Inference of distributional parameters of compositional samples containing nondetects

Ricardo A. Olea[1]

[1]US Geological Survey, Reston, USA; *olea@usgs.gov*

Abstract

Low concentrations of elements in geochemical analyses have the peculiarity of being compositional data and, for a given level of significance, are likely to be beyond the capabilities of laboratories to distinguish between minute concentrations and complete absence, thus preventing laboratories from reporting extremely low concentrations of the analyte. Instead, what is reported is the detection limit, which is the minimum concentration that conclusively differentiates between presence and absence of the element. A spatially distributed exhaustive sample is employed in this study to generate unbiased sub-samples, which are further censored to observe the effect that different detection limits and sample sizes have on the inference of population distributions starting from geochemical analyses having specimens below detection limit (nondetects). The isometric logratio transformation is used to convert the compositional data in the simplex to samples in real space, thus allowing the practitioner to properly borrow from the large source of statistical techniques valid only in real space. The bootstrap method is used to numerically investigate the reliability of inferring several distributional parameters employing different forms of imputation for the censored data. The case study illustrates that, in general, best results are obtained when imputations are made using the distribution best fitting the readings above detection limit and exposes the problems of other more widely used practices. When the sample is spatially correlated, it is necessary to combine the bootstrap with stochastic simulation.

Key words: Detection limit, censored data, statistic, bootstrap, correlation, probability distribution.

1 Introduction

Geochemistry is becoming increasingly important in the earth sciences and engineering because of the success that laboratory analyses are having for providing clues about past processes and helping to assess the future in environmental studies. Considering that geochemical data in general are a minute sample of large systems, statistical rather than deterministic analysis should be considered in trying to infer any of the characteristics of the parent population. Proper statistical analysis of geochemical data is different from the analysis of other types of data because of the often simultaneous influence of the following complicating factors:

- There is no technology to detect analytes to the level of a single molecule. Usually instruments cannot properly distinguish noise from true readings for all concentrations above the resolution of interest. This limitation results in nondetects, or values intermediate between indisputable absence and a limit of detection reported by the laboratory.
- Relative abundance of analytes is typically reported as concentrations, which are conventional ways to denote contribution to a whole divided into components adding to a constant, customarily one hundred or a million (Pawlowsky-Glahn et al., 2006). This additivity of analytical concentrations to a constant makes them a type of variables known as compositional data comprising spuriously correlated components varying in the simplex.
- Finally, specimens in a geochemical analysis commonly have an associated geographical location. When analyzed as a whole, often close to a small value there is another small value, and another rather large value is rather the rule than the exception near a large value. When that is the case, it is said that the values are spatially correlated.

The analysis and modelling of spatially correlated data are the subject of geostatistics (e.g. Chilès and Delfiner, 1999; Olea, 1999).

Although John Aitchison was not the first one to publish in the subject of data adding to a constant, he is the author of the book providing the best general treatment on the subject (Aitchison, 1989, 2003). Two collections of papers containing more recent developments are those by Pawlowsky-Glahn (2005) and Buccianti et al. (2006).

For years, researchers at the University of Girona in Spain have been about the only ones dealing with compositional data and problems derived from analytical resolution, but they have been working in a different direction than estimation of distributional parameters. Logarithmic transformations, such as the one in Equation 2, are central to processing of compositional data. When nondetects are recorded as rounded zeroes, they prevent taking logarithms when the rounded zeroes are in the numerator and perform the operation of division when the rounded zeroes go into the denominator. Several replacement strategies using small values have been proposed conditioned to not disturbing the covariance matrix of the composition, strategies that work satisfactory well when the proportion of zeroes is low, say, no more than 10% (Martín-Fernández et al. 2003; Martín-Fernández and Thió-Henestrosa, 2006; Howel, 2007). Palarea-Albalarejo et al. (2007) have recently released a more general approach to rounded zeroes that is a parametric formulation based on the expectation-maximization algorithm, a form of maximum likelihood.

One of the most complete statistical treatments of nondetects in the earth sciences is the one by Helsel (2005), but his book does not go into the compositional character of the data or the relevance of the geographical location of the specimens sent to the laboratories, topics that so far have been sparingly addressed. There have been a few papers devoted to stochastic assessment in connection to environmental remediation mapping, but none treat the surveys as compositional data or deal with estimation of population distributional parameters. Pardo-Igúzquiza and Chica-Olmo (2005) do their mapping based on the estimation method of empirical maximum likelihood kriging (Pardo-Igúzquiza and Dowd, 2005); Rathbun (2006) applies a Robbins-Monroe algorithm; Fridley and Dixon (2007) employ the data augmentation method combined with a Markov chain Monte Carlo algorithm.

One of the problems with the correct treatment of data with nondetects has been the reluctance of practitioners to analyze the data according to the best available methods (Huybrechts et al., 2002), even after disregarding closure and spatial correlation. Thus, a first objective of this paper is to emphasize once more that, even under the simplest of the circumstances, not all ways to extract information from geochemical samples with nondetects are equally effective. Part of the problem to convince practitioners to abandon certain methods is that there are no general formulas describing how much better one method is relative to another. Geochemists prepare standard samples to test analytical methods; statisticians have synthetically generated data for similar purposes. As long as the general characteristics of a synthetic dataset follow those of real systems, not much is lost in terms of not dealing with actual measurements coming from the real world. The big advantage of synthetic samples, of course, is to know all answers.

The main objective of this study is the preparation of the distribution for parameters customarily used to infer the characteristics of a population, inference that is necessarily uncertain because of the size of the sample and the existence of nondetects. Trying to avoid cluttering the exposition, the study starts by analyzing two samples with one detection limit and without regard to the sampling location, but the compositional character of the analyte is taken into account from the outset. These restrictions are removed in the second part of the exposition by analyzing multiple detection limits and the effect of spatial correlation for the best method to handle nondetects according to the analysis in the first part.

PART A ANALYSIS OF SAMPLES WITH A SINGLE DETECTION LIMIT, IGNORING SAMPLING LOCATION

2 Methodology

2.1 Isometric logratio transformation

Compositional data involving *D* variables adding to *C* define the simplex S^D,

$$S^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D]; \quad x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = C \right\}, \tag{1}$$

space where the standard methods of statistics do not apply. Several transformations have been formulated over the years to bring the data into a real space and then be able to apply classical multivariate statistics. Here, I employ the isometric logratio (ilr) transformation characterized by relating angles and distances in the simplex to angle and distances in the real space (Egozcue et al., 2003).

I concentrate the attention to one attribute X_1, amalgamating all the others into X_2. In such case, the ilr transformation is:

$$ILR = \frac{1}{\sqrt{2}} \log \frac{X_1}{X_2}. \tag{2}$$

All calculations will be performed in the ilr space, backtransforming final results when necessary.

2.2 Simple bootstrap method

The main interest of this method is the calculation of population distributional parameters from a sample and their associated uncertainty, which comes from the fact the sample size is smaller than the population size. In general, one possible way to associate levels of uncertainty to the estimation of a parameter is by finding the distribution for all possible unbiased samples of the same size as the sample at hand. One indirect way of doing it avoiding the actual collection of multiple samples is to generate synthetic samples from the only empirical sample actually available. This is the aim of the bootstrap method (Diaconis and Efron, 1984; Chernick, 2008). The general steps are:

1. Collect an empirical sample of size *N*.
2. Randomly pick *N* measurements from the empirical sample. Some values may be taken more than once, while others not at all.
3. Calculate and save as many statistics for the synthetic sample generated by the bootstrap resampling.
4. Go back to step 1 and repeat the process at least 1,000 times.
5. Stop

2.3 Nondetect bootstrap method

Simple bootstrap is not applicable to samples containing nondetects because it does not take into account a second source of uncertainty: nondetects are not accurate measurements, but instead fuzzy information within an interval. Essential to model that uncertainty in Step 2 is the imputation of a value. Thus the modified steps to account also for the compositional character of the data are:

1. Collect an empirical sample of size *N*.
2. Randomly pick *N* values from the empirical sample. Some values will be above detection limit, some others below.
3. Replace all nondetects by imputing a likely value.
4. Make an ilr transformation of all values.
5. Calculate and save as many statistics of interest for each bootstrap resample.
6. Go back to step 1 and repeat the process at least 1,000 times.
7. Backtransform results.
8. Prepare summaries.
9. Stop

Several imputation criteria have been employed over the years for obtaining distributional parameters. Among the most popular ones, here I consider the following:

- Assign a value that is a fraction of the detection limit.
- Draw at random a value from a uniform distribution between zero and the detection limit.
- Draw at random a value from the tail below the detection limit of a distribution best fitting the values above detection limit.

The first option is by far the one followed most commonly, despite repeated warning from a minority of practicians to stay away from it because of lack of justification and poor results unless the number of nondetects is a minute fraction of the sample size (Helsel, 2006). The temptation of the practice is clearly its simplicity. Two of the most common options are to substitute nondetects by 0.5 or 0.7 times the detection limit (Lee and Helsel, 2005). An important objective of this study is to join efforts to stop the practice of substitution by showing the dangers of the approach through comparison to true answers.

The general idea of the other two imputation methods is to randomly replace the nondetects with values drawn from a probability distribution. We will see that arbitrary selection of the uniform distribution can provide better results than substitution. Again, the main appeal of the uniform distribution is simplicity because the parameters are automatically determined by the condition that the values have to be positive and below the detection limit. Systematically good results are obtained only when the distribution models the actual tail of the distribution below detection limit, distribution that is rarely uniform.

Most methods attempting to model the actual distribution of data below detection limit are generalizations of maximum likelihood methods originally formulated to analyze data without nondetects. The generalization involves factoring the likelihood function into two parts, one depending on the detection limit and the other being the usual term involving the actual measurements (Huybrechts et al., 2002). Some of these methods have their origins in the related field of statistics of missing data (Little and Rubin, 2002).

