

Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it

Dennis R. Helsel *

U.S. Geological Survey, P.O. Box 25046, MS 964, Lakewood, CO 80225, USA

Received 3 January 2006; received in revised form 17 April 2006; accepted 18 April 2006

Available online 5 June 2006

Abstract

The most commonly used method in environmental chemistry to deal with values below detection limits is to substitute a fraction of the detection limit for each nondetect. Two decades of research has shown that this fabrication of values produces poor estimates of statistics, and commonly obscures patterns and trends in the data. Papers using substitution may conclude that significant differences, correlations, and regression relationships do not exist, when in fact they do. The reverse may also be true. Fortunately, good alternative methods for dealing with nondetects already exist, and are summarized here with references to original sources. Substituting values for nondetects should be used rarely, and should generally be considered unacceptable in scientific research. There are better ways. Published by Elsevier Ltd.

Keywords: Nondetect; Detection limit; Censored data; Statistics

1. Introduction

In his satire “Hitchhiker’s Guide To The Galaxy”, Douglas Adams wrote of his characters’ search through space to find the answer to “the question of Life, The Universe and Everything”. In what is undoubtedly a commentary on the inability of science to answer such questions, the computer built to process it determines that the answer is 42. There is beauty in a precise answer – a totally arbitrary, but precise, answer.

Environmental scientists often provide a similar answer to a different question – what to do with “nondetect” data? Nondetects are low-level concentrations of organic or inorganic chemicals with values known only to be somewhere between zero and the laboratory’s detection/reporting limits. Measurements are considered too imprecise to report as a single number, so the value is commonly reported as being less than an analytical threshold, for example “<1”. Long considered second class data, nondetects

complicate the familiar computations of descriptive statistics, of testing differences among groups, and of correlation coefficients and regression equations.

The worst practice when dealing with nondetects is to exclude or delete them. This produces a strong upward bias in all subsequent measures of location such as means and medians. After exclusion, comparisons are being made between the mean of the top 20% of concentrations in one group versus the top 50% of another group, for example. This provides little insight into the original data. Excluding nondetects removes the primary signal that should be sent to hypothesis tests – the proportion of data in each group that lies above the reporting limit(s), the shift producing the difference between 20% and 50% detects.

The most common procedure within environmental chemistry to deal with nondetects continues to be substitution of some fraction of the detection limit. This method is better labeled as “fabrication”, as it reports and uses a single value for concentration data where a single value is unknown. Within the field of water chemistry, one-half is the most commonly used fraction, so that 0.5 is used as if it had been measured whenever a <1 (detection limit

* Tel.: +1 303 2365340; fax: +1 303 2361425.

E-mail address: dhelsel@usgs.gov

of 1) occurs. For air chemistry, one over the square root of two, or about 0.7 times the detection limit, is commonly used. Douglas Adams might have chosen 0.42. Studies 20 years ago found substitution to be a poor method for computing descriptive statistics (Gilliom and Helsel, 1986). Subsequent justifications for using one-half the reporting limit when data follow a uniform distribution (Hornung and Reed, 1990) only considered estimation of the mean. Any substitution of a constant fraction of reporting limits will distort estimates of the standard deviation, and therefore all (parametric) hypothesis tests using that statistic. This is illustrated later using simulations. Also, justifications such as these have never considered errors due to changing reporting limits arising from changing interferences between samples or similar causes. Substituting values tied to those changing limits introduces a signal into the data that was not present in the media sampled. Substituted values using a fraction anywhere between 0 and 0.99 times the detection limit are equivalently arbitrary, equivalently precise, equivalently wrong.

Examples of substitution of fractions of the detection limit for nondetects abound in the scientific literature. McCarthy et al. (1997) computed descriptive statistics of organic compounds in relatively uncontaminated areas. They employed substitution of a ‘sliding scale’ fraction of the detection limit, setting the fraction to be a function of the proportion of nondetects in the data set. The accuracy and value of their resulting statistics is unknowable. Another scientist using different fractions to provide values for nondetect data would get different results. Similarly, Tajimi et al. (2005) computed correlation coefficients after substituting one-half the detection limit for all nondetects. They found no correlations between dioxin concentrations and the factors they investigated. Was this because there were none, or was it the result of their data substitutions? Barringer et al. (2005) tested for differences in mercury concentrations of groundwater in areas of differing land use. Were their results due to concentrations actually found in the aquifer, or to the fact that one-half the detection limit was substituted for some nondetects, while other nondetects were simply deleted? Finally, Rocque and Winker (2004) substituted random values between zero and the detection limits in order to compute sums and test hypotheses. How would those results have changed if different random values had been assigned?

