Averaging Soil-Contaminant Concentration Data

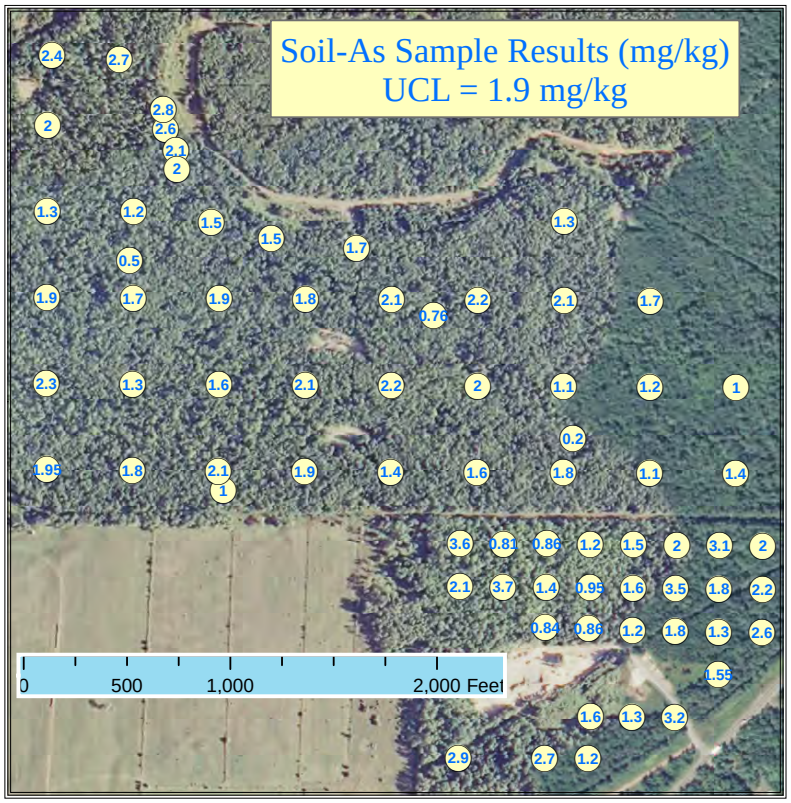
DNR's RR Program has created a guidance document, *Averaging Soil-Contaminant Concentration Data* (RR-991), to help responsible parties and consultants understand the data analysis that the agency recommends when performing soil averaging. This guidance document is available for public comment at http://dnr.wi.gov/news/input/guidance.html until March 9, 2015.

Averaging the results of individual soil samples to determine the risk of direct contact with contaminated soil at an individual property is permissible in certain situations. If done appropriately, the averaging of soil contaminant concentrations is allowed for comparison to the contaminant's direct-contact soil residual contaminant level (RCL).

Comments and questions regarding this guidance document may be directed to Resty Pelayo at aristeo.pelayo@wisconsin.gov or 608-267-3539.

Averaging Soil Contaminant Concentration Data Under Ch. NR 720, Wis. Admin. Code

RR-991 **February 2015**



Wisconsin Department of Natural Resources
P.O. Box 7921, Madison, WI 53707
dnr.wi.gov, search "brownfield"

Table of Contents

I. Introduction 3

A. Purpose of Guidance 3

B. Disclaimer, Contacts and Revisions 3

C. Statistical Averaging - Is It Right for Your Site? 3

II. Averaging Soil Contaminant Concentration Data 4

A. Overview 4

B. Guidance Objective and Definition of Average 4

C. Using ProUCL Software from EPA 5

III. Examples of Averaging 6

Summary of Examples 6

Example 1: Simulation of a Mixture Population 7

Example 2: Separating a Mixture Population 12

Example 3: Misstep in UCL Calculation 14

Example 4: Straightforward Comparison between the DC-RCL and the UCL 15

IV. Use of Background Threshold Values 16

V. Conclusion 17

I. Introduction

A. Purpose of Guidance

The purpose of this document is to provide Department of Natural Resources (DNR) staff, responsible parties and other interested individuals with guidance on the application of administrative rules related to data averaging and the cleanup of soil contamination under NR 720, Wis. Admin. Code. This guidance document identifies a few key tools in the U.S. Environmental Protection Agency's (EPA's) ProUCL statistical software package for use in averaging soil data. For more information contact Resty Pelayo, DNR, at aristeo.pelayo@wisconsin.gov or (608) 267-3539.

B. Disclaimer, Contacts and Revisions

This document is intended solely as guidance and does not contain any mandatory requirements except where requirements found in statute or administrative rule are referenced. This guidance does not establish or affect legal rights or obligations and is not finally determinative of any of the issues addressed. This guidance does not create any rights enforceable by any party in litigation with the State of Wisconsin or the Department of Natural Resources. Any regulatory decisions made by the Department of Natural Resources in any matter addressed by this guidance will be made by applying the governing statutes and administrative rules to the relevant facts.

This guidance will be updated as needed. Questions, comments and concerns may be sent to: Judy Fassbender - RR/5, DNR, P.O. Box 7921, Madison, WI 53707. Phone: (608) 266-7278, E-mail: judy.fassbender@wisconsin.gov [cc: aristeo.pelayo@wisconsin.gov]

C. Statistical Averaging - Is It Right for Your Site?

Per NR 720.07(2)(b), Wis. Admin. Code, alternatives to point-to-point comparison to soil cleanup standards require DNR approval. This document explains how one approvable alternative – statistical averaging – may be developed to determine if applicable direct-contact soil cleanup standards have been met. Statistical averaging of sample data is an alternative approach, and it may not be appropriate at some sites. It is recommended that you contact your DNR project manager before deciding whether or not to proceed with an averaging approach.

Data from at least 10 (ten) sample points is generally necessary before considering averaging. The data should be from a single statistical population. Data from mixed populations (e.g., mixed background and hot spot data) requires processing to separate the different populations. Using averaging to avoid remediating hot spots is not acceptable. It is strongly recommended that you submit a statistical averaging proposal early to DNR, long before submitting a case closure request (along with a NR 720 soil cleanup standard/report technical assistance fee) to avoid delays in the review and approval process.

A good understanding of statistics is necessary to complete the requisite data analysis for averaging. For data analysis, DNR recommends using EPA's ProUCL statistical software, which is available to DNR and consultants at no charge.

II. Averaging Soil Contaminant Concentration Data

A. Overview

Discrete point-to-point comparison is the primary mechanism to determine if the contaminant concentration in a soil sample exceeds standards. Soil standards are termed soil residual contaminant levels (RCLs) in ch. NR 720, Wis. Admin. Code. For direct-contact (DC) RCLs, however, there is an alternative to point-to-point comparison that DNR can approve: Averaging.

The underlying concept is that, over time, the average is more representative of the concentration that would actually be contacted at the site. Hence, averaging, when done appropriately, can be approved under NR 720.07(2)(b), Wis. Admin. Code as an alternative to the conventional point-to-point comparison to the DC RCL.

The "average" for a particular contaminant is defined by the 95% upper confidence limit for the mean (UCL) of the contaminant concentrations in soil samples. The aim of using the UCL is to provide a reasonable confidence that the true site average will not be underestimated. It is fundamentally important that the UCL be computed using a data set representing a single statistical soil population, and not from a mixture of populations.

Averaging soil concentrations may be appropriate where the contaminant is widespread, relatively consistent in concentrations and at levels close to the applicable RCL for that compound. This guidance explains how to average measured soil sample contaminant concentrations to determine whether the calculated DC RCL standard has been exceeded. It includes several examples to illustrate important statistical concepts and to show the utility of a few tools available for calculating UCLs using EPA's ProUCL software.

