Mathematical Geology, Vol. 25, No. 1, 1993

An Objective Replacement Method for Censored Geochemical Data[1]

Richard F. Sanford,[2] Charles T. Pierson,[2] and Robert A. Crovelli[2]

Geochemical data are commonly censored, that is, concentrations for some samples are reported as "less than" or "greater than" some value. Censored data hampers statistical analysis because certain computational techniques used in statistical analysis require a complete set of uncensored data. We show that the simple substitution method for creating an uncensored dataset, e.g., replacement by $\frac{3}{4}$ times the detection limit, has serious flaws, and we present an objective method to determine the replacement value. Our basic premise is that the replacement value should equal the mean of the actual values represented by the qualified data. We adapt the maximum likelihood approach (Cohen, 1961) to estimate this mean. This method reproduces the mean and skewness as well or better than a simple substitution method using $\frac{3}{4}$ of the lower detection limit or $\frac{4}{3}$ of the upper detection limit. For a small proportion of "less than" substitutions, a simple-substitution replacement factor of 0.55 is preferable to $\frac{3}{4}$; for a small proportion of "greater than" substitutions, a simple-substitution replacement factor of 1.7 is preferable to $\frac{4}{3}$, provided the resulting replacement value does not exceed 100%. For more than 10% replacement, a mean empirical factor may be used. However, empirically determined simple-substitution replacement factors usually vary among different data sets and are less reliable with more replacements. Therefore, a maximum likelihood method is superior in general. Theoretical and empirical analyses show that true replacement factors for "less thans" decrease in magnitude with more replacements and larger standard deviation; those for "greater thans" increase in magnitude with more replacements and larger standard deviation. In contrast to any simple substitution method, the maximum likelihood method reproduces these variations. Using the maximum likelihood method for replacing "less thans" in our sample data set, correlation coefficients were reasonably accurately estimated in 90% of the cases for as much as 40% replacement and in 60% of the cases for 80% replacement. These results suggest that censored data can be utilized more than is commonly realized.

KEY WORDS: substitution method, qualified data, lognormal distribution, environmental science.

INTRODUCTION

Geochemical data are commonly qualified or censored. A qualified or censored dataset is any dataset that contains one or more qualified or censored values. A

[1]Received 20 May 1991; accepted 20 February 1992.

[2]U.S. Geological Survey, M.S. 905, Denver Federal Center, Denver, Colorado 80225.

qualified or censored value is any value reported as less than or greater than a censor point. For geochemical data, the censor point is typically the analytical limit of detection. Commonly, qualified values are referred to as "less thans" or "greater thans." Censoring typically results when analytical techniques are not sensitive enough to detect small quantities of an element or when the technique is so sensitive that large concentrations overwhelm the detection system. Censored data hampers statistical analysis because some important statistical techniques require a complete set of uncensored data. The raw data must therefore be processed to replace "less thans" and "greater thans" with unqualified values. Typically an arbitrary constant is the replacement value. For example, workers who use the STATPAC program of the U.S. Geological Survey often substitute $\frac{3}{4}$ of the detection limit for "less thans" and $\frac{4}{3}$ of the upper limit for "greater thans" (VanTrump, 1977). In some cases, a value of 0 is used (Miesch, 1976b). The decision faced by the investigator is then: what percentage of replacements is justifiable without introducing significant errors? Rules-of-thumb have been that 10–20% replaced is the maximum allowable. If too many replacements are made, then accuracy is sacrificed; if too few replacements are made, then valuable data are not utilized. We present here an objective method to determine the replacement value, and we show that this method allows greater utilization of censored data with minimal loss of accuracy in estimation of descriptive statistics and in factor analysis.

Our basic premise is that the replacement value should be equal to the mean of the actual values represented by the qualified data. We adapt the maximum likelihood approach (Cohen, 1961) to estimate this mean.

The tendency of geochemical data to obey a lognormal distribution is well known (De Wijs, 1951; Krige, 1951, 1960; Sichel, 1952, 1966; Miesch and Riley, 1961; Miesch, 1967, 1976b). For such data, maximum likelihood estimation methods can closely reproduce descriptive statistics such as the mean and standard deviation (Sichel, 1952, Cohen, 1959, 1961, 1976; Krige, 1960; Cohn, 1988). Various methods of estimating descriptive statistics have been reviewed and evaluated by Gilliom and Helsel (1986), Helsel and Gilliom (1986), and Helsel and Cohn (1988). We propose that such methods can be adapted for determining a replacement value in statistical analysis, and that replacement using the maximum-likelihood estimate of the mean is more objective and accurate than the commonly used simple substitution methods. Our numerical experiments on a representative set of geochemical data and our theoretical analysis shows this method to be superior to common methods in reproducing the mean, standard deviation, and correlation coefficients of the uncensored data.

METHODS

In essence, our method consists of estimating the mean of the entire lognormal population (normal distribution), using a maximum likelihood method. This estimated mean is then used to calculate an estimated replacement value for the qualified values.

For clarity in the following discussion, we define some terms and corresponding symbols here. Concentration is denoted by x, and the detection limit or censor point, by x_d. The replacement value, x_r, is the number to be substituted in place of each of the qualified values, that is, for the "less thans" where $x < x_d$ and for the "greater thans" where $x > x_d$. The replacement factor, r_x, is the ratio of the replacement value divided by the detection limit for a particular constituent. By definition,

$$r_x \equiv \frac{x_r}{x_d} \tag{1}$$

For example, $\frac{3}{4}$ is a commonly used replacement factor, and $\frac{3}{4}$ times the detection limit is the corresponding replacement value. Because the detection limit may be different for different elements, the corresponding replacement value may also vary, although the replacement factor is a constant. We use the term "qualified" to refer to an individual sample. We use the term "censored" to refer to a data set. Thus a censored data set contains qualified data. "Detection limit" is synonymous with "censor point" used by other authors.

The sample data set is from a mineral resource study in the San Juan Mountains of southwestern Colorado (Sanford et al., 1987a,b). The 265 samples consist of dacitic to rhyolitic, unmineralized, and mineralized igneous rocks and of epithermal polymetallic veins. Sampling was concentrated in areas exhibiting evidence of mineral potential, and consequently veins and mineralized rocks are over-represented. Although random sampling is preferable statistically, our purpose is to improve methods of statistical analysis of actual datasets which are typically not collected according to rigorous statistical procedures. For our numerical experiments on replacement of values below the lower detection limit, we selected the nine elements that were at least 93% uncensored: Al, Ca, Cu, Fe, Li, Mn, Na, Pb, and U. For numerical experiments on replacement of data censored at the upper limit, we selected the five least censored elements, Al, Fe, Li, Mn, and Pb. Elements were analyzed by inductively-coupled Ar-plasma emission spectroscopy (ICP), except for U, which was analyzed by delayed neutron activation (DNAA) (Sanford et al., 1987a). Because several different geochemical processes control the distribution and concentration of these elements, the selected elements are representative of a wide range of elements. For example, Pb and Cu are concentrated during epithermal mineralization and are representative of Ag, As, Au, Sb, Zn, etc.; Al, Ca, and Na are controlled by primary igneous processes and by hydrothermal alteration and proxy for K, Mg, Ti, V, etc.

