Facts Sheet for ProUCL 4.0
A Statistical Software Package for Environmental Applications for Data Sets With and Without Nondetect Observations

Exposure assessment, risk assessment and management, and cleanup decisions at potentially polluted sites are often made based upon the mean concentrations of the contaminants of potential concern (COPCs). Typically, the mean concentration of a COPC at a contaminated site is unknown, and is frequently estimated by the sample mean based upon the data collected from the site areas under investigation. In order to address the uncertainties associated with the estimates of the unknown mean concentrations of the COPCs, appropriate 95% upper confidence limits (UCLs) of the respective unknown means are used in many environmental applications including the estimation of exposure point concentration (EPC) terms. The Technical Support Center (TSC), EPA Las Vegas, NV developed ProUCL Version 3.0 software package (EPA, 2004) to support risk assessment and cleanup decisions at contaminated sites based upon full data sets without nondetect observations (NDs). For data sets without NDs, ProUCL 3.0 has several parametric and nonparametric UCL computation methods as described in the revised EPA UCL Guidance Document for Hazardous Waste Sites (EPA, 2002a).

The Need for an Upgrade of ProUCL 3.0

Nondetect observations are inevitable in most data sets collected from the various environmental applications. The ProUCL 4.0 software package is an upgrade of ProUCL 3.0 software package, and provides several statistical methods that can be used on left censored data sets with nondetect observations potentially having multiple detection limits (DLs). ProUCL 4.0 is especially developed to address the various statistical issues arising in exposure and risk assessment studies, and also in background and site evaluation and comparison applications. All capabilities of ProUCL 3.0 have been retained in ProUCL 4.0. The TSC, EPA Las Vegas has revised the Background Guidance Document for CERCLA sites (EPA, 2002b). The revised background document includes some exploratory graphical methods to pre-process a data set, and a couple of new chapters describing the computations of parametric and nonparametric upper limits that are used to estimate the background level contaminant concentrations or background threshold values (BTVs), and other not-to-exceed values based upon data sets with and without ND observations.

It is noted that the methods to compute upper limits to estimate BTVs and not-to-exceed values are not easily available in any of the available software packages, especially for data sets with nondetect observations. ProUCL 4.0 can be used to compute various parametric and nonparametric upper limits often used to estimate environmental parameters of interest including the EPC terms, BTVs, and other not-to-exceed values. The BTVs and not-to-exceed values are also used for screening of contaminants of potential concern (COPCs). Typically, upper confidence limits (UCLs) are used to estimate the EPC terms; upper prediction limits (UPLs), upper tolerance limits (UTLs), or upper percentiles are used to estimate the BTVs and not-to-exceed values. ProUCL 4.0 can be used to compute those upper limits based upon full uncensored data sets without NDs and left-censored data sets with NDs having multiple DLs.

Additionally, ProUCL 4.0 offers several parametric and nonparametric single sample and two sample hypotheses testing approaches used in background versus site comparison studies. Those hypotheses testing approaches can be used on data sets with NDs and without NDs. ProUCL 4.0

also offers some useful graphical displays including histograms, multiple quantile-quantile (Q-Q) plots, and side-by-side box plots for data sets with and without ND observations. The graphical displays provide additional insight and information contained in data sets that cannot be revealed by the use of estimates (e.g., 95% UCLs) and test statistics such as goodness-of-fit (GOF) test statistics, t-test statistic, Rosner test, and various other statistics. In addition to providing information about the data distributions (e.g., normal, lognormal, gamma), the graphical Q-Q plots are very useful to identify potential outliers and the presence of mixture samples (if any) in a data set. Side-by-side box plots and multiple Q-Q plots are quite useful to visually compare two or more data sets such as site versus background contaminant concentrations, monitoring well (MW) concentrations, and so on. Therefore, it is desirable and suggested that the conclusions derived using estimates (e.g., 95% UCL) and test statistics (e.g., t-test) should always be supplemented with graphical displays.

ProUCL 4.0 serves as a companion software package for the *UCL* Computation Guidance Document for Hazardous Waste Sites (EPA, 2002a) and the Background Guidance Document (currently under revision) for CERCLA Sites (EPA, 2002b). Most of the statistical and graphical methods described and recommended in these two EPA guidance documents have been incorporated in ProUCL 4.0. It should be noted that ProUCL 4.0 also has some parametric and nonparametric single sample hypotheses approaches that may be used to compare site mean concentrations (or some site threshold value such as an upper percentile) with some average cleanup standards, C_s (with a not-to-exceed limit, A_0) to verify the attainment of cleanup levels (EPA, 1989, and EPA, 2006) after some remediation activities have been performed at potentially impacted site areas. Several of the statistical methods as incorporated in ProUCL 4.0 can be used in groundwater (GW) monitoring applications (EPA, 1992).

Two reference guides: 1) ProUCL 4.0 User Guide and 2) ProUCL 4.0 Technical Guide have also been developed for ProUCL 4.0 software package. The User Guide describes and illustrates the uses of the various menu items and options as incorporated in ProUCL 4.0. The ProUCL 4.0 Technical Guide describes the theory (with references) behind the statistical methods as incorporated in ProUCL 4.0. These two documents can be downloaded from the EPA website for ProUCL 4.0. ProUCL 4.0 also provides Online Help for the various methods available in ProUCL 4.0.

Data Requirements

Statistical methods (e.g., upper limits) as incorporated in ProUCL 4.0 (and also in other software packages such as SAS and Minitab) assume that the user has collected an adequate amount of data of good quality, perhaps using appropriate data quality objectives (DQOs) as described in EPA, 2006. However, many times (e.g., using the available historical data, or due to budgetary and time constraints), it may not be possible to collect data sets based upon specified performance measures (e.g., decision errors) and other DQOs. It is noted that many times, administrators and decision makers do not want to collect many samples, especially background samples. Therefore, when it may not be possible to collect adequate amount of data using DQOs (EPA, 2006), Chapter 1 of the two ProUCL 4.0 reference guides can be used to determine the minimum sample size requirements associated with the various estimation and hypotheses testing approaches available in ProUCL 4.0. The suggested minimum sample size requirements as described in Chapter 1 are made based upon the practical applicability of the procedures incorporated in ProUCL 4.0. Those suggestions are particularly useful when the data are sparse and it may not be feasible to collect additional data based upon DQOs. However, it should be

pointed out that for more accurate (reduced bias) estimates and reliable (increased precision) results, whenever possible, it is desirable to collect adequate amount of data, perhaps using DQOs with specified performance measures.