Statisticians use the term “censored data” for data sets where specific values for some observations are not quantified, but are known to exceed or to be less than a threshold value. Techniques for computing statistics for censored data have long been employed in medical and industrial studies, where the length of time is measured until an event occurs such as the recurrence of a disease or failure of a manufactured part. For some observations the event may not have occurred by the time the experiment ends. For these, the time is known only to be greater than the experiment’s length, a censored “greater-than” value. Methods for computing descriptive statistics, testing hypotheses,

and performing correlation and regression are all commonly used in medical and industrial statistics, without substituting arbitrary values. These methods go by the names of “survival analysis” and “reliability analysis”. There is no reason why these same methods could not also be used in the environmental sciences, but to date, their use is rare.

Two early examples using methods for censored data in environmental applications are Millard and Deverel (1988) and She (1997). Millard and Deverel (1988) pioneered the use of two-group survival analysis methods in environmental work, testing for differences in metals concentrations in the groundwaters of two aquifers. Many nondetected values were present, at multiple detection limits. They found differences in zinc concentrations between the two aquifers using a survival analysis method called a score test. Had they substituted one-half the detection limit for zinc concentrations and run a *t*-test, they would not have found those differences (Helsel, 2005b). She (1997) computed descriptive statistics of organics concentrations in sediments using a survival analysis method called Kaplan-Meier, the standard procedure in medical statistics. Means, medians and other statistics were computed without substitutions, even though the data contained 20% nondetects censored at eight different detection limits. Substitution would have given very different results. More recently, Baccarelli et al. (2005) reviewed a variety of methods for handling nondetects in a study of dioxin exposure. They found that imputation methods designed for censored data far outperformed substitution of values such as one-half the detection limit. Other examples of the use of survival analysis methods for environmental data can be found in Helsel (2005b).

The goal of this paper is to clearly illustrate the problems with substitution of arbitrary values for nondetects. Methods designed expressly for censored data are directly compared to results using arbitrary substitution of values for nondetects when computing summary statistics, regression equations, and hypothesis tests.

2. Methods

Statisticians generate simulated data for much the same reasons as chemists prepare standard solutions – so that the conditions are exactly known. Statistical methods are then applied to the data, and the similarity of their results to the known, correct values provides a measure of the quality of each method. Fifty *X*, *Y* pairs of data were generated for this study with *X* values uniformly distributed from 0 to 100. The *Y* values were computed from a regression equation with slope = 1.5 and intercept = 120. Noise was then randomly added to each *Y* value so that points did not fall exactly on the straight line. The result is data having a strong linear relation between *Y* and *X* with a moderate amount of noise in comparison to that linear signal.

The noise applied to the data represented a “mixed normal” distribution, two normal distributions where the second had a larger standard deviation than the first. All

of the added noise had a mean of zero, so the expected result over many simulations is still a linear relationship between X and Y with a slope = 1.5 and intercept = 120. Eighty percent of data came from the distribution with the smaller standard deviation, while 20% reflected the second distribution's increased noise level, to generate outliers. The 50 generated values are plotted in Fig. 1A.

The 50 observations were also assigned to one of two groups in such a way that group differences should be discernible. The mean, standard deviation, correlation coefficient, regression slope of Y versus X , a t -test between the means of the two groups and its p -value for the 50 generated observations in Fig. 1A were then all computed and stored. These "benchmark" statistics are the target values to which later estimates are compared. The later estimates are made after censoring the points plotted as gray dots in Fig. 1A.

Two detection limits (at 150 and 300) were then applied to the data, the black dots of Fig. 1A remaining as detected

values with unique numbers, and the gray dots becoming nondetects below one of the two detection limits. In total, 33 of 50 observations, or 66% of observations, were censored below one of the two detection limits. This is within the range of amounts of censoring found in many environmental studies. Use of a smaller percent censoring would produce many of the same effects as found here, though not as obvious or as strong.

