Optimising Archaeologic Ceramics XRF Analyses

Bergman, Jakob; Lindahl, Anders

Published in:
Proceedings of the 6th International Workshop on Compositional Data Analysis

2015

Citation for published version (APA):
Bergman, J., & Lindahl, A. (2015). Optimising Archaeologic Ceramics XRF Analyses. In S. Thió-Henestrosa, & J. A. Martín Fernández (Eds.), *Proceedings of the 6th International Workshop on Compositional Data Analysis* (pp. 29-37) http://dugi-doc.udg.edu/handle/10256/150

Total number of authors:
2

Optimising Archaeologic Ceramics XRF Analyses

J. Bergman[1] **and A. Lindahl**[2]
[1]Dept. of Statistics, Lund University, Sweden; *jakob.bergman@stat.lu.se*
[2]Dept. of Geology, Lund University, Sweden

Abstract

We present the first results of an experiment which is aimed at ultimately producing recommendations for analysing archaeologic ceramics specimens using hand held XRF analysis devices. In the experiment we study the effects of different measurement durations, different number of measured points, and three different types of surface treatments (breakage, polished, grounded) when analysing ceramics specimens, while controlling for nine different types of clay and three different types of temper (no temper, sand, rock), in total almost 1 000 analysed points. For each measurement, the proportions of 36 different elements and all other elements are estimated. In those cases with multiple measurements of a specimen, the compositional centre of the measurements is calculated. A complicating issue in the analysis is the large number of parts found to be below detection limit; 13 elements have more than 50 % of the measurements below detection limit and for more than half of those (almost) all measurements are below detection limit. We try nine different strategies for imputing the values. Each estimated elemental composition is compared to a reference estimate using the simplicial distance. The log distances are finally analysed using analysis of variance with main and interaction effects. We find that the different surface treatments have the greatest effect on the distances: grounded specimens yield the most accurate estimates and polished surfaces the least. We also find a significant effect of increasing the number of measured points, but less effect of increasing the duration of the measurements.

Key words: Archaeologic XRF analyses, Archaeometric experiment, Ceramics analysis, Elemental composition analysis, Simplicial distance.

1 Introduction

X-ray fluorescence (XRF) analysis, using hand held devices, has gained increased popularity among archaeologist during the recent years, primarily because of its portability and relatively low cost. The analysis produces an estimate of the elemental composition of a specimen. There is, however, not any real knowledge or agreement of how the analyses should be done to obtain the best results. For how long time should a specimen be analysed? How many points on a specimen should be analysed? Which would be preferable, to analyse one point for four minutes or two different points for two minutes each?

To obtain a good measurement one also needs to consider the type of surface that is being analysed. When an archaeologic artefact is encountered, the surface has usually been exposed to various chemical and mechanical interactions with surrounding materials, changing the elemental composition of the surface. Hence the surface might not be representative of the rest of specimen. To overcome this one could ground the specimen to a fine powder which would mix all parts of the specimen and also remove any effect that large grains might have on the analysis. Another option would be to break off a small piece of the specimen to create a fresh breakage surface gaining access to the interior of the specimen. A third more controllable option would be to remove a part of the surface of the specimen by polishing it with a suitable tool. An important question is how the choice of treatment will affect the analysis. Is one alternative preferable to the others?

In an attempt to answer the questions above, we present some first results of an experiment in which we study the effects of number of points measured, measurement duration, and treatment of the surface. The design of the experiment is described in more detail in Section 2 and the results of the experiment are presented in Section 3.

2 Experimental design

Nine different, comercially available, clays were purchased. The clays are listed in Table 1. Each clay was partitioned into three parts and different types of temper was applied, i.e. different materials were added to the clays to control for shrinkage as is commonly done in pottery. Sand was added to the first partition, to the second partition crushed rock was added, and to the third partition no temper was added. From the in all 27 different clay partitions, small samples were produced resembling potsherds and fired in a modern kiln.

Table 1: The nine different clays used in the experiment.