Averaging is generally straightforward if it can be established that there is a single population being represented by the observed concentrations. One key goal in any site investigation is the delineation of soil "hot spots" for remediation. As such, straightaway averaging of soil concentrations without considering the data distribution(s) may defeat that goal.

Averaging is not appropriate as the sole method for addressing sites where areas of significant soil contamination may exist. Simple averaging cannot be relied on for addressing hot spots or source areas on a property. Summing the concentrations and then dividing by the number of samples to get the arithmetic mean may cause concentrations that are significantly higher than the applicable RCL to be "averaged away" by the more numerous low or non-detectable levels in the soil sample data set. This is not acceptable.

B. Guidance Objective and Definition of Average

This guidance document provides general background information on the process and assumptions that need to be considered when statistically evaluating data. It identifies the use of a graphical Q-Q plot, outlier test and histogram plot to preliminarily categorize data before averaging is performed. It is not intended to be a detailed discussion on statistical applications.

The generic word “average” can refer to the mean, median, mode, or some other central or typical value. To be more precise, we define average here as the 95% upper confidence limit for the mean (95% UCL or simply, UCL).

The UCL is particularly dependent on the number of samples (or sampling size). As the sampling size increases, the difference between the UCL and the true mean decreases. The UCL provides a reasonable confidence that the true site average will not be underestimated when it is being compared to the RCL.

An assumption in this guidance is that the individual responsible for analyzing the soil data has a good working knowledge of statistics and statistical applications, like knowing the use of and the difference between the 95% UCL and the 95th percentile statistics.

C. Using ProUCL Software from EPA

The critical first step in environmental data analyses is in determining the presence of multiple populations or outliers in the data set. Recognizing when two or more populations exist is important because any calculated statistic will be meaningless if, for example, that statistic comes from unseparated “apples and berries” data. Likewise, recognizing outlying data is important because the statistics that include outliers will be distorted (often inflated) for the main dominant population. Further investigation of the outliers may be needed to determine the reasons for their occurrences (e.g., cross-contamination, contaminated locations). After separating out the different populations and having outlier-free subsets, the computation of the appropriate decision statistics (e.g., UCLs) for each identified population becomes a relatively straightforward task.

Given the technical scrutiny and range of options available in EPA’s ProUCL software, DNR recommends its use when statistically evaluating environmental data to determine if applicable soil cleanup standards have been met. It is available free to DNR and environmental professionals. The software, its documentation and training webinars can be found at: www.epa.gov/osp/hstl/tsc/software.htm

The ProUCL tools described in this document help compute appropriate decision statistics, particularly the UCL. Other statistical tools and tests in ProUCL, besides the ones identified in this document, may also be acceptable but they have not been evaluated by DNR in detail and therefore are not discussed.

The focus of this document is on calculating the UCL (“average”) of the data set being evaluated, particularly pinpointing situations where taking the average may be inappropriate. In general, the use of ProUCL to calculate the UCL requires at least 10 sample points, so if the site data set is smaller, the statistical approaches described here will likely not be usable. It is always a good idea to talk with your DNR project manager before deciding to use averaging at a site.

III. Examples of Averaging

Summary of Examples

The following pages include four examples with data sets. The statistical evaluation of the data sets is intended to illustrate the tools used in calculating UCLs so that it can be determined if the applicable direct contact RCL has been met. The procedures in the examples, if proposed for use at a site, are approvable as an alternative approach in NR 720.07(2) (b), Wis. Admin. Code.

- Example 1 simulates a mixture of data composed of random samples from two normal populations ("A" and "B") having different means. When graphed using the normal probability plot or ProUCL's quantile-quantile (Q-Q) plot tool, the mixture is depicted by 2 distinguishable straight-line segments, with A's data falling nicely into one line, separable from B's data falling in another.
- Examples 2 and 3 have actual site data where the Q-Q plot, outlier test and histogram chart are used in resolving the data sets as composed of subsets of several populations.
- Example 4 shows a straightforward RCL-UCL comparison. It is first established that the data set comes from a single population, and then the UCL is determined. If the UCL does not exceed the RCL, then that data set (even if there were point exceedances) meets the soil standard.

If you are reading a printed copy of this document and want to view the charts and images in a larger format go to the online PDF version of this document, at dnr.wi.gov/files/PDF/pubs/rr/RR991.pdf, and zoom in.

Example 1 - Simulated MIXTURE of POPULATIONS

		A: Normally-Distributed Population w/ u = 8; sd = 2	B: Normally-Distributed Population w/ u = 20; sd = 8		Simulated "Mixed" Population
u (true mean)-->		8	20		
sd (true std. dev.)-->		2	8		
		Sample Set A	Sample Set B		"Mixed" A and
1		11.08	14.62		11.08
2		7.00	21.36		7.00
3		6.82	26.05		6.82
4		8.77	20.99		8.77
5		6.92	17.86		6.92
6		10.19	21.80		10.19
7		4.82	9.03		4.82
8		11.35	27.15		11.35
9		4.89			4.89
10		4.94			4.94
11		6.83			6.83
12		9.58			9.58
13		7.15			7.15
14		6.50			6.50
15		5.87			5.87
16		6.93			6.93
17		3.16			3.16
18		8.34			8.34
19		7.84			7.84
20		5.72			5.72
21		10.48			10.48
22		8.81			8.81
23		5.82			5.82
24		9.49			9.49
25		6.82			6.82
26		8.73			8.73
27		4.07			4.07
28		7.59			7.59
29		9.00			9.00
30		5.86			5.86
31					14.62
32					21.36
33					26.05
34					20.99
35					17.86
36					21.80
37					9.03
38					27.15
	Min:	3.16	9.03		3.16
	Max:	11.35	27.15		27.15
Sample Mean:		7.38	19.86		10.01
Sample Std. Dev.:		2.08	5.95		6.05

A "mixture" sample size 38 was generated from 2 normal populations "A" and "B." Population A has a true mean (u) of 8 and a true standard deviation (sd) of 2. Population B has u=20 and sd=8. (Think "A" for apples and "B" for berries, so the mixture is data from incompatible apples and berries.)

Random samples of size (n) 30 were drawn from "A," providing a range from 3.16 to 11.35. Fewer random samples of size 8 were drawn from "B," providing a wider range from 9.03 to 27.15, indicating some overlap between "A" and "B."

The data were analyzed using ProUCL. Specifically, graphical normal probability quantile-quantile ("Q-Q") plots were rendered to get an insight into the mixture.

As shown in Box 2 for this example, the Q-Q plot is useful in identifying the presence of mixture populations (*i.e.*, data from several populations) potentially present in a data set. However, once mixed, it becomes difficult to correctly categorize all the data, whether they belong to "A" or "B", especially for the data points in their overlap region.

For this example, inspection of the Q-Q plot shows an overlap region around the value 12, such that data larger than 12 can be categorized as coming from "B" and values lower to "A." However, by picking 12 as a threshold value separating A from B, one data point from "B" with value 9.03 will be miscategorized. Nevertheless, with the Q-Q plot, we were able to not only identify 2 populations, but even with just a visual inspection also able to correctly categorize almost all the data.

If mixture of populations is indicated by the Q-Q plot, separation into subsets is necessary to get meaningful statistics from ProUCL for the different populations.