All data below the lower detection limit of 150 were changed to <150 . Half of the data between 150 and the higher detection limit of 300 were randomly chosen and recorded as <300 . In order to mimic laboratory results with two detection limits, up to half of the values at the lower limit of <150 were randomly selected and changed to <300 . The result is censored data that can be considered a "best-case scenario" in regard to the censoring process. In practice, data may come from multiple laboratories, and laboratories differ in their protocols for setting

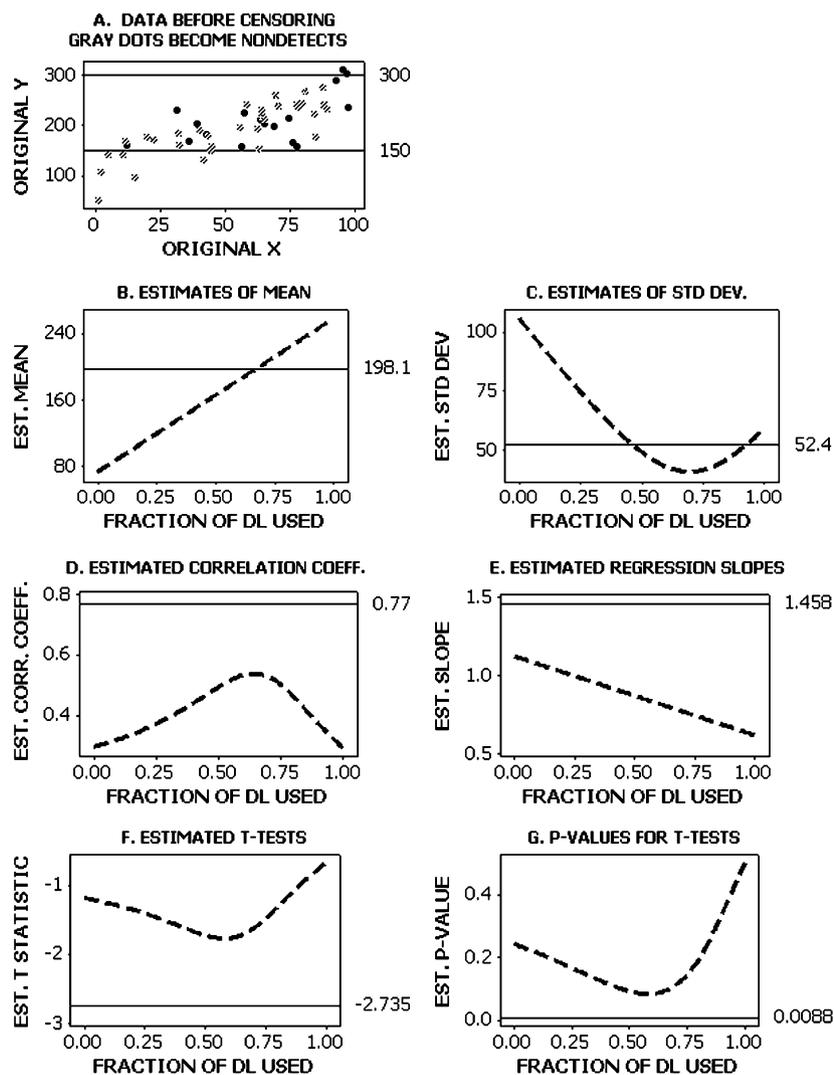


Fig. 1. (A) Generated data used. Horizontal lines are detection limits. (B)–(G) Estimated values for statistics of censored data (Y) as a function of the fraction of the detection limit (X) used to substitute values for each nondetect. Horizontal lines are at target values of each statistic obtained using uncensored data.

detection limits. Within a single lab, detection limits change over time due to changes in methods, protocols, and instrument precision. Interferences by other chemicals cause detection limits for the chemical of interest to change from sample to sample. Laboratories sometimes use biased reporting conventions known as “insider censoring”, in which all values measured as <150 are instead reported as <300 while values measured between 150 and 300 are reported as single numbers (Helsel, 2005a). All of these practices introduce more movement of detection limits than is being considered here. Substituting a fraction of these variable limits for nondetects may introduce patterns into the data that were not there originally, or obscure patterns that were (Helsel, 2005b). Therefore, substituting values as a function of detection limits that change due to these influences is likely to produce less accurate results in practice than those reported here for substitution based on two static detection limits.

