Journal of Environmental Statistics

December 2011, Volume 2, Issue 2. http://www.jenvstat.org

Estimation of Multiple Trace Metal Water Contaminants In the Presence of Left-Censored and Missing Data

Heather J. Hoffman
The George Washington University

Robert E. Johnson
Virginia Commonwealth University

Abstract

Environmental data often include left-censored values reported to be less than some limit of detection (LOD). While simple imputation of a specific value such as LOD/2 is commonly implemented in practice, maximum likelihood methods accounting for censoring provide an alternate way of analyzing such data. Concentration levels of trace metal contaminants in water are typically modeled with normal or lognormal distributions. The corresponding maximum likelihood estimates (MLEs) of means and variances in univariate analyses can be obtained from standard software packages; however, a multivariate analysis may be more appropriate when multiple measurements are taken from the same entity. For example, the overall contamination level of freshwater streams may be represented by a linear combination of several dissolved trace metal amounts present within. Especially in less polluted areas, one or more of these levels may fall below the LOD. We propose a multivariate method that provides MLEs of mean and unstructured covariance parameters corresponding to a multivariate normal or lognormal distribution in the presence of left-censored and missing values. In conducting hypothesis tests and estimating functions of MLEs with appropriate standard errors, we apply this multivariate method to trace metal concentration data collected from freshwater streams across the Commonwealth of Virginia.

Keywords: trace metal concentrations, water quality, limit of detection, maximum likelihood estimation, multivariate normal distribution.

1. Introduction

Complications arise in the analysis of environmental samples containing potentially hazardous chemicals due to the presence of various pollutants at trace levels that cannot reliably be dis-

tinguished from 0 and, therefore, are reported as results that lie numerically below a certain limit of detection (LOD). The vast array of univariate statistical methods presented in the literature for estimating measures of centrality and variability in the presence of these nondetects can be partitioned into three distinct categories: substitution, parametric and nonparametric methods. Lacking in the literature, however, are extensions of these univariate estimation techniques to multivariate estimation, which are needed, for instance, in chemical monitoring problems. See, for example, Gilliom and Helsel (1986), Sanford *et al.* (1993), and Farnham *et al.* (2002). With such multivariate data, estimation is further complicated by the presence of multiple censoring. Multivariate extensions of univariate estimation techniques that account for nondetects are required. We present a multivariate maximum likelihood method for estimating parameters of multivariate normal and lognormal models that appropriately accounts for the proportion of data falling below the LOD.

Multivariate assay data typically appear in environmental applications. We may examine the effects of environmental toxins on the health of a population by measuring the cumulative amount of trace metals in streams with industrial outfalls or particulates in the atmosphere of a specific site, such as a national park. Missing data in this context can be categorized into two groups: completely missing and nondetected. Whereas some data may be completely missing as a result of a flawed assay or data entry errors, other data may be reported only as falling below a specified LOD. Our primary focus is estimation of parameters of multivariate assay distributions, such as means and variances, and functions of these parameters in the presence of missing data.

2. Environmental Application

The concentration levels of certain dissolved trace metals in freshwater streams across the Commonwealth of Virginia are compared to the worldwide standard using a well-defined index function and our proposed multivariate method. The Virginia Department of Environmental Quality (VDEQ) provided the data used in this application, which can be found in the supplemental materials (VDEQ_Data).

It is of particular interest to determine the quality of Virginia's water resources throughout the state to discover their true usefulness (VDEQ 2003, p. 1). The methodology used to assess water quality should not underestimate contamination levels thereby compromising public health and the environment. Neither should it overestimate contamination levels so that local industry is unfairly restricted.

With an estimated 51,021 miles of perennial rivers and streams, Virginia has a total combined flow of approximately 25 billion gallons of freshwater per day (VDEQ 2008, p. 6.1-1). By the Federal Clean Water Act and the Virginia Water Quality Monitoring Information and Restoration Act, the state is required to assess and report on the quality of these state waters (VDEQ 2008, p. 1.1-1). In addition to the EPA regulations, each state implements environmental water quality standards that permit water body usages—including drinking, swimming, farming, fish production, or industrial processes—and quantify the safety associated with such uses by measuring acceptable levels of the pollutants present within (US EPA 1997, p. 49). Of the more than 15,951 miles of rivers and streams studied in 2008, approximately 34 percent have high water quality that satisfies such designated uses, while the other 66 percent do not fully support designated uses (VDEQ 2008, p. 1.1-4). Past

studies reveal that unsatisfactory water quality results primarily from disregarding the E. coli bacteria standards—largely from agricultural practices, urban runoff, leaking sanitary sewers, failing septic tanks, domestic animals, and wildlife (VDEQ 2008, p. 1.1-4).

As part of the VDEQ's ProbMon program, freshwater samples are collected and submitted to the VA Division of Consolidated Laboratory Services in Richmond, VA for analysis (VDEQ 2003, p. 29). For the application at hand, we use a subset of these data to determine whether the collective concentration levels of the dissolved trace metals copper (Cu), lead (Pb) and zinc (Zn) in freshwater streams across Virginia are significantly different from the reported worldwide standard, while correcting for the water hardness as a function of calcium (Ca) and magnesium (Mg).

Specifically, we estimate an index function of the concentrations of the trace metals Cu, Pb and Zn that represents a ratio or difference in the mean contamination level of the given independent freshwater streams in Virginia relative to the worldwide standard. Since the water quality standard for each trace metal is a function of Ca and Mg, we view the problem as follows.