After determining that the logarithms of the concentrations were more normally distributed than the original units, we converted the data for the selected elements to $\log_{10}$. The remainder of the paper will refer to $\log_{10}$ data unless otherwise stated. The transformations between log data, y, and original data, x, are as follows: for any x,

$$y \equiv \log_{10} x, \qquad x = 10^{y} \tag{2}$$

for the replacement value, x_r

$$y_r \equiv \log_{10} x_r, \qquad x_r = 10^{y_r} \tag{3}$$

and for the detection limit, x_d,

$$y_d \equiv \log_{10} x_d, \qquad x_d = 10^{y_d} \tag{4}$$

Taking the log of both sides of Eq. (1) yields the transformed replacement factor, r_y,

$$r_y \equiv \log_{10} r_x = \log_{10} x_r - \log_{10} x_d = y_r - y_d, \qquad r_x = 10^{r_y} \tag{5}$$

Thus, the log of the replacement factor is equal to the *difference* between the log of the replacement value and the log of the detection limit.

A variety of alternative transformations could be used instead of $\log_{10} x$, for example $\log_{10} (\pm x \pm \alpha)$, where α is a constant determined from the data set (Chayes, 1954; Miesch, 1967, 1982). The purpose of any transformation is to convert the variable x to one that has a more normal distribution so that the theory of normal distributions can be applied. In many cases other transformations can be superior to $\log_{10} x$ in normalizing data. However, we use $\log_{10} x$ in this paper, because it is simple to compute, it is a significant improvement over arbitrary simple substitution methods, and it achieves a satisfactory accuracy in our numerical experiments.

We then made the necessary small number of replacements to the sample data set to yield a completely uncensored reference set which we used in subsequent numerical experiments. Because the number of replacements was small, these corrections have a negligible effect on the means, standard deviations, and correlation coefficients. No replacements were needed for Al, Fe, and Pb, which are 100% uncensored. For the censored data sets of Ca, Cu, Li, and Na, the method of Cohen (1961) was used to estimate the true means. The replacement values were then calculated using this mean according to the method described below. For U, two values below the detection limit were assigned $\frac{3}{4}$ of the detection limit, and six samples having unreported values were given the value of the mean of the data after making the above replacement. For Mn, we substituted $\frac{3}{4}$ of the detection limit for the two "less thans" and determined the replacement value for the nine "greater thans" according to our adaptation of Cohen's method, described below. The effect of these substitutions on the descriptive statistics was negligible.

We then obtained the univariate statistics and correlation coefficients for the dequalified sample data set using the U.S. Geological Survey STATPAC program (VanTrump, 1977). These results served as the standard for judging the performance of various replacement methods. We then modified the data set

to simulate detection limits at 10, 20, 30, 40, 50, 60, 70, and 80% "less thans" and 5, 10, 20, and 30% "greater thans." For each of these modified sets, we replaced the artificially qualified values with: (1) a replacement value equal to $\frac{3}{4}$ of the lower detection limit for "less thans" and $\frac{4}{3}$ of the upper limit for "greater thans," and (2) a replacement value calculated using Cohen's (Cohen, 1961) maximum likelihood method. Replacement values using Cohen's method were calculated according to Eq. (7) and the procedure described next. The replacement values are calculated to give the dequalified datasets the same means as those estimated by Cohen's method. For all of these sets we found the univariate statistics and correlation coefficients, which we compared to those of the reference set.

Our adaptation of Cohen's maximum likelihood method is as follows. Our basic premise is that the replacement value should be equal to the mean of the actual values that are reported as qualified. We use a maximum likelihood approach (Cohen, 1961) to estimate the true mean of the whole dataset, then this result is used to estimate the true mean of the data reported as qualified values. Because our artificially censored sample datasets are singly censored, we use the method of Cohen (1961) to estimate the mean of the whole dataset, $\hat{\mu}$, including qualified data. (For datasets that are multiply or progressively censored we would use extensions of this method [Cohen, 1976; Helsel and Cohn, 1988]). We also estimate the mean of the unqualified data, $\hat{\mu}_u$. (The symbol, ^, designates estimated values.) The mean of the whole dataset, $\hat{\mu}$, estimated using Cohen's method, times the total number of samples, n, is equal to the (unknown) mean of the qualified data, $\hat{\mu}_q$, times the number of qualified samples, n_q, plus the mean of the unqualified samples, μ_u (which is known), times the number of unqualified samples, n_u:

$$n\hat{\mu} = n_q\hat{\mu}_q + n_u\mu_u \tag{6}$$

Solving for $\hat{\mu}_q$ yields the estimated mean of the qualified data,

$$\hat{\mu}_q = \frac{n\hat{\mu} - n_u\mu_u}{n_q} \tag{7}$$

Our basic assumption is that this estimated mean of the qualified data is the preferred replacement value, that is (in log units),

$$y_r = \hat{\mu}_q \tag{8}$$

Using Eq. (3), the replacement value in original units is

$$x_r = 10^{\hat{\mu}_q} \tag{9}$$

Just as we could have used other transformations than $\log_{10} x$, we could have used other maximum likelihood methods to estimate the mean of the uncensored

dataset, for example, those of Sichel (1952) or Krige (1960). We use Cohen's method for simplicity and convenience.

RESULTS

Numerical Experiments of Censoring at Low Concentrations

In this section, we first evaluate the accuracy of the maximum likelihood and simple substitution methods in estimating the true mean and skewness from the artificially censored datasets. Then we calculate the true replacement values and show that they are a function of the proportion of replacements and the standard deviation. Finally we evaluate three methods of obtaining correlation coefficients and show that the maximum likelihood method is preferred.

