Ceramic investigation. How to perform statistical analyses.

Abstract

The aim of this article is to summarize and organize the statistical methodologies used for the statistical analysis towards ceramic investigation and in particular study of ceramic provenance. An update and review of all related methodologies is provided during the presentation of a typical statistical analysis. The presentation is given in a step-by-step process and emphasis is on interpretation of the intermediate and final results. The analysis attempts to cover the following:

- What issues to examine in a preliminary analysis
- Data transformation
- Cluster Analysis
- Clustering assessment
- Data dimension reduction methods as part of a clustering visualization and assessment
- Outliers and small groups
- Mixed-mode analysis
- Cluster characterization and discriminating factors
- Classification

1. Introduction

This paper is a part of a larger project, a series of papers, aiming to provide a guide to a researcher in the study of ceramic materials, ideally in every single stage of the study. Statistical methods in general are used throughout the process of an archaeological survey, from the initial stage of planning the survey and sampling until the stage of data collection. However, with the term statistical analysis in the title of this paper we focus on the last part of the survey, when the data have already been collected. In particular, we refer to the specific problem of analyzing the data obtained from an archaeological survey aiming to answer questions with respect to provenance, as a part of reconstruction of the past. Most of the statistical methodologies used for this problem fall within the branch of multivariate statistical analysis. This is due to the fact that the data collected for a provenance study in Archaeology consist of a multivariate data matrix, where the columns, i.e. the variables of the problem are in general correlated and a univariate study would be inadequate. Methods from multivariate statistical analysis would be in particular, data reduction methods, clustering and classification.

Quantitative methods in Archaeology, in general, have a long history. It is widely accepted by the community of Archeology scientists that statistical theory can be a valuable tool, sometimes necessary, in order to plan and organize the survey, handle the volume of data

collected, verify or clarify scientific hypotheses and make statistical inference based on the data at hand. First publications of statistical methods applied to problems in Archaeology appear the decade of 1960s (Binford, 1964) and more specific publications to provenance and multivariate statistical techniques appear in the mid of 1970s (Bieber et al., 1976). Baxter (2008) provides a thorough review on use of Mathematics and Statistics in the last fifty years in Archaeometry.

The paper with title 'The awful truth about statistics in Archaeology' by D.H. Thomas in 1978 is an indicator of the great expansion of statistical methodology used in Archaeology over the recent years before the date of its publication. The author although admits that undoubtedly statistical theory, when correctly used, can assist the researcher to efficiently and subjectively derive results on the archaeological questions, he discusses the misuses and in some cases abuses of Statistical techniques. The paper is organized in sections with titles 'the good', 'the bad' and 'the ungly' presenting in each section from harmless to more serious mistakes in statistical analysis applied to Archaeology. It is mentioned that the instruction of his editor was to 'shake things up in a pleasant way' and the author definitely accomplished this purpose.

Since then, Archaeological scientists have gain more experience in using quantitative tools, they are aware of the methodologies appropriate for their analysis, they are more educated and they have access to statistical/computer packages that can assist towards the implementation of such methodologies. This article does not attempt to present some new statistical methodologies –although a review on various choices in each case is given- but to provide some guidance with respect to the sequence of the steps an analysis and the implementation. The instruction of my editor was to present 'a tutorial approach –solving problem oriented' and my contribution to have 'an educational character'. I hope this is accomplished. The paper although technical has kept Mathematics to a minimum and emphasis is on 'how', 'why' and 'when' we use each method.

In particular, with respect to its content the article is organized as follows. In section 2, the issues we examine during a preliminary analysis are presented. Questions such as why a preliminary analysis is important and how we can use any conclusions made at this stage in the subsequent steps of the analysis, are answered. Section 3, is the main part of the paper, reviewing various approaches for clustering, illustrated in simulated data. Section 4 lists the tools and the steps the researcher can follow if he wishes to conduct an assessment of the clustering result(s). Within this section, as part of the clustering assessment, methods for data dimension reduction are also presented. Moreover, a presentation and proposed solution of special problems one has to deal with in cluster analysis, such as existence of outliers is given. In section 6 a brief presentation of the classification problem and possible approaches is listed. Finally, a summary of the steps of the analysis is listed in the last section.

2. Preliminary analysis. Data manipulation and transformation

The data collection should normally be a part of a more general process which includes the questions of the research. These questions, the purpose of the research project, are set at an

early stage of this project and the data collection is adjusted so that the data at hand will include information sufficient for answering the questions. Moreover, the type of the data collected needs to be in accordance with the methods of data manipulation which are going to be used. Alternatively, the methods of analysis need to be adapted to the type of data that are meaningful for the purpose of the analysis.

For example a key question in Archaeology is provenance. Are we able to identify which artifacts have common provenance and which differentiate from others? Is it possible to determine the factors which differentiate the distinct groups and learn from this process either with respect to the characterization of a source, or the technology used towards ceramic production? If provenance is the question of the analysis, data carrying information about the characteristics of an origin are necessary, such that the chemical composition. This will lead to a number of continuous variables in statistical terms, i.e. quantitative variables which can be measured on a scale and take infinite number of values. Those continuous variables will be the input data for a cluster analysis for example. Apart from chemical composition, other types of variables may be available, such as data collected from typology or microscopic study of the specimens. These variables will be categorical, nominal or ordinal. For example, variables taking values 'yes' or 'no' to the question if a certain mineral is present or not to a specimen, or variables taking values 1, 2, 3 and 4 where level 1 corresponds to the 'no presence' outcome, 2 to the outcome 'few', 3 for 'moderate' and 4 for 'plenty'. If we assume that this type of data also carry information about the provenance, methods of clustering which accommodate categorical variables would be appropriate to pursue or methods solely dedicated for categorical data.

The first stage of the analysis is to have the questions clearly stated. Moving to the numerical variables the preliminary analysis can include univariate summary statistics for each of the variables and univariate of bivariate plots for pairs. The preliminary analysis in general can give an idea of the data at hand. For example, range of values for a variable, shape of the distribution, existence of specimens with extreme values and possibly an obvious pattern, e.g a bi-modal distribution for one or more variables may suggest to the existence of two distinct groups. The preliminary analysis will not be sufficient to answer questions of the research project, but is essential and has an auxiliary role in the proceed analysis.

More analytically, for continuous variables a preliminary analysis can include summary statistics, i.e. mean, variance, standard deviation, median, first and second quantile, minimum and maximum value. Graphs, such as histograms for univariate variables will give evidence for symmetry or not, existence of a long tail due to extreme high values on this variable for instance, bi-modality etc. Moreover, it will point out a specimen with extreme value on that particular variable, a fact that needs to explore further. For example, if a statistical method which assumes normality is going to be used, a transformation in the log-scale will be suggested before the analysis in the case of a long right tail at the variable distribution.

Scatter plots for pairs of variables will also provide similar information, but moreover will add information about the relation between the two variables. Ellipsoidal shape of a scatter plot instead of spherical will indicate significant level of correlation. An obvious pattern in the data with respect to distinct groups may also be apparent. However, as it is the case for any simple or complex statistical analysis with multivariate data, the difficulty is the number of variables available, usually they are too many, and the contradict information that they may contain. Moreover, by studying the variables one by one the part of information due to the correlations among the variables is left out. A bivariate analysis is definitely superior that a univariate study, but the number of possible pairs is even higher and a selection of informative pair or pairs is essential.

As an example, we use simulated data for giving measurements on five chemical elements. The data set of total size 60 is comprised from three groups, where each group is generated using the lognormal distribution and parameters, i.e. means and covariance matrices differ among the three groups. Figure 1(a) plots a histogram of one of the element measurement. A long right tail is apparent and log transformation could transform the data producing a more symmetric plot allowing a symmetric model (e.g. normal distribution) to be assumed for the statistical methodology. Figure 1(b) is a scatter plot in two dimensions, plotting the measurements of two elements and using different symbols and colors for the three groups. From figures 1(a) and 1(b) we draw immediate results that the data presents heterogeneity, in particular high values in some of the variables and different type of correlation for at least two subsets of the data as indicated from the two selected variables in Fig. 1(b). However, this preliminary analysis does not form an analysis or enough evidence to draw any conclusions on certain structure regarding distinct groups in data.

