Environmental Science & Technology

Article

pubs.acs.org/est

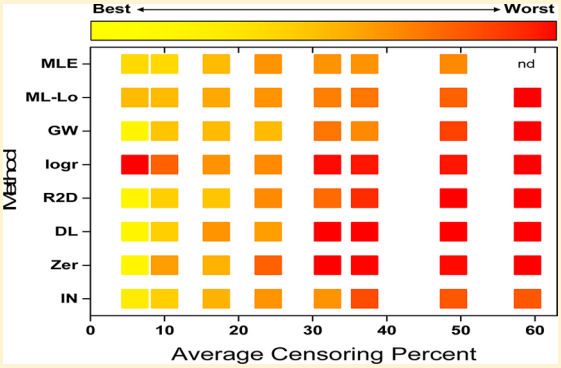
Evaluation of Statistical Treatments of Left-Censored Environmental Data Using Coincident Uncensored Data Sets. II. Group Comparisons

Ronald C. Antweiler*

U.S. Geological Survey, 3215 Marine Street, Boulder, Colorado 80309, United States

Supporting Information

ABSTRACT: The main classes of statistical treatments that have been used to determine if two groups of censored environmental data arise from the same distribution are substitution methods, maximum likelihood (MLE) techniques, and nonparametric methods. These treatments along with using all instrument-generated data (IN), even those less than the detection limit, were evaluated by examining 550 data sets in which the true values of the censored data were known, and therefore "true" probabilities could be calculated and used as a yardstick for comparison. It was found that technique "quality" was strongly dependent on the degree of censoring present in the groups. For low degrees of censoring (<25% in each group), the Generalized Wilcoxon (GW) technique and substitution of $\sqrt{2}/2$ times the detection limit gave overall the best results. For moderate degrees of censoring, MLE worked best, but only if the distribution could be estimated to be normal or log-normal prior to its application; otherwise, GW was a suitable alternative. For higher degrees of censoring (each group >40% censoring), no technique provided reliable estimates of the true probability. Group size did not appear to influence the quality of the result, and no technique appeared to become better or worse than other techniques relative to group size. Finally, IN appeared to do very well relative to the other techniques regardless of censoring or group size.



INTRODUCTION

Environmental data sets are frequently "left-censored", indicating that some values are less than the limit of detection (<LOD) for the analytical methods employed. These data sets are problematic because values of censored data are known only to reside in a range between zero and the LOD, thus complicating any statistical treatments used to summarize, compare, or regress the data. Researchers have employed several schemes to treat such data sets ranging from substitution for < LOD data based on some prescribed rule[1−3] through modeling the data set (or a portion thereof) according to an assumed statistical distribution (parametric techniques),[4−7] to nonparametric techniques which generally rely only on ranked information.[8−10] Other authors have suggested the use of raw instrument data below the detection limit.[11,12]

The comparison of two groups of uncensored data is statistically straightforward (relying on the *t* test or the nonparametric Kruskal−Wallis test), but the same cannot be said for censored data sets because part of the data are not explicitly known. Previous works evaluating group comparisons on censored data[13−16] have either reviewed older papers or evaluated techniques based on simulated or truncated data. Although there is a widespread condemnation of substitution techniques among these papers, there is no consensus regarding whether maximum likelihood estimation (MLE) or nonparametric techniques are to be preferred. A number of studies have proposed new methods using variations of established techniques,[17−21] but these also establish no clear consensus. This paper reevaluates the most common statistical techniques for group comparison of censored data by using real, censored environmental data sets where the true values of the censored data are known because they were also measured by more sensitive analytical techniques. The current work will not evaluate multiple imputation techniques; it also will not investigate "interval censored" or "right censored" data sets.

STUDY DESIGN

The basic study design is similar to that employed previously,[22] and is summarized here. Each sample from every data set was analyzed for up to 55 inorganic elements comprising practically all that naturally occur. Many of these elements were coincidentally determined using two different chemical analytical techniques, one of which was far more sensitive than the other; this gave rise in one case to an uncensored data set and in the other to a censored one. Because of this, the censored data set could be "completed" by replacing its nondetects with their corresponding uncensored values from the more sensitive analysis. Details of how the censored data

Received: May 14, 2015
Revised: September 24, 2015
Accepted: October 22, 2015
Published: October 22, 2015

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439−13446

sets were completed can be found in the Supporting Information (SI).

Each data set examined in this paper consists of two groups of censored data upon which a group comparison was made. Because the data are censored, the traditional techniques (the *t* test and Kruskal–Wallis, depending on the statistical distribution of the data set) cannot be employed, necessitating the use of censored statistical techniques. However, the "completed" data set *is* uncensored (by construction), and therefore, those traditional techniques can be used on it to calculate a "true" probability, p_{true}, that the two groups originated from the same distribution. Each p_{true} therefore serves as a yardstick to evaluate the p_{test} values from the censored statistical methods.

In order to determine p_{true} for a given "completed" data set, the statistical distribution type of that data set (normal, log-normal or other) needed to be determined first. This was done by applying the Anderson-Darling statistic (AD) for the normal and log-normal distributions on it. If AD < 0.787 for either of these (corresponding to a probability of ∼0.05), the data set was assigned to that distribution; if AD ≥ 0.787 for both the normal and log-normal distributions, the underlying distribution was deemed to be "other". The software package Minitab was used to calculate AD (which is largely independent of the number of samples). Once the distribution type of the "completed" data set was determined, p_{true} was calculated using the *t* test if it were normal, the *t* test on log-transformed data if log-normal, or the Kruskal–Wallis test otherwise. The "true" distribution type and p_{true} for each data set are tabulated in SI.

Turning to the determination of p_{test} values for censored data sets, the majority of the statistical treatments evaluated here represent the most commonly cited in environmental sciences literature. These can be broadly divided into five different categories (the bold font is the abbreviation used here):

(1) Substitution methods in which censored data are replaced according to a specific rule. Those examined were: Replacing all censored values with (a) zero (**Zer**); (b) the LOD (**DL**); (c) half the LOD (**HDL**); (d) $\sqrt{2}/2$ times the LOD (**R2D**); and (e) a random number between 0 and the LOD (**Rand**).

