Non-detect bootstrap method for estimating distributional parameters of compositional samples revisited: a multivariate approach

J. Palarea-Albaladejo[1], J. A. Martín-Fernández[2] and R. A. Olea[3]

[1]Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ, UK, javier@bioss.ac.uk
[2]Dept. Informàtica i Matemàtica Aplicada, UdG, Campus Montilivi, Edifici P-IV. E-17071, Girona, Spain
[3]US Geological Survey, Reston, VA 20192, USA

Abstract

Bootstrap resampling is an attractive, computationally-intensive approach for estimating population parameters and their associated uncertainties. Values below detection limit—also referred to as non-detects—frequently arise particularly when dealing with multivariate geochemical concentrations, making the estimation of distributional parameters—mean, median, percentiles—a difficult challenge. The bootstrap method can be used repeatedly for analyzing resampled versions of the original data set. This way it is possible to estimate univariate distributional parameters while also capturing the additional uncertainty due to missing information. Within this approach, a method must be chosen to substitute non-detects with appropriate values given the compositional nature of the data. This idea was first introduced by Olea (2008) in the previous CoDaWork'08 meeting. Making use of the isometric log-ratio transformation and analyzing one variable at a time, he proposed a univariate bootstrap procedure where the distributional parameters of geochemical components were modeled from bootstrap resamples considering different criteria to impute non-detects. After conducting a sensitivity analysis on both proportion of non-detects and sample size, the study concluded that when drawing randomly a value from the extrapolated tail below the detection limit of the distribution best fitting the complete data—usually the log-normal distribution for geochemical data—the bootstrap estimates turned out to be more accurate than those obtained using simple imputation methods. Rather than analyzing each variable separately, here we make a step further to get the most of the covariance structure of the data set, extending the univariate approach for replacing non-detects to a multivariate setting. As a test bench, a number of data sets containing non-detects are artificially generated from real geochemical data and used to evaluate the performance of different replacement methods within the bootstrap process. First results show improved results when non-detects are replaced by random values drawn from a conditional truncated additive logistic model.

1 Introduction

The presence of trace elements in concentrations below a certain detection limit (BDL) is a practical problem that often arises when analysing experimental samples, particularly in the natural sciences. Values below detection limits are then reported by the laboratories as indicators of limiting concentrations at which the analyte may be present, but it cannot be detected—hence the name *non-detects*. The number of non-detects is mainly related to current technical—or even economical—limitations of the laboratories. Also important is how "clean" the samples are in a particular study. A large amount of analytes may interfere with the determination of the concentration of the element of concern and, hence, the detection limit. All of this implies that detection limits can be regarded as a dynamical issue as, for a certain element, they may vary depending on the laboratory and on the time the samples were analysed. In fact, it is not unusual to have several detection limits associated with a single element. Following Olea (2008), the non-detects problem is depicted here in the context of inferring univariate distributional parameter from compositional data sets applying the bootstrap method. The constrained nature of compositional data determines the way non-detects have to be dealt with. This work seeks to be a first approach for extending the idea proposed in Olea (2008) by incorporating available multivariate information.

From a statistical point of view, non-detects represent a particular missing data problem: data are non-observed but the detection limit provides information as an upper limit, that is, as a left censoring point for the data distribution. Depending on the severity of the censoring, distributional parameters may not reflect the characteristics of the true underlying data distribution, so it is necessary to have a procedure to recover as realistically as possible the missing information. Any method aimed at

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

replacing compositional non-detects by sensible values must meet desirable requirements such as scale invariance and subcompositional coherence (Aitchison, 1986, postscript to the 2003 printing: p. 2). They also must not alter the relative relationship between those elements without non-detects. The simple idea of leaving all the elements containing BDL values out of the data analysis is generally regarded as suboptimal, as it implies getting rid of valuable information. One might think that since non-detects are unusually small values, they are not truly relevant. Nevertheless, they could have a major influence on the results, and even greater when working within the log-ratio framework as it is required for proper compositional data analysis (Aitchison, 1986). Other popular strategy such as just replacing non-detects with a constant value—commonly a fraction of the detection limit, e.g. 0.5DL or 0.7DL—may appear to work relatively well when the number of non-detects is very small, but they introduce bias and variability underestimation as the amount of non-detects increases (Helsel, 2006). Moreover, given the constant-sum constraint of compositional data, these proposals may distort the covariance structure of the data if suitable adjustments are not made in those elements free of non-detects (Martín-Fernández et al., 2003). As a result, debatable statistical outputs can be produced. An alternative approach is to exploit the statistical characteristics of the data distribution above the detection limit to provide estimates for the non-detects. Although more computationally intensive, these methods relying on some kind of data modelling are generally more reliable and provide better results, as will be shown below.