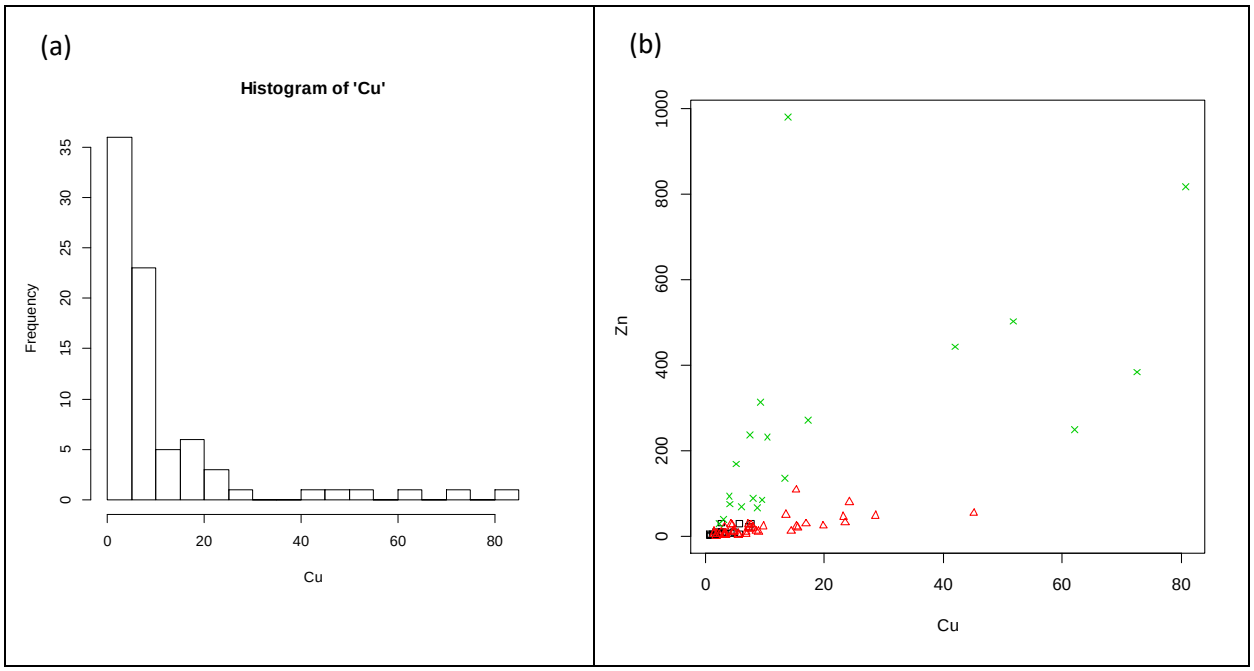

Figure 1. (a) A Histogram of univariate measurements of one selected element. (b) A scatter plot of bivariate data set. Color and symbol shape correspond to the groups.

Data Transformation

Data transformation plays an important role in subsequent analysis. If provenance is the problem of interest, a method of multivariate statistics within the framework of clustering will be applied and any result of cluster analysis will be visualized through principal component analysis (PCA) or any other statistical multivariate technique aiming to data reduction. Performance of either clustering or data reduction techniques will depend on data transformation. For example, for PCA in particular, it is a fundamental result that if the data variable variances differ, the first components of a PCA analysis will be dominated from the variable(s) with large variance. This is against the PCA performance since the first components instead of explaining the majority of data heterogeneity will solely explain the variability that data present at one or two variables, the variable with dominant variances.

For implementing cluster analysis, as seen in Section 3 analytically, both factors of data transformation and measure of distance will be essential for the analysis. Transformation of data primarily holds for continuous data in Archaeology, i.e. data resulting from the chemical composition of specimens. This analysis will provide us with a large number of variables, each one corresponding to one chemical element or their oxides. The measurements are either counts i.e. absolute measures or percentages out of 100% of weight. These data where the measurements for the set of variables sum to 100% for each sample are called fully compositional. Compositional data in general are called either the fully compositional or data that can be considered as a subset (with respect to variables) of fully compositional data.

Transformations usually implemented in analysis of Archaeometric data are (i) standardization or scale transformation in zero mean and variance one, (ii) log-transformation (iii) log and scale, (iv) log-ratio transformation for compositional data.

In particular, standardization will be suggested, among other cases, for a PCA visualization as mentioned before, log-transformation will be appropriate and improve the performance of either a data reduction/visualization technique or a clustering method if the data present asymmetry. It is proposed from the literature (see for example Bieber et al., 1976 and Bishop and Neff, 1989) logarithms to be taken with base 10. This transformation will result to a set of variables that will have nearly equal variances (a characteristic that standardization also guarantees) and variables measured in percentages or ppm are transformed to measurements which weight almost the same. Towards the implementation of log transformation special care is required for measurements that are zero.

Zero measurement naturally results in compositional data when the measured value is below the detection threshold. A possible treatment in this case would be to substitute zero with the threshold or with a value α smaller than the threshold (Beardah et al, 2003) or substitute zero with $\alpha/(p-1)$ with α a small number and p the number of variables (Aitchison, 1986). Other proposal is to adjust the remaining variable measurements when a zero is replaced by α. More specifically if x_{ij} is the measurement of i sample for the j variable, x_{ij} is replaced with $(x_{ij} - ax_{ij})/100$ (see Pawlowsky-Glahn, 2002). Another approach is to impute those values based on the remaining measurements (see Palarea-Albaladejo et al. 2015). Most of the statistical/computational packages include imputation techniques.

The log-ratio transformation is aiming to take into account the condition that compositional data have that they add up to a constant number. It is proposed by Aitchison (1986) and adopted by Buxeda I Garrigos (1999), Martin – Fernandez et al. (2015) and Pawlowsky-Glahn and Buccianti (2011) amongst others. If $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the vector of measurements of artefact i for the set of p variables, the log-ratio transformation of $x_{ij}, (j = 1,2, \dots, p)$ is $\log(x_{ij}/g(\boldsymbol{x}_i))$ where $g(\boldsymbol{x}_i) = (x_{i1}\, x_{i2} \cdots\, x_{ip})^{1/p}$ (centered log-ratio transformation). The merit of the log-ratio transformation is that the condition which holds for the data have is taken into account. Transforming the data in log-ratio scale such way the distances among vectors are expressed in a geometry that represents better their relation. There is however a discussion in the literature for choosing log-ratio and simply standardize the data for compositional data. When using log-ratio transformation an element with small absolute values, but relatively high variance is promoted in comparison to other elements with higher absolute presence and more relative to the question of the analysis, e.g. provenance. See Baxter (2001) and Baxter et al (2006) for some examples.

Nevertheless, if a structure is apparent in the data, both transformation will suggest this same structure and differences caused by practical problems as mentioned or presence of outliers will need further examination. It is important at this stage to mention that a PCA plot, in

either transformation will not be a method to conclude about a certain structure in the data, even if this it is apparent. A clustering technique is the appropriate method to answer this question and PCA will only act complimentary to a clustering technique confirming the validity of a clustering result. Other methods for assessing a clustering technique are also available and a common practice is to reckon in all them. Moreover, a two dimensional PCA plot, based on the first two principal components will not count for the 100% of the data heterogeneity and therefore it can serve as a data visualization method, as the closest view of the data we can have in two dimensions, but yet it is not the complete information. Some theoretical and practical considerations on PCA are presented in section 3.