(Box 1 for Example 1)

Example 1: Simulation of a Mixture Population

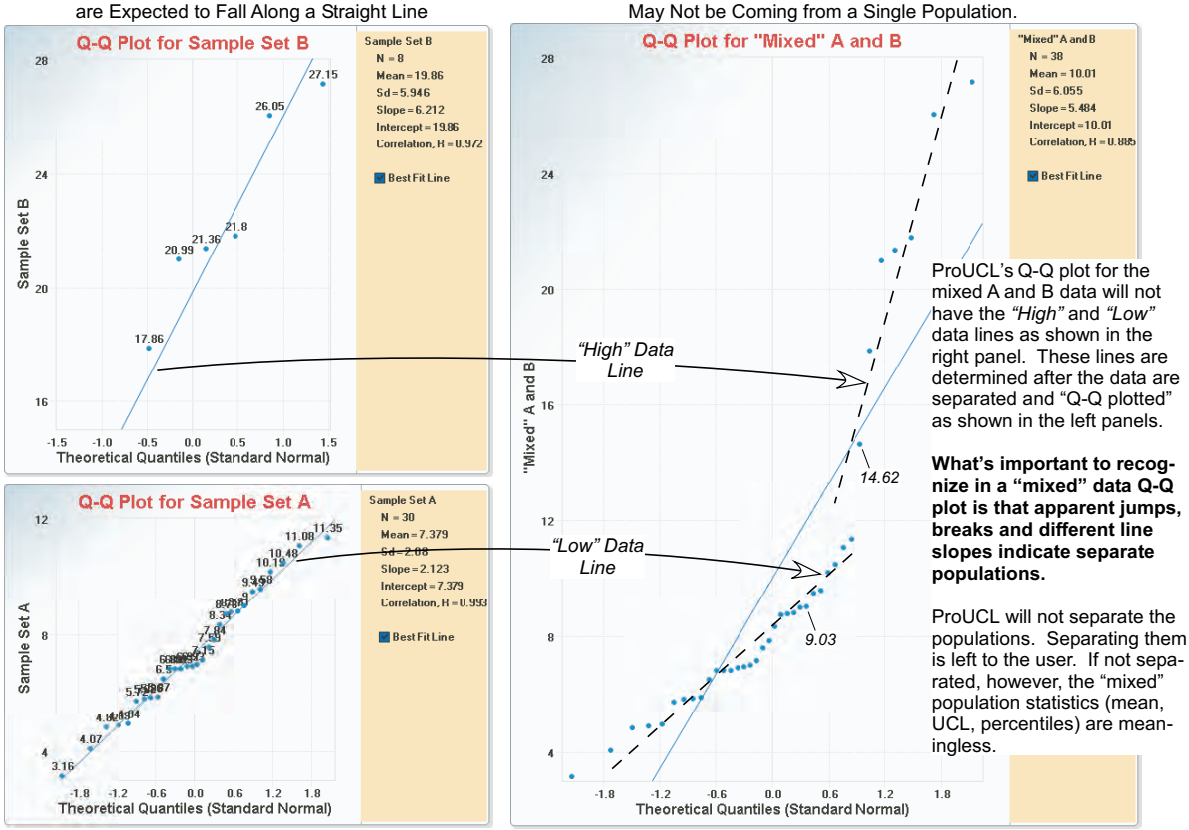
For this simulation, we assume perfect knowledge for 2 normal populations ("A" and "B") with different true means and standard deviations. Population A has a true population mean (u) of 8 and a population standard deviation (sd) of 2. Population B has u=20 and sd=8. Schilling *et al.* [2002] identified several criteria for when data from 2 normal populations are separable. Our assumed populations qualify by having their true means (8 and 20, respectively) separated by at least the sum of their standard deviations (2+8=10). So when we have enough samples from both populations, we will expect two separate bell curves when we plot the histogram of their mixed data. In Example 1, we simulate the case where we have fewer samples from B (the population with the larger mean). This may more often be the scenario in many real site situations where because the contamination is small volumetrically and areally, fewer samples are collected in those areas relative to the uncontaminated area of the site. In Example 1, we create a mixture of the data from both populations, but with relatively fewer B samples so as not to easily see two bell curves.

The data table in Box 1 for Example 1 shows our randomly-generated sample values from Population A (sample size n: 30; sample mean $\bar{x}$: 7.38), Population B (n: 8; $\bar{x}$: 19.86) and the Mixture A and B (n: 38; $\bar{x}$: 10.01). The first thing to note is that with the mixture population, we have calculated a mixture sample mean ($\bar{x}$: 10.01), a numerically "in-between" statistic that is not particularly relevant to A (with true mean $u = 8$) nor B ($u = 20$).

Example 1 (continued) - Simulated MIXTURE of POPULATIONS

Normally-Distributed Data in a Q-Q Plot are Expected to Fall Along a Straight Line

When Q-Q Plot Indicates Not a Single Line, but Possibly 2 or More, Data May Not be Coming from a Single Population.



ProUCL's Q-Q plot for the mixed A and B data will not have the *"High"* and *"Low"* data lines as shown in the right panel. These lines are determined after the data are separated and "Q-Q plotted" as shown in the left panels.

What's important to recognize in a "mixed" data Q-Q plot is that apparent jumps, breaks and different line slopes indicate separate populations.

ProUCL will not separate the populations. Separating them is left to the user. If not separated, however, the "mixed" population statistics (mean, UCL, percentiles) are meaningless.

(Box 2 for Example 1)

The question is: Once mixed, is there a statistical tool to distinguish "A" samples from "B" samples? The answer is: Yes, and the best distinguishing tool is to render a quantile-quantile or Q-Q plot for the mixed data. Data points from "A" will fall along a line, and data points from "B" will fall along another line segment in the Q-Q plot. The right panel of Box 2 for Example 1 has the Q-Q plot for the mixed data. In this Q-Q plot, the whole range of the mixture data set is displayed, arranged from the lowest to highest value along the vertical y-axis. The Q-Q plot's horizontal x-axis has the theoretical normal quantile values. Having the larger mean, the "B" data set plots higher (along the "High" Data Line), separate from the "A" data ("Low" Data Line).

Q-Q Plot. The Q-Q plot is also known as the normal probability plot. This graphical tool is used in many statistical data analyses to gain insight into data sets collected in the different fields of biology, medicine, geochemical prospecting and environmental studies. The probability plot approach has been used in geochemical prospecting for over 50 years [e.g., see *Carranza*, 2008, for published references dating back to 1959] to identify outliers and geochemical anomalies, to examine potential multiple populations, and to define the breaks or inflection points at which to subdivide the data set into subsets representing those populations. The apparent jumps and breaks in the Q-Q plot suggest the presence of multiple statistical populations [e.g., *Fowlkes*, 1979]. The normal probability plot or its equivalent cumulative probability plot has been used to separate background levels from contaminated levels [e.g., *Fleischhauer and Korte*, 1990]. The inflection

points on the Q-Q graph indicate an interpretable breakdown of the distribution into separable populations with different means [e.g., *Singh et al.*, 1994]. However, the identification of the exact break points (or threshold values for the different populations) in a Q-Q plot is highly subjective, requires experience, and so is not a trivial task. Historical and theoretical reviews of decomposing finite mixture distributions can be found in *Holgersson and Jorner* [1978], *Sinclair* [1991] and *Everitt* [1996]. For this guidance, we limit our discussion to the graphical and visual fit of straight lines to Q-Q plot segments. Rather than delve into advanced routines (e.g., determining the intersection for 2 line segments or minimizing the kurtosis for each population), the approach here will be more towards visually inspecting inflection points when picking a potential threshold value that separates one population from another. In many practical applications, this visual approach may suffice in determining statistically valid conclusions.

