Determination of Statistically Significant Increases over Background Concentration Levels

May Raad

HDR, 2540 Research Park Drive, Lexington, KY 40511

CONFERENCE: 2017 World of Coal Ash – (www.worldofcoalash.org)

KEYWORDS: background concentration levels, statistically significant increase, outliers, non-detects, parametric and non-parametric upper prediction limits

ABSTRACT

HDR is currently supporting its power clients to comply with the Environmental Protection Agency’s Hazardous and Solid Waste Management System; Disposal of Coal Combustion Residuals from Electric Utilities; Final Rule (CCR Rule). The groundwater sampling and analysis requirements of Part 257.93 form a critical step in the compliance process; however, the implementation is often a challenge since multiple statistical methods are possible. It is not readily apparent which method is best for testing for statistically significant increases over background levels and how the issues of non-detects, parametric versus non-parametric sample distributions and small sample sizes are tackled. HDR has developed a methodology to efficiently assess the background sample data in line with the rules found in Part 257.93 and produce site-specific test statistics based on the background concentration levels using the upper prediction limit (UPL) for detection monitoring (Part 257.93(f)(3)). HDR has selected the UPL since this statistic has many intuitive features and can adjust for the increase in the site-wide false positive rate during detection monitoring. Attributes of HDR’s robust methodology include the statistical methods to detect for normal, lognormal and gamma distributions and methods to impute the values of non-detects with substantially less bias in the estimates than when the simple substitution method is used. The methodology draws on sound statistical principles and addresses the challenges due to small sample sizes, outliers, varying distributional forms of the samples, serial correlation, and trends to ultimately produce test statistics which can flag statistically significant increases over background to the best extent possible given the available data.

INTRODUCTION

Part 257.93 of the Hazardous and Solid Waste Management System; Disposal of Coal Combustion Residuals from Electric Utilities; Final Rule (CCR Rule) describes the sampling and analytical requirements to ensure monitoring results can provide accurate
representation of groundwater quality at the background and downgradient wells. A key requirement in the CCR Rule is that background sample results are statistically analyzed for the objective of testing results from downgradient wells for statistically significant increases (SSI) over background levels. An important point, which is not readily apparent in the CCR Rule, is that the statistically derived background levels are themselves components of statistical tests of hypothesis or statistical conjectures of random events.

The objective of the detection monitoring phase of the CCR Rule is to determine if the groundwater samples from downgradient wells exhibit higher than expected concentrations of Appendix III constituents compared to background concentrations due to possible impacts from the CCR unit(s). The statistical analysis used to achieve this objective builds toward testing the ‘null’ hypothesis that there are no downgradient impacts versus the ‘alternative’ hypothesis that there may be downgradient impacts caused by the CCR unit(s). If there is sufficient statistical evidence to reject the null hypothesis in favor of the alternative for an Appendix III constituent, then there is an SSI indicating the potential presence of groundwater impacts. The CCR rule (257.94(e(1)(2))) then specifies that the owner or operator of the CCR unit must establish an assessment monitoring program within 90 days of detecting such an increase—provided the owner or operator can not rule out another source as causing the SSI.

HDR has developed a robust methodology that draws from the methods presented in the U.S Environmental Protection Agency’s (U.S. EPA) “Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Unified Guidance” (Unified Guidance) and research documents “The Lognormal Distribution in Environmental Applications” and “ProUCL Version 5.0.00 Technical Guide” to efficiently and effectively assess background sample results in order to test for SSIs at downgradient wells using the upper prediction limit (UPL). The CCR Rule provides several statistical testing options such as control charts, parametric or non-parametric analysis of variance tests (ANOVA), and both upper prediction and tolerance limits. These available tests can detect for statistical differences between two different samples (e.g., background and downgradient samples); however, certain conditions must be met for the test results to be defensible. Given the variable nature of background concentrations, the small sample sizes, and the need for multiple future comparisons to be made at the downgradient wells, it is often a challenge for assumptions to be met.

For example, control charts are a visually appealing means to flag SSIs, but for the results to be actionable, the data must have normal or lognormal distributional properties. If the distributional properties of a sample fall into the non-parametric realm, which is often the case in groundwater samples, then non-parametric control chart method options are limited; however, research in this area has been recently growing.¹ Until recently, control chart methods did not address the site-wide false positive rate problem when testing multiple well-constituent pairs. Monte Carlo simulations may be necessary for practitioners to derive the optimal control chart limits (e.g., SCL, k, h) and re-testing strategies particular to their background and downgradient sample sizes.³
The use of ANOVA assumes stringent assumptions that both background and downgradient distributions follow the same distributions of either normal or lognormal and have equality of variances. If these assumptions fail, then non-parametric ANOVA methods are the recourse to test for differences in populations resulting in a loss of statistical power to detect for an increase in concentrations when an increase has actually occurred. Even if normality or lognormality conditions hold for the ANOVA, a large number of multiple comparisons from the downgradient wells can also cause the test to lose statistical power.\textsuperscript{13}

Upper tolerance limits (UTLs) and UPLs flag if outliers are from population distributions different from that of background. The Unified Guidance has recommended that the UTLs are more appropriate statistics to develop ‘fixed’ reference values that can be used as a type of groundwater protection standard (GWPS) where one does not currently exist.\textsuperscript{13}

UTLs are different from UPLs because they lack the statistical properties that allow practitioners to modify the tests in order to control the site-wide false positive rate. As multiple comparisons of downgradient sample results are made to the background levels, the chance of falsely declaring an SSI at the site increases. Without careful planning for the detection monitoring phase of the CCR Rule, the Unified Guidance’s recommended site-wide false positive rate of 10 percent\textsuperscript{13} may be exceeded, causing needless and costly investigations into the apparent SSI, and may also shift the owner or operator of the CCR unit into the assessment monitoring phase.

