Environ. Sci. Technol. **2008,** *42,* 3732–3738

Evaluation of Statistical Treatments of Left-Censored Environmental Data using Coincident Uncensored Data Sets: I. Summary Statistics

RONALD C. ANTWEILER* AND HOWARD E. TAYLOR

U.S. Geological Survey, 3215 Marine Street,
Boulder, Colorado 80303

Received June 1, 2007. Revised manuscript received February 25, 2008. Accepted February 29, 2008.

The main classes of statistical treatment of below-detection limit (left-censored) environmental data for the determination of basic statistics that have been used in the literature are substitution methods, maximum likelihood, regression on order statistics (ROS), and nonparametric techniques. These treatments, along with using all instrument-generated data (even those below detection), were evaluated by examining data sets in which the true values of the censored data were known. It was found that for data sets with less than 70% censored data, the best technique overall for determination of summary statistics was the nonparametric Kaplan–Meier technique. ROS and the two substitution methods of assigning one-half the detection limit value to censored data or assigning a random number between zero and the detection limit to censored data were adequate alternatives. The use of these two substitution methods, however, requires a thorough understanding of how the laboratory censored the data. The technique of employing all instrument-generated data—including numbers below the detection limit—was found to be less adequate than the above techniques. At high degrees of censoring (greater than 70% censored data), no technique provided good estimates of summary statistics. Maximum likelihood techniques were found to be far inferior to all other treatments except substituting zero or the detection limit value to censored data.

Introduction

Environmental data sets frequently contain values below the limit of detection by the analytical techniques employed. These values—referred to as "left-censored" data—indicate only that sample concentrations are less than some number, creating a multitude of problems for scientists and policy makers. Researchers have utilized many approaches to treat left-censored data, ranging from data substitution, the replacement of censored numbers according to a prescribed rule (*1–5*) through modeling the data set (or a portion thereof) based on a known statistical distribution (*6–9*), to nonparametric methods that utilize only rank information (*10–13*). Evaluations of these approaches have used both published "real" environmental data and generated (simulated) data that have attempted to model the "real world". In most cases involving real environmental data, however, the "true" values of the censored data were unknown. In studies with simulated or generated data, a statistical distribution had to be assumed a priori which may not have adequately modeled the "real world". It is the main purpose of this paper to re-evaluate the best known statistical techniques that have been utilized by employing censored real environmental data sets in which the "true" values were known because they were measured.

Only summary statistics on censored data are considered in this paper. The mean, median, 25th (Q1), and 75th (Q3) percentiles were chosen as representative of location statistics; the standard deviation (SD) and interquartile range (IQR) were used as representative of spread statistics.

Study Design

The a priori assumption underpinning all environmental data is that when a laboratory measures a quantity of interest in a sample it is measuring in that sample a real environmental value that is independent of the analytical technique used to obtain it. Thus, for example, when dissolved vanadium is measured by two scientifically reputable analytical techniques, the assumption is made that although the values obtained may differ slightly, each technique is measuring the *same* true quantity. Because different analytical techniques have differing sensitivities, they will inevitably have differing levels of detection and therefore potentially differing levels of censoring for a given set of samples. This is the essential tool exploited in this paper: samples were coincidentally analyzed by different analytical techniques, giving rise on one hand to a censored data set and on the other to an uncensored one, with the underlying assumption that although there might be analytical variability between the analyses, each was measuring the same real quantity. Thus, a censored data set could be "completed" by replacing censored values with their uncensored counterparts from the more sensitive analysis, thereby creating the opportunity to measure how the best-known statistical treatments for censored data compare with each other and, more importantly, with the true value. For example, using the "completed" data set, the true value of the mean of that data set can be calculated, which can then be compared with the values estimated for it by using various statistical treatments on the original, censored data set.

To evaluate the main statistical treatments used for censored data (discussed below), a data set pair containing one censored line and one uncensored (and usually far more sensitive) line for the same element was selected. The censored data within the censored data set were replaced with data from the uncensored line, leading to a "complete" uncensored data set which served as the control. A statistical parameter was calculated on both the control and, using a given statistical treatment, the original censored data sets. The treatment was then evaluated according to its agreement relative to the true value derived from the control data set.

The treatments evaluated were some of the more popular used in the literature. (1) Substitution methods are techniques whereby values are assigned to censored data according to a specified rule. The specific rules evaluated were (letters in parentheses denote the shorthand code used for this method): (a) Replacing all censored values with zero (Zero); (b) Replacing all censored values with their respective detection limits (DL); (c) Replacing all censored values with one-half their respective detection limits (Half); (d) Assigning to each censored datum a randomly chosen value between 0 and the detection limit (Rand).

(2) Maximum likelihood estimation (ML) techniques are methods that rely on knowing the underlying statistical

* Corresponding author fax: (303) 541-3084; e-mail: antweil@usgs.gov.