3 Data

3.1 Exhaustive sample

The study is based on a synthetically generated dataset that is supposed to exhaust possibilities of taking measurements (Deutsch and Journel, 1998). As seen in Figure 1, this exhaustive sample contains 2500 measurements that closely follow a lognormal distribution, which is the distribution most closely followed by geochemical analytes (Helsel, 2005). As the attribute X_1 is in ppm, the value of X_2 in Equation 2 is $1000000 - X_1$.

3.2 Empirical samples

Two datasets were extracted from the exhaustive sample to emulate two unbiased samples of significantly different sizes (Fig. 2–3).

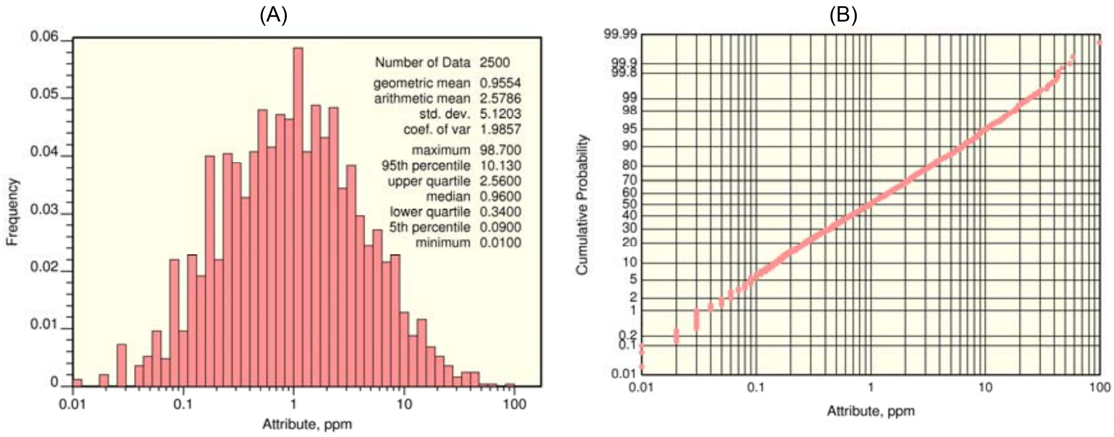

Figure 1: Exhaustive sample: (A) histogram; (B) cumulative distribution.

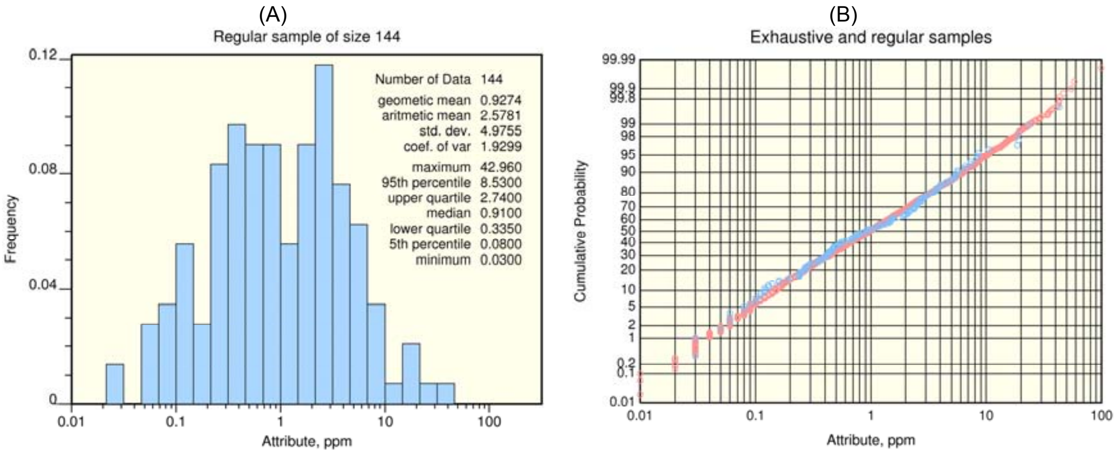

Figure 2: Empirical sample of size 144: (A) histogram; (B) cumulative distribution for the sample of size 144 in blue and for the exhaustive sample in orange.

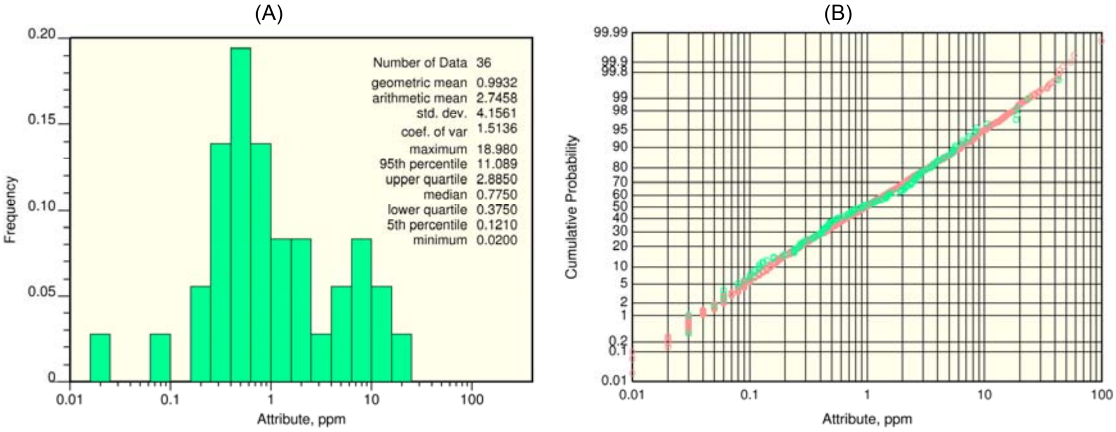

Figure 3: Empirical sample of size 36: (A) histogram; (B) cumulative distribution for the sample of size 36 in green and for the exhaustive sample in orange.

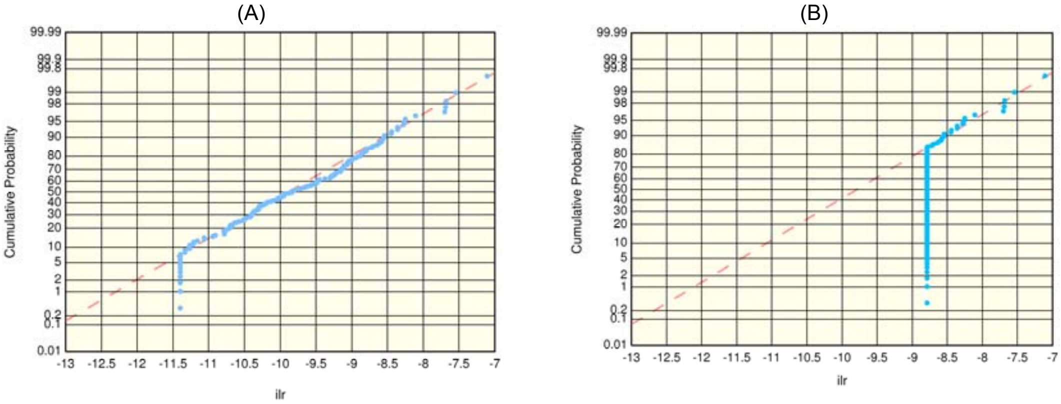

Figure 4: Sample of size 144 after censoring, showing as a segmented line the best model fitting the data about the detection limit, which is denoted by the vertical alignment of points: (A) detection limit of 0.1 ppm; (B) detection limit of 4 ppm.

4 Estimation of parameters

The two empirical samples of sizes 144 and 36 were artificially censored to investigate the efficiency of the three imputations methods in Section 2.3 for estimating five distributional parameters by the nondetect bootstrap method in the same section. These parameters are: geometric mean, 5th percentile, lower quartile, median, and upper quartile. The selected detection limits were 0.1, 0.4, 1, and 4 ppm. They were arbitrarily selected trying to generate four significantly different proportions of nondetects.

4.1 Estimation using the sample of size 144

Figure 4 shows the cumulative distribution for the sample of size 144 after censoring according to detection limits of 0.1 and 4 ppm. With only 6% of the values censored by the 0.1 ppm detection limit, Figures 5 and 6 show that any imputation method provides good results in the sense that the center of the distribution is not far from the true value, even though the distributions for the 5th percentiles for the substitution approach have erratic increases. No estimation was prepared for the other parameters because they do not feel the effect of the censoring. Remember that all calculations are done in the ilr space.

Figures 7–8 render the results for the estimation of the fifth percentile and the upper quartile for the highest detection limit of 4 ppm. There is now significant discrepancy in the effectiveness of the imputation methods. The clear winner is the approach of drawing from the lower tail of a distribution fitted to the data about detection limit, distribution that in our case is a normal distribution of the ilr values, which overall represents a lognormal model of the original attribute.

Backtransformed, complete summaries for all four detection limits are in Tables 1–4. Estimations for quantiles above the detection limit are systematically omitted for being insensitive to the censoring. In these tables, a value in red denotes the worst estimator for the given sample size and detection limit relative to exhaustive sample parameter. A blue value, on the contrary, indicates the best estimate.

Results clearly show the drawbacks of the substitution method when all values below a quantile are nondetects:

- The distribution for the parameter is a Dirac delta function, implying certainty about the parameter under estimation.
- The value of the estimate is equal to the substitution value, thus it can be right only when the parameter is equal to the substitution value.
- All quantiles for a proportion smaller than the fraction of nondetects are equal, which is seldom the case in actual practice. For example, in Table 4, the fifth percentile, the lower quartile and the median are equal to 2.0 when the decision is to substitute the nondetect by half it value, and they change to 2.8 if the rule is replacement by 0.7 times the detection limit.