METHODOLOGY TO DETERMINE STATISTICALLY SIGNIFICANT INCREASES OVER BACKGROUND CONCENTRATION LEVELS USING THE UPPER PREDICTION LIMIT

As with any statistical test, the UPL is no different, certain assumptions must be met for the tests to be valid. Issues related to small sample sizes, outliers, presence of non-detects, spatial variability, trending patterns in background sample data, and the distributional form of measurements must all be addressed prior to selecting the correct form of the UPL. In actuality, this process can be used to produce defensible results for ANOVA, control chart, tolerance, or prediction interval procedures provided their respective statistical assumptions have been met. The series of steps to assess the background data prior to producing the UPLs is termed the preliminary data analysis (PDA).

While the PDA is broken down into sequential steps, it is actually an iterative process. At any step, a new finding may initiate a re-check from one or all of the previous steps. As more background samples are collected over time, a point is reached where the owner or operator of a CCR unit may want to update background levels and take advantage of the added accuracy that larger background samples can provide.
The PDA initially evaluates the sample sizes from the background wells. After the collection and analysis of the minimum eight background samples, by October 17, 2017, we study sample characteristics such as percentage of non-detects, presence of outliers, serial correlation of observations over time, seasonality, spatial variability, and trending patterns over time with or without non-detects. The distributional forms of either normal, lognormal, or gamma are then selected based on goodness-of-fit tests if the data supports parametric distributional forms; otherwise, we assume no discernable distribution based on the data and use non-parametric methods to test the hypothesis of whether or not groundwater downgradient of the CCR unit(s) has been impacted.

Once we have reached an understanding of the sample’s characteristics and know which distributions are applicable, we proceed to produce the UPLs at a particular significance level (i.e., the test significance level). The exact value for this significance level is a function of the number of downgradient wells at the site, the number of constituents to be tested, the number of statistical evaluations to be completed in a specific time period (typically a year), and whether re-testing strategies are incorporated into the design such that the aggregation of the false positive rates over all the multiple comparisons does not exceed 10 percent, (i.e., the cumulative significance level).

BACKGROUND SAMPLE SIZE CONSIDERATIONS

As a general rule of thumb, sufficient samples for the purpose of statistical testing should have a minimum of 25-30 independent samples under central limit theorem assumptions. Realistically, the availability of such numbers from sampling background wells at CCR sites may not be possible. The EPA’s Unified Guidance has acknowledged this reality, and has made the recommendation that the minimum sample size per constituent of concern from a field area known to be free of anthropogenic impacts is eight to ten samples. The CCR rule has incorporated this guidance in its rules as a minimum requirement for background sample sizes (see Part 257.94(b)).

Technically, a minimum of eight samples will support the estimation of UPLs; however, the presence of censored observations or non-detects restricts the usefulness of small samples. An observation is considered censored as a result of the limitation of the measurement process. Typically, censored observations in water quality samples are ‘left-censored,’ meaning values are assumed to fall within a certain range of concentration values such as 0 to the method detection limit (MDL).

Table 1 below summarizes the impact the sample size and the level of censorship on the sample has on the process to test for SSIs over background levels using the UPLs.
Table 1: Impact on Sample Size and Percentage of Non-Detects on Testing for Statistically Significant Increase over Background Levels

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>0 -&lt; 50% Non-detects</th>
<th>50 -&lt; 100% Non-detects</th>
<th>100% Non-detects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than eight</td>
<td>Insufficient sample size quality, select largest detected value</td>
<td>Insufficient sample size quality, select largest detected value</td>
<td>DQR*</td>
</tr>
<tr>
<td>Eight or more</td>
<td>Conduct PDA to determine distributional form of measurements prior to producing the UPLs</td>
<td>Conduct PDA using non-parametric forms of tests, produce non-parametric UPLs</td>
<td>DQR</td>
</tr>
</tbody>
</table>

*Double quantification rule

For those samples with less than eight observations, if 50 percent to less than 100 percent of the eight measurements consists of non-detects, then the amount of information from the pool of sample results is limited. In these situations, the largest detected value can be used to represent the critical value for testing an SSI as a placeholder until more samples are collected. Regardless of the background sample size, when 100 percent of the measurements are non-detects, then the double quantification rule (DQR) can be used to test for an SSI. According to the Unified Guidance, 2009, “A confirmed exceedance is registered if any well-constituent pair in the ‘100% non-detect’ group exhibits quantified measurements (i.e., at or above the reporting limit [RL]) in two consecutive sample and resample events.”