Denote the respective concentrations of the five metals Cu, Pb, Zn, Ca, and Mg by $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$, and assume that $\mathbf{X} \sim LN_5(\boldsymbol{\eta}, \mathbf{T})$—multivariate lognormal—with correlation matrix $\mathbf{G}$, where $\boldsymbol{\eta}$ and $\mathbf{T}$ denote the (lognormal) mean vector and covariance matrix, respectively. In order to calculate a measure of centrality, ideally we would like to normalize the variables of interest. One approach is to divide the individual metal concentrations by their respective water quality standard, which is a function of the corresponding hardness factor. To calculate hardness (mg/L $CaCO_3$) from Ca and Mg, we use the equation

$$h = H(X_4, X_5) = 2.497X_4 + 4.118X_5 \tag{1}$$

with units reported in mg/L (Ledo de Medina 2000, p. 58). The respective water quality standards for Cu, Pb and Zn, denoted as f_1, f_2, and f_3, respectively, are then given by

$$\begin{aligned} f_1(X_4, X_5) &= h^{0.8545} e^{-1.702} \\ f_2(X_4, X_5) &= h^{1.273} e^{-3.259} \\ f_3(X_4, X_5) &= h^{0.8473} e^{0.884} \end{aligned} \tag{2}$$

with units reported in μg/L (VDEQ 2009, p. 20, 25, 30). Note that the hardness factor must fall within the closed interval [25 mg/L, 400 mg/L] in order to use these equations.

For this application, the geometric mean—which approximates the median of the lognormal distribution—is a preferable measure of centrality as compared to the arithmetic mean. So, the index of interest is

$$g(\mathbf{X}) = \left[\prod_{i=1}^{3} \frac{X_i}{f_i(X_4, X_5)} \right]^{\frac{1}{3}}. \tag{3}$$

The logarithm of the index in (3) is given by

$$\ln[\, g(\mathbf{X}) \,] = \frac{1}{3} \sum_{i=1}^{3} \ln(X_i) - \frac{1}{3} \sum_{i=1}^{3} \ln[\, f_i(X_4, X_5) \,]. \tag{4}$$

Note that the formulation in (4) can be interpreted as the average difference of the collective log-transformed concentration levels of Cu, Pb and Zn in the freshwater streams of Virginia relative to the worldwide standard, whereas (3) designates an *average* ratio of the two on the observed scale. A 95% confidence interval about the mean of (4) is thus expected to include 0 if no difference in the trace metal concentrations exists between the freshwater streams of Virginia and the worldwide standard, whereas a 95% confidence interval about the mean of (3) indicating no difference would include 1. In order to construct such confidence intervals, we estimate the expected value and standard error of the index function $g(\mathbf{X})$ given in (3).

The given data consist of the concentration levels of these five dissolved trace metals from 184 independent randomly selected sites in freshwater streams across Virginia. Observe that Cu, Pb and Zn concentrations are conveniently reported in μg/L of water, whereas Ca and Mg concentration are suitably reported in mg/L of water, which happen to be the correct units required for the index of interest. Notably, the freshwater sites are selected with weightings to give approximately equal numbers of sites in each of five sampling strata determined by the Strahler Order, a size standard used by the U.S. EPA. VDEQ's probabilistic monitoring uses a random tessellation stratified survey design to select stream sample sites (VDEQ 2003, p. 13-16). Since the measurements are taken at different times, the presence of multiple LOD values is possible for each trace metal.

Using these data, it is of interest to estimate measures of central tendency and of variability for each trace metal and estimate the expected value and standard error of the index function that measures the overall contamination level in freshwater streams throughout Virginia. Since a number of the recorded levels fall below the LOD assigned to each metal, our proposed method is utilized to estimate the summary statistics of interest. Imputation—of one-half of the LOD—is employed for comparison.

3. Statistical Approach

3.1. Background

As seen in the environmental science literature, the two most commonly applied methods employed in the presence of left-censored data are also the simplest: deletion and substitution. Especially in biogeochemical studies, it is common practice to interpret the data using traditional statistical tests after substituting one-half of the LOD for nondetected values. Deleting—ignoring—the nondetects overestimates the true mean levels and underestimates variation while ignoring some information. Simulation studies by Singh and Nocerino (2002) reveal that substituting one-half of the LOD for nondetects yields a biased estimate of the mean with the highest variability in comparison to other methods used to calculate summary statistics for censored data. Rather than using a simplistic technique that only accounts for information in the observed data and in the value of the LOD, one should conduct analyses that additionally use information contained in the proportion of data that fall below the LOD.

Estimation of summary statistics for censored data is among the most widely researched topics in the literature where nondetected data appear. While the majority of the publications reference the same complication brought about by the presence of nondetects, they present differing opinions as to which method yields the best results. See, for example, Helsel (1990), Helsel (2005), Travis and Land (1990), Rao *et al.* (1991), and Singh and Nocerino (2002).

The predominant methods currently employed to deal with nondetects include:

1. simple substitution methods, which entail replacing censored observations by a constant and treating them as observed,
2. distributional or parametric methods, which involve fitting a given statistical distribution to the data with left-censoring and calculating the summary statistics from the estimated distribution via MLE for censored samples, and
3. nonparametric methods, which assign ranks to the measure—typically assigning the smallest ranks to the censored values.

It is more advisable to account for left-censored observations by correcting the estimation method as opposed to correcting the sample. A vast amount of research presented in the literature recommends applying parametric methods— specifically MLE—adapted for left-censored observations. See, for example, Helsel (1990), Koo *et al.* (2002), and Zhao and Frey (2006). With this approach, we estimate the summary statistics of interest based on the characteristics of the assumed distribution.