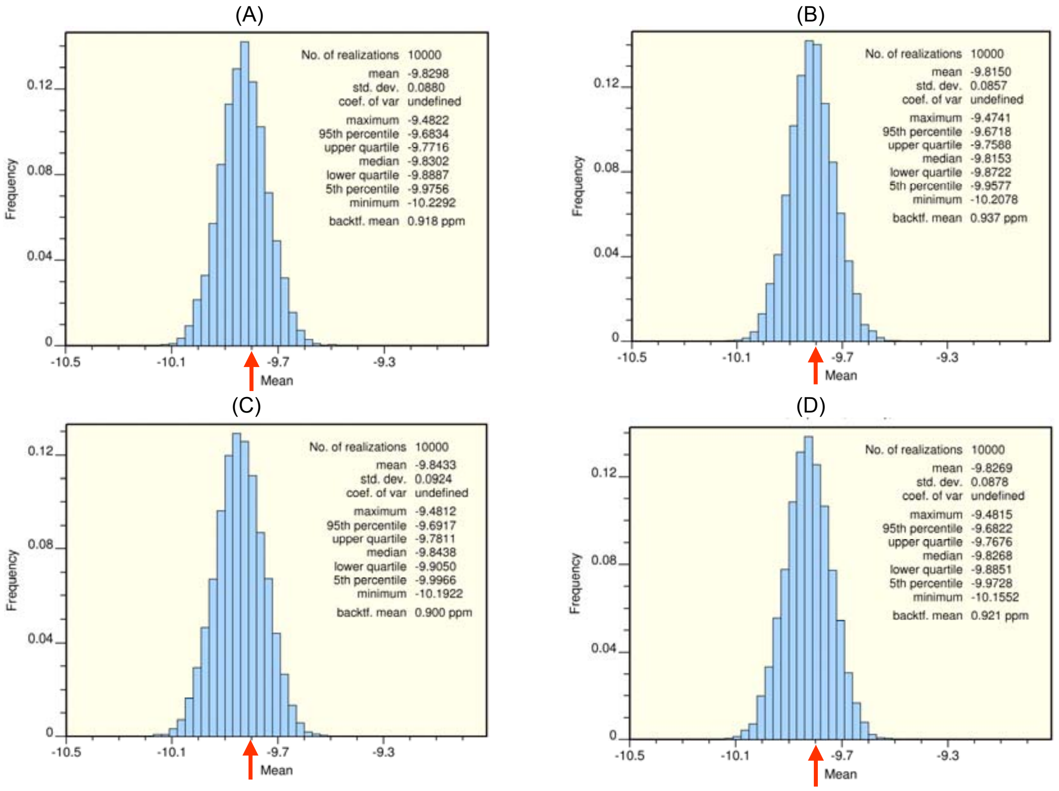

Figure 5: Mean according to four imputation methods using the sample of size 144 and a detection limit of 0.1 ppm, with the arrow denoting mean for the exhaustive sample: (A) substitution of nondetects by 0.5 times the detection limit; (B) substitution by 0.7 times the detection limit; (C) imputation using uniform distribution; (D) normal imputation of ilr values.

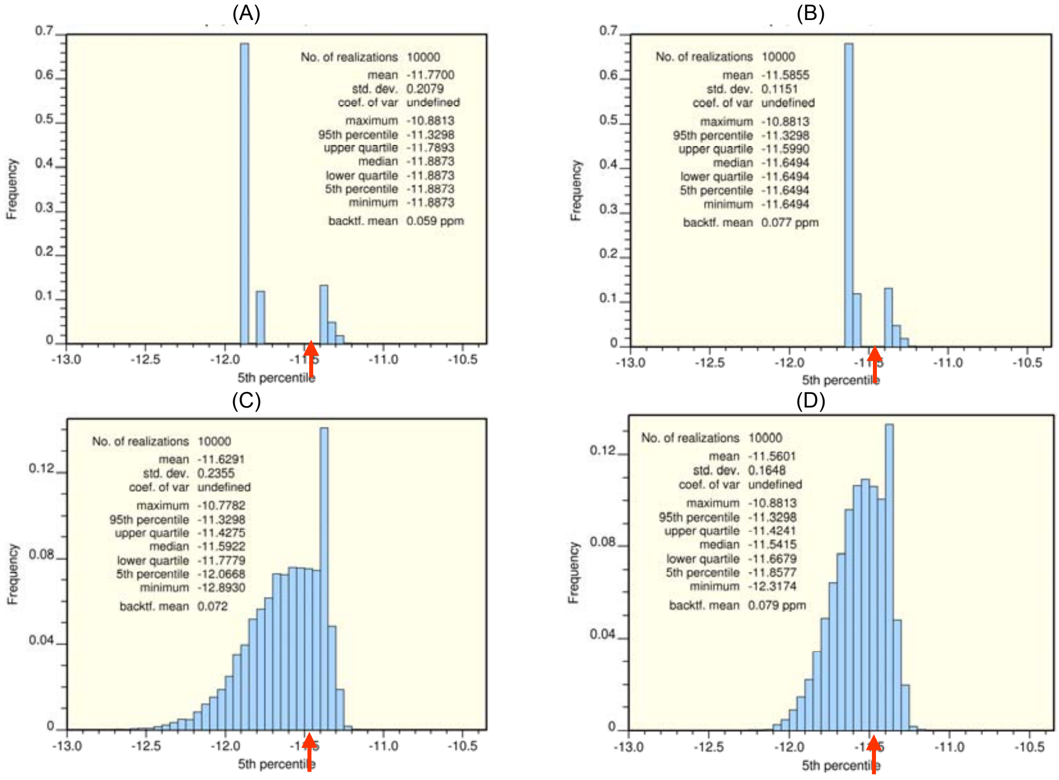

Figure 6: Fifth percentile according to four imputation methods using the sample of size 144 and a detection limit of 0.1 ppm, with the arrow denoting the fifth percentile for the exhaustive sample: (A) substitution of nondetects by 0.5 times the detection limit; (B) substitution by 0.7 times the detection limit; (C) imputation using uniform distribution; (D) normal imputation of ilr values.

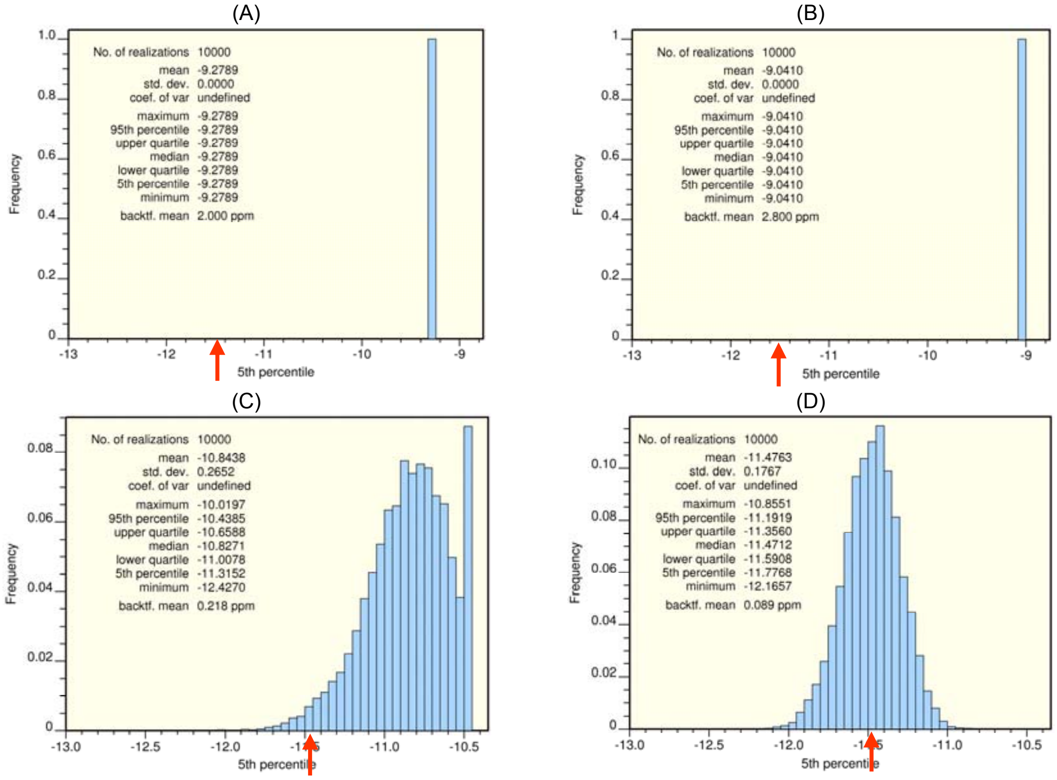

Figure 7: Fifth percentile according to four imputation methods using the sample of size 144 and a detection limit of 4 ppm, with the arrow denoting the same percentile for the exhaustive sample: (A) substitution of nondetects by 0.5 times the detection limit; (B) substitution by 0.7 times the detection limit; (C) imputation using uniform distribution; (D) normal imputation of ilr values.