2 Bootstrapping compositional samples containing non-detects

The bootstrap (*e.g.* Efron and Tibshirani, 1993; Chernick, 2008) is a computer-based resampling method that treats the collected data as a pseudo-population. A number of *bootstrap resamples*—usually the same size as the original—are randomly drawn by resampling with replacement from the empirical distribution of the original data set, and univariate statistics of interest (mean, median, percentiles, and so on) can be then computed from each one of them. Finally, *bootstrap estimates* of univariate distributional parameters, summarising bootstrap sample-wise statistics, and their associated uncertainly measures can be derived in a straightforward way. This iterative process allows estimation of the sample distribution of almost any statistic we may be interested in.

Here we divert from the ordinary bootstrapping as the existence of non-detects requires procedures to deal with partly censored data. It is important to note that non-detects add an additional source of variability which will be reflected in resulting bootstrap standard errors and confidence intervals. Additionally, the compositional nature of our data requires that those procedures meet compositional principles. Note that a simpler alternative could be to first have the non-detect treated and then initiating the bootstrap, but in such a case the non-detects extra variability would not be incorporated into the modeling.

Given a data set $\mathbf{X} = (x_{ij})_{N\times D}$—$N$ samples of a D-part compositional random vector $\mathbf{x} = [x_1, \ldots, x_D]$—containing non-detects, the non-detect bootstrap scheme can be outlined as follows:

1. Randomly sample rows with replacement from $\mathbf{X}$ to get bootstrap resample $\mathbf{X}_b$ of size $N \times D$.
2. Replace non-detects in $\mathbf{X}_b$ by estimated values $\rightarrow \mathbf{X}_b^*$.
3. Compute and save univariate statistics of interest from $\mathbf{X}_b^*$.
4. Repeat 1–3 B times.
5. Compute bootstrap distributional summary estimates.

Originally, in the context of geochemical data and considering a single analyte, Olea (2008) showed how replacing non-detects at point 2 with random values from the left tail of an univariate log-normal—using the detection limit as cut-off value—provides more accurate results than the popular constant-value substitution. Now, we assess that strategy against other model-based replacing approaches considering a multi-analyte framework. Particularly, five different imputation criteria will be considered:

1. Simple substitution: non-detects replaced by 0.7DL.

2. Random uniform: non-detects are replaced by random values from an uniform probability distribution in (0, DL).

3. Random univariate log-normal: non-detects are replaced by random values BDL from an univariate log-normal distribution.

4. alr-EM algorithm: following Palarea-Albaladejo and Martín-Fernández (2008), non-detects are replaced, within a *Expectation-Maximization* full-data parameters updating loop, by conditional expected values from an additive logistic normal (ALN) model.

5. Random conditional ALN model: non-detects are replaced by random conditional values from a right-truncated ALN model.

The detection limit DL is not required to be the same for all components in any case and all procedures are designed to generate values below it. Model parameters for methods 3 and 5 are simply estimated from the observed data. The novelty of methods 4 and 5 is the incorporation of multivariate information from other components in the estimation of non-detects. In other words, they generate values taking into account the correlation structure of the components. In consequence, these methods are not likely to stand out in those situations where the components are nearly uncorrelated or the correlations are low. Note that methods 4 and 5 require a log-ratio transformation of the data for the multivariate normal model be used in real space. The additive log-ratio (alr) transformation (Aitchison, 1986) is used here for that purpose because it allows easily moving the DL information between original and transformed space. Note that the results are invariant with respect to the alr-denominator chosen (Palarea-Albaladejo and Martín-Fernández, 2008). Particularly for method 5, random values from the univariate conditional right-truncated normal distributions are generated using the alr-transformed DLs as truncation points. Then, the results are transformed back into the simplex to get the univariate statistics of interest from $\mathbf{X}_b^*$ in the original units.