In ProUCL's Graph, we rendered the Q-Q plot of the mixture A and B data to show that with their different slopes, the "high" and "low" data lines are distinguishable and that their dissimilarity can be the basis for their separation. Upon close inspection of the Q-Q plot for the mixture, it seems apparent that the threshold value (on the y-axis) separating A and B is around 12, such that values less than 12 may be categorized as belonging to "A," while values larger may be categorized as "B" sample points. This simple procedure of visually picking 12 did not resolve the categorization of all the data. Out of B's original 8 sample points (see Box 1 for Example 1), there are now after mixing only 7 that we can resolve as "B" results because B's lowest value (9.03) was less than 12. Nevertheless, we have accomplished most of the "un-mixing," such that when we proceed to calculate the statistics for two (2) separate subsets (below-12 and above-12), we will get more meaningful unmixed statistics for the 2 separate populations.

Outlier Tests. Another tool that may be useful in separating A from B is the outlier tests under ProUCL's "Statistical Tests." The outlier test is particularly helpful when the mixture data set is dominated by samples from a main population with only a few data from another – which aptly describes our mixture A and B data. The result from the outlier test may help corroborate the result from our analysis using the Q-Q plot. One important consideration in testing for outliers is that it should be performed on raw data, as the cleanup decision needs to be made based upon concentration values in the raw scale and not in the log-scale. ProUCL's outlier test tool applies 1 of 2 "classical" outlier tests, either the Rosner's or Dixon's test, depending on how many data points are present. ("Modern" robust outlier tests exist, but they are not tools in ProUCL.)

ProUCL carries out the Rosner's test when the sample size $n \geq 25$; otherwise, ProUCL applies the Dixon's test. For the mixture A and B data (n: 38), the Rosner's outlier test was performed. The outliers identified by the test are in Box 3 for Example 1. The Rosner's test can be used to identify up to 10 outliers. The test procedure is described more fully in *Gilbert* [1987]. In ProUCL, the number of suspected outliers is specified before the Rosner's test is run. Specifying how many outliers are suspected is an important step in the Rosner's test. Counting suspected outliers from the Q-Q plot can help in initially deciding how many there are. If more than 10 are suspected, it is procedurally possible to delete so as to exclude the most obvious outlying data to reduce to 10 or less the number of suspected points subjected to Rosner's test. For a suspected extreme outlier, a Rosner test statistic is computed by getting the difference between the outlier and sample mean and dividing the difference by the sample standard deviation. If this

Example 1 (continued) - Simulated MIXTURE of POPULATIONS

Rosner's Outlier Test for "Mixed" A and B

Mean	10.01
Standard Deviation	6.055
Number of data	38
Number of suspected outliers	10

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	10.01	5.974	27.15	38	2.87	3.01	3.36
2	9.543	5.412	26.05	33	3.05	3	3.34
3	9.084	4.704	21.8	36	2.703	2.99	3.33
4	8.721	4.229	21.36	32	2.989	2.98	3.32
5	8.349	3.667	20.99	34	3.447	2.97	3.3
6	7.966	2.953	17.86	35	3.35	2.954	3.284
7	7.657	2.397	14.62	31	2.905	2.938	3.268
8	7.432	2.066	3.16	17	2.068	2.922	3.252
9	7.575	1.94	11.35	8	1.946	2.906	3.236
10	7.444	1.837	11.08	1	1.979	2.89	3.22

For 5% significance level, there are 6 Potential Outliers

Potential outliers are:

27.15, 26.05, 21.8, 21.36, 20.99, 17.86 ← Identified Outliers

For 1% Significance Level, there are 6 Potential Outliers

Potential outliers are:

27.15, 26.05, 21.8, 21.36, 20.99, 17.86

ProUCL Outlier Test

When analyzing a data size of 25 or more, the Rosner's test in ProUCL can iteratively check if there is at most 10 outliers in the data set. This may help separate out mixture populations.

For the simulated mixture (A *and* B data), the test identified the 6 largest values (all from B) as outliers. But even after this test, with only 6 outliers identified, we failed to exclude the 7th suspected point (14.62 also from B) from the "A" data set.

In the Q-Q plot, 14.62 falls in B ("*High*" data line). Its exclusion from A will be based not on the outlier test, but on the Q-Q plot.

Nevertheless, excluding identified outliers before proceeding with ProUCL calculations will provide for more meaningful statistics than if the outliers were included for the "A" data set.

(Box 3 for Example 1)

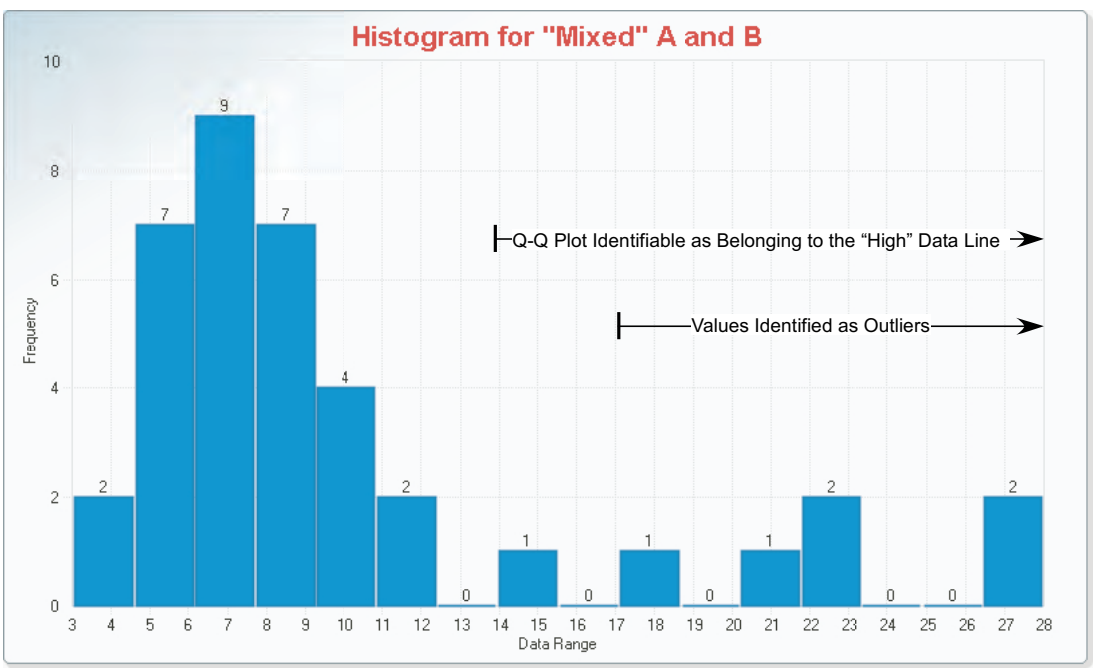
test statistic is greater than a certain critical value, then the conclusion is that the extreme outlier is indeed an outlier; otherwise, it is not. However, the presence of intermediate (less extreme) outlier(s) can inflate the sample standard deviation (divisor in the test statistic), making the test-value statistic small. This, in turn, can make the test inconclusive about the extreme outlier. In other words, the presence of other outliers in between the extreme value and the main population may "mask" the fact that the extreme value is an outlier. To minimize masking, the Rosner's test proceeds by excluding the extreme value and computing a new test statistic for the next extreme outlier. The procedure is repeated for how many suspected outliers we initially specified.