A partial listing of the statistical and graphical methods as incorporated in ProUCL 4.0 is given as follows. The details of the various statistical and graphical procedures with illustrating examples can be found in the User Guide and the Technical Guide associated with ProUCL 4.0.

ProUCL Version 4.0 Capabilities

All of the capabilities of ProUCL 3.0 have been retained in ProUCL 4.0. It is anticipated that ProUCL 4.0 will serve as a companion software package for: 1) *UCL* Computation Guidance Document for Hazardous Waste Sites (EPA, 2002a), and 2) Background Guidance Document (currently under revision) for CERCLA Sites (EPA, 2002b). Several statistical and graphical methods for data sets with and without ND observations have been incorporated in the upgraded ProUCL 4.0 software package. Some of those capabilities are listed in the following paragraphs.

Group Option

ProUCL 4.0 provides a "Group" option. An appropriate Group-ID variable representing the various groups such as different site areas of concern (AOC) or monitoring wells (MWs) should be available in the data sheet. Using this option, graphical displays and statistical analyses can be performed separately for each of the group represented by the Group-ID variable. This group graph option is very useful to perform visual multiple comparison (multiple Q-Q plots, side-by-side box plots) of the various groups (e.g., AOCs, MWs) identified by the Group-ID variable. The details of this option are given in ProUCL 4.0 User Guide.

Graphical Methods

ProUCL 4.0 has several graphical methods including multiple quantile-quantile (Q-Q) plots, side-by-side box plots, and histograms. These graphical methods can be used on data sets with and without nondetect observations. A typical Q-Q plot (normal, gamma, lognormal) is often used to visually assess the data distribution of the COPCs. A Q-Q plot also provides important information about presence of potential outliers and multiple populations that may be contained in a data set. For data sets with NDs, ProUCL 4.0 can be used to generate Q-Q plots based upon regression on order statistics (ROS) methods including the robust ROS method. The graphical displays of multiple Q-Q plots and side-by-side box plots are useful to visually compare the concentrations of two or more populations, some of which are listed as follows:

- Site versus background populations (areas)
- Surface versus subsurface concentrations
- Concentrations of two or more AOCs or MWs

Goodness-of-Fit (GOF) Test Methods

ProUCL 4.0 has GOF tests for normal, lognormal, and gamma distributions for data sets with and without nondetect observations. The following GOF tests to assess normality or lognormality of a data set are available in ProUCL 4.0.

GOF Tests to Assess Normality or Lognormality for Full Data Sets without ND Observations

- Informal graphical Q-Q plot (normal probability plot) and histogram.
- Shapiro-Wilk (SW) test for sample sizes less than or equal to 50.
- Lilliefors test for larger sample sizes such as greater than 50.

GOF Tests to Assess Normality or Lognormality for Left-Censored Data Sets with NDs and Multiple Detection Limits

- ProUCL 4.0 can be used to perform normal GOF tests as mentioned above (for full data) on data sets consisting of ND values. Specifically, normal or lognormal GOF tests can be performed using detected data values only.
- The normal or lognormal GOF tests can also be performed on data sets (detected values and extrapolated NDs) obtained using one of the regression on order statistics (ROS) methods. The details of constructing Q-Q plots (normal and lognormal) and performing ROS on data sets with multiple DLs are given in ProUCL 4.0 Technical Guide.
- The three ROS methods available in ProUCL 4.0 are the normal ROS, lognormal ROS (also known as robust ROS), and Gamma ROS methods.
- ProUCL 4.0 can be used to generate additional columns (with suitable headings assigned by ProUCL 4.0) of data consisting of the detected data and extrapolated nondetect data.

<u>Goodness-of-Fit Test for Gamma Distribution</u>

Gamma GOF Tests for Full Data Sets without ND Observations

- Informal graphical quantile-quantile (Q-Q) plot (gamma probability plot) and histogram.
- Kolmogorov-Smirnov test for sample sizes in the range 4-2500 (critical values computed using Monte Carlo simulations) and values of the estimated shape parameter, k, in the interval [0.01, 100.0].
- Anderson-Darling test for sample sizes in the range 4-2500 (critical values computed using Monte Carlo simulations) and values of the estimated shape parameter, k, in the interval [0.01, 100.0].

Gamma GOF Tests for Left-Censored Data Sets with NDs and Multiple Detection Limit

- ProUCL 4.0 can be used to perform gamma GOF tests on data sets consisting of ND values. Specifically, gamma GOF tests can be performed on data set consisting of only detected data.
- The gamma GOF tests listed above can also be used on data sets (detected values and extrapolated NDs) obtained using one of the regression on order statistics (ROS) methods as incorporated in ProUCL 4.0. The details of constructing gamma Q-Q plots and performing ROS on data sets with multiple detection limits are given in ProUCL 4.0 Technical Guide.

Summary Statistics

- For full data sets without NDs, ProUCL computes and lists all relevant descriptive summary statistics for raw and log-transformed data.
- For data sets with NDs, ProUCL computes simple summary statistics using only detected data values for raw or log-transformed data.

Note: Summary statistics option does not compute and lists the estimates of the population parameters. Those estimates are computed and listed by the 'UCL' and 'Background' options of ProUCL 4.0.

Estimates of Population Parameters

- Computes the maximum likelihood estimates (MLEs) and minimum variance unbiased estimates (MVUEs) of the various population parameters such as the mean, standard deviation, quantiles, coefficient of variation (CV), skewness, and also the MLEs of the shape parameter k and scale parameter θ of a gamma distribution. These estimates (e.g., MLE, MVUE) are shown when the menu items Background and UCL are used to compute the upper limits.