3 The Fort Union data set

With the goal of comparing bootstrap results according to the non-detects treatment applied, a real data set originally free of non-detects was kindly provided by geochemists at the U.S. Geological Survey. The data consists of $N = 229$ samples of the concentration (in ppm) of $D = 5$ minor elements $[Cr, Cu, Hg, U, V]$ in carbon ashes from the Fort Union formation (Montana, USA). Actually, this vector of elements represents a subcomposition of a much larger composition available, and the data are not closed to a constant sum. Note that originally all the values are above the detection limit. Table 1 summarizes the ordinary descriptive univariate statistics. It can be seen that Cr, Hg and U have the smaller concentrations, whilst V exhibit the higher.

	min	p5	p25	geo mean	median	p75	max
Cr	0.72	1.24	2.35	4.07	3.71	7.07	28.8
Cu	16	26.4	37	49.54	47	67	203
Hg	0.14	0.26	0.48	0.74	0.71	1.17	5.77
U	0.21	0.50	0.91	1.53	1.36	2.36	17.4
V	10	25	49	70.05	70	130	500

Table 1: Fort Union univariate descriptive statistics: concentrations in ppm.

The biplot on clr-coordinates (Aitchison and Greenacre, 2002)—Figure 1—reveals that clr(Hg) and clr(U) have the higher relative variability as they have the longer rays. On the other hand, the vertices in Figure 1 lie far apart from each other and the minimum value in the variation matrix (Aitchison, 1986, p.76)—Table 2—is equal to 0.2965, which corresponds to the log-ratio variance between Cu and V. In consequence no strong geochemical associations are found in the data set, although care must

	Cr	Cu	Hg	U	V
Cr	0	0.553	1.084	0.475	0.346
Cu	0.553	0	0.522	0.493	0.297
Hg	1.084	0.522	0	0.895	0.946
U	0.475	0.493	0.895	0	0.512
V	0.346	0.297	0.946	0.512	0

Table 2: Fort Union variation matrix.

be taken on conclusions drawn from the clr-biplot given the modest percentage (76%) of variability explained by the first two axes. However the poor relationship between elements is also reflected in the correlation matrix of the alr-transformed data set (not reproduced here), where the highest value obtained is equal to 0.68 between log-ratios $\ln(Cu/V)$ and $\ln(Hg/V)$. Given the above, no large differences between the results from methods 3 and 5 are expected.

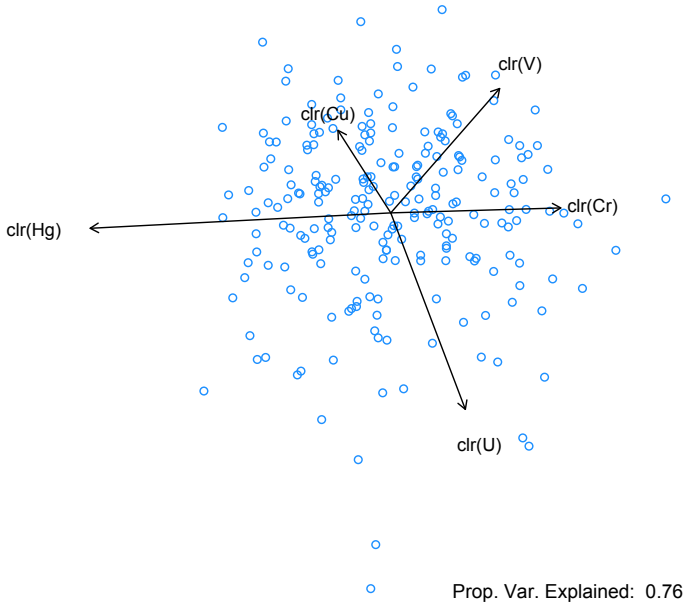


Figure 1: Fort Union data: clr-biplot.

From the original Fort Union data, a collection of synthetic data sets with different distribution of non-detects was generated. Geochemists provided us with a set of current reference detection limits for each element (*DL range* column in Table 3) and also with those detection limits particularly applied to the Fort Union data set (*FU DLs* column in Table 3). From this, we considered 6 different levels of non-detects—Low (< 5%), Moderate (5-15)%, Medium (15-25)%, ...—and detection limits were accordingly established for each element at each level as shown in Table 3.