Another practice would be to take ratios of variables instead of raw measurements in variables as an attempt to explain dilution and alleviate its effect. This dilution may be the result of different proportion of temper added to the paste towards the production of ceramic. The paste source may be the same, but different proportion of temper may result in different composition. One way to cancel out this effect if to work with ratios, i.e. instead of measurements $(x_{i1}, x_{i2}, \dots, x_{ip})$ for the i artefact in elements p elements, the ratios $(x_{i1}/x_{ik}, x_{i2}/x_{ik}, \dots, x_{ip}/x_{ik})$ can be used for the analysis, where x_{ik} is the measurement at a chosen element k. For a detailed discussion see Baxter (2001).

3. Statistical methods for provenance

Cluster Analysis (CA) is the most widely used method of multivariate statistical analysis in Archaeology (Baxter, 2008). It is a term to include any statistical method, now-days any machine learning as well, seeking for similarities among observations, based on a number of variables, and identify groups consisted of observations that have common characteristics. Those groups are called clusters. CA is applied to Archaeometric data in Archaeology aiming to identify artefacts that share common characteristics in composition and therefore make inference about these with respect to provenance, technology and draw subsequent conclusions on economic and social relation of past societies.

A plethora of clustering techniques is available and this is in practice one of the reasons the task it can be challenging. CA is typically used when after a preliminary analysis of the data, we suspect that there is heterogeneity within the dataset and it cannot be assumed as a sample coming from a unique population. This can be apparent either from univariate or bivariate plots (Fig. 1), summary statistics or plotting the data using any data reduction technique (e.g. PCA) and detect some pattern in the data, e.g. bi-modality or different type of relation among observations (e.g. Fig 1b). Another charactering of CA that make the task challenging, is that no prior knowledge is of group membership is assumed. Any CA method is classified within the unsupervised statistical learning techniques. With ‘unsupervised’ we refer to the fact that there

is no prior information available to either assist towards the group formation or the assessment/comparison of clustering techniques.

Moreover, the range of statistical methods available for CA varies with respect to the statistical assumptions they make, the complexity of the methodology and the computational demand. Most importantly, they may also vary in special characteristics, such as the effect of presence of outliers in clustering procedure and the properties they impose in resulting groups, e.g. some method impose spherical shape into clusters or they assume equal size groups. Therefore, for a scientist of Archaeology, or any other area, some knowledge of the theory behind each technique is recommended and in any case an exhaustive exploration of clustering techniques would be necessary before drawing any conclusions. Some literature in CA among others, are: Everitt et al (2011), Baxter (2015) and Papageorgiou (2018). For a practical guide to implementation CA see Kassambara (2017).

The statistical clustering techniques are classified into four categories with respect to their approach towards the problem. Namely, the categories are: Hierarchical, optimization or partitioning, model based and density based. From those four categories hierarchical and partitioning methods are most often used in Archaeology mainly because of the ease in implementation and lack of statistical hypothesis they assume. Most of the statistical methodologies may handle both continuous and categorical data, but it is more common in Archaeology to use the chemical composition (continuous) data only for obtaining the groups and verify or compare those with the information available from categorical data, such as mineralogical data. The main reason why this is the case is that mineralogical data are less frequently recorded in a manner that invites quantitative analysis and they are considered semi-quantitative. If quantitative discrete data are available, an analysis to the combined data, mixed-mode as they called, is appropriate. A discussion on mixed-mode data analysis follow the presentation of the clustering techniques and methodologies with the relative references will be provided at this stage.

Hierarchical Clustering

Hierarchical clustering is an approach of clustering that is algorithmic and is based on distances or equivalently similarities that can be calculated among observations based on the set of variables that contribute to the analysis. There are two 'symmetric' ways to implement hierarchical clustering. The agglomerative, according which the algorithm initiates from the situation that each of the n observations form a separate cluster and algorithm proceeds by merging observations until all observations are located in one single group and the divisible hierarchical clustering where the starting scenario is the inverse, i.e. all n observations form a single group and the algorithm proceeds by repetitive partitions until all observations are separated. In both cases, the whole procedure, from the one extreme situation until the other,

is completed and we decide about the number of clusters and the observations' cluster membership by inspection.

Two factors determine the hierarchical method. (a) The distance measure and (b) the linkage method. Distance or dissimilarity measure for continuous variables, can be any distance metric. The distances between all possible pairs of observations are calculated using the same chosen metric. Assuming that the data matrix consists of measurements of n observations on p chemical elements, an nxp matrix, calculating the distances among all possible pairs will lead to an $n \times n$ symmetric matrix called dissimilarity matrix, usually denoted as D. Table 1 lists the most widely used distance metrics as a measure of dissimilarity for continuous variables. At each stage of the algorithm an updated dissimilarity matrix is calculated.

Measure	***Definition***
Euclidean	$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$
Squared Euclidean	$d_{ij} = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2$
City block or Manhattan	$d_{ij} = \lvert x_{i1} - x_{j1} \rvert + \lvert (x_{i2} - x_{j2}) \rvert + \cdots + \lvert x_{ip} - x_{jp} \rvert$
Minkowski	$d_{ij} = \sqrt[m]{(x_{i1} - x_{j1})^m + (x_{i2} - x_{j2})^m + \cdots + (x_{ip} - x_{jp})^m}$
Maximum or Chebyshev	$d_{ij} = \max_k \lvert x_{ik} - x_{jk} \rvert$
Pearson	$d_{ij} = (1 - \varphi_{ij})/2$, where φ_{ij} is the Pearson correlation between data vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Table 1. Distance measures for continuous variables.

Updating will be necessary as merging for example in agglomerative clustering will occur and the number of clusters will change. Linkage is the method that the algorithm will use in calculations of distances among clusters or clusters and observations. Such calculations will arise at intermediate stages of the algorithm. If we further denote by x_{im} the (i, m) element of the data matrix , ($i = 1,2, \dots, n$ and $m = 1,2, \dots, p$), i.e. the measurement of the i-th observation on the m-th chemical element, and d_{ij} the elements of matrix D, Table 2 lists the definition of the most frequently used in practice linkages. It is worth mentioning that all these definitions are relatively simple, therefore not time consuming, and they can be implemented in most of the statistical packages. At each step of the agglomerative algorithm, units or clusters that correspond to the smallest distance are merged. The adopted choices of both metric and linkage remain the same for all intermediate steps of the algorithm.

Linkage	*Definition of inter-group distance of groups A and B*
Single Linkage or Nearest Neighbor	$d(A,B) = \min_{\substack{i\in A \\ j\in B}} d(i,j)$
Complete Linkage or Furthest Neighbor	$d(A,B) = \max_{\substack{i\in A \\ j\in B}} d(i,j)$
Average Linkage	$d(A,B) = \bar{d}$, where $\bar{d}$ the average distance $\bar{d} = \frac{1}{n_A n_B}\sum_{i\in A}\sum_{j\in B} d(i,j)$ n_A, n_B the group sizes of A and B respectively.
Centroid	$d(A,B) = d(\bar{x}(A) - \bar{x}(B))$ where $\bar{x}$(A) is the point $\bar{x}(A) = \frac{1}{n_A}\left(x_1(A), x_2(A), \dots, x_p(A)\right)$ and called centroid of group A.
Ward's	The increase in total error sum of squares (ESS) the merge of A and B will cause. If d_e is the Euclidean distance, ESS of a group U with centroid $\bar{x}(U)$ is defined as $ESS(U) = \sum_{i\in U} d_e(x_i - \bar{x}(U))$

Table 2. Agglomerative linkages.

Linkages listed in Table 2, have some properties it is useful to know, especially when assessing a clustering result. Single linkage is prone to the chaining phenomenon, a negative property where distant groups may be merged due to the existence of two neighbour measurements. . On the other hand it useful in identifying outliers. Complete, average and ward's methods are used more often in practice and they produce compact clusters. However, they all have problems in practice when ellipsoidal clusters occur in the data -a case quite common in Archaeometry.

Hierarchical clustering has as a property that once a merging of two observations or groups occurred during an iteration this will remain until the end of the algorithm. From the practical point of view, towards the implementation of hierarchical clustering, choices on: data transformation as discussed in section 2, distance metric and linkage method have to be made. It is apparent that having more than one option for each one of the three factors the number of possible combinations is quite large.

