Censored Geochemical Data Analyses for Non-Scientists*

Richard B. Shepard, PhD†

Contents

†Applied Ecosystem Services, Inc.
Troutdale, OR
503-667-4517
rshepard@appl-ecosys.com

1 Introduction

Many projects or operation involve geochemistry: chemicals in water, sediments, soils, or rocks. Most people are not concerned with chemicals like magnesium sulfate or sodium chloride, but they are seriously concerned with toxins that effect human and environmental health. These toxins can be inorganic metals such as arsenic, cadmium, and zinc or organic compounds such as polychlorinated dioxins, furans, and biphenyls, and pesticides.

The consequences of incorrectly analyzed and interpreted geochemical data can range from fines through remedial actions to cease-work orders. It is important to appreciate the potential for environmental and economic harm, to recognize when geochemical data are incorrectly analyzed, and understand how they should be analyzed.

The three points of this document are:

1. What censored chemical concentration values are and how they commonly are mistreated.

2. How to recognize when censored values are correctly incorporated into environmental data analyses, and to reject wrong data analyses.

3. Why familiarity with these issues are important, especially when the issues involve environmental policies or regulatory actions.

This paper will not teach you statistics or ecology, only to recognize how nondetected chemical values should be analyzed in permit applications, compliance reports.

1.1 Censored data defined

Censored data are values that cannot be quantified to the standard of 99% confidence. Most people know of censored values from medical survival and mechanical failure reports. These are right-censored values because they are larger or longer than the last quantified value. Consider testing of a new light bulb filament to determine whether it lasts longer than the currently used filament. The company set up 15 bulbs of each type and turned on the electricity, recording the length of time each remains lit. After 48 hours someone decided that sample size was too small so an addition 20 bulbs of each type were added. Six weeks (1008 hours) after the first bulbs were turned on the experiment was ended. By then, many of each type of bulb had burned out and each duration recorded. Some from the second group still burned and their duration recorded as "greater than 980 hours," because they were still burning after 1008-48 hours of use. A few bulbs from the original group were still burning and their durations recorded as "greater than 1008 hours."[1] The question to be answered is whether there is a difference in mean or median survival time of each filament type. How this question would be answered with the two different censored values is beyond this document's coverage.

Environmental chemical concentrations cannot have negative values and for each element, instrument, analyst, sample size, and other factors there is a concentration value below which the "signal" of the concentration value is lost in the electronic "noise" of the measuring instrument. These results are called "less-thans," "nondetects," or "left-censored" data. There are three limits related to left-censored[2] environmental chemical data found in regulations and reports: method detection limit, quantitation limit, and reporting limit (RL).

Method detection limit The smallest concentration of a chemical element that the instrument and analytical chemist can qualitatively identify.

Quantitation limit The smallest concentration of a chemical element that the chemist can quantify with 99% confidence.

Reporting limit Either the method detection or quantitation limit. The analytical laboratory's reports show it as a number preceded with the less-than (<) sign; e.g., <0.5 mg/L.

[1] As an aside, the 4 watt Livermore Centennial light bulb at Fire Station #6 in Livermore, CA, has been burning continuously for 115 years.

[2] Commonly referred to as just "censored" because it is only the left side of the distribution that is affected.

2 How Censored Data Are Incorrectly Treated

There are two categories of wrong approaches to environmental censored data: ignore it and substitute an arbitrary value.

2.1 Dropping censored data

Ignoring censored chemical data means to drop all these values from all statistical analyses, summary descriptions, comparison of groups of data, forecasting future states. Doing this loses valuable information and will result in bad decisions.

Consider a data set of 100 samples of arsenic (As) collected at a single location on a regular schedule. There will certainly be some censored values in the set; perhaps as few as 5% or as many as 70–80%. In CERCLA guidelines the EPA told data analysts to drop censored values if they constituted 15% or less of the data set. Fortunately, demands like this are slowly changing and becoming more realistic. The loss of 15% of data will definitely raise the mean and median values and reduce the variability.

When the percentage of censored values is high (up to 70–80% for metals is not unusual) dropping them from consideration is a gross error. The mean, median, and standard deviation of the data are higher than reality. When the chemical cannot be detected (it may be totally absent or cannot be quantified) the knowledge is important to operators, regulators, and others.

2.2 Substitution of arbitrary values

There are three substitutions in EPA, state, and Tribal regulations for analyzing environmental data: zeros, reporting limit, or some value between those two. All are completely arbitrary, have no basis in science or statistics, and distort analytical results by adding information not present in the original data (unlike dropping censored values which subtracts valuable information). These, too, are gradually being replaced with correct methods.

The reasons substitution are wrong are:

1. Censored values are unknown. Therefore, any substituted value is arbitrary and wrong by definition.

2. There is no scientific, statistical, or mathematical basis for assuming that all these unknown values should be assigned to the same number.

2.2.1 Set all censored values to zero

When all censored values are set equal to 0 the mean and median are artificially lowered because the sum of the detected values are divided by the total number of observations, not just the detected ones. By adding zeros to the analysis you are adding information to

the analyses that is not present in the original data set; for example, that the number of censored values reflect the absence of the chemical. This is an unwarranted assumption because the actual values are unknown.

2.2.2 Set all censored values to the RL

The opposite extreme to setting all censored values to 0 is setting them equal to the reporting limit. The information this adds to the original data is a set of concentration values right at the lowest value of 99% confidence in measuring the amount of chemical in the sample. This is not the case since the actual concentrations are unknown and not likely to be the same value. Adding all the RL values when calculating the mean and medium of the data artificially increases them, and variability is decreased.

2.2.3 Set all censored values to a value between 0 and the RL

Historically, this has been the favored method of the EPA, states, and Tribes. The most common wrong value used is half the distance between 0 and the RL, but for air quality the EPA directed the use of the $\sqrt{2}$, approximately 0.7, instead. (Douglas Adams would suggest that 0.42 is the correct proportion.) Any substituted value is arbitrary and capricious and totally wrong.