(2) Maximum likelihood estimation (**MLE**) techniques, which are schemes relying on knowledge of the underlying distribution from which the data are arrived. Uncensored data are used to calculate parameters which represent the best fit to the distribution, from which the group comparison probabilities are calculated.[15]

(3) Nonparametric methods, which generally rely only on the ranks of the data. The most common ones for group difference evaluation are (a) the Mann–Whitney U test (**MW**) in which all data less than the highest LOD are given the same rank; (b) the log-rank test adapted for left-censoring (**logrk**) which is appropriate for skewed data;[23] and (c) the Generalized Wilcoxon test (**GW**) which is a weighted version of the Gehan test,[15] and which accommodates multiple LODs.

(4) Regression on Order substitution test (**ROS**): By first assuming an underlying distribution, Regression on Order methods[24] are used to determine location and spread statistics of each group which are then used with traditional distribution-specific *t* tests to determine a probability; it should be pointed out that using regression on order statistics in this way is contrary to what the developers and promoters of it intended (Helsel, Pers. comm.).

(5) Using instrument-derived values (**IN**) from the laboratory for all data less than LOD; these numbers, which the laboratory normally discards and replaces with the LOD prior to releasing the data, have been suggested by some as containing useful information.[11,12,25,26]

Because **IN** and all the Substitution techniques assign values to the censored data, p_{test} probabilities can be calculated for each of these in three ways, namely using a normal *t* test, a log-normal *t* test or Kruskal–Wallis. This gives rise to 17 separate technique probabilities, distinguished by appending "**-No**", "**-Lo**", or "**-KW**" to the appropriate method (Table 1).

Table 1. 32 Statistical Tests (in Bold) Evaluated for Group Comparisons[a]

	assumed distribution:			
type	normal	lognormal	other	dependent on AD*
substitution				
zero:	**Zer-No**		**Zer-KW**	**Zer**
det.lim.:	**DL-No**	**DL-Lo**	**DL-KW**	**DL**
det.lim./2:	**HDL-No**	**HDL-Lo**	**HDL-KW**	**HDL**
det.lim.*$\sqrt{2}/2$:	**R2D-No**	**R2D-Lo**	**R2D-KW**	**R2D**
Rand number:	**Rand-No**	**Rand-Lo**	**Rand-KW**	**Rand**
parametric				
max. likelihood:	**MLE-No**	**MLE-Lo**		**MLE**
nonparametric				
Mann–Whitney:				**MW**[b]
log rank:				**logr**[b]
Gen.Wilcoxon:				**GW**[b]
hybrid				
ROS substitution:	**ROS-No**	**ROS-Lo**		**ROS**
instrument				
instrument:	**IN-No**	**IN-Lo**	**IN-KW**	**IN**

[a]The final column indicates (excluding nonparametric techniques) that the value for lognormal or other was used depending on what the Anderson-Darling estimate (AD*) of the distribution type was. [b]No assumed distribution.

Similarly, for both MLE and ROS techniques, p_{test} can be determined assuming either a normal or log-normal distribution. In what follows, a method to which is appended "**-No**" indicates that the data were treated as though the underlying distribution was normal and the test was applied on that basis. For example, "**MLE-No**" indicates that the p_{test} probability was determined using MLE methods assuming a normal distribution (regardless of what the underlying distribution might be). Similarly, a method to which "**-Lo**" is appended indicates that the data were treated as though the underlying distribution was log-normal and the test was applied on that basis; and methods to which "**-KW**" was appended indicates that no distribution was assumed, and that the Kruskal–Wallis test was applied to the data set. Altogether this gives rise to 24 p_{test} probabilities, 17 for the substitution methods, 4 for MLE and ROS, and 3 for the nonparametric methods (Table 1).

A reasonable approach followed by a researcher with censored data would be to attempt to determine the distribution type of that data set and then apply the appropriate

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439–13446

test based on the determination. Because part of the data are unknown (being censored), the AD statistic thusly calculated is an estimate (labeled here as AD*). Given these constraints, this same approach was used here by calculating an AD* value for the normal and log-normal distributions from each of the censored data sets. As with the "true" determination, if AD* < 0.787 for either the normal or log-normal distribution, the censored data set was estimated to have this type; otherwise it was labeled as "other". Once the underlying distribution-type was estimated, a technique probability was determined taking into account this information. Thus, for example, **R2D** (without the appendix "**-No**", "**-Lo**", or "**-KW**") represents the probability given by using either **R2D-No**, **R2D-Lo**, or **R2D-KW** depending on what the estimated distribution was. This consideration therefore gives an additional 8 p_{test} probabilities (**Zer**, **DL**, **HDL**, **R2D**, **Rand**, **MLE**, **ROS**, and **IN**); no probability could be calculated for **MLE** or **ROS** if the distribution was estimated to be "other". In summary, for methods which have an appendix, the scheme was applied with no consideration of the distribution; for the methods without an appendix (not including the nonparametric methods), the estimated distribution determined the probability that was selected. For reference, Table 1 lists all 32 of the statistical methods evaluated.

EXPERIMENTAL SECTION

550 distinct data sets, each comprised of censored data from a given analytical technique and "completed" by data from a more sensitive analytical technique, were used to evaluate the 32 statistical treatments. Data consist of inorganic analyses from studies conducted by the US Geological Survey during 1999–2014 comprising over 10000 samples from different media, including dissolved, "whole-water" (samples composed of both dissolved and particulate matter), sediment, vegetation, and animal tissue. Each group comparison was made as an environmentally logical step regarding the data in question. For example, in some cases the comparison was between the dissolved zinc in lakes within two National Parks; in others, it was between copper concentrations in the tops vs the bottoms of sediment cores; in yet others it was between chromium concentrations in two different strains of corn. Thus, the group comparisons which were used to evaluate the statistical techniques were not randomly chosen, but represent real, environmentally relevant comparisons which either were made or could have been made as part of the various studies from which the data originated.

All samples were analyzed by ICP-MS,[27,28] ICP-AES,[29,30] and/or IC;[31] of the 550 group comparisons, there were 55 distinct analytical lines used. Group sizes ranged from a minimum of 15 (7 in one group and 8 in the other) to 510 (261 and 249). Censoring levels ranged from 2.4% to 89.7% censored. Details of the above information are presented in the SI.