10.1021/es071301c Not subject to U.S. Copyright. Publ. 2008 Am. Chem. Soc.
Published on Web 04/16/2008

TABLE 1. Results (Top Half) and Evaluations (Bottom Half) of the Censoring Treatments for V292[a]

Statistic	True	Zero	DL	Half	Rand	ML-no	ML-lo	NP-KM	ROS	Lab
PARAMETER RESULTS										
Mean	0.684	0.573	0.786	0.680	0.683	0.685	0.668	0.680	0.685	0.634
Median	0.486	0.428	0.643	0.500	0.513	0.685	0.494	0.505	0.505	0.486
Q1	0.342	0.000	0.395	0.350	0.327	0.189	0.293	0.328	0.319	0.271
Q3	0.758	0.745	1.000	0.745	0.817	1.181	0.834	0.795	0.804	0.786
SD	0.725	0.786	0.714	0.728	0.736	0.735	0.608	0.748	0.730	0.768
IQR	0.416	0.745	0.605	0.395	0.490	0.992	0.542	0.466	0.485	0.515
TREATMENT EVALUATIONS										
Mean	-	-16.2	15.0	-0.6	-0.1	0.3	-2.3	-0.5	0.2	-7.2
Median	-	-11.9	32.2	2.9	5.6	41.0	1.6	3.9	3.9	0.0
Q1	-	-100.0	15.5	2.3	-4.4	-44.6	-14.5	-4.0	-6.6	-20.7
Q3	-	-1.7	32.0	-1.7	7.8	55.9	10.1	4.9	6.1	3.8
SD	-	8.4	-1.6	0.4	1.5	1.4	-16.2	3.1	0.7	5.9
IQR	-	79.2	45.5	-5.0	17.9	138.5	30.3	12.1	16.7	23.9

[a] All values are in μg/L. The numbers in the evaluation section of the table are the percent deviations, D (defined in eq 1), from the true values as given in the top half of the table. The colors indicate the "quality" of the result: Yellow (best, $|D| \leq 5\%$), followed in turn by light green ($5\% < |D| \leq 10\%$), tan ($10\% < |D| \leq 15\%$), pink ($15\% < |D| \leq 20\%$), and red ($|D| > 20\%$).

distribution from which the data are derived. Uncensored data are then used to calculate fitting parameters that represent the best fit to the distribution, and from these the various statistical parameters can be calculated. ML techniques are sensitive to outliers and do not perform well if the data do not follow the assumed distribution (*14*). ML was evaluated by assuming for the data the following distributions: (a) a normal distribution (ML-no); (b) a log-normal distribution (ML-lo).

(3) Nonparametric methods: the Kaplan–Meier technique (NP-KM). Nonparametric methods rely only on the ranks of the data and make no assumptions about the statistical distribution from which they originate. The standard nonparametric technique is the Kaplan–Meier method. Because it is nonparametric, Kaplan–Meier tends to be insensitive to outliers (a frequent occurrence in environmental data), which results in it generally working well with smaller data sets.

(4) Regression on order methods (ROS) are techniques that calculate summary statistics with a regression equation on a probability plot. These methods have become a popular alternative with the recent publication of a book by Helsel (*14*). The specific ROS technique evaluated herein is the more robust form advocated by Helsel, which uses uncensored data whenever possible and assumes a distribution only for censored data (*7, 14*).

(5) Instrument-generated data (Lab). In addition to the above, several researchers (*15–18*) have suggested that a superior approach would be to request that the analyzing laboratory provide all instrument-generated numbers including those less than the censoring level and therefore use a "complete" and uncensored data set from which all the traditional statistical treatments could be applied. According to the above papers, this approach—using these numbers below the detection limit instead of dealing with the problems which censored data engender— would be superior to any of the above methods. We chose to examine this claim by evaluating data sets containing all instrument-generated data including that below the detection limits. This technique we elected to name "Lab".

Experimental Section

Forty-four distinct data sets—comprised of censored data from a given analytical technique "completed" by data from a more sensitive analytical technique—were used to evaluate the statistical treatments. These contain inorganic data originating from roughly 5000 samples analyzed over the last 6 years by a research laboratory within the U.S. Geological Survey. Samples of different media were used and included both surface- and groundwater samples, lake sediment, and plant and animal tissues originating from locales throughout the western United States. The data sets varied in terms of number of samples (between 34 and 841) and percentage of censoring (from 13.7% to 94.5% censored) in addition to the media type.

Each sample was analyzed by two separate analytical techniques: inductively coupled plasma-mass spectrometry ICPMS (*19, 20*) and inductively coupled plasma-atomic emission spectroscopy ICPAES (*21–23*). For every analytical run, a detection limit, DL, was calculated for each analyte, whereby laboratory blanks were analyzed between 10 and 20 times per analytical run: DL = SD $t_{0.05,n}$, where SD is the standard deviation of the blanks and $t_{0.05,n}$ is the t-statistic at the 95% confidence level for n degrees of freedom (*24*). If a sample had a nominal concentration less than DL, its value was replaced by "< DL". ICPMS analyses are typically far more sensitive than ICPAES, giving rise to ICPMS detection limits which can be between 10 and 10,000 times lower than ICPAES for the same element.