For "less than" replacements, our method yields mean values closer to the true mean than the simple substitution method ($\frac{3}{4}$ of the detection limit) for seven of the nine elements tested (Fig. 1). In this and subsequent figures, the true value of the quantity plotted (mean, skewness, replacement factor, etc., depending on the figure) is at the left side of the diagram where the percentage of "less thans" replaced is ≈ 0. The calculated value of the quantity plotted is shown as a function of the percentage of replacements. Ideally, the calculated values should plot as a horizontal line having an intercept equal to the true value. In all cases, the simple substitution method for "less thans" yields a higher estimated mean. For Mn, the maximum likelihood method underestimates the mean by the same amount that the simple substitution method overestimates it. The only element whose mean is better estimated by the simple substitution method is U. These two elements, Mn and U, are characterized by several samples having extremely high concentrations and therefore by large positive skewness (≥ 1.0) (Fig. 2).

Skewness is consistently better estimated by the maximum likelihood method than by the simple substitution method (Fig. 2). For both methods, the skewness increases with more replacements. This is an unavoidable result of replacing all the qualified data by a single replacement value.

For both mean and skewness, the deviation from the true value generally increases gradually with the proportion of replacements (Figs. 1 and 2). For as much as 40% replacements, the error in the mean and skewness by the maximum likelihood method is fairly small. From 40 to 80% the error in the estimated mean is still small for the more normally distributed elements, but the elements having more skewed distributions have more serious errors in the estimated means.

From the original unmodified data sets, we determined the true or empirical replacement factor. This factor is the mean of the "less thans" when we assume fictitious censor points at 10, 20, 30, etc. percent replacement. The empirical

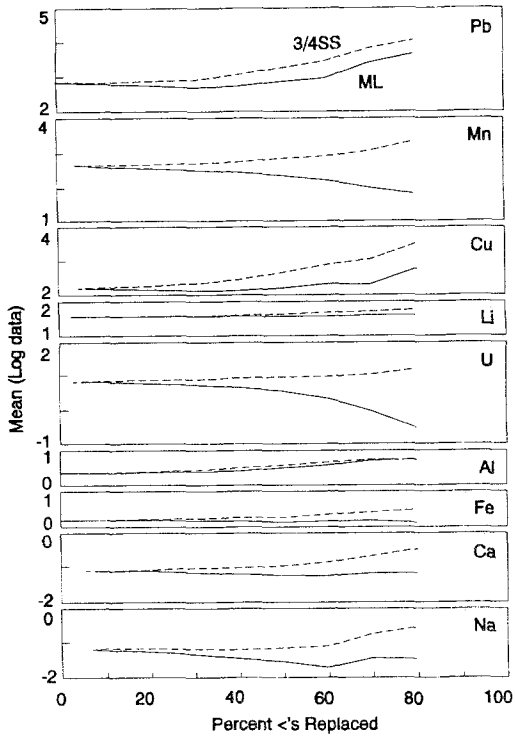

Fig. 1. Means of the data after replacement by the maximum likelihood method (ML, solid lines) and the $\frac{3}{4}$-detection-limit simple substitution method ($\frac{3}{4}$SS, dashed lines) vs. proportion of replacements for the artificially censored reference set.

replacement factor decreases with increasing proportion of replacement and with increasing standard deviation (Fig. 3). The reason for this relationship is discussed in the section below on theoretical replacement values. Exceptions to this rule result from data that significantly violate the assumption of normality, particularly when the data are significantly skewed. For example, the curves in Fig. 3 are sequentially ordered downward according to increasing standard deviation, with the exception of U. The curve for U ($\sigma = 0.88$) should be between the curves for Na ($\sigma = 0.84$) and Pb ($\sigma = 1.16$). However, the U data is positively skewed more than the other data (Fig. 2), and consequently the U curve is displaced to higher replacement values.

The simple substitution method using $\frac{3}{4}$ times the detection limit consistently overestimates the replacement value, and this overestimation is worse with more replacements. The maximum likelihood method of estimating the replacement

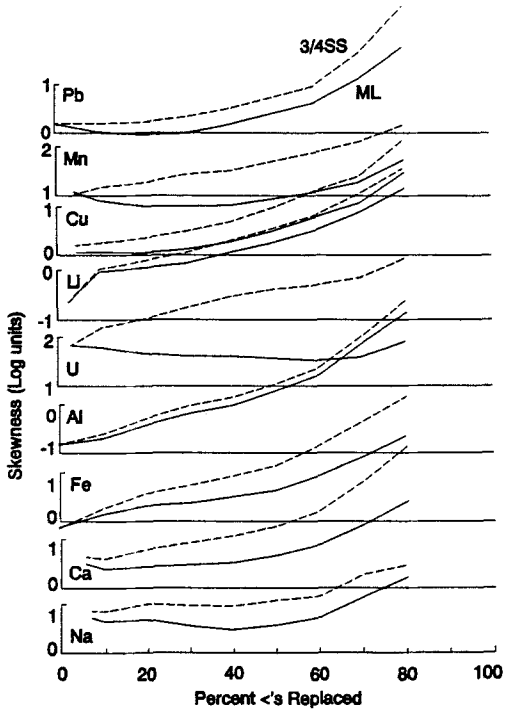

Fig. 2. Skewness of the data after replacement by the maximum likelihood method (ML, solid lines) and the $\frac{3}{4}$-detection-limit simple substitution method ($\frac{3}{4}$SS, dashed lines) vs. proportion of replacements for the artificially censored reference set.

factor accounts for both the variation with standard deviation and with proportion of replacements. Thus, the maximum likelihood method is clearly superior to the simple substitution method.

Further, if a simple substitution method must be used, then an empirical factor is superior to the arbitrary $\frac{3}{4}$-factor commonly used. The replacement factor should vary depending on the proportion of replacements and should not be a constant such as $\frac{3}{4}$. If the standard deviations for individual elements are unknown, then a mean factor has to suffice. The mean of the empirical replacement factors for the nine elements is shown by the curve labeled "Mean" in Fig. 3. This constant, when multiplied by (or added to, in $\log_{10}$ units) the detection limit, yields on average the true mean of the censored samples. An empirical replacement factor should be read from the "Mean" curve for the appropriate percent of replacements. The particular factor may vary with different datasets, but for our particular set, average replacement factors are, in