The determination of which of the 32 tests is best cannot be made with only one data set. Each test provides an estimate, p_{test}, of p_{true} which may by chance be close to it. Thus, it is desirable to examine many censored data sets with different conditions to see which test generally gives p_{test} values that are "closer" to p_{true} than the others. Because each data set generates a different p_{true}, whose value ranges between 0 and 1, how one defines "closer" becomes of great importance; it was therefore necessary to establish a metric M to quantify this. Initially, the difference in probabilities, $M_{test} = p_{test} - p_{true}$, was proposed, but was found to be inadequate by considering that for $p_{true} = 0.40$, a value of $M_{test} = 0.10$ would be considered a small difference (i.e., $p_{test} = 0.50$ is virtually the same as p_{true}), but for $p_{true} = 0.01$, this same value of $M_{test} = 0.10$ is very large (i.e., $p_{test} = 0.11$, which is vastly different than p_{true}). Next, the inverse of the Z-distribution, $\Phi^{-1}(p)$ (i.e., the z-score associated with the given probability) was considered. For all probabilities between 0.001 and 0.20, this appeared to work intuitively well, yet outside of these bounds it failed: consider $p_{true} = 0.0001$ and $p_{test} = 0.000001$, for which $\Phi^{-1}(p_{true}) = -3.72$, $\Phi^{-1}(p_{test}) = -4.75$; the difference between these two z-scores is very large, yet pragmatically the two probabilities are virtually equivalent. Finally, it was decided to create the metric from the z-score whenever the probability was in the range 0.001–0.20, but to alter it when outside so that differences in probability in these regions returned small differences in the metric. The altered z-score of a given probability p, $altz(p)$, is defined as

$$altz(p) = \begin{cases} -108605p^2 + 518.4p - 3.50 & p < 0.001 \\ \Phi^{-1}(p) & 0.001 \le p \le 0.20 \\ -0.3167p^2 + 0.855p - 0.9983 & p > 0.20 \end{cases} \quad (1)$$

By this definition, $altz(0) = -3.50$, and $altz(1) = -0.46$. The metric, M_{test}, used as the yardstick to evaluate a statistical scheme, test, is defined as

$$M_{test} = altz(p_{test}) - altz(p_{true}) \quad (2)$$

Some features of M are (1) M is bounded, i.e., $-3.04 \le M_{test} \le +3.04$; these extremes occur if $p_{test} = 0$ and $p_{true} = 1$ or vice versa. (2) If $M_{test} = 0$, then $p_{test} = p_{true}$. (3) For $M_{test} > 0$, the true probability is smaller (i.e., more likely significant) than the test probability. The metric as created satisfies the condition that it implies the same degree of "closeness" regardless of whether the value of p_{true} is 0.0005 or 0.68. More information on M is provided in the SI.

Each group comparison generates an M_{test} value for every statistical test, and thus considering all 550 group comparisons, it is appropriate to consider the set of all values of M_{test}, denoted as $\{M_{test}\}$. The objective of comparing test1 to test2 has thus been converted into comparing $\{M_{test1}\}$ to $\{M_{test2}\}$ by regarding each as a statistical set upon which location and spread statistics can be calculated. However, because each $\{M_{test}\}$ is highly non-normal, statistics such as the mean and standard deviation are not appropriate. Pragmatically more useful yardsticks are the median absolute deviation, D_{50}, defined as $D_{50} = \text{median}|M_{test}|$, and the 90th percentile absolute deviation, D_{90}, defined as $D_{90} = \text{Percentile}_{90}|M_{test}|$. D_{50} indicates that half the time a treatment's $altz$ will be within D_{50} of the true value of $altz$, which in turn indicates that half the time a given treatment's probability is within a value $altz^{-1}(D_{50})$ of the true probability. Bias, defined nonparametrically as Bias = median$\{M_{test}\}$, indicates whether the method tends to over- or underestimate the true probability. These three statistics are the tools used to evaluate the 32 tests.

RESULTS

A method's quality was expected to be strongly dependent on the amount of censoring, so this aspect was investigated first. Accordingly, data sets were subdivided into bins called "Areas" based on the degree of censoring of each group (and so that each "Area" had roughly the same number of data sets) (Figure

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439–13446

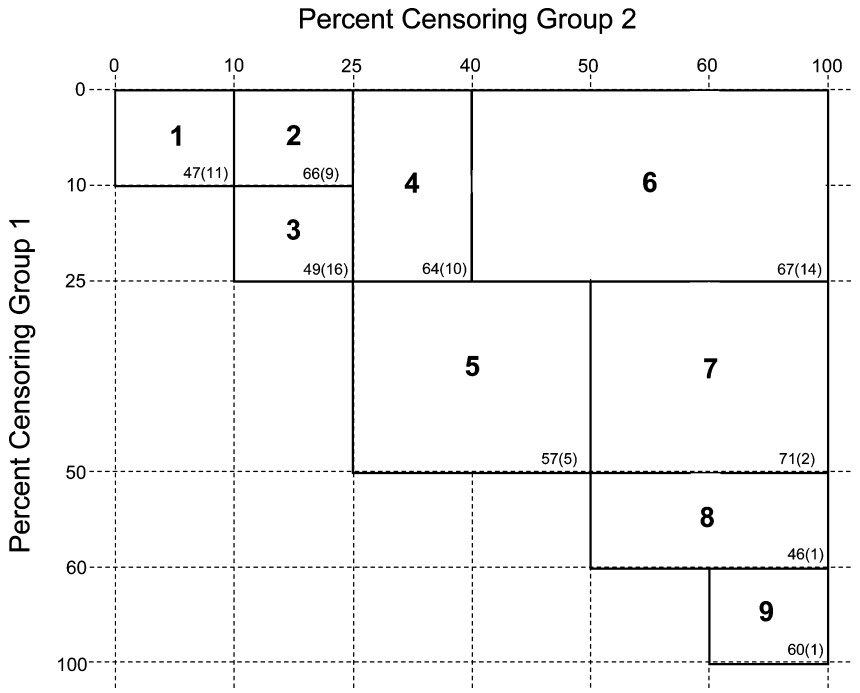


Figure 1. Setup used for the evaluation of the relationship between statistical technique and degree of censoring. Each rectangle is referred to as an "Area", denoted by the bold number in its center. Smaller numbers in the corners of the rectangles represent the number of group comparisons in that "Area", followed by (in parentheses) the number of comparisons for **MLE** and **ROS**. Group 1 is the group with the smaller percentage of censoring. Data sets estimated to be normal are not included (see the SI).