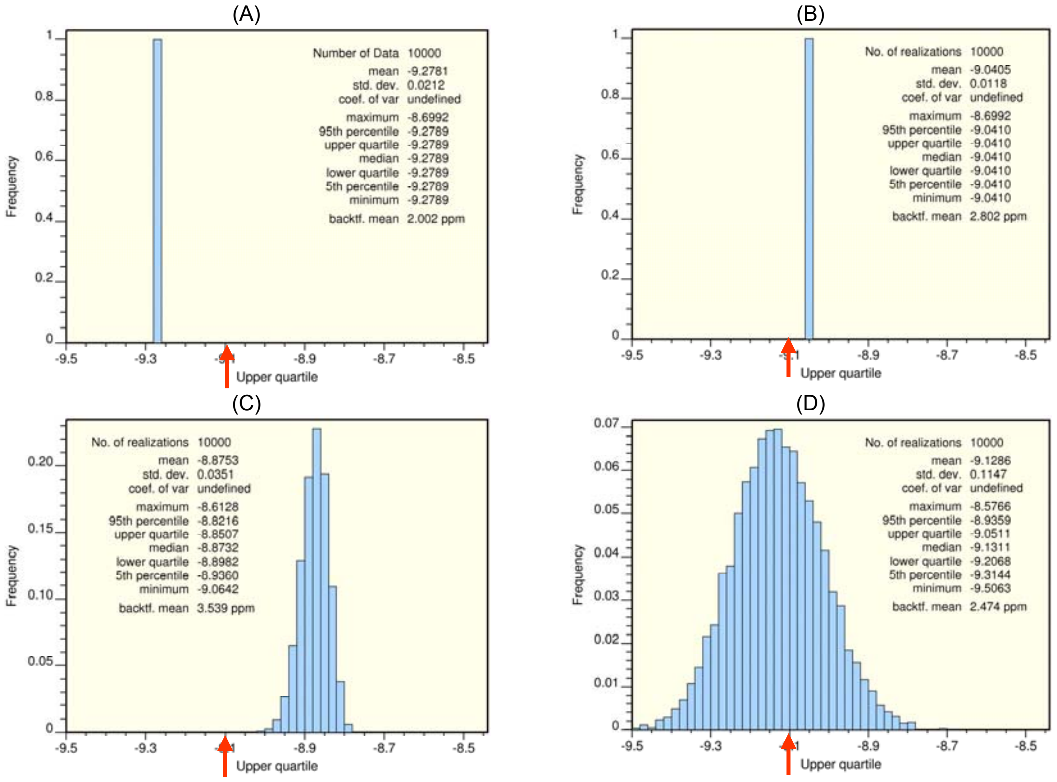

Figure 8: Upper quartile according to four imputation methods using the sample of size 144 and a detection limit of 4 ppm, with the arrow denoting upper quartile for the exhaustive sample: (A) substitution of nondetects by 0.5 times the detection limit; (B) substitution by 0.7 times the detection limit; (C) imputation using uniform distribution; (D) normal imputation of ilr values.

Table 1. Parameter estimation with sample of size 144 and detection limit of 0.1 ppm, which results in 9 nondetects (6%).

	Geo. mean	*p*5
Exh. sample	0.955	0.090
0.5**DL*	0.918	0.059
0.7**DL*	0.937	0.077
Uniform	0.900	0.072
Lognormal	0.921	0.079

Table 2. Parameter estimation with sample of size 144 and detection limit of 0.4 ppm, which results in 40 nondetects (28%).

	Geo. mean	*p*5	*q*1
Exh. sample	0.955	0.090	0.340
0.5**DL*	0.999	0.200	0.232
0.7**DL*	1.097	0.280	0.303
Uniform	0.918	0.066	0.351
Lognormal	0.954	0.087	0.346

Table 3. Parameter estimation with sample of size 144 and detection limit of 1 ppm, which results in 75 nondetects (52%).

	Geo. mean	*p*5	*q*1	*q*2
Exh. sample	0.955	0.090	0.340	0.960
0.5**DL*	1.208	0.500	0.500	0.631
0.7**DL*	1.479	0.700	0.700	0.808
Uniform	1.058	0.088	0.471	0.983
Lognormal	0.996	0.092	0.366	0.943

Table 4. Parameter estimation with sample of size 144 and detection limit of 4 ppm, which results in 121 nondetects (84%).

	Geo. mean	*p*5	*q*1	*q*2	*q*3
Exh. sample	0.955	0.090	0.340	0.960	0.960
0.5**DL*	2.501	2.000	2.000	2.000	2.000
0.7**DL*	3.319	2.800	2.800	2.800	2.802
Uniform	1.931	0.218	1.167	2.350	3.539
Lognormal	0.961	0.089	0.365	0.951	2.474

4.2 Estimation using the sample of size 36

Figure 9 shows the sample of size 36 censored to 4 ppm. Tables 5–8 contain complete results for different censoring of the same sample, which are graphically displayed in Figures 10–11 only for two of the parameters for the largest detection limit of 4 ppm. Comparison of results in Figures 7–8 for the uniform and lognormal distribution with equivalent results in Figures 10–11 substantiates the following remarks:

- As the sample size decreases, the standard deviation for the parameter distribution increases;
- As the number of observations available to do the distribution fitting decreases, bias of the estimator may be significant,

results that are both typical of parameter estimation in general, regardless of the existence of nondetects.

Table 5. Parameter estimation with sample of size 36 and detection limit of 0.1 ppm, which results in 2 nondetects (6%).

	Geo. mean	*p*5
Exh. sample	0.955	0.090
0.5**DL*	0.997	0.079
0.7**DL*	1.016	0.099
Uniform	0.981	0.066
Lognormal	1.001	0.082

Table 6. Parameter estimation with sample of size 36 and detection limit of 0.4 ppm, which results in 9 nondetects (25%).

	Geo. mean	*p*5	*q*1
Exh. sample	0.955	0.090	0.340
0.5**DL*	1.030	0.200	0.287
0.7**DL*	1.121	0.280	0.347
Uniform	0.956	0.054	0.363
Lognormal	0.977	0.072	0.357

Table 7. Parameter estimation with sample of size 36 and detection limit of 1 ppm, which results in 21 nondetects (58%).

	Geo. mean	*p*5	*q*1	*q*2
Exh. sample	0.955	0.090	0.340	0.960
0.5**DL*	1.208	0.500	0.500	0.558
0.7**DL*	1.470	0.700	0.700	0.751
Uniform	1.010	0.058	0.399	0.842
Lognormal	0.903	0.061	0.289	0.740

Table 8. Parameter estimation with sample of size 36 d detection limit of 4 ppm, which results in 121 nondetects (84%).

	Geo. mean	*p*5	*q*1	*q*2	*q*3
Exh. sample	0.955	0.090	0.340	0.960	2.740
0.5**DL*	2.774	2.000	2.000	2.000	2.636
0.7**DL*	3.604	2.800	2.800	2.800	3.375
Uniform	2.182	0.175	1.195	2.463	4.052
Lognormal	1.995	0.300	0.946	1.892	3.760

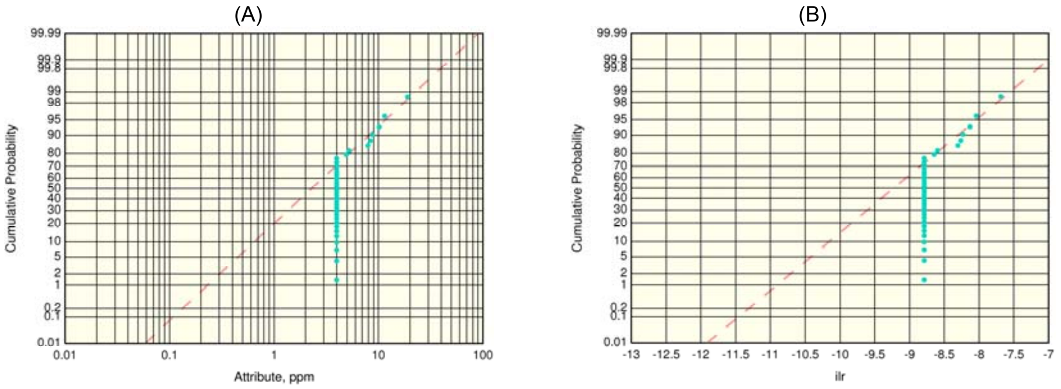

Figure 9: Cumulative distribution for the sample of size 36 after emulating a detection limit of 4 ppm: (A) attribute space; (B) ilr space.

5 Discussion

Table 9 and Figure 12 summarize the relative performance of the four imputation methods: substitution by 0.5 and 0.7 times the detection limits, and random drawing from uniform and lognormal distribution. It is clear that in this study the performance of replacement by 0.7 times the detection limit is undeniably inferior and that of the lognormal fitting is quite superior, to the point that lognormal fitting was the only method never to come up with the worse estimate in the 28 estimations.

Table 10 and Figure 13 illustrate in more detail the problem of bias that can result from a small sample.

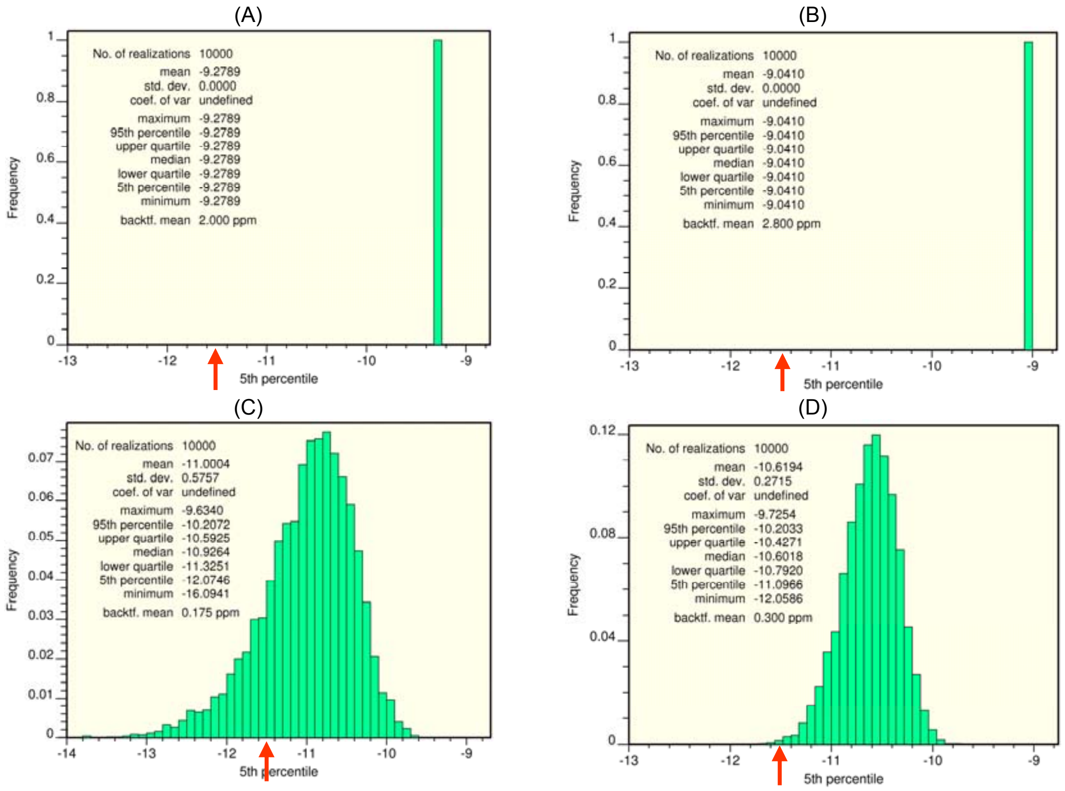

Figure 10: Fifth percentile according to four imputation methods using the sample of size 36 and a detection limit of 4 ppm, with the arrow indicating the fifth percentile for the exhaustive sample: (A) substitution of nondetects by 0.5 times the detection limit; (B) substitution by 0.7 times the detection limit; (C) imputation using uniform distribution; (D) normal imputation of ilr values.