For samples of eight or more, imputation methods such as Kaplan-Meier or regression on order statistics (ROS) can be used to fill in the gaps created when samples have less than 50 percent of the observations as censored. Even if the owner or operator of a CCR unit has the benefit of a large number of background samples, the presence of non-detects will impact how the sample results can be analyzed statistically. If 50 percent or more of the sample is censored, then non-parametric techniques should be used to conduct the statistical analysis including the production of UPLs. Imputation of the non-detects in this situation using the above mentioned techniques is not recommended. The practice of substituting a non-detect with either the face value or half the value of the method detection limit (MDL), or the reporting limit (RL) leads to bias in interpreting test results and should not be done.

During the initial detection monitoring phase commencing in October 2017, owners and operators of the CCR unit will continue to collect samples from the background and downgradient wells on a semi-annual basis (Part 257.94(c)). Over time, the background sample size will increase. The CCR Rule does not require that background concentration levels be updated at a specific time interval; however, it is typically in the best interest of the owner or operator of the CCR unit to update background concentration levels on a periodic basis. The periodicity of the update is dependent on having sufficient ‘new’ background samples to compare to the baseline background.
samples from the initial detection monitoring phase. If the ‘new’ background samples exhibit similar distributional properties as the ‘baseline’ background samples using statistical tests, such as the test for similar population means using the student’s t-tests or similar ranks such as the Kruskal-Wallis test, then the test statistics from the background concentration levels can be re-computed using the entire pool of background samples. While the minimum sample size for statistical testing is eight based on the guidelines in the Unified Guidance, 2009 and the CCR Rule, for practical purposes given the semi-annual scheduling of the sampling events, an update could be done after at least four sampling events to cover a two year period. The two year period can provide enough additional data for statistical testing purposes and cover variability related to possible temporal effects. Updates to the background concentration levels will improve the accuracy of the tests used to assess if there are downgradient impacts caused by a CCR unit.

DETECTING OUTLIERS

A statistical outlier is defined as a value originating from a different statistical population than that of the rest of the sample pool. Outliers or observations not derived from the same population as the rest of the samples violate the basic statistical assumption of identically-distributed measurements. If an outlier is suspected, an initial step is to plot the observations over time. The visual display of both the detected and non-detected data can make apparent anomalous values or patterns in the data.

Two statistical tests are available to test for possible outliers. Dixon’s outlier test is appropriate for data series with sample sizes less than 25, and Rosner’s outlier test is applicable to those with a sample size of 25 or larger. These outlier tests assume that the rest of the data except for the suspect observation(s) are normally distributed.

In conjunction with Dixon’s and Rosner’s tests, additional analyses can be performed using box-plots and quantile-quantile (Q-Q) plots. Box-plots represent the central tendency, spread and outliers from a sample of observations. The ‘box’ contains a line in the middle or the 50th percentile, which represents the value where 50 percent of the observations fall below. The lower edge of the box represents the 25th percentile and the upper edge represents the 75th percentile. The difference between the 75th and the 25th is termed the inter-quartile range or IQR. The IQR is used to illustrate the ‘whiskers’ of the box-plot. The value for the upper whisker line is calculated by multiplying the IQR by 1.5 and then adding that value to the 75th percentile. Correspondingly, the value for the lower whisker line is 1.5 times the IQR less the 25th percentile. Observations below or above the whiskers are considered statistical outliers. Values that are higher or lower than three times the IQR are flagged and considered severe outliers.

Q-Q plots provide graphical means of gauging how closely the data follow the probability distribution of interest, such as the normal, lognormal, or gamma. Q-Q plots are interpreted by determining the degree to which the points follow the straight line, which represents the distribution of interest. Gaps in the series of points and/or jumps above or below the straight line can indicate outliers or multiple distributions.
Figure 1 and Figure 2 demonstrate visual aids that help a practitioner determine if this particular constituent, with 31 observations from samples collected over a period of time, has outliers. Rosner’s test indicated that the values of 0.199 and 0.165 were possible outliers at the 1% level of significance. The two dots over the top whisker have been flagged as outliers using the box-plot method for outlier detection. The red one is a ‘severe’ outlier and the orange one a ‘moderate’ outlier. The Q-Q plot of the observed values compared to the theoretical distribution line assuming a lognormal distribution indicates that the two highest values are not markedly different from other observations that have deviated from the line. This distribution is non-parametric. In this particular case, all values from the sample remained to characterize background distributions.

Figure 1: Box-Plot with Outliers
If outliers are found from testing, the outliers should be investigated further by the project scientist(s). If they are correct values and collected under standard, consistent protocols, they should likely remain in the data series. However, if there are uncertainties about the quality of the data, the outliers can be dropped before proceeding. Some distributions naturally have anomalously low or high values. The subsequent tests for distribution types should find the best fitting distribution that can explain the anomalous values. Dropping outliers simply based on statistical test results, especially during the early phases of the CCR Rule groundwater monitoring program when sample sizes are small, may unintentionally change the assumed distribution that explains the data. As additional background samples are collected over time, outlier test results may change and earlier observations thought to be outliers may no longer be outliers.