Table 2. "Area" Evaluations of 16 Statistical Techniques Using D_{50} and D_{90}[a]

Area	MLE Lo	IN KW	DL KW	HDL KW	R2D KW	IN	Zer	DL	HDL	R2D	Ran	ROS	MLE	MW	logr	GW
								D_{50}								
1	0.24	0.14	0.10	0.10	0.09	0.14	**0.12**	0.10	0.10	0.09	0.15	**0.14**	*0.12*	**0.17**	0.50	0.10
2	*0.23*	0.15	0.15	0.15	0.16	0.13	0.22	0.15	0.14	0.14	0.18	*0.17*	0.36	**0.26**	0.35	0.17
3	0.23	**0.22**	0.16	0.21	0.19	0.15	**0.21**	0.18	0.17	0.16	0.25	*0.48*	**0.12**	0.36	**0.34**	0.21
4	*0.28*	**0.23**	**0.24**	**0.30**	**0.25**	0.24	0.33	0.24	0.30	**0.23**	**0.31**	*0.31*	*0.26*	**0.37**	*0.33*	0.19
5	**0.39**	**0.41**	**0.45**	**0.41**	**0.37**	**0.43**	**0.41**	**0.53**	**0.41**	**0.43**	**0.54**	*na*	*na*	**0.81**	**0.52**	**0.34**
6	0.27	0.30	0.65	0.42	0.39	0.31	0.42	0.70	0.47	0.39	0.41	*0.27*	*0.24*	**0.39**	0.54	0.34
7	**0.35**	**0.41**	**0.57**	**0.48**	**0.58**	**0.41**	**0.56**	**0.57**	**0.48**	**0.59**	**0.49**	*na*	*na*	**0.38**	**0.54**	**0.38**
8	**0.60**	**0.42**	**0.74**	**0.71**	**0.92**	**0.43**	**0.73**	**0.74**	**0.71**	**0.92**	**0.74**	*na*	*na*	**0.87**	**0.72**	**0.68**
9	**0.94**	**0.84**	**0.71**	**0.79**	**0.75**	**0.84**	**1.00**	**0.71**	**0.79**	**0.75**	**0.65**	*na*	*na*	**1.12**	**0.88**	**0.92**
								D_{90}								
1	0.80	0.40	0.38	0.42	0.38	0.54	**0.42**	0.32	0.49	0.34	0.58	**0.49**	*0.36*	**1.69**	1.43	0.40
2	*0.92*	0.67	0.55	0.84	0.68	0.67	0.95	0.55	0.84	0.69	0.80	*na*	*na*	**1.20**	1.10	0.67
3	1.11	1.04	**0.97**	1.00	1.00	0.98	**0.95**	0.97	1.00	1.00	1.08	*1.05*	**0.88**	1.36	**1.58**	0.96
4	*0.88*	**1.13**	**0.84**	**1.61**	**1.03**	1.23	2.05	0.88	1.73	**1.21**	**1.28**	*1.33*	*1.10*	**1.17**	*1.17*	0.86
5	**1.95**	**1.29**	**1.81**	**1.52**	**1.71**	**1.77**	**1.90**	**1.81**	**1.52**	**1.65**	**1.73**	*na*	*na*	**2.72**	**1.79**	**1.86**
6	1.79	1.46	2.53	1.82	1.86	1.46	1.75	2.53	1.90	1.95	1.77	*2.45*	*1.62*	**1.87**	1.69	1.43
7	**1.17**	**0.97**	**1.84**	**1.30**	**1.39**	**0.97**	**1.58**	**1.84**	**1.32**	**1.39**	**1.41**	*na*	*na*	**1.54**	**1.45**	**1.30**
8	**2.19**	**1.90**	**2.49**	**2.15**	**2.02**	**1.90**	**2.42**	**2.49**	**2.15**	**2.02**	**2.38**	*na*	*na*	**2.53**	**2.40**	**2.30**
9	**2.44**	**1.93**	**2.40**	**2.25**	**2.24**	**1.93**	**2.58**	**2.40**	**2.25**	**2.24**	**2.45**	*na*	*na*	**2.67**	**2.31**	**2.45**

[a]These are generally the methods with the lowest values for D_{50} and D_{90}. Complete information on all 32 techniques is given in SI. Bright yellow shading indicates the technique that had among the best values for that "Area"; light yellow shading indicates the values that were a suitable alternative. Red shading indicates progressively worse values (darkest red are the worst values). No shading indicates the values were in the middle. Bold numbers indicate that the Bias was greater than 0.04, whereas italics indicate that the Bias was less than −0.04. *na* indicates that there were insufficient data to make an evaluation. Data presented do not include those data sets estimated to be normally distributed (see SI).

1). Thus, for example, Area 4 represents all data sets in which group 1 had 0−25% censoring and group 2 had 25−40% censoring; in the figure, numbers in the corner of Area 4 indicate that there were 64 such data sets, and that 10 of these were estimated to be log-normal.

The performances using D_{50} and D_{90} of 16 techniques are shown as a function of the degree of censoring in Table 2 and eight of these are shown graphically in Figure 2. Considering D_{50} (Figure 2A), although most of the methods exhibited a strong dependence on censoring, both **logr** and **ROS** are almost independent until the highest censoring levels; however, at low censoring levels these two are much poorer choices than other techniques. Results for the other six techniques shown here, the substitution method **R2D**, two instrument data