- For data sets with NDs, ProUCL 4.0 also computes parametric (e.g., normal MLE) and nonparametric (Kaplan Meier (KM), Bootstrap) estimates of population mean, variance, and standard error of the mean. These statistics do not represent simple summary statistics. Therefore, these estimates (e.g., MLE, KM) are shown when the menu items Background and UCL are used to compute the upper limits.

Upper Confidence Limits (UCLs) to Estimate Exposure Point Concentration Terms

A 95% UCL of the unknown population arithmetic mean, μ_1, of a COPC is used to estimate the EPC term and also to determine the attainment of cleanup standards. It should be noted that gamma distribution is often better suited to model positively skewed environmental data sets than the lognormal distribution. For positively skewed data sets, the default use of a lognormal distribution often results in impractically large UCLs, especially when the data sets are small (Singh, Singh, and Iaci, 2002). In order to obtain accurate and stable UCLs of practical merit, other distributions such as a gamma distribution should be used to model positively skewed data sets. ProUCL, Version 4.0 has procedures to perform the gamma goodness-of-fit tests and to compute UCLs of the population mean, and various other limits based upon gamma distributed data sets with and without nondetect observations. ProUCL 4.0 also has several bootstrap methods (e.g., percentile bootstrap, bias corrected bootstrap, bootstrap-t) to compute UCLs of the mean for data sets with and without ND observations.

For full data sets without NDs and for left-censored data sets with ND observations, ProUCL 4.0 can compute several parametric and nonparametric UCLs with a confidence coefficient (*CC*) specified in the interval [0.5, 1.0) including the commonly used *CC* level 0.95. ProUCL 4.0 can compute parametric UCLs for normal, lognormal, and gamma distributions. It is noted that in environmental applications (e.g., estimation of EPC), a 95% UCL of mean is used, therefore, ProUCL makes recommendations only for an appropriate 95% UCL (s) which may be used to

estimate the EPC term. The basis and theoretical justification for those recommendations are summarized in Singh and Singh (2003) for full data sets without ND observations.

UCLs for Full Uncensored Data Sets without ND Observations

1. Student's-t UCL: to be used for normally (or at least approximately normally) distributed data sets. Student's-t UCL is available for all confidence coefficients, (1-α) in the interval [0.5, 1.0).
2. Approximate Gamma UCL: to be used for gamma distributed data and is typically used when k hat (ML estimate of the shape parameter, k) is greater than or equal to 0.5. Approximate gamma UCL is available for all confidence coefficients (1-α) in the interval [0.5, 1.0).
3. Adjusted Gamma UCL: to be used for gamma distributed data sets and should be used when k hat is greater than 0.1 and less than 0.5. Adjusted gamma UCL is available only for three confidence coefficients: 0.90, 0.95, and 0.99.
4. H-UCL based upon Land's H-statistic: to be used for lognormally distributed data sets. In ProUCL, H-UCL is available only for two confidence coefficients: 0.90 and 0.95. ProUCL can compute H-UCL for samples of size up to 1001.
   Caution: For highly skewed data sets, the use of H-UCL should be avoided as the H-statistic often results in unrealistically large, impractical and unusable H-UCL values. ProUCL provides warning messages and recommends the use of alternative UCLs for such highly skewed lognormally distributed data sets.
5. Chebyshev (MVUE) UCL: to be used for lognormally distributed data sets. This UCL computation method uses the MVU estimates of the standard deviation of the mean and of other parameters of a lognormal distribution. Chebyshev (MVUE) UCL is available for all confidence coefficients, (1-α) in the interval [0.5, 1.0).
6. Central Limit Theorem (CLT) based UCL: to be used when the sample size is large.
7. Adjusted-CLT (adjusted for skewness) UCL: may be used for mildly skewed data sets of large sizes.
8. Modified-t statistic (Adjusted for skewness) based UCL: may be used for mildly skewed data.
   Caution: UCLs listed in 6, 7, and 8 do not provide adequate (e.g., 95%) coverage when the data are moderately to heavily skewed, even when the sample size is large such as greater than 50.
9. Chebyshev (Mean, Sd) UCL: based upon the sample mean and standard deviation, Sd.
10. Jackknife UCL for mean (same as Student's-t UCL).
11. Standard Bootstrap UCL.
12. Bootstrap-t UCL.
13. Hall's Bootstrap UCL.
14. Percentile Bootstrap UCL.
15. Bias-corrected accelerated (BCA) Bootstrap UCL.

UCLs Based Upon Left Censored Data Sets with ND Observations

In order to compute UCLs, one has to first obtain estimates of population mean, standard deviation, and standard error of the mean based upon data sets with single or multiple detection limits. ProUCL 4.0 has a couple of estimation methods such as the ROS methods and Kaplan-Meier (KM) method that can handle multiple detection limits. The following methods for estimation of population mean and the standard deviation have been incorporated in ProUCL 4.0.

- Maximum likelihood method (MLE) (Cohen (1991)) – Single DL
- ROS Methods for normal, gamma, and lognormal distributions –Multiple DLs
  Note: ProUCL 4.0 can be used to generate columns consisting of detected data and extrapolated NDs obtained using a ROS method (normal, lognormal, and gamma).
- Kaplan-Meier (KM) method (Kaplan-Meier (1958)) – Multiple DLs
- Winsorization method
- DL/2 substitution (DL/2) method – not a recommended method. *The DL/2 method is included for historical reasons only.*

Note on the Use of DL/2 and Other Substitution Methods

- The use of DL/2 (and DL) method is not recommended in statistical procedures that may be used in decision-making processes. Therefore, it is suggested to avoid the use of the DL/2 method (and other substitution methods such replacement of NDs by '0', 'DL') to estimate the EPC terms and BTVs.
- Also, the use of the substitution methods is not recommended in hypothesis testing approaches.
- However, the substitution methods such as the DL/2 method may be used in graphical and exploratory methods to gain visual information about the data distributions and outliers. Several graphical methods (e.g., boxplots, Q-Q plots) based upon DL/2 method are available in ProUCL 4.0.