3 Correct Methods for Censored Data

3.1 Introduction

The first thing an environmental data analyst does with a new data set is describe and summarize the data. This is the basis of how to analyze the data for forecasting, comparing to other data sets, correlating, and determining cause and effect. Plots are more valuable than tables of data (or data summaries) because they allow everyone—including non-technical audiences—to see the patterns and important parameters of the data.

The most effective illustration of the data is the box-and-whisker plot (boxplot; Figure 1). The boxplot displays the important attributes of the data set that describe and characterize it: the minimum and maximum values (the ends of the whiskers), the range and distribution of the data, skewness (the position of the median in the interquartile range (IQR) box, and any outliers at the extremes of the distribution (as open circles beyond the ends of the whiskers). The IQR includes the most common 50% of the data and is approximately equivalent to ±1 standard deviation (σ) of a normal distribution. The ends of the whiskers are approximately 3 $\sigma$from the median, or less than 1% of all values.

An example uses a subset of 24 measurements of arsenic (As) in Manoa Stream on Oahu. The first step is examining the distribution of the data using a boxplot (Figure 2). It is easily seen that the majority of values (13, or 54%) are well below the analytical chemical laboratory's reporting limit (RL) of 2.0 mg/L. The median value is 0.8 mg/L, the

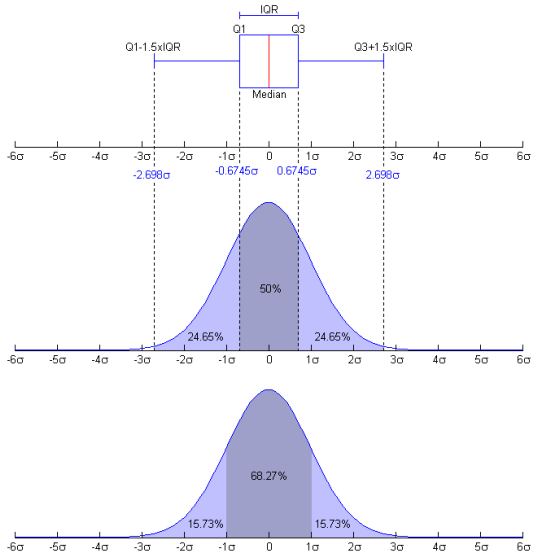

Figure 1: A boxplot and its relationship to the familiar normal curve.

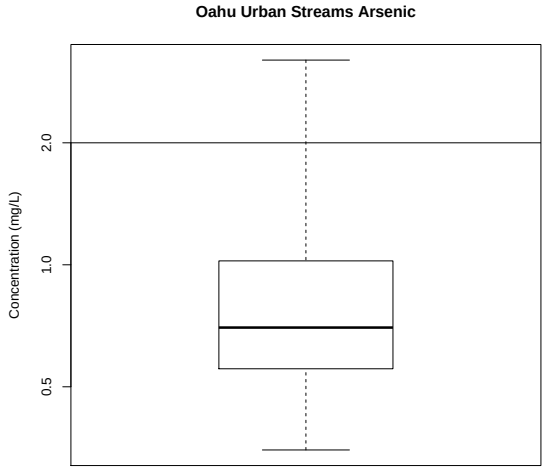

Figure 2: Distribution of arsenic (As) concentrations in Manoa Stream on the Hawaiian island of Oahu. Most of the values are censored below the reporting limit (RL) of 2.0 mg/L.

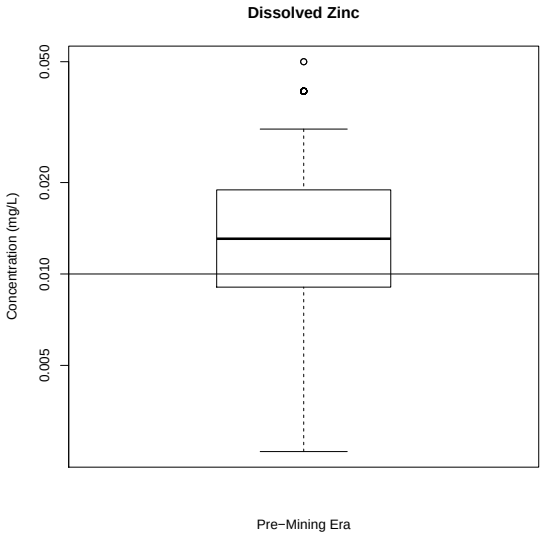

Figure 3: Baseline concentrations of zine in surface and ground water samples.

mean value is 0.9 mg/L, and the standard deviation is 0.66 mg/L. While this distribution of concentration values is not unusual for arsenic, not all metals have concentration distributions as extreme. Figure 3 shows the baseline distribution of several hundred zinc concentrations in surface and ground waters.

Another example illustrates how variable metal concentrations can be in different areas. Figure 4 demonstrates the information value of the boxplot. It is apparent that fish liver cadmium concentrations in the Colorado Plateau physiographic region are extremely low (only about 30% of observations detected) and the median value about 0.4 mg/L while in the Southern Rocky Mountain physiographic region 75% of all observations can be detected, the median value is approximately 4 mg/L and some concentrations approach 100 mg/L.

3.2 Descriptive statistics

Descriptive statistics summarize the data being analyzed by calculating the values displayed on the boxplot, including the median or mean, the range, and the variability (the IQR, deviance, or standard deviation). With very small data sets (generally fewer than 25 individual values) this summary is the percentage of censored data and further analyses use nonparametric methods. Contrary to common perceptions, nonparametric statistical models are scientifically valid and statistically robust.

There are three types of geochemical data and all can be used to produce correct data-analytical results.