3. Results

Fig. 1B–G illustrate the results of estimating a statistic or running a hypothesis test after substituting numbers for nondetects by multiplying the detection limit value by a fraction between 0 and 1. Estimated values for each statistic are plotted on the Y axis, with the fraction of the detection limit used in substitution on the X axis. A fraction of 0.5 on the X axis corresponds to substituting a value of 75 for all <150 s, and 150 for all <300 s, for example. On each plot is also shown the value for that statistic before censoring, as a “benchmark” horizontal line. The same information is presented in tabular form in Table 1.

Estimates of the mean of Y are presented in Fig. 1B. The mean Y before censoring equals 198.1. Afterwards, substitution across the range between 0 and the detection limits (DL) produces a mean Y that can fall anywhere between 72 and 258. For this data set, substituting data using a fraction somewhere around 0.7 DL appears to mimic the uncensored mean. But for another data set with different characteristics, another fraction might be “best”. And 0.7 is not the “best” for these data to duplicate the uncensored standard deviation, as shown in Fig. 1C. Something larger or smaller, closer to 0.5 or 0.9 would work better for that

statistic, for this set of data. Performance will also differ depending on the proportion of data censored, as discussed later. Results for the median (not shown) were similar to those for the mean, and results for the interquartile range (not shown) were similar to those for the standard deviation. The arbitrary nature of the choice of fraction, combined with its large effect on the result, makes the choice of a single fraction an uncomfortable one. As shown later, it is also an unnecessary one.

Substitution results in poor estimates of correlation coefficients (Fig. 1D) and regression slopes (Fig. 1E), much further away from their respective uncensored values than was true for descriptive statistics. The closest match for the correlation coefficient appears to be near 0.7, while for the regression slope, substituting 0 would be best! With data having other characteristics, the “best” fraction will differ. Correlation coefficients, regression slopes, and their p -values should be considered particularly suspect when values are substituted for nondetects, especially if the statistics are found to be insignificant.

The generated data were split into two groups. In the first group were data with X values of 0–40 and 60–70, while the second group contained those with X values from 40 to 60 and then 70 and above. For the most part, values in the first group plotted on the left half of Fig. 1A, and the second group plotted primarily on the right half. Because the slope change is large relative to the noise, mean Y values for the two groups should be seen as different. Before the data were censored, the two-sided t -statistic for the test of whether the mean Y values were equal was -2.74 , with a p -value of 0.009. This is a small p -value, so before censoring the means for the two groups are determined to be different.

Fig. 1F and G and Table 1 report the results of two-group t -tests following substitution of values for nondetects. The t -statistics never reach as large a negative value as for the uncensored data, and the p -values are therefore never as significant. At no time do the p -values go below 0.05, the traditional cutoff for statistical significance. Results of t -tests after using substitution, if found to be insignificant, should not be relied on. Much of the power of the test has been lost, as substitution is a poor method for recovering the information contained in nondetects. Fig. 1F and G show a strong drop-off in performance when the best choice of substituted fraction, which in practice is always unknown, is not chosen.

Clearly, no single fraction of the detection limit, when used as substitution for a nondetect, does an adequate job of reproducing more than one of the statistics in Fig. 1. This study should not be used to pick 0.7 or some other fraction as “best”; different fractions may do a better job for data with different characteristics. The process of substituting a fraction of the detection limits has repeatedly been shown to produce poor results in simulation studies (Gilliom and Helsel, 1986; Singh and Nocerino, 2002; and many others – see Helsel, 2005b for a list). As demonstrated by the long list of research findings and this simple

Table 1
Statistics and test results before and after censoring data as nondetects

Procedure	Before censoring	Range after using substitution	Using censored methods
Estimating mean	198.1	72–258	191.3
Estimating std. dev.	52.4	41–106	54.0
Correlation coeff.	0.77	0.29–0.54	0.55
Regression slope	1.46	0.62–1.12	1.46
t -statistic	-2.74	-1.8 to -0.68	-1.81
p -value for t	0.009	0.08–0.50	0.07

Data in the middle two columns are also shown in Fig. 1. The right column reports the results of tests expressly designed to work with censored data, without requiring substitution for nondetects.

study, substitution of a fraction of the detection limit for nondetects should rarely be considered acceptable in a quantitative analysis. There are better methods available.

