Quartile	246
Second Largest	272	Median	251.5
Maximum	275	Third Quartile	261
Mean	252	SD	13.34
Coefficient of Variation	0.0529	Skewness	-0.563
Mean of logged Data	5.528	SD of logged Data	0.0539
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.309	d2max (for USL)	2.644
Normal GOF Test			
Shapiro Wilk Test Statistic	0.931	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.916	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.167	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.181	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Background Statistics Assuming Normal Distribution			
95% UTL with 95% Coverage	282.8	90% Percentile (z)	269.1
95% UPL (t)	275.3	95% Percentile (z)	273.9
95% UPL for Next 5 Observations	286	99% Percentile (z)	283
95% UPL for Mean of 5 Observations	263.2	95% USL	287.3

The output generated by the program reiterates the results of a normal hypothesis test on the sample data and confirms that they represent a normally distributed population at 95% confidence (5% significance). The last four lines of the table summarize three different decision thresholds calculated from the background statistics that the Idaho Department of Environmental Quality (DEQ) recommends can be used in compliance decisions, where appropriate:

Upper tolerance limit (UTL) at 95% confidence level, with 95% coverage = 282.8 ppm
UPL at 95% confidence and 5 future observations = 286.0 ppm
Upper simultaneous limit (USL) at 95% confidence and all future observations = 287.3 ppm

Consistent with the definitions and compliance goals of these thresholds (Section 4, "Statistical Determination of Water Quality Degradation"), the UTL is less than the UPL which is less than the USL. This is typical for *well-behaved* symmetrical data sets. The UTL is the lowest because it has a built-in tolerance for a specified percentage of exceedances that are to be expected to occur in all future samples (5%, in this example). The USL is the highest threshold because it does not allow for *any* exceedance in any future sampling event. Unlike the UTL and USL, a UPL applies to the next specified number (k) of future sampling events in which no exceedance is expected. In this case, the choice of k = 5 represents a tradeoff between a much lower UPL for

k = 1 (Table K2, third line from bottom) and a higher threshold for larger k values. In this case, therefore, the UPL's intermediate threshold represents a balance between the compliance goals of the UTL and USL.

The importance of choosing the correct distribution for calculating a decision threshold can be illustrated by incorrectly assuming that this data set is nonparametric. Under that stipulation, the nonparametric UTL, UPL, and USL thresholds are 274.9, 274.3, and 275, respectively (Table K3), which are significantly lower than the thresholds for a normal population. If used to assess future compliance, the lower thresholds could lead to an unacceptable number of false positives (incorrect determinations of noncompliance) and increased monitoring costs.

Table K3 summarizes the UTL, UPL, and USL values that were calculated from the same background data set, considering a gamma and lognormal distribution (which are equally good statistical representations of the sample data) as well as a nonparametric distribution and the normal distribution option shown in Table K2. The similarity of the decision threshold magnitudes for this data set indicates that the thresholds are insensitive to the choice of population distribution, a situation that lends confidence to the outcome of any statistical analysis.

Table K3. Decision thresholds calculated from the same background data for different assumed population distributions. Where ProUCL uses two or more estimation methods, the estimates are averaged in the right-most column.

Data appear Normally Distributed at 5% Significance Level	Estimates	Averages
95% UTL with 95% Coverage		282.8
95% UPL for Next 5 Observations		286.0
95% USL		287.3
Data appear Gamma Distributed at 5% Significance Level		
95% WH Approx. Gamma UTL with 95% Coverage	284.2	284.3
95% HW Approx. Gamma UTL with 95% Coverage	284.4	
95% WH UPL for k = 5	287.8	287.9
95% HW UPL for k = 5	288.0	
95% WH USL	289.1	289.3
95% HW USL	289.4	
Data appear Lognormally Distributed at 5% Significance Level		
95% UTL with 95% Coverage		285.0
95% UPL for Next 5 Observations		288.7
95% USL		290.2
For Nonparametric Background Statistics		
95% UTL with 95% Coverage	275.0	274.9
95% Percentile Bootstrap UTL with 95% Coverage	275.0	
95% BCA Bootstrap UTL with 95% Coverage	274.6	
95% UPL		274.3
95% USL		275.0
WH = Wilson-Hilferty approximation		
HW = Hawkins-Wixley approximation		

K.4. Example Using ProUCL 5.0: Strongly Skewed Data

In contrast to the preceding example, the choice of distribution for highly skewed data is a critical consideration in deriving statistically defensible decision thresholds. The data set evaluated in Appendix D, section D.4 illustrates the impact that an inappropriate choice can have when calculating decision thresholds from highly skewed data. That data set is markedly non-normal, but the goodness-of-fit test results indicate that the data fit both a gamma and a lognormal distribution at a 95% confidence level.

Table K4 summarizes the calculated decision thresholds for the gamma and lognormal options. In contrast to the outcome in Table K3, the calculations for the skewed data set are very sensitive to the choice of population distribution: the lognormal decision thresholds are 200%–300% higher than their gamma counterparts.

The root of the problem lies in transforming the data values to their logarithms: the transformation tends to mask the presence of extreme data values and, as discussed in Appendix D, section D.4, this inflates the standard deviation of the fitted lognormal distribution which in turn inflates the calculated decision thresholds. In contrast, the statistics of a gamma distribution are much less sensitive to outliers, so that decision thresholds based on the gamma option are not overly influenced by outliers (EPA 2009, 2013a, 2013b).

This example underscores the importance of avoiding the lognormal option when modeling a data set in which the gamma distribution is a statistically valid alternative, particularly when the background data set is strongly skewed and/or contains multiple outliers that may not be representative of the background (uncontaminated) population. The gamma distribution is far less sensitive to the presence of extreme values and whenever possible should be used to model highly skewed data sets.

Table K4. Comparison of decision thresholds calculated from log-transformed background data that should have been modeled with a gamma distribution.

Data are Not Normal at 5% Significance Level						Estimates	Averages
				95% UTL with	95% Coverage		~~161.7~~
				95% UPL for k = 5			~~174.5~~
					95% USL		~~180.8~~
Data appear Gamma Distributed at 5% Significance Level							
	95% WH Approx. Gamma UTL with				95% Coverage	245.5	265.8
	95% HW Approx. Gamma UTL with				95% Coverage	286.1	
				95% WH UPL for k = 5		290.5	319.7
				95% HW UPL for k = 5		348.8	
					95% WH USL	314.2	348.5
					95% HW USL	382.7	
Data appear Lognormally Distributed at 5% Significance Level							
				95% UTL with	95% Coverage		816.5
				95% UPL for k = 5			1245.0
					95% USL		1528.0
For Nonparametric Background Statistics				**(data can be parametrically described)**			
				95% UTL with	95% Coverage	~~169.8~~	~~169.8~~
	95% Percentile Bootstrap UTL with				95% Coverage	~~169.8~~	
	95% BCA Bootstrap UTL with				95% Coverage	~~169.8~~	
					95% UPL		~~168.2~~
					95% USL		~~169.8~~
WH = Wilson-Hilferty approximation							
HW = Hawkins-Wixley approximation							

References

Davis, C.B. and R.J. McNichols. 1987. "One-sided Intervals for at Least *p* of *m* Observations from a Normal Population on Each of *r* Future Occasions." *Technometrics*. 29: 359–370.

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

EPA (United States Environmental Protection Agency). 2013a. *ProUCL 5.0 Software and User Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

EPA (United States Environmental Protection Agency). 2013b. *ProUCL 5.0 Technical Guide*. *http://www.epa.gov/osp/hstl/tsc/software.htm*.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Appendix L. Interim Decision Thresholds in the Presence of a Secular Trend

L.1. Introduction

Trending data present a special problem in ground water monitoring. Statistically, trending data cannot be used to estimate a mean and variance because these quantities change with time, so that many of the methods outlined in this document cannot be applied. This appendix provides guidelines that can be used to evaluate statistically significant changes in trend, specifically, to provide guidance for cases where (1) background water quality cannot be established because upgradient monitoring wells exhibit secular trends in water quality, (2) site practices are being modified to bring the system into compliance and downgradient water quality is or will be affected, or (3) and an interim compliance threshold is required for wells exhibiting a positive secular trend.

In case 1, the facility is affected by secular trends originating off site and over which the regulated entity may have no control. In case 2, previous permitted use of a facility may have caused downgradient wells to exceed the primary and/or secondary standards (IDAPA 58.01.11.200) or alternative concentration limits (ACLs) for the constituent of concern (COC); if so, the regulated entity may be implementing operational changes in response and future water quality trends may change as site background approaches a new steady state condition. In cases 1 and 2, the method described in sections L.2 and L.3 should be followed to monitor the trend until procedures described in section L.4 can be applied and an appropriate decision threshold (Appendices H–K) calculated.

In case 3, the existence of a positive trend suggests that the downgradient well may be out of compliance at some point in the future so that an interim decision threshold needs to be established to determine if and when that happens. The methodology in section L.4 is suggested for this purpose, but it places a substantial burden on the regulated entity to justify the statistical and hydrogeological rationale on which the method relies.

L.2. Procedure for Setting a Decision Threshold to Monitor a Trend

The assumption is made that if ground water is not currently in a steady state condition, then it is approaching a steady state condition because upgradient land uses or practices at the facility have stabilized. During this transition time, data still need to be collected, and limits are required to ensure that the approved practices are causing the water quality to continue to trend towards a future steady state condition. The following method is to be used for setting limits for background during the transition time.

1. Each year, in the first quarter, determine if the system is trending or in steady state (Mann-Kendall test).
    a. The facility should recheck annually because it is likely that a trend will change with time as a new steady state condition is approached. For example, the COC will begin to level off as it approaches steady state.
    b. Once the COC has defined a statistically steady state condition, use the last 12 data points to establish tolerance or prediction intervals described in the appendices.