As a result, 12 different scenarios were set out giving rise to 12 data sets $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{12}\}$ with percentage distribution of non-detects shown in Table 4.

			Low	Moderate	Medium	Medium-High	High	Very high
	DL range	FU DLs	(< 5%)	(5-15)%	(15-25)%	(25-50%)	(50-67%)	(> 67%)
Cr	0.1-9	0.1	1 (3.1%)	1.5 (8.7%)	2 (20.5%)	2.5 (28.8%)	5 (62.3%)	6 (69.4%)
Cu	1-500	20	25 (3.9%)	30 (8.7%)	35 (16.6%)	45 (45.0%)	60 (65.1%)	80 (86.0%)
Hg	0.005-0.02	0.01	0.25 (4.8%)	0.35 (14.8%)	0.4 (18.8%)	0.6 (38.0%)	0.75 (53.3%)	1 (69.4%)
U	0.1-1	0.15; 1	0.4 (2.6%)	0.7 (9.6%)	0.85 (20.1%)	1 (30.1%)	1.5 (54.2%)	4.5 (91.3%)
V	0.1-1	0.9	25 (3.9%)	30 (7.9%)	35 (23.1%)	70 (37.6%)	75 (63.8%)	120 (72.9%)

Table 3: Fort Union detection limits (left) and imposed detection limits by element-level combination (right). In parenthesis the percentage of BDL values.

Note that we decided to let Vanadium not to have values BDL in order to facilitate the use of

replacement methods requiring log-ratio transformation, particularly methods 4 and 5 described in Section 2 making use of the alr transformation. On the other hand, given that the data used here are a subcomposition of concentrations in ppm they are not in fact closed to a constant sum, as required to come back from the real space to the simplex using the alr inverse transformation. For addressing this issue we could fill the gap up to 1,000,000 by adding a residual—really an estimated residual—to those samples containing non-detects. We could also just transform the data into percentages or proportions dividing by the associated row sums—that is, closing them to 100 or 1—but then losing the original scale. However, our objective when replacing non-detects will be to approximately recover the original data set in ppm without altering the relative structure of the data. This can be achieved without adding any residual term as long as the replacement method does not alter the ratios between the elements. The alr- and inverse alr-transformation process embedded in methods 4 and 5 will itself close the data to a constant sum, say 1. In order to get the data set expressed back in the original ppm scale we can convert those replaced non-detects as follows. Let y_j be the value of a replaced non-detect after closure (as returned by the inverse alr-transformation), and y_k, $k \neq j$, the corresponding value for any other non-replaced element. Then the replaced non-detect in the original scale x_j^* can be recovered as

$$x_j^* = y_j \frac{x_k}{y_k},$$

where x_k is the value of element k in the original scale, taking advantage of the fact that the relative ratios between elements are preserved. In addition, this strategy for not carrying out an a priori closure of the data also prevent us from transforming the given detection limits according to a particular closure constant.

	% values BDL					
BDL scenarios	Cr	Cu	Hg	U	V	Total
1. Some low and nothing	3.06	3.93	4.8	0	0	2.40%
2. All low	3.06	3.93	4.8	2.62	0	2.90%
3. Some moderate, rest low	8.73	8.73	14.85	2.62	0	7.00%
4. All moderate	8.73	8.73	14.85	9.61	0	8.40%
5. Some medium, rest moderate-low	20.52	16.59	14.85	2.62	0	10.90%
6. Some medium, rest moderate	20.52	16.59	18.78	9.61	0	13.10%
7. All medium	20.52	16.59	18.78	20.09	0	15.20%
8. Some med-high	28.82	44.98	18.78	9.61	0	20.40%
9. Some med-high, rest medium	28.82	44.98	37.99	20.09	0	26.40%
10. All medium-high	28.82	44.98	37.99	30.13	0	28.40%
11. Some high	62.88	65.07	37.99	20.09	0	37.20%
12. All High	62.88	65.07	53.28	54.15	0	47.10%

Table 4: Twelve synthetic non-detects scenarios and associate percentage of non-detects.