Many elements were analyzed multiple times by ICPMS and ICPAES: for example, copper was analyzed using ICPMS at two different isotopes (63 and 65) and using ICPAES at three different spectral emission wavelengths (224.700, 324.752, and 327.393 nm). Throughout the rest of this paper, the term "lines" has been employed to represent either isotopes via ICPMS analyses or spectral wavelengths via ICPAES. For each data set pair, one line—typically that on the ICPMS—was completely uncensored while the other had a proportion of censored data. Detection limits for censored lines changed in response to differing instrumental conditions, leading to multiple censoring levels for all studied data sets.

The various statistical parameters were calculated using each censoring technique, and these were compared with the values calculated for the "true" data and registered as a percent deviation or bias, D

$$D = (b - t) \times 100/t \tag{1}$$

where b is the observed treatment value and t is the "true" value. Although this is the typical formulation used to calculate bias, it has the disadvantage that for a fixed value of t, D has a range from −100 to $+\infty$.

Results

A specific data set pair, that of dissolved vanadium in surface water samples from the Bear−Yuba drainage basins in

TABLE 2. Data Sets and Analytical Lines Used To Evaluate the Various Censoring Techniques[a]

Line	Total	% Censored	Line	Total	% Censored
Surface Water - Dissolved			**Fish Filet**		
Al 394.401	841	54.5	Ba 455.403	95	13.7
Al 396.153	841	19.4	Cu 324.752	95	41.1
Co 228.616	821	92.2	Mn 257.310	92	23.9
Co 230.786	821	94.5	Ni 231.604	34	73.5
Cu 324.752	841	69.1	Sr 460.716	94	35.1
Cu 327.393	841	77.6	Ti 334.940	94	77.7
Ni 231.604	825	30.5	**Tree Trunk**		
Ni 232.003	825	57.3	Cr 357.869	194	67.5
V 292.402	821	32.2	Cu 224.700	118	29.7
Zn 206.200	841	29.4	Cu 327.393	118	47.5
Zn 68*	841	23.5	Cu 324.752	118	55.9
Ground Water - Dissolved			Ni 231.604	166	27.1
As 193.696	158	62.0	Ni 232.003	166	61.4
Co 228.616	98	91.8	**Lichen**		
Cu 324.752	125	93.6	Cu 327.393	77	87.0
Mn 257.310	140	22.9	SiO_2 288.158	64	26.6
Mn 260.568	140	55.0	Sr 460.716	77	24.7
Fish Liver			Ti 336.121	77	22.1
Al 394.401	68	19.1	V 292.402	62	67.7
Ba 493.408	68	20.6	**Lake Sediment**		
Cd 214.440	75	52.0	As 188.980	711	90.6
Cd 226.502	75	29.3	B 208.957	713	16.7
Ni 231.604	42	73.8	Co 230.786	726	34.3
Sr 460.716	75	73.3	Ni 232.003	729	48.1
V 292.402	72	76.4			

[a] Numbers in the total column represent the total number of samples (both censored and uncensored). Within data sets, lines are sorted alphabetically. Surface water samples originated from the Bear−Yuba River drainages in northern California (unpublished data); groundwater samples were from Grand Canyon springs (*34*); fish liver and filet samples originated from National Parks in the western United States (*35*); tree trunk samples came from northern and central New Mexico (*36*); Lichen samples originated from National Parks in the western United States (*35*); lake sediment samples came from lakes in the western United States (*35*). All lines represent the wavelength (in nm) used for analysis on the ICP-AES, with the exception of the starred entry (Zn 68), which represents the isotope used on the ICP-MS analysis.

northern California (unpublished data), is examined in detail first; thereafter, the results from the remaining 43 lines are summarized.

The dissolved vanadium data set contained 821 samples and consisted of the censored ICPAES spectral line of 292.402 nm (designated as V292) and the uncensored ICPMS ^{51}V isotope (designated as V51). The V292 line had 264 censored samples, 32% of the total. To create an uncensored control data set, censored data were completed using equivalent samples from V51.

Values for the representative summary statistics (listed above) were calculated for the completed V292 line (designated hereafter as the "true" value) and for each of the four substitution techniques, the two ML methods, the nonparametric Kaplan–Meier, the ROS technique, and the instrument-generated data (Lab). Table 1 presents the true values and treatment results (in the top half) and their evaluation by comparison with the true values (in the bottom half). A color-coding scheme is employed to indicate the quality of the results, with yellow representing the best (being within 5% of the true value), grading through green, brown, and pink to red, representing the worst (values more than 20% away from the true value).