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439−13446

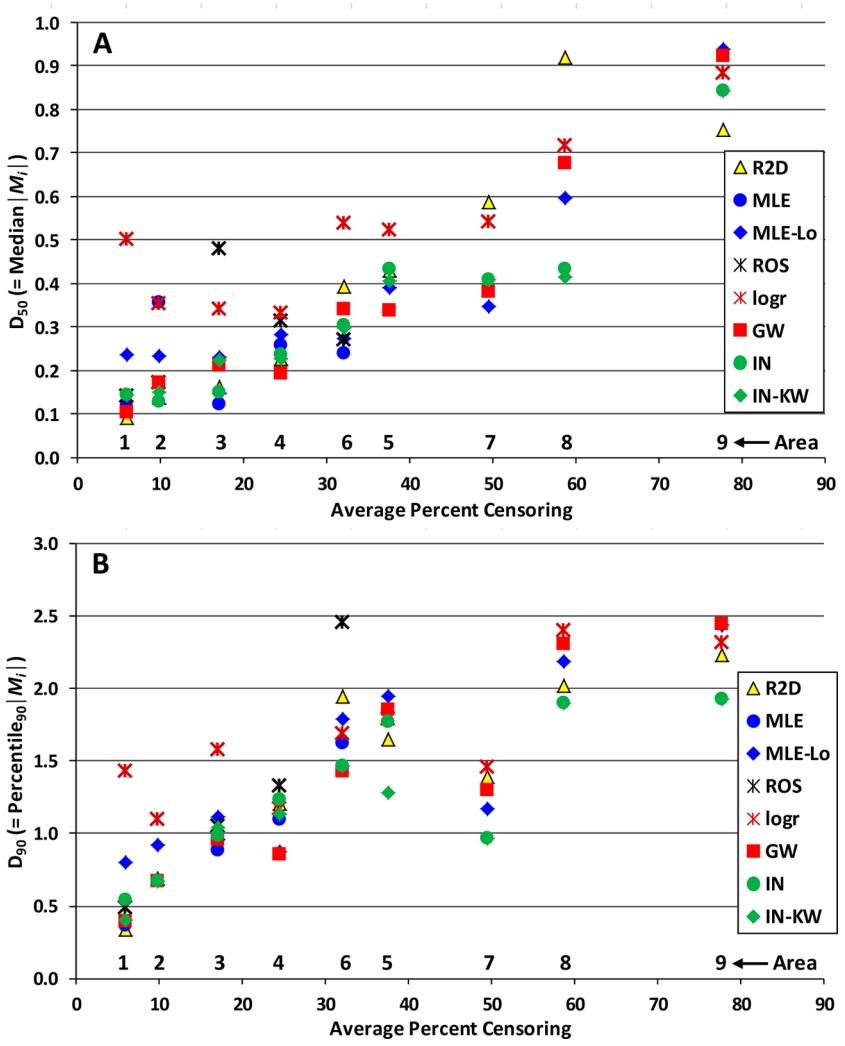


Figure 2. Behaviors of eight techniques as a function of average percent censoring. Panel A: D_{50}; panel B: D_{90}. Numbers above the *x*-axis denote the "Area". The "average percent censoring" is the median value of censoring percentages for all groups contained in that "Area". Data sets estimated to be normally distributed are not included (see the SI).

methods (**IN** and **IN-KW**), two maximum likelihood techniques (**MLE** and **MLE-Lo**) and the nonparametric method **GW**, all generally clump together, at least so long as the average censoring level is less than 40%. Thereafter, there is an increasingly large spread among them with **R2D** becoming much worse; above 60% censoring (overall), no technique does well. The finding that there is relatively little to distinguish any of the techniques plotted on Figure 2A at lower censoring levels (except **ROS** and **logrk**) is surprising and suggests that in spite of its widespread condemnation, **R2D** is as viable as any other method at low levels of censoring.

Figure 2B shows how the methods perform using D_{90} as a gauge. At low censoring levels (Area 1, < 10% censoring in each group), four techniques (**R2D**, **MLE**, **GW** and **IN-KW**) all had D_{90} values less than 0.40. To relate this to probabilities, suppose that the "true" probability is 0.04, implying a z-score of −1.75; a D_{90} value of 0.40 indicates a range of z-scores from −1.35 to −2.15, implying a range of probabilities from 0.016 to 0.088. Thus, nine times out of ten each of the four techniques above would return a probability in this range. In contrast, **MLE-Lo** has a D_{90} value of 0.80 indicating that nine times of ten the **MLE-Lo** probability falls in the much wider probability range of 0.0054 to 0.171. If a higher censoring level is considered, say, Area 3 (10−25% censoring in each group), the best technique (**MLE**) had a D_{90} value of 0.88, indicating that it would return a wider range than **MLE-Lo** did in Area 1. The point is that, with only moderate amounts of censoring (Area 3 or higher), there is relatively little assurance that even the best technique will actually provide a probability which is close to the true value.

Figure 2 demonstrates that most techniques lose quality with increasing average censoring percentage and also that the average amount of censoring is generally a good gauge of that quality (as opposed to the individual censoring levels in each of the two groups). Although only eight of the 32 techniques are shown, these are generally the best methods. Blindly choosing to assume a normal or log-normal distribution (all those with a **-No** or **-Lo** suffix) almost never gave good results relative to other methods although **MLE-Lo** was at times comparable in quality (Table 2). Methods which blindly used Kruskal−Wallis ("**-KW**") gave results which were usually similar to their "parent" methods (i.e., those without a suffix). Full information for all methods is included in SI.

Figure 3 presents a visual aid to understanding the above results for 12 of the "best" methods based the average censoring percentage. Each of the methods is color-coded based on its average D_{50} value ranging from yellow to red. For low

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439−13446

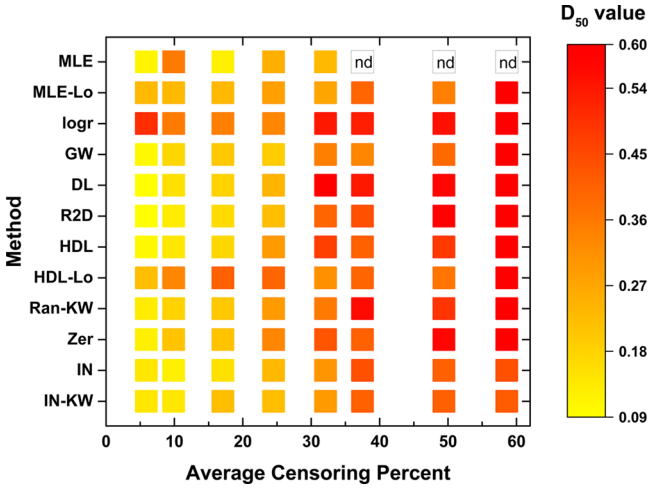


Figure 3. Performances of 12 techniques using the D_{50} statistic in relation to the average percent censoring. "nd" means there were insufficient data to evaluate.

censoring, **MLE**, **GW**, **DL**, **R2D**, and **HDL** all provide the best results; by 25% censoring, the best methods are **GW**, **R2D**, and **IN-KW**, yet the best D_{50} value here is considerably worse than for 5% censoring. By 50% censoring, no method is good, although **IN** and **IN-KW** are slightly better than others.