DETECTING SPATIAL VARIABILITY

The CCR Rule specifies that a minimum of one upgradient well should be installed to derive background levels ((Part 257.91 (c)(1)). For instance, when the hydrogeology and site formation suggests multiple upgradient wells should be installed (or if existing, included in the sampling program), the assumption is that concentrations of constituents measured at these wells over time when pooled represent an estimate of overall well field conditions for those constituents. This assumption implies the variability of the concentrations per well is comparable and the values are independent of each other. A way to test this assumption is to study the spatial variability of the observations across the wells. Spatial variability exists when the distribution or pattern of concentration measurements changes between well locations, either from natural or anthropogenic factors. To evaluate the potential for spatial variability in background wells, side-by-side box plots can be used for each constituent at each well. If the variation is significant
from a visual perspective, both parametric and non-parametric ANOVA tests can be used to flag the differences in mean or median levels to corroborate the visual cues. An investigation by the project scientist(s) would then be performed to determine if the variation is from natural\(^1\) or anthropogenic sources.

The Unified Guidance, 2009 suggests if natural variation is markedly apparent across the background wells, then testing resorts to intrawell testing at the downgradient wells;\(^{13}\) however, wording in the CCR Rule implies an expectation that the selected locations of the background and downgradient wells will yield observations that are spatially invariant across the wells.

Intrawell testing is not practical for existing CCR unit sites. The majority of CCR units in the United States are existing landfills and unlined impoundments, and have been so for many years. For unlined landfills and impoundments, it is unlikely there are no impacts downgradient of the CCR unit(s) and that the groundwater quality immediately downgradient of the waste boundary could be considered representative of groundwater conditions prior to the construction of the CCR unit(s). If practitioners determine spatial variability is an issue for any reason at the upgradient, background wells, then a re-evaluation of the area should be performed to determine which of the wells may need to be excluded from the background analyses.

Figure 3 provides an example of a constituent sampled at eight different wells. The width of the boxes and the relatively consistent alignment of the median values across the wells suggests spatial invariability at this site even with the presence of outliers. Statistical test results from ANOVA and Kruskal-Wallis methods showed there was no evidence of different distributions among the wells.

\(^1\) Natural spatial variability refers to a pattern of changing mean levels in groundwater associated with normal geochemical behavior unaffected by human activities such as variation in contents of constituents in the subsurface and variation in geochemical conditions resulting in different solubility of the constituents.
SERIAL CORRELATION

Serial correlation describes a common issue with data collected over time, as done with groundwater sampling, whereas the values themselves are related to each other over various time intervals. Sources for serial correlation in groundwater samples can be due to temporal or seasonal effects related to the timing of the sample collection. The statistical functions used to characterize the background sample and to test if an SSI has occurred assume the background observations sampled over time are independent of each other. To test if the background samples are serially correlated, several statistical tests are available. For parametric distributions, the sample autocorrelation function flags if adjacent correlations of various lags are statistically significant or not. For non-parametric tests of randomness, various tests are available such as the ordinary run test, the sign test and the Bartel’s rank test which is the rank version of the Von Neumann’s Ratio Test.\textsuperscript{13}

A main reason why we may observe serial correlation in background samples is from the seasonal changes as evidenced from varying temperatures and precipitation. These
changes may impact water quality in a predictable and cyclical manner over time. The study of water quality changes over time is focused on the ability to discern true trends through regression analysis amidst the cyclical nature of the data, its “seasonality”, or any type of temporal effect. The correct use of these regression analyses rests on the crucial assumption that regression errors or residuals arising from the model fitting are independent of each other. This is often not the case with data that is seasonal or temporal in nature. If serial correlation exists either in its lagged form (e.g., a value at time ‘t’ is correlated to the value at time t-1) or by seasons (values in the spring of one year are correlated to spring values from previous years), then the sample autocorrelation function (ACF) test will identify the behavior. To better understand the type of seasonality (e.g., quarterly, bi-annually, etc.), which could factor into the observed variability of a time series, a visual inspection using side-by-side box plots of the data as a function of time is recommended.

To illustrate how the ACF can be used to detect the presence of serial correlation, the ACF is plotted by time lag and is demonstrated in a ‘correlogram’ denoted in Figure 4 using data from a constituent with 58 observations sampled across seven wells on a quarterly schedule (selected months within a quarter varied). The ACF value of 0.349 at lag 4 is statistically significant at the 5 percent level of significance. There may be a seasonal effect as values selected at a particular month are correlated to values observed four months prior (e.g., values measured in August are related to values measured in April).