On a case-by-case basis where there is a new, mutually agreed-upon mean concentration for a particular COC, the standard deviation should be based on the (nontrending) background concentration in the last 12 sampling events.

2. If the Mann-Kendall test shows the system has not achieved a new steady state, estimate the trend by the Sen's slope method outlined in section L.3.

   In cases where a limit is needed for comparison with the trending data, use the 100(1-α)% lower confidence limit for an increasing trend and the 100(1-α)% upper confidence limit for a decreasing trend. The difference between the next measurement and its previous measurement should be within these limits. In the case of an increasing trend, the regulated entity should use any exceedance as a warning that the current practices being used to reach a new steady state condition may not be adequate. An exceedance of the confidence limit during the transition period will be addressed on a case-by-case basis.

L.3. Nonparametric Sen's Slope Method to Estimate Trend

Sen's trend estimator is particularly useful for ground water monitoring: it is simple to compute; robust to outliers, missing data, and nondetects (Gibbons 1994); and it can be applied to as few as 2 years of quarterly data. The following fabricated example shows how to calculate Sen's slope for a trending ground water total dissolved solids (TDS) time series. The data values in Table L1 have been rearranged from those in Table B1 to serve as an example of a positive data trend.

Table L1. Fabricated well B1 data.

Time ID	Fabricated Well B1
Year 1—1st quarter	228
Year 1—2nd quarter	210
Year 1—3rd quarter	216
Year 1—4th quarter	248
Year 2—1st quarter	235
Year 2—2nd quarter	274
Year 2—3rd quarter	240
Year 2—4th quarter	259
Year 3—1st quarter	285
Year 3—2nd quarter	258
Year 3—3rd quarter	305
Year 3—4th quarter	290

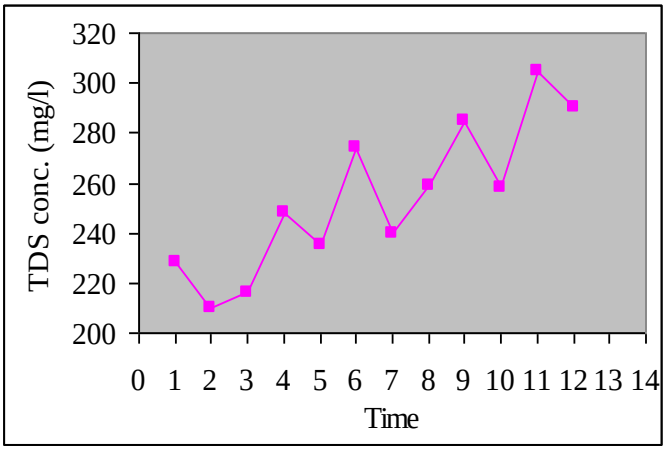


Figure L1. Concentration versus time plot for fabricated well B1 data.

Step 1: Lay out the concentration data in temporal order.

Step 2: Calculate individual slopes Q_i for each pair of monitoring events i and i', where i' follows i,

$$Q = \frac{x'_{i} - x_i}{i' - i}$$

and where x'_i and x_i are the measured concentrations during those monitoring events. Examples of the first few individual slope calculations are listed below:

$Slope^1 = (210-228)/(2-1) = -18$

$Slope^2 = (216-228)/(3-1) = -6$

$Slope^3 = (248-228)/(4-1) = 6.67.$

The total number of individual slope comparisons is N'=n(n-1)/2. In this example, N'=(12)(11)/2=66.

Time period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
TDS conc. (mg/L)	**228**	**210**	**216**	**248**	**235**	**274**	**240**	**259**	**285**	**258**	**305**	**290**
Q_i		-18.0	-6.00	6.67	1.75	9.20	2.00	4.43	7.13	3.33	7.70	5.64
			6.00	19.0	8.33	16.0	6.00	8.17	10.7	6.00	10.6	8.00
				32.0	9.50	19.3	6.00	8.60	11.5	6.00	11.1	8.22
					-13.0	13.0	-2.67	2.75	7.40	1.67	8.14	5.25
						39.0	2.50	8.00	12.50	4.60	11.7	7.86
							-34.00	-7.50	3.67	-4.00	6.20	2.67
								19.0	22.5	6.00	16.3	10.0
									26.0	-0.50	15.3	7.75
										-27.0	10.0	1.67
											47.0	16.0
												-15.0

Step 3: Rank the N' individual slopes from smallest to largest. In this example, the ranking results are shown in the following table (midpoint ranks and their Q_i values are highlighted):

Q_i	Rank	Q_i	Rank	Q_i	Rank	Q_i	Rank
-34.0	1	3.33	18	7.75	35	11.7	52
-27.0	2	3.67	19	7.86	36	12.5	53
-18.0	3	4.43	20	8.00	37	13.0	54
-15.0	4	4.60	21	8.00	37	15.3	55
-13.0	5	5.25	22	8.14	39	16.0	56
-7.50	6	5.64	23	8.17	40	16.0	56
-6.00	7	6.00	24	8.22	41	16.2	58
-4.00	8	6.00	24	8.33	42	19.0	59
-2.67	9	6.00	24	8.60	43	19.0	59
-0.50	10	6.00	24	9.20	44	19.3	61
1.67	11	6.00	24	9.50	45	22.5	62
1.67	11	6.00	24	10.0	46	26.0	63
1.75	13	6.20	30	10.0	46	32.0	64
2.00	14	6.67	31	10.6	48	39.0	65
2.50	15	7.13	32	10.7	49	47.0	66
2.67	16	7.40	33	11.1	50		
2.75	17	7.70	34	11.5	51		

Step 4: The trend estimate, *S*, is the median of the individual slopes. If N' is odd, S is the middle slope, $Q_{(N'+1)/2}$; if N' is even, $S=1/2*(Q_{(N'/2)}+Q_{(N'+2)/2})$. In this example, the estimated slope of the data trend is the average of the 33rd and the 34th Q_i values (highlighted in yellow). Therefore $S=(Q_{33}+Q_{34})/2= (7.4+7.7)/2=7.55$ milligrams per liter (mg/L). In other words, the TDS concentration increased during the 3-year monitoring period at an average rate of 7.55 mg/L per quarter.

Step 5: Calculate the variance of the original data using the following formula (Kendall 1975):

$$\text{var}(S) = \frac{1}{18}[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p-1)(2t_p+5)]$$

where n is the sample size, q is the number of data values that have ties, and t_p is the number of occurrences of each tie in Table L1. In this example, there are no tied data values, so n=12, q=0 and the t_p are all zero. Therefore,

$$\sum_{p=1}^{q} t_p(t_p-1)(2t_p+5)] = 0$$

and

$$\text{var}(S) = \frac{1}{18}[(12)(11)(29)\text{-}0] = 212.67$$

Step 6: To define a confidence limit for a future decrease in the absolute rate of change (either positive or negative), calculate the lower confidence limit (LCL) for an increasing trend or the upper confidence limit (UCL) for a decreasing trend.

$L.C.L$:

$$M_1 = \frac{N' - Z_{1-\alpha}\sqrt{\mathrm{var}(S)}}{2}$$

$U.C.L$:

$$M_2 = \frac{N' + Z_{1-\alpha}\sqrt{\mathrm{var}(S)}}{2}$$

where Z is the score of a standard normal population with mean = 0 and standard deviation = 0, and M_1 and M_2 are the rank orders for the ranked individual slopes in step 3. If M_1 and M_2 are not integers, interpolate from the nearest neighboring ranked slopes. For example, if $M_1 = 3.7$, the nearest neighboring ranks are 3 and 4, and the weighted average of the individual slopes from rank 3 and rank 4 is used. Therefore, since 3.7 is closer to 4, the individual slope at rank 4 is given more weight, so that the LCL = $(4 - 3.7)*Q_3 + (3.7 - 3)*Q_4$.

In the above example, S is positive indicating an increasing trend. To detect a statistically significant future decrease in this slope, we are interested in defining a LCL. The 95% LCL for the estimated slope is

$$M_1 = \frac{66 - 1.65\sqrt{212.67}}{2} = 21.01$$

and

$Q_{21.01} = 0.99*Q_{21} + 0.01Q_{22} = 4.6$ mg/L

That is, at 95% confidence, the LCL indicates that slopes computed with future quarterly data would have to be lower than 4.6 mg/L per quarter before a statistically significant slope change could be declared. Note that this LCL is a confidence limit for the slope *of the trend*; it is NOT used to compare individual future data points.

This methodology provides a means to determine when a change in site practices (or natural or off-site hydrologic conditions) has a statistically significant effect on the slope of an existing trend. In practice, it is advisable to consider multiple future data points, specifically by computing slopes on subsets of data (e.g., biannually, if quarterly data are being collected). As an example, consider the data in Table L2 that represents two additional years of fabricated monitoring information on well B1. Figure L2(A) shows the estimated slope of the historic trend as well as its 95% LCL, computed as above. As shown in Figure L2(B), even a full year of additional data would decrease the overall four-year slope only slightly, and the 95% UCL of all four years' data (7.6 mg/L/qtr) would still bracket the historic slope. As shown in Figure L2(C), the latest two years of data reveal a slope that is well below the LCL of the historic slope. Repeating the slope calculation after a second year of additional data become available shows a

further decrease in the slope of the biannual data set (Figure L2(D)). In addition, the 95% UCL of the latest two years' slope (2.91 mg/L/qtr) no longer brackets the historic slope, so at 95% confidence, it can be concluded that by year 5 the slope of the recent trend has decreased in a statistically significant manner relative to the historic trend.

When monitoring a secular trend for possible changes in slope, the Idaho Department of Environmental Quality (DEQ) recommends using the above procedure and annually recompute the slope of the most recent two years of quarterly data. Once the latest two-year slope is shown to fall outside both the UCL and LCL of the historic slope, then the secular trend is deemed to have changed in a statistically significant manner.