When might substitution be acceptable? Research scientists tend to use chemical analyses with relatively high precision and low detection limits. These chemical analyses are often performed by only one laboratory, and detection limits stay fairly constant. Research data sets may include hundreds of data points, and in comparison our 50 observations appears small. For large data sets with a censoring percentage below 60% nondetects, the consequences of substitution should be less severe than those presented here. In contrast, scientists collecting data for regulatory purposes rarely have as many as 50 observations in any one group; sizes near 20 are much more common. Detection limits in monitoring studies can be relatively high compared to ambient levels, so that 60% or greater nondetects is not unusual. Multiple detection limits arise from several common causes, all of which are generally unrelated to concentrations of the analyte(s) of interest. These include using data from multiple laboratories, varying dilutions, and varying sample characteristics such as dissolved solids concentrations or amounts of lipids present. Resulting data like that of She (1997) with eight different detection limits out of 11 nondetects is quite typical. In this situation, the cautions given here must be taken very seriously, and results based on substitution severely scrutinized before publication. Reviewers should suggest that the better methods available from survival analysis be used instead.

Is there a censoring percentage below which the use of substitution can be tolerated? The short answer is “who knows?” The US Environmental Protection Agency (USEPA) has recommended substitution of one-half the detection limit when censoring percentages are below 15% (USEPA, 1998). This appears to be based on opinion rather than any published article. Even in this case, answers obtained with substitution will have more error than those using better methods (see Helsel and Cohn, 1988; She, 1997; and other references in Helsel, 2005b). Will the increase in error with substitution be small enough to be offset by the cost of learning to use better, widely available methods of survival analysis? Answering that question depends on the quality of result needed, but substitution methods should be considered at best “semi-quantitative”, to be used only when approximate answers are required. Their current frequency of use in research publications is certainly excessive, in light of the availability of methods designed expressly for analysis of censored data.

3.1. Statistical methods designed for censored data

Methods designed specifically for handling censored data are standard procedures in medical and industrial studies, and have been applied to the environmental sciences by Helsel (2005b). Results for the current data using one of these methods, maximum likelihood estimation (MLE), are reported in the right-hand column of Table 1.

MLE assumes that data have a particular shape (or distribution), which in Table 1 was a normal distribution, the familiar bell-shaped curve.

The right-hand column of Table 1 shows that methods designed for censored data produce values for each statistic that are as good or better than the best of the estimates produced by substitution. These methods accomplish this without substituting arbitrary values for nondetects. Instead, MLE fits a distribution to the data that matches both the values for detected observations, and the proportion of observations falling below each detection limit. The information contained in nondetects is captured efficiently by the proportion of data falling below each detection limit. The correlation coefficient reported is the “likelihood r ” coefficient, computed by comparing the MLE solutions that best fit the data with and without the X variable. If errors decrease and the fit improves by including the X variable, a significant likelihood correlation coefficient is produced. The traditional Pearson’s r correlation coefficient is the uncensored analogue to the likelihood r coefficient (Allison, 1995).

MLE can be used to compute hypothesis tests between groups of censored data (Helsel, 2005b). Multiple detection limits can be incorporated. No substituted values are used. Instead, likelihood ratio tests determine whether splitting the data into groups provides a better fit than leaving all the data as one group. If so, the test is significant and the means differ among the groups. Results for an MLE version of a two-group test are reported in the t -test row of Table 1. That statistic provides a closer approximation to the uncensored statistic than any of the substitution results. The test statistic is not as significant as was a t -test prior to censoring. The difference between the MLE test results and those prior to censoring is a measure of the loss of information caused by changing numerical values into values known only as less than the detection limits.

Maximum likelihood methods generally do not work well for small data sets (fewer than 30–50 detected values), in which one or two outliers throw off the estimation, or where there is insufficient evidence to know whether the assumed distribution fits the data well (Singh and Nocerino, 2002; Shumway et al., 2002). For these cases, nonparametric methods that do not assume a specific distribution and shape of data would be preferred. See Helsel (2005b) for a full list of nonparametric procedures for censored data, including variations on the familiar rank-sum test and Kendall’s τ correlation coefficient.

3.2. Availability of software for censored data methods

All of the maximum likelihood methods shown here, and equivalent nonparametric tests, are found in the “survival analysis” or “reliability analysis” sections of standard statistics software, including Minitab, SAS, Stata or S-Plus. MLE routines are generally coded to handle nondetects, called “left-censored data” by statisticians, as well as right-censored “greater-thans” more common to the

medical/industrial fields for which they were designed. Nonparametric methods in software are currently coded to use only greater-thans. A transformation called “flipping” (Helsel, 2005b) allows these nonparametric methods to use the nondetects of environmental sciences. None of these methods are available in Excel, although there are add-on packages that compute some of them.