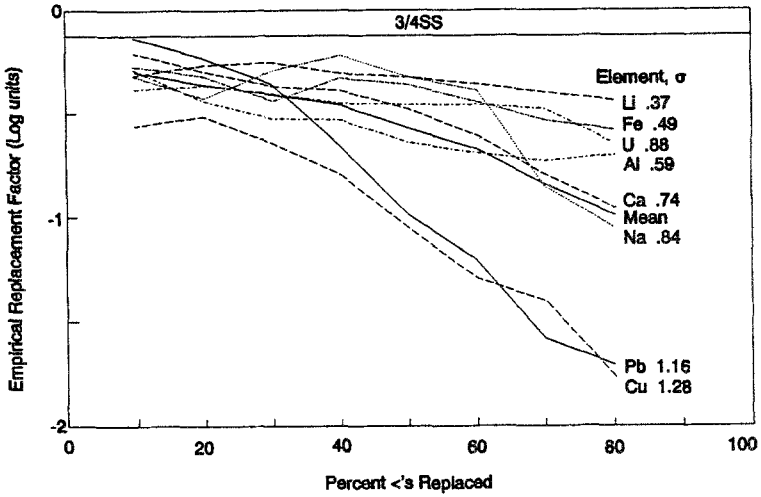

Fig. 3. Replacement factor as a function of replacements. The replacement factors plotted are calculated from the known means of the artificially censored samples and are thus the "true" replacement factors. Note that the replacement factor decreases with increasing replacements and standard deviation (σ). The replacement factor for the $\frac{3}{4}$ simple substitution method ($\frac{3}{4}$SS) plots as a horizontal line at log ($\frac{3}{4}$) = -0.12, which is systematically higher than the true factor.

original units as a function of percent replacement: for 10%, 0.55; for 20%, 0.45; for 30%, 0.40 and so on.

Correlation coefficients for the 28 pairs of elements were computed using: (1) the unqualified data only, no replacements, (2) "less thans" replaced by $\frac{3}{4}$ of the detection limit, and (3) "less thans" replaced using our maximum likelihood method. The pairs Al–Pb and Al–U illustrate typical results (Fig. 4). The maximum likelihood method yields significantly better correlation coefficients than the simple substitution method using $\frac{3}{4}$ of the detection limit. Not making any replacements consistently yields the worst results and sometimes (e.g., Fig. 4b), with a large proportion of censored data, indicates apparently significant correlations having the wrong sign! This consequence is explained below.

The accuracy of the three methods in reproducing the correlation coefficients is evaluated by the root mean square error (rmse; Fig. 5a), and by a scoring system (Fig. 5b). The two tests rate the performance by the amount of the deviation and by the number of correct correlations, respectively. The rmse is equal to the square root of the mean of the squares of the differences between the estimated and actual correlation coefficients.

The scores are calculated by counting the number of correlations that fall

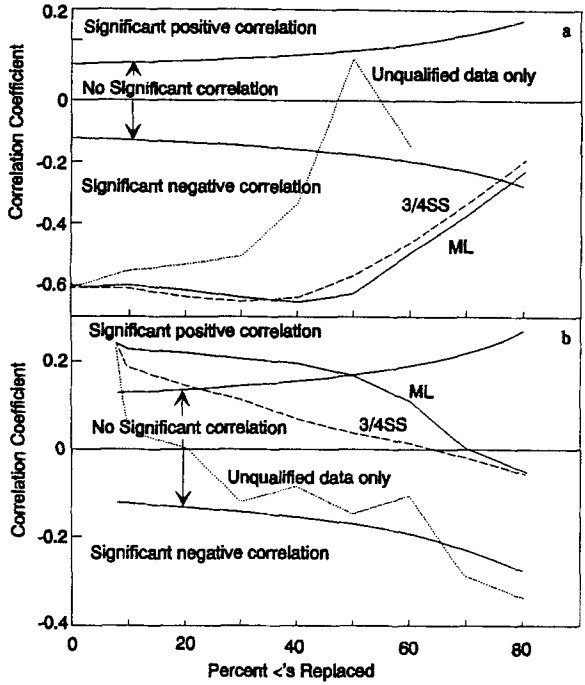

Fig. 4. Two examples of correlation coefficients calculated from artificially censored data using uncensored samples only, the $\frac{3}{4}$ simple substitution method ($\frac{3}{4}$SS), and the maximum likelihood method (ML). (a) Al vs. Pb, (b) Al vs. U. For most of the 28 element paris, the maximum likelihood method yields the correlation coefficient closest to the known value. Significance is calculated at the 95% confidence level according to Crow et al. (1960, p. 241).

in the correct category (significant-positive, not significant, or significant-negative as determined from the uncensored reference dataset), summing over all the 28 pairs, and normalizing to 100. For example, in Fig. 4a, all three methods yield the correct significant-negative correlation from 10 to 40%, so each method scores 1 point; at 50–60%, the unqualified-only method scores 0, whereas the two replacement methods each score 1 point; at 70%, there are too few pairs to compute a correlation coefficient for the unqualified-only method, but the two replacement methods score 1 point each; at 80%, both replacement methods yield an incorrect "not significant," and both score 0. This process is repeated for each element. The scores are summed for all elements and the result normalized to 100.

The rmse and scores are consistently better for the maximum likelihood method than for the simple substitution method, and both replacement methods

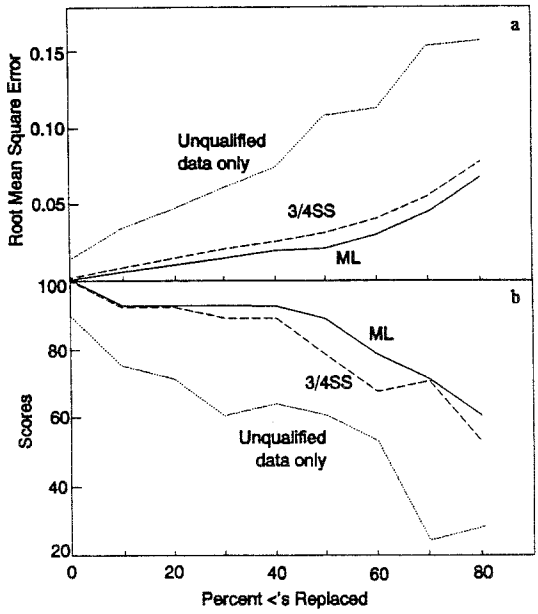

Fig. 5. Summary of accuracy in estimating the correlation coefficient by the three methods. (a) root mean square error (rmse), (b) numerical scores. Correlation coefficients from unqualified data only are unreliable. The maximum likelihood method (ML) typically yields more accurate results than the $\frac{3}{4}$ simple substitution method ($\frac{3}{4}$SS).

are much better than ignoring the qualified values altogether. The differences in rmse among the methods increase with more replacements, although the scores maintain the same relationships beyond about 40%. For as much as 40% replacements, the maximum likelihood method estimates the correct category for 26 out of 28, or 93%, of the correlations (Fig. 5b). By comparison, the simple substitution method is correct in 89% of correlations, and the qualified-only method correctly predicts only 64% of correlations. At 80% replacement, the corresponding scores are 61, 54, and 29% correct, respectively.

The maximum likelihood method is clearly preferable to the simple substitution method and to neglecting unqualified data. For maximum likelihood replacement on our sample data, as much as 40% replacement yields a relatively small error in correlation coefficient and is correct in better than 90% of the cases. Replacement of as much as 80% qualified data yields useful results in 60% of the cases. These results suggest that censored data can be utilized more than is commonly realized.