4 Results

The non-detect bootstrap scheme outlined in Section 2 was applied to each data set in $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{12}\}$ considering the 5 different replacement methods for non-detects. For each bootstrap resample—$B = 1000$ were generated—the distributional parameters geometric mean, median and percentiles p5, p25 and p75 were computed for each chemical element.

For a better visualisation of the results, the bootstrap distribution of each distributional parameter for each method was obtained by kernel density estimation based on a Gaussian kernel. Figures 2 and 3 show the kernel density distribution of the parameter p5 for each combination of chemical element, imputation method and scenario. Analysing these distributions and their usual bootstrap statistics—bootstrap mean, standard deviation and percentiles p25 and p97.5—the following general remarks are made:

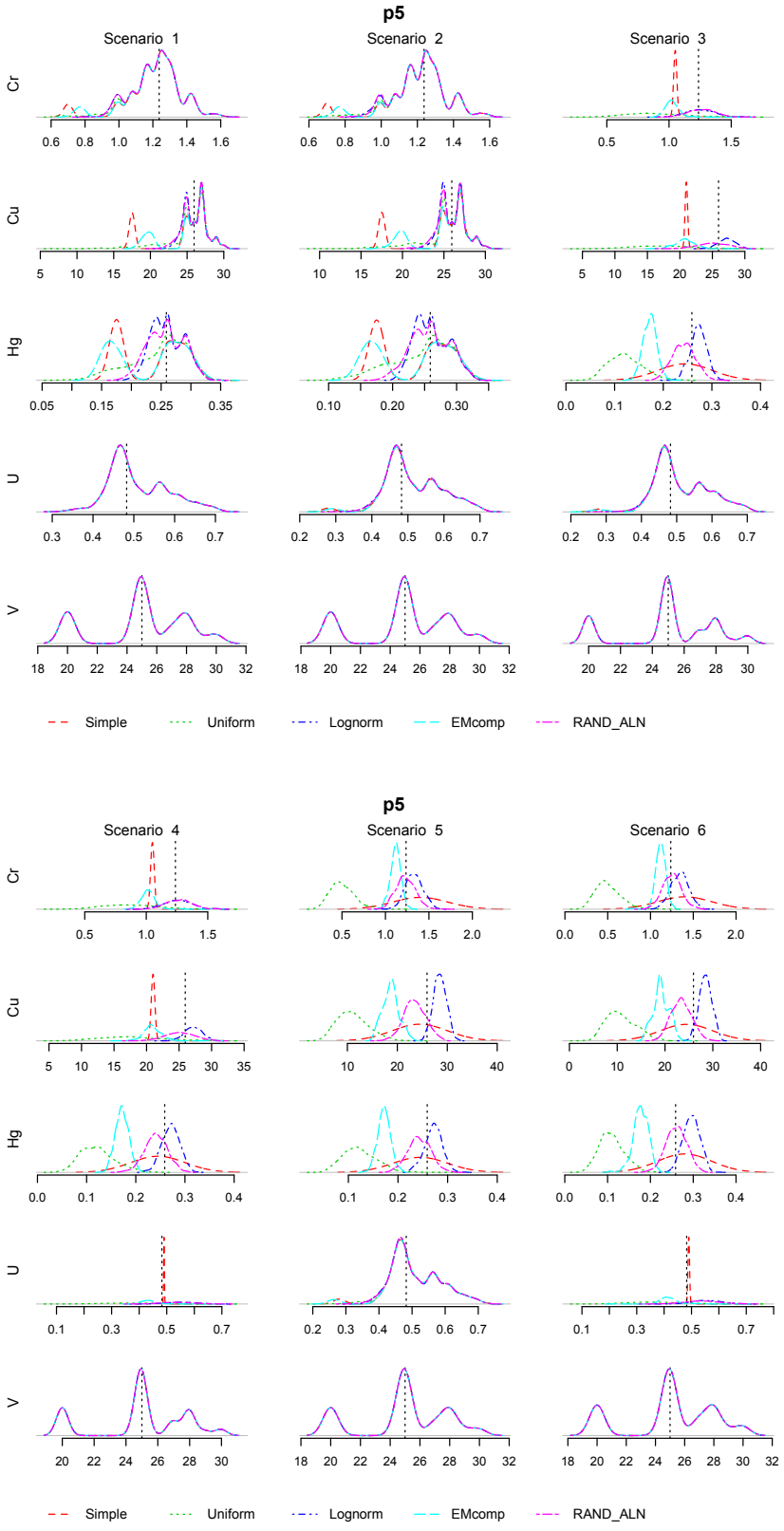
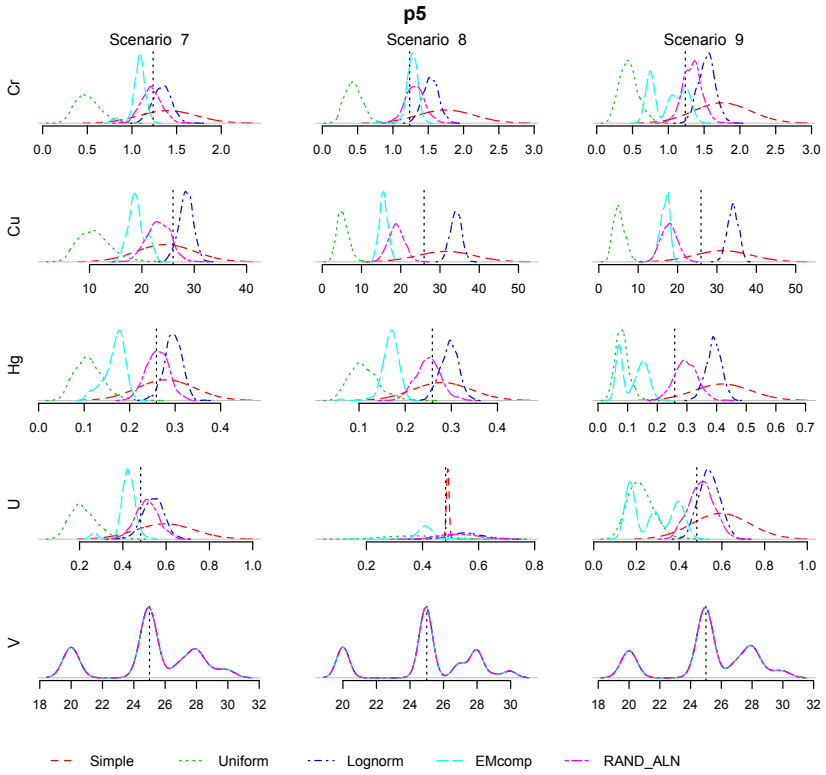