Macros or scripts that either add functionality to software or save steps in setting up these analyses are available through the internet. Macros for Minitab statistical software that were produced to follow the textbook by Helsel (2005b) are available no cost at: <http://www.practical-stats.com/nada>. Scripts (named NADA for R) for the R statistics package, which runs on PCs, Macintosh and Unix computers, are available through the Comprehensive R Archive Network (CRAN) at <http://www.r-project.org/> or at the Enviro-R software page on the Source Forge archive site, <http://enviro.r.sourceforge.net/>. However, though free, R is a complex software package. Those not already familiar with it will find that it takes about the same amount of effort to master it as required by the software for modeling and spatial analysis familiar to environmental scientists.

4. Conclusions

When a fraction of the detection limits is used to substitute (fabricate) values for nondetects, resulting estimates of correlation coefficients, regression slopes, hypothesis tests, and even simple means and standard deviations are inaccurate and irreproducible. They may be very far from their true values, and the amount and type of deviation is unknown. Given the current expense and technological sophistication of sampling equipment and chemical analyses, fabricating values for data at the end of a study is nowhere close to being “state of the science”. It wastes the considerable expense of data collection and salaries by producing inconclusive and potentially incorrect results. Better methods than substitution are available for estimating descriptive statistics, performing hypothesis tests, and computing correlation coefficients and regression equations. Using these methods should provide better, more accurate scientific interpretations.

References

- Allison, P.D., 1995. *Survival Analysis Using the SAS System: A Practical Guide*. SAS Institute, Inc., Cary, NC.
- Baccarelli, A., Pfeiffer, R., Consonni, D., et al., 2005. 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* 60, 898–906.
- Barringer, J., Szabo, Z., Kauffman, L., Barringer, T., Stackelberg, P., Ivahnenko, T., Rajagopalan, S., Krabbenhoft, D., 2005. Mercury concentrations in water from an unconfined aquifer system, New Jersey coastal plain. *Sci. Total Environ.* 346, 169–183.
- Gilliom, R.J., Helsel, D.R., 1986. Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques. *Water Resour. Res.* 22, 135–146.
- Helsel, D., 2005a. Insider censoring: distortion of data with nondetects. *Hum. Ecol. Risk Assess.* 11, 1127–1137.
- Helsel, D., 2005b. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. John Wiley, New York.
- Helsel, D.R., Cohn, T., 1988. Estimation of descriptive statistics for multiply censored water quality data. *Water Resour. Res.* 24, 1997–2004.
- Hornung, R.W., Reed, L.D., 1990. Estimation of average concentration in the presence of nondetectable values. *Appl. Occup. Environ. Hyg.* 5, 46–51.
- McCarthy, L., Stephens, G., Whittle, D., Peddle, J., Harbicht, S., La Fontaine, C., Gregor, D., 1997. Baseline studies in the Slave River, NWT, 1990–1994; Part II. Body burden contaminants in whole fish tissue and livers. *Sci. Total Environ.* 197, 55–86.
- Millard, S.P., Deverel, S.J., 1988. Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits. *Water Resour. Res.* 24, 2087–2098.
- Rocque, D.A., Winker, K., 2004. Biomonitoring of contaminants in birds from two trophic levels in the North Pacific. *Environ. Toxicol. Chem.* 23, 759–766.
- She, N., 1997. Analyzing censored water quality data using a non-parametric approach. *J. Am. Water Resour. Assoc.* 33, 615–624.
- Shumway, R.H., Azari, R.S., Kayhanian, M., 2002. Statistical approaches to estimating mean water quality concentrations with detection limits. *Environ. Sci. Technol.* 36, 3345–3353.
- Singh, A., Nocerino, J., 2002. Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometr. Intell. Lab.* 60, 69–86.
- Tajimi, M., Uehara, R., Watanabe, M., Oki, I., Ojima, T., Nakamura, Y., 2005. Correlation coefficients between the dioxin levels in mother’s milk and the distances to the nearest waste incinerator which was the largest source of dioxins from each mother’s place of residence in Tokyo, Japan. *Chemosphere* 61, 1256–1262.
- USEPA, Office of Research and Development, 1998. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis*. EPA/600/R-96/084, Washington, DC.