No.	Clay type	Description	Firing temp.
1	Earthenware	black	970–1040 °C
2	Earthenware	red, 25 % grog 0.2 mm	up to 1220 °C
3	Earthenware	pale red, mix of natural blue and red clay	
4	Earthenware	white	1020–1140 °C
5	Earthenware	white, 25 % grog 0–0.5 mm	1000–1280 °C
6	Stoneware	white	1000–1300 °C
7	Earthenware	red	1000–1150 °C
8	Stoneware	black, 40 % grog 0–0.5 mm	1220–1260 °C
9	Earthenware	red, all lime has been washed/removed	

From each of the 27 different types of potsherds three replicates was then prepared for analysis in one of three ways: one potsherd was broken to create a breakage surface commonly found in archaeologic ceramic samples, one potsherd was polished using a diamond polishing disc to give a "perfect", smooth surface, and one potsherd was grounded to a fine power to give a complete mixture of the sample removing any differences between the surface and the interior of the potsherd. This produced in total 81 different samples. The elemental composition of each sample was then analysed at 1 and 5 points, during 60 and 380 seconds, yielding in theory 972 measurements. However, due to the human factor the number of points analysed were in a few cases four or six instead of five, yielding in total 971 measurement. The analysis was done using a portable XRF device providing measurements of 36 elements plus a “Balance” containing all other elements.

2.1 Measurements below detection limit

Looking at the measurements it was noted that a fairly large amount of measurements were below the detection limit (BDL). The number of BDL measurements for each element is given in Table 2. It should be noted that five elements (chlorine, cobalt, selenium, antimony, and bismuth) have more than 99 % BDL measurements, and silver, cadmium, and tin have all more than 90 % BDL. Furthermore, magnesium, nickel, copper, tungsten, and gold have more than 50 % BDL measurements.

Table 2: The number of measurements below detection limit (BDL) for each element. In total the analysis comprises 971 measurements. Note that all measurements of selenium and bismuth were below detection limit, and all but one of antimony. Also chlorine, cobalt, silver, cadmium, and tin had more than 90 % measurements below detection limit.

Si	Ti	Al	Fe	Mn	Mg	Ca	K	P	S	V	Cr
0	0	0	1	432	634	1	0	410	274	24	83
Ni	Cu	Zn	Rb	Sr	Y	Zr	Nb	Ba	Pb	Th	Cl
656	530	25	0	0	1	1	1	11	21	85	965
Co	As	Se	Mo	Ag	Cd	Sn	Sb	W	Au	Bi	U
968	315	971	221	931	922	918	970	658	854	971	422

It is of course problematic to analyse data with such a large amount of measurements BDL. At least two main strategies are conceivable: elements could be excluded or measurements could be imputed. As in all such cases, it becomes a question of retaining information but not altering the data too much. Imputing data is of course not a problem when only a limited number of measurements are imputed, but one can question the reasonableness of imputing almost all values. We have chosen to try different ways of excluding and imputing in order compare the effects of the different strategies. The procedure was done in two steps. First three data sets were created with all elements with more than 50 %, 90 %, and 99 % BDL, respectively, removed. In those cases where elements with observed measurements where excluded, the observed measurements were added to the balance. Secondly, three imputation schemes were implemented to each data set. A non-parametric imputation with 0.65 of the detection limit and a model-based lognormal with either fixed or random imputation values (Palarea-Albaladejo and Martín-Fernández 2013). All imputation was done using the functions `multRepl` and `multLN` in the R package *zCompositions* (Palarea-Albaladejo and Martín-Fernández 2014). To provide some sort of comparisons of the effects of the imputation, the first two principal components for the nine data sets are plotted in Figure 1(A) and third and fourth are plotted in Figure 1(B). The plots in Figure 1(B) indicate that the major difference between the data sets might be between only retaining elements with less than 50 % BDL and keeping more elements. On can clearly see in Figure 1(A) that the random imputation adds more variation, as intended.