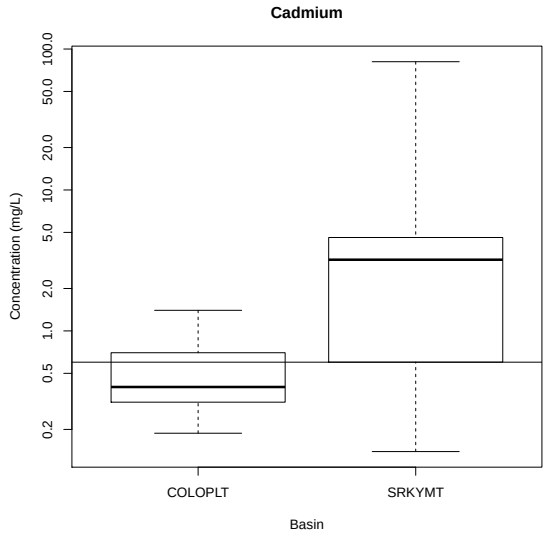

Figure 4: Cadmium concentrations in fish livers in two different physiographic regions in Colorado: Colorado Plateau and Southern Rocky Mountains.

3.2.1 Binary methods

Binary methods have only two values: uncensored and censored. The analysis uses the relative proportions of each to produce useful information. Here is a small data set of 11 values:

$$<1 \quad <1 \quad 3 \quad <5 \quad 7 \quad 8 \quad 8 \quad 8 \quad 12 \quad 15 \quad 22$$

To apply the binary method the data are recoded as either less than (LT) of the highest reporting limit or greater than or equal to (GE) for uncensored data. For the small data set there are with its highest reporting limit (RL) of 5 there are four Lt's and 7 GE's:

<1	<1	3	<5	7	8	8	8	12	15	22
LT	LT	LT	Lt	GE	GE	GE	GE	GE	GE	GE

The binary method summary is stated this way: 7 of the 11 data values (64%) are equal to or exceed the concentration of 5; i.e., were detected. There is no mean, median, or mode computed.

3.2.2 Ordinal methods

Ordinal methods use the order (or ranking) of the reported values rather than the values themselves. Tied values each have the average of their values. These methods can also

be used on non-numeric data as long as the values can be ranked (e.g., vary small, small, moderate, large, very large). Censored values are ranked smaller than uncensored values above the threshold. For the small data set above the ranking is:

Data:	< 1	< 1	3	< 5	7	8	8	8	12	15	22
Ranks:	2.5	2.5	2.5	2.5	5	7	7	7	9	10	11

The ranks of tied values are themselves tied as the median of the ranks they would have had had they not been tied. This preserves the sum of the ranks, a statistic used in many nonparametric tests. The four values equal to or less than 5 mg/L sum to 10 so the rank for each is the median of 4, or 2.5. Similarly, the three values of 8 mg/L are each assigned the median rank (7) in the sequence. Applying a nonparametric statistical model for descriptive summaries (see Section 3.2.5 below), comparison of two data sets, correlations, and regressions will yield accurate results without the errors that would result if arbitrary values had been substituted for the censored values. Using percentiles of the ranks shows that the median value is 8 mg/L and the 75^{th} percentile is 12 mg/L, equivalent to 1 standard deviation above the mean in a normal distribution.

The following analytical methods all use the actual values in the data rather than a characteristic of each value.

3.2.3 Maximum likelihood estimation model

Maximum likelihood estimation (MLE) assumes that there is a probability distribution that best fits the uncensored values. This makes MLE a parametric statistical model. The quantified values and the proportion of censored values are fit to different distributions and the means and standard deviations are optimized by a likelihood function. The distribution that has the maximum likelihood of describing the data set is used to report the parameters of the data: mean, median, standard deviation, and so on. For geochemical data the lognormal distribution is almost always the best fit.[3] Less frequently, transforming the raw data by taking the square root of each will fit a normal distribution. MLE methods produce poor estimates with small samples (< 50 observations). An example of the MLE method is in Section 3.1 and Figure 2.

3.2.4 Kaplan-Meier survival model

Kaplan-Meier (K-M) survival analyses are nonparametric because they do not fit the data to a probability distribution to obtain mean or variance values. Survivorship analyses have been used for decades in medical and industrial studies. Drug and other medical treatments are conducted for a fixed length of time. Cure or other indicator of efficacy for each patient is recorded. Those waiting for the desired effect when the study ends is recorded as censored; that is, the cure occurrence time is unknown. Industrial applications

[3]The raw concentration values are not normally distributed, but the logarithms of these values are.

Table 1: Summary statistics of the Oahu urban stream arsenic concentrations based on Kaplan-Meier (K-M) survival analysis.

Mean	Standard Deviation	Standard Error	First Quartile	Median	Third Quartile
0.949	0.807	0.165	0.500	0.700	0.900

determine the time to failure (see Section 1.1 for an example) usually reported as the mean time between failures (MTBF).

In both of these applications of survival analysis the unknown quantities are at the right side of the survival curve; that is, they are right-censored data. Geochemical data are the mirror image as the unknown values are on the left side of the distribution curve; this means they are left-censored data. Statistical software, both proprietary and free, use models for only right-censored survival analysis. To apply the K-M model to environmental geochemistry requires transforming the raw data from left- to right-censored by "flipping" the data set. This is done by selecting a concentration value greater than the largest quantified concentration, then subtracting each raw value from it. The flipped data now fulfill the requirements of the statistical software's survival methods and results in a plot of the probability of observing each concentration in the range of the data set, along with 95% confidence intervals. The K-M survival model of the Manoa Stream arsenic concentration values are shown in Figure 5 with the summary statistics for these data in Table 1.

3.2.5 Regression on order statistics

Regression on order statistics (ROS) has a parametric method that calculates summary statistics of a data set using a regression model on a probability distribution plot. When the data set has more than 50 observations the fully parametric ROS is less efficient than MLE methods and offers little advantage. There is a robust ROS implementation that uses sample data as much as possible and imputes (calculates) values to fill the censored portion of the distribution. The robust form of ROS has advantages over the more restrictive parametric assumptions of MLE and is better used with small samples—fewer than 50 values—where MLE models are inaccurate. Understanding the implementation details of applying ROS are not necessary to recognize when its use is appropriate for the available data. Analyzing the Manoa Stream arsenic data using ROS produces the regression plot shown in Figure 6 with summary statistics in Table 2.