ProUCL 4.0 can compute several parametric and nonparametric UCLs with a confidence coefficient (*CC*) specified in the interval [0.5, 1.0) including the commonly used *CC* level 0.95. However, since in most environmental applications (e.g., estimation of EPC), a 95% UCL of mean is used, therefore, ProUCL 4.0 makes recommendations for the most appropriate 95% UCL (s) that may be used to estimate the EPC terms based upon data sets with ND observations. The theory behind those recommendations can be found in Singh, Maichle, and Lee (EPA, 2006). Using the estimates of mean and standard deviation, or extrapolated NDs obtained using one of the ROS methods listed above, ProUCL 4.0 computes UCLs of the means using the following methods.

- Tiku's UCL method (Tiku (1967 and 1971)) – Single DL
- Ad hoc UCL methods using Student's t-statistic on ML estimates and KM estimates
- Ad hoc UCL methods based upon Land's H-statistic – Single DL
- Gamma UCL – Bootstrap UCL on gamma ROS
- Nonparametric Chebyshev UCL based upon KM estimates
- Bootstrap (percentile, standard bootstrap, bootstrap t, and bias-corrected accelerated (BCA)) methods on ROS methods and KM estimates.

Upper Limits to Estimate Background Level Threshold Values (BTVs) or Not-to-Exceed Values

ProUCL 4.0 can be used to compute several parametric and nonparametric upper limits that are used to estimate the BTVs or not-to-exceed values for data sets with NDs and without NDs. These upper limits include: upper prediction limits (UPLs), upper tolerance limits (UTLs), and upper percentiles. Some of the nonparametric methods such as the Kaplan-Meier (Meier, 1958) method and ROS methods are applicable on left-censored data sets having multiple detection limits. The background statistics as incorporated in ProUCL 4.0 are particularly useful when individual site observations from some impacted site areas (perhaps after some remediation activities) are to be compared with BTVs to determine if adequate amount of remediation and cleanup has been performed yielding remediated site concentrations comparable to background level concentrations; that is if the site concentrations can be considered as coming from (or approaching to) the population of background concentrations.

The process of comparing individual site observations with BTVs or some other not-to-exceed values is also used for screening purposes (e.g., before performing any cleanup and assessment) to identify the COPCs, and to determine if site areas under study need further sampling and remediation actions. Specifically, the process of comparing onsite data with the BTVs may help the working crew, project team, or the decision makers to take immediate decisions if more remediation and more onsite sample collection need to be performed at the site areas under investigation.

The first step in establishing site specific background level contaminant concentrations for site related hazardous pollutants is to perform background sampling to collect appropriate number of samples from the designated site specific background areas or some agreed upon site reference areas. An appropriate DQO process (EPA, 2006) may be followed to collect an adequate number of background samples. It is desirable to collect at least 10-15 background samples to compute reliable estimates of BTVs. Furthermore, it is suggested not to use estimated BTVs and not-to-exceed values based upon background data sets of sizes smaller than 8-10. Once, an adequate amount of background data have been collected, the next step is to determine the data distribution. This can be achieved by using exploratory graphical tools (quantile-quantile (Q-Q) plots and histograms) as well as formal GOF tests as incorporated in ProUCL 4.0.

Once the data distribution of a background data set has been determined, one can use parametric or nonparametric statistical methods to compute background statistics. A review of the environmental literature reveals that one or more of the following statistical limits are used to compute the background statistics; that is to determine and estimate background level contaminant concentrations. Collectively, these statistics represent estimates of the background threshold values (BTVs). The BTVs are estimated by statistics representing values in the upper tail (e.g., 95% upper percentile, 95% UPL) of the background data distribution. Typically, a site observation (preferably based upon a composite sample) in exceedance of a BTV (e.g., UPL, upper percentile) can be considered as coming from a site area (location), which might have been impacted by the site-related activities. In other words, such a site observation may be considered as exhibiting some evidence of contamination at that site area (location) due to site related activities. For data sets with NDs, the BTVs can be estimated using upper limits based upon KM

estimates. Some of the statistical limits used to estimate the BTVs for data sets with and without NDs as incorporated in ProUCL 4.0 are listed as follows.

1. Upper Percentiles (e.g., 95%, 99%) for data sets without and with NDs (e.g., based upon KM estimates)
2. Upper Prediction Limit (UPL) for a future (site observation) observation (using KM or other estimates for data sets with NDs)
3. UPL for future k (e.g., next k or k site observations) observations
4. Upper Tolerance limits (UTLs) - Upper Confidence Limits for Upper Percentiles
5. Upper percentiles, UPLs, UTLs based upon data obtained using ROS methods – data with NDs
6. IQR Upper Limit (upper end of the upper whisker in a Box Plot)
7. UPL and UTL based upon resampling bootstrap
8. UPL based upon Chebyshev inequality
9. UTL based upon bootstrap methods for data sets with NDs
10. BTVs using nonparametric methods based upon higher order statistics (Conover, 1999)

Note: *The behavior of the exploratory IQR based upper limit as an estimate of a BTV is not well studied. This limit should be used with caution to estimate the BTVs or not-to-exceed values.*

It should be noted that background versus site comparisons based upon the BTVs are performed when not enough site data (e.g., < 4-6 observations) are available to perform traditional two sample comparisons using hypotheses testing approaches such as t-test, Wilcoxon Rank Sum test, and Gehan test. When enough site data are available (e.g., at least 8-10, more are preferable), it is preferable to use hypotheses testing approaches to compare site data with BTVs or not-to-exceed values. Thus, in the absence of adequate amount of site data, individual point-by-point site observations are compared with some BTVs to determine the presence or absence of contamination due to site related activities. This method of comparing site versus background level contamination is particularly helpful to use after some sort of remediation activities have taken place at the site; and the objective is to determine if the remediated site areas have been remediated enough to the background level contaminant concentrations.