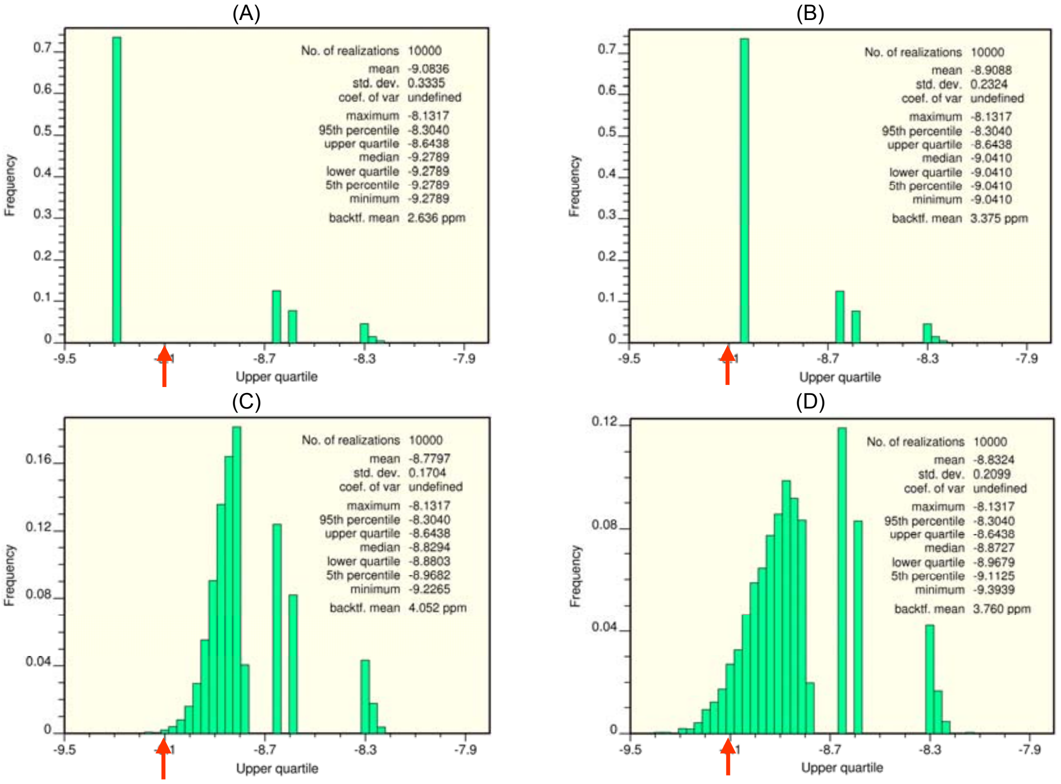

Figure 11: Upper quartile according to four imputation methods using the sample of size 36 and a detection limit of 4 ppm, with the arrow indicating the upper quartile for the exhaustive sample: (A) substitution of nondetects by 0.5 times the detection limit; (B) substitution by 0.7 times the detection limit; (C) imputation using uniform distribution; (D) normal imputation for ilr values.

Table 9. Relative performance of four imputation methods used in nondetect bootstrap.

Method	Worst	Best
0.7**DL*	20	1
0.5**DL*	5	2
Uniform	3	5
Lognormal	0	20

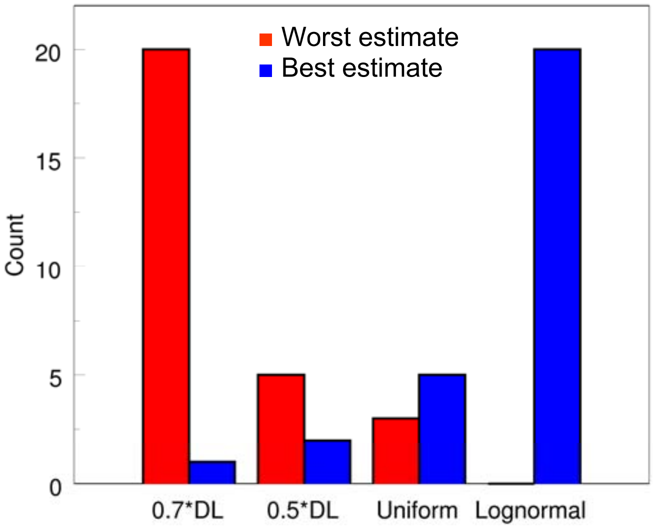

Figure 12: Relative performance of imputation methods.

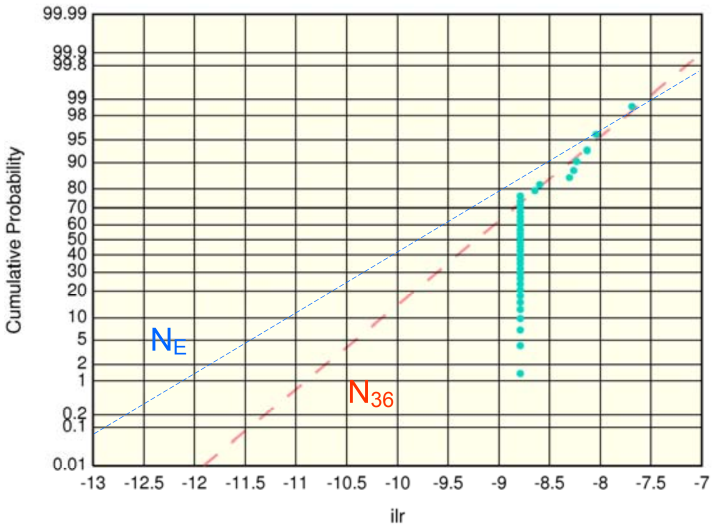

Figure 13: Lognormal imputation for the sample of size 36 and detection limit of 4 ppm. The green dots denote values reported by the laboratories, with those at -8.8 actually being values below detection limit. The red segmented line is the model based on the values above detection limit and the blue line the model based on the exhaustive sample.

Table 10. Parameters of the exhaustive sample and those obtained using lognormal models fitting two different datasets.

	Geo. mean	*p*5	*q*1	*q*2	*q*3
Exh. sam.	0.955	0.090	0.340	0.960	2.74
N_{36}(-9.2,0.77)	1.995	0.300	0.946	1.892	3.76
N_E(-9.75,1.0)	1.185	0.084	0.402	1.105	3.29

The box-and-whisker diagrams in Figure 14 are intended to analyze bias. It can be seen that the distribution for the fifth percentile is slightly biased toward smaller values for detection limits up to 1 ppm. As confirmed by the blue diagram, this negligible bias goes back to the uncensored sample, so it is not an imputation problem. The bias in the result for 4 ppm, however, does break the pattern in terms of significance and sign of the bias.

Figure 15 illustrates an unavoidable law of statistics: all other factors being equal, as the sample size decreases, the uncertainty of the estimator increases.

PART B EXPANDED ANALYSES

It has been established in Part A that fitting a distribution using the information above detection limit is the most efficient way to learn about the parent population. The following sections expand the findings to multiple detection limits and spatially correlated samples. The discussion is restricted to imputation by model fitting.

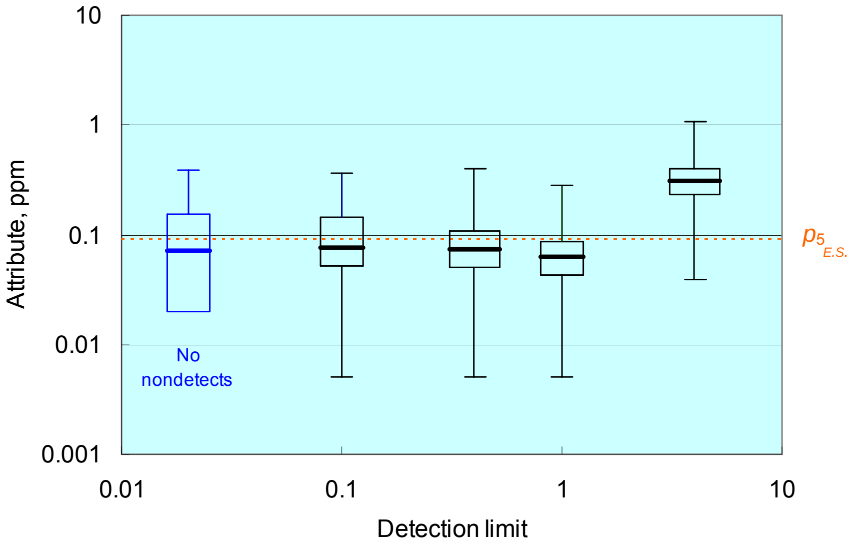

Figure 14: Box-and-whisker diagrams for estimation of the fifth percentile using the sample of size 36 and lognormal imputation.