The Rosner's test on mixture A and B identified the 6 largest values as outliers (Box 3). How the test identified the outliers requires some description. For the first extreme observation (27.15), its test-statistic value (2.87) is less than the critical value (either 3.01 or 3.36), so initially we cannot conclude it as an outlier. Had we specified only 1 (rather than 10) suspected outlier before the test is run, the test would not have identified this extreme point as an outlier. In the next iteration, the program discards the 1st extreme point and calculates a 2nd test value for the 2nd extreme point (26.05). The 2nd test value (3.05) is greater than the 5% critical value (3.), but is less than the 1% critical value (3.34). So had we specified only 2 suspected outliers, the Rosner's test at 5% significance would have identified the 2nd extreme value (26.05) as an outlier, but the test at 1% would not. Upon proceeding and by the 6th test statistic, it becomes apparent why the first

few more-extreme data points (27.15, 26.05, …, 21.36) had low test-value statistics. The sample standard deviations for their test values were inflated because there were other outliers remaining in the data set. Those intermediate outliers had obscured (or masked) the presence of the more extreme outlier. By the 7th test and on, the Rosner's test did not result in any more outliers being identified. The Rosner's test did not identify the 7th suspected point (14.62) as an outlier, even though based on the Q-Q plot, it is a point to be included in the "B" (not "A") subset. When a value is in close proximity to the inflection point between 2 populations, it is not uncommon for a data point to be incorrectly classified. This is typically not a major concern because the more extreme data points have already been identified as outliers. So even though the outlier test did not corroborate all 7 values from the Q-Q plot as non-"A" data (nor the 8 original simulation B sample data), what is important in the outlier test result is that it identified the presence of values not belonging to A, the main population represented by the samples.

Histogram. A third tool that can help distinguish populations is by graphing the data histogram. The rendered histogram for the mixture A and B is shown in Box 4 for Example 1. ProUCL allows for either decreasing or increasing the number of categories or bins for the histogram. Generally, the utility of the histogram may be limited because of the number of samples (typically, too few for either of the populations), the proximity between the population means or the unequal standard deviations for the populations. How the histogram is rendered is affected by the number of bins. With the sample histogram we can only try to capture the shape of the population distribution, so with too few bins, the shape is lost. And then with too many bins, the shape is likewise lost. Nevertheless, the histogram's practical use is that it should illustrate the conclusions derived from the Q-Q plot and the outlier test.

Example 1 (continued) - Simulated MIXTURE of POPULATIONS



The histogram supports the Q-Q plot analysis of having 2 populations with "low" and "high" values, respectively. Values under 12 are likely from the "low" population, while values above 12 are likely from the "high" population. Since the outlier test showed the highest 6 values as outliers, at a minimum, they should be dropped from calculating the statistics for the main grouping of the data.

(Box 4 for Example 1)

Example 2: Separating a Mixture Population

One of the inherent assumptions required to compute the UCL for the mean is that the data set under consideration originated from a single statistical population [EPA, 2013, ProUCL Version 5.0.00 Technical Guide, p. 8]. To avoid calculating statistics that may be meaningless, it is important to be able to recognize if multiple populations may be present in the soil data set we have. Correctly characterizing the distribution of the sample data set provides for robust statistics for use in any site decision. A normally distributed data set with a few high observations– such as those expected from hot spots– can be incorrectly modeled by the lognormal distribution with the outliers conveniently hidden by the lognormal assumption. An incorrect assumption on the data distribution, either because of the presence of outliers or the fact that the data set is from two or more populations, may result in misleading statistics. Data from a site investigation generally consist of samples from clean (uncontaminated) and contaminated parts of the site. As such, it is reasonable to think of the data set as a mixture of (at least) 2 populations with the idea of separating them out first before any statistical analysis. Population partitioning is not a trivial exercise, so a published example is included here to illustrate what steps may be involved.

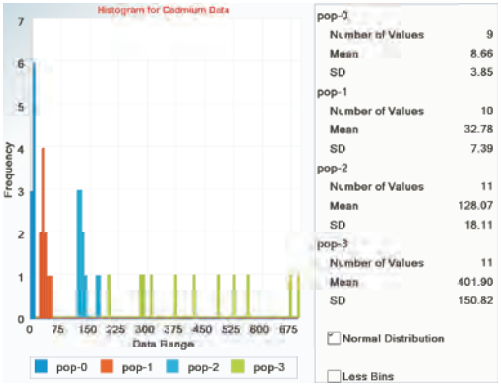
Example 2 has a mixture data set from a Superfund site where, even after the removal of the most obvious extreme outliers (column Cd-Pop-4), not just 2, but 4 component populations (with varying degrees of contamination levels) have been identified and separated by a partitioning technique developed by *Singh et al.* [1994]. The histogram in Box 1 for Example 2 displays the different identified populations; each population is represented by a unique color in the histogram. It is apparent from the colors that the spread or width of each population varies, highlighting why the histogram may not be the best initial tool to separate out the populations. A bin width of 50 may be suitable for the very wide data shown in green (Pop-3), but it will then lump together the 2 narrower populations with the lowest means in dark blue (Pop-0) and red (Pop-1).

The Q-Q plot for all the data is shown in the right panel of Box 2 for Example 2. This initial Q-Q plot can generally provide an idea about the number of populations present. *Singh et al.*'s [1994] decomposition of the data relied heavily on the Q-Q plot, and the gaps and straight line segments identifiable in the Q-Q plot. Each line segment in the initial Q-Q plot represents data from a separable normal population. The subset of data falling along a line segment is considered a component population and its separate Q-Q plot constructed. These "component" Q-Q plots are shown in the left panels. An interesting aspect of this analysis is that the lowest identified component can be thought of as coming from background. Hence, the decomposition procedure can be followed for estimating the statistics for site-specific background level at remediation sites, specifically in identifying a background threshold value (BTV). Once determined, any concentration greater than the BTV can be concluded as exceeding background.

Example 2 - Mixture Data Decomposition (Published Analysis of Cd Concentrations)

Count	Cadmium-All	CATEGORIES	Cd-Pop-0 / LowestLevels	Cd-Pop-1	Cd-Pop-2	Cd-Pop-3	Cd-Pop-4/ HighestLevels
1	26.2	Pop-1		26.2			
2	27.55	Pop-1		27.55			
3	445.01	Pop-3				445.01	
4	30.77	Pop-1		30.77			
5	486.31	Pop-3				486.31	
6	513.79	Pop-3				513.79	
7	112.81	Pop-2			112.81		
8	159.3	Pop-2			159.3		
9	1300.	Pop-4					1300.
10	6.68	Pop-0	6.68				
11	33.72	Pop-1		33.72			
12	35.01	Pop-1		35.01			
13	10.99	Pop-0	10.99				
14	22.05	Pop-1		22.05			
15	830.94	Pop-4					830.94
16	125.07	Pop-2			125.07		
17	40.84	Pop-1		40.84			
18	345.52	Pop-3				345.52	
19	384.8	Pop-3				384.8	
20	183.04	Pop-3				183.04	
21	2300.	Pop-4					2300.
22	1500.	Pop-4					1500.
23	260.27	Pop-3				260.27	
24	32.09	Pop-1		32.09			
25	166.16	Pop-2			166.16		
26	31.68	Pop-1		31.68			
27	12.39	Pop-0	12.39				
28	614.53	Pop-3				614.53	
29	639.52	Pop-3				639.52	
30	116.24	Pop-2			116.24		
31	119.43	Pop-2			119.43		
32	111.6	Pop-2			111.6		
33	10.29	Pop-0	10.29				
34	1.68	Pop-0	1.68				
35	3.34	Pop-0	3.34				
36	10.47	Pop-0	10.47				
37	11.74	Pop-0	11.74				
38	10.32	Pop-0	10.32				
39	122.3	Pop-2			122.3		
40	283.03	Pop-3				283.03	
41	265.08	Pop-3				265.08	
42	125.49	Pop-2			125.49		
43	131.06	Pop-2			131.06		
44	47.9	Pop-1		47.9			
45	119.34	Pop-2			119.34		
		Counts:	9	10	11	11	4
		Max:	12.39	47.9	166.16	639.52	2300.