For location statistics, substituting zero (Zero) or the detection limit (DL) or using maximum likelihood estimation assuming a normal distribution (ML-no) tended to give poor results. For example, ML-no estimates that the last quartile (Q3) is 1.18 μg/L, while the true value is 0.76 μg/L, a bias of 56%. On the other hand, of the techniques that were tested, the one that gave the *best* results was substituting one-half the detection limit (Half): all four of its estimates of the location statistics were within 5% of the true value. Use of Half has been discouraged in the literature as being unreliable (*7, 13, 14, 25, 26*), yet the data here indicate otherwise. In addition to Half, Kaplan–Meier (NP-KM), ROS, and substituting a random value (Rand) gave good results, though not as good as Half. Maximum likelihood estimation assuming a log-normal distribution (ML-lo) performed less well in its estimates of Q1 and Q3, being biased by −14.5% and 10.1%, respectively. Finally, use of the instrument-generated numbers (Lab) gave estimates of a mean which were 7% low and an estimate of Q1 which was more than 20% low.

For spread statistics, the best censoring technique was again Half, with both estimates being within 5% of the true value. Rand, NP-KM, and ROS did well in estimating the standard deviation, but all of them were only mediocre in their estimates of IQR; ML-lo did poorly, being almost 17% too low in its estimate of the standard deviation and more than 30% too high in its estimate of the IQR. Finally, Lab was within 6% of the true value in estimating the standard deviation but more than 20% high in estimating the IQR.

Additional Data Sets

The above analysis was performed on 43 additional data sets (Table 2). As above, the true values of the summary statistics were determined by completing the censored data with uncensored data from a more sensitive uncensored line. Values for each statistic were calculated for every statistical treatment and compared with the true values. Because so much information was generated, the results are summarized below; additional information is available in the Supporting Information. The three treatments Zero, DL, and ML-no were almost always far inferior to the others in terms of bias (eq 1) and consequently are largely ignored in the ensuing discussions.

The data were first examined to determine if media type, data set size, or percentage of censoring affected the results. Media type and data set size each had only minor effects on the treatment results: there was no evidence that one treatment was favored over another because of these. Therefore, all lines regardless of sample type or number of samples were analyzed together. The percentage of censored data, however, did show a discrimination among treatments, and consequently, this was used as the explanatory or independent variable in the discussion which follows.

Because each censored line had multiple detection limits (rather than a single, fixed value), some of the results which follow may seem to be the result of a typographical or mathematical error but are, in fact, correct. For example, consider the following data set, which is 33% censored: $\{<2, 2, 3, 6, <7, 8\}$. If the Zero technique is used, the substituted data set becomes $\{0, 0, 2, 3, 6, 8\}$, which has a median of 2.5 and a Q3 of 6. If the DL technique is used, the substituted data set becomes $\{2, 2, 3, 6, 7, 8\}$, which has a median of 4.5 and a Q3 of 7. Thus, for this example, Zero, DL, and Half all will have different estimates of the median and Q3, in spite of the fact that only 33% of the data are censored. This simple example should be borne in mind with the results below: for example, it is frequently the case that Half, Zero, DL, and Rand have nonzero biases in their estimates of the median even for data sets containing less than 50% censored data.

Figure 1 plots in six panels the deviation (or bias) from the true value of the treatments Half, Rand, ML-lo, NP-KM, ROS, and Lab against percent censoring; each panel represents a different summary statistic. Among these statistics, there is a tendency for most of the treatments to cluster near zero bias so long as the percent censoring is less than 70% (the areas to the left of the lines on Figure 1). For example, for the mean (A), with the exception of at most one point, Half, Rand, NP-KM, and ROS all had deviations within 20% of the true value; for Q3 (D), with the exception of one point,