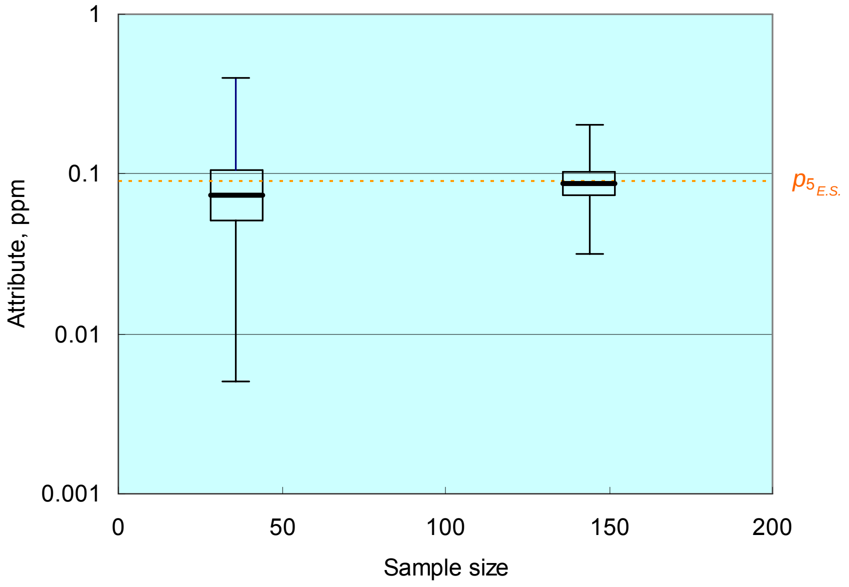

Figure 15: Box-and-whisker representation of distribution of fifth percentile distribution for the two sample sizes used in this study employing lognormal imputation.

6 Multiple detection limits

Figure 16 emulates the values for the sample of size 36 that may result after sending the specimens to two different laboratories with different detection limits. If one keeps track of the laboratories, it is possible to notice a mixture of samples from both laboratories above the largest detection limit and only values from the most precise laboratory below the largest detection limit. For an artificially censored dataset like the one in Figure 16, it is also possible to observe that below a detection limit other than the smallest one, values could be below the other detection limits. In the case of Figure 16, for example, some of the values shown below 4 ppm indeed are below 0.4 ppm.

Fortunately, maximum likelihood can operate for any number of detection limits. Figure 17 shows results for the estimation of four parameters using the sample in Figure 16. As expected, addition of detection limits increases the standard deviation of the estimate, which can be confirmed by comparing the distributions in Figures 17A and 17C with those in Figures 10D and 11D. In terms of bias, as shown in Table 11, the situation is intermediate to those considering the two detection limits separately. None of the parameters obtained using the sample in Figure 16 has the worst bias and the estimate for the fifth percentile is better than when the censoring is done separately.

7 Spatial correlation

The exhaustive sample characterizes an attribute actually varying in a two-dimensional space, but so far, for rhetorical reasons, the influence of sample location has been postponed trying to clarify the influence of other factors first and also because proper handling of specimen location is more demanding and not always necessary, as will be shown below.

Table 11. Distributional parameters for sample of size 36 and detection limits of 0.4 and 4 ppm.

	Geo. mean	*p*5	*q*1	*q*2	*q*3
Exh. sam.	0.955	0.090	0.340	0.960	2.74
D. limit 0.4	0.977	0.072	0.357	0.758	2.608
D. limit 4.0	1.995	0.300	0.946	1.892	3.760
D. lim. 0.4 & 4	1.274	0.077	0.484	1.262	3.362

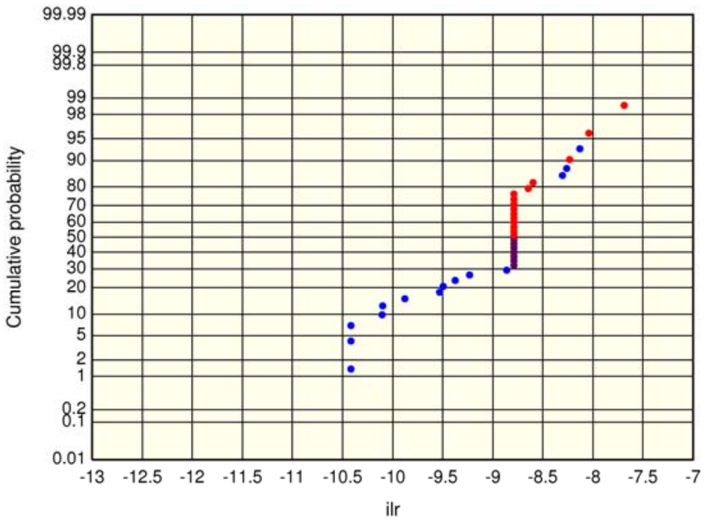

Figure 16: The sample of size 36 censored to detection limits of 0.4 and 4 ppm. The 14 specimens sent to the more precise laboratory are denoted by blue dots. Purple dots indicate specimens sent to the less precise laboratory that would have remained below detection limit if they had been sent to the more precise laboratory. The rest of the 22 specimens are posted as red dots.

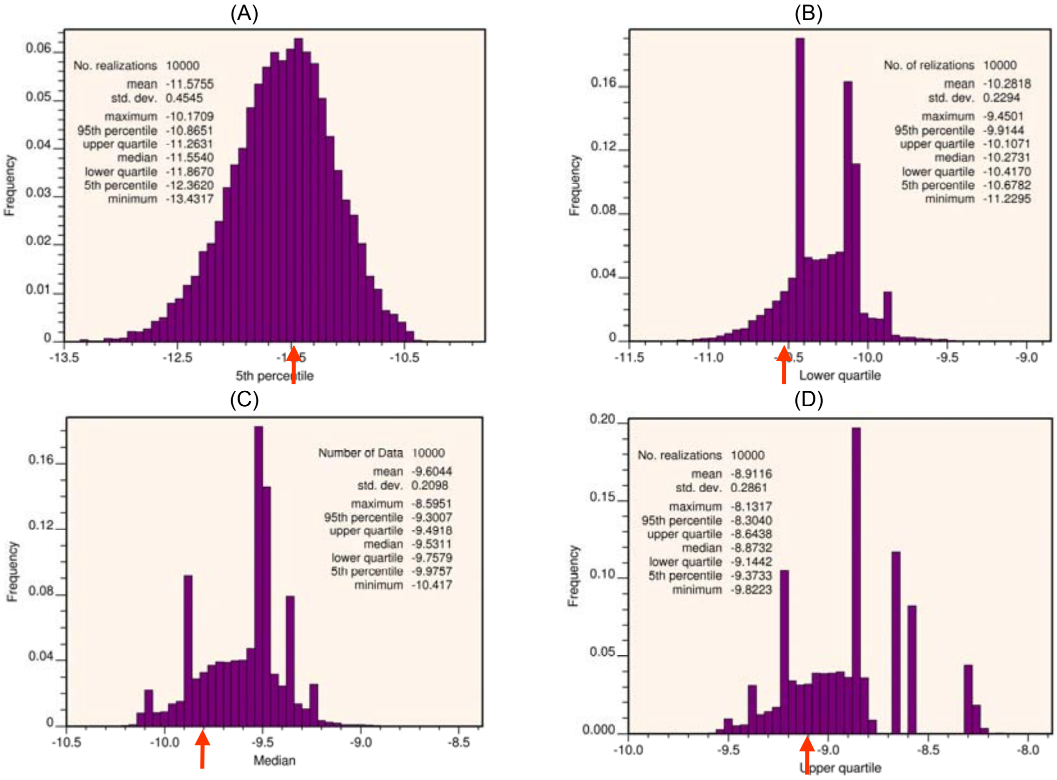

Figure 17: Four parameter distributions for the sample in Figure 16, with the arrow indicating the true value according to the exhaustive sample: (A) fifth percentile; (B) lower quantile; (C) median; (D) upper quartile.

7.1 The samples revisited

A map is the conventional way to graphically display an attribute with geographical distribution. Figures 18–19 show that the samples of size 36 and 144 indeed have a geographical distribution and how they were extracted from the exhaustive sample.

7.2 Spatial correlation

Relevance of geographical location relates to the resample step in bootstrap. When location is ignored, each value is drawn from the sample distribution with replacement and assuming the outcomes are independent, that is, previous drawings have no influence on future drawings. This assumption is contrary to the common notion of mapping by which transition from a high value to a low value is gradational, so that in proximity of a high value there will be another high value and close to a small value it is more likely to encounter another low value. One common quantitative way to objectively evaluate such spatial continuity is to prepare a semivariogram. The semivariogram, $\gamma(\mathbf{h})$, is the expected value of square differences between two observations, $Z(\mathbf{s})$and $Z(\mathbf{s}+\mathbf{h})$, as a function of distance, $\mathbf{h}$. If N_h is the number of pairs $\mathbf{h}$ units apart, the following is an unbiased estimate for the semivariogram:

$$\gamma(\mathbf{h}) = \frac{1}{2N_h}\sum_{i=1}^{N_h}\left[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\right]^2 \qquad (3)$$

It is not possible, however, to estimate semivariograms directly for a sample with nondetects. One way around the problem is to transform the data to indicators. For a threshold c, the indicator is:

$$i(\mathbf{s};c) = \begin{cases} 1, \, if \;\; z(\mathbf{s}) \le c \\ 0, \text{ otherwise} \end{cases} \qquad (4)$$