3.2.6 Comparing the three methods

While each of the three methods above produce results those are not exactly the same numbers. A comparison of the median, mean, and standard deviation calculated by each method is shown in Table 3. In practical terms, the differences are meaningless because

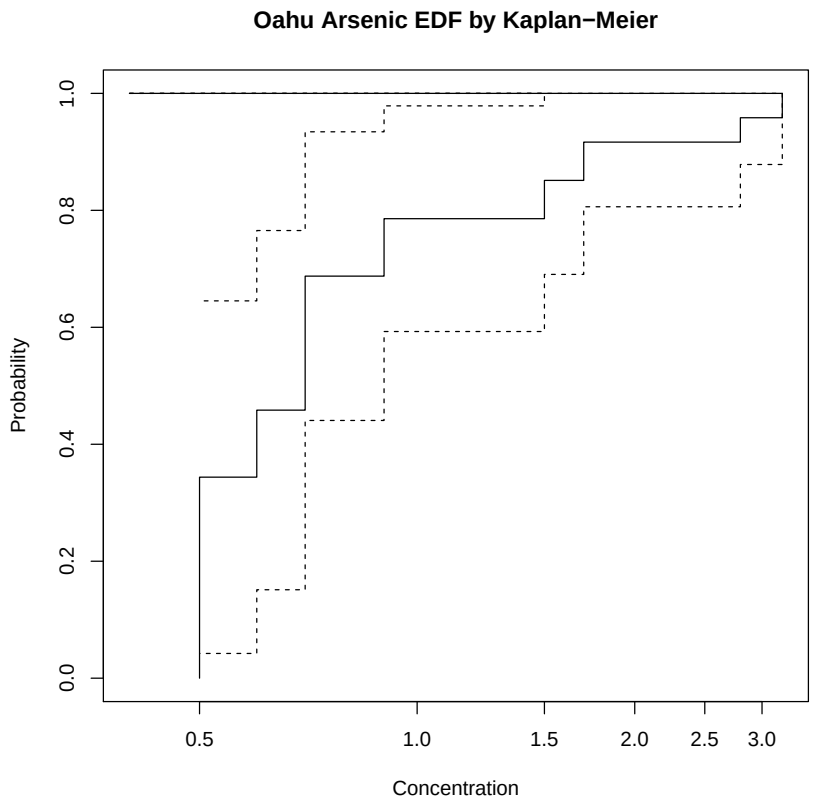

Figure 5: The results of analyzing arsenic concentrations in Manoa Stream on the Hawaiian island of Oahu using Kaplan-Meier survival analysis after flipping the raw data.

Table 2: Summary statistics of the Oahu urban stream arsenic values analyzed by regression on order statistics (ROS).

Sample size	Number Censored	Mean	Median	Standard Deviation
24	13	0.700	0.972	0.718

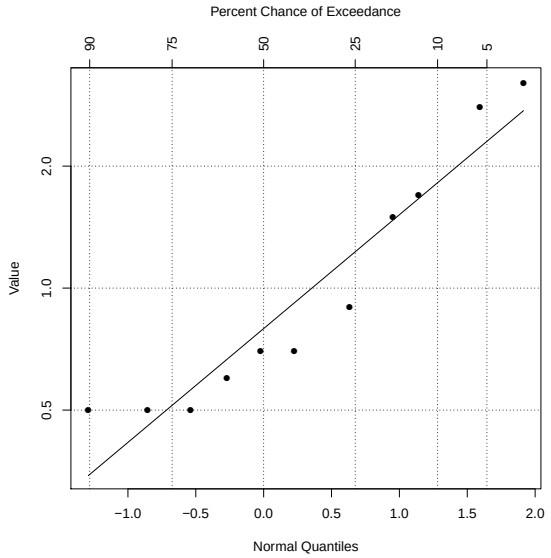

Figure 6: Oahu urban stream arsenic concentrations plotted with a linear regression of the response values relative to normal quantiles as the explanatory variable. While the regression line extends to the lower left corner no censored values less than 5 mg/L are below that concentration.

Table 3: The median, mean, and standard deviation of the Oahu urban stream arsenic concentration values as calculated by Kaplan-Meier survival analysis (K-M), maximum likelihood estimation (MLE), and regression on order statistics (ROS).

	Median	Mean	Standard Deviation
K-M	0.700	0.949	0.807
MLE	0.777	0.945	0.656
ROS	0.700	0.970	0.718

natural ecosystems, particularly aquatic ones, are highly variable. Any one of the three statistical models will results in technically sound and legally defensible results.

There are, however, four data characteristics that strongly influence the results produced by each of these methods:

Sample size MLE methods work much better with larger samples, generally >50 reported observation values.

Transformation bias When data are transformed from a normal distribution to a lognormal distribution, the mean and standard deviation are not the same because the scales are different. Untransformed data are on a scale using single integers while logarithmic scales use powers of 10. Not accounting for this difference will produce incorrect results. Robust ROS and robust MLE methods attempt to overcome this bias.

Robustness It is essential to start any geochemical data analysis by determining the best fitting probability distribution for each specific data set. Assuming that the data are normal, or lognormal, without testing to see if that assumption is correct can lead to large errors and any policy or regulatory decisions based on them will be less effective than they could be.

Details of method computation and terminology Both the fully parametric and robust ROS methods have the same name (ROS); this can be confusing. For MLE older reports might have used a table lookup method while all modern reports should use a computer solution of a likelihood function. Some reports incorrectly named substitution as "imputation". The term imputation applies to a statistical method to fill in missing data; substitution inserts arbitrary values for missing data.

