Webinar 1: ProUCL A to Z

59.: sorry, where do I get the data file?
* Moderator Jean Balent(privately): ProUCL can be downloaded for free from https://www.epa.gov/land-research/proucl-software
The sample dataset can be downloaded from https://clu-in.org/conf/tio/ProUCLAtoZ1/Zn-Cu-two-zones-NDs.xls

X 62.: How did you load the data file? File --> ??
See ProUCL 5.1 User Guide - Chapter 2: Entering and Manipulating Data

X 64.: datasets sometimes use -999 for missing values. would this work?
ProUCL perceives -999 as a numerical value and will be included in computations. All blank cells, cells with alphanumeric strings and characters and large number denoted as 1E31 will be accounted as missing values and ignored from the computations. Refer to ProUCL 5.1 User guide, Chapter 2 for more details.

65. This appears to be a different data set then was provided on the webinar page
* Moderator Jean Balent(privately): They updated the dataset back in Dec. It should be updated on the website.

X 66.: JUSTTO CONFIRM__Are missing data excluded by ProUCL from the Data set? In other words, you dont need to do the work, the program does it for you?
All blank cells, cells with alphanumeric strings and characters and large number denoted as 1E31 will be accounted as missing values and ignored from the computations. Refer to ProUCL 5.1 User guide, Chapter 2 for more details.

X 67.: Does ProUCL have a specific excel layout format needed for running the basic stats?
ProUCL User Guide Chapter 2 provides detailed guidance on formatting and importing data in ProUCL.

70.: Why did the course description insist up[on having the latest version of ProUCL installed (and having downloaded the data set...) if the presentation is gonna mopve too fast to serve as any kind of a tutorial?
* Moderator Jean Balent(privately): Thank you for the feedback. I will let the instructors know to try to slow down. Please know the entire session is being recorded so that you can play it back online. With the recording you can pause and restart the recording as needed which may be more helpful.

72.: My file when loaded does not show what yours does
* Moderator Jean Balent(privately): Thanks for reporting that. Are you using this dataset? The sample dataset can be downloaded from https://clu-in.org/conf/tio/ProUCLAtoZ1/Zn-Cu-two-zones-NDs.xls

75: FULLIRIS with NDs Presentation is the correct file?
* Moderator Jean Balent(privately): No, the correct dataset is here https://clu-in.org/conf/tio/ProUCLAtoZ1/#tabs-4

77.: The website has a link to the Iris flower dataset, not the one that is being used.
* Moderator Jean Balent(privately): Thanks for the question. The speakers updated the dataset to be used in Dec. Did you download the version listed here https://clu-in.org/conf/tio/ProUCLAtoZ1/#tabs-4

X 78.: do you have to include a certain number under the column for analyte in which you call out a ND in the corresponding ND column??
ProUCL 5.1 User Guide Chapter 1.11 and Chapter 2.8 discuss handling of non-detect observations.

X 80.: would you ever put the mdl (ex <0.0132) in the column insted of zero?
ProUCL understands any character string as missing value. You need to enter a numeric value to be included in

computations and create an additional variable to designates each entered value as detect or non-detect. ProUCL 5.1 User Guide Chapter 1.11 and Chapter 2.8 discuss handling of non-detect observations.

X 81.: Is the value in columns 1 and 2 the DL value when ND?
It can be either DL or reporting limit, depending on how data are reported. ProUCL 5.1 User Guide Chapter 1.11 and Chapter 2.8 discuss handling of non-detect observations.

X 85.: Can you export graphs?
ProUCL User Guide Chapter 17.1 provides instructions on how to copy and save graphs.

87.: Outlier tests - in ProUCL assume normality without outliers
* Moderator Jean Balent(privately): Thanks for the comment.

88.: It is incorrectly assumed that outlier tests assume normality on data with outliers. It is recommended to use graphical displays with outlier tests. A data set with outliers very seldom follow sa normal distribution.
* Moderator Jean Balent(privately): Thanks for the comment.