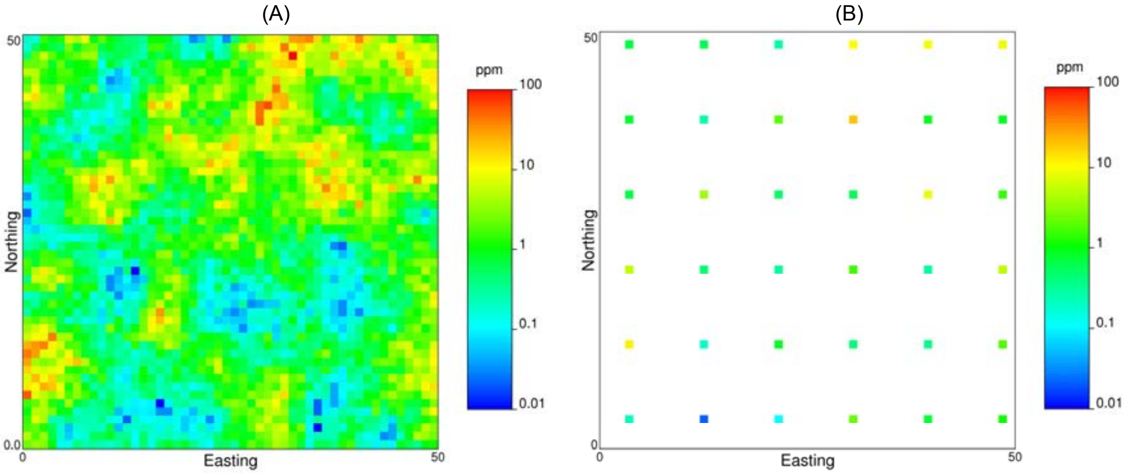

Figure 18: Posting of the exhaustive sample and the sample of size 36, which was obtained by considering every other 8th measurement: (A) exhaustive sample; (B) sample of size 36.

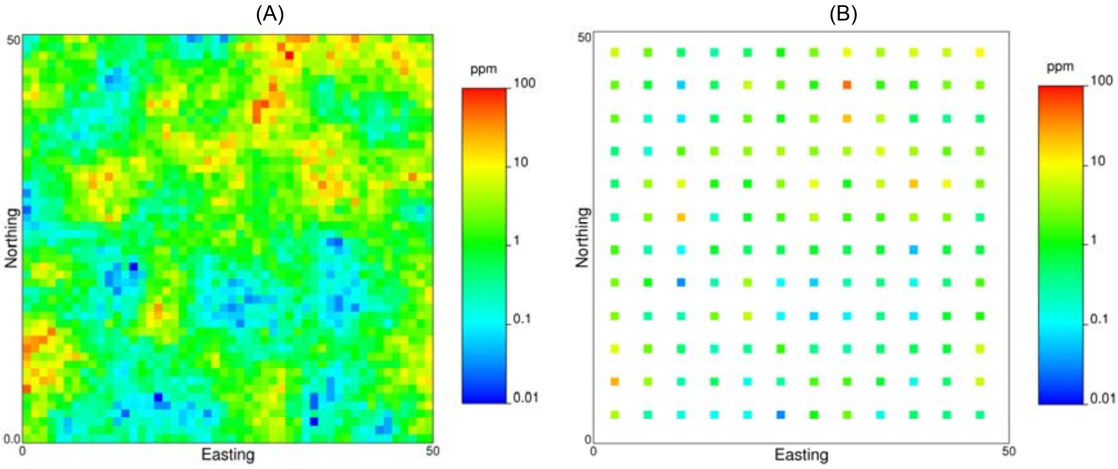

Figure 19: Geographical relationship between sampling sites in the exhaustive sample and sample of size 144: (A) exhaustive sample; (B) every other third measurement, resulting in the sample of size 144.

Figure 20 renders the semivariograms for both of our samples considering a threshold of 1 ppm. The indicator semivariograms are typical of samples with and without spatial correlation. Figure 20B is a typical semivariogram for a spatially correlated attribute, characterized by a positive function increasing from a distance of zero and then stabilizing. The distance for the beginning of the stabilized values is called the range and the value at which the semivariogram remain constant is the sill. When, like in Figure 20A, the ascending part is missing, the sample is not spatially correlated and is said to follow a pure nugget effect model. Note than the same exhaustive sample can generate spatially correlated and not correlated samples depending on the sampling interval.

Despite the exhaustive sample being indeed correlated, a sparse sample like the ones in Figures 3 and 18 is not spatially correlated. Hence, statistical analysis ignoring spatial location is valid.

7.3 Nondetect bootstrap for spatially correlated samples

If the sample is indeed spatially correlated, nondetect bootstrap in Section 2.3 needs further retooling because the resamples resulting from Steps 2 and 3 are not going to be spatially correlated. Disregarding such correlation becomes an unacceptable departure from the relevant characteristics of the sample.

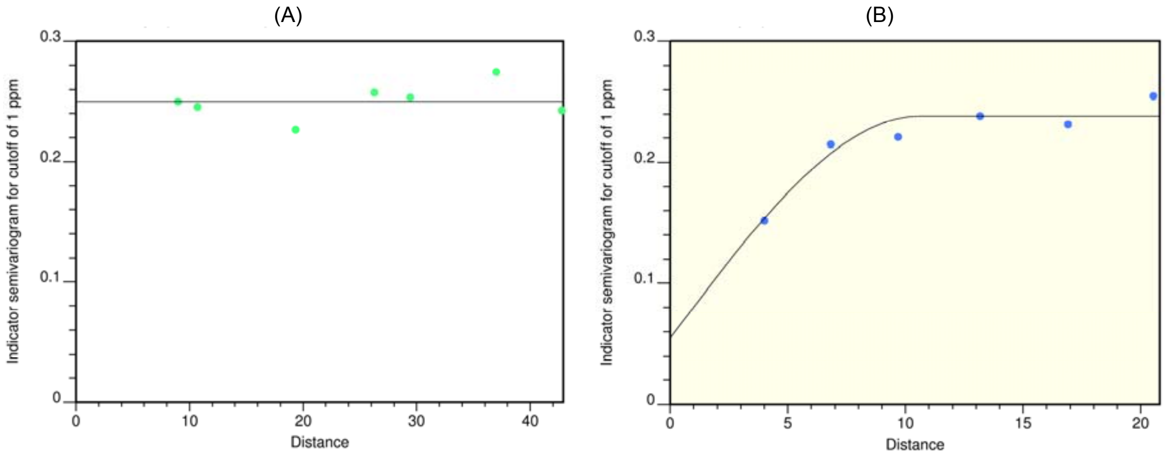

Figure 20: Indicator semivariograms for a threshold of 1 ppm: (A) sample of size 36; (B) sample of size 144.

The following is a procedure to properly handling preparation of distributional parameters for correlated samples with nondetects. The key for generating resamples with the same semivariogram as the sample is to generate stochastic realizations by sequential indicator simulation (Deutsch and Journel, 1998) and then resample from such realizations instead of doing it directly from the sample, as formulated below. Despite apparent radical departure from the procedure in section 2.3, keep in mind that the only significant change is to have resamples with the same semivariogram as the sample.

1. Collect an empirical sample of size *N*.
2. Make an ilr transformation.
3. Define at least 5 thresholds above the detection limits and model the indicator semivariograms.
4. Use the values above detection limit to fit a distribution model.
5. Employ the model to define at least 8 points of the cumulative frequency function.
6. Employing the semivariograms and cumulative frequency points, generate an unconditional realization using sequential indicator simulation.
7. Read the value of the realization at the sampling locations.
8. Use this new artificial sample to calculate and save as many statistics of interest.
9. Go back to step 1 and repeat the process at least 1,000 times.
10. Backtransform the results.
11. Stop

7.4 Spatially correlated distributional parameters

Figure 21 has been included as a way to confirm the fact that the resamples honor now not only the sample distribution but the sample semivariogram as well. In this particular case, the indicator semivariogram is close to being the indicator semivariogram for a threshold equal to the median.

Figure 22 displays the distributions for four parameters. Main change in the new distributions is an increase in the standard deviation relative to processing improperly ignoring such correlation, effect that can be appreciated by inspecting Tables 12 and 13.

Table 12. Distributional parameters for sample of size 144 containing 75 nondetects (52%), employing lognormal imputation, and considering spatial correlation.

	Geo. mean	*p*5	*q*1	*q*2
Exh. sample	0.955	0.090	0.340	0.960
No correlation	0.996	0.092	0.366	0.943
Correlated	0.977	0.087	0.357	0.937

Table 13. Standard deviations for distributions in Table 12.

	Geo. mean	*p*5	*q*1	*q*2
No correlation	0.084	0.177	0.111	0.128
Correlated	0.162	0.346	0.191	0.221