3.3 Summing data with censored observations

Environmental chemical data are often summed, most frequently by adding monthly quantities to estimate the total amount of a pollutant or contaminant discharged over the course of a year. These estimates add quantities which each have the same weight of importance so the addition is straight forward. It is a more complicated summation when each value is weighted unequally and these weighted values summed to a total quantity. This is the case with ecological risk assessments (ERA) at Superfund sites under CERCLA in accordance with EPA guidelines.[4]

The purpose of an ecological risk assessment is to calculate an overall numeric measure of the effects of chemicals such as dioxins, furans, and PCBs on animals. Each of these pollutants is a class of similar chemicals rather than a single chemical, and each specific chemical (called a congener) has a different toxicity to different organisms. Summarizing the general toxicity of all congeners in the class on an organism assumes that

[4]See, for example, Guidelines for Ecological Risk Assessment. 1998. EPA/630R95/002F.

the individual congener toxicities are additive and allows calculation of a toxicity equivalent concentration (TEC). One application of TECs is setting recommendations for fish consumption by people which means that how TECs are calculated can have significant ecological and economic effects.

Because each dioxin, furan, and biphenyl congener has a different toxicity to organisms each congener's toxicity is related to what is accepted as the most toxic congener: 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD). This is done using a toxicity equivalent weighting factor (TEF) determined by a group of experts for each class of organism.[5] TCDD has a TEF of 1 while less toxic congeners have TEFs closer to 0. The concentration of each congener in the medium (sediments or soils) is multiplied by its TEF to obtain the TEC for that congener. The individual TECs are added to calculate the overall TEC.

It is not unusual for a congener's concentration to be below the analytical chemical laboratory's reporting limit, so how these censored concentration values are included makes a big difference in the value of the overall TEC. The only recommendation the EPA offers on the subject is to substitute both 0 and the reporting limit to calculate the range of possible concentration values. Substitution of any arbitrary value for censored ones has been shown to be wrong. And substitution is unnecessary because there is a simple process to avoid it with censored values in the data set.

The sum and the mean are related: the mean is the sum divided by the number of values summed. To calculate an overall TEC, multiply the mean concentration of congeners (using MLE, ROS, or K-M when censored values are included) by the number of congeners to obtain the sum. This avoids substitution of arbitrary values, reporting the extremes of a range in which the actual TEC would be found, and is supported by robust mathematics.

4 Comparing Two Groups

Comparing two groups is a frequent requirement under all environmental regulations. Are there differences between baseline and operational periods? Are there differences between sites in the same or different areas? Has a remedial action changed the concentrations of chemicals of concern? In some cases, testing is done in only one direction because one group (the "control") is expected to be either better (effective remedial action) or worse (operational period) than the "treatment" group; these comparisons use a one-sided test. In other cases there is no *a priori* expection of a difference or its direction so the comparison uses a two-sided test.

With no censored values in either set comparisons use the two-sample parametric *t*-test or the nonparametric Mann-Whitney test. When either (or both) data sets contain censored data options include binary methods, ordinal nonparametric methods, and both parametric and nonparametric survival analysis methods as described in Section 3.2.

[5] Van den Berg, M. et al. 1998.Toxic Equivalency Factors (TEFs) for PCBs, PCDDs, PCDFs for Humans and Wildlife. Environmental Health Perspectives **102**(12):775–792.

Table 4: The contingency table with the number of detected and nondetected values in each of the two data sets.

	Detects	Nondetects
Set 1	7	4
Set 2	3	9

4.1 Nonparametric methods

It is not unusual to have only a few geochemical concentration values when information must be provided to a regulator or other stakeholder. Despite the paucity of available data decisions must be made based on this limited amount of information.

In Section 3.2.1 a set of 11 values was separated into two categories: those less than the highest reporting limit and those equal to or greater than that reporting limit. Adding a second small data set allows comparing the two using binary and ordinal methods. The first data set has the highest reporting limit of 5.

< 1	< 1	3	< 5	7	8	8	8	12	15	22
LT	LT	LT	LT	GE	GE	GE	GE	GE	GE	GE

The second data set has 12 observations and the same highest reporting limit of 5. The question is whether these two data sets are significantly different.

<1	<1	2	3	3	<5	<5	<5	<5	7	8	10
LT	LT	LT	LT	LT	LT	LT	LT	LT	GE	GE	GE

4.1.1 Binary methods

The percentages of uncensored observations (quantified values above the highest reporting limit) in two or more groups can be tested to determine whether they are the same or significantly different using a test of proportions or contingency table test. The second data set has 3 out of 12 uncensored observations. Is 25% significantly different from 64% with these two sets of data?

Table 4 contains the number of detected and nondetected observations in each data set that is the basis for the statistical test of similarity. The test of proportionality results in $p = 0.85$, much greater than the standard of $p = 0.05$. Therefore, the null hypothesis that the sets are the same cannot be rejected. It seems counter-intuitive that two sets of data that appear so different are not significantly statistically different, which is why correctly incorporating censored geochemical data can make a major difference in environmental policies and regulations. This same approach can be used to test proportions in three or more groups.

4.1.2 Ordinal methods

The Mann-Whitney test is the nonparametric equivalent of the parametric *t*-test for two samples. This test is directly applied to censored data when all censored values are ranked lower than the uncensored values and considered to be tied in rank. The first step is to jointly rank the two data sets. Using the above two sets the 23 values have these ranks:

Set 1 data:	< 1	< 1	3	<5	7	8	8	8	12	15	22
Ranks:	7	7	7	7	14.5	17.5	17.5	17.5	21	22	23

Set 2 data:	<1	<1	2	3	3	<5	<5	<5	<5	7	8	10
Ranks:	7	7	7	7	7	7	7	7	7	14.5	17.5	20

All 13 observations below 5 are ranked as tied and assigned the same rank, 7, the median of 1–13. The uncensored values are also ranked individually (except for ties). The Mann-Whitney test sums the ranks in one data set. If the sum is unexpectedly low or high the null hypothesis that the groups are the same is rejected; if the sum of that data set is moderate (a function of the size of the set), the null hypothesis is not rejected.[6] With the two data sets above, the Mann-Whitney test produces a probability (p value) of 0.052 after correcting for ties so the null hypothesis that the groups are the same cannot be rejected. This approach is technically sound and legally defensible, while substituting arbitrary values for the censored data then applying a parametric *t* test will produce results dependent upon the substitution value chosen.