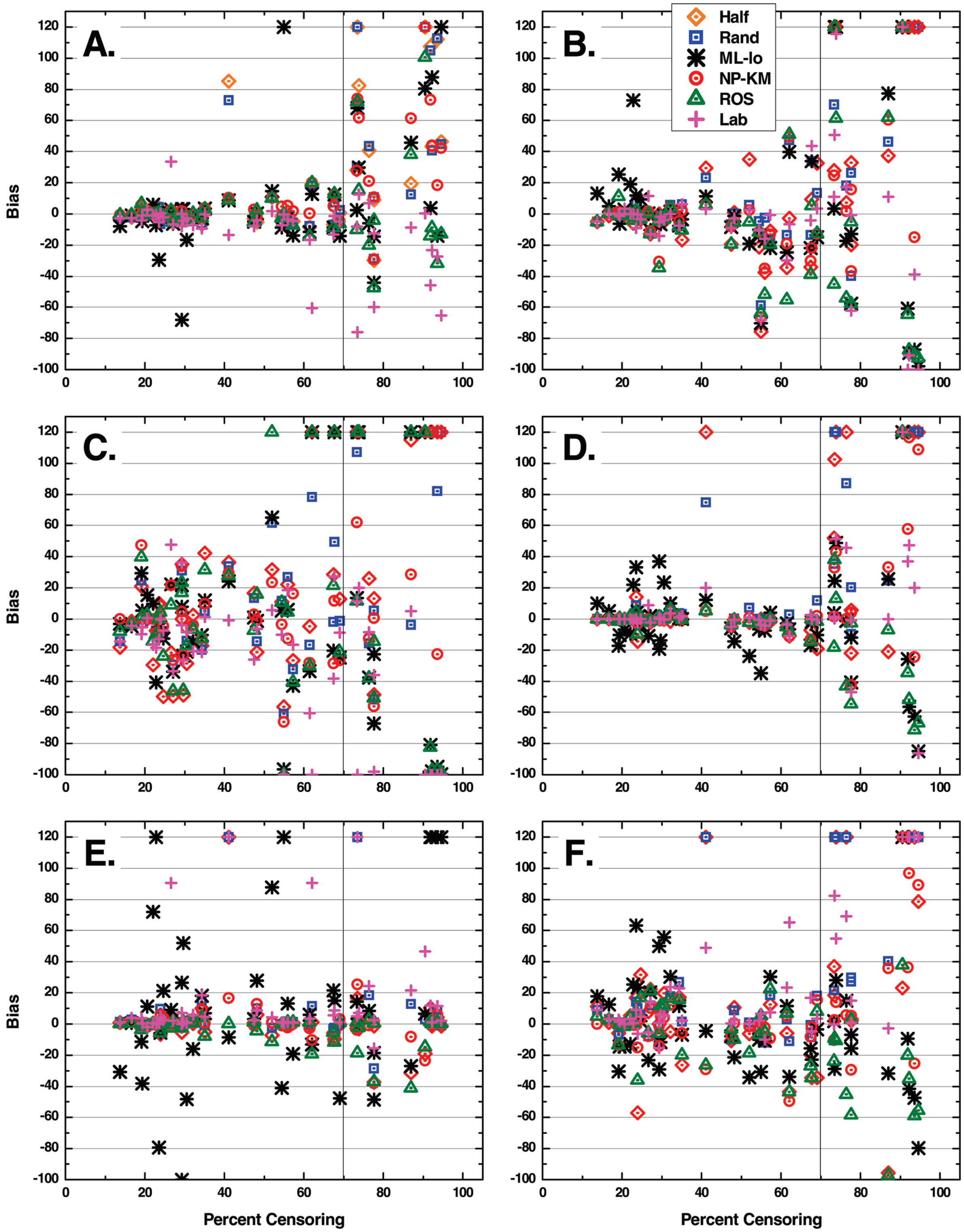

FIGURE 1. Evaluation of six treatment techniques (Half, Rand, ML-lo, NP-KM, ROS, and Lab) for estimation of (A) Mean, (B) Median, (C) Q1, (D) Q3, (E) SD, and (F) IQR. Biases larger than 120% were assigned a value of 120%.

all treatments except ML-lo were within 20% of the true value. For the median (B), Q1 (C), and IQR (F), there is far more scatter yet the values still cluster around zero bias. In each of these cases, ML-lo tends to be farthest from the true value—this is especially true for the standard deviation (E)—with the other five treatments mostly being indistinguishable from each other. Above 70% censoring (the areas to the right of the lines in figure 1), the graphs tend to "explode" both positively and negatively, indicating that there is little hope of predicting the true value of a statistic using any of the treatments.

The data for the mean were examined in a slightly different way in Figure 2. On the *y* axis are plotted the number of lines for which the treatment mean was more than 5% away (either positively or negatively) from the true mean, i.e., the "number of failures". On this graph, the lower the curve, the better the treatment is at estimating the true value. For example, there were nine "failures" at 70% censoring for the treatment Half,

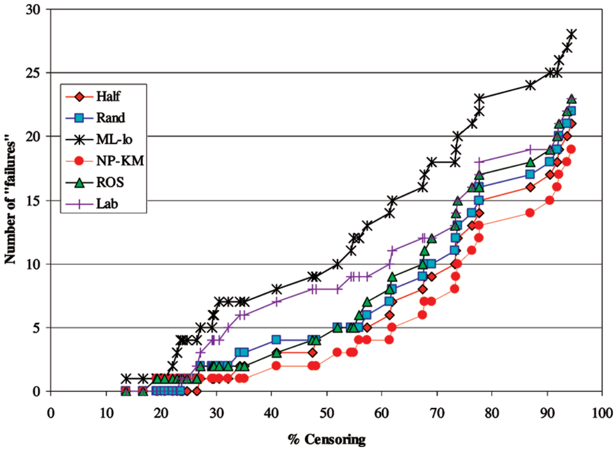

FIGURE 2. Evaluation of the treatment means in comparison with the true mean. A "failure" (on the *y* axis) is a treatment value that is more than 5% away from the true mean, i.e., |*D*| > 5%, where *D* is from eq 1.

indicating that for lines with less than 70% censoring, there were nine times in which the Half estimate of the mean was more than 5% away from the true mean. Because there were 32 lines in this study with less than 70% censored data, this indicates that the estimate of the mean provided by Half was within 5% of the true value of the mean 23/32 or 72% of the time. In contrast, NP-KM was within 5% of the true value 78% of the time, ROS was 63%, and NP-lo was only 44%.

To quantify the above observations, a scoring scheme was created to rate the treatments. First, data sets were divided into two classes: Class 1 consisted of those data sets with less than 70% censoring (containing 32 lines), and Class 2 consisted of those with more than 70% censored data (containing 12 lines). In both classes for each parameter, the median bias for each treatment was calculated; this allowed for a determination of whether treatments generally gave higher or lower values than the true value. In addition, the median of the absolute value of all biases was calculated to determine which treatment was generally nearest to the true value. The results are presented in Table 3.