Theoretical Replacement Values at Low Concentrations

Figure 3 shows that the empirical replacement factor for "less thans" decreases with increasing proportion of replacements. The $\log_{10}$ replacement values range from -0.1 to -0.6 at 10% replacement to -0.4 to -1.8 at 80% replacement (in original units, 0.8 to 0.25 at 10% and 0.4 to 0.016 at 80%). There is also a pronounced tendency toward lower replacement factors with greater standard deviation. In order to understand these empirical relationships, we compared them with the relationships derived from probability theory and based on the normal distribution. What are the expected replacement factors assuming a normal distribution, and how do they compare with the empirically determined replacement factors?

We now assume that the concentration, y, expressed in log units, is a normally distributed random variable. We define the following quantities:

μ: Mean of y

σ: Standard deviation of y

z: Standardized normal variable, $z = (y - \mu)/\sigma$.

The standard normal distribution is given by the probability density function, $\varphi(z)$ (from Korn and Korn, 1968, pp. 629–631),

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp -\left(\frac{z^2}{2}\right) \tag{10}$$

The cumulative distribution function, $\Phi(z)$, is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp -\left(\frac{z^2}{2}\right) \tag{11}$$

which can be expressed by the commonly tabulated error function erf; from Korn and Korn, 1968, p. 630,

$$\Phi(z) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right] \tag{12}$$

$\Phi(z)$ varies from 0 to 1 and, for our purposes, is the proportion of qualified values replaced. The expected value or mean of the "less thans," i.e., the expected replacement value, μ_r, is

$$\mu_r = \mu - \frac{\sigma}{\sqrt{2\pi}} \left\{ \frac{\exp\left[-\left(\frac{z_d^2}{2} \right) \right]}{\Phi(z_d)} \right\} \tag{13}$$

where

$$z_d \equiv \frac{y_d - \mu}{\sigma} \tag{14}$$

By setting

$$y_r = \mu_r \tag{15}$$

we say that the desired replacement value (in log units) here is the expected value or mean of those concentrations less than the detection limit. This equation is analogous to Eq. (8), but using the theoretical conditional mean instead of the empirical mean of the qualified values. Transforming from log units, as in Eq. (3) and (9), yields the theoretical replacement value in original units,

$$x_r = 10^{\mu_r} \tag{16}$$

By substituting μ_r for y_r in Eq. (5) we obtain the expected transformed replacement factor, r_y,

$$r_y \equiv \mu_r - y_d \tag{17}$$

Now, rearranging the definition of z_d, and solving for y_d yields

$$y_d = z_d\sigma + \mu \tag{18}$$

Substituting μ_r from Eq. (13) and y_d from Eq. (18) into Eq. (17) yields the expected replacement factor as a function of the cumulative distribution function,

$$r_y = -\frac{\sigma}{\sqrt{2\pi}} \left\{ \frac{\exp\left[-\left(\frac{z_d^2}{2} \right) \right]}{\Phi(z_d)} \right\} - z_d\sigma \tag{19}$$

As in Eq. (5), transforming yields the expected replacement factor in original units, r_x,

$$r_x \equiv 10^{r_y} \tag{20}$$

Equation (19) expresses the theoretical relationship between the replacement factor and the proportion of replacements for the normal distribution. It is the theoretical relationship sought at the beginning of this section. We can now compare the empirically determined replacement factor with the theoretically

calculated replacement factor. By varying z_d and σ and treating them as variables, the replacement factor is now only a function of z_d and σ. Note that z_d represents standardized values; therefore z_d can be simply denoted by z, and r_y is not a function of μ. This is the identical conclusion drawn from the empirical data above. A plot of r_y vs. $\Phi(z)$ contoured for different values of σ shows both the decrease in the replacement factor with increasing replacements and the lower replacement factors for greater standard deviation (Fig. 6). The empirical curves (Fig. 3) closely resemble these theoretically derived curves in form. Thus, the theoretical analysis of the ideal normal distribution confirms the results from empirical analysis of the sample data.

The reason for these relationships are illustrated in Fig. 7. Figure 7a compares replacement values and replacement factors for two different detection limits given the same normal distribution. Because of the shape of the probability distribution function, the replacement value (y_r) increases more slowly than the detection limit (y_d). Consequently, the replacement factor (r_y), which is the difference between the two (see Eq. 5), increases in magnitude as the detection limit increases. The replacement factor for "less thans" is always less than the detection limit, so r_y is always negative. Thus, the replacement factor becomes more negative with higher detection limits, all else being equal.

Figure 7b shows the same detection limit (y''_d) as in Fig. 7a (y_d) and the same mean, but the standard deviation is larger. This causes the curve below the mean and below the detection limit to extend to lower values of y for the

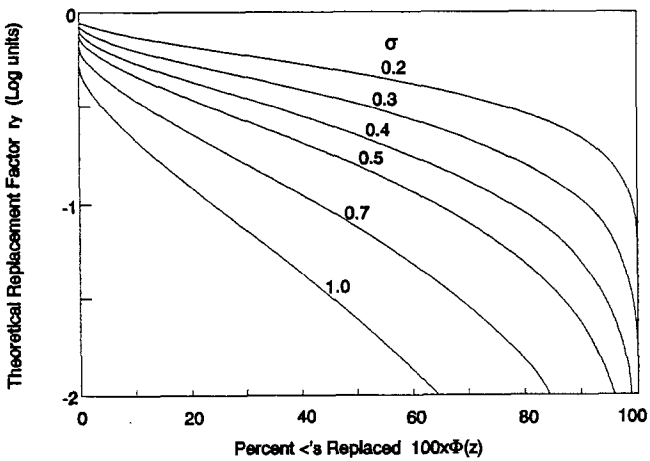

Fig. 6. Plot of theoretical replacement factor (Eq. 19) vs. proportion of replacements or cumulative distribution function (Eq. 11) contoured for standard deviation. Note that the form of the curves mimics that of the empirical curves in Fig. 3.