4.1.3 Maximum likelihood estimation

The use of uncensored values and the proportion of censored observations to calculate summary statistics was described in Section 3.2.3. The parametric MLE can be used to determine whether two data sets are the same.

In the late 1980s 113 ground water samples were collected from two areas in the San Joaquin Valley in California, 69 from the Alluvial Fan and 49 from the Basin Trough.[7] Both chemicals had censored observations, 31 for copper (27%) and 20 for zinc (17%); see Figure 7. One question was whether there was a statistically significant difference in zinc concentrations between the two areas. A MLE method tested the null hypothesis that the mean values of concentrations in both areas were the same. The frequentist paradigm for hypothesis testing can only reject the null hypothesis with a probability $p < 0.05$. The analysis calculated $p = 0.94$; therefore, the mean values of the two areas are not significantly different.

[6] Failing to reject the null hypothesis that the two samples are the same does not mean accepting that they are the same. While the difference might seem subtle or trivial it is meaningful and one of the drawbacks of this statistical paradigm.

[7] Millard, S.P. and S.J. Deverel. 1988. Nonparametric statistical methods for comparing two sites based on multiple nondetect limits. Water Resources Research **24**(12):2087--2097.

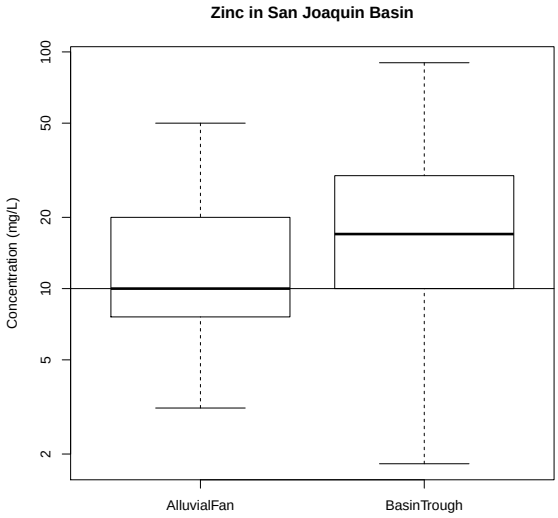

Figure 7: The distribution of zinc (Zn) concentrations in the Alluvial Fan and Basin Trough areas of the San Joaquin Valley in the late 1980s.

4.1.4 Nonparametric methods

There are several analytic methods that do not require fitting a probability distribution to the data sets. Among these are ordinal methods and several varieties of survival analyses.

4.2 Censored values contain valuable information

Some scientists have suggested that censored observations carry no information; therefore, they contribute nothing, or little, to analyses of the data. This is an incorrect assumption. From a strictly nontechnical perspective, the higher the percentage of censored values in the data set the less likely adverse impacts will occur. Using data, such as the zinc concentrations from the Alluvial Fan zone of the San Joaquin Valley ground water, a series of analyses with different percentages of uncensored observations found that increasing the percentage of censored values (such as by changing 10 mg/L to < 10 mg/L) strengthened the information signal in the data and produced a lower p value. This result rejected the null hypothesis that the two zones did not have different mean concentrations even though the Alluvial Fan data now had a higher percentage of uncensored values than did the Basin Trough area.

Several regulatory guidance documents, including some from the EPA, recommend that statistical tests not be run on data sets with a high percentage of censored values. There is little justification in the statistical literature for such recommendations as the mathematicians continue to improve the models. Statistical models that efficiently extract information from censored data, such as the Kruskal-Wallis test (the multigroup equivalent

of the Mann-Whitney test), use the information contained in the censored values in the data. There is no reason to limit use of these tools to inform environmental policy and regulatory decisions.

5 Comparing Three or More Groups

There are many situations when geochemical data from three or more locations (or times) need to be compared. Regulators or others might ask whether the means, medians, or the probability of observing a detected concentration of a chemical of interest are the same for all groups. Or, if there are differences, which groups differ. It is common for the chemicals of interest be below several laboratory reporting limits which makes how they are analyzed very important.

When all data are uncensored comparison among the groups is done with the parametric analysis of variance (ANOVA) or the equivalent nonparametric Kruskal-Wallis test. In the usual situation where censored values are found in all data sets the same methods described in the previous section can be extended to include more groups. Differences among means can be identified using parametric maximum likelihood methods and nonparametric Mann-Whitney tests determine whether the distributions of values in the groups differ.

In these situations not correctly incorporating censored data in the analyses could result in ineffective policy decisions, in regulatory fines and unnecessary and expensive corrective actions, or environmental litigation.

An example uses concentrations of trichloroethylene (TCE) in shallow ground waters of Long Island, NY.[8] Water samples were obtained from wells in areas of three residential housing densities: low, medium, and high. TCE sources were expected in each residential area because the chemical had been used in the past as a septic system cleaner and as a solvent in residential and light-industrial uses. The researchers wanted to learn if the concentrations of TCE were similar in all three residential density areas. Water samples were sent to different laboratories for analysis, and analytical precision improved over time, so four different reporting limits are in the three data sets: 1, 2, 4, and 5 µg/L. Figure 8 shows the observations in the residential areas using the commonly used histogram. It is important to know that 78.5% of the observations were censored; their distribution by density category is shown more clearly in Figure 9. Note that all observations in the low density areas are censored and concentrations above the 5 µg/L reporting limit are small proportions in the medium and high density areas (Table 5). This suggests that there might be differences among the three densities.

[8] Eckhardt, D.A., W.J. Flippse and E.T. Oaksford. 1989. Relation Between Land Use and Ground-Water Quality in the Upper Glacial Aquifer in Nassau and Suffolk Counties, Long Island, New York. US Geological Survey, Water Resources Investigation Report 86-4142.