For data sets with less than 70% censoring (Class 1), Half, Rand, NP-KM, and ROS were unbiased for all parameters. ML-lo and Lab were both slightly biased for a few parameters, and, as expected, Zero, DL, and ML-no had severe biasing. When considering the ability of a treatment to estimate a value close to the true value, NP-KM was best with only one parameter, Q1, with a median estimate which was more than 5% away from the true value. ROS, Half, Rand, and Lab were slightly worse, doing less well on Q1 and IQR. ML-lo was substantially worse than these, ranking about the same as DL; Zero and ML-no did the poorest.

In highly censored data sets (Class 2), all treatments were poor, although ML-lo and Lab were best when considering bias alone. In considering closeness to the true value, ROS and Lab were both better in general than ML-lo, yet all were poor. All four substitution techniques and NP-KM did substantially worse than these. In summary, for highly censored data sets, no treatment did well at predicting any parameter except the standard deviation. Additional information is presented in the Supporting Information.

Discussion

The above data analysis indicates that for data sets with less than 70% censoring, the best overall technique is NP-KM, with ROS, Rand, and Half being acceptable though inferior alternatives. Of interest is that neither of the two maximum likelihood techniques did well, and from the results of this study they cannot be recommended for use.

Although these findings would seem to suggest an endorsement of the two substitution techniques as easy-to-use alternatives, it cannot be stressed enough that they are not panaceas. The differences between a detection limit as defined in this paper and a quantitation limit in a production facility have not been considered here: Most users must deal with quantitation limits and do not have access to data sets with detection limits. There are some laboratories which report double the value of their censoring level while at the same time leaving data between their detection limits and its doubled value as detected observations. Thus, for example, if the detection limit is 1 and there are two measured values of 0.8 and 1.4, the laboratory would report these as <2 and 1.4 (*27*). In a situation like this, the conclusions reported above may no longer be valid, especially concerning the substitution techniques; in the example above, substituting one-half the detection limit would be equivalent to the DL treatment evaluated throughout this study, and this has been shown to be unreliable. In general, the substitution techniques which appear to work so well in the our study will not necessarily work well for data sets with quantitation limits which attempt to avoid both so-called "false positives" and "false negatives". However, if the user has access to laboratory detection limits prior to adjustments for "false negatives", then the substitution techniques should provide reasonable alternatives to NP-KM or ROS.

Although the best known techniques for computation of summary statistics on censored data are discussed here, there are a number of other treatments which have not been evaluated, among them being restricted maximum likelihood procedures (*28*), maximum likelihood procedures using the expected maximization algorithm (*29, 30*), and Winsorization methods (*31*). Multiple imputation (MI) (*32, 33*), a technique to determine statistics on data sets with missing data, was briefly investigated on three data sets (the "case study" data of V292 and two others). In all of these, the MI estimates were poor relative to NP-KM, ROS, Half, and Rand, and because MI is time consuming and tedious to use, it was decided to not evaluate it for all the remaining lines.

The use of laboratory-generated data below the detection limit (Lab) has been shown to be inferior to NP-KM and certainly no better than ROS, Half, or Rand (compare Figures 1 and 2 and Tables 1 and 3). There is an additional compelling reason to avoid using this approach: data generated by a laboratory are not necessarily exclusively greater than zero. In environmental data sets which have traditionally been thought to be at least nominally log-normally distributed, zero and negative numbers definitely would present problems beyond their physical impossibility. Because of all of these issues, Lab cannot be recommended as a treatment technique to deal with censored data.

Summary

Nine different statistical techniques commonly used to determine summary statistics in left-censored data sets were evaluated by examining how closely the treatment's estimate of a given parameter matched the "true" value of that parameter. The nine techniques were as follows: (1) substituting zero for censored data (Zero); (2) substituting the detection limit value for censored data (DL); (3) substituting one-half the detection limit value for censored data (Half); (4) substituting a randomly generated number between zero and the detection limit value for censored data (Rand); (5) maximum likelihood estimation assuming a normal distribution (ML-no); (6) maximum likelihood estimation assuming a log-normal distribution (ML-lo); (7) the nonparametric Kaplan–Meier technique (NP-KM); (8) regression on order

TABLE 3. Scoring for Treatments for Data Sets with Less Than 70% Censoring (Top Half) and Greater Than 70% Censoring (Bottom Half)[a]