Figure 2: Fort Union data: kernel density estimation for distributional parameter p5 in scenarios 1–6.

- For any method and any scenario the bootstrap distributions of the univariate parameters of V, the element without non-detects, are exactly the same.
- Conversely, the estimation of p5 is nearly always affected since its value is closely related with

the presence of non-detects in the chemical elements.



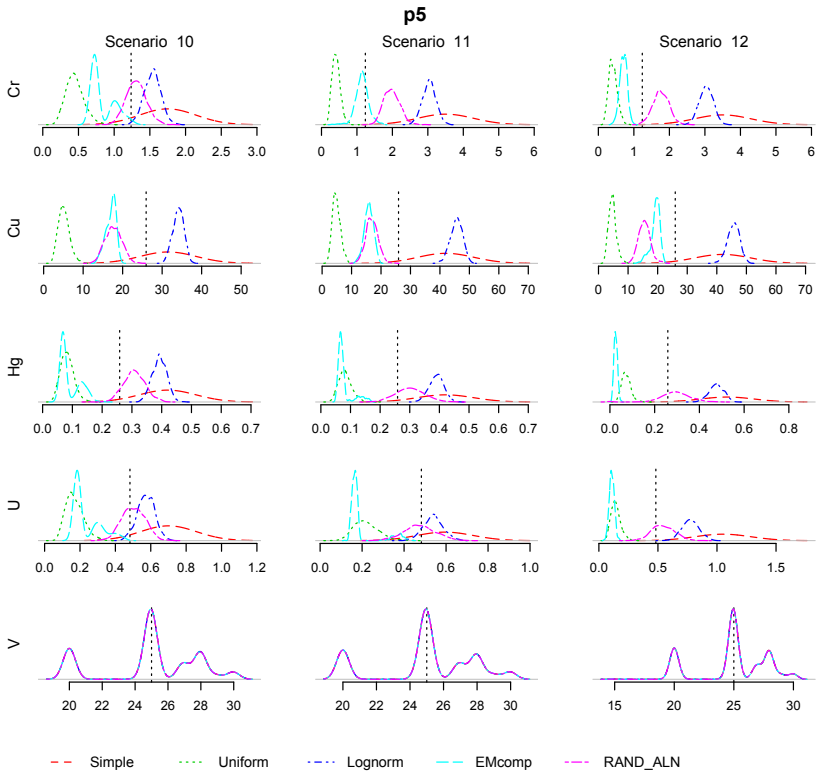


Figure 3: Fort Union data: kernel density estimation for distributional parameter p5 in scenarios 7–12.

- In most of the scenarios no impact is observed in the estimates of larger percentiles: p25, median, p75, and p95. Only when the number of non-detects is very high—say scenarios 10 to 12—the

estimations are slightly different over the five methods.

- The estimation of the geometric mean is nearly always affected given that this measure deals with all the values of the data distribution.

Because the crucial information in relation to the performance of the five methods relies on the estimation of the smaller percentiles, we only present here the detailed results for the 5th percentile. From the bootstrap estimates and the plots in Figures 2 and 3, the evaluation of the methods in the twelve scenarios may be summarized as follows:

- Across all the scenarios it is observed that simple substitution and random uniform methods tend to underestimate the parameter p5. Only for scenarios 8 to 12 the simple substitution method overestimates the true value. The distributions produced by the random uniform method are systematically located left of those from the simple substitution method. When the number of non-detects increases the random uniform method produces distributions mostly located around the half of the DL, whereas those from the simple substitution method are located by the 70% of the DL.

- The random log-normal method sometimes underestimates the true values, however it predominantly tends to overestimate p5. In fact, in most of the cases, the distributions provided by this method are located at the right-hand side of the plots. This suggests that it usually replaces non-detects by values close to the DL.

- The method based on the alr-EM algorithm provides distributions more concentrated around the mean than those produced by the random conditional ALN method. In addition, in most of the scenarios, the alr-EM distributions are to the left of the random conditional ALN distributions. This behavior could have been anticipated because the alr-EM method replaces non-detects by expected values and then tends to underestimate the variability as the amount of non-detects rises. On the other hand, the random conditional ALN imputations reproduce better the data variability.

- Despite the poor correlation structure of the FU data set, we could say that the random conditional ALN method, which generalizes the method proposed in Olea (2008), provides in general better results than the other approaches.

5 Concluding remarks

In this work we suggest an extension of the univariate non-detect bootstrap method proposed in Olea (2008) to a multivariate framework so as to take advantage of the information available in the correlation structure of the data set. Two alternative multivariate imputation strategies for replacing non-detects within the bootstrap iterations are proposed: the alr-EM algorithm (Palarea-Albaladejo and Martín-Fernández, 2008) and the random conditional ALN model. From our simulation results it can be stated that the multivariate approach based on the random conditional ALN model represents an improvement relative to a similar approach modeling each variable separately. This result is even more remarkable considering the low correlations between the elements comprising the data set used as a basis for the study. We have also observed that the alr-EM algorithm, which has been a reliable method for dealing with BDL values in compositional data sets, does not appear to suit well within the bootstrap scheme.

Acknowledgments

This research has been supported by the Spanish Ministry of Science and Innovation under the project "CODA-RSS" Ref. MTM2009-13272; by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2009SGR424.

We are grateful for the suggestions of Lawrence Drew (USGS), Gloria Mateu-Figueras (UdG), and Santiago Thió -Henestrosa (UdG) for valuable comments that help to improve an earlier version of the paper.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and M. Greenacre (2006). Biplots for compositional data. *Journal of the Royal Statistical Society, Series C, 51*(4), 375–392.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers.* Wiley Interscience, Hoboken (USA). 369 p.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

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.

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

Olea, R. A. (2008). Inference of distributional parameters from compositional samples containing nondetects. *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop.* Universitat de Girona, ISBN 84-8458-272-4, http://ima.udg.es/Activitats/CoDaWork08/. May 27-30, CD-ROM.

Palarea-Albaladejo, J. and J. A. Martín-Fernández (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computer & Geosciences 34*(8), 902–917.