This mixture data set was from a Superfund site and published in an article by Singh *et al.* [1994].* The article's lead author (Anita Singh) is also the lead author of ProUCL. The histogram of the data (without the 4 highest levels) is shown below with different colors for the different populations.



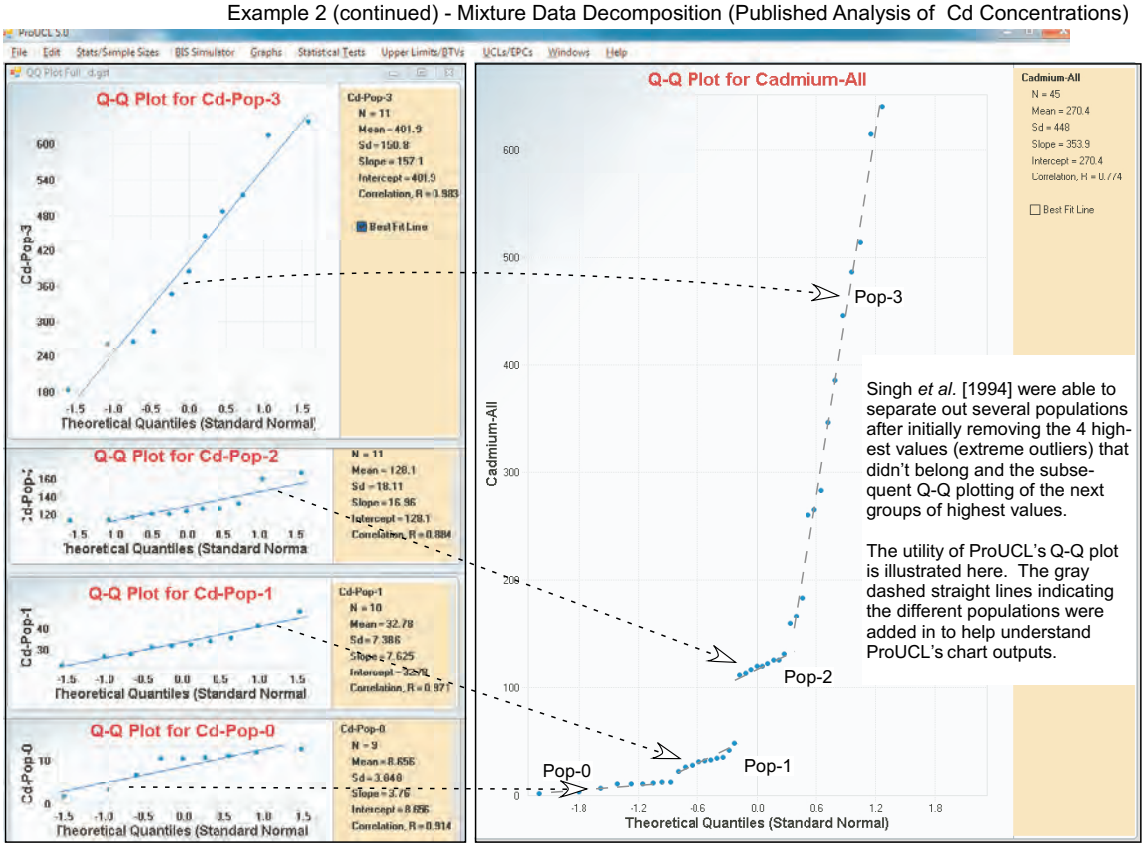
After initially removing the very highest values (Cd-Pop-4 column in the table), the authors were able to separate 4 other populations (Cd-Pop-0 thru -3) by a stepwise procedure of removing high values and checking the Q-Q plots of both the remaining group and the removed high group.

The graphical analysis using ProUCL is shown in Box 2 for this example. Several Q-Q plots are shown; the largest plot in the right panel is for all 45 data points. With this initial Q-Q plot of all the data, it becomes apparent that certain data groupings formed straight-line segments. Because each separate component population is expected to lie along a straight line in a Q-Q plot, data falling along the distinct line segments having different slopes in the initial Q-Q plot are interpreted as representing samples from separable populations. The smaller Q-Q plots in Box 2 are for the different populations identified in the mixture.

* Singh, A., A. K. Singh, and G. Flatman, 1994. Estimation of background levels of contaminants, *Mathematical Geology, Vol. 26, No. 3*, p. 361-388.

(Box 1 for Example 2)

Example 2 (continued) - Mixture Data Decomposition (Published Analysis of Cd Concentrations)



Singh *et al.* [1994] were able to separate out several populations after initially removing the 4 highest values (extreme outliers) that didn't belong and the subsequent Q-Q plotting of the next groups of highest values.

The utility of ProUCL's Q-Q plot is illustrated here. The gray dashed straight lines indicating the different populations were added in to help understand ProUCL's chart outputs.

(Box 2 for Example 2)

Example 3: Misstep in UCL Calculation

Soil-PAH (polycyclic aromatic hydrocarbon) samples were collected for direct-contact evaluations (sample depths between 0 and 4 feet) at an industrial site that was a bulk petroleum storage terminal for over 100 years. Today, it is no longer a bulk terminal, and several non-residential buildings and vacant lots are now situated at this redeveloped site. It was found that BaP (benzo[a]pyrene) was the most-frequently encountered PAH to exceed its direct-contact soil residual contaminant level (RCL) of 0.211 mg/kg allowed at industrial sites. The soil-BaP data are tabulated in Example 3.

The table did not include some data (empty cells) as a decision regarding them (excavation) was already in place. The table shows that over half of the BaP data are "NDs." (In a ProUCL input table, "ND" means entering the lab detection limit under the BaP column followed by a "0" flag in the d_BaP column.) Hence at the outset, based on the frequency of NDs, there is already an indication of at least 2 populations. Example 3 is provided here to show how an inappropriate UCL estimate can be misleading.

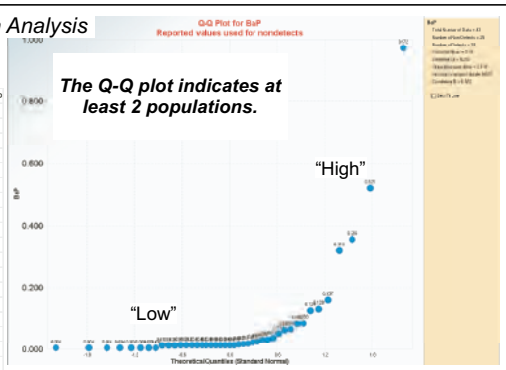
The 4 highest concentrations (shown in red in the table) are greater than BaP's RCL of 0.211 mg/kg. Without regard to the presence of multiple populations, plugging the data into ProUCL

Example 3 - Misstep in UCL Calculation

ProUCL Data Analysis

Data

Count	BaP	d_BaP
1	0.356	1
2	0.972	1
3	0.0039	0
4	0.0038	0
5	0.0172	1
6	0.0039	0
7	0.128	1
8	0.0141	1
9	0.0622	1
10	0.124	1
11	0.0038	0
12	0.0039	0
13	0.004	0
14	0.0038	0
15	0.521	1
16	0.0037	0
17	0.0515	1
18	0.318	1
19	0.0822	1
20	0.025	1
21	0.0182	1
22		
23	0.0834	0
24	0.157	0
25		
26	0.0135	0
27	0.0654	1
28	0.0167	0
29	0.0133	0
30	0.0137	0
31	0.0138	0
32		
33	0.013	0
34	0.031	1
35	0.0365	1
36	0.0136	0
37	0.0136	0
38	0.0141	0
39		
40		
41	0.0138	0
42	0.0138	0
43	0.0142	0
44	0.0222	1
45	0.0304	1
46	0.0132	0
47	0.0131	0
48		
49	0.0125	0



Rosner's Outlier Test for BaP

The outlier test indicates that the 9 highest values are outliers.