	Treatment								
Statistic	Half	Rand	ML-lo	NP-KM	ROS	Lab	DL	Zero	ML-no
CLASS 1: DATASETS WITH LESS THAN 70% CENSORING									
Median Bias									
Mean	-0.7	-0.8	-1.6	0.2	-0.5	-3.8	12.1	-15.0	-0.6
Median	-2.2	-0.6	-2.3	-3.5	-4.4	-2.4	19.4	-12.4	59.8
Q1	0.2	-1.5	-2.2	-0.1	0.9	-9.3	35.8	-100.0	-30.3
Q3	-0.1	0.0	-4.9	0.0	-0.2	0.0	0.0	-0.5	50.0
SD	0.1	0.3	4.9	0.2	0.1	2.5	-1.7	5.5	0.0
IQR	0.0	2.2	-7.5	-0.4	-1.0	3.1	-5.4	20.7	70.9
Median Absolute Bias ("Closeness")									
Mean	2.7	3.2	6.1	2.0	2.7	4.1	12.1	15.0	2.6
Median	4.3	4.9	10.8	4.7	5.3	4.0	19.8	12.4	59.8
Q1	21.6	14.3	17.8	10.7	16.9	15.5	35.8	100.0	80.7
Q3	0.5	0.4	10.0	0.1	1.5	0.7	4.8	0.5	50.0
SD	0.9	1.1	20.3	1.3	1.1	2.5	2.7	5.5	1.1
IQR	10.8	8.2	19.0	4.5	9.0	5.3	14.2	20.7	70.9
CLASS 2: DATASETS WITH GREATER THAN 70% CENSORING									
Median Bias									
Mean	44.7	44.2	16.7	43.1	-9.6	-17.8	117.1	-41.5	45.1
Median	120.0	120.0	-15.0	90.1	-49.5	5.1	120.0	-100.0	120.0
Q1	120.0	113.5	-30.1	91.0	-26.2	-99.1	120.0	-100.0	-59.2
Q3	120.0	120.0	-6.5	33.0	-26.3	10.8	120.0	-100.0	120.0
SD	-1.4	1.6	7.5	-0.5	0.2	9.3	-1.7	2.4	-1.0
IQR	57.8	120.0	-12.6	9.8	-29.6	75.8	120.0	-100.0	120.0
Median Absolute Bias ("Closeness")									
Mean	44.7	44.2	36.9	43.1	17.6	17.9	117.1	41.5	45.1
Median	120.0	120.0	82.2	90.1	63.3	56.5	120.0	100.0	120.0
Q1	120.0	113.5	96.4	91.0	96.9	109.1	120.0	100.0	120.0
Q3	120.0	120.0	33.3	37.8	40.6	41.3	120.0	100.0	120.0
SD	4.2	2.5	22.8	2.6	1.9	11.8	20.3	3.4	2.5
IQR	87.2	120.0	28.6	27.5	36.5	75.8	120.0	100.0	120.0

[a] All values are the deviation or bias, D, from the true value as defined in eq 1. The colors indicate the "quality" of the result: Yellow (best, $|D| \leq 5\%$), followed in turn by light green ($5\% < |D| \leq 10\%$), tan ($10\% < |D| \leq 15\%$), pink ($15 < |D| \leq 20\%$), and red (worst, $|D| > 20\%$). Values of $|D|$ that were greater than 120% were replaced by 120%.

statistics (ROS); and (9) using instrument-generated data (Lab). Each of these treatments was evaluated for six different parameters: the mean, median, first quartile, third quartile, standard deviation, and interquartile range. Each treatment was evaluated for all of these parameters on 44 distinct data sets comprised of inorganic analytes in surface waters, ground waters, fish filets and livers, tree trunks, lichen samples, and lake sediment. These data sets ranged in size from 34 to 841 samples and in degree of censoring from 13.7% to 94.5% censored.

It was found that the sample type and number of samples had little effect on the quality of the results for the various treatments but that the degree of censoring had a large effect. For sample sets with less than 70% censored data, the best technique was NP-KM: not only was it least biased, but it also provided closest estimates of the parameters. ROS, Half, and Rand were acceptable though inferior alternatives to NP-KM; in general, they were unbiased estimators of the statistics, but they tended to provide estimates which were farther from the true value than NP-KM. The use of Half and Rand as general techniques, however, must be accompanied by a thorough understanding of the censoring schemes employed by the laboratory. The worst treatments were Zero and DL, with both maximum likelihood techniques being only marginally better than these: consequently, maximum likelihood is not recommended as an adequate tool for estimating summary statistics. Lab was marginally worse than ROS, Half, and Rand and far less superior than NP-KM and should therefore be avoided. For sample sets with greater than 70% censoring, there were no good techniques, although both ML-lo and Lab were least biased; no treatment provided good estimates of any parameter except perhaps the standard deviation.

Acknowledgments

We gratefully thank Dale Peart, David Roth, and Terry Plowman, who assisted in collection of the data, and Kirk Nordstrom, Dixon Landers, Larry Benson, Dennis Helsel, and four anonymous reviewers, who all reviewed the manuscript and offered helpful criticism. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Supporting Information Available

Detailed information about each of the 44 data sets that was summarized in the Results section and information about the specific software commands used to generate the results in this paper and two ancillary topics regarding creation of the control data sets are available. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Baccarelli, A.; Pfeiffer, R.; Consonni, D.; Pesatori, A. C.; Bonzini, M., 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.

(2) 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.