Table L2. Two additional years of fabricated well B1 data.

Time ID	Time Period	TDS (mg/L)
Year 4—1st quarter	13	294
Year 4—2nd quarter	14	305
Year 4—3rd quarter	15	290
Year 4—4th quarter	16	305
Year 5—1st quarter	17	309
Year 5—2nd quarter	18	295
Year 5—3rd quarter	19	306
Year 5—4th quarter	20	294

Notes: total dissolved solids (TDS); milligram per liter (mg/L)

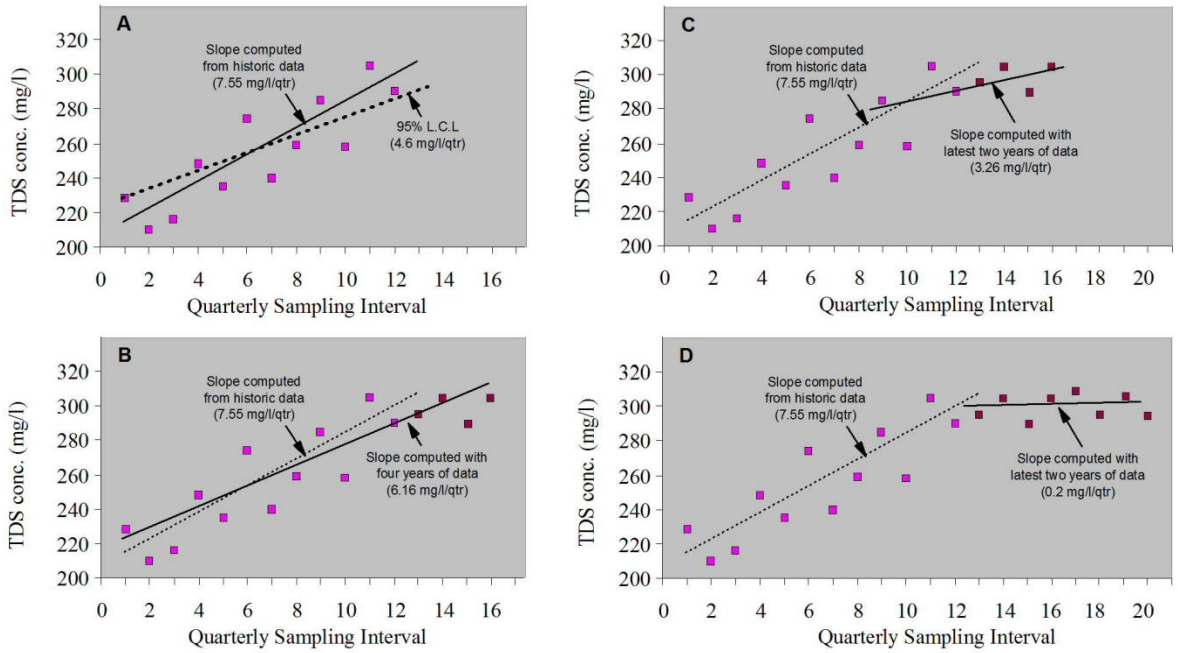


Figure L2. An example of applying a biannual slope recalculation procedure to identify changes in slope as future data is collected.

L.4. Interim Process for Estimating a Decision Threshold in the Presence of a Trend

The existence of a positive trend indicates that a well will be out of compliance at some point in the future and a decision threshold is required to determine when that situation occurs. However, the decision thresholds described in Appendices H–K are only valid for nontrending data. Before one can be computed, the trend must be quantified and removed so that background statistics and a decision threshold can be calculated.

Doing so, however, requires that detrending can be justified in both a statistical and a hydrogeological sense. Specifically, the trend must be expected to remain unchanged for some time into the future. If site practices are being modified or if aquifer conditions such as recharge or dewatering rates change, there is no guarantee that the historic trend will continue indefinitely. To compute a decision threshold based on detrended data, it is up to the regulated entity to annually justify that the historic trend still applies (using Section L.3 "Statistical Characterization of Ground Water Quality" methodology). This requirement is necessary to ensure the integrity of the decision threshold that will be used for determining compliance. When the trend changes in a statistically significant manner, then the detrending procedure may need to be modified (e.g., by using a new trend slope) and further compliance decisions put on hold until a new, stable trend (or a new background level) is established.

To compute an interim decision threshold based on detrended data, DEQ recommends the use of intrawell methodology, specifically, a Shewart-CUSUM control chart (Appendix N, section N.2). The following procedure, based on Gibbons (1994), pertains to an individual downgradient well's time-series data; it assumes that (i) the trend is linear and statistically significant (e.g., as in Appendix F) and (ii) data detrending is justified. The decision logic is outlined in Figure L3 and described in detail in the following steps.

I. Quantify the trend and the decision thresholds:

   1. Detrend the data using the following procedure.

      i. Determine the regression slope and intercept, a_0 and b, of the time-series data.

      ii. Detrend the raw data, x_i, using a_0 and b and the equation

      $$x_i^{detrended} = a_0 + [x_i - (a_0 + bt) \quad] \qquad \text{(Gibbons 1994, equation. 8.5)}$$

      iii. Compute the trend residuals ($x_i - x_i^{detrended}$) for each data value and check that their mean is zero (or acceptably close).

   2. As described in Appendix N, section N.2, calculate control chart limits using the detrended data (the $x_i^{detrended}$ values in Appendix N, section N.2).

   3. Plot a Shewart-CUSUM control chart using the *undetrended* data, x_i; declare an out-of-compliance situation if future measurements exceed either of the control chart thresholds.

II. Annually reevaluate the trend and decision thresholds:

1. Using the procedure in Appendix L, section L.2, determine if the trend of the past two year's data remain within the bounds of the historic trend. That is, ensure that the slope of the most recent trend is within the 95% upper and lower confidence limits determined for the historic trend.

2. If the trend of the most recent data is within the historic trend, then continue to use the existing control limits in the control chart.

3. If the most recent data deviates from the historic trend, then recalculate a_0 and b for the latest two years of quarterly data (or minimum 8 data values) and update the control chart decision thresholds (steps I.1 to I.3, above). The previous year's data should also be compared against the revised decision threshold to ensure that the well was in compliance during the previous year.

4. Repeat steps II.1 through II.3 annually to document that the current year's data can justifiably be detrended.

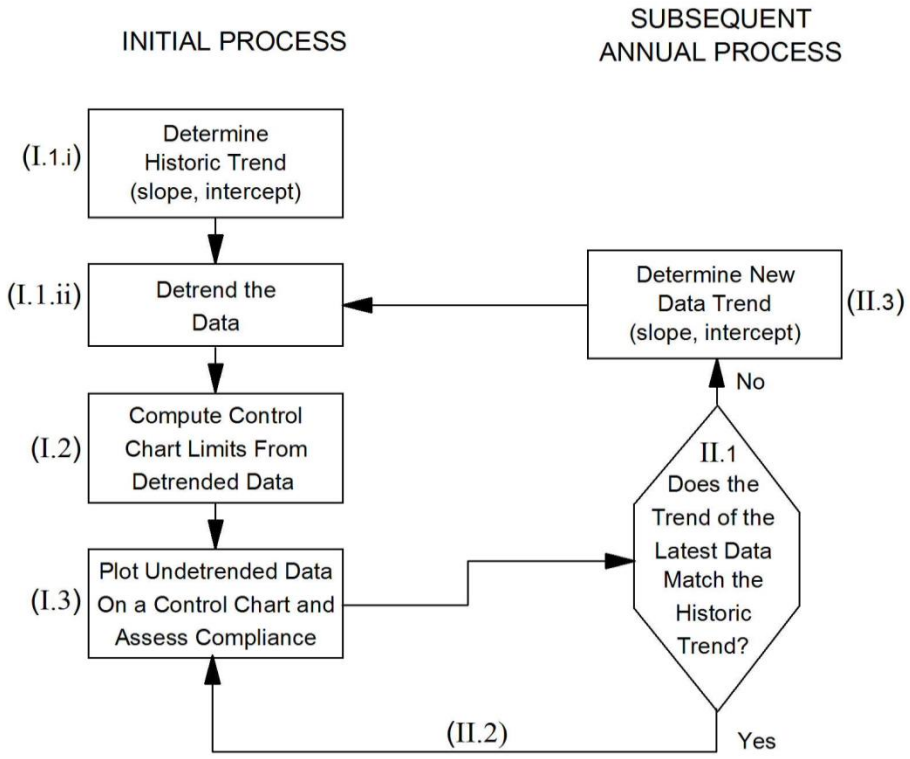


Figure L3. Recommended procedure for estimating interim decision thresholds in the presence of a trend (based on an approach outlined by Gibbons (1994) and the control chart methodology described in Appendix N). Parentheses indicate corresponding steps in the narrative process described in the text.

References

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Kendall, M.G. 1975. *Rank Correlation Methods*. 4th ed. London: Charles Griffon.

Appendix M. Example Scenario for an Existing Wastewater Reuse Facility With No Chemical Impact

The majority of the wastewater reuse facilities using this guidance will probably be in this category. In addition to existing facilities, new facilities where previous land uses have altered the downgradient ground water at compliance wells from an ambient condition also fall into this category. As with other facilities, the first steps are to conduct descriptive statistics on the constituents of concern (COCs) for each background well (Section 3, "Statistical Characterization of Ground Water Quality," Figure 2).

Following the initial descriptive statistical documentation (Appendix B) and an evaluation of data independence (Appendix C), the distribution of the each COC should be determined (Appendix D) and its temporal behavior evaluated for statistically significant secular trends and seasonal pattern (Appendix E). The concentration versus time diagrams will likely indicate whether there is a cyclic nature to the data, but seasonality must be statistically demonstrated. Ideally, at least 3 years of quarterly data should be available for this analysis (wherein each quarter is tested in the same month). Some of the variation may be due to changing land uses (e.g., nearby agricultural activities and river and canal flows) as well as true seasonal effects such as precipitation patterns, evapotranspiration. The preferred method for determining seasonal stationarity is the nonparametric Kruskal-Wallis test (Appendix E). Once seasonality has been tested for and possibly removed, the resulting data sets should be tested for secular trends using the recommended nonparametric Mann-Kendall test (Appendix F).