The sequence of hierarchy in merging or partitioning of a hierarchical clustering is represented graphically into a plot called dendrogram. The existing clustering, if any, will result by 'cutting' the dendrogram at a certain height. We illustrate the procedure using some simulated data.

Example 1. This first example is consisted from a 95x9 data matrix. The data have been generated by assuming that there are three groups each one distributed by a lognormal distribution with

different sample sizes and parameters. The sample sizes for the three groups are 15, 20 and 40 respectively.

Using standardization as a transformation for the data Figure 2 presents the dendrograms for average and complete linkages. Different software packages may use different ways of presenting the dendrogram, but the idea is the same. The sequence of the hierarchical clustering algorithm is captured in one plot and based on that we would like to answer the question 'how many clusters are in the data'? Or in other words, at which height we need to 'cut' the tree? The suggested height, at which we cut the tree is where the compact branches of the dendrogram will remain intact. For example, for the dendrogram according to complete linkage in Figure 2, cutting the tree at height 6 would be sensible, as the three compact braches of the tree remain intact. Deciding the number of clusters or the height one cuts the tree is not always straightforward. There might be two different heights, equally possible, or none. Since the problem is unsupervised and true clustering is not available to compare with and conclude, the suggestion is to try all possible scenarios and assess each one of them.

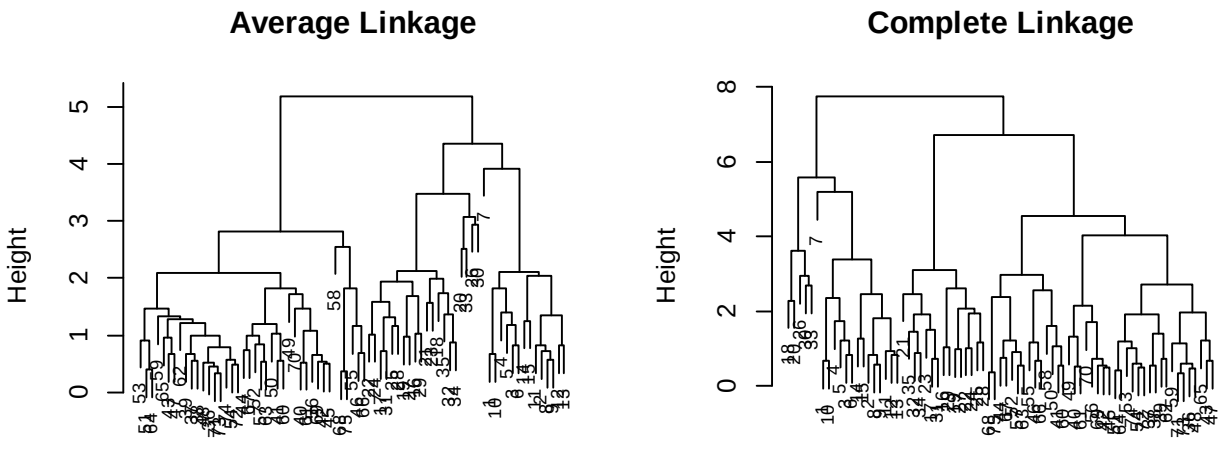

Figure 2. Dendrograms for the standardized simulated data using average and complete linkage

Apart from the dendrogram which is available from a hierarchical clustering, a number of indices are available in the literature dedicated to suggest the optimal number of groups (Charrad et al, 2014). Figure 3 plots the bar plot of 30 available indices implemented to the simulated data under study under the Average Linkage. The number of groups equal to three is the dominant choice.

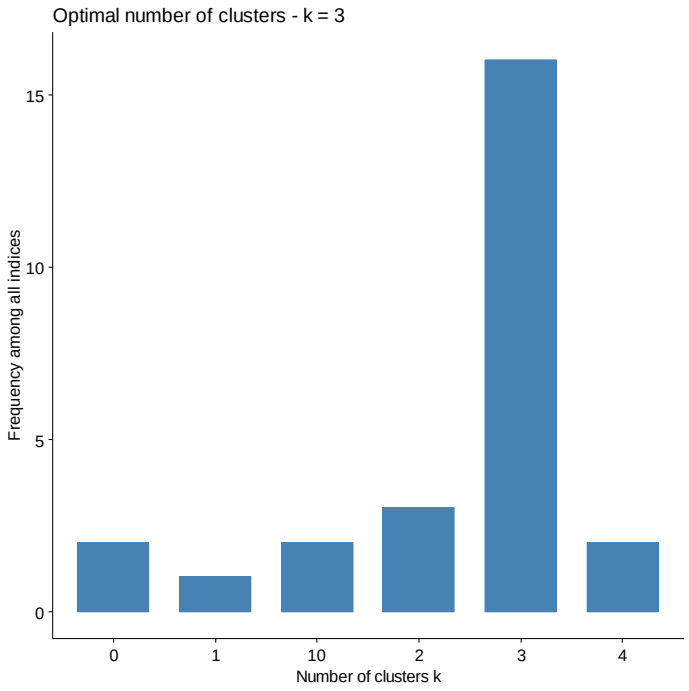

Figure 3. Optimal number of clusters for the simulated data under Average Linkage.

Taking into account both dendrograms and indices, let's assume that we decide to cut both trees at the height were the three braches would be suggested, e.g. height 4 for the average linkage and height 6 for the complete. Figure 4 plots the resulting clusters coloured using different colour for each cluster.

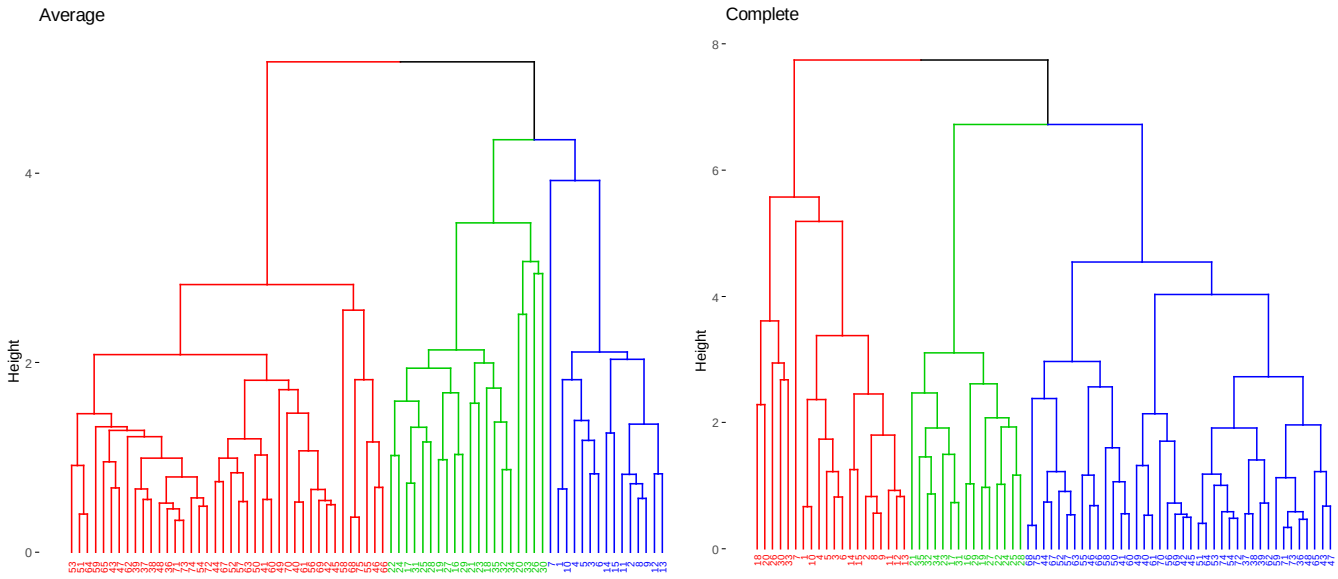

Figure 4. Average and complete linkage dendrograms with the proposed clusters given in different colors.