Over the years, numerous authors such as Afifi and Elashoff (1967), Hocking and Smith (1968), and Morrison (1971) have derived MLEs of parameters of multivariate normal distributions in the presence of missing observations; however, none of these approaches were directly applied to left-censored data, which is required in the presence of nondetects. In this paper, we fill this gap with methods based on the fully parametric MLE associated with the multivariate normal and lognormal distributions.

3.2. Methods

We begin with a discussion on the construction of the multivariate normal distribution likelihood function, which is the backbone of our multivariate MLE tool. Noting that some of the data are left-censored, the most complex component of the likelihood is the cumulative distribution function (CDF). We use the delta method to estimate standard errors of the MLEs as well as means and variances of composite functions. See Casella and Berger (2001, sec. 5.5.4) for details.

Defining the likelihood function

In order to extend the univariate MLE approach to a multivariate setting, an MLE tool is developed that constructs the appropriate log-likelihood function accounting for left-censored values and maximizes this function using the Newton-Raphson optimization method. In constructing the tool, we give considerable attention to two important factors: starting values and efficiency issues.

Suppose we have a set of data consisting of p variables measured on each of n independent subjects, where $p < n$. Let $\mathbf{Y_1}, \ldots, \mathbf{Y_n}$ denote $p \times 1$ random vectors, i.e. $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{ip})'$ for $i = 1, \cdots, n$, belonging to a multivariate distribution with probability density function (PDF) $f(\cdot)$ and CDF $F(\cdot)$. For the i^{th} subject, we arrange the variables so that the first r_i are observed and the next m_i are left-censored. The remaining variables, if any, have missing values—assumed to be missing completely at random—and, thus, contribute nothing to the likelihood function. Let $s_i = r_i + m_i \leq p$. To simplify notation, let $Y_i^{(1)}, \ldots, Y_i^{(r_i)}$ represent the

r_i observed variables and $Y_i^{(r_i+1)}, \ldots, Y_i^{(s_i)}$ represent the m_i left-censored values corresponding to the i^{th} subject. Then the contribution of the i^{th} subject to the likelihood function is given by

$$L_i(\boldsymbol{\theta}) = F\left(LOD_i^{(r_i+1)}, \ldots, LOD_i^{(s_i)} \;\middle|\; y_i^{(1)}, ..., y_i^{(r_i)}\right) f\left(y_i^{(1)}, ..., y_i^{(r_i)}\right), \tag{5}$$

where$F(\cdot\,|\cdot)$ is the conditional CDF, and $LOD_i^{(j)}$ is the LOD corresponding to the j^{th}variable of the i^{th} subject—which accommodates different LODs on the same measure that could result from different labs or LOD change after equipment calibration. For n independent subjects, the full likelihood function is simply the product of the individual likelihoods for each subject provided in (5), written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} F\left(LOD_i^{(r_i+1)}, \ldots, LOD_i^{(s_i)} \;\middle|\; y_i^{(1)}, ..., y_i^{(r_i)}\right) f\left(y_i^{(1)}, ..., y_i^{(r_i)}\right). \tag{6}$$

The log-likelihood function, which leads to more tractable computations, is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \ln\left[F\left(LOD_i^{(r_i+1)}, \ldots, LOD_i^{(s_i)} \;\middle|\; y_i^{(1)}, ..., y_i^{(r_i)}\right) \right] + \ln\left[f\left(y_i^{(1)}, ..., y_i^{(r_i)}\right) \right] \right\} \tag{7}$$

Transforming the cumulative distribution function

When data from one sample source contain some censored measures, its contribution to the likelihood function is through a multivariate normal CDF. In a univariate dimension, several authors have published relatively good approximations to the normal CDF, such as the trapezoidal or Simpson's rules, and callable functions are readily available in standard statistical software. The reader is referred to Bagby (1995), Shore (2004), Shore (2005), and Bowling *et al.* (2009) for further details. Extensions of such methods to multivariate dimensions, however, lack resolution. Existing algorithms are increasingly computationally intense as the number of censored variables increases.

Our approach combines a set of transformations on the multivariate normal CDF devised by Genz (1992) with numerical quadrature. Genz's transformations simplify the multiple integration problem by transforming infinite integrals to integrals over a unit hyper-cube.

Suppose a single observation is censored on each of m variables so that construction of the likelihood requires computation of an m-variate normal CDF. Define the random vector $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_m})' \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We begin with the multivariate normal CDF given by

$$F(\mathbf{a}) = \frac{1}{(2\pi)^{m/2}\,|\boldsymbol{\Sigma}|^{1/2}} \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_m} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\} dy, \tag{8}$$

where $\mathbf{a} = (a_1, \ldots, a_m)'$ is the vector of upper limits of integration, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ is the mean vector, and $\boldsymbol{\Sigma}$ is the $m \times m$ symmetric positive definite variance-covariance matrix.

After conducting a series of tests involving integral transformations, Genz concludes that accurate multivariate normal probabilities can be computed relatively quickly for applications having up to ten variables (Genz 1992, p. 148). With this in mind, Genz proposes a series of transformations to convert the infinite integral to an integral over the unit hyper-cube. More details are provided in the supplemental materials (GenzDetails).

Computing the likelihood function

Using his transformations, Genz (1992) applies the Monte-Carlo method to approximately evaluate the CDF. Necessity requires us to take a different approach. We employ a Newton-Raphson maximization method which requires multiple computations of the multivariate CDF over small regions. For the method to converge, approximations of the CDF over these regions should constitute a smooth surface. The Monte-Carlo method approximations correspond to spikes over the region due to random sampling error. Numerical quadrature provides a smooth surface of estimates. We selected the Legendre-Gauss quadrature rule as the numerical integration method to approximately evaluate the CDF. Specifically, we chose the 8-point Legendre-Gauss quadrature rule as a reasonable compromise between accuracy and efficiency.