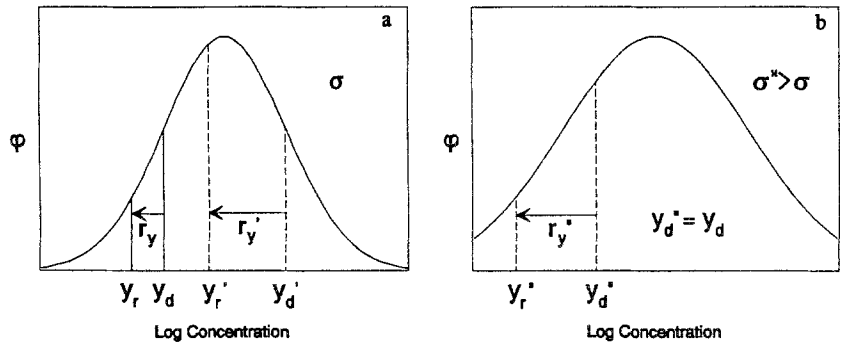

Fig. 7. Normal probability density function showing effect of variation in detection limit (y_d) and standard deviation (σ) on replacement value (y_r) and replacement factor (r_y) assuming log normal distribution. Logarithm of replacement factor is always negative for "less thans" (see Eq. 5) and has greater absolute value for both higher detection limit (a, variables with '), and larger standard deviation (b, variables with ").

same values of φ compared to Fig. 7a. Thus the replacement factor is more negative with larger standard deviation.

Empirical and theoretical replacement factors show some discrepancies. There is a systematic bias toward higher empirical replacement factors for a given percent replacement and standard deviation compared to the theoretical replacement factors. This bias is at least partly due to the tendency of the real data to be positively skewed, which tends to increase the calculated replacement factor, as discussed above. Another factor may be that x, the concentration variable, physically cannot exceed 100% (Chayes, 1960). Thus, at high concentrations, x behaves less like a random variable and begins to violate the governing assumptions of this analysis. Other deviations from normality, such as multiple modality may also account for the discrepancies between actual and theoretical replacement factors.

Figure 8 illustrates the effects of various replacement methods. The calculation of correlation coefficients from unqualified data only should always be avoided, because the unqualified data are an arbitrary subset of the whole dataset and are not necessarily representative. As shown here and in Figs. 4 and 6, the deviations from the true correlation coefficients become increasingly severe with more replacements. For our dataset, the maximum likelihood method of replacement performs as well as replacement by the true mean replacement value. Despite differences in the linear regression coefficients, the correlation factors are remarkably accurate, as shown by the similar r-values.

In spite of deviations from normality, the correlation coefficients calculated from the maximum likelihood method are similar enough to the true correlation coefficients to be useful for many purposes. This observation, based on our

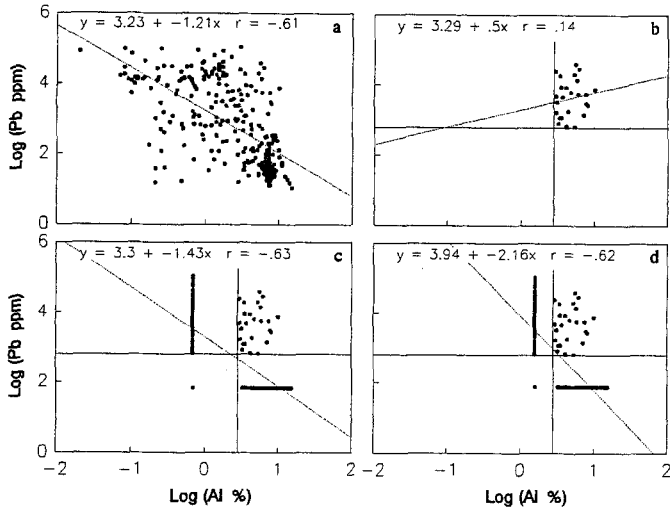

Fig. 8. Comparison of correlation between two variables using various substitution methods: (a) Original uncensored data, (b–d) same data assuming 50% censoring, i.e., a censor point of 0.4523 for Al and 2.812 for Pb, (b) uncensored data only (censored data ignored), (c) censored data replaced by empirical, actual mean of censored data, (d) censored data replaced by mean estimated using Cohen's method and our Eq. 7.

dataset, should be tested using other datasets. Because R-mode factor analysis depends on correlation coefficients, the results of R-mode results should have similar confidence as the correlation coefficient analysis. However, this assumption also deserves further verification.

Censoring at High Concentrations

Because of the smaller number of elements suitable for testing in our data set and the fact that censoring at high concentrations is less common than censoring at low concentrations (Miesch, 1967), we present only a brief analysis of "greater than" replacements. The mean of the concentrations of each element is reproduced closely by both the maximum likelihood method and $\frac{4}{3}$ times the detection limit for our dataset. Skewness is consistently better by the maximum likelihood method. The maximum likelihood method consistently reproduces the replacement factor better than the simple substitution method, and the difference between the methods increases with more replacements (Fig. 9). The empirically-determined replacement factor increases with more replacements; for example, the mean of the replacement values for the five elements increases from 0.18 in $\log_{10}$ units (1.5 in original units) at 5% replacement to 0.4 in $\log_{10}$ units (2.5 in original units) at 30% replacement. The maximum likelihood method

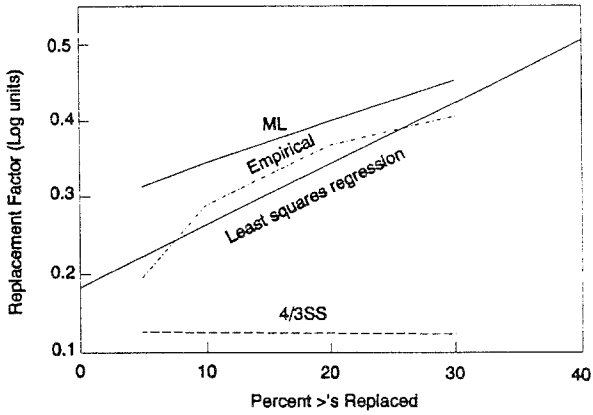

Fig. 9. Comparison of maximum likelihood (ML) and $\frac{4}{3}$ simple substitution methods ($\frac{4}{3}$SS) with empirical replacement factors for "greater than" replacements. Note that the $\frac{4}{3}$ simple substitution method consistently underestimates the factor, and that the factor increases with more replacements.

correctly predicts this increase in replacement factor, however it tends to overestimate the replacement factor.

Correlation coefficients for the ten pairs of elements are equally good for the maximum likelihood and simple substitution methods (Fig. 10). The root mean square error is slightly better for the maximum likelihood method (Fig. 10a), however, the number of correlations having the correct significance is slightly better by the simple substitution method (Fig. 10b). In all cases, the uncensored data alone yields significantly worse results.

Theoretical analysis of "greater thans" is similar to that of "less thans." One additional fact that must be considered is that the sum of the concentrations is physically limited on the high end. The limit in original units is 100% or 1 million parts per million. In log units, the corresponding limits are 2 and 6, respectively. This limitation is usually not serious for "less thans," which typically have values much less than the physical limit and, since $\log_{10}(0) = -\infty$, theoretically have no lower limit when expressed in log units. However, as x and y approach their upper physical limits, as they commonly do for "greater thans," the fit of the normal distribution will be more approximate, because the right tail extends indefinitely to the right of the physical limit.