Typically, a site observation (possibly based upon composite samples) in exceedance of a background threshold value can be considered as coming from a contaminated site area that may have been impacted by the site-related activities. In other words, such a site observation may be considered as exhibiting some evidence of contamination at the site due to site related activities. In case of an exceedance of the BTV by a site location, some practitioners like to verify the possibility of contaminated site location by re-sampling (collecting 2-3 additional samples) that location, and comparing the sampled value(s) with the BTV.

<u>Hypothesis Testing Approaches</u>

Both single sample and two sample parametric and nonparametric hypotheses testing approaches are available in ProUCL 4.0. The hypotheses testing approaches as incorporated in ProUCL 4.0 can be used on full data sets without any ND observations, and on left-censored with nondetect

data values. Form 1, Form 2, Form 2 with substantial difference, and two-sided alternative hypotheses approaches (EPA, 2002b) are available in ProUCL 4.0. It is desirable to collect adequate amount of data of good quality from the populations under investigation using appropriate DQOs (EPA, 2006). In case, data sets cannot be collected using DQOs, it is suggested to follow the minimum sample size requirements as described in Chapter 1 of the ProUCL Technical Guide and User Guide. Some single sample and two sample hypotheses testing approaches as available in ProUCL 4.0 are listed as follows.

Single Sample Hypotheses Testing Approaches

One Sample t-Test: Based upon the sampled site data, this test is used to compare the site mean, μ, with some specified cleanup standard, $C_{s,}$ where the cleanup standard, $C_{s,}$ represents an average threshold value, say μ_0. The Student's t- test (or a UCL of mean) is often used (assuming normality of site data or when site sample size is large such as larger than 30, 50) to determine the attainment of cleanup levels at a polluted site, perhaps after some remediation activities. This test should be used on data sets without any ND observations.

One Sample Sign Test or Wilcoxon Signed Rank (WSR) Test: These two tests are nonparametric tests and can also handle nondetect observations provided all nondetects (e.g., associated detection limits) fall below the specified threshold value, C_s. These tests are used to compare the site location (e.g., median, mean) with some specified cleanup standard, $C_{s,}$ representing the similar location parameter.

One Sample Proportion Test or Percentile Test: When a specified cleanup standard, $A_{0,}$ such as a PRG or a BTV represents an upper threshold value (e.g., not-to-exceed value, compliance limit) of a contaminant concentration distribution rather than the mean or median concentration value, μ_0, of the contaminant concentration distribution, then a test for a proportion or a percentile (equivalently a UTL 95%-95% or UTL 95%-90%) may be used to compare the site proportion, P, of exceeding (by site observations) the threshold value, A_0 with some pre-specified proportion, P_0, of exceedances of A_0 by site observations. This test is especially useful when the data set consists of many ND observations. However, this test also assumes that all ND observations lie below the Compliance Limit, A_0.

Two Sample Hypotheses testing Approaches

Typically, two sample hypotheses testing approaches are used for site versus background comparisons, for comparisons of two or more site areas of concern (AOCs), or for comparison of contaminant concentrations of two or more monitoring wells (MWs), provided enough data are available from each population under evaluation. Two sample hypotheses testing approaches as incorporated in ProUCL 4.0 are listed as follows.

1. Student's Two Sample t-Test to compare means - with equal dispersions - Parametric Test
2. Satterthwaite Two Sample t-Test to compare means - with unequal dispersions - Parametric Test

3. F Test to compare two variances (dispersions) – Parametric Test
4. Wilcoxon-Mann-Whitney (WMW) Test to compare two locations, comparability of two continuous distributions – Nonparametric Test
5. Quantile Test to compare the upper tails of two continuous distributions - Nonparametric Test
6. Gehan Test to compare two locations - Nonparametric Test

T-tests and F-test assume normality of the data sets under comparison. Some details of these approaches are described in ProUCL 4.0 Technical Guide. It should be noted that Gehan test, WMW test and Quantile test are also available for data sets with NDs. Gehan's test is specifically meant to be used on data sets with multiple detection limits. The Quantile test is a nonparametric test and is useful to detect a shift in the right tail of the site data distribution. The Quantile test when used in parallel with the Wilcoxon Mann Whitney (WMW) test provides the user with stronger evidence to make decisions about the comparability of site and background distributions, leading to more reliable conclusions whether the site has attained remediation levels or not. It is suggested that for best results, both WMW test and Quantile tests should be used on the same data set.

Note on Comparability of Data Sets

The samples collected from the two (or more) populations under comparisons should all be of the same type obtained using similar analytical methods and apparatus. In other words, the collected site and background samples should be all discrete or all composite (obtained using the same number of discrete samples, same design and pattern), and be collected from the same medium (soil) at similar depths (e.g., all surface samples or all subsurface samples) and time (e.g., during the same quarter in groundwater applications) using comparable (preferably same) analytical methods. Some good soil sample collection methods and sampling strategies are described in EPA, 2003 guidance document.

Note on Influence of Outliers and Use of Lognormal Distribution

Typically, in environmental data sets collected from impacted sites or monitoring wells (MWs), an outlier represents an observation coming from a potentially contaminated site location. This is especially true, when the data are collected from a site specific background area. The outlying observations need to be identified before computing the background statistics (and other estimates and test statistics) as outliers when present distort all statistics of interest, which in turn may lead to incorrect remediation and cleanup decisions for the site under investigation. For an example, inclusion of an outlier may distort the t-test statistic resulting in distorted and incorrect decision errors (Type 1 or Type 2 errors), which can lead to incorrect conclusion about the hypotheses testing. The incorrect decisions may adversely affect the human health and the environment.

The main objective of using a statistical procedure is to model the majority of the data representing the main dominant population, and not to accommodate a few low probability outliers that may yield inflated and impractical statistics, results, and incorrect conclusions. For

an example, background threshold values (BTVs) and exposure point concentration (EPC) terms should be estimated by reliable statistics (and not distorted statistics) obtained using data sets representing the main dominant population under study (e.g., site, background). The low probability high outlying values contaminate the underlying left-censored or uncensored full data set from the population under study. The inclusion of outliers in a background data set needs to be justified before performing other relevant statistical analyses including the estimation of BTVs. If possible, all interested parties should be involved in decision making about the disposition (inclusion or exclusion) of outliers in a background data set. Typically, outlying locations (if any) with elevated concentrations need separate investigation.