Extending the quadrature method to a multiple integral of m dimensions is relatively straightforward. Using Genz's transformed version of the multivariate normal CDF and the formulation of the n-point Legendre-Gauss quadrature rule presented by Hildebrand (1956, sec. 8.5), the approximation to the integral is given by

$$F(\mathbf{b}) = e_1 \int_0^1 e_2(u_1) \int_0^1 e_3(u_1, u_2) \cdots \int_0^1 e_m(u_1, u_2, \ldots, u_{m-1}) du_{m-1} \cdots du_2 du_1 \qquad (9)$$

where $\tilde{w}_i = \frac{1}{2} w_i$and $\tilde{a}_i = \frac{1}{2}(1 + a_i)$, and $\{a_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$ are the abscissas and weights, defined over $[-1,\ 1]$, for the n-point Legendre-Gauss rule.

Computations were performed using SAS/IML® software. Custom modules were created for Legendre-Gauss quadrature and for computation of the likelihood function. The SAS/IML module NLPNRA was used to perform the Newton-Raphson procedure. All programs written with SAS software are provided as supplemental materials (Part1-MACROS, Part2-MAIN). A user's manual is also provided (User_Manual).

Evaluating the method

We evaluated the precision and accuracy of our method in producing estimates of multivariate normal and lognormal parameters using bias and mean squared error (MSE). We simulated multivariate normal data—and lognormal data by transformation—for evaluation and comparison with the imputation method.

The performance of the proposed multivariate MLE method is analyzed using six groups (I-VI) of 1000 simulated data sets, each of sample size 25, consisting of four-dimensional multivariate normal values. The first three groups I, II and III are assigned correlations of 0.3 between each pair of variables and respective censoring percents of 0%, 10% and 25% for three of the variables. The remaining three groups IV, V and VI are assigned correlations of 0.7 between each pair of variables and respective censoring percents of 0%, 10% and 25% for three of the variables. The fourth variable is assumed to have no censored values. Without loss of generality, we assign a mean of 0 and a variance of 1 for each of the variables.

In order to generate a data set with, say, 10% of the values censored for a variable, we find the 10th quantile of the standard normal distribution, $d = \Phi^{-1}(0.10)$. We consider any value within the simulated data set falling below d a left-censored value, censored at d.

We simulated standard normal data but can consider nonstandard values by the simple transformation $y = \sigma x + \mu$. Notice that changing the value of the normal means from 0 to μ would not change the relative bias or relative mean square error. We consider only 0 as the mean

value—in the normal scale—from here on. We may consider scenarios with arbitrary σ^2 by multiplying each simulated value by σ. This tactic varies the LOD to accommodate the change in variance so that the percent censoring remains constant. The LOD is thus σd in the standard normal scale and $(\exp(d))^{\sigma}$in the standard lognormal scale.

In addition to the six scenarios I-VI described above, we consider three values of the coefficient of variation in the lognormal scale, $\text{CV} = \exp(\sigma^2) - 1$: 0.3, 0.5, and 0.7. This is equivalent to normal-scale variances of $\sigma^2 = 0.086,\ 0.223,$ and 0.399, respectively.

The imputation method entails imputing one-half of the LOD for the censored observations and then computing the means and unbiased variances of each of the variables individually, as well as the Pearson correlations between the variables. Specifically, we begin with lognormal data and convert to normal data by log transformation. Then we impute by replacing the normal values falling below the LOD with $\ln(LOD) - \ln(2)$, which is analogous to replacing the original lognormal values below LOD with $LOD/2$. Note that when the percentage of values below the LOD remains constant, the LOD value shifts to the right as the variance decreases. Notably, it is this increase in the LOD value that increases the bias of the imputation method.

The multivariate MLE method requires no imputation; rather, the censoring is accounted for in construction of the likelihood function. We derive the MLEs of the lognormal parameters from the multivariate normal parameters using the invariance property.

Evaluation results

The complete set of the method evaluation results is provided as supplemental material (MethEvalRslts). We present here in Table 1 a representative subset that is typical of the complete set of results. Specifically, in Table 1 we provide the ratio of imputation MSE to multivariate MLE MSE for the mean, variance and correlation estimates obtained from the set of simulated data with assigned variances of 0.223 and correlations of 0.7 for 0%, 10% and 25% censoring—groups (IV), (V), and (VI). For each of the parameters, notice that the ratios are all greater than 1 with the exception of the correlations, meaning that the imputation method had higher MSE than the multivariate MLE method overall. Since we are not performing any imputation with 0% censoring, the departure from 1 seen with the normal variance parameters is simply due to the fact that the imputation method yields an unbiased estimate of the variance whereas the MLE method yields a biased estimate of the variance, but with smaller MSE. Moreover, observe how the ratio increases as the percentage of censoring increases.

4. Environmental Application: Results

For each trace metal, a summary of the number of streams with levels observed, below the corresponding LOD value, or missing is provided in Table 2. Note the high percentage of nondetected values for Pb and Zn in comparison to Cu, Ca and Mg. Even more surprisingly, notice that none of the data is completely missing!

A breakdown of the frequency of nondetects within each stream is provided in the footnote of Table 2. While the majority of the observations are censored on two or fewer metals, four streams do have four or five levels nondetected, requiring the multivariate MLE tool to approximate four- and five-variate normal CDFs.