2.2 Assessing the accuracy of the measurements

In order to assess the accuarcy of the measurement a reference (or “truth”) is needed. During the summer of 2013, we made an agreement with a colleague in Germany who had access to analytical

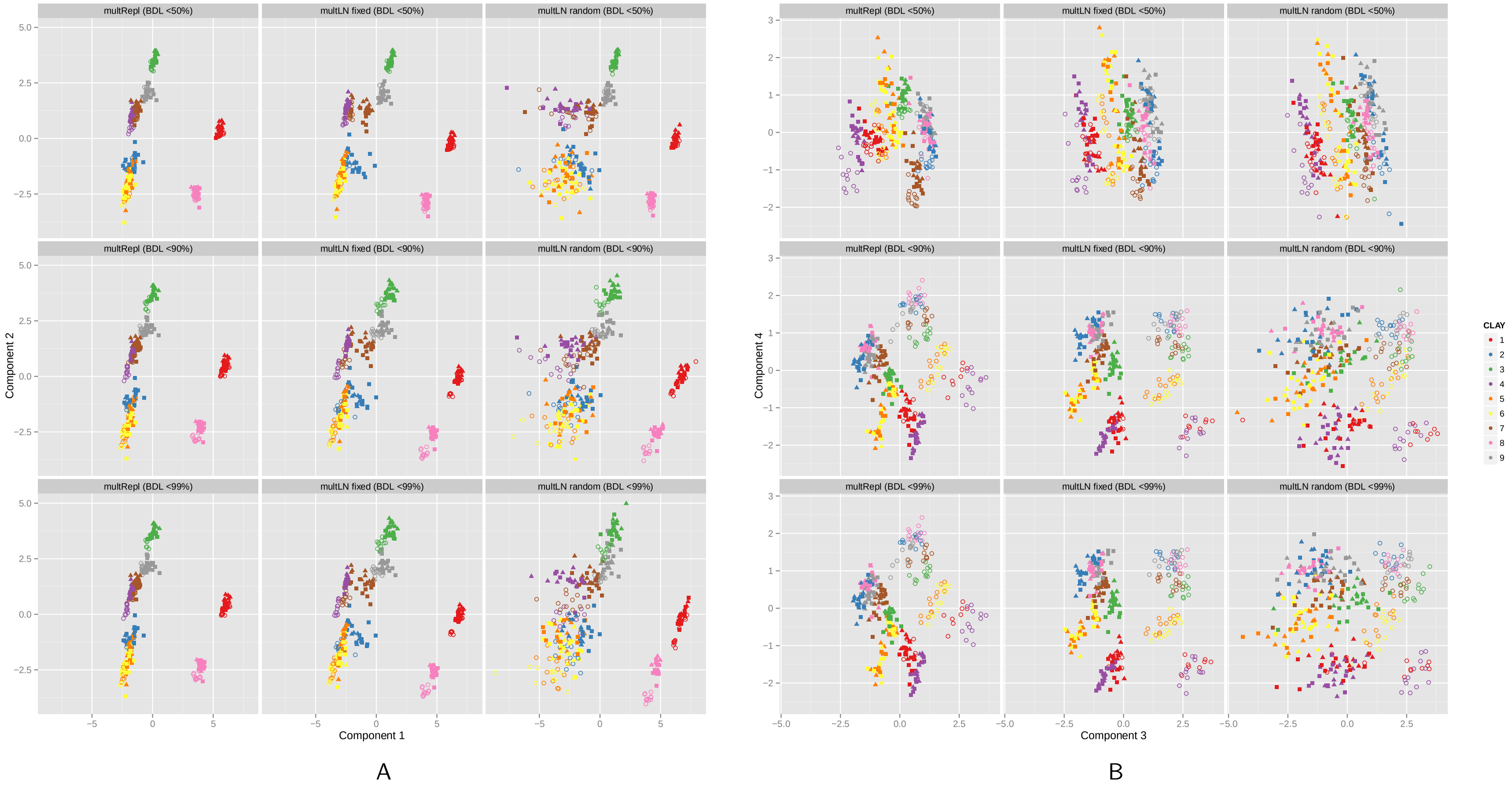

Figure 1: The (A) first and second and (B) third and fourth principal components from the compositional PCA analysis of the nine different data sets, with different colours for the different clays (see colourbar on the left) and different symbols for the different treatments: breakage surface (■), polished surface (▲), grounded (○). The results are rather similar; the main difference seems to be what number of elements is retained, not how the values are imputed. One can see that the random imputation induces more variation, which of course is to be expected. (The signs of some components have been reversed for some data sets for ease of comparison.)