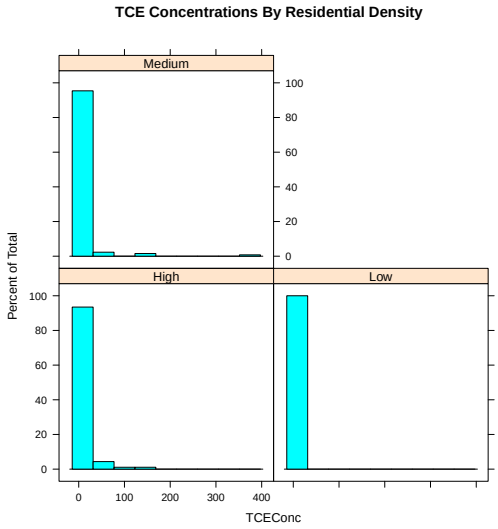

Figure 8: Histograms of TCE concentrations in shallow ground waters surrounded by different densities of residential housing in Long Island, NY.

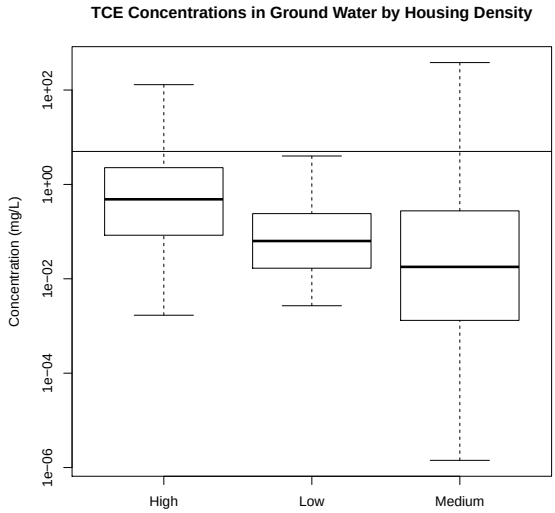

Figure 9: Boxplots of TCE concentrations by residential housing density in Long Island, NY.

Table 5: Percentage of detected TCE concentrations in Long Island ground waters.

Land Use Density	Low	Medium	High
Percent Detected	0	9	20

5.1 Substituting for censored observations

If the data analyst ignores the high percentage of censored values by substituting the reporting limit of 5 µg/L an analysis of variance (ANOVA) test for differences in the means yield $p = 0.547$. This value means that about 55% of the time these results are expected and the null hypothesis of equal means cannot be rejected. This bias of artificially high means for all groups introduces a pattern that does not exist in the data and distorts the results.

5.2 Using all data to compare groups

The nonparametric equivalent of the ANOVA test is the Kruskal-Wallis test. Because the question asked of the data is whether the TCE concentrations are similar in the three residential housing densities, the data will be recoded so all censored values have the same rank, for example -1, for an overall test of differences. When the Kruskal-Wallis rank sum test is run on the three data sets without substituting an arbitrary value for the censored ones the resulting $p = 0.01$, corrected for ties. Because 0.01 is much less than 0.05 the null hypothesis that all three means are the same is rejected.

Going one step further, determining which one (or more) area is different from the others, is accomplished using a maximum likelihood estimation (MLE) to compute a censored regression. This agrees with the Kruskal-Wallis test as the resulting probabilities are 0.007 and 0.003 for the low and medium density sites; therefore, all three densities are significantly different from each other (as shown in Figure 9). This is certainly valuable information for informing policy and regulatory decisions that would have been missed if the very high percentage of censored values been lumped as all having the value of the reporting limit (5 µg/L).

6 Correlations

6.1 Introduction

Correlations are associations between variables, not cause and effect. The correlation strength is measured by a correlation coefficient in the range of -1 (one variable increases when the other decreases) to +1 (both variables change in the same direction). When there is no relationship between the two variables the correlation coefficient is 0. Correlations are (or should be) commonly applied to environmental laws and regulations, particularly when seeking trends in chemical and biological observations over time. It may be necessary to remove seasonal effects and otherwise process the data prior to determining whether there is any consistent increase or decrease over time of the variable plotted on the Y axis.

Table 6: The four possible combinations of "high" and "low" for pairs of variables X and Y.

X	Y
High	High
High	Low
Low	High
Low	Low

6.2 Correlation coefficients

There are three correlation coefficients used in ecology and environmental science: Pearson's *r*, Spearman's rho (ρ), and Kendall's tau (τ).

Pearson's *r* is the traditional parametric correlation coefficient commonly seen in environmental technical reports by consultants and agency staff. The three quantities needed to calculate this correlation coefficient are difficult to obtain from environmental data so it should not used in this context.

Spearman's rho (ρ) is a nonparametric correlation coefficient calculated using ranks rather than the actual values. When there is a single reporting limit and ranks can be unequivocally computed this correlation coefficient is acceptable. However, multiple reporting limits are very common in geochemical data and ranking values such as <1, a detected 3, and <5 is subjective so Spearman's rho should not be used.

Kendall's tau (τ) is also a nonparametric correlation coefficient that is commonly applied to trends and can be used with censored data. It is computed by comparing conforming pairs of X, Y data (i.e., they move in the same direction) with nonconforming pairs of X, Y data (those moving in opposite directions). Kendall's tau (τ) can be used with data containing multiple reporting limits.

6.3 Example

Correlation coefficients can also be computed using a contingency table with binomial data when values are coded by assigning 0 to all censored values and 1 to all uncensored values. The test is whether the frequency of detection changes from one group to another. To compute the binomial correlation coefficient phi (φ) the variables X and Y each is classified as low or high; there are four combinations of these two categories (Table 6). The correlation coefficient φ will have the highest positive value with both variables are high and both are low and the largest negative value when each variable is in a different category.

To illustrate this method a subset of concentrations of the herbicide atrazine in ground waters is used to prepare a contingency table that is then analyzed using the proportion-

Table 7: Subset of atrazine concentrations in ground water.