X 89.: what if you are trying to separate a contaminated area fron naturally occurring background? Shouldn't one remove the "high" outliers as they might be representing "contamination"?
The outcome of outlier test indicates suspected extreme values that may misrepresent the population and need to be investigated before removed from the data set. These values may indicate more variability than expected, such as extreme population values, on-site hot-spots, spots with high naturally occurring concentrations or multiple soil types in background area. Please note that not removing outliers or removing false outliers both lead to distorted estimates of population parameters. Therefore 5 step approach explained in EPA guidance QA G-9S Data Quality Assessment: Statistical Methods for Practitioners should be followed:

1. Identify extreme values that may be potential outliers;
2. Apply statistical test;
3. Scientifically review statistical outliers and decide on their disposition;
4. Conduct data analyses with and without statistical outliers; and
5. Document the entire process.

X 92.: Nevermind I understand what the grouping variable means now
ProUCL 5.1 User Guide Chapter 1.8 provides information on use of grouping variables.

93.: It's incorrect to assume that high concentrations aren't naturally occuring (aren't part of the background population).
* Moderator Jean Balent(privately): Thanks for the comment!

95.: can you share contact info for the speakers for additional questions?
* Moderator Jean Balent(privately): The speakers are listed at https://clu-in.org/conf/tio/ProUCLAtoZ1/#tabs-2

96.: there is no email address for Anita Singh who is the person who I'd like to connect with
* Felicia Barnett: If you wish to ask something of Anita. Please contact me with the questions and I will try to get it to her.

X 102.: to what extent is ProUCL valid for use on data with a spatial trend?
ProUCL is generally not going to be the best tool for spatial data as trend analysis is limited to a single independent variable generally reserved for time.

X 103.: Does US EPA still endorse or promote geostatistics software such as the 1989 vintage GeoEAS?
EPA as a matter of policy does not endorse or approve of any commercial products. It is up to user to determine if

any product is qualified and correct for their specific need.

X 104.: Can someone help me find the setup.zip in the overall zip file? I'm trying to follow along
ProUCL downloads are available here: https://www.epa.gov/land-research/proucl-software

X 105.: Is there a way to determine if the statistical output made use of your nondetect values?
Most of the statistical outputs will have a table discussing the non-detects as well as how their values were handled or imputed. However, as long as you select "With NDs" from the dropdown of whatever statistical tool you are using within ProUCL it should be utilizing them.

X 106.: What indicators would you see in the data that would lead to you Log transform the data?
Log transformation is commonly used to deal with skewed data and is performed to achieve data symmetry. However, this approach has some major problems. Refer to ProUCL 5.1 Technical guide Chapter 4.2.2 for further discussion on problems with data transformation in environmental applications.

One example would be in trend analysis seeing a curve over time that is not very linear might get you to try that transformation. If your constant variance assumptions for your residuals of OLS regression seem to be being violated a log transformation is a nice place to start there as well.

X 107.: ...or if it treated the NDs as actual values.
Answer in #105

X 108.: how do you set min and max NDs?
You don't necessarily "set" min and max NDs in ProUCL. The max ND value returned on the boxplot simply represents the highest ND value included in your graphed data. ProUCL 5.1 User Guide Chapter 1.11 and Chapter 2.8 provides more information on how ProUCL handles of non-detect observations.

X 109: mine does not show the "cleanup standards line?
Refer to ProUCL 5.1 User Guide page 92 – Options_Boxplot for instructions how to add a reference line to box plot.

X 110.: When the output provides a "Suggested UCL to Use" but also specifies "Warning: Recommended UCL exceeds the maximum observation," would you recommend that we use the suggested UCL or the maximum observation (assuming we cannot take more samples)?
UCL > max indicates that there are issues with the data set. We recommend consulting a statistician since use of more advanced statistical methods that are beyond the capability of ProUCL may be needed to better support the decision making.