equipment of greater accuracy to analyse the 27 different clay-temper combinations. To date we have not received the results, but hopefully the will arrive in the near future. However, we still needed a reference, so we decided to use the results we had. It was deemed that the grounded samples would provide the best estimates, and theory (and common sense?) suggested that longer duration would also provide better measurements. So, for every clay-temper combination the centre composition (Aitchison 1989)

$$\mathcal{C}\left(g(x_{11},\dots,x_{n1}),\dots,g(x_{1D},\dots,x_{nD})\right),$$

where $\mathcal{C}(\boldsymbol{x}) = (x_1,\dots,x_D)/\sum_{i=1}^{n} x_i$, i.e. the closure operation, and $g(x_1,\dots,x_n) = (x_1\cdots x_n)^{1/n}$, i.e. the geometric mean, of all the 380 seconds measurements was calculated. This was done separately for the nine different imputation schemes, thus obtaining nine different reference sets.

3 Analysis

For each combination of clay, temper, treatment, number of measured points, and measurement duration, we calculate the compositional centre of the measurements. Thus, for one measured point we keep that measurement and for five points we calculate the centre of the five points. This is repeated for all imputation schemes resulting in nine data sets of 324 compositional estimates. For each estimate we calculate the simplicial distance (Aitchison 1983, p. 64)

$$d_S(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{D}\left(\log\frac{x_i}{g(\boldsymbol{x})} - \log\frac{y_i}{g(\boldsymbol{y})}\right)},$$

where $g(\cdot)$ denotes the geometric mean, from the corresponding reference composition, i.e. the composition of that combination of clay and temper, resulting in nine sets of 324 distances. The calculations are done using the R package *compositions* (van den Boogaart, Tolosana, and Bren 2014).

The logarithm of the distances are analysed using analysis of variance. (The logarithm of distances are used as the distances are only positive and are expected to have a skew distribution. The decision is further strengthened by the fact that Box-Cox transformations indicate an optimal $\lambda \approx 0.2$, i.e. fairly close to 0.) We model the effect of different treatments, number of points and measurement duration including all two-way interactions, controlling for different clays and temper.

In Table 3 we present results of the analysis of variance for one of the imputation schemes, the fixed lognormal with less than 50 % BDL. It can be noted that the clearly most significant factor is the treatment, i.e. if the measurement was done on a breakage surface, a polished surface, or on the grounded sample. The least significant factor is the duration of the measurement. The results are similar for the other imputation schemes. The only difference is that the duration becomes significant when the number of elements is increased.

To get an idea of how the distances differ for different factor levels, we calculate the change in estimated mean value for the various combinations of treatment, number of points and measurement duration compared to the baseline of one measured point for 60 seconds on a breakage surface. The values are presented in Table 4. The shortest distances are found when the samples are grounded and the longest for the polished surfaces, with the breakage surfaces in-between. An interesting observation is that, whereas the accuracy of the measurements are improved with increased measurement duration for the grounded sample, the accuracy deteriorates with increased measurement duration for polished surfaces. For increased number of measurements the, the accuracy is improved for 380 seconds duration for both breakage and polished surfaces but deteriorated for the shorter duration.

Figure 2 shows normal QQ plots of the residuals for each of the nine different analyses (imputation

Table 3: The results of the analysis of variance for lognormal fixed imputation retaining only elements with less than 50 % BDL. The results for the other imputation schemes are similar. The treatment (breakage, polished or grounded) is the most significant factor and the duration of measurements the least significant. The main difference between the imputation schemes is that duration is significant when more elements are retained, but not when only elements with less than 50 % BDL are retained.

Factor	Df	Sum Sq	Mean Sq	F value	p-value
Clay	8	4.478	0.560	0.9154	0.5038
Temper	2	1.616	0.808	1.3212	0.2683
TREATMENT	2	272.422	136.211	222.7604	0.0000***
DURATION	1	1.880	1.880	3.0746	0.0805
POINTS	1	10.266	10.266	16.7887	0.0001***
TREATMENT:DURATION	2	19.024	9.512	15.5560	0.0000***
TREATMENT:POINTS	2	9.652	4.826	7.8924	0.0005***
DURATION:POINTS	1	3.476	3.476	5.6848	0.0177*
Residuals	304	185.887	0.611		

Significance codes: 0 *** 0.001 ** 0.01 * 0.05

Table 4: Changes in estimated mean value of the log distances for the various combinations of treatment, number of points and measurement duration. The changes are compared to the baseline of one measured point for 60 seconds on a breakage surface. The estimates come from the analysis of the lognormal fixed imputation data set retaining only elements with less than 50 % BDL.