K-means

Another category of clustering techniques is partitioning or optimization method (see Everitt et al, 2011). According to this approach the groups are formed applying a partition to the data into a certain number of groups, using an optimization criterion. There is no hierarchy associated to the group membership, as in hierarchical clustering, and number of groups in the data has to be known in advance. The most known method in this category is k-means which uses as optimization criterion the minimum sum squared error. K-means as a method is very popular in Archaeology.

Baxter (2015) states that this popularity is mainly because it is readily understood, is perceived as being geared to archaeological needs, and was rendered accessible at a time when computational resources were limited compared to what is now available. Baxter in the same paper proposes variations of k-means method. This is because it is well established in the literature that k-means tends to produce spherical shape clusters and k-means is more appropriate for equally sized and spherical shape clusters (see Baxter 2015, Banfield and Raftery 1993, Papageorgiou et al 2001 among others). This is not the case in Archaeometric data, especially for geochemical data collected for provenance studies, where ellipsoidal shape and unequal size clusters exist. Another disadvantage of k-means is the effect of the presence of outliers.

Therefore, although popular, k-means may be used as one method, but needs to be cross examined with other techniques and check the validity of the proposed clustering. Within the framework of partitioning methods, k-medoids and Trimmed k-means are possible alternatives with improved behavior for both disadvantages of k-means method (Steinley, 2006).

Model-Based methods

Model-based clustering assumes a model, i.e. statistical distribution; to describe each distinct group in the data. The data matrix is assumed to be generated from a mixture of distributions and each component of the mixture captures one group. Fitting the mixture model to the data and estimating the parameters for every component will allow the scientist to identify the clusters and assign the observations to one of the resulting clusters based on the maximum posterior probability (see Banfield and Raftery, 1993). The estimation of parameters is implemented by Expectation Maximization (EM) algorithm, a computational statistical technique for deriving the maximum likelihood estimator. As a consequence, model-based clustering method is more demanding with respect to computational effort and it can be implemented via a statistical package. The computational demand and technical problems will be more intense as the dimension of the data observation is increase. The use of specialized statistical package to implement model-based cluster may be the reason for a limited use of model-based clustering in Archaeological applications.

The advantages of model-based clustering are many. The probabilistic model that is assumed provides the framework to assume ellipsoidal, different size or different orientation clusters. It is

therefore more general and more appropriate for compositional data. Moreover, estimating the model parameters results to a fully defined probabilistic model and statistical inference is now feasible. Statistical tests for comparison the model fitting are available and can be used in order to decide the number of components, i.e. the number of clusters in the dataset.

Other clustering methods, for example density-based methods, fuzzy clustering, kernel clustering are also available in the literature. Moreover, recent methods especially designed to cope with the presence of outliers, high dimensionality and large-scale data have also been proposed (Xu and Wunsch II, 2008).

Example 1 (continued). K-means: Let us assume that as suggested from dendrograms in Figure 2 and the majority of indices in Figure 3 the number of clusters is three. Implementing k-means with k=3 and using the same transformation for the data results into three quite different clusters in comparison with those in hierarchical and the true known origin since the data are simulated. More analytically, the first two clusters of 15 and 20 observations are merged to one cluster, and the third cluster of 40 observations is divided into two smaller.

Model-based: Implementing model-based clustering without any prior information about the number of clusters in the data the optimal model based to BIC criterion was the model with three components and variable volume, shape and orientation. Figure 5 plots the BIC value for a range of number of components (one to nine) and a range of possible models. All seventy five observations are correctly assigned, i.e in complete agreement with the true origin.

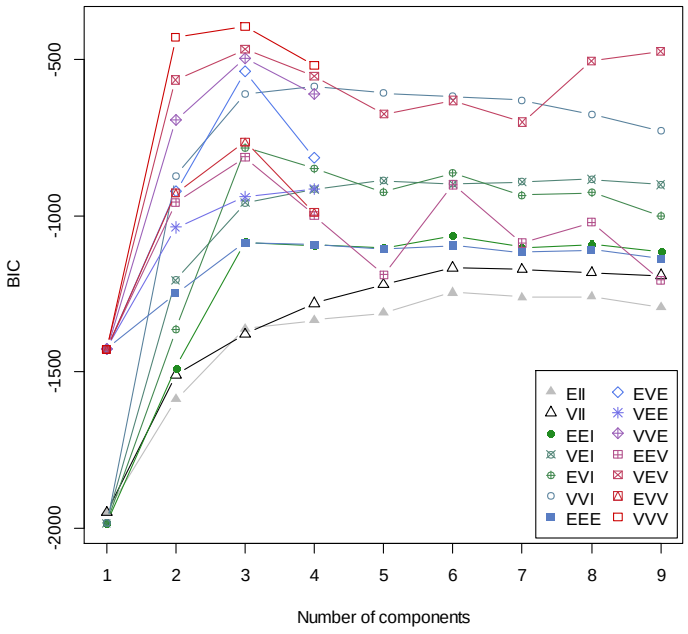

Figure 5. BIC criterion for model selection in model-based clustering.

4. Compare clustering results and assess their validity

The clustering methods and possible modifications of each method, together with data transformation choices, lead to a great number of possible clustering. Clustering in general, is an unsupervised technique, meaning that the true classification is not known in advance and there is not available a labelling for the observations indicating the cluster they belong to. Lacking of this information, assessment of the plethora of possible clustering results that may be available for the same case study may be a very difficult task. The difficulty is to decide which clustering is more valid compared to others or which clustering is more valid in general and therefore more probable to represent the true classification. In this paragraph we focus on how we compare clustering results obtained from different methods or choices in implementation and secondly how we assess a specific clustering with respect to its validity.

Correlation and Similarity Indices for clustering comparison.

One way to compare two or more clustering with respect to their similarity is to calculate the correlation of the corresponding dendrograms. Cophenetic and Baker correlation coefficients are appropriate measures (Saraçli, S. et al. 2013). Adjusted Rand index (Hubert and Arabie, 1985) is a popular measure of clustering agreement and other indices used for the same purpose are: Rand, Fowlkes and Mallows, Wallace and Jaccard index (Gordon, 1998 and Shotwell, 2013). Adjusted Rand index is taking values from 0 to 1 and the closer to one is the better agreement between the two clustering results. Such measures are useful, especially when a large number of clustering techniques is considered because if the majority of some clusterings agree, then it is more probable this particular clustering to hold in reality.

Example 1 (continued). Let us consider the two clustering results as presented in Figure 4, clustering from k-means with k=3 and clustering as suggested from the optimal model on model-based clustering. Comparison between the two hierarchical methods produces 0.90 and 0.77 Cophenetic and Baker correlation coefficients respectively. These results suggest that the two dendrograms are similar, but not dentical. This comparison can be extended to as many dendrograms are considered. Table 3 lists the adjusted Rand and Fowlkes and Mallows indices for the clusterings under consideration. Based on these coefficients we can conclude that hierarchical average and model-based clustering give almost identical result, complete is close, but bot identical to the first two, whereas k-means clustering has a relatively low degree of similarity with all other clustering results. In a real analysis situation, the list of clustering results under consideration is expected to be much longer than four. If one particular clustering is systematically apart from the majority, it would suggest that this clustering is weak. If on the contrary, a clustering is confirmed with its comparison with many others, in the sense of agreement, this clustering is a strong candidate.

	Adjusted Rand				Fowlkes and Mallows			
	Average	**Complete**	**k-means**	**Model-based**	**Average**	**Complete**	**k-means**	**Model-based**
Average	1				1			
Complete	0.89	1			0.93	1		
k-means	0.46	0.46	1		0.66	0.66	1	
model-based	0.97	0.87	0.47	1	0.98	0.92	0.66	1

Table 3. Adjusted Rand and Fowlkes-Mallows indices calculated for the four clustering results on simulated data.