If the Mann-Kendall test shows no temporal trend for background ground water quality data, then the methodology in Appendix G should be used to determine whether data from multiple background wells can be pooled. If the Mann-Kendall test shows that there is a temporal trend, then an alternative method needs to be followed to define a decision threshold for future monitoring (Appendix L).

After defining background ground water quality for each COC, decision thresholds for future monitoring are set. In most cases, the existing facility will have altered background ground water quality, and the process outlined in Appendix J can be used to set parametric prediction levels for future interwell comparisons for the COCs. To do so, the data set must (1) exhibit no temporal trends, (2) have no statistically significant seasonal effects or be corrected for seasonality, and (3) be parametrically distributed. Appendix K makes the same assumptions except that the data distribution is nonparametric. In either case, site conditions are such that downgradient water quality has been affected by the facility, requiring that interwell statistical methods be applied.

Wherever possible, site conditions should be evaluated to determine if interwell comparisons are justified. For example, in situations where background ground water quality is highly variable, or aquifer heterogeneity makes it difficult or impossible to decide which upgradient well(s) should be compared to a downgradient well, then intrawell comparison procedures or modifications to those suggested in this document should be considered. For example, can background data in downgradient wells be filtered of outliers (Appendix N, section N.4) that may represent existing site impacts, prior to applying intrawell comparisons? Or could alternative methods for setting decision thresholds be used, such as Shewart-CUSUM control charts (Gibbons 1994)? Such a

decision may prove to be far more defensible for an existing facility than trying to force interwell comparisons where hydrogeologic conditions do not warrant them.

The single greatest advantage to using intrawell methods (including variants such as Shewart-CUSUM charts) is that decisions are solely based on the statistical behavior of COCs in individual wells rather than between wells whose upgradient versus downgradient hydrogeologic relationship may be suspect or unknown. In all cases, the use of intrawell comparisons at existing facilities are justified only if the regulated entity can demonstrate that the data set(s) to be considered as *background* for the COCs in downgradient wells have either not been affected by the facility's prior operations or appropriately filtered of suspected contamination influences (e.g., outliers).

References

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Appendix N. Applying Intrawell Analysis at Existing Facilities When Interwell Methods are Inadvisable

At existing facilities, large variations in natural background water quality across a site can make it difficult to identify hydrogeologically appropriate pairs of wells for interwell comparison. This problem that can be exacerbated by very slow ground water flow rates between wells. This increases the difficulty of identifying true exceedances from other confounding influences in an interwell comparison. For these reasons intrawell comparison, where it can be justified, is the method of choice for compliance monitoring (Gibbons 1994).

At existing facilities intrawell comparison is preferred over interwell analysis whenever possible. Specifically, if historical background data in a downgradient well demonstrates that the constituents of concern (COCs) have not been impacted by the facility's operations, then the use of intrawell methods may be justified for detection monitoring.

To apply an intrawell monitoring method at an existing facility, the regulated entity must demonstrate that preexisting contamination was not present in the downgradient wells during the historical period selected for establishing the background level. The intrawell method that the Idaho Department of Environmental Quality (DEQ) suggests for monitoring preexisting facilities is the combined Shewhart-CUSUM control chart method. It is capable of detecting both immediate and gradual releases and is applicable to data sets containing up to 75% nondetects. It combines the power of the Shewhart control chart method, which is ideal for rapid detection of large releases, and the Cumulative SUM method that is sensitive to gradual releases. Data must be temporally independent, so that quarterly data are recommended, and should be screened for outliers or other evidence of preexisting impacts by the facility.

N.1. Demonstrating that Intrawell Comparison is Appropriate for Site-Specific Conditions

The regulated entity should provide evidence that COCs in downgradient wells have not been affected by the facility. For example, based on an evaluation of historical data (outlier screening, seasonality, trends), the regulated entity may be able to demonstrate that a window of time exists for defining background COC levels for each well in the monitoring network. Outlier detection is addressed in Appendix N, section N.4. To check trend and seasonality of the historical data, refer to Appendices E and F.

Alternatively, groups of upgradient and downgradient wells can be tested for statistical similarity (using methods of Appendix G) to identify those downgradient wells whose water quality is statistically indistinguishable from upgradient wells and whose future data could be analyzed using either interwell or intrawell methods.

For each downgradient well that has not been affected by facility impacts, screen the historical COC data and remove outliers to establish an historic statistical baseline (e.g., Gibbons 1994, section 8.4.3). Justify using an intrawell comparison by demonstrating that no COCs have yet been detected in the downgradient well and that other indicator constituents show no significant trends (ASTM 1998; Cal EPA 2001).

N.2. Apply a Shewhart-CUSUM Control Chart Method to Detect Future Changes in Water Quality

The Shewhart-CUSUM control chart procedure is a widely used intrawell comparison method that the United States Environmental Protection Agency (EPA) recommends for identifying a statistically significant increase in chemical concentrations at a single monitoring location (EPA 1989, 1992, 2009; ASTM 1998; ITRC 2006; URS 2003; Gibbons 1994, 1999). DEQ recommends 12 background samples (i.e., 3 years of quarterly data) are needed to compute a standardized difference value and control limits against which subsequent measurements from the same well are compared. The method has been applied in an evaluation mode at various sites (e.g., Chou 2004) and has performed well. Because the method is sensitive to both gradual (long-term) and sudden (short-term) increases, it allows for detection of facility impacts at different spatial and temporal scales. The method is applicable to data that are independent and normally distributed; hence a well's historic background data should be evaluated for temporal independence, or the analysis should be restricted to data that have been collected no more frequently than quarterly.

The procedure can be implemented as follows: Let x_i be a series of independent background observations i = 1, 2…., n (n = 12 at minimum). Let x_j be a series of future monitoring measurements j = 1, 2, 3….. . Then, using the background and future data, the following steps are applied. For additional detail, see Gibbons (1994). Alternatively, a slightly modified procedure is described in Chapter 20.2 of EPA's Unified Guidance (EPA 2009).

1. Check the background and future data for temporal independence.
2. Use the background data (x_i) to compute $\bar{x}$ and s as estimates of the mean μ and standard deviation σ, respectively.
3. Define three parameters (all in units of standard deviation) for the control chart:

   SCL—Shewhart control limit
   h—CUSUM control limit
   k—half the shift in standard deviation to be detected rapidly

   For ground water quality monitoring, experience has shown that values of SCL=4.5, h=5.0 and k=1 are appropriate. Other values may be used depending on the sampling scheme and sample size (Gibbons 1994; EPA 2009).

4. For each future data value, compute its standard normal deviate, z_j:

$$z_j = \frac{x_j - \bar{x}}{s}$$

5. Compute the CUSUM statistic S_j for each future data value:

$$S_j = \max[0, (z_j - k) + S_{j-1}] \; ; \; S_0 = 0.$$

If either z_j exceeds the Shewart control limit (SCL) **or** if S*j* exceeds the CUSUM control limit h, then a potential exceedance is indicated, triggering verification resampling that is temporally

independent of the initial sampling within a time frame based on consideration of site-specific ground water flow conditions and consultation with DEQ. Verification resampling corroborates the exceedance only if the verification measurement also exceeds either SCL or h. If an exceedance is not corroborated by verification resampling, then those data can be used in step 2 to update the background data set for future comparisons.

The power of the control chart method arises because the SCL is capable of detecting large, rapid deviations from background, whereas the CUSUM portion of the test is sequential: a small positive concentration shift over the preceding time period will slowly aggregate in the CUSUM statistic and eventually cause it to exceed the CUSUM control limit. Thus, the combined Shewhart-CUSUM method has the ability to detect rapid as well as gradual releases from a monitored facility.

N.3. Example

For this example, we assume that the fabricated data in the table have been screened and outliers removed, have been corrected for seasonality, and are free of any secular trend.

Table N1. Background total dissolved solids measurements.

Background sample number	Background TDS, mg/L
1	259
2	228
3	240
4	216
5	285
6	235
7	290
8	274
9	290
10	228
11	216
12	248

Notes: total dissolved solids (TDS); milligram per liter (mg/L)

$\bar{x}$ = 251, s = 28.1, n= 12

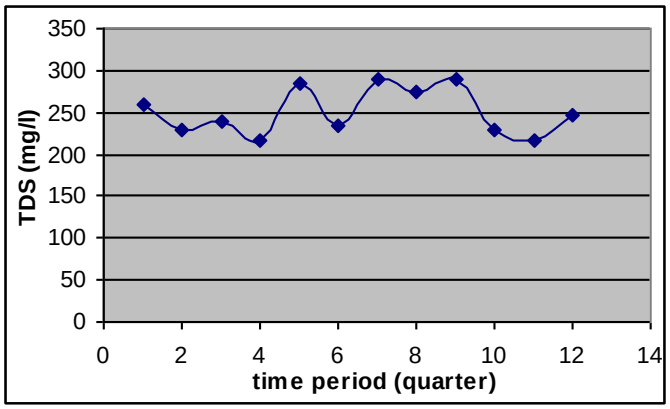


Figure N1. Historical (background) total dissolved solids concentrations versus time.

In this example, set h=5, SCL= 4.5, k=1 and calculate z_i, z_i-k and S_i as outlined in Appendix N, section N.2. The calculated limits are summarized in Table N2 and plotted in Figure N2, demonstrating that both z_i and S_i are within specified limits. Therefore, the additional year's monitoring data confirm that the system remains in compliance.

Table N2. Calculated test statistics for current year's monitoring data.