It should be noted that the objective is to compute reliable background statistics based upon the majority of a defensible background data set representing the dominant background population. In the process of estimating the BTVs, it may not be desirable to accommodate a few low probability outlying observations (if any) by using a lognormal distribution (Singh, Singh, and Iaci, 2002). The use of a lognormal distribution often accommodates outliers and multiple populations, which in turn yields inflated UCLs and background statistics such as UPLs, percentiles, and UTLs.

The proper identification of multiple outliers is a complex issue based upon robust statistical methods, and is beyond the scope of ProUCL 4.0. For details of the robust outlier identification procedures, refer to Barnett and Lewis (1994), and Singh and Nocerino (1995). A more complicated problem arises when the collected background data set may represent a potentially mixture data set including observations from some of the site areas. The occurrence of mixture samples is quite common in many environmental applications. This is especially true when data sets are collected from large federal facilities (e.g., Navy Sites). For such cases, the underlying data set may consist of samples from the background areas as well as from some other potentially contaminated site areas. In this situation, first, one has to separate the background observations from the other site related observations. After the background data set has been properly extracted from a potentially a mixture sample, one can proceed with the computation of background statistics as available in ProUCL 4.0.

Appropriate population partitioning techniques (e.g., see Singh, Singh, and Flatman (1994)) can be used to extract a background data set from a potentially mixture data set. However, the population partitioning methods are beyond the scope of ProUCL 4.0. It should be noted that some of those methods will be available in Scout (EPA, 2000) software which is currently under revision and upgrades. For methods as incorporated in ProUCL, it is assumed that one is dealing with a sample from a "single" population representing a valid site-related background data set. Therefore, before using statistical methods to compute the various limits such as UCLs, UTLs, and UPLs, it is suggested that the user pre-processes the data set to identify potential outliers and mixture populations (if any).

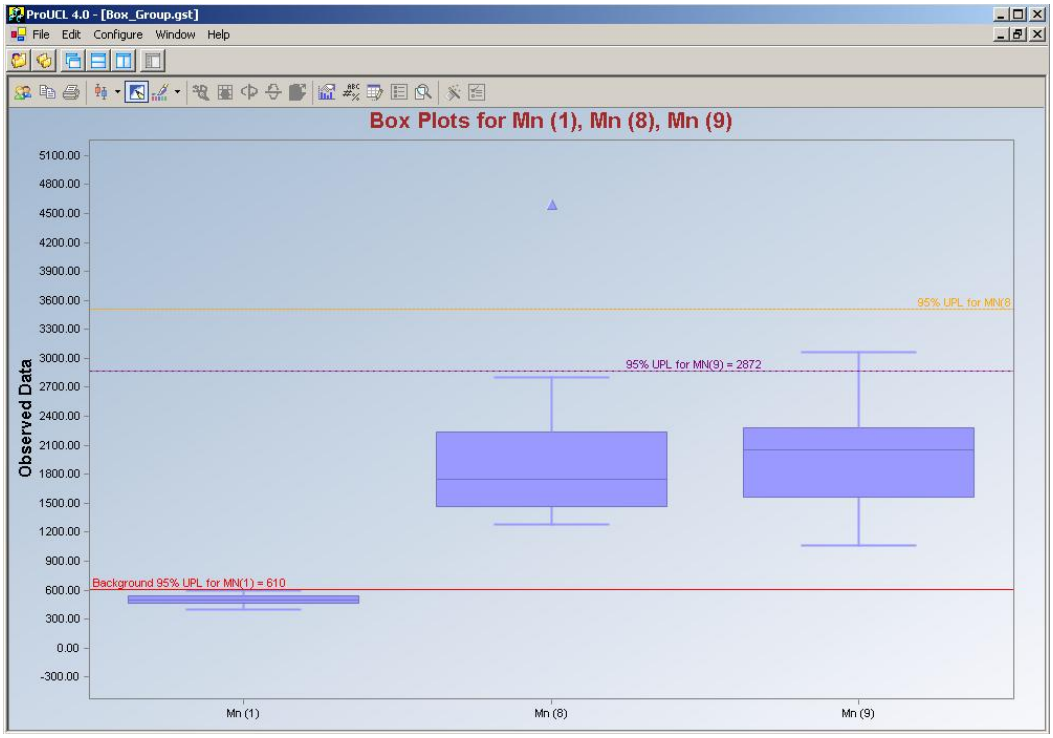
Outlier Tests

ProUCL 4.0 has a couple of classical outlier test procedures, such as the Dixon test and the Rosner test. Additionally, ProUCL 4.0 software has exploratory graphical methods including