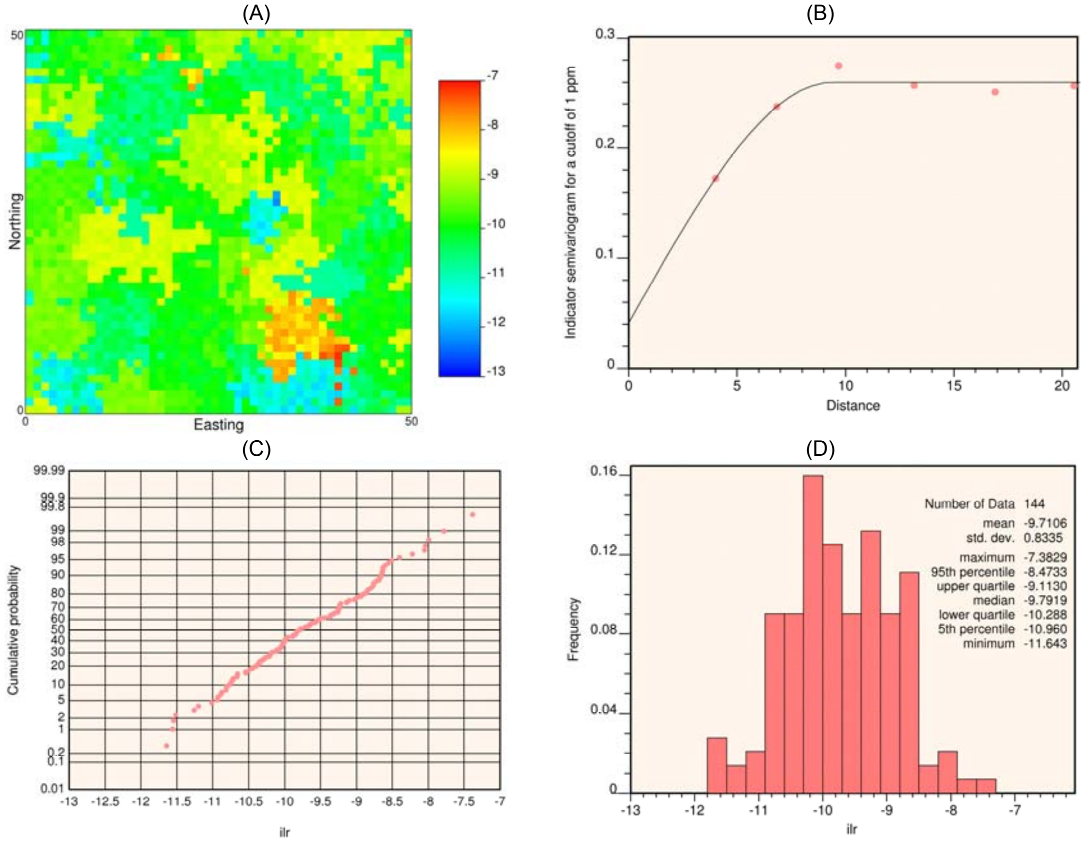

Figure 21: First pass for the bootstrap in Section 7.3 using the sample of size 144 censored at 1 ppm: (A) Map of first realization; (B) 1 ppm indicator semivariogram for the 144 bootstrap resample obtained by sampling the realization at the sites posted in Figure 19B; (C) cumulative probability for the first resample; (D) histogram for the first resample.

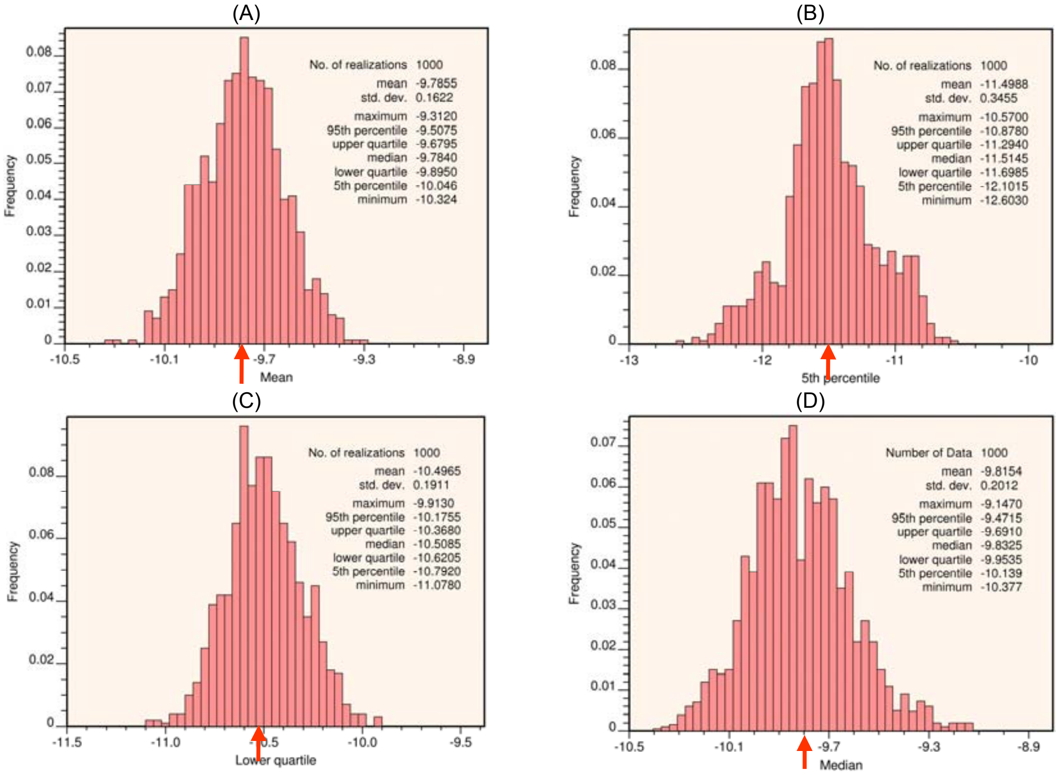

Figure 22: Distributional parameters for the sample of size 144 and a detection limit of 1 ppm: (A) mean; (B) fifth percentile; (C) lower quartile; (D) median.

8 Conclusions

Bootstrap was successfully used to predict distributional parameters of a parent population employing two samples with various levels and types of censoring. Ordinarily, geochemical data always have a geographical distribution. The first step in properly implemented bootstrapping is investigation of the spatial correlation for deciding the type of resampling. If values in the sample are not correlated, it is acceptable to employ a bootstrap consisting of randomly drawing values with replacement from the sample distribution followed by imputation by randomly drawing from the lower end of a distribution model based on the data about detection limits. The parametric model applies only to the nondetects; no model is used handling the observations above detection limit.

If the sample is spatially correlated, it is necessary to first generate an unconditional stochastic realization and then do the resampling from the realizations at the sampling locations. This approach assures that the resamples not only follow the same distribution as the sample, but also follow its semivariogram.

Dealing with only one detection limit and assuming independence, one of the most general conclusions is that estimation of distributional parameters by the bootstrap method tends to become more sensitive to the imputation method used to handle nondetects as the proportion of nondetects increases.

Unless the proportion of nondetects is minor and under the lower quantile of interest, the practice of substituting values below detection limit by a fraction of such limit is highly inaccurate, regardless of the particular fraction. Random imputation from a uniform distribution between zero and the detection limit is superior to any form of substitution, yet still significantly inferior to random imputation according to a model consistent with the information about detection limit.

Consideration of multiple detection limits is not a problem for any of the three imputation methods.

The standard error of the parameters grows when considering spatial correlation, by reducing the sample size, or by increasing the proportion of nondetects.

Although all conclusions in this study are based on a single numerical population closely following a lognormal distribution, results should be quite general because most analytes tend to be lognormally distributed (Helsel, 2005) and some conclusions are not bound by the nature of the distribution. In addition, the two improvement to the bootstrap method proposed in sections 2.3 and 7.3 do not require the use of a lognormal model. The formulations are quite general, calling for any distribution best fitting the data above the detection limit. It just happened, by the judicious selection of the exhaustive sample, that all censored samples followed lognormal distributions.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall Ltd; reprinted in 2003 at Caldwell, NJ: Blackburn Press, 416 p.

Buccianti, A., G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors (2006) *Compositional data analysis from theory to practice*. London: The Geological Society, Special Publications 264, 212 p.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researcher*. Hoboken, NJ: Wiley Interscience, second edition, 369 p.

Chilès, J.-P. and P. Delfiner (1999). *Geostatistics: modeling spatial uncertainty*. New York: Wiley-Interscience, 695 p.

Deutsch, C. V. and A. G. Journel (1998). *GSLIB: Geostatistical Software Library and User's Guide*. New York: Oxford University Press, second edition, 384 p.

Diaconis, P. and B. Efron (1983). Computer-intensive methods in statistics. *Scientific American* 248(5), 116–130

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35(3), 279–300.

Fridley, B. L. and P. Dixon (2007). Data augmentation for a Bayesian spatial model involving censored observations. Environmetrics 18(2), 107-123.

Helsel, D. R. (2005). *Nondetects and data analysis*. Hoboken, NJ: Wiley-Interscience, 250 p.

Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* 65(11), 2434–2439.

Howel, D. (2007) Multivariate data analysis of pollutant profiles:PCB levels across Europe. Chemosphere 67(7), 1300–1307.

Huybrechts, T., O. Thas, J. Dewulf, and H. van Langenhove (2002). How to estimate moments and quantiles of environmental data sets with non-detect observations? A case study of volatile organic compounds in marine water samples. *Journal of Chromatography A* 975, 123–133.

Lee, L. and D. Helsel (2005). Statistical analysis of water quality data containing multiple detection limits; S-language software for regression on order statistics. *Computers & Geosciences* 31, 1241–1248.

Little R. J. A. and D. B. Rubin (2002) *Statistical Analysis with Missing Data*. New York: Wiley-Interscience, second edition, 408 p.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35(3), 253–278.

Martín-Fernández, J. A. and S. Thió-Henestrosa (2006). Rounded zeros: some practical aspects for compositional data. In: A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn, editors, *Compositional data analysis from theory to practice*. London: The Geological Society, Special Publications 264, 191–201.

Olea, R. A. (1999). *Geostatistics for Engineers and Earth Scientists*. Norwell, MA: Kluwer Academic Publishes, 313 p.

Palarea-Albaladejo, J., J. A. Martín-Fernández, and J. Gómez-García (2007). *A parametric approach for dealing with compositional rounded zeros*. *Mathematical Geology* 39(7), 625–645.

Pardo-Igúzquiza, E. and P. Dowd (2005). Empirical likelihood kriging: the general case. *Mathematical Geology* 37(5), 477–492.

Pawlowsky-Glahn, V., editor, (2005). Special issue in advances in compositional data. *Mathematical Geology* 37(7), 671–850.

Pawlowsky-Glahn, V. and J. J. Egozcue (2006). Compositional data and their analysis: an introduction. In: A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn, editors, *Compositional data analysis from theory to practice*. London: The Geological Society, Special Publications 264, 1–10.

Rathbun, S. L. (2006) Spatial prediction with left censored observations. *Journal of Agricultural, Biological, and Environmental Statistics* 11(3), 317–336.