(3) Liu, S.; Lu, J.-C.; Kolpin, D. W.; Meeker, W. Q. Analysis of environmental data with censored observations. *Environ. Sci. Technol.* **1997**, *31* (12), 3358–3362.

(4) Lynn, H. S. Maximum likelihood inference for left-censored HIV RNA data. *Stat. Med.* **2001**, *20* (1), 33–45.

(5) 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.

(6) 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.

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

(8) Kroll, C. N.; Stedinger, J. R. Estimation of moments and quantiles using censored data. *Water Resour. Res.* **1996**, *32* (4), 1005–1012.

(9) Kuttatharmmakul, S.; Massart, D. L.; Coomans, D.; Smeyers-Verbeke, J. Comparison of methods for the estimation of statistical parameters of censored data. *Anal. Chim. Acta* **2001**, *441* (2), 215–229.

(10) Buhamra, S. S. The analysis of VOCs survey data from residences in Kuwait. *Environmetrics* **1998**, *9* (3), 245–253.

(11) Pajek, M.; Kubala-Kukusì, A.; Banasì, D.; Braziewicz, J.; Majewska, U. Random left-censoring: A statistical approach accounting for detection limits in x-ray fluorescence analysis. *X-Ray Spectrom.* **2004**, *33* (4), 306–311.

(12) She, N. Analyzing censored water quality data using a nonparametric approach. *J. Am. Water Resour. Assoc.* **1997**, *33* (3), 615–624.

(13) Singh, A.; Nocerino, J. Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemom. Intell. Lab. Syst.* **2002**, *60* (1–2), 69–86.

(14) Helsel, D. R. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*; John Wiley and Sons Hoboken, NJ, 2005; 250 p.

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

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

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

(18) Porter, P. S.; Ward, R. C. Estimating central tendency from uncensored trace level measurements. *Water Resour. Bull.* **1991**, *27* (4), 687–700.

(19) 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 p.

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

(21) Garbarino, J. R.; Taylor, H. E. An inductively-coupled plasma atomic-emission spectrometric method for routine water quality testing. *Appl. Spectrosc.* **1979**, *33*, 220–225.

(22) Mitko, K.; Bebek, M. ICP-OES determination of trace elements in salinated water. *At. Spectrosc.* **1999**, *20*, 217–223.

(23) Mitko, K.; Bebek, M. Determination of major elements in saline water samples using a dual-view ICP-OES. *At. Spectrosc.* **2000**, *21*, 77–85.

(24) Skogerboe, R. K.; Grant, C. L. Comments on the definition and terms sensitivity and detection limit. *Spectrosc. Lett.* **1970**, *3*, 215–219.

(25) Helsel, D. R. Less than obvious: Statistical treatment of data below the detection limit. *Environ. Sci. Technol.* **1990**, *24*, 1766–1774.

(26) Sharma, M.; Agarwal, R. Maximum likelihood method for parameter estimation in non-linear models with below detection data. *Environ. Ecol. Stat.* **2003**, *10* (4), 445–454.

(27) Helsel, D. R. Insider censoring: Distortion of data with nondetects. *Hum. Ecol. Risk Assess.* **2005**, *11*, 1127–1137.

(28) Persson, T.; Rootzen, H. Simple and highly efficient estimators for a Type I censored normal sample. *Biometrika* **1977**, *64*, 123–128.

(29) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., Ser. B* **1977**, *39*, 1–38.

(30) Gleit, A. Estimation for small normal datasets with detection limits. *Environ. Sci. Technol.* **1985**, *19*, 1206–1213.

(31) Gilbert, R. O. *Statistical methods for environmental pollution monitoring*: Van Nostrand Reinhold: New York, 1987.

(32) Li, K. H.; Raghunathan, T. E.; Rubin, D. B. Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *J. Am. Stat. Assoc.* **1991**, *86* (416), 1065–1073.

(33) Schafer, J. L. *Analysis of Incomplete Multivariate Data*; Chapman and Hall: London, 1997.

(34) Monroe, S. A.; Antweiler, R. C.; Hart, R. J.; Taylor, H. E.; Truini, M.; Rihs, J. R.; Felger, T. J. *Chemical characteristics of ground water discharge along the south rim of Grand Canyon in Grand Canyon National Park, Arizona, 2000–2001.* U.S. Geological Survey Scientific Investigations Report 2004-5146, 2005; 57 p.

(35) Landers, D. H.; Simonich, S. L.; Campbell, D. H.; Erway, M. M.; Geiser, L. H.; Jaffe, D. A.; Kent, M. L.; Schreck, C. B.; Blett, T. F.; Taylor, H. E. *Western Airborne Contaminants Assessment Program Research Plan. EPA/600/R-03/035*; U.S. Environmental Protection Agency, Office of Research and Development, NHEERL, Western Ecology Division: Corvallis, OR, 2003.

(36) Durand, S. R.; Shelley, P. H.; Antweiler, R. C.; Taylor, H. E. Trees, chemistry, and prehistory in the American Southwest. *J. Archaeol. Sci.* **1999**, *26* (2), 185–203.

ES071301C