Using the original lognormal data as given and the log-transformed normal data, summary

Table 1: Relative MSE of Estimators: Imputation to Mulitvariate MLE

Normal				**Lognormal**			
	% Censored				% Censored		
Parameter	**0**	**10**	**25**	**Parameter**	**0**	**10**	**25**
μ_1	1.00	1.61	2.63	η_1	1.02	1.10	1.41
σ_1^2	1.05	4.02	4.46	τ_1^2	1.68	1.33	1.41
μ_2	1.00	1.57	2.66	η_2	1.02	1.09	1.41
σ_2^2	1.08	4.08	4.64	τ_2^2	1.73	1.61	1.36
μ_3	1.00	1.64	2.67	η_3	1.01	1.11	1.43
σ_3^2	1.05	4.23	3.91	τ_3^2	1.75	1.26	1.33
μ_4	1.00	1.00	1.00	η_4	1.02	1.00	1.00
σ_4^2	1.08	1.05	1.08	τ_4^2	1.69	1.41	1.37
ρ_{12}	1.00	1.25	1.42	γ_{12}	1.36	1.21	1.10
ρ_{13}	1.00	1.16	1.00	γ_{13}	1.40	1.16	0.76
ρ_{14}	1.00	1.07	1.11	γ_{14}	1.41	1.28	1.25
ρ_{23}	1.00	1.17	1.07	γ_{23}	1.40	1.17	0.80
ρ_{24}	1.00	1.08	1.13	γ_{24}	1.39	1.33	1.21
ρ_{34}	1.00	1.04	0.91	γ_{34}	1.47	1.22	0.95

In the simulation, the variances were set to 0.223 (equivalent to a lognormal CV of 0.5) and the correlations were set to 0.7. Relative MSE values greater than 1 indicate that the imputation method has higher MSE than the multivariate MLE method.

Table 2: Frequency of Observed and Nondetected Values for Each Trace Metal

Metal	**N**	**# Observed (%)**	**LOD**	**# Nondetected (%)**
Cu	184	175 (95.1)	0.1 μg/L	9 (4.9)
Pb	184	40 (21.7)	0.1 μg/L	144 (78.3)
Zn	184	113 (61.4)	1.0 μg/L	71 (38.6)
Ca	184	179 (97.3)	0.5 mg/L	1 (0.5)
			1.0 mg/L	4 (2.2)
Mg	184	166 (90.2)	0.5 mg/L	2 (1.1)
			1.0 mg/L	16 (8.7)

This dataset contained no missing values. Note that 33 (17.9%) of the streams had 0 nondetected trace metals, 72 (39.1%) had 1, 68 (37.0%) had 2, 7 (3.8%) had 3, 2 (1.1%) had 4, and 2 (1.1%) had 5 nondetected trace metals.

Table 3: Summary Statistics of the Raw (and Log-transformed) Data

Metal	**Cu (μg/L)**	**Pb (μg/L)**	**Zn (μg/L)**	**Ca (mg/L)**	**Mg (mg/L)**
Total N	184	184	184	184	184
Observed N	175	40	113	179	166
Mean	0.58	0.27	3.45	12.41	4.18
	(-0.79)	(-1.51)	(0.94)	(1.99)	(1.04)
Variance	0.28	0.04	14.23	199.51	21.90
	(0.43)	(0.34)	(0.47)	(1.03)	(0.67)
95% CI	0.50, 0.66	0.20, 0.33	2.74, 4.15	10.33, 14.49	3.46, 4.89
	(-0.88, -0.69)	(-1.69, -1.32)	(0.81, 1.07)	(1.84, 2.14))	(0.92, 1.17)
Minimum	–	–	–	–	–
	(–)	(–)	(–)	(–)	(–)
Lower Quartile	0.27	–	–	3.40	1.40
	(-1.31)	(–)	(–)	(1.22)	(0.34)
Median	0.40	–	1.35	5.45	2.00
	(-0.92)	(–)	(0.30)	(1.70)	(0.69)
Upper Quartile	0.66	–	2.52	14.35	4.55
	(-0.42)	(–)	(0.92)	(2.66)	(1.52)
Maximum	4.00	1.01	29.00	64.90	35.00
	(1.39)	(0.01)	(3.37)	(4.17)	(3.56)

All the observed and nondetected values–Total N–were used to obtain the five-number summary statistics. Only the observed values–Observed N–were used to caculate means, variances, and 95% CIs.

statistics of only the observed subset of the data are computed and reported in Table 3, including the arithmetic means, variances and 95% confidence intervals about the means.

Additionally, the typical five-number-summary—minimum, lower quartile, median, upper quartile, and maximum—of the entire data set—observed and nondetected—is provided in Table 3 where applicable; e.g., since more than 75% of the Pb data is nondetected, we can accurately report the maximum only. Non-applicable cells are indicated with a dash (–).

We used probability plots for both the original concentration values and the log-transformed concentration values to assess the normality of the data. The probability plots corresponding to Cu and Zn are provided in Figure 1. Notice how the log transformation normalizes the data. Seeing as the given data for each metal are positively skewed, a log-transformation is deemed appropriate to normalize the data by making it more symmetrical and homoscedastic. Here we focus on estimating the parameters of the normal and lognormal distributions.