![Figure 4: Correlogram Plot of the Sample Autocorrelation Function by Time Lag](image)

To further investigate this possible seasonal impact, a side-by-side plot of the concentrations by the four seasons is a good visual tool. The plot in Figure 5 shows
variability across the seasons. Both ANOVA and Kruskal-Wallis tests for differences in means or medians across the seasons resulted in statistically different populations by season. If serial correlation is present in the data, techniques are available to reduce their impacts such as deseasonalizing, differencing observations by lagged intervals or changing the frequency of sampling.8,9,13

**Figure 5 Side-by-side Box-plots Which Demonstrate Seasonality**

With the initial minimum of eight samples for the CCR Rule initial detection monitoring phase, it will be difficult to test for serial correlation and seasonality. Sample sizes of 25-30 or even more than 50 are recommended for the sample autocorrelation function results to be unbiased.2 Sampling events after October of 2017 are planned on a semi-annual basis. Even then, variations on the schedule are possible depending on the sites and whether or not assessment monitoring is needed. Several years of collection may be needed before sufficient data exists to test for serial correlation or seasonality using the ACF. Until sufficient samples are collected, the non-parametric test equivalents can be used, as required for the CCR Rule initial detection monitoring phase.
IMPUTATION OF NON-DETECTS

While efforts are made to collect and measure samples at regular intervals over time at the background and downgradient wells, not every observation is available due to either the inability of the laboratory test or method to detect a concentration from the sample, or a sample was not taken at its scheduled time. These instances of unknown concentrations are a challenge for the statistical analysis of water quality data. For most constituents being tested, there are established method detection limits (MDLs) below which the equipment or process can not accurately measure. In those cases, the value for the MDL is recorded in the database and that observation is flagged as being below the MDL. From a statistical perspective, those observations are treated as censored values. Within the water quality research field, censored values are also referred to as non-detects while measureable values are referred to as detects.

As soon as censoring appears in significant numbers (more than 15 percent), parametric distribution assumptions often fail and application of these tests can produce biased results. Historically, scientists and practitioners have used simple substitution methods for censored values. The value of the MDL or a fraction of the MDL (often ½) is used to represent the actual value. These simple substitution approaches introduce bias into the statistical analysis of water quality. Even though this approach has been deemed to be inferior since the 1990’s, it continues to be used today by mainstream practitioners. 3,4,6,10

Significant progress has been made in this area of research, specifically in the area of imputation of non-detects. Statistical tests and user-friendly, inexpensive software have been developed to detect trends and estimate background levels with the presence of censoring. The research is timely with the growth in water quality databases and changes in analytical methods over time. Unlike fifty years ago, practitioners are more likely to deal with multiple values of MDLs within a constituent’s series.

Imputation methods such as the regression on order statistics (ROS), Kaplan-Meier method and maximum likelihood estimation regression offer the advantage of a substitution method for non-detects that does not give all samples the same value, thereby reducing the impact of bias. ROS is a simple imputation method that fills in non-detect data on the basis of a probability plot of detects allowing practitioners to test best fits under normal, lognormal, and gamma probability distribution assumptions.10,13

The Kaplan-Meier method is the standard method for computing statistics with censored data for medical and industrial statistics, and has been successfully adapted to analyze water quality data in the presence of censoring. 6,13 The hybrid variation of the Kaplan-Meier method factors in the assumed distributional properties including normal, lognormal, or gamma distributions. This method can accommodate data with multiple MDLs, along with both known and unknown distribution types.10

With respect to analyzing trends over time using samples with non-detects, the maximum likelihood estimation (MLE) regression technique solves for the mean and
standard deviation most likely to have produced both the set of detected and censored or non-detected observations. It uses information from the detected observations and the proportions falling below each MDL (there may be more than one) to fit the regression curve.

**Trend Analysis**

The samples from background wells represent water quality conditions that exhibit natural variability and are unaffected by anthropogenic activities. As such, the measurements taken at regular intervals over time are expected to demonstrate a steady or stationary time series. The data from the background wells should be tested to determine whether trends exist (values steadily increasing or steadily decreasing in a statistically significant manner). Depending on the presence of non-detects or seasonality, either the MLE Regression method (with or without non-detects) or the Mann-Kendall non-parametric trend test is used to detect trends. While there are a multitude of methods to test for trends, such as the Seasonal Kendall regression test\(^5\), Cochrane-Orcutt procedure and ordinary least squares,\(^9\) we have found the MLE and Mann-Kendall test methods to be versatile in their applicability to water quality data. MLE is suitable for samples with known distributional patterns such as normal, lognormal, or gamma, and has the added benefit of being able to accommodate samples with different values of MDLs. The Mann-Kendall test is suitable for data series with no discernable distributions, and if there is censorship, that there is only one value of the MDL. If the data series exhibits significant seasonal effects, deseasonalizing can be done prior to running either the parametric or non-parametric tests for trends. While there are a multitude of techniques to deseasonalize, the Unified Guidance, 2009 provides a straight-forward method to remove the effect of seasonality by subtracting the mean concentration for a period from the concentrations sampled in that period, then adding back in the overall grand mean concentration.\(^{13}\) For example, if samples were taken on a monthly basis, then samples taken in the month of January would be reduced by the average concentration from all January observations. Then the average concentration over all observations would be added to that amount. The same procedure would be done to concentrations sampled each following month.\(^{13}\)

For owners or operators of CCR units that will not reach sufficient background sample sizes (25-30) and consistent periodicity of sampling events, it may not be possible to meet the data assumptions for running parametric trend analysis such as the MLE or the seasonal Kendall regression methods. However, we recommend testing for trend using the Mann-Kendall trend analysis.