Current Year's Sampling Data	Measured TDS (mg/L)	z_i	z_i-k	S_i
1st quarter	258	0.26	-0.74	0.0
2nd quarter	305	1.93	0.93	0.9
3rd quarter	289	1.36	0.36	1.3
4th quarter	268	0.61	-0.39	0.9

Notes: total dissolved solids (TDS); milligram per liter (mg/L)

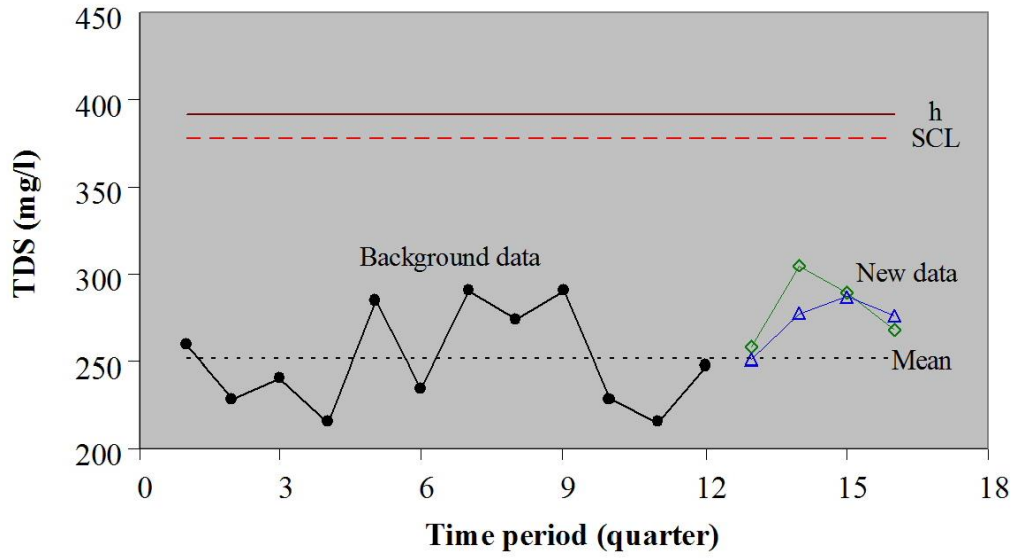


Figure N2. Comparison of latest monitoring results to historical data and specified control limits.

N.4. Detection of Outliers in Background Data

The following discussion outlines the steps necessary to detect outliers using Dixon's method. Dixon's test can be used when the number of suspected outliers is small. If m outliers are suspected, all m tests must be performed regardless of the outcomes of the previous m-1 test. If the m^{th} test exceeds the critical value, all m outliers must be rejected. If data are not normal in original scale, proper transformation should be applied. Once the data are transformed, the following steps then should be applied.

1. Sort the data from lowest to highest, denoted by $x_{(i)}$ where i=1 to n.
2. Calculated the average of the data, $\bar{x}$.
3. Calculated $\left|x_{(i)} - \bar{x}\right|$ for each observation and sort the difference from largest to smallest.
4. Identify suspected outliers and their number, m.
5. Calculate Dixon's statistics using following formula (Gibbons 1994) for the m outliers, starting from the most extreme value.

N	Highest value	Lowest value
3–7	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$	$\frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$
8–10	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$	$\frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$
11–13	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$	$\frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$
14–25	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}}$	$\frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}$

6. Compare the statistic to following tabulated critical values (Gibbons 1994) and draw conclusions.

N	5% level	1% level	N	5% level	1% level
3	.941	.988	14	.546	.641
4	.765	.889	15	.525	.616
5	.642	.780	16	.507	.595
6	.560	.698	17	.490	.577
7	.507	.637	18	.475	.561
8	.554	.683	19	.462	.547
9	.512	.635	20	.450	.535
10	.477	.597	21	.440	.524
11	.576	.679	23	.421	.505
12	.546	.642	24	.413	.497
13	.521	.615	25	.406	.489

Using the same fabricated data, we assume that there is one more observation in the historical data, the 13th measurement with total dissolved solids (TDS) equal to 380 milligrams per liter (mg/L). Applying the steps outlined above results in Table N3. The ascending sorted observations $x_{(i)}$ is shown in column 3. The mean of the data, $\bar{x}$, equals 260.7 mg/L. Therefore, $|x - \bar{x}|$, sorted for each observation and its corresponding measured values are shown in columns 4 and 5. The suspected number of outliers is 3 (m=3, the highest TDS and the two lowest TDS in the data set). Using Dixon's formula in step 5 for n=12, starting with the most extreme value TDS=380 mg/L, Dixon's statistic is (380-290)/(380-216)=0.549. It is significant at 5% level but not at 1% level comparing to critical values in the step 6 table. Continuing with the lowest TDS=216, using the same approach, Dixon's statistic is 0.162, not significant at 5% level and 1% level. Therefore, the observation with TDS=380 mg/L can be rejected and observations with TDS=216 should be retained for intrawell comparison.

Table N3. Background total dissolved solids measurement with fabricated outliers.

Background Sample number	TDS, mg/L	Sorted TDS, $x_{(i)}$	Sorted $\lvert x - \bar{x} \rvert$	Corresponding Background TDS, mg/L	Dixon's Statistic
1	259	$x_{(8)}$	119.3077	380	0.549
2	228	$x_{(3)}$	44.69231	216	0.162
3	240	$x_{(6)}$	44.69231	216	0.162
4	216	$x_{(1)}$	32.69231	228	
5	285	$x_{(10)}$	32.69231	228	
6	235	$x_{(5)}$	29.30769	290	
7	290	$x_{(11)}$	29.30769	290	
8	274	$x_{(9)}$	25.69231	235	
9	290	$x_{(12)}$	24.30769	285	
10	228	$x_{(4)}$	20.69231	240	
11	216	$x_{(2)}$	13.30769	274	
12	248	$x_{(7)}$	12.69231	248	
13	**380**	$x_{(13)}$	1.692308	259	

Notes: total dissolved solids (TDS); milligram per liter (mg/L)

References

ASTM (American Society for Testing and Materials). 1998. *Standard Guide for Developing Appropriate Statistical Approaches for Groundwater Detection Monitoring Programs*. Designation: D 6312-98. West Conshohocken, PA: ASTM.

Cal EPA. (California Environmental Protection Agency). 2001. *Guidance Document - Monitoring Requirements for Permitted Hazardous Waste Facilities.* California EPA, Department of Toxic Substances Control, Hazardous Waste Management Program. *http://www.dtsc.ca.gov/HazardousWaste/upload/HWMP_Guidance_Monitoring-Requirements.pdf.*

Chou, C.J. 2004. *Evaluation of an Alternative Statistical Method for Analysis of RCRA Groundwater Monitoring Data at the Hanford Site*. Richland, WA: Pacific Northwest National Laboratory. PNNL-14521.

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Gibbons, R.D. 1999. "Use of Combined Shewart-CUSUM Control Charts for Ground Water Monitoring Applications." *Ground Water*. 37(5): 682–691.

ITRC. (Interstate Technology & Regulatory Council). 2006. *Evaluating, Optimizing, or Ending Post-Closure Care at MSW Landfills Based on Site-Specific Data Evaluations*. Washington, DC: Interstate Technology & Regulatory Council, Alternative Landfill Technologies Team. *www.itrcweb.org*.

URS Group. 2003. *Implementation of Alternative Measures Industrial Waste Lagoon, Tooele Army Depot, Tooele, UT; Final System Non-operation Test Proposal*. Bethesda, MD: URS Group, Inc.

Appendix O. Statistical Concepts

O.1. Statistical Notation Used

Throughout this document, certain mathematical symbols are reserved for quantities related to sample size such as the number of observations, number of years of sampling, and frequency of sampling within the year. Other symbols will be used to denote the sample mean, standard deviation, and other sample-based statistics. For reference, some of the frequently used symbols are summarized in Table O1.

Unless stated otherwise, the symbols $x_1, x_2, \ldots, x_N$ are used in this guidance to denote a chemical concentration measurement in each of N ground water samples taken at regular intervals during a specified period of time. The subscript indicates the order in which the sample was drawn (e.g., x_1 is the first or oldest measurement while x_N is the N^{th} or latest/newest measurement). Collectively, the set of values used (represented by x's) is referred to as a data set, and in general x_i will be used to denote the i^{th} measurement in the data set.

Table O1. Summary of statistical notation used.

Symbol	Definition
x_i	Constituent concentration measurement for the i^{th} ground water sample
m_j	The number of years for which data were collected (the types of analysis discussed in this guidance usually will be performed with 2 to 3 full years of data)
k	The number of sample measurements per year (for quarterly data, n=4). This is also referred to as the number of *seasons* per year.
N	The total number of sample measurements (for m complete years of data, N = n x m).
$\bar{x}$	The mean (or average) of the chemical measurements in a sample of size N.
s	The standard deviation of the chemical measurements in a sample of size N.
s^2	The variance of the chemical measurements in a sample of size N.

O.2. Terminology and Definitions

A **population** is the set of all possible measurements of interest in the real world. For example, an aquifer's nitrate concentration represents a population of all possible aliquots of water that could be collected from that aquifer and analyzed for nitrate.