Points	Duration		Treatment	
		Breakage	Polished	Grounded
1	60 s	0	0.26998	−0.91252
	380 s	0.14817	0.86613	−1.49240
5	60 s	0.04198	0.41486	−1.54592
	380 s	−0.22417	0.59669	−2.54012

schemes). The plots indicate that the residuals have a slightly skewed distribution possibly violating the normality assumption.

Apparently, the treatment has the greatest impact on the distances, and especially whether or not the sample was grounded. Since grounded samples were used to create the elemental reference compositions, we rerun the analyses without the grounded samples, i.e. with only breakage and polished surfaces. In all the nine data sets treatment and the interaction between treatment and duration are the only significant effects. As the results are similar for all data sets, we provide as an illustration in Table 5, the estimated changes in mean values of the log distances for the various combinations of treatment, number of points and measurement duration for the same data set as above, i.e. the lognormal fixed imputation with less than 50 % BDL. We note that the breakage surface still provides shorter distances than the polished surface. An interesting observation is that for breakage surfaces, the mean distance decreases when either the number of points or the measurement duration is increased, but when both are increased not much is gained.

4 Conclusions and future research

We have investigated how the accuracy of the elemental composition analysis of clay specimens, using a hand held XRF analysis device, is affected by different types of surface treatments, different number of points measured, and different measurement durations. Our prior belief was that the

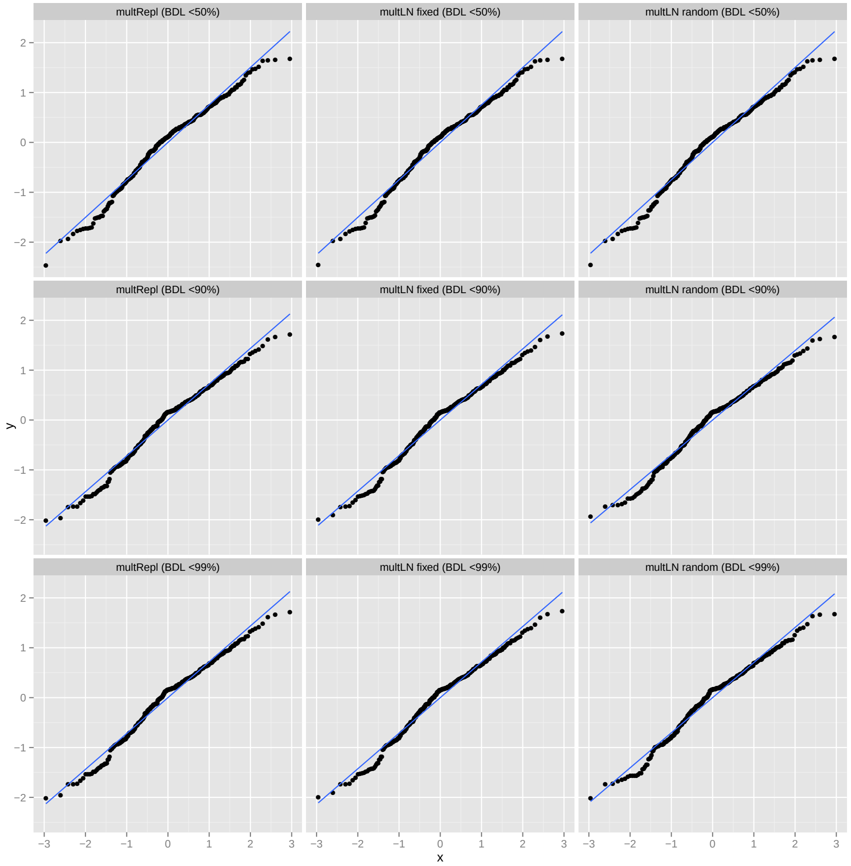

Figure 2: Normal QQ plots of the residuals for the nine different imputation schemes.

more points and the longer the duration, the better the accuracy. We also believed that a grounded specimen would produce the most accurate measurements. These suppositions are confirmed by the analyses.