Figure 6 below is an example of fitting a maximum likelihood estimation regression model under lognormal distributional assumption to samples collected from seven wells for a particular constituent over a period from 2011 to 2016. The green dotted line represents the fit of the model to both the detected and non-detected observations. The red line represents the fit of only the detected observations. The overall trend is declining and has been tested as being statistically significant. This may be a temporary situation as the number of wells included in the testing has increased over time. As
sampling is performed on the fixed set of wells going forward a steady state may appear.

Figure 6: MLE Regression Result

AIDS TO INTERPRETING LINEAR TRENDS

In situations where there is sufficient data to declare trends in the background samples, further analysis can be done to breakdown the trends in a piece-wise fashion. The aforementioned regression models assume that the relationship between water quality concentration levels and time are linear while in reality, changes in water chemistry over time are highly complex and variable, and are more likely to be non-linear in nature.

The piece-wise polynomial model has proven useful in circumstances when changes in trend may occur within the time series for a constituent. It provides another line of evidence that may be performed should geologic conditions or other factors indicate shifts in trends may have occurred. This approach attempts to find an appropriate mathematical model to express the relationship between the constituent’s values and the sampling dates by using piece-wise regressions. Two types of piece-wise models are used to study the trends: linear-linear model and linear-linear-linear model.

The linear-linear regression model assumes and identifies one structural break in a constituent’s data series, in which the two portions of the data separated by the break point follow two different trends as modeled by two different linear equations. Similarly,
the linear-linear-linear model attempts to identify two structural breaks to separate three different linear trends.

While this approach is informative, it is not able to account for non-detects in a sample. Hence, it’s recommended to implement the piece-wise polynomial models and regression models complementary to one another. The piece-wise models are applied mainly as a visual guide when confirming if all or a subset of the background samples can be used to represent background.

The same set of samples used to demonstrate the MLE regression method in Figure 6 is used to run a linear-linear-linear piece-wise regression model. Figure 7 shows that after a rise and then drop in concentrations after 2013 (most likely due to new wells being introduced to the sampling network), there is a statistically significant shift in the trend in the early part of 2016. Continued monitoring of the trend is recommended as more samples are collected to determine if the background samples from this particular site are in a steady-state. When using linear regression techniques on what are essentially time series data on water quality samples, practitioners need to keep in mind that the most recent trends in concentrations are the best predictors of the future concentration trends.

![Figure 7: Linear-linear-linear Piece-wise Regression Result](image-url)
Another consideration in the decision process regarding trends in background samples is that if the number of samples in the period of record is low (<10 samples) and over a short duration (< 2 years), then an observed trend in the time series may not necessarily indicate changes to the natural variability in the water quality of the background wells. In fact, the trend could be part of the natural variation; and if the time series were longer, a steady state could be observed for concentrations of constituents of interest.

SELECTING THE APPROPRIATE DISTRIBUTION FOR THE BACKGROUND SAMPLES

At this point, the background samples have been explored and analyzed to evaluate the data’s potential to represent background concentrations levels. Any values that may not be representative of background, such as certain outliers or a run of background samples trending apart from the main body of observations, have been removed, provided that investigations showed that the values in question arose as a result of collection or measurement errors or anthropogenic causes. The next step prior to estimating the UPL for detection monitoring is to determine the distribution that describes the patterns in the data. Since many tests make an explicit assumption concerning the distribution represented by the sample data, the form and type of distribution must be checked using a goodness-of-fit test. A goodness-of-fit test assesses how closely the observed sample data resembles a proposed distributional model, and whether the sample data closely resembles the tails of that distributional model. The models under consideration for groundwater quality samples are normal, lognormal, or gamma distributions. These distributions are part of the parametric family of distributions. For samples whose distributional patterns do not conform to any of these three distributions, non-parametric statistical tests are appropriate. The ProUCL software from the EPA provides a means to test for the goodness-of-fit for either the normal, lognormal, or gamma distributions from which to base a particular type of UPL.

The CCR Rule describes three types of distributions: normal, lognormal and non-parametric. Since the 1990s, significant research has been conducted to evaluate the benefits of using the gamma distribution to model water quality samples. Researchers have shown that the heavy dependence on the lognormal for highly skewed distributions to model groundwater samples or other environmental data can lead to biased characterizations of the samples of interest, especially for values in the upper ranges of the distribution. More specifically, if the standard deviation of the log-transformed data is greater than one and the sample size of the data set is less than 30 to 50, the upper limit statistics produced using log-normal distribution assumptions are unstable and impractically large. HDR has included testing for the gamma distribution to evaluate its effects on the upper limits and potential SSIs at the sites, and ultimately to allow for a more accurate characterization of background conditions.

The CCR Rule is not entirely prescriptive in nature and allows subject matter expertise at the site to tailor the groundwater monitoring requirements for the unit and site conditions. Since the rule allows flexibility to take advantage of future technology or
research improvements that can aid in the accurate site-specific testing for statistically significant increases over background concentration levels, tests for the gamma distribution as well as normal and lognormal distributions should be considered in order to fairly characterize background concentration samples. Through its varied assignments, the research findings from Singh and Singh, 2013 have been empirically corroborated in HDR’s production of UPLs for water quality monitoring programs. Even when the goodness-of-fit tests identify the lognormal distribution as fitting the sample data, when the standard deviation of the log-transformed data is higher than one, the upper prediction limits can be unrealistically high and hence, open to questioning and debate.