The dependence of the results on the sample size was investigated next. As with censoring, sample size is two-dimensional and therefore was broken into bins of roughly equal size to determine if the most relevant factor was the average group size, the smaller group or something else. Because results are strongly dependent on censoring, each of the factors of group-size and censoring were divided into four bins, thus giving a 4 × 4 arrangement and resulting in 16 separate bins to consider (Table 3). For clarity, these group size-censoring combinations are hereafter called "Collections" and labeled with a number and a letter to identify the group size and censoring, respectively -- for example, 2D represents group sizes of 31−50 and ≤50 and censoring levels of ≥40% and ≥40%. Because of the number of increased bins (16), only one "Collection" had more than 10 log-normal data sets; consequently, no determination of the dependence of **MLE** and **ROS** on group size was made for these.

Table 3. Composition of the "Collections" Used in the Assessment of the Importance of Group Size on Statistical Treatment[a]

"Collection"	group sizes	pct. censoring	# tests	# paramT
1A	≤30, ≤30	<20, <20	20	1
1B		<40, 20−40	29	5
1C		<40, ≥40	36	4
1D		≥40, ≥40	40	1
2A	≤50, 31−50	<20, <20	27	9
2B		<40, 20−40	25	6
2C		<40, ≥40	29	4
2D		≥40, ≥40	40	0
3A	≤40, >50	<20, <20	43	12
3B		<40, 20−40	33	5
3C		<40, ≥40	38	6
3D		≥40, ≥40	42	1
4A	>40, >50	<20, <20	28	6
4B		<40, 20−40	31	3
4C		<40, ≥40	23	5
4D		≥40, ≥40	43	1

[a]"Collection" numbers (the first character in the "Collection" title) refer to the group sizes; "Collection" letters (the second character in the "Collection" title) refer to the percent censoring. "# ParamT" refers to the number of tests in which the distribution was estimated to be log-normal. Data presented do not include those data sets estimated to be normally distributed (see the SI).

The D_{50} statistic was used to evaluate various statistical methods within each "Collection". Table 4 summarizes the results, with red shading indicating poor performance and yellow or light yellow indicating better; it is unsurprising that all methods which blindly assumed a normal distribution did poorly relative to the other tests for all Collections. Across all non-D "Collections" (i.e., at least one group <40% censoring), the best three methods were **IN-KW**, **IN**, and **GW**, with **R2D** close behind. In agreement with the above findings, if only and A and B collections were considered, **R2D** becomes among the best.

In contrast to censoring percentage, there did not appear to be any dependence of results on group size, that is, statistical techniques performed equally good (or bad) regardless of the group size. In addition, increasing group size did not appear to favor one technique over another. These statements must be tempered by the fact that neither MLE nor ROS were evaluated because of an insufficient number of data sets which were estimated to be log-normal.

DISCUSSION

As stated above, both the distribution-type of the completed data set and that of its coincident censored counterpart were determined. This latter was done because researchers cannot know a priori the distribution-type of a censored data set and must therefore estimate it, thus giving rise to discrepancies between them. Table 5 shows that the success ratio between the estimate of the distribution and its "true" type is poor. Of the 23 data sets estimated to be normally distributed, only 10 actually were normal; similarly, of the 69 estimated to be lognormally distributed, only 44 actually were. Failure to be able to predict the true distribution is a cautionary tale for all methods which rely on the distribution type, in this case, **MLE** and **ROS**, and represents a strong reason why MLE performs so well with modeled (simulated) data but not necessarily with real data. It further suggests that nonparametric techniques probably are "safer". Indeed, with highly censored data sets (in both groups), the ability of censored statistical methods to be able to accurately determine the distribution type is increasingly worse. Environmental researchers might be tempted to assume a given data set is log-normal and thereby apply MLE with this assumption; this course is not recommended.

In spite of the lack of theoretical underpinnings for the Substitution method **R2D**, judicious use of this can lead to answers which are generally no worse and sometimes better than traditionally sanctioned methodologies such as **MLE** and **GW**. This is especially so when the degree of censoring is small (less than 25% censoring). Thus, the wide-scale dismissal of all substitution techniques[10,15,32,33] appears unwarranted at least for group comparisons with these conditions; substitution of $\sqrt{2}/2$ times the detection limit (**R2D**) gives answers of the same level of quality as **MLE** or **GW**.

Finally, if they are available, using instrument values below the detection limit gives among the best answers when applied to group comparisons (Tables 2 and 4 and Figures 2 and 3),

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439−13446

Table 4. Summary of Evaluations of the 32 Statistical Tests for the Different "Collections" (Dins of Group Sizes and Percent Censoring) Using the D_{50} Statistic[a]