Briefly, without presenting the detailed derivation, if y is a normal random variable, the expected value or mean for "greater thans" is given by

$$\mu_r = \mu + \frac{\sigma}{\sqrt{2\pi}} \left\{ \frac{\exp\left[-\left(\frac{z_d^2}{2} \right) \right]}{1 - \Phi(z_d)} \right\} \tag{21}$$

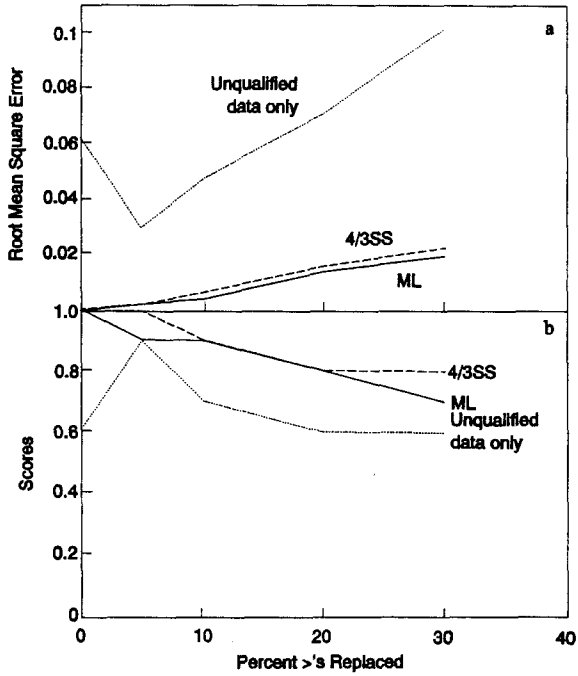

Fig. 10. Summary of performance in estimating the correlation coefficient by the three methods for "greater than" replacements, (a) root mean square error, (b) numerical scores (see text for explanation of score). Correlation coefficients from unqualified data only are unreliable. The maximum likelihood method (ML) yields slightly better rmse than the $\frac{4}{3}$ simple substitution method ($\frac{4}{3}$SS), but the scores are slightly better for the $\frac{4}{3}$ simple substitution method.

which is analogous to Eq. (13) for "less thans." The resulting expression for the replacement factor is then

$$r_y = \frac{\sigma}{\sqrt{2\pi}} \left\{ \frac{\exp\left[-\left(\frac{z_d^2}{2} \right) \right]}{1 - \Phi(z_d)} \right\} - z_d\sigma \tag{22}$$

which is analogous to Eq. (19) for "less thans." All the other relations can be derived following the steps used for "less thans." Equation (22) is illustrated graphically in Fig. 11, which is analogous to Fig. 6. This equation shows that for "greater than" replacements, the replacement factor increases with more replacements and with larger standard deviation, and the mean does not affect

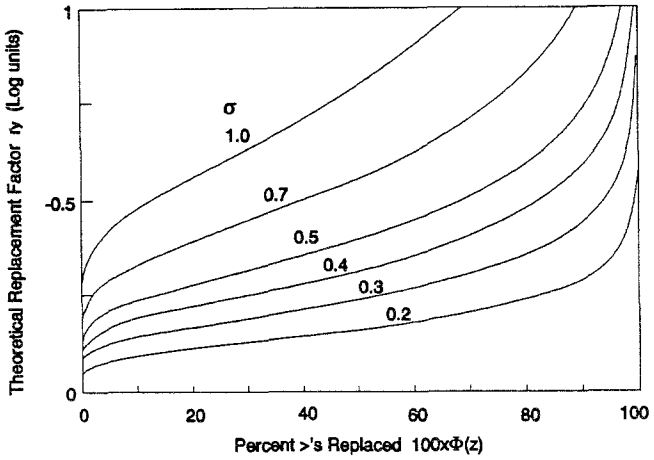

Fig. 11. Plot of theoretical replacement factor for "greater thans" (Eq. 22) vs. proportion of replacements or cumulative distribution function (Eq. 11) contoured for standard deviation. Note that form of the curves mimics that of the empirical curve in Fig. 9.

the relationship between the replacement factor and the proportion of replacements, in agreement with the empirical data (Fig. 9).

RECOMMENDATIONS

Based on our theoretical analysis and empirical testing on one dataset, we recommend the following procedure when preparing a censored dataset for correlation coefficient and factor analysis. We assume the data have been converted to $\log_{10}$ units. For each element:

1. If the data are censored on one end, then skip to step 2. If the data are doubly censored, determine whether the "less thans" or the "greater thans" have less effect on the mean. This can be done approximately by multiplying the number of censored values times the respective detection limit. Typically the lower values will have less effect on the mean. Replace these values with an empirical replacement factor ($\log_{10}$ units) plus the detection limit ($\log_{10}$ units). An investigator can determine reasonable empirical replacement values for his own data if some elements in his dataset are uncensored. The relationship between replacement factor and percent replacements can be determined by setting artificial detection limits and calculating the mean of the values below these limits, as we described above. If this is unfeasible, an empirical replacement factor may be read from the curve for the mean in Fig. 3 or 9, which are based on our dataset; for example, with 10% "less thans," a replacement factor of

−0.3 (equivalent to 0.55 in original units); or with 10% "greater thans," a replacement factor of 0.25 (equivalent to 1.8 in original units). The maximum allowable estimated replacement value for "greater thans" is 100% or 1 million parts per million. Now the data can be treated as though it is censored on one end only.

2. Determine the replacement factor using the method for either singly or multiply-censored data (Cohen, 1959, 1961, 1976). Substitute the estimated mean of the whole dataset from Cohen's method, μ, the mean of the unqualified data, μ_u, and the number of qualified, unqualified and total samples, n_q, n_u, and n, into Eq. (7) to obtain the estimated replacement value, μ_q, which is the estimated mean of the qualified data.

3. Create one data set with elements having as much as 40% replacements. Create another set with elements having as much as 80% replacements. Our results suggest that the set having as much as 40% replacement will yield results that are accurate at the 90% confidence level; for as much as 80% replacement, the confidence level is about 60%.