Number of data 43
Number of suspected outliers 10

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	0.078	0.172	0.972	2	5.198	3.07	3.41
2	0.0667	0.106	0.521	15	4.415	3.06	3.4
3	0.0453	0.0763	0.356	1	4.074	3.05	3.39
4	0.0376	0.0585	0.318	18	4.79	3.04	3.39
5	0.0304	0.0374	0.157	23	3.389	3.03	3.37
6	0.0271	0.0314	0.128	7	3.211	3.018	3.355
7	0.0243	0.0269	0.124	10	3.301	3.006	3.342
8	0.0216	0.0213	0.0834	22	2.901	2.994	3.329
9	0.0196	0.0186	0.0822	13	3.327	2.982	3.314
10	0.018	0.0155	0.0654	25	3.055	2.97	3.3

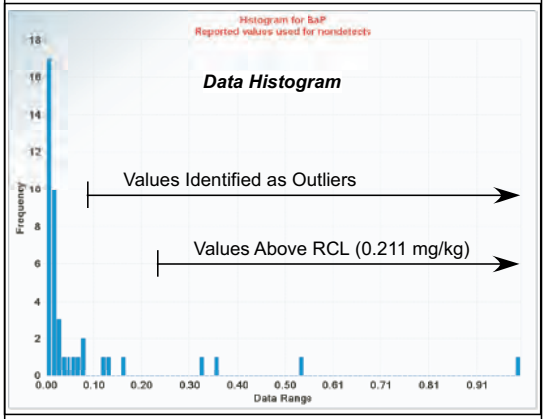
For 5% significance level, there are 10 Potential Outliers
Potential outliers are:
0.972, 0.521, 0.356, 0.318, 0.157, 0.128, 0.124, 0.0834, 0.0822, 0.0654

For 1% Significance Level, there are 9 Potential Outliers
Potential outliers are:
0.972, 0.521, 0.356, 0.318, 0.157, 0.128, 0.124, 0.0834, 0.0822

Use in Site Decision

Tabulated on the left are soil-BaP (mg/kg) data collected at an industrial site. Empty cells had data but weren't included here because a remediation decision had been reached for them. Data in red are the remaining 4 highest concentrations. Without checking whether different populations may be present, plugging the data in ProUCL resulted in 2 UCL estimates, both of which were less than 0.2 mg/kg, or below the soil-BaP residual contaminant level (RCL) of 0.211 mg/kg. This "averaging" led to a proposal of no further remediation even if RCL exceedances still remained.

Upon further analysis, however, both the Q-Q plot and the outlier test showed why the averaging and estimated UCLs may not be appropriate.



Conclusion: The 4 highest concentrations must not be included with the population of lower values. Since their levels exceed the RCL, they will still need to be addressed.

would result in UCL estimates that are less than 0.2 mg/kg. However, the Q-Q plot for the data shows the "disconnect" between the low values and the high values; the latter as a group has a distinctly steeper slope.

The Q-Q plot strongly suggests that the high values should not be grouped together with the low values. In addition to the Q-Q plot, because the majority of the data is "ND", it is fairly easy to show that the 4 highest values (all >0.211 mg/kg RCL) are outliers. Hence, the conclusion after our statistical analysis is that more scrutiny is needed where the 4 highest levels are located. They are RCL exceedances that cannot be included with the main population that are less than the RCL.

In this example, we illustrate that in many practical applications, it is not necessary that we proceed to determine the UCLs for every statistical population. To make a decision about averaging and RCL exceedances, it may likely suffice to show that concentrations exceeding the RCL standard are not from the same statistical population as those not exceeding the RCL.

Example 4: Straightforward Comparison between the DC RCL and the UCL

With this last example, Example 4 (on p. 16), we illustrate how it is possible that some data may exceed the DC RCL, but that no further remediation may be necessary because the UCL for the data set did not exceed the RCL. The prerequisites to such comparison include: 1.) Establishing statistically that the UCL was estimated for a dataset coming from a single population, and 2.) Estimating the UCL with no outliers present. When these prerequisites are met, then the UCL (rather than individual point concentrations) can be directly compared to the RCL. When the UCL does not exceed the RCL, then the RCL is met. Outliers exceeding the RCL, or other data from another population whose UCL exceeded the RCL, will still need to be addressed either with further investigation or remediation.

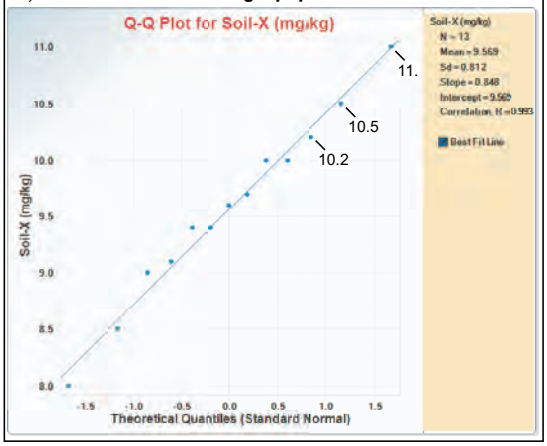
The data set for Example 4 has 13 point concentrations, with the highest 3 greater than the RCL (10 mg/kg) for contaminant X. The Q-Q plot shows that all 13 points fall nicely along a single straight line. Because the data set has less than 25 points, the Dixon's (not Rosner's) outlier test was used to determine whether outliers were present. The outlier test yielded no outliers in the dataset. So all of the data were then used to calculate the UCL. The resulting UCL was less than the RCL (10). This means that even though 3 individual points were above the RCL, no further work at the site is necessary because the average did not exceed the RCL standard.

Example 4 - Straightforward RCL-UCL Comparison

Data for Contaminant **X**
(RCL for X is **10** mg/kg.)
Red font indicates RCL exceedance.

Sample#	Soil-**X** (mg/kg)
1	8.5
2	9.7
3	10.
4	9.
5	8.
6	10.
7	9.1
8	10.5
9	11.
10	10.2
11	9.4
12	9.6
13	9.4

*1.) Q-Q Plot indicates a **single population**.*



*2.) Outlier test results in **no outlier**.*

Dixon's Outlier Test for Soil-X (mg/ kg)

Number of Observations = 13
10% critical value: 0.467
5% critical value: 0.521
1% critical value: 0.615

1. Observation Value 11 is a Potential Outlier (Upper Tail)?

Test Statistic: 0.316

For 10% significance level, 11 is not an outlier.
For 5% significance level, 11 is not an outlier.
For 1% significance level, 11 is not an outlier.

2. Observation Value 8 is a Potential Outlier (Lower Tail)?

Test Statistic: 0.400

For 10% significance level, 8 is not an outlier.
For 5% significance level, 8 is not an outlier.
For 1% significance level, 8 is not an outlier.