	1A	1B	1C	1D	2A	2B	2C	2D	3A	3B	3C	3D	4A	4B	4C	4D
#	20	29	36	40	27	25	29	40	43	33	38	42	28	31	23	43
#P	1	5	4	1	9	6	4	0	12	5	6	1	6	3	5	1
Tests																
IN-No	0.26	0.54	0.31	0.43	0.25	0.54	0.59	0.61	0.48	0.72	0.75	1.22	0.48	0.42	1.02	1.00
Zer-No	0.39	0.45	0.35	0.52	0.24	0.37	0.68	0.69	0.52	0.85	0.63	1.07	0.78	0.48	1.39	0.79
DL-No	0.33	0.34	0.47	0.44	0.25	0.49	0.80	0.65	0.39	0.82	1.07	1.20	0.42	0.71	0.79	1.01
HDL-No	0.38	0.41	0.28	0.43	0.28	0.47	0.56	0.64	0.42	0.67	0.92	1.17	0.54	0.44	0.68	0.79
R2D-No	0.36	0.36	0.29	0.45	0.27	0.55	0.76	0.64	0.41	0.74	1.03	1.16	0.46	0.57	0.65	0.91
Ran-No	0.40	0.49	0.27	0.41	0.29	0.46	0.72	0.72	0.49	0.79	0.98	1.27	0.65	0.52	0.64	0.78
ROS-No	0.42	0.32	0.43	0.58	0.35	0.36	1.14	0.60	0.37	0.39	0.80	1.24	0.43	0.42	0.39	1.05
MLE-No	0.40	0.46	0.39	0.43	0.35	0.45	0.46	0.66	0.33	0.62	0.61	1.12	0.50	0.32	0.62	0.59
IN-Lo	0.31	0.37	0.32	0.52	0.26	0.28	0.50	0.65	0.20	0.34	0.54	0.80	0.28	0.37	0.50	0.73
DL-Lo	0.28	0.19	0.34	0.25	0.24	0.26	0.52	0.60	0.42	0.32	0.93	1.26	0.21	0.45	0.39	0.94
HDL-Lo	0.30	0.26	0.29	0.29	0.28	0.28	0.71	0.57	0.30	0.35	0.72	1.23	0.36	0.62	0.37	0.67
R2D-Lo	0.25	0.18	0.26	0.28	0.24	0.29	0.48	0.55	0.24	0.31	0.82	1.14	0.25	0.59	0.54	0.83
Ran-Lo	0.29	0.24	0.35	0.60	0.50	0.23	0.46	0.57	0.58	0.43	0.95	1.06	0.51	0.65	1.06	0.97
ROS-Lo	0.37	0.35	0.28	0.40	0.23	0.55	0.57	0.78	0.36	0.68	0.69	0.95	0.42	0.35	0.76	0.75
MLE-Lo	0.23	0.20	0.21	0.41	0.24	0.25	0.49	0.66	0.26	0.19	0.41	0.96	0.22	0.51	0.58	0.57
IN-KW	0.22	0.22	0.37	0.40	0.17	0.14	0.50	0.58	0.21	0.24	0.48	0.74	0.10	0.12	0.41	0.41
Zer-KW	0.26	0.36	0.37	0.53	0.16	0.16	0.77	0.66	0.19	0.34	0.55	1.21	0.14	0.14	0.81	0.69
DL-KW	0.13	0.16	0.38	0.26	0.13	0.22	0.89	0.71	0.18	0.27	0.80	1.26	0.06	0.26	0.41	0.98
HDL-KW	0.16	0.27	0.38	0.21	0.10	0.22	0.85	0.70	0.21	0.18	0.55	1.08	0.09	0.19	0.60	0.72
R2D-KW	0.16	0.20	0.31	0.20	0.13	0.22	0.71	0.68	0.19	0.34	0.67	1.00	0.09	0.19	0.81	0.82
Ran-KW	0.17	0.28	0.34	0.57	0.13	0.27	0.71	0.54	0.21	0.26	0.54	0.96	0.11	0.28	0.68	0.46
IN	0.22	0.27	0.33	0.38	0.14	0.15	0.50	0.58	0.19	0.23	0.44	0.74	0.10	0.12	0.61	0.50
Zer	0.26	0.36	0.37	0.53	0.16	0.16	0.77	0.66	0.19	0.34	0.55	1.21	0.14	0.14	0.81	0.69
DL	0.13	0.22	0.36	0.26	0.12	0.30	0.85	0.71	0.18	0.27	0.84	1.26	0.05	0.25	0.65	0.98
HDL	0.15	0.22	0.36	0.21	0.12	0.21	0.83	0.70	0.24	0.18	0.76	1.08	0.09	0.19	0.60	0.72
R2D	0.15	0.16	0.26	0.20	0.10	0.21	0.71	0.68	0.18	0.31	0.73	1.00	0.08	0.19	0.81	0.82
Ran	0.17	0.33	0.30	0.60	0.16	0.27	0.80	0.54	0.23	0.27	0.60	0.94	0.10	0.28	0.41	0.44
MW	0.36	0.47	0.27	0.54	0.17	0.21	0.64	0.86	0.24	0.38	0.63	1.53	0.20	0.28	1.29	0.91
logr	0.34	0.32	0.37	0.59	0.23	0.25	0.89	0.70	0.60	0.38	0.54	0.83	0.48	0.66	0.78	0.62
GW	0.16	0.28	0.28	0.50	0.12	0.20	0.57	0.80	0.25	0.19	0.36	1.07	0.07	0.15	0.68	0.62

[a]Bright yellow shading indicates the technique that had among the best values for that "Collection"; light yellow shading indicates the values that were a suitable alternative; no shading indicates the values were in the middle; red shading indicates progressively worse values (darkest red are the worst values). # is the number of total group comparisons in the "Collection"; #P is the number of group comparisons estimated to be log-normal. Data presented do not include those data sets estimated to be normally distributed (see SI). For clarity, Collections 1*x* were small groups ranging up to 4*x* (large groups); Collections *x*A were low censoring ranging up to *x*D (high censoring).

Table 5. Number of Datasets Falling into Each Category of "True" Distribution (As Determined by the Anderson-Darling Statistic on the "Completed" Dataset) and Assumed Distribution (Estimated from the Censored Dataset)

		true distribution			
		normal	lognormal	other	total
estimated distribution	normal	10	9	4	23
	lognormal	0	44	25	69
	other	48	138	272	458
	total	58	191	301	550

especially if Kruskal–Wallis is used indiscriminately on these instrument values. As stated in an earlier publication,[22] instrument values can contain their own set of problems, principally being that negative numbers may be present, yet in spite of these limitations, **IN** and **IN-KW** were among the best techniques evaluated. It is difficult to say why this should be so, and the answer is likely to be dependent on both the laboratory protocols and the analytical procedures involved. Many laboratories, in an effort to prevent "false positives" (detecting a compound which is not really there), raise their detection limits; they have decided it is better to err on the side of caution in terms of detecting the presence of an analyte. It may be that, in so doing, they are actually discarding real information present below the censoring level.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.5b02385.

This addendum contains detailed explanations of (1) the structure of the paper; (2) the notation and terminology used; (3) the "completions" of the data sets; (4) summary information on the 550 data sets; (5) the metric, *M*; (6) full results for the "Area" evaluations; and (7) the normally distributed data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: antweil@usgs.gov.