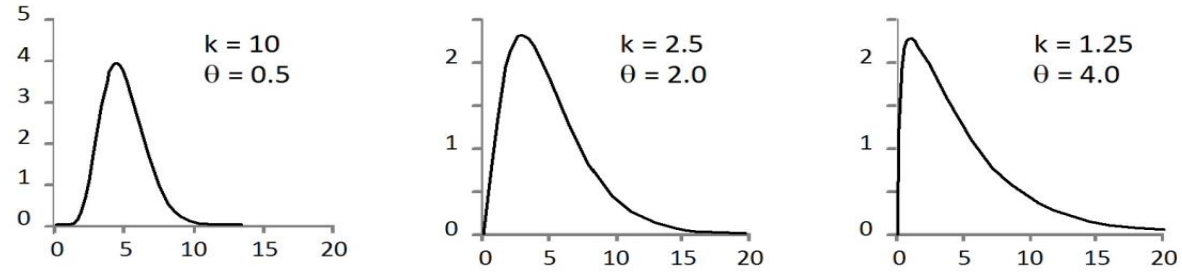
A **sample** is a set of measurements collected from the population in a manner that attempts to be representative and unbiased. In the preceding example, a sample might consist of 20 measurements (the **sample size**) collected at randomly chosen wells from across the aquifer.
A **parameter** is a numerical measure of the characteristic of interest in the population being sampled. Typical parameters are the population mean (μ), variance (σ^2) or standard deviation (σ), and proportion (p). Parameter values are usually unknown.

An **estimate** is a numerical measure of a parameter derived from a sample. Commonly used sample estimates are its mean ($\bar{x}$), variance (s^2) or standard deviation (s), and proportion ($\hat{p}$).

Inference is the process applied to estimate a population's parameter(s) from the sample. The parameters are usually the targets of our interest but, because it is impossible or prohibitive to collect every measurement from the population, they are usually unknown. There are two possible approaches to making inferences: estimation and hypothesis testing. The first answers the question, "what is the value of any given parameter?" and the second answers the question, "does the parameter meet a certain value or condition?" In ground water analysis, corresponding questions might be "what is the nitrate concentration in the ground water?" or "does the nitrate concentration meet State standards?"

Data independence is the most basic requirement of statistical inference. All measured values in a sample are assumed to be random. In a time-series sense, a measurement must not depend on—or affect—any prior or future measurement. In a spatial sense, a measurement made in one well must be independent of those made in other wells. Data that violate this requirement (e.g., as in replicate measurements collected over a short time span in the same well) carry redundant information that biases the calculation and/or inference of any statistical quantity based on it. Appendix C provides further information.

The sample **distribution** is the frequency or probability of occurrence of measured values. In ground water analysis, two commonly encountered distributions are the **normal distribution** (bell-shaped curve) and the **lognormal distribution** (a right-skewed curve that takes on a normal shape when values are logarithmically transformed). The **gamma distribution** is a third type, representing a family of functions whose shape can vary from symmetrical to highly right-skewed depending on its parameter values. It is a particularly useful parametric distribution because it can be used to represent both bell-shaped and right-skewed data sets, including lognormally distributed data. Its skewness and mean are described by a shape parameter (k) and a scale parameter (θ), where the distribution's mean is defined as the product of the two, $\mu = k*\theta$. Three different shapes of the gamma distribution are shown below, all having the same mean:



Transformation of the data to a normal distribution via a mathematical function is commonly done on environmental data prior to statistical analysis. For example, by converting data values of a lognormally distributed data set to their logarithmic equivalents, the resulting transformed data set would have a normal distribution.

Estimates derived from different samples of the same population are known as the **sampling distribution**. Estimates derived from a random sample of the population (e.g., $\bar{x}$) are also random; different samples generate different estimates. Many common statistical methods are based on a knowledge of, or the assumed characteristics of, the sampling distributions. One of the most famous is the **central limit theorem,** which says that the sampling distribution of the

mean of many independent random samples is normal regardless of the underlying distribution of the population that was sampled.

O.3. Methods for Describing a Distribution

Data need to be summarized in order to make meaningful interpretations and to pose testable hypotheses. Data summarization can be graphical or numerical. Graphical methods emphasize the shape of the distribution and numerical methods emphasize its central tendency and dispersion.

Graphical Methods

Histogram—summarizes the frequency distribution of a data set by displaying the number of observations that fall in defined intervals. The numbers of intervals commonly suggested is 5–15, but the number of intervals can greatly affect the visual appearance of the distribution. The Y-axis can either be the number of occurrences or the percentage of total occurrences in each interval. Figure O1 is an example of a histogram.

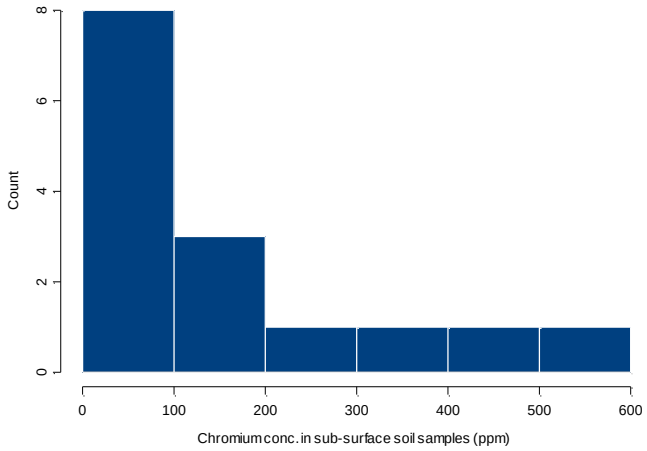


Figure O1. Example histogram.

Box plot—a simple, graphical representation of a data distribution. Figure O2 shows an example box plot. A typical box plot has the following elements: the box, representing the center half of the data bounded by the **interquartile range** (the 25% and 75% quantiles, also known as the first and third quartiles); the line within the box, representing the median; and the whiskers, representing the upper and lower adjunct values. Generally, the upper adjunct value is 1.5 times the interquartile range (IQR) above the 75% quantile and the lower adjunct value is 1.5 times the IQR below the 25% quantile. The IQR is robust to extreme values but cannot describe the overall nature of the dispersion. The interquantile range and quantiles are described in the paragraph on dispersion, below.

Observations beyond the upper and lower adjunct values are extreme values; an extreme value in a data set is NOT necessarily a bad observation that needs to be removed. However, outliers can be discarded from the data set with adequate justification. Box plots summarize a sample's central tendency, spread, skewness, and extreme values. Side-by-side box plots are a useful tool for comparing the distributions of different samples or groups of measurements.

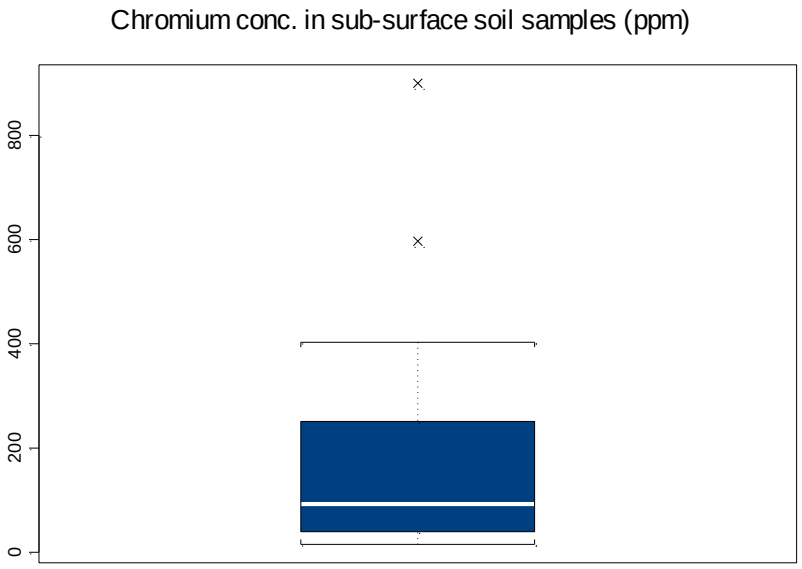


Figure O2. Example box plot.

Time-series plot—shows the values of observations on the y-axis versus corresponding points in time when they were collected on the x-axis. Equally spaced time points are desirable. A time-series plot is useful for examining general trends over time and evaluating seasonal or cyclical patterns and disrupting events (such as the effect of a drought year on water quality). If the data points are not collected in equal time intervals, it is important to reflect the interval width between time points in the plot. Otherwise, the apparent visual trend could be misleading. Figure O3 provides an example of a time-series plot.

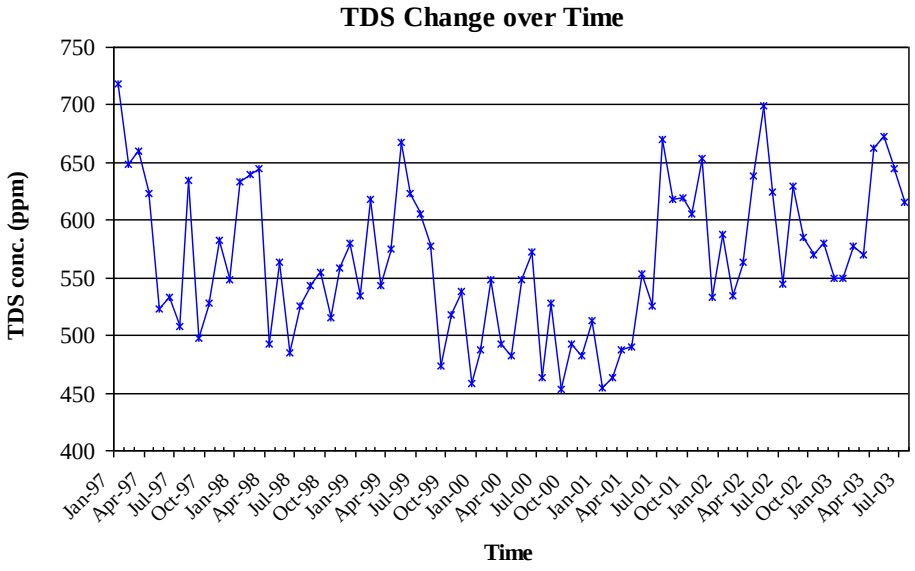


Figure O3. Example time-series plot.