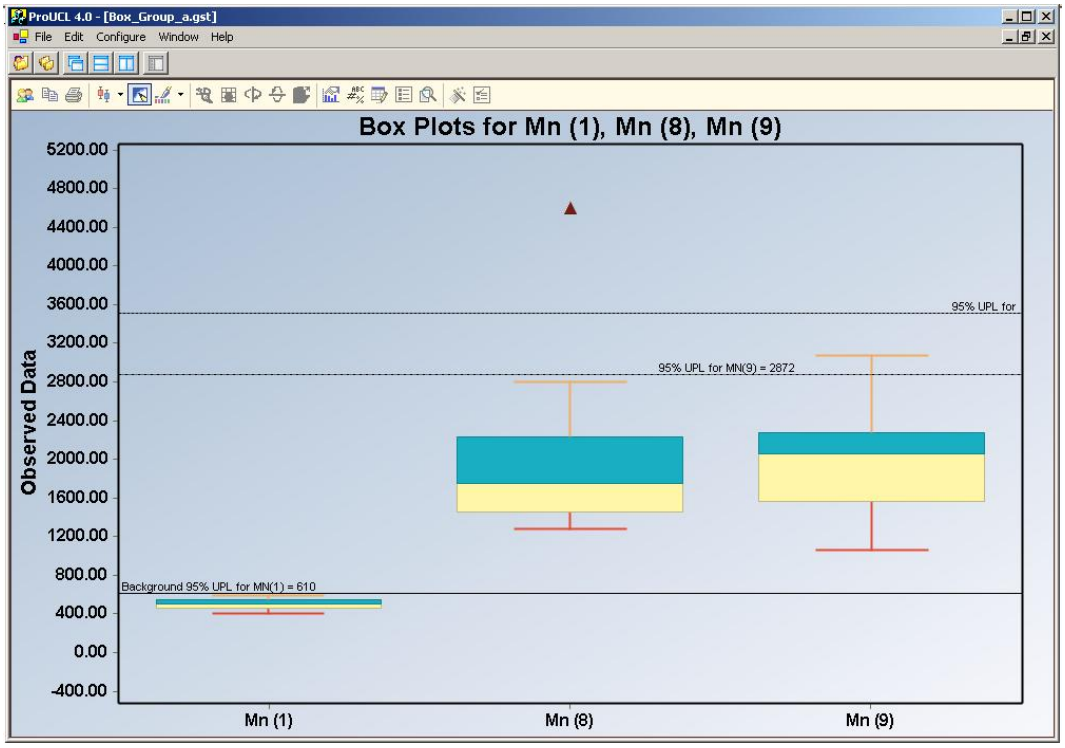
quantile-quantile (Q-Q) plots, box plots, and histograms. The graphical displays of Q-Q plot and box plot are also useful to visually identify outliers that may be present in a data set. It is noted that, the classical test statistics such as the Dixon test and the Rosner test get distorted by the presence of the same outlying observations that those tests are supposed to identify. Therefore, those test statistic (Dixon and Rosner) results should always be supplemented by the graphical displays to confirm the presence of outliers (and potential multiple populations) in a data set. Alternately, the use of robust and resistant outlier identification methods (Singh and Nocerino, 1995) is recommended to identify outliers. The robust outlier identification methods are beyond the scope of ProUCL 4.0.

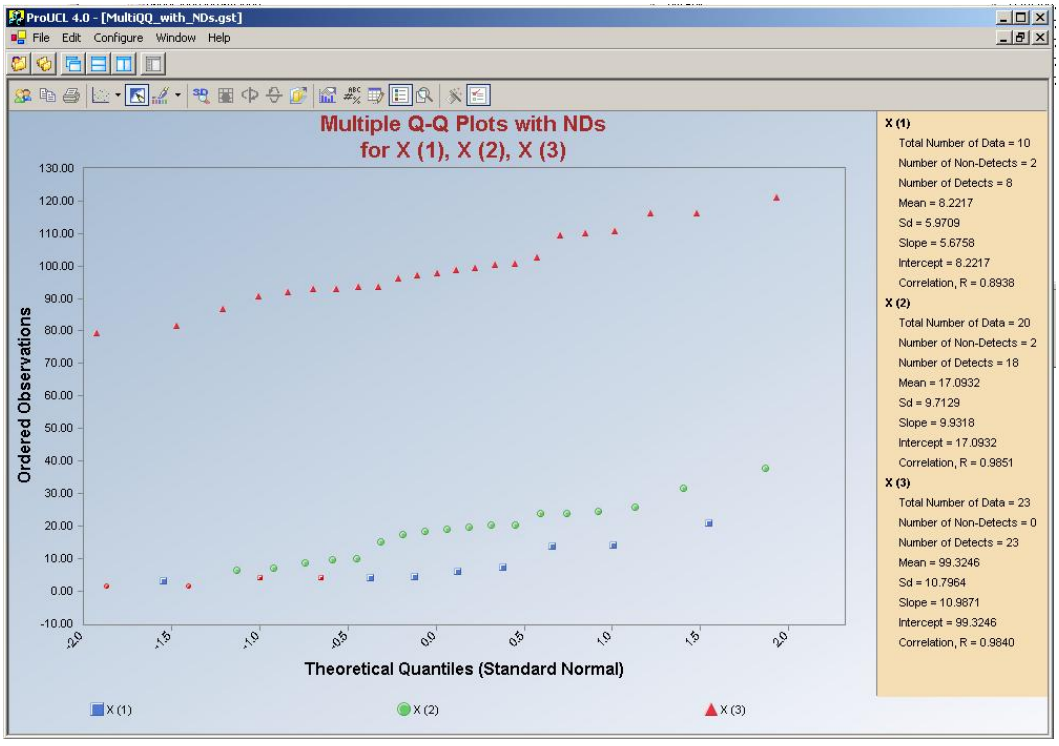
The proper disposition of outliers to include or not to include outliers in the computation of various statistics should be determined by the project team, site experts, and the decision makers involved in the project. In an effort to determine the influence of outliers on the statistics of interest, it is suggested to compute the various statistics based upon data sets with and without the outliers. This extra step should help the project team in determining the proper disposition of outliers. These issues have also been discussed in detail in ProUCL 4.0 Technical Guide.