Table 2 has the 95 percent UPLs for one to four independent comparisons, which have been calculated using EPA’s freeware, ProUCL, for ten different water quality constituents from a confidential client. Goodness-of-fit tests have determined that the series follows lognormal distributional patterns. UPLs under gamma distributional properties produced using two different methods called Wilson-Hilferty (WH) and Hawkins-Wixley (HW), and under the normal distributional properties have also been produced. The five constituents with a standard deviation of greater than one are listed on the left, while the remaining five with standard deviations of one or less are listed on the right.
Table 2: 95 Percent UPLs from Background Constituent Samples of Varying Skewness for Testing One to Four Future or Independent Samples

<table>
<thead>
<tr>
<th>Highly Skewed Lognormal Distribution</th>
<th>Moderate to Low Skewed Lognormal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td>UPL(1)</td>
</tr>
<tr>
<td>Gamma HW</td>
<td>1,793</td>
</tr>
<tr>
<td>Gamma WH</td>
<td>1,812</td>
</tr>
<tr>
<td>Normal</td>
<td>2,187</td>
</tr>
<tr>
<td><strong>Beryllium, (ug/l)</strong></td>
<td>Lognormal</td>
</tr>
<tr>
<td>Gamma HW</td>
<td>0.552</td>
</tr>
<tr>
<td>Gamma WH</td>
<td>0.560</td>
</tr>
<tr>
<td>Normal</td>
<td>0.699</td>
</tr>
<tr>
<td><strong>Chromium, (ug/l)</strong></td>
<td>Lognormal</td>
</tr>
<tr>
<td>Gamma HW</td>
<td>11.820</td>
</tr>
<tr>
<td>Gamma WH</td>
<td>12.130</td>
</tr>
<tr>
<td>Normal</td>
<td>14.420</td>
</tr>
<tr>
<td><strong>Nitrates/Nitrites, (mg/l)</strong></td>
<td>Lognormal</td>
</tr>
<tr>
<td>Gamma HW</td>
<td>0.943</td>
</tr>
<tr>
<td>Gamma WH</td>
<td>0.995</td>
</tr>
<tr>
<td>Normal</td>
<td>0.919</td>
</tr>
</tbody>
</table>

A review of the UPL values from the highly skewed distributions shows that as the number of multiple comparisons increases, which is likely the case under CCR Rule given that a minimum of three downgradient wells are required, the lognormal UPLs deviate dramatically from the comparable limits with other distributional assumptions (over 50 percent larger when testing four independent samples). On the other hand, for moderately skewed distributions, the tails of the lognormal are comparable to those in the gamma and normal distributions showing the UPLs are of similar order of magnitude.

**THE UPPER PREDICTION LIMIT**

A primary goal of water quality sampling is to determine whether or not a given sample or collection of samples contain values that are considered statistically significant.
increases (i.e., outliers) based on the characteristics and distributional patterns from some baseline or background set of samples. If so, then the collection of samples can be considered to have values higher than in the background, and may imply impacts at the sample location from an anthropogenic source. Under the context of this paper, this source would be the CCR unit. In general, the determination of SSIs is done by computing the UPL from the observations in the background set of samples. The UPL represents the upper boundary of a prediction interval for an independently obtained observation(s) (or an independent future observation(s)). The upper limit is computed because the concern is generally for exceedances greater than one would expect from the background values. Note that pH (an Appendix III constituent) may require both upper and lower prediction limits.

The formulation of the prediction limit may vary slightly with the particulars of the test to be made and the characteristics of the data involved such as the type of distribution, the sample size, presence of non-detects, and level of skewness, but under normal distribution assumptions with no non-detects, the formula for the prediction limit for \( k \) future or independent observations is:

\[
UPL = \bar{x} + t_{1-\alpha/k,n-1} S \sqrt{1 + \frac{1}{n}}
\]

Where

- \( \bar{x} \) = background sample mean
- \( S \) = background standard deviation
- \( t \) = Student’s t with \( 1-\alpha \) degrees of freedom
- \( \alpha \) = Type I (false positive) error rate
- \( n \) = number of observations in the background dataset
- \( k \) = number of independent or future obtained samples to be predicted.\(^{10}\)

ProUCL is widely used to estimate the appropriate UPL given the sample size (see Table 1 above), type of parametric or non-parametric distribution that a constituent’s background samples follow, the level of censorship, and the level of skewness.

A different algorithm is available in ProUCL to compute UPLs for normal, lognormal, and gamma distributions for samples with or without non-detects. UPLs with non-detects are based on the data series with imputed values using either ROS or hybrid KM methods for imputation. If no discernable distribution is found, then the non-parametric UPL is produced. A variation on non-parametric UPLs is Chebyshev’s non-parametric UPL, which may be a good choice for non-parametric samples with less than 30 observations. It produces reasonable, conservative, and stable UPLs.