Scatter plot—used to examine the relationship between two variables, x and y. Each point on the scatter plot represents a pair of measurements of x and y from the same source (e.g., concentrations of total dissolved solids [TDS] and nitrate-nitrogen [NO_3-N] in the same well

sample). Usually we are interested in determining if there is a linear or nonlinear correlation between the two variables. Figure O4 is an example scatter plot.

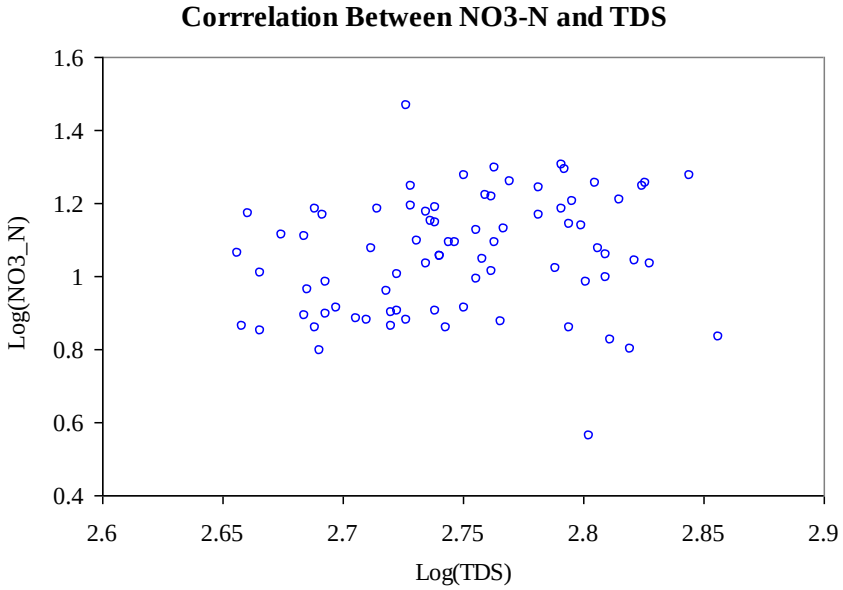


Figure O4. Example scatter plot.

Numerical Methods

Central tendency is a distribution's *center of mass*. Common measures of central tendency include the mode, median, and mean.

Mode—the most frequently occurring value in a data set. Distributions can have more than one mode (e.g., bimodal and trimodal).

Median—the middle value of a data set. It is the 50th percentile of a distribution, in which half of the observations are less, and half are greater, than the median value. In a data set whose N observations are arranged from smallest to largest, the location of the median is (N+1)/2 from the bottom of the list (or the average of the two middle observations).

Mean or **arithmetic mean** or **average—**the sum of N observations divided by N. The goal of statistical inference is to estimate the population mean (μ) from the sample mean. The sample mean ($\bar{x}$) is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$$

where N is total number of observations in the sample. The mean is sensitive to extreme values in a given data set and therefore may not always represent a distribution's central tendency. The median is robust to extreme values and thus is a better measure of central tendency in skewed distributions. For a symmetric distribution, the mode, mean, and median are the same; for

skewed distributions, they are different. The graphs in Figure O5 show their relationships in various distributions.

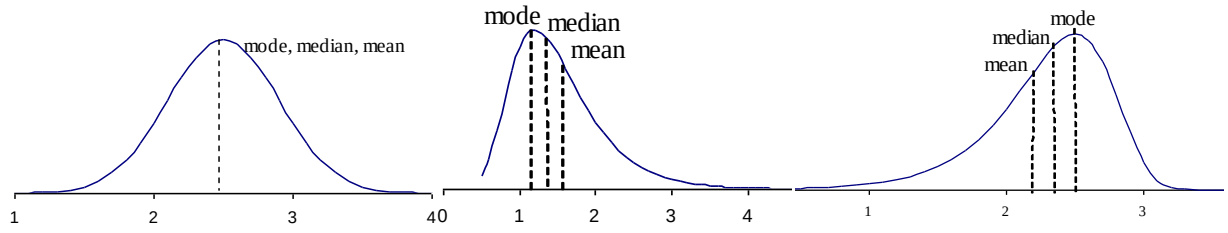


Figure O5. Mode, median, mean for various distributions (from left to right: symmetric, positively skewed, and negatively-skewed).

Dispersion is the **spread or variability** around the central tendency. Common measures include the range, IQR, variance, and standard deviation. **Range** is the difference between the largest and smallest values in a data set. Although it is simple to calculate, it is least useful in describing dispersion since it reflects the extreme values. A better measure involves the use of **quantiles**. The p^{th} quantile of a data set is the value that p percent of the observations are less than or equal to. The most commonly used quantiles are the 25^{th} (Q1 or first quartile), 50^{th} (Q2 or median) and 75^{th} (Q3 or third quartile).

The **variance** of a sample data set is the average of the squared deviations of the observations from the mean. It is calculated as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

The **standard deviation** (SD) is the square root of the variance, $s = \sqrt{s^2}$. For a normal distribution, the following empirical rule applies: 68% of the measurements are within one SD of the mean, 95% are within two SDs, and 99% are within three SDs of the mean.

$\bar{x}$ and s or s^2 are the most commonly used descriptive statistics for a distribution's central tendency and dispersion. However, they are most appropriate for symmetric distributions because they are sensitive to extreme values. To adequately characterize a skewed distribution, the range, Q1, median, and Q3 should be reported.

Skewness is the third moment of a distribution and measures its asymmetry, defined as

$$\boldsymbol{Skewness} = \frac{\sum(\boldsymbol{y_i} - \boldsymbol{y})^3}{(\boldsymbol{n}-\boldsymbol{1})^3}$$

Where y is the sample mean of n measurements and the y_i are the sample measurements. Skewness is zero for a symmetric distribution and either **positively skewed** (skewed-to-the-right) or **negatively skewed** (skewed-to-the-left) for asymmetric distributions. In a positively skewed distribution, the measurements tend to cluster around smaller values and tail toward larger

values. The coefficient of skewness is an alternate measure of skewness, defined as 3*(mean - median)/std. deviation.

Kurtosis is the fourth moment of a distribution and measures the sharpness of its peak. Kurtosis for a normal distribution is equal to 3.0 (zero in some statistical packages that subtract 3 from calculated moment to make kurtosis equal to zero for a normal distribution). A kurtosis greater than 3.0 (zero) indicates a distribution that is more sharply peaked than a normal distribution. The example in Figure O6 shows different example distributions, one with a kurtosis greater than 1 (red,), one with zero kurtosis (blue, a normal distribution) and one with a kurtosis less than 0 (green) (all values after subtracting 3 as just described).

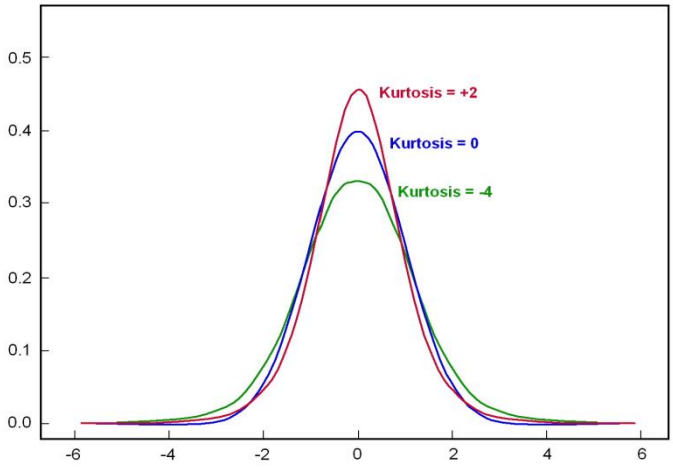


Figure O6. Example of three distributions with various degrees of kurtosis (peakedness).

O.4. Inference: Decision Thresholds

The first common type of statistical inference is designed to answer, "*what is the value* of a given parameter?" Some decision thresholds that are commonly applied in ground water analysis are the prediction limit, tolerance limit, and confidence limit. An **upper prediction limit** is the value, based on existing measurements and a specified level of confidence, below which the next *k* future measurements are expected to fall. An **upper tolerance limit** is the value defined at a particular confidence level for a specified percentage of all future measurements**.** In contrast, a **confidence limit** defines a permissible range for a specified population parameter (e.g., the mean) at a specified level of confidence. All three limits are calculated from historical background data on a constituent-by-constituent basis and are used to compare future measurements to determine whether the sampled population has changed (as by contamination).

Prediction and tolerance intervals are applied for compliance sampling events in detection, assessment, and monitoring programs and can be used for establishing background-based ground water protection standards (EPA 2009). Confidence intervals are often applied for comparing measurements to a ground water standard that is based on a mean or median value (Virginia DEQ 2003). Before such intervals are calculated, the background sample distribution should be checked for normality or lognormality, stationarity, and data independence.

Upper prediction limit (UPL):

UPL = $\bar{x} + Ks$

where *K* is the one-sided normal tolerance factor that can be found tabulated in various ground water monitoring guidance documents (Gibbons 1994; EPA 2009). As Gibbons (1994) has pointed out, *K* must be calculated for a specified statistical model that includes a verification sampling protocol, background sample size, and *k*, the number of relevant[††] future measurements that will be compared within a specified time period. Appendix J provides an example table of *K* factors. If any constituent of concern exceeds the UPL during the comparison time period, then standard practice is to conduct further sampling according to the verification sampling protocol, to verify the exceedance.

Upper tolerance limit (UTL):

UTL = $\bar{x} + Ks$

where *K* is the one-sided normal tolerance factor defined for a specified fraction (the *coverage*, e.g., 95%) of all future comparisons (Gibbons 1994; EPA 2009). A specified number of exceedances are allowed as long as their total number is no greater than the specified percentage of comparisons made since the UTL was set (e.g., 5% for 95% coverage).

Upper confidence limit (UCL):

UCL = $\bar{x} + t_{n-1,\alpha} \frac{s}{\sqrt{n}}$

where $\bar{x}$, *n* and *s* are the average, number and standard deviation of the background data, respectively, and t is the t-statistic with n-1 degree of freedom at a 1-α upper-tail confidence.