Screen Shots Generated By ProUCL 4.0



(from MW89)





From (censor by grps)

X (1)

Some Non-Parametric Statistics	
Number of Valid Samples	10
Number of Unique Samples	9
Minimum	3.202
Maximum	20.777
Second Largest	14.138
Mean	8.2217
First Quartile	4
Median	5.347
Third Quartile	13.98575
SD	5.9709101
Variance	35.651767
Coefficient of Variation	0.7262379
Skewness	1.2868074
Mean of Log-Transformed data	1.9005085
SD of Log-Transformed data	0.6522793

Data Follow Appr. Gamma Distribution at 5% Significance Level

Non-Parametric Background Statistics	
90% Percentile	20.1131
95% Percentile	20.777
99% Percentile	20.777

95% UTL with 90% Coverage	
Order Statistic	10
Achieved CC	1
UTL	20.777
95% BCA Bootstrap UTL with 90% Coverage	20.777
95% Percentile Bootstrap UTL with 90% Coverage	20.777
95% UPL	20.777
95% Chebyshev UPL	35.518622
Upper Limit Based upon IQR	28.964375

Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV

X (2)

Some Non-Parametric Statistics	
Number of Valid Samples	20
Number of Unique Samples	19
Minimum	1.5
Maximum	37.867
Second Largest	31.565
Mean	17.09315
First Quartile	8.787
Median	18.794
Third Quartile	23.9455
SD	9.7128685
Variance	94.339814
Coefficient of Variation	0.5682316
Skewness	0.1567791
Mean of Log-Transformed data	2.5790218
SD of Log-Transformed data	0.8911249

Data appear Normal at 5% Significance Level

Non-Parametric Background Statistics	
90% Percentile	30.9882
95% Percentile	37.5519
99% Percentile	37.867

95% UTL with 90% Coverage	
Order Statistic	20
Achieved CC	1
UTL	37.867
95% BCA Bootstrap UTL with 90% Coverage	31.565
95% Percentile Bootstrap UTL with 90% Coverage	37.867
95% UPL	37.5519
95% Chebyshev UPL	60.476088
Upper Limit Based upon IQR	46.68325

Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV

From (censor by grps)

Computer Requirements to Operate ProUCL 4.0

Minimum Hardware Requirements

- Intel Pentium 1.0 GHz
- 50 MB of hard drive space
- 512 MB of memory (RAM)
- CD-ROM drive
- Windows 98 or newer. ProUCL was thoroughly tested on NT-4, Windows 2000, and Windows XP Operating Systems. Limited testing has been conducted on Windows ME.

Software Requirements

ProUCL 4.0 has been developed in the Microsoft .NET Framework using the C# programming language. As such, to properly run ProUCL 4.0, the computer using the program must have the .NET Framework pre-installed. The downloadable .NET files can be found at one of the following two Web sites:

- http://msdn.microsoft.com/netframework/downloads/updates/default.aspx
  Note: *Download .Net version 1.1*

- http://www.microsoft.com/downloads/details.aspx?FamilyId=262D25E3-F589-4842-8157-034D1E7CF3A3&displaylang=en

The first Web site lists all of the downloadable .NET Framework files, while the second Web site provides information about the specific file (s) needed to run ProUCL 4.0. Download times are estimated at 57 minutes for a dialup connection (56K), and 13 minutes on a DSL/Cable connection (256K).

Installation

ProUCL 4.0 can be downloaded from TSC website at http://www.epa.gov/nerlesd1/tsc/tsc.htm. The same website can be used to download ProUCL 4.0 User Guide, Technical Guide and Factsheet. The website contains download and usage instructions.

Find More Information About ProUCL

The TSC website at http://www.epa.gov/nerlesd1/tsc/tsc.htm provides additional information. EPA technical issue papers used in the development of ProUCL are also available at the TSC website. For additional information, contact:

Felicia Barnett, (HSTL)
US EPA, Region 4
61 Forsyth Street, S.W.
Atlanta, GA 30303-8960
barnett.felicia@epa.gov
(404) 562-8659
Fax: (404) 562-8439

References

U.S. Environmental Protection Agency (EPA). 1989. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*, EPA 230/02-89-042, Washington DC.

USEPA 1992. *Statistical Analysis of Ground-water Monitoring Data at RCRA Facilities*. Addendum to Interim Final Guidance. Washington DC: Office of Solid Waste. July 1992.

U.S. Environmental Protection Agency (EPA) 2002a. *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites*, OSWER 9285.6-10.

USEPA. 2002b. *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites*. EPA 540-R-01-003-OSWER 9285.7-41. September 2002.

U.S. Environmental Protection Agency (EPA) 2004. *ProUCL Version 3.1, a Statistical Software*, National Exposure Research Lab, EPA, Las Vegas Nevada. ProUCL 3.0 can be freely downloaded from the EPA website: http://www.epa.gov/nerlesd1/tsc/tsc.htm

USEPA. 2006. *Data Quality Assessment: Statistical Methods for Practitioners,* EPA QA/G-9S. EPA/240/B-06/003. Office of Environmental Information, Washington, D.C. Download from: http://www.epa.gov/quality/qs-docs/g9s-final.pdf.

Barnett, V. and Lewis, T. 1994.*Outliers in Statistical Data.* Third Edition. John Wiley.

Conover, W. J. 1999. *Practical Nonparametric Statistics*. Second Edition. John Wiley.

Gerlach, R.W., and Nocerino, J. M. (2003). *Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples*, EPA/600/R-03/027, November 2003

Kaplan, E.L. and Meier, O. 1958. *Nonparametric Estimation from Incomplete Observations*. Journal of the American Statistical Association, Vol. 53. 457-481.

Scout. 2002. *A Data Analysis Program*, Technology Support Project. USEPA, NERL-LV, Las Vegas, NV.

Singh, A. and Nocerino, J.M. 1995. *Robust Procedures for the Identification of Multiple Outliers*. Handbook of Environmental Chemistry, Statistical Methods, Vol. 2.G, pp. 229-277. Springer Verlag, Germany.

Singh, A., Singh, A.K., and Flatman, G.T. 1994. *Estimation of Background Levels of Contaminants,* Math Geology. Vol. 26, No. 3, pp. 361-388.

Singh, A, Singh, A.K., and Iaci, R.J. 2002. *Estimation of the Exposure Point Concentration Term Using a Gamma Distribution*, EPA/600/R-02/084.

Singh, A. and Singh, A.K. (2003). Estimation of the Exposure Point Concentration Term (95% UCL) Using Bias-Corrected Accelerated (BCA) Bootstrap Method and Several other methods for Normal, Lognormal, and Gamma Distributions. Draft EPA Internal Report.

Singh, A., Maichle, R., and Lee, S. 2006. *On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets With Below Detection Limit Observations. ,* EPA/600/R-06/022, March 2006.