We can conclude that a grounded specimen is the most important factor in obtaining an accurate measurement. We are actually a bit surprised by the large differences between the three surface treatments. The polished surface should provide an optimal surface for the XRF device, but turns out to yield the worst results. A possible explanation could be that the samples have been polluted by substances from the polishing disc, even though this seems unlikely. Another possible explanation could be that the polished surface prevents the analyst from avoiding the large grains of temper etc., which is easily done with a breakage surface. It is, however, a gratifying result that the breakage surface does so well. Grounding and, to a slightly lesser extent, polishing are both destructive treatments, which are often not an option for an archaeologist. One reason for the popularity of the handheld XRF device is that it can be used on artefacts in e.g. museums without damaging or even removing them. Grounding the specimen is thus an optimal but perhaps more theoretical strategy.

At least for breakage surfaces and grounded samples, the most accurate measurements are obtained when using both five points and a measurement duration of 380 seconds per point. (This is not the case for polished surfaces, which is rather puzzling.) However, from a practitioner's point of view, it should be noted that five measurements each during 380 seconds means that the total time of the analysis is more than 30 minutes, not counting the time setting up the equipment,

Table 5: Changes in estimated mean value of the log distances for the various combinations of treatment, number of points and measurement duration. The changes are compared to the baseline of one measured point for 60 seconds on a breakage surface. The estimates come from the analysis of the lognormal fixed imputation data set retaining only elements with less than 50 % BDL.

Points	Duration	Treatment	
		Breakage	Polished
1	60 s	0	0.26998
	380 s	−0.12913	0.58883
5	60 s	−0.23531	0.13757
	380 s	−0.22416	0.59670

preparing the specimen and moving the specimen between the analyses. Since time is limited (for most of us), the question becomes whether it is preferable to analyse a specimen at only one point for 380 seconds or at five points for 60 seconds each? (In each case the total time of analysis is about six minutes.) Our findings clearly show that measuring five points during 60 seconds yields more accurate estimates than one point during 380 seconds. It should be noted though, that for breakage and polished surfaces a single measurement of 60 seconds gives more accurate estimate when the grounded samples are included in the analysis.

This has been a first report from an ongoing experiment. It is of course highly unsatisfactory to use the same measurements that are analysed to estimate the reference values. We are therefore eagerly looking forward to getting new independently estimated reference values. In this paper an analysis of 971 measurements was presented. In total, the experiment to date consists of more than 1 800 measurements and more than 139 hours of XRF analysis device running time. In order to obtain a balanced experiment, about half of the measurements were excluded, as not all combinations of the factors are currently measured. It remains thus to complete the measurement sequence. This will also allow us to identify any threshold values in number of points and measurement duration: How much is gained in accuracy when increasing the number of points from one to three compared to increasing the number of points from three to five? Is there an optimal combination of number of points and measurement duration? It also remains as future research to investigate why the breakage and polished surfaces yielded more accurate measurements when measured only once for 60 seconds, than compared to five measurement for 60 seconds and to one measurement for 380 seconds. Is there a reason for this, or is it some sort of artefact of the extremely strong treatment effect?

Our conclusions in this experiment so far are that an archaeologist intending to do an elemental analysis of a ceramic specimen using a hand held XRF analysis device should ground the specimen if possible, and if not possible find a fresh breakage surface, and analyse the specimen at five points for 60 seconds each or, time permitting, for 380 seconds each, and finally calculate the compositional centre of the measurements.

References

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika 70*(1), 57–65.

Aitchison, J. (1989). Measures of location of compositional data sets. *Mathematical Geology 21*(7), 787–790.

van den Boogaart, K. G., R. Tolosana, and M. Bren (2014). *compositions: Compositional Data Analysis.* R package version 1.40-1.

Palarea-Albaladejo, J. and J. A. Martín-Fernández (2013). Values below detection limit in compositional chemical data. *Analytica Chimica Acta 764*, 32–43.

Palarea-Albaladejo, J. and J. A. Martín-Fernández (2014). *zCompositions: Imputation of Zeros and Nondetects in Compositional Data Sets*. R package version 1.0.3.