If the data are lognormally distributed, all of the above limits should be calculated on a log-transformed scale and compared with data that have also been log-transformed.

O.5. Inference: Hypothesis Testing

The second common type of statistical inference is aimed at answering the question, "does the population parameter *meet a specific condition or value*?" For example, does the mean NO_3-N concentration of the upgradient wells around a land application facility exceed the 10 milligrams per liter (mg/L) Idaho Ground Water Quality standard? The question is assessed by examining the sample characteristics relative to a statistical hypothesis concerning the population's characteristics.

A statistical hypothesis is a statement about a parameter in a population and a **hypothesis test** is a formal procedure for comparing the sample data with a hypothesis whose truth we want to

[††] Measurements that can be used to detect an exceedance. The number of future comparisons is defined on the basis of number of constituents of concern examined per well, number of wells, sampling frequency and duration of the comparison time period. Appendix J provides details.

assess (Moore and McCabe 1998). The results of a test are expressed in terms of a probability that expresses how well the hypothesis agrees with the data.

A hypothesis test involves four steps:

1. State the hypotheses and confidence level: a null hypothesis (H_0) is the statement being tested; an alternative hypothesis (H_A) is the statement we will accept should H_0 be rejected. The significance of the test, α, is the complement of the confidence level ($1-\alpha$) and indicates the strength of the evidence against the null hypothesis. The smaller the α, the less the chance of falsely rejecting H_0.
2. Choose and compute the test statistic. A test statistic provides a quantifiable measure for deciding between H_0 and H_A. Some examples are the t statistic and the F statistic.
3. Find a p-value based on the test statistic. The p-value is the lowest significance level at which H_0 can be rejected (or the probability of obtaining a test statistic as extreme as or more extreme than that calculated from the sample, if H_0 were true).
4. State the conclusions based on a decision rule: (a) if the p-value is less than α, then reject H_0 and accept H_A; (b) if the p-value is greater than or equal to α, then we cannot reject H_0 based on information provided in the data set. Both decisions are made at an α significance level ($1-\alpha$ confidence level).

Two types of errors are associated with any hypothesis test, a Type I (false-positive) error occurs when H_0 is falsely rejected; a Type II (false-negative) error occurs when H_0 is falsely accepted. For example, if the null hypothesis, H_0, asserts that ground water is not contaminated, a Type I error would lead to the claim that contamination exists when it actually does not. A Type II error would lead to the claim that ground water is not contaminated when it actually is. The risk of committing a Type I or Type II error is α or β, respectively, and they are complimentary: specifying a low value of α means accepting a high β; $\alpha = 0.05$ is usually considered an acceptable trade-off between the two risks. Just as ($1-\alpha$) is the confidence level of a hypothesis test; ($1-\beta$) is the **power** of the test (the likelihood of identifying contamination if it is present). Type II errors are more likely for small sample size, so β should be considered at the time of sampling design. Sample size should be large enough to achieve a power of 0.8 or above.

O.6. Sample Size

Sample size affects both estimation and hypothesis testing. For estimation, it determines the estimate's precision, and for hypothesis testing, it affects the power of the test.

Sample size depends on the type of statistical test chosen and also on measurement precision. A large sample size is almost always desirable. Having many observations can make trivial differences detectable. The goal of determining sample size in a statistical study is to find the number of samples that will provide adequate yet practically feasible evidence to draw meaningful conclusions relative to the goals of the study. It is always good practice to state the problem first and then set up decision rules to address the problem.

For ground water analysis, 8 to 12 background samples should be available for determining decision thresholds and for making interwell comparisons. The samples must be statistically independent and representative of seasonal and spatial variability at the site. For this reason, the 8 to 12 samples preferably should be collected quarterly over a 2- to 3-year period in a well. For

interwell comparisons with two upgradient wells reflecting statistically indistinguishable chemistry, 1 year of quarterly data for each well is required (if the two wells' chemistries are different, then 2 years of quarterly data at each well should be available). Statistical analysis can be conducted with smaller data sets, but smaller sample size usually leads to such wide prediction intervals that no meaningful conclusions can be drawn. The statistical requirements of the various analysis methods should be understood so adequate numbers of samples are collected prior to analyzing the data.

O.7. Nonparametric Methods

Before choosing nonparametric methods, it should be recognized that data sets having normal or lognormal distributions should be analyzed with parametric methods. Parametric methods are more powerful because the actual values of the measurements are used in the analysis. Parametric methods assume some knowledge of the shape of the distribution (i.e., normal or lognormal) and use the measured data values to estimate population parameters. For example, the t-test is a parametric method for bell-shaped distributions (either in original scale or transformed scale) that are centered at μ with a dispersion of σ.

Sample distributions that do not have normal or lognormal form can be analyzed with nonparametric methods, which do not require assumptions about the form of the population distribution. The only requirement is that the population distribution be continuously valued; additionally, if two populations are to be compared, then they should have the same shape.

It is common that sample size is inadequate to determine whether a particular distribution is parametric, or that the number of nondetects is too large to determine the form of the distribution. In such cases, nonparametric methods can be used both for establishing background levels and for hypothesis testing. These methods do not require assumptions about the form of the distribution and can be used to estimate parameters or to test a hypothesis.

Some common nonparametric methods used in ground water analysis are based on ranks of the data but not the actual values. Data values are ordered from lowest to highest and ranked according to their position in the ordered list. Commonly used rank analysis methods include the **Wilcoxon signed-rank test** (nonparametric equivalent of the one-sample t-test), **Wilcoxon rank-sum test *or* Mann-Whitney's test** (nonparametric equivalent of the two-sample t-test), **Kruskal-Wallis test** (nonparametric form of the multiple-sample ANOVA test) and **nonparametric regression**. As with parametric methods, statistical independence of the observations is required for all nonparametric methods.

The **Kruskal-Wallis test for seasonality** can be used to test for the presence of significant seasonal fluctuations in a time-series data set. **Mann-Kendall's test** for trend shows if a significant secular trend exists. **Sen's test** estimates the slope of the trend, regardless of the presence of missing observations or variable sampling time intervals.

Nonparametric prediction and tolerance limits are based on the maximum values observed in N background measurements, where sample size depends on confidence level and future comparison strategy (Appendix I and K). Confidence level in turn is a function of the number of future comparisons (k) and the exceedance verification sampling plan. These methods require

very large background sample sizes if k is large or if α is small, so that trade-offs are usually required.

Bootstrap and Jacknife resampling methods are recently developed nonparametric methods for making statistical inference. Basically, the original sample data set is randomly resampled thousands of times and statistics of interest recomputed each time. The calculated statistics from all resampled data sets are used to estimate the relevant sampling distributions. In the Bootstrap method, resampling is conducted with replacement (of size N, the original sample size). In the Jacknife method, resampling systematically leaves out one value from the original data set each time (sample size = N-1). Unfortunately, small sample size is a major limitation because the resampling method assumes that the original data set is representative of the underlying population. These methods are computer-intensive but demonstrate growing potential for environmental statistical analysis. Their technical aspects are beyond the scope of this document. The Idaho Department of Environmental Quality leaves it to the regulated entity to choose the methods that best fulfill the objectives of the statistical analysis but retains the right to ask for alternative methods to be used if they prove to be more appropriate.

References

EPA (United States Environmental Protection Agency). 2009. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Unified Guidance.* Washington, DC: EPA. EPA 530/R-09-007.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*. New York, NY: John Wiley & Sons.

Moore, D. S. and G.P. McCabe. 1998. *Introduction to the Practice of Statistics*. NewYork, NY: W.H. Freeman and Company.

Virginia Department of Environmental Quality. 2003. *Data Analysis Guidelines for Solid Waste Facilities.* Richmond, VA: Virginia DEQ.

This page intentionally left blank for correct double-sided printing.

Appendix P. Summary of Revisions

Updates

Figure 3—Modify flow chart logic for handling nondetects to conform to EPA's guidance.

Figure 4—Statistical assumptions must be consistent with the site conceptual model.

Section 4.4.1—Included Upper Simultaneous Limits, their definition, and usage.

Section 5.2.3—Clarified verification versus confirmation resampling data.

Appendix B, Figure. B1—Replaced with ProUCL box plot example.

Appendix D—Importance of distribution; rearranged Shapiro-Wilk calculation and added ProUCL examples.

Appendix E.3—Removing seasonality reduces data variance and lowers decision thresholds.

Appendix F.2—This section was deleted in favor of a more illustrative ProUCL example.

Appendix L—Corrections and added interim thresholds for trends.

Additions

Section 3.4.1—Limited Annual Sampling (4 equally spaced samples during the sampling season).

Section 4.4.1—Interwell Tolerance Limits (added discussion on Interwell UTL).

Section 4.4.2—Interwell Prediction Limits (clarified to maintain consistency with EPA guidance).

Section 4.4.3—Interwell Simultaneous Limits (to maintain consistency with EPA guidance).

Section 4.5—Verification Resampling (clarifies recommended "1-of-3" retesting process).

Section 4.6—Trending Data (detrending and interim decision threshold procedures).

Appendix D—Figure D1 (formalize the decision logic for choosing a distribution).

Appendix D—Two ProUCL examples (do not use lognormal option for strongly skewed data).

Appendix F.2—ProUCL example (removing a secular trend and its effect on variance).

Appendix K.3—ProUCL example (UPLs, UTLs, and USLs calculated for near-normal data).

Appendix K.4—ProUCL example (impact on decision thresholds of strongly skewed data).

Appendix L.4—Added control chart process (to estimate decision threshold in trending data).

Appendix L, Figure. L.3—Proposed decision logic (to estimate interim threshold for upward data trend).

Appendix O—Defined the gamma distribution (with examples).