4. Perform correlation coefficient calculations, factor analysis, etc. using the replaced data.

CONCLUSIONS

We propose that an objective method based on maximum likelihood is superior to an arbitrary simple-substitution method for replacement of censored data. One such maximum likelihood method (Cohen, 1961) reproduces the mean and skewness as well or better than a simple substitution method using $\frac{3}{4}$ of the lower detection limit or $\frac{4}{3}$ of the upper detection limit. Theoretical analysis and empirical testing on a typical dataset suggest that this method has general applicability as an objective replacement method. For a small proportion of "less than" substitutions, a simple-substitution replacement factor of 0.55 is preferable to $\frac{3}{4}$; for a small proportion of "greater than" substitutions, a simple-substitution replacement factor of 1.7 is preferable to $\frac{4}{3}$. For more than 10% replacement, a mean empirical factor may be used (Figs. 3 and 9). Theoretical and empirical analysis shows that the replacement factors for "less thans" decrease with more replacements and larger standard deviation; those for "greater thans" increase with more replacements and larger standard deviation. The maximum likelihood method reproduces these variations, whereas any simple substitution method assumes a fixed replacement factor. Thus the errors induced by simple substitution are increasingly severe with more replacements and greater standard deviation. Correlation coefficients are also computed more accurately using our adaptation of the maximum likelihood method. Correlation coefficients should never be computed from the unqualified data only. Using the maximum likelihood method for replacing "less thans" in our sample dataset, correlation

coefficients were reasonably accurately estimated in 90% of the cases for as much as 40% replacement and in 60% of the cases for 80% replacement. We recommend that other datasets be tested to determine better the robustness of the maximum likelihood method and the sensitivity of the correlation coefficients to variations in the replacement value.

Our observations compare favorably with the results of Helsel and Cohn (Cohn, 1988; Helsel and Cohn, 1988), who concluded that $\frac{1}{2}$ times the detection limit is better than other simple substitution factors for estimating descriptive statistics, but that the maximum likelihood method is superior to any simple substitution method. We have extended their analysis by showing that the replacement factor varies with both the proportion of replacements and the standard deviation, and by demonstrating that the correlation coefficients are better estimated using an adaptation of the maximum likelihood method rather than a simple substitution method. Future investigations should test the application of the probability plotting method for the estimation of replacement values. Our method can also be combined with transformations (Miesch, 1967) other than $\log_{10} x$ to better normalize the data and obtain still better statistics.

ACKNOWLEDGMENTS

Discussions with David Grundy, Dennis Helsel, and Charles Spirakis were helpful. Ronald Tidball, Aldo V. Vecchia, Donald Rimstidt, and an anonymous person thoughtfully reviewed the manuscript.

REFERENCES

Chayes, F., 1954, The lognormal distribution of the elements: A discussion: Geochim. et Cosmochim. Acta, v. 6, p. 119–120.

Chayes, F., 1960, On correlation between variables of constant sum: J. Geophys. Res., v. 65, p. 4185–4193.

Cohen, A. C., 1959, Simplified estimators for the normal distribution when samples are singly censored or truncated: Technometrics, v. 1, p. 217–237.

Cohen, A. C., 1961, Tables for maximum likelihood estimates: Singly truncated and singly censored samples: Technometrics, v. 3, p. 535–541.

Cohen, A. C., 1976, Progressively censored sampling in the three parameter log-normal distribution: Technometrics, v. 18, p. 99–103.

Cohn, T. A., 1988, Adjusted Maximum Likelihood Estimation of the Moments of Lognormal Populations from Type I Censored Samples: U.S. Geological Survey, Open-file Report 88-350, 34 p.

Crow, E. L., Davis, F. A., and Maxfield, M. W., 1960, Statistics Manual: Dover Publications, New York, 288 p.

De Wijs, H. J., 1951, Statistics of Ore Distribution: Geologie en Minjbouw, 13E Jaargang Nieuw Serie, p. 365–396.

Gilliom, R. J., and Helsel, D. R., 1986, Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques: Water Res. Res., v. 22, p. 135–146.

Helsel, D. R., and Cohn, T. A., 1988, Estimation of descriptive statistics for multiply censored water quality data: Water Res. Res., v. 24, p. 1997–2004.

Helsel, D. R., and Gilliom, R. J., 1986, Estimation of distributional parameters for censored trace level water quality data, 2. Verification and applications: Water Res. Res., v. 22, p. 147–155.

Korn, G. A., and Korn, T. M., 1968, Mathematical Handbook for Scientists and Engineers: McGraw-Hill, New York, 1130 p.

Krige, D. G., 1951, A statistical approach to some basic mine valuation problems on the Witwatersrand: J. Chem. Metall. Mining Soc. S. Afr., v. 52, p. 119–139.

Krige, D. G., 1960, On the departure of ore value distributions from lognormal models in South African gold mines: J. S. Afr. Inst. Mining Metall., v. 61, p. 231–244.

Miesch, A. T., 1967, Methods of Computation for Estimating Geochemical Abundance: U.S. Geological Survey Professional Paper 574-B, 15 p.

Miesch, A. T., 1976a, Sampling designs for geochemical surveys—Syllabus for a short course: U.S. Geological Survey Open-File Report, 76-772, 140 p.

Miesch, A. T., 1976b, Geochemical survey of Missouri—Methods of sampling, laboratory analysis and statistical reduction of data: U.S. Geological Survey Professional Paper 954-A, 39 p.

Miesch, A. T., 1982, Estimation of the geochemical threshold and its statistical significance: J. Geochem. Exp., v. 16, p. 77–104.

Miesch, A. T., and Riley, L. B., 1961, Basic statistical methods used in geochemical investigations of Colorado Plateau uranium deposits: HIMMP Trans. (Mining), v. 220, p. 247–251.

Sanford, R. F., Korzeb, S. L., Seeley, J. L., and Zamudio, J. A., 1987a, Geochemical Data for Mineralized Rocks in the Lake City Area, San Juan Volcanic Field, Southwest Colorado: U.S. Geological Survey, Open-file Report 87-54, 213 p.

Sanford, R. F., Grauch, R. I., Hon, K., Bove, D. J., and Grauch, V. J. S., 1987b, Mineral resources of the Redcloud Peak and Handies Peak Wilderness Study Areas, Hinsdale County, Colorado: U.S. Geological Survey, Bulletin 1715-B, 40 p.

Sichel, H. S., 1952, New methods in the statistical evaluation of mine sampling data: London, Inst. Mining and Metall. Trans., v. 61, p. 261–288.

Sichel, H. S., 1966, The estimation of means and associated confidence limits for small samples from lognormal populations: J. S. Afr. Inst. Mining and Metall., Symposium: Mathematical Statistics and Computer Applications in Ore Valuation, p. 106–123.

VanTrump, G., Jr., 1977, The U.S. Geological Survey RASS-STATPAC system for management and statistical reduction of geochemical data: Comp. Geosci., v. 3, p. 475–488.