The CCR Rule requires multiple, independent comparisons from the set of downgradient well-constituent pairs be tested for statistically significant increases over background concentration levels on a semi-annual basis. The UPL offers the advantage of controlling the Type I error rates when testing an exact number of multiple independent or future observations. This relates to the assumption that the set of
observations at the site’s downgradient wells, after background levels have been established, follow distributions similar to those found in background. In addition, the mathematical underpinnings of the UPL under re-testing strategies are well established and easily implemented while those of other statistics such as UTLs and control charts are not.

For demonstration purposes, consider a site that has four downgradient wells. The concentration of a particular constituent is measured at each of these wells during a sampling event. A practitioner would have to test if at least one of the four downgradient samples exceeded the UPL to determine if a statistically significant increase over background levels has occurred. According to the CCR Rule, a statistically significant increase has occurred at the site if the concentration of at least one of the Appendix III constituents has exceeded its respective test statistic at any one of the monitoring wells. See Part 257.93 h(1)(2).

The practitioner has determined that a 5 percent test significant level should be used to determine the critical value of the UPL using the background data. The 95 percent UPL represents a concentration value with a low probability of 5 percent and a confidence coefficient of 0.95 of being exceeded when compared to concentration from an independent or future sample taken from the downgradient well. While there is still a 5 percent chance the value will exceed that UPL and be from the same distribution as background, we accept that chance, and if we observe a value higher than the UPL, we state that a statistically significant increase over background concentration levels has occurred.

This 5 percent is the level of the Type I error we are willing to make. However, if multiple independent or future samples are being compared at that sampling event to determine site-wide status, then the chance that at least one of those samples will exceed the UPL and still be from the same background will increase. The Type I error rate is then higher than 5 percent. Sometimes depending on the background sample size and the number of downgradient comparisons to be made, this Type I error rate can be as high as 20 percent or more. In this situation, adjustments are made to the UPL formula to mitigate the increase in the Type I error as more comparisons are made.

Figure 8 provides a demonstration of the use of the 95 percent UPL for multiple comparisons of a constituent measured in units of mg/l. The background sample consists of 68 observations of which nine are non-detects. The median concentration is 0.017 mg/l. Non-detects were replaced using the regression order statistics method under lognormal distributional properties. The 95 percent UPL for one independent comparison (0.084) may yield more false positives than expected if conducting multiple tests. The UPLs for two, three, and four comparison tests are 0.116, 0.138 and 0.155 mg/l, respectively.

Four downgradient samples with measurements of 0.05, 0.06, 0.021, and 0.10 mg/l are collected at the site. The correct test statistic to use is the 95 percent UPL for four future or independent observations (UPL(4)). Since all four of these samples had concentrations less than 0.155 mg/l, no statistically significant increase has occurred. However, if the UPL for one future or independent sample (UPL(1)=0.084s) was used to conduct the test, then a statistically significant increase would have been incorrectly
recorded and the decision that impacts at the downgradient site was evident would be in error.

**Figure 8: Demonstration of Upper Prediction Limits by Number of Site-wide Comparisons**

**SUMMARY**

The statistical methods reviewed in this paper are all accessible across the commercially available statistical software packages and through freeware such as R (https://www.r-project.org/), ProUCL from the U.S. EPA or the Seasonal Kendall regression tests from the US Geological Survey. HDR’s robust methodology has allowed its practitioners to efficiently move through the statistical decision-making process to test if independent samples are statistically different from the background levels for its clients. The framework allows practitioners to document the process in a transparent and defensible manner.

Figure 9 below succinctly summarizes HDR’s approach as described in this paper. If after running the PDA, certain value(s) are deemed not part of background, the background sample is revised and the PDA repeated. Depending on the quality of the background sample in terms of quantity and the incidence of censorship, tests for parametric or non-parametric distributions are conducted before determining which form of the UPL to use. When a set of samples has 50 percent or more of its values as non-detects, then non-parametric methods are chosen to produce the UPLs. If a sample has fewer than the minimum eight observations as specified in the CCR Rule, then the
largest detected value is used as the placeholder test statistic until more samples can be collected and background updated. If all observations are non-detects, the double quantification rules can be used to flag for a statistically significant increase in independent samples.

**Figure 9 Decision Logic Framework for Determining Appropriate UPLs from Background Samples**

While this demonstration of HDR’s approach has worked for its past engagements, it is not the only approach. Other engagements may require a different sequence of statistical diagnostic tests that are more appropriate for their site-specific conditions. This paper demonstrates that one can make sense of the multitude of guidance reports and books, such as the Unified Guidance, 2009, Gibbons’ Statistical Methods for Groundwater Monitoring, 1994 and Helsel’s article More Than Obvious: Better Methods for Interpreting Nondetect Data, 2005, to develop a framework suitable for data collected at CCR unit sites while satisfying the requirements in the CCR Rule. As well, the paper has highlighted advancements in imputing values for samples with non-detects using methods such as regression order statistics and Kaplan-Meier, the benefits of using the gamma distribution to model certain samples of background values, and the versatility of the UPL for testing if one or multiple downgradient samples are from different populations than found in background.
REFERENCES