*3.) UCL (9.97) is **not greater** than RCL (10).*

General Statistics

Total Number of Observations	13	Number of Distinct Observations	11
		Number of Missing Observations	0
Minimum	8	Mean	9.569
Maximum	11	Median	9.6
SD	0.812	Std. Error of Mean	0.225
Coefficient of Variation	0.0848	Skewness	-0.227

Normal GOF Test

Shapiro Wilk Test Statistic	0.99	**Shapiro Wilk GOF Test**
5% Shapiro Wilk Critical Value	0.866	Data appear Normal at 5% Significance Level
Lilliefors Test Statistic	0.11	**Lilliefors GOF Test**
5% Lilliefors Critical Value	0.246	Data appear Normal at 5% Significance Level

Data appear Normal at 5% Significance Level

Assuming Normal Distribution

95% Normal UCL		**95% UCLs (Adjusted for Skewness)**	
95% Student's-t UCL	9.971	95% Adjusted-CLT UCL (Chen-1995)	9.924
		95% Modified-t UCL (Johnson-1978)	9.968

Suggested UCL to Use

95% Student's-t UCL	9.971

Note: Suggestions regarding the selection of a 95% UCL are provided to help the user to select the most appropriate 95% UCL.

Some post-remediation confirmation samples showed soil-X contaminant still exceeding the RCL: 3 largest concentrations > RCL of 10 mg/kg.

Statistical analysis demonstrates the following:
1.) A single population (Q-Q plot has correlation coefficient R > 0.99.)
2.) No outliers (Dixon's test shows neither of the extreme value is an outlier.)
3.) UCL < RCL.

ProUCL will calculate several UCL estimates based on several different distributions ProUCL assumes for a given set of data. Then it will provide a suggested UCL near the bottom of the output. For soil-X, a normal distribution sufficiently explains the data set; hence, its normal-distribution UCL can be used to compare to the RCL.

Conclusion: Even though there were individual samples that exceeded the RCL, the UCL for the data set was less. Hence, the statistical analysis indicates that they can remain at the site.

IV. Use of Background Threshold Values

Another option for statistically evaluating site data is to initially establish a background threshold value (BTV) for the contaminant of concern. This option may be appropriate when the contaminant is either naturally occurring or perhaps attributable to atmospheric deposition. This typically includes metals and PAHs. ProUCL provides several tools to analyze for the background threshold value (BTV). The BTV is typically set by the upper tolerance limit of the 90th or 95th percentile, or even the maximum of an acceptable background reference dataset. So the BTV is a statistic of higher-order than the UCL for the mean.

To illustrate what a BTV is compared to a UCL, we go back to the population partitioning shown in Example 2. After population partitioning, the background subset can be separated from other populations. In Example 2, the column "Cd-Pop-0" has the subset comprising the background data set. Its maximum concentration (12.39) is the BTV. The UCL value (11) is less than BTV. (Because there are less than 10 data points in Cd-Pop-0, running ProUCL to determine the UCL for the data set will give a warning about the small sample size. The UCL of 11 is calculated assuming the normal t-distribution for the data points.) Utilizing population partitioning to establish site-specific BTV can be a useful option at sites where the threshold level, like for the naturally-occurring background concentrations, may be set as the soil cleanup standard.

After a statewide background sampling in Wisconsin, a different method was used in establishing the statewide surficial soil-arsenic BTV because of the amount of NDs in the data set. The methodology for establishing a BTV is contained in the ProUCL Technical Guide when NDs (not detected) are plentiful in the background dataset. The USGS [*Stensvold*, 2012] had used ProUCL for evaluating the data during the development of a report to determine the distribution and variation of surficial soil-arsenic in Wisconsin. Nearly one-third (211 out of 664 sample results) of the soil-arsenic samples were NDs (< 1 mg/kg). The data set was evaluated for outliers and the data points identified as outliers were removed so as not to bias the higher-order statistics. The DNR Remediation and Redevelopment (RR) program defined the statewide BTV for soil-arsenic as the highest value from the dataset once the outliers have been removed. That value can then be categorically accepted as not exceeding background. The guidance document (RR-940) explaining how the statewide BTV for arsenic was established is available at: dnr.wi.gov/files/PDF/pubs/rr/RR940.pdf. The RR guidance has a table with the UCL (2.4 mg/kg) and BTV (8 mg/kg) for the outlier-free statewide soil-As background data set.

The statewide BTV definition used for arsenic (i.e., maximum without outliers) was extended to the other metals included in the USGS report. These statewide BTVs for arsenic and other metals are incorporated in the spreadsheet of RCLs available from the RR program. Since both RCLs and BTVs are available in the same spreadsheet, it is worth mentioning how they need to be applied when it comes to site UCLs. In practical application, a site's UCL can be compared to an RCL; however, that UCL should not be compared to the statewide BTV.

It is re-emphasized here that the UCL is not the same as the BTV. The comparison of a 95% UCL of one population with a 95th percentile or the maximum of another population (as the BTV can be defined) cannot be considered fair and reasonable as these limits (UCL and BTV) represent different parameters [*U.S. EPA*, 2013, p. 20].

A site-wide UCL for a contaminant is not needed for comparison to its respective statewide BTV. Each individual data point from the site should be compared to the statewide BTV to determine if that data point exceeds background or not. To the degree that an individual point is less than the BTV, that sample result would be considered background.

V. Conclusion

The use of statistics to evaluate environmental data requires a good working knowledge of statistics. Specialized assistance may be needed to appropriately use statistical averaging to comply with NR 720, Wis. Admin. Code.

The averaging approach outlined in this guidance document may be useful in some situations, where data is plentiful. However many sites will typically have insufficient data for statistical analysis to be used, and it will simply be unnecessary at many more sites where a reasonable and environmentally sound remedy can be identified through other methods.

It is strongly recommended that the Responsible Party discuss potential options with the DNR

Project Manager, before taking action, to help ensure the selected method is appropriate for the data set being evaluated. For more information contact Resty Pelayo at aristeo.pelayo@wisconsin.gov or (608) 267-3539.

VI. References

Carranza, E. J. M., *Geochemical Anomaly and Mineral Prospectivity Mapping in GIS, Handbook of Exploration and Environmental Geochemistry, Vol. 11* (M. Hale, Editor), The Netherlands: Elsevier B.V., 351 pp., 2009.

Everitt, B. S., An introduction to finite mixture distributions, *Statistical Methods in Medical Research, Vol. 5, No.2*, p. 107-127, 1996.

Fleischhauer, H. L., and N. E. Korte, Formulation of cleanup standards for trace elements with probability plots, *Environmental Management. Vol. 14, No. 1*, p. 95-105, 1990.

Fowlkes, E. B., Some methods for studying the mixture of two normal (lognormal) distributions, *Journal of the American Statistical Association, Vol. 74, No. 367*, p. 561-575, 1979.

Gilbert, R. O., *Statistical Methods for Environmental Pollution Monitoring*, New York: Van Nostrand Reinhold, 320 pp., 1987.

Holgersson, M., and U. Jorner, Decomposition of a mixture into normal components: A Review, *International Journal of Bio-Medical Computing , Vol. 9, No. 5*, p. 367-392, 1978.

Schilling, M. F., A. E. Watkins, and W. Watkins, Is human height bimodal?, *The American Statistician, August 2002, Vol. 56, No. 3*, p. 223-229, 2002.

Sinclair, A. J., A fundamental approach to threshold estimation in exploration geochemistry: Probability plots revisited, *Journal of Geochemical Exploration, Vol. 41*, p. 1-22, 1991.

Singh, A., A. K. Singh, and G. Flatman, Estimation of background levels of contaminants. *Mathematical Geology, Vol. 26, No. 3*, p. 361-388, 1994.

Stensvold, K. A., Distribution and variation of arsenic in Wisconsin surface soils, with data on other trace elements: *U.S. Geological Survey Scientific Investigations Report 2011–5202*, 41 pp., 1 app., 2012.

U.S. EPA, *ProUCL Version 5.0.00 Technical Guide, Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations*, U.S. Environmental Protection Agency Office of Research and Development,Washington, DC, EPA/600/R-07/041, 273 pp., 2013.