After calculating as many as possible correlation/similarity indices for the derived clusterings, we proceed with examining the source of disagreement. For the numerical example, if we calculate the classification table between the four examined clusterings using labels 1, 2 and 3 to denote the three clusters, Table 4 lists the results. The degree of similarity between Average and Complete hierarchical clusterings is explained because they only differ at the cluster membership of five observations. Average linkage classifies those in cluster 2, while complete linkage in cluster 1. The data here are simulated and we know that three groups of sizes 15, 20 and 40 have been generated. It is easy to confirm here that average linkage agrees with the grouping as simulated. Model-based which scores high in similariry with hierarchical clusterings, differs with the Average linkage in classification of one out of seventy five observations. For k-means, the two smaller groups 1 and 2 are merged to one, and the third group of forty observations is split into two others. This is the more distant to the true clustering among the four proposals.

		Complete			k-means			Model-based		
		1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
	1	15	0	0	0	15	0	14	1	0
Average	**2**	5	15	0	0	20	0	0	20	0
	3	0	0	40	21	0	19	0	0	40

Table 4. Cluster membership comparison for average and complete linkage dendrograms.

Internal and External cluster validation.

Apart from comparing the clustering results to conclude on which classification is more dominant, a validation of a specific clustering would be valuable to assist in deciding which analysis is the analysis that produce trustworthy results, closer to reality. Clustering validation or assessment is not easy nor has a unique solution. Two approaches are proposed here as dominant for the problem. The first approach is based on specialized indices or measures that

measure the quality of clustering. Secondly, a graphical representation of a clustering result under consideration against the data may be used.

Silhouette coefficient (Rousseeuw, 1987) is the most popular measure to assess the internal validity of a clustering. Silhouette coefficient is a measure which shows how well the objects lie within the clusters they are assigned to. The silhouette coefficient takes values from minus one to one and a value close to one corresponds to a well clustered observation whereas a negative value indicates the worst situation with respect to clustering.

Example 1 (continued). Figure 6, plots the Silhouette coefficients for all the observations of the simulated dataset at Example 1, according to the clustering results presented in Figure 3. The average score of Silhouette coefficient is higher for the Average Linkage (0.52) in comparison with Complete (0.49) indicating a relatively better clustering. Moreover the Silhouette coefficient for five observations according to complete linkage is negative, i.e. these observations are not well clustered. It is easy to verify that these are the same five observations that the two clustering results assign in different groups as listed in Table 4. The Silhouette coefficient for k-means method is 0.31, a lower value than 0.52 for the Average linkage, a fact that advocate for the inferior performance of k-means in this certain example.

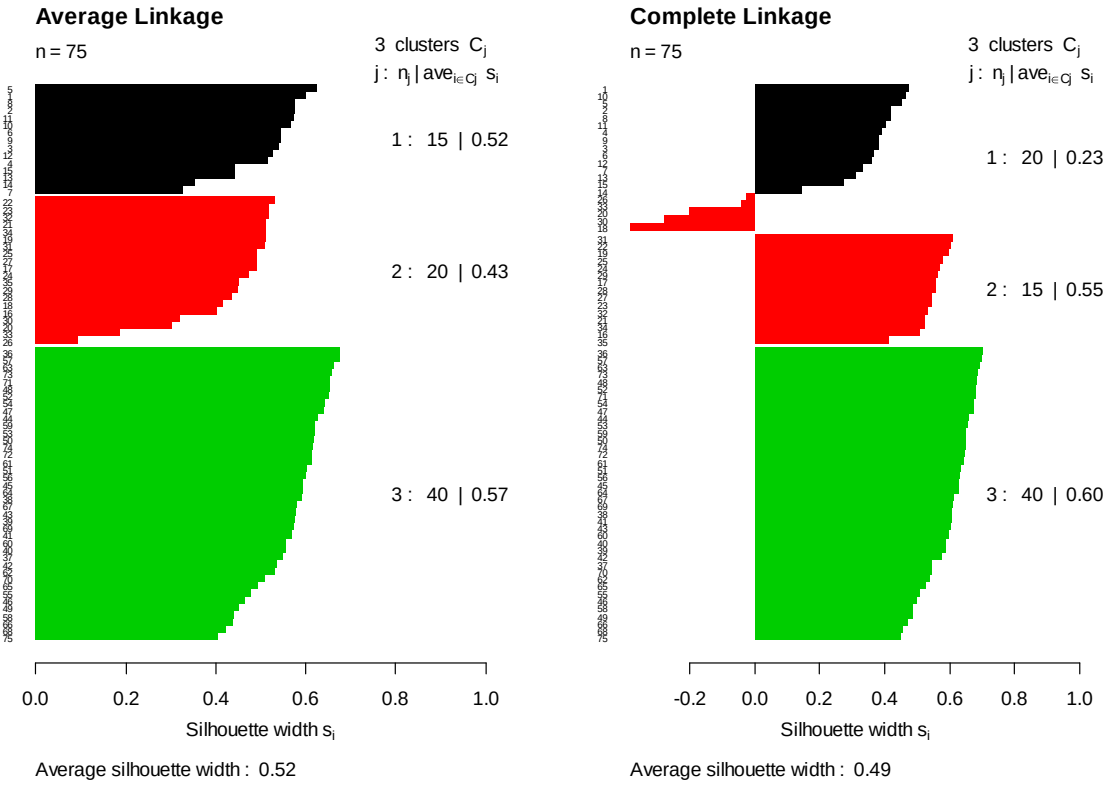

Figure 6. Silhouette coefficients for a three cluster result according to average and complete linkage.

Lambda-Wilk's test is another statistic that in this context can provide a measure of compactness within each proposed cluster and how well clusters are separated from each other.

The smallest value for Lambda-Wilk's test the better separation among the groups. For the data at Example 1, Lambda-Wilk's test for all considered methods are very small, within the range of 0.02 to 0.05.

The second approach towards the assessment of a clustering result is a graphical method. A clustering proposal is visualized by plotting the data using the labeling provided from the clustering. The plot can either give grounds to accept a clustering when this gives a sensible result when plotted, or on the contrary, it can be used to explain why a clustering result performs poorly. In order to plot a clustering result on data the need of a lower dimension coordinates system is required. For this reason, a technique of data dimension reduction is necessary and Principal Component Analysis is the most frequently used in Archaeometry.

Principal Component Analysis

Principal Component Analysis (PCA) (see for example Jolliffe, 2002 as a general reference and Rogers et al. 2016 for a practical guide) is a data transformation technique which exploits the correlation among the data variables, in order to construct a new set of variables, the Principal Components (PC) with some good properties. The first property is that the two data sets are equivalent with respect to the amount of information included as this measured from the total variability of the data points.

The advantage of PCA is that the variability reconstruction is such that there are no correlations among the principal components, i.e. the new set of variables is uncorrelated which means that a univariate study of all variables will reveal all included information. Moreover, there is a hierarchy is Principal Components importance. The leading ones explain the majority of the total variability allowing us to retain only a subset of PCs and explain a significant percentage of the total variability based only on those PCs. Especially in applications, such as Archaeometry, that data variables are highly or moderately correlated, two or three of the leading PCs can explain a sufficient amount of the total data variability. Another characteristic of PCs is that they are linear combinations of the original variables. This very simple form of transformation enables the scientist to interpret the PCs with respect to the original variables. This is based on the loadings, as they called the coefficients of the linear combinations of PCs.

Before applying PCA in archaeometric data, transformations as discussed in Section 2 will be necessary. If raw data are used, and not logged or log-ratio transformed the standardization of the data in zero mean and variance one will be needed. This is because if the variances among the data variable vary a lot, the leading PCs will be dominated from those with large variance. In this case the leading PCs will not explain the majority of the data set variability across all variables, but will explain the variability of the data with respect to the variables with large variance only. Especially in compositional data where the measurements for some oxides account around 50% of the total composition and for some trace elements the measurements are particles in millions

their corresponding variance will inevitably vary and standardization will be essential. If log transformation or log-ratio transformation results to a data set that the variables have comparable variances, a further standardization is not necessary, at least with respect to PCA performance. Another practical issue that may arise in PCA implementation is the presence of outliers. We discuss this issue at the dedicated to outliers paragraph as they affect not only PCA, but also cluster analysis performance.