X. 111.: Is there a rule of thumb for determining normality vs skewness?
Normality can be assessed using the provided GOF tests in ProUCL as well as checking qq plots. Skewness is more of a sliding scale. No skew is 0 positive numbers represent a right skew with negative representing a left skew. I start caring about it around -1 or 1 but that's not a hard and fast rule.

X 113.: Can Pro-UCL process J-coded values?
W There is no special processing step for J-coded values in ProUCL.

X 114.: How does it perform intrawell statistics?
In general analysis of interwell / intrawell data requires more advanced statistical approaches. ProUCL has only limited capabilities for such analysis.
ANOVA (Chapter 13 in ProUCL User Guide and Chapter 9 in Technical Guide) One-way ANOVA is used to compare means (or medians) of multiple groups such as comparing mean concentrations of areas of concern and to perform inter-well comparisons.

Trend Analysis (Chapter 14 in ProUCL User Guide and Chapter 10 in Technical Guide) The ordinary least squares (OLS) regression model, trend tests, and time series plots are used to identify upwards or downwards trends potentially present in constituent concentrations identified in wells over a certain period of time. The Trend Analysis module performs the M-K trend test and Theil-Sen (T-S) trend test on data sets with missing values; and generates trend graphs displaying a parametric OLS regression line and nonparametric T-S trend line. The Time Series Plots option can be used to compare multiple time-series data sets.

115.: My spreadsheet gives me an error when I try to load it. Not sure what I'm doing wrong.
116.: Also, just a note, makes it easier if you put the units in the header so you don't have to guess.
X 117.: You said character strings indicate a blank cell - so how is the Zone column set up, or how are columns based on characters okay?
Strings in numeric variables such as concentrations will be dismissed as missing. However, columns designating grouping variables will not do the same. For example, a vector [1,2,3,"test"] used as a grouping variable would produce 4 different groups whereas. If you used the same vector for numerical variable (e.g. concentrations) instead of using it as a grouping variable, ProUCL will understand "test" as missing value.

X 118.: For a site with many wells, can I replace zone with well ID?
The header name doesn't matter much; you'll just need to select whatever column you desire as your grouping variable.

X 119.: What is the smallest number of observations (i.e. detections and non-detections) required for ProUCL?
This varies depending on the statistical task at hand. Please refer to the ProUCL user or tech guides for specifics on a given task.

X 120.: Are the censored data being plotted on the Q-Q plot using robust methods, like the Kaplan-Meier method?
They are plotted as inverted triangles but do not receive a different imputed value.

X 121.: How do you get 95%UCL of the 99th percentile?
That is a 95-99UTL and using the methods discussed in the 3rd presentation you will be able to produce it.

X 122.: When UCL>maximum, would you agree that UCL is a better estimate of the exposure point concentration (EPC) than the max?
First check your sample sizes against test assumptions as this can occur when your sample sizes do not meet minimum requirements. If that does not help your problem, dealing with it on a site/problem level basis is recommended.

X 123 Does ProUCL include censored goodness-of-fit tests?
The goodness of fit tests in ProUCL are conducted on detected data.

X 124.: can you please demonstrate how to use the Trend Analysis
This was demonstrated in 2nd webinar of ProUCL series: Trend Analysis

X 125.: What can one conclude when there is nothing in the GOF "Conclusion ..." cells?
When Conclusion cells do not have any content, a message will display at the very bottom of the output: "Data do not follow a discernible distribution at (0.05) Level of Significance". We recommend consulting statistician regarding further steps in data analysis.

X 126.: When would you decide NOT to use your NDs? (Ii.e. run without NDs)
Please refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations.

X 127.: what does it mean to make the grouping variable "zone"?

That was just the name of the variable I chose to group on. The variable name is not relevant, just use the column you want your data grouped by. ProUCL 5.1 User Guide Chapter 1.8 provides detail information on use on grouping variables.