June	September	Rank in June	Rank in September
0.38	2.66	10	10
0.04	0.63	8	8
<0.01	0.59	2.5	7
0.03	0.05	6.5	5
0.03	0.84	6.5	9
0.05	0.58	9	6
0.02	0.02	5	4
<0.01	0.01	2.5	3
<0.01	<0.01	2.5	1.5
<0.01	<0.01	2.5	1.5

Table 8: Atrazine concentrations in June and September with each pair classified as low or high.

		June	
		Low	High
September	Low	2	0
	High	2	6

ality test (Table 7).[9] Table 8 displays the counts of pairs of low values (< 0.01) and high values (detected, GTE 0.01) for the atrazine data in Table 7. For these data, the correlation coefficient φ is calculated to be 0.61. Whether this is significant depends on the question being asked. If the question is whether only a positive correlation exists between June and September, the probability $p = 0.027$ and the answer is there is a positive correlation between the two months. If both positive and negative correlations are of interest the probability $p = 0.54$.

There are also correlation methods for data using ordinal and MLE methods for singular or multiply censored data.

7 Regression and Trends

7.1 Introduction

Regression methods examine the relationships between response variables and explanatory variables; that is, they quantify cause-and-effect relationships. In environmental data

[9] Junk, G.A., R.F. Spalding, and J.J. Richard. 1980. Areal, vertical, and temporal differences in ground water chemistry: II. Organic constituents. Journal of Environmental Quality **9**:479–483.

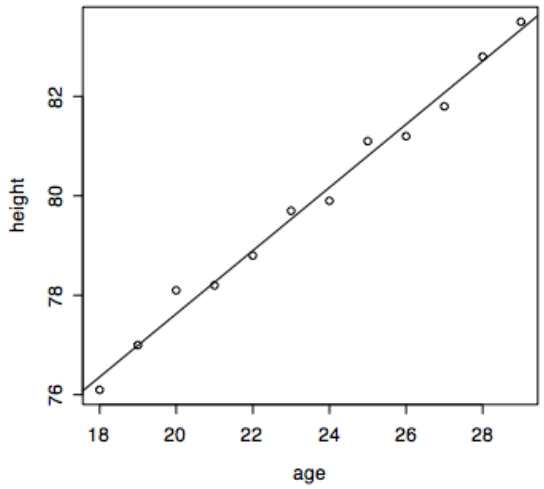

Figure 10: Example of a linear regression relating height to age.

analyses under the CWA, ESA, NEPA, and other laws regression models can provide insights on that inform policy and regulatory decisions.

Linear regressions are very familiar; they have a line through the data which represent the mean value of a response variable on the Y axis to the value of an explanatory variable on the X axis (Figure 10). Rarely, if at all, can untransformed geochemical data be correctly fit by a linear regression model. When there are censored values in the data set the parametric linear regression does not produce correct answers.

There is a robust nonparametric analog to the parametric linear regression that can be used with censored data; it is recommended for use with all environmental geochemical data.[10] This method is the Theil-Sen regression (Figure 11). The slope is estimated by the nonparametric Kendall's tau correlation coefficient. This line is a linear median (not mean) and is not influenced by outlying observations. The significance text for Theil-Sen slope is the same as the significance test for Kendall's tau. When the explanatory variable on the X axis is time, the significance test of the Theil-Sen slope tests for a temporal trend.

As with previous sections, there are regression methods suitable for censored data in binomial and MLE formats. If there are censored values in both the response (Y axis) and explanatory (X axis) variables the Theil-Sen line is the preferred method to apply to the data. For binomial data logistic regression models are appropriate.

[10]Helsel, D.R. and R.M. Hirsch. 2002. Statistical Methods in Water Resources. Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Chapter A3, Hydrologic Analysis and Interpretation. 524 pp.

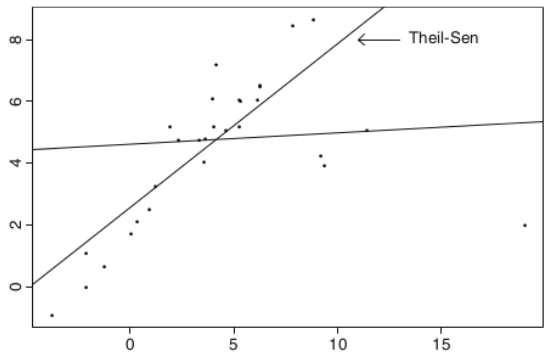

Figure 11: Parametric linear regression line (below) misses most of the positive associations captured by the nonparametric Theil-Sen estimator.

7.2 Example

In Section 5 above, a categorical approach was used to determine if there were differences in Long Island ground water TCE concentrations in three density categories of residential housing. Here the question to be answered is the probability of detecting TCE in shallow ground waters (the response variable) using population density and/or depth to the water table as explanatory variables. A maximum likelihood method was applied to the data and the likelihood ratio test applied to the population density and water table depth together to determine if the two factors significantly effected the probabilities of TCE concentrations being uncensored (≥ 5 µg/L). The results of a binary logistic regression is that water table depth does not significantly determine TCE concentrations ($p = 0.44$), but population density does significantly affect the levels of TCE in shallow ground waters ($p = 0.002$).

There are other regression and trend statistical models that are robust and produce technically sound and legally defensible explanatory and predictive answers to environmental effects of geochemicals.

8 Summary

You are not ecologists or environmental data analysts, but you rely on correct statistical analyses of geochemical data and interpretation of the results based on established ecological theory. Therefore, it is essential that you recognize when censored values are mis-handled with resulting distorted and incorrect results. Regardless of the setting: pre-application planning; permit application preparation; permit compliance issues with regulators and others; or litigation, understanding the issues involved with censored geochemical data is critical. Knowing there are robust and valid approaches to analyze cen-

sored geochemical data makes you a better informed consumer of reports prepared by technical consultants, regulators, and resource agency staff.