Assuming that the log-transformed data follow the multivariate normal distribution, we apply the multivariate MLE tool to obtain the MLEs of the normal means, variances and correla-

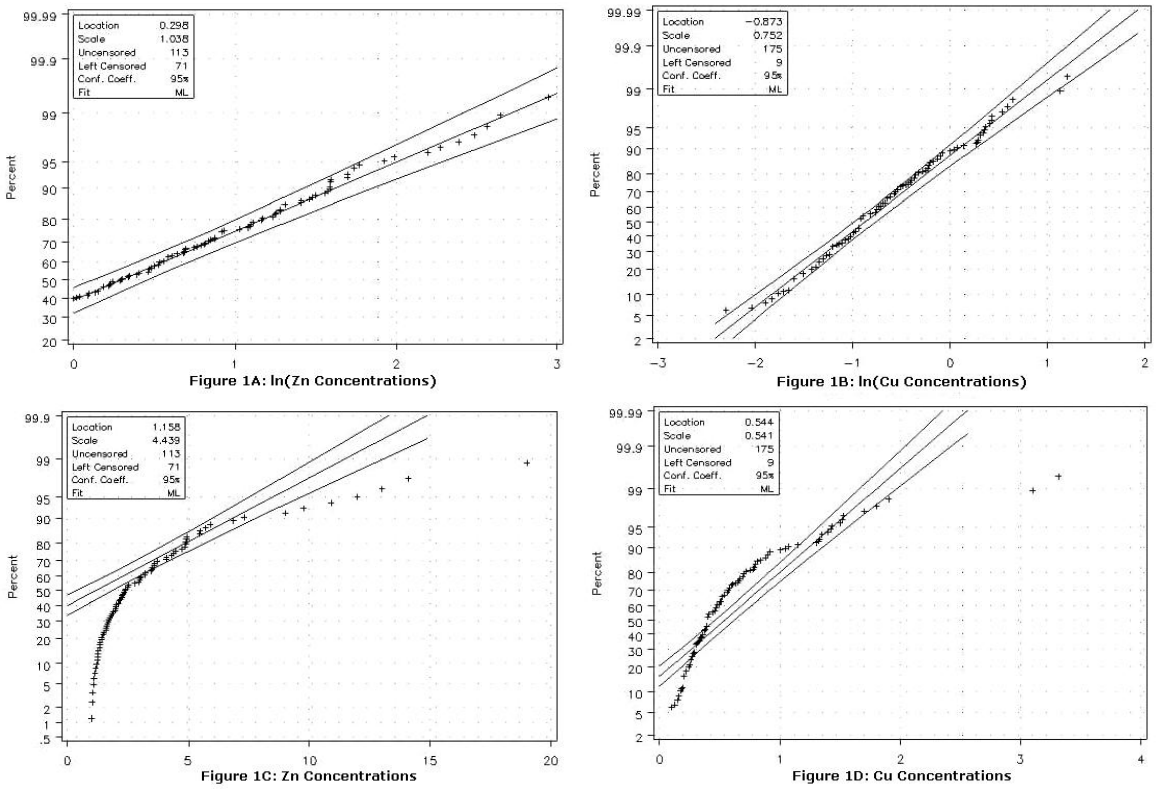

Figure 1: Censored Normal Probability Plots for Zinc (Zn) and Copper (Cu) Concentrations: A. Log-Transformed Zn Concentrations, B. Log-Transformed Cu Concentrations, C. Zn Concentrations, and D. Cu Concentrations

tions. The results are summarized below.

For all of the trace metals, the normal mean, variance and/or correlation parameter estimates from the imputation and multivariate MLE methods are included in Table 4, along with corresponding standard errors. Note the order of the trace metals is given by $Y_1 = \ln(X_1) =$ Cu, $Y_2 = \ln(X_2) =$ Pb, $Y_3 = \ln(X_3) =$ Zn, $Y_4 = \ln(X_4) =$ Ca, and $Y_5 = \ln(X_5) =$ Mg. Utilizing the invariance property of MLEs, we obtain the corresponding lognormal estimates with appropriate standard errors.

Using the methodology introduced in section 2, we estimate the index function (3) using the two sets of parameter estimates for comparative purposes. The results are summarized in Table 5. Since the value of $g(\mathbf{Y})$ must be positive, the confidence intervals about $g(\mathbf{Y})$ should be bounded below by 0. We report the negative values here, however, for illustrative purposes.

5. Discussion

Importantly, we emphasize a major advantage that the multivariate approach has over the imputation approach at this point. In considering a univariate approach, realize that any observation with a completely missing value is thrown out of the analysis. In contrast, a

Table 4: Normal and Lognormal Parameter Estimates (with Standard Errors)