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

The author would like to acknowledge that little of this work would have occurred but for the pioneering work of Dennis Helsel. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439–13446

REFERENCES

(1) Succop, P. A.; Clark, S.; Chen, M.; Galke, W. Imputation of data values that are less than a detection limit. *J. Occup. Environ. Hyg.* **2004**, *1* (7), 436−441.

(2) Baccarelli, A.; Pfeiffer, R.; Consonni, D.; Pesatori, A. C.; Bonzini, M.; Patterson, D. G., Jr.; et al. Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the seveso chloracne study. *Chemosphere* **2005**, *60* (7), 898−906.

(3) Ganser, G. H.; Hewett, P. An accurate substitution method for analyzing censored data. *J. Occup. Environ. Hyg.* **2010**, *7*, 233−244.

(4) Huybrechts, T.; Thas, O.; Dewulf, J.; Van Langenhove, H. How to estimate moments and quantiles of environmental data sets with non-detected observations? A case study on volatile organic compounds in marine water samples. *J. Chromatog. A* **2002**, *975*, 123−133.

(5) Cohn, T. A. Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data. *Water Resour. Res.* **2005**, *41* (7), 1−13.

(6) Hewitt, P.; Ganser, G. H. A comparison of several methods for analyzing censored data. *Ann. Occup. Hyg.* **2007**, *51*, 611−632.

(7) Jain, R. M.; Caudill, S. P.; Wang, R. Y.; Monsell, E. Evaluation of maximum likelihood procedures to estimate left censored observations. *Anal. Chem.* **2008**, *80*, 1124−1132.

(8) Feigelson, E. D.; Nelson, P. I. Statistical methods for astronomical data with upper limits. I. Univariate distributions. *Astrophys. J.* **1985**, *293*, 192−206.

(9) She, N. Analyzing censored water quality data using a non-parametric approach. *J. Am. Water Resour. Assoc.* **1997**, *33* (3), 615−624.

(10) Fievet, B.; Vedova, C. D. Dealing with non-detect values in time-series measurements of radionuclide concentration in the marine environment. *J. Environ. Radioact.* **2010**, *101*, 1−7.

(11) Porter, P. S.; Ward, R. C. Estimating central tendency from uncensored trace level measurements. *J. Am. Water Resour. Assoc.* **1991**, *27* (4), 687−700.

(12) Cressie, N. Limits of detection. *Chemom. Intell. Lab. Syst.* **1994**, *22*, 161−163.

(13) Clarke, J. U. Evaluation of censored data methods to allow statistical comparisons among very small samples with below detection limit observations. *Environ. Sci. Technol.* **1998**, *32* (1), 177−183.

(14) Brankov, E.; Rao, S. T.; Porter, P. S. Identifying pollution source regions using multiply censored data. *Environ. Sci. Technol.* **1999**, *33*, 2273−2277.

(15) Helsel, D. R. *Statistics for Censored Environmental Data using Minitab and R*; John Wiley and Sons: New York, 2012; p 324.

(16) Zhang, D.; Fan, C.; Zhang, J.; Zhang, C. H. Nonparametric methods for measurements below detection limit. *Statist. Med.* **2009**, *28*, 700−715.

(17) Millard, S. P.; Deverel, S. J. Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits. *Water Resour. Res.* **1988**, *24* (12), 2087−2098.

(18) Slymen, D. J.; de Peyster, A.; Donohoe, R. R. Hypothesis testing with values below detection limit in environmental studies. *Environ. Sci. Technol.* **1994**, *28*, 898−902.

(19) Zhong, W. *Statistical approaches to analyze censored data with multiple detection limits*; Ph.D. Dissertation; University of Cincinnati, 2005; p 98.

(20) Fu, L.; Wang, Y.-G. Nonparametric rank regression for analyzing water quality concentration data with multiple detection limits. *Environ. Sci. Technol.* **2011**, *45*, 1481−1489.

(21) Yu, J.; Vexler, A.; Hutson, A. D.; Baumann, H. Empirical likelihood approaches to two-group comparisons of upper quantiles applied to biomedical data. *Stat. Biopharm. Res.* **2014**, *6* (1), 30−40.

(22) Antweiler, R. C.; Taylor, H. E. Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environ. Sci. Technol.* **2008**, *42*, 3732−3738.

(23) Lee, E. T.; Wang, J. W. *Statistical Methods for Survival Data Analysis*; John Wiley and Sons: New York, 2003; p 513.

(24) Helsel, D. R.; Cohn, T. A. Estimation of descriptive statistics for multiply censored water quality data. *Water Resour. Res.* **1988**, *24* (12), 1997−2004.

(25) Porter, P. S.; Ward, R. C.; Bell, H. F. The detection limit. *Environ. Sci. Technol.* **1988**, *22* (8), 856−861.

(26) Lambert, D.; Peterson, B.; Terpenning, I. Nondetects, detection limits, and the probability of detection. *J. Am. Stat. Assoc.* **1991**, *86*, 266−277.

(27) Garbarino, J. R.; Taylor, H. E. Inductively coupled plasma-mass spectrometric method for the determination of dissolved trace elements in natural water. *U.S. Geological Survey Open-File Report 94−358* **1996**, 88.

(28) Taylor, H. E. *Inductively coupled plasma-mass spectrometry - practices and techniques*; Academic Press: San Diego, 2001; p 294.

(29) Mitko, K.; Bebek, M. ICP-OES determination of trace elements in salinated water. *Atom. Spectros.* **1999**, *20*, 217−223.

(30) Mitko, K.; Bebek, M. Determination of major elements in saline water samples using a dual-view ICP-OES. *Atom. Spectros.* **2000**, *21*, 77−85.

(31) Brinton, T. I.; Antweiler, R. C.; Taylor, H. E. Method for the determination of dissolved chloride, nitrate and sulfate in natural water using ion chromatography. *U.S. Geological Survey Open-File Report 95−426A* **1996**, 16.

(32) Huston, C.; Juarez-Colunga, E. Guidelines for computing summary statistics for datasets containing non-detects. 2009, Online citation: http://bvcentre.ca/files/research_reports/08-03GuidanceDocument.pdf.

(33) Ogden, T. L. Handling results below the level of detection. *Ann. Occup. Hyg.* **2010**, *54*, 255−256.

DOI: 10.1021/acs.est.5b02385
Environ. Sci. Technol. 2015, 49, 13439−13446