128.: In addition to determine normality/skewness of data sets, Q-Q plots are very useful to compare data from multiple populations. One can use Q-Q plots to compare Cu of alluvian fan and basin trough

X 129.: This sample dataset consists of nondetects with values at multiple detection limits. Can this software be used for data with a single detection limit (i.e. per chemical)?
Yes. Please refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations.

X 130.: I thought most recent EPA guidance did not support DL/2 due to bias?
If your percentage of NDs is greater than 15% statistical methods such as Kaplan Meier are preferred, but in cases of <15% NDs where there is a reason to believe DL/2 or another arbitrary placement is appropriate, those methods may be used. Please also refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations and ProUCL 5.1 Technical Guide for information on analysis of data sets with non-detects with specific statistical methods

131.: For older tests such as Rosner test and Dixon test - they assumed normality to use critical value - this is the reason - robust statistical methods such as Bi-weight and other robuyst outlier methods are preferred.
132: I would like you to mention -that if you want to use Rosner and Dixon - do not say normality is required. As I said - a data set with outliers does not follow a normal distribution

X 133.: DL/2 is not current EPA guidance for background. ProUCL notes this with messages that state "DL/2 is not a recommended method. DL/2 provided for comparisons and historical reasons." Speaker should correct his statement.
Generally speaking, you will be using statistical methods such as Kaplan Meier when you have an ND percentage between 15% and 50% of the data. However, when your ND percentage is <15% use of an arbitrary placement such as DL/2 or 0 could be warranted given specific site/analyte scenarios.

X 134.: USEPA's 2009 Statistical Analysis of Groundwater Monitoring Data at RCRA facilities Unified Guidance document has the 15% and 50% Non-Detect Rule on how to handle non-detect concentrations. Can the speaker talk to how ProUCL is used with this rule?
Please refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations and ProUCL 5.1 Technical Guide for information on analysis of data sets with non-detects with specific statistical methods.

X 135.: When including nondetects in hypothesis tests, what values is the software using for nondetects? i.e. how are nondetects being treated mathematically?
Please refer to ProUCL 5.1 Technical Guide chapter 6 for information on analysis of data sets with non-detects with specific statistical methods.

X 136.: For the two-sample hypothesis testing in ProUCL 5.1, why is the quantile test no longer included?
Because the quantile test is comparing a set of data to a single value not another set of data

X 137.: Any tips for whether to use Gehan's test or Tarone-Ware test for data sets with NDs?
Please refer to ProUCL 5.1 Technical Guide Chapters 6.9.2 and 6.9.3

X. 138.: When will this presentation be available again?
It is being recorded and is available on the EPA website any time.

X 139.: can you change the name of the files in the navigation panel?

Please refer to ProUCL 5.1 User Guide Chapter 2. 6 Saving Files.

X 140.: Can you provide an example using date fields and seasons trend analysis
Trend analysis is discussed in webminar Trend Analysis.

X 141. Please provide recommendations for analysis and potential removal of nondetected data that may be outliers.
Just like detects, non-detects suspected of being outliers should be scrutinized in the ways discussed in the presentation

X 142.: Can it distinguish between MDL and PQL for non-detects?
No, it cannot. Please refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations.

X 143.: For NDs, is the value listed under Cu = to the lab ND.
This depends how data were reported. Please refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations.

X 144.: Am I understanding your correctly that the outlier tests included in ProUCL are not valid for nonparametric data?
The outlier tests built into ProUCL assume normality of the data without the outliers in question.

X 145.: What is your opinion regarding how to handle non detects: 1/2 the DL or use PRo UCL?
Please refer to ProUCL 5.1 User Guide Chapter 1.11 for detailed discussion on samples with non-detect observations and ProUCL 5.1 Technical Guide for information on analysis of data sets with non-detects with specific statistical methods.

146.: How are there values for non-detects?
X 147: Can this method also be applied to groundwater data?
I'm unclear what "this method" is referring to, but tests can be run for any application as long as the data meets test requirements such as sample size or distribution, and the test is applicable to the problem.