Example 1 (continued). Using the simulated data of Example 1, we apply PCA on the standardized data and the first PC explains 57.1% of the total variation, while the second 19.8%. A graph on the two first PC will combine 77% of the total variability of the original data. This is a vast amount of the variability of the original data and it could be extremely useful in visualization of a clustering result, but yet a remaining 23% may be informative as well. This is why a plot of the first against the third PC may be also needed to be checked. Figure 6 plots the simulated data of dimension nine, at a two dimensional space using the first two principal components. The PC analysis on standardized, log-transformed with basis 10 and log-ratio transformed data is plotted and observations are labeled with respect to the true origin, since the data are simulated.

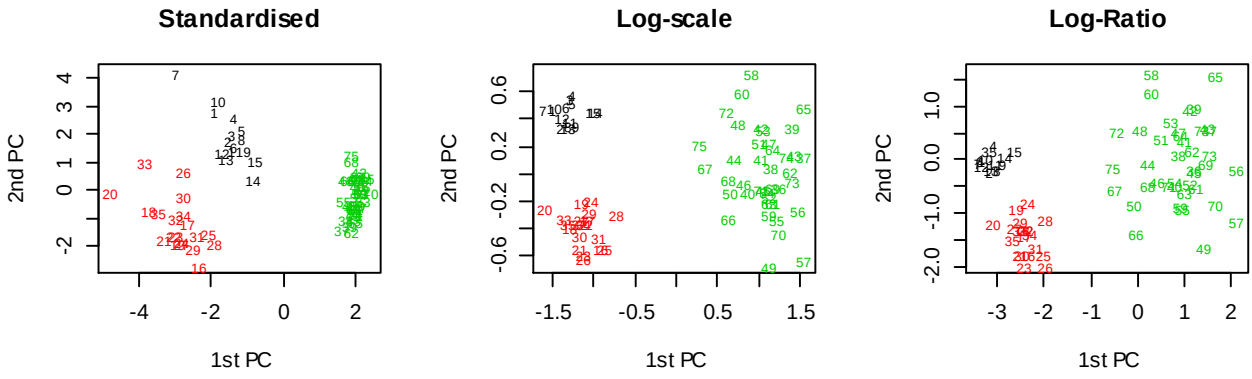

Figure 6. Data plotted in the first two PCs for standardized data (left) and log-ratio transformed data (right).

To use PCA as a graphical way of clustering assessment, we plot all under consideration clustering results into the first two PCs. Figure 7 plots clusterings listed in Table 4. In practice this list may be quite long. It is easy in our working example to verify that Average linkage and model-based method propose a compact result that captures the heterogeneity. According to Complete linkage, five observations seem to have been misclassified and k-means produces the less sensible result because the very compact group at the right of the plot is separated into two smaller groups.

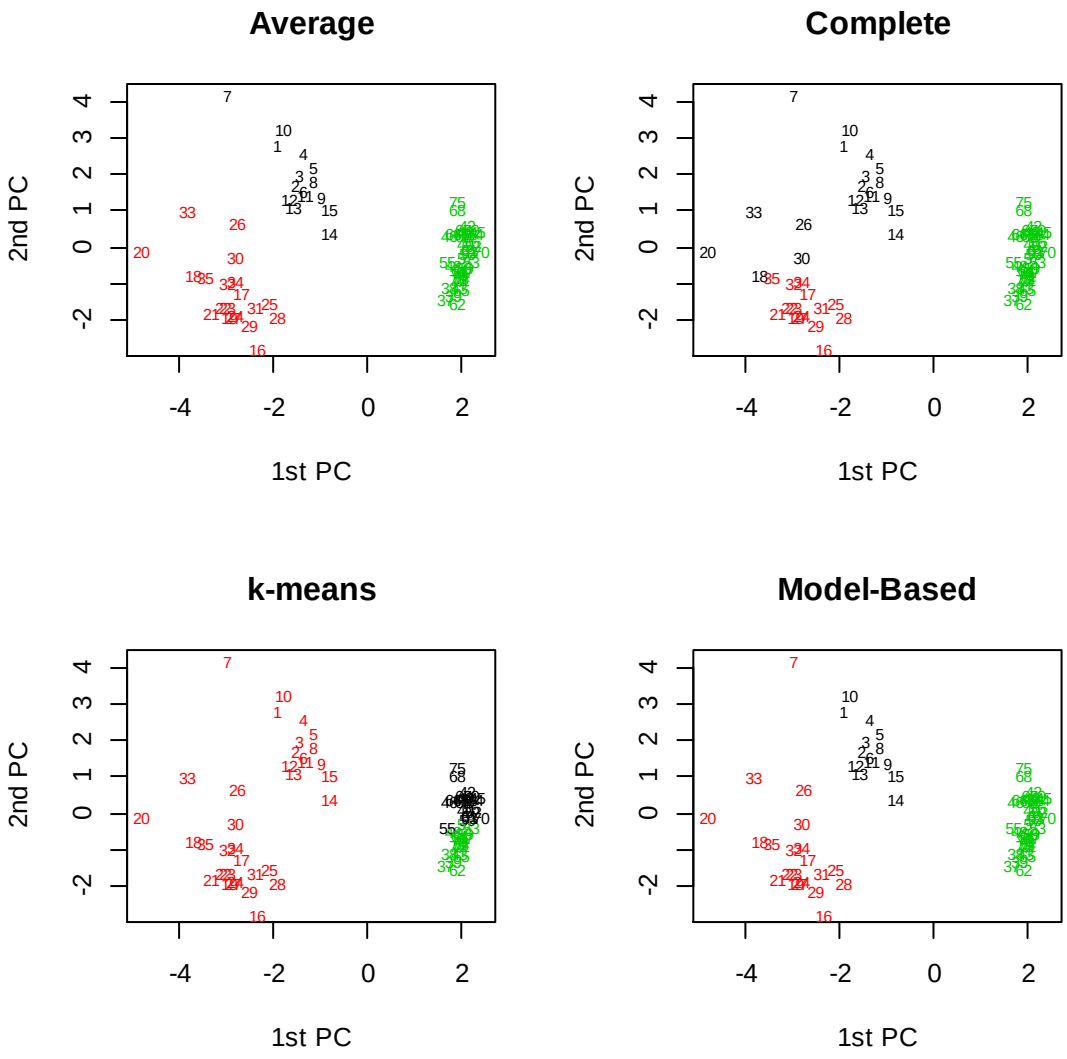

Figure 7. Clustering results from previous analysis on simulated data of Example1 plotted on the first two PCs.

PCA can also be used for characterizing the derived provenance groups. This can be done by using the loadings of PCA with respect to the original data. For the example, the loadings for the first two PCs are given in Table 5. Graphically, the loadings together with a scatter plot of the observations into the two first principal components are plotted in a biplot. The biplot of the simulated data is plotted in Figure 8. By inspection of Table 5 and Figure 8, we conclude that, for example, measurements in group 1 are characterized by higher values in CaO, Na_2O and BaO. Group 2 is characterized by higher values in MgO, MnO and K_2O and at the same time lower values in Al_2O_3 and TiO_2. The special characteristics of group 3, or in other words the factors which discriminate this group from the others is the high measurements in Al_2O_3 and TiO_2. Figure 8

presents the box-plots of three selected chemical elements for the simulated dataset, using the labels of the group membership as derived from Average Linkage which coincides with the true clustering. It is apparent that these selected elements, among others, can play the role of discriminating factors of the three groups.

	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	MnO	BaO
PC1	0.30	-0.39	-0.40	-0.22	-0.32	-0.42	0.32	-0.42	-0.04
PC2	0.42	0.23	-0.22	0.52	0.40	-0.09	0.26	-0.04	0.46

Table 5. Loadings of the first two PCs.