Parameter	**Impute (S.E.)**	**Multi MLE (S.E.)**	**Parameter**	**Impute (S.E.)**	**Multi MLE (S.E.)**
μ_1	-0.893 (0.065)	-0.873 (0.056)	η_1	0.556 (0.038)	0.554 (0.035)
σ_1^2	0.642 (0.082)	0.566 (0.061)	τ_1^2	0.276 (0.083)	0.233 (0.054)
μ_2	-2.672 (0.047)	-3.274 (0.195)	η_2	0.097 (0.009)	0.085 (0.011)
σ_2^2	0.450 (0.072)	1.619 (0.385)	τ_2^2	0.017 (0.006)	0.029 (0.018)
μ_3	0.311 (0.067)	0.280 (0.088)	η_3	2.309 (0.223)	2.322 (0.232)
σ_3^2	0.928 (0.082)	1.125 (0.166)	τ_3^2	10.776 (3.933)	11.213 (4.404)
μ_4	1.911 (0.080)	1.914 (0.081)	η_4	12.083 (1.088)	12.301 (1.258)
σ_4^2	1.210 (0.131)	1.191 (0.129)	τ_4^2	197.855 (32.621)	346.592 (121.397)
μ_5	0.863 (0.070)	0.874 (0.070)	η_5	3.813 (0.353)	3.729 (0.313)
σ_5^2	0.896 (0.101)	0.884 (0.099)	τ_5^2	20.966 (6.247)	19.749 (5.865)
ρ_{12}	0.376 (0.063)	0.471 (0.072)	γ_{12}	0.272 (0.063)	0.325 (0.064)
ρ_{13}	0.351 (0.071)	0.385 (0.067)	γ_{13}	0.196 (0.070)	0.286 (0.055)
ρ_{14}	0.267 (0.086)	0.246 (0.070)	γ_{14}	0.141 (0.094)	0.170 (0.052)
ρ_{15}	0.325 (0.068)	0.324 (0.068)	γ_{15}	0.106 (0.066)	0.248 (0.056)
ρ_{23}	0.492 (0.064)	0.541 (0.079)	γ_{23}	0.578 (0.133)	0.371 (0.079)
ρ_{24}	-0.249 (0.074)	-0.392 (0.086)	γ_{24}	-0.179 (0.032)	-0.138 (0.027)
ρ_{25}	-0.254 (0.064)	-0.418 (0.090)	γ_{25}	-0.171 (0.031)	-0.164 (0.033)
ρ_{34}	-0.130 (0.082)	-0.139 (0.077)	γ_{34}	-0.103 (0.049)	-0.068 (0.035)
ρ_{35}	-0.028 (0.085)	-0.005 (0.080)	γ_{35}	-0.034 (0.057)	-0.003 (0.046)
ρ_{45}	0.847 (0.023)	0.861 (0.020)	γ_{45}	0.795 (0.040)	0.786 (0.023)

Table 5: Estimates of Index Function $g(\mathbf{Y})$ with Approximate 95% Confidence Intervals

	Impute	**Multi MLE**
Estimated Mean	0.085	0.080
Estimated Variance	0.004	0.003
Standard Error	0.061	0.058
95% CI about $g(\mathbf{Y})$	(-0.036,0.206)	(-0.035,0.195)
95% CI about $ln(g(\mathbf{Y}))$	(-∞,-1.581)	(-∞,-1.632)

multivariate method utilizes the information provided by the other variables in the analysis for that particular observation while taking the covariance structure into consideration. For instance, suppose we want to estimate the population mean and standard deviation of a function, denoted $g(\cdot)$, of our data. With complete data we could compute $g(\cdot)$ for each observation and then directly find its sample mean and standard error. Conversely, $g(\cdot)$ cannot be directly calculated for any observation that has levels of Cu, Pb, Zn, Ca, or Mg completely missing or nondetected. With estimates of the population means, variances, and covariances for all measures, we can indirectly estimate the mean and standard deviation of the desired function.

The results reveal that the choice of method does not alter the conclusion for this particular application. Regardless of the method for analysis, we conclude that there is sufficient evidence to suggest that the collective concentration levels of the trace metals Cu, Pb and Zn in freshwater streams across Virginia are significantly different from the reported worldwide standard, while correcting for the water hardness as a function of Ca and Mg. In fact, it appears that freshwater streams in Virginia are less contaminated with these trace metals than other freshwater streams around the world. Note that for these data the LOD values are relatively small as compared with the values of the observed data. We would have predicted that imputation performs better than the MLE method. Consider comparing the LOD values of each metal reported in Table 2 to the corresponding mean levels of the observed subset of the data only in Table 3. In doing so, notice that for the metals Cu, Ca and Mg—which have censoring percentages less than 10%—the LOD values are small relative to the corresponding means. While the LOD values for Pb and Zn are also smaller than their respective means, this difference is less dramatic than for Cu, Ca and Mg. Moreover, Pb and Zn have much higher percentages of censored values (78% and 39%, respectively); and so, the means estimated from the imputation method for Pb and Zn could have a larger bias, since the upper bound on the bias is much higher.

As previously cited, statistical analyses used to evaluate the trace element chemistry of groundwaters are typically complicated by the presence of many trace metals at concentrations below the LOD. Thus, Farnham *et al.* (2002) sought a new imputation approach that utilizes the best value for substitution. In performing Monte Carlo simulation experiments with a mixture multivariate model to test the performance of the substituted values 0, LOD and $LOD/2$, Farnham *et al.* (2002) not surprisingly showed that substitution of $LOD/2$ yields superior results compared to 0 and LOD; however, the performance of all substitution methods declines when the percentage of values below the LOD exceeds 25%.

6. Conclusion

The multivariate MLE tool is the optimal method in that it provided more accurate and precise estimates; however, the computational time can be excessive when large numbers of variables are being analyzed.

We have provided a tool that can be applied to a variety of environmental data to obtain the MLEs of multivariate normal parameters in the presence of left-censored and missing data. Estimation of the parameters of left-censored multivariate normal data does not, however, end with the proposed tool. Rather, this tool simply opens the door to new issues to explore. We assume that the variables are missing completely at random (MCAR), but can this approach be adapted for nonignorable missing data? For instance, values may be missing due to some specific problems existing in the assay itself that prohibit one from obtaining an observed value, such as some external contaminant or mechanical malfunction in the instrument. We resorted to Legendre-Gauss quadrature in estimating the multivariate normal CDF with our multivariate MLE tool. Alternate methods may show improved efficiency in comparison. Additionally, one could consider mixtures of distributions as opposed to only a single multivariate normal for the entire set of data.

In summary, we proposed a method that provides the MLEs of mean and unstructured (co)variance parameters corresponding to a multivariate (log)normal distribution in the presence of left-censored and missing values. The resulting estimates may be used to approximate and make inferences about one or more composite functions of the measures.