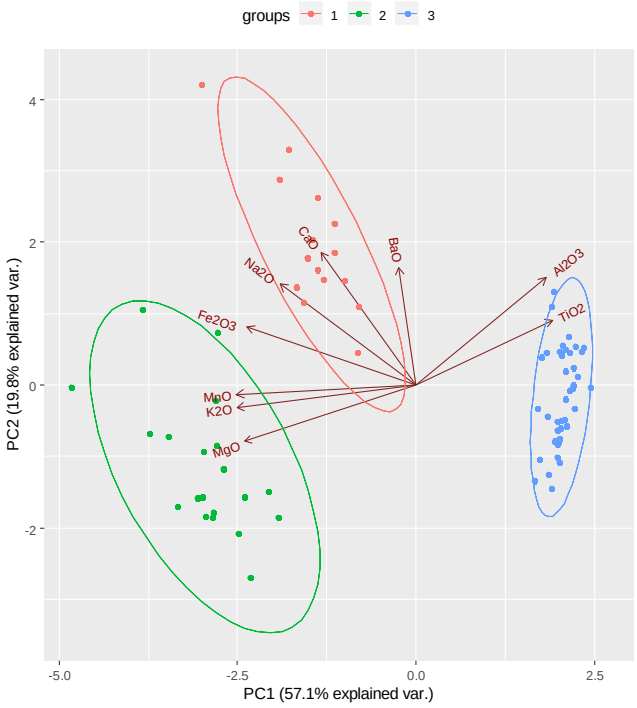

Figure 8. Biplot of the first two PCs obtained from the simulated data of Example 1.

Summarizing, PCA can be used as a first graphical test to investigate and possibly detect chemical compositional structure in the data. If this is present a cluster analysis would be a sensible next step of the analysis. Labeling cannot be possible at this stage, but still a certain structure, if present, can be detected. This suspected structure can further pursued by a cluster analysis. Moreover, if a structure is present and coherent all possible transformations would be able to capture this. If a structure is suggested only with some data transformations usually means something.

PCA at a later stage of the analysis can be used for cluster validation and lastly, PCA can assist to derive conclusions on the special characteristics that each group possesses.

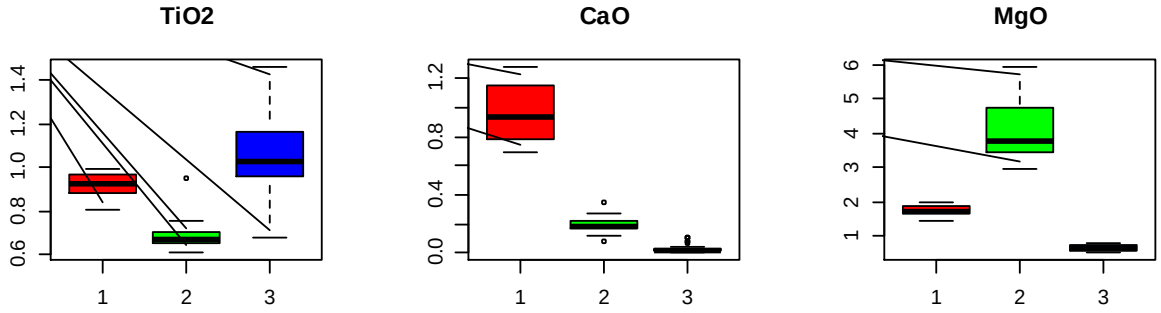

Figure 8. Box-Plots of selected original measurement variables.

Apart from PCA technique used as a data projection in to a lower dimension system of coordinates, other statistical learning or machine learning techniques can also be used. For example, factor analysis, multidimensional scaling.

Categorical Data and Mixed-mode data

Apart from the geochemical continuous data, other source of information leading to quantitative discrete data may also be available. Such information may result from petrographic examinations, e.g. optical microscopy examination of thin-sections.

To accommodate this type of information in the analysis, if quantitative, one can proceed with two ways. The information provided from discrete data can be used complimentary. A cluster analysis based on continuous data can be conducted independently and a cross examination from groups suggested from the discrete data can be used as confirmation or explanation of the cluster findings obtained from the chemical information only. Alternatively, but less often used in practice, one can merge both types of variables to a single, integrated analysis, that takes into account both types of data continuous and discrete. This type of analysis is introduced in archaeological application from Baxter et al. (2008) with the name mixed-mode analysis. The merits of such analysis are that it is more informative and avoids the problem of contradictive results suggested from two separate cluster analysis.

Baxter et al. (2008) methodology for the mixed-mode approach of the problem is based on a generalization of Gower coefficient of similarity and weighting the contribution of continuous and discrete data. This weighting aims to avoid the domination of binary data over the continuous

towards clustering. Papageorgiou and Moustaki (2005) propose latent class models as a model-based method appropriate to answer the cluster problem while the input variables can be continuous, binary, categorical ordinal or nominal. They successfully implement this methodology to an Archaeological dataset. Recently, Androulakis et al 2018 proposed a model-free approach for the mixed-mode problem based on a modification of Gower coefficient of similarity.

However, although methodologies have proposed in the literature for dealing with the problem of different type of information adopting the integrated approach, in practice the use of this approach is limited. This is because special software, sometimes not freely available, is necessary or they are computational methods and therefore are demanding in technical and/or computational effort.

Outliers

Another characteristic of Archaeometric data is the presence of outliers, i.e. observations that deviate from the remaining sample. This may be due a deviate measurement of this observation in one or two variables, or a multivariate outlier where the difference of this measurement is more general and it may be due to an unexpected relation among the measured variables for this particular observation. For example, two or more variables may be strongly positively related, i.e. strong positive correlation, and the measurements on these variables indicate negative relation for a particular data point. If a plot of the data at these dimensions was possible, the bulk of data points would be expected to form an ellipsoidal (not spherical shape) due to the strong correlation, and the outliers would lie outside this ellipsoidal. Detection of univariate or multivariate outlier is essential for a statistical analysis because most of the methodologies presented in previous paragraphs are affected from the presence of outliers and their performance degenerate.

Detection of univariate variables is relatively straightforward problem and univariate techniques, such as boxplots, histograms etc. can identify those outliers, or better extreme values on certain variables. Detection of multivariate outliers is however a more challenging problem and a number of approaches have been proposed in the literature (see Filzmoser et al. 2005 and Rousseeuw et al 1990 amongst others). The most commonly used technique to identify multivariate outliers in the literature is the Mahalanobis distance. Assuming a set of p-dimensional observations $\boldsymbol{x}_i = \left(x_{i1}, x_{i2} \dots, x_{ip}\right),\ i = 1,2, \dots, n$ Mahalanobis distance for observation i is defined as

$$d_M(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})' \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})}$$

where $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are the mean and variance covariance matrix of the theoretical population that has generated the sample and $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ are some estimates of those parameters based on the data. Usual estimates for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are the sample mean and sample variance covariance matrix $\bar{\boldsymbol{x}}, \boldsymbol{S}$ respectively. If the population distribution is multivariate normal the distribution of d_M^2 can be proved to be chi-square with p degrees of freedom, X_p^2. Under this assumption any observation that has a d_M value at the outer, say 95% probability area, for the X_p^2 distribution is characterized as extreme and therefore outlier. Note that d_M expression takes into account all measured variables, although the distribution is univariate, and therefore conclusions on multivariare outliers can be obtained. For example, if $p = 14$ the cut-off point for a 95% confidence ellipsoidal is $X_{14,0.95}^2 = 23.68$. According this measure, any data point $\boldsymbol{x}_i$ for which the square of $d_M(\boldsymbol{x}_i)$ is greater than 23.68 is considered as outlier.

However, estimates $\bar{\boldsymbol{x}}, \boldsymbol{S}$ are sensitive to the presence of outliers themselves, or in statistical terminology, they are not robust as estimates when one or a small set of observations deviate from the rest. This has as a result Mahalanobis distance in turn to be affected from the presence of outliers. Since, its practicality as a measure, is