Acknowledgements

We would like to thank the following collaborators who assisted us with this research by reviewing an earlier draft of this manuscript and offering insightful recommendations: Dr. Kellie Archer, Dr. Chris Gennings, Dr. James Mays, and Dr. Greg Miller. In addition, we would like to thank Dr. Donald Smith and Dr. Roger Stewart of the Virginia Department of Environmental Quality for supplying the dissolved trace metals data for the application.

References

Afifi AA, Elashoff RM (1967). "Missing Values in Multivariate Statistics II: Point Estimation in Simple Linear Regression." *Journal of the American Statistical Association*, **62**, 10–29.

Bagby RJ (1995). "Calculating Normal Probabilities." *The American Mathematical Monthly*, **102**(1), 46–49.

Bowling SR, Khasawneh MT, Kaewkuekool S, Cho BR (2009). "A Logistic Approximation to the Cumulative Normal Distribution." *Journal of Industrial Engineering and Management*, **2**(1), 114–127.

Casella G, Berger RL (2001). *Statistical Inference*, chapter 5.5.4, pp. 240–244. Second edition. Australia, Duxbury Press.

Farnham IM, Singh AK, Stetzenbach KJ, Johannesson KH (2002). "Treatment of Nondetects in Multivariate Analyses of Groundwater Geochemistry Data." *Chemometrics and Intelligent Laboratory Systems*, **60**, 265–281.

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**, 141–150.

Gilliom RJ, Helsel DR (1986). "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 1. Estimation Techniques." *Water Resources Research*, **22**(2), 135–146.

Helsel DR (1990). "Less Than Obvious: Statistical Treatment of Data Below the Reporting Limit." *Environmental Science and Technology*, **24**(12), 1766–1774.

Helsel DR (2005). *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. New Jersy, John Wiley and Sons, Inc.

Hildebrand FB (1956). *Introduction to Numerical Analysis*, chapter 8.5, pp. 323–325. New York, McGraw-Hill.

Hocking RR, Smith WB (1968). "Estimation of Parameters in the Multivariate Normal Distribution With Missing Observations." *Journal of the American Statistical Association*, **63**, 159–173.

Koo JW, Parham F, Kohn MC, Masten SA, Brock JW, Needham LL, Portier CJ (2002). "The Association Between Biomarker-Based Exposure Estimates for Phthalates and Demographic Factors in a Human Reference Population." *Environmental Health Perspectives*, **110**(4), 405–410.

Ledo de Medina H (2000). "Chemical Parameters." In LML Nollet (ed.), *Handbook of Water Analysis*, chapter 5, p. 58. New York, Marcel Dekker, Inc.

Morrison DF (1971). "Expectations and Variances of Maximum Likelihood Estimates of the Multivariate Normal Distribution Parameters With Missing Data." *Journal of the American Statistical Association*, **66**(335), 602–604.

Rao ST, Ku JY, Rao KS (1991). "Analysis of Toxic Air Contaminant Data Containing Concentrations Below the Limit of Detection." *Journal of the Air and Waste Management Association*, **41**, 442–448.

Sanford RF, Pierson CT, Crovelli RA (1993). "An Objective Replacement Method for Censored Geochemical Data." *Mathematical Geology*, **25**, 59–80.

Shore H (2004). "Response Modeling Methodology (RMM): Current Distributions, Transformations, and Approximations as Special Cases of the RMM Error Distribution." *Communications in Statistics (Theory and Methods)*, **33**(7), 1491–1510.

Shore H (2005). *Approximating the CDF of the Standard Normal*, volume 8 of *Quality, Reliability and Engineering Statistics*, chapter 19.4.2, pp. 325–330. Singapore, World Scientific Publishing Company.

Singh A, Nocerino J (2002). "Robust Estimation of Mean and Variance Using Environmental Data Sets With Below Detection Limit Observations." *Chemometrics and Intelligent Laboratory Systems*, **60**, 69–86.

Travis CC, Land ML (1990). "Estimating the Mean of Data Sets With Nondetectable Values." *Environmental Science and Technology*, **24**(7), 961–962.

US EPA (1997). "Terms of Environment Glossary, Abbreviations and Acronyms." *Technical Report 175-B-97-001*, United States Environmental Protection Agency. URL http://www.epa.gov/glossary.

VDEQ (2003). "The Quality of Virginia Non-Tidal Streams: First Year Report." *VDEQ Technical Bulletin WQA/2002-2001*, Office of Water Quality and Assessments, Virginia Department of Environmental Quality. URL http://www.deq.state.va.us/probmon/pdf/report1.pdf.

VDEQ (2008). "Virginia Water Quality Assessment." *Integrated Report 305(b)/303(d)*, Virginia Department of Environmental Quality. URL http://www.deq.virginia.gov/wqa/ir2008.html.

VDEQ (2009). "Virginia Water Quality Standards." *Technical Report Regulation 9 VAC 25-260*, State Water Control Board, Virginia Department of Environmental Quality. URL http://www.deq.virginia.gov/wqs/documents/WQS_eff_20Aug2009.pdf.

Zhao Y, Frey HC (2006). "Uncertainty for Data With Non-Detects: Air Toxic Emissions From Combustion." *Human and Ecological Risk Assessment: An International Journal*, **12**(6), 1171–Ű1191.

Affiliation:

Heather J. Hoffman
Department of Epidemiology and Biostatistics
The George Washington University
Washington, DC 20037
E-mail: sphhjh@gwumc.edu
URL: http://www.gwumc.edu/sphhs/departments/epibio/

http://www.jenvstat.org
Submitted: 2010-06-25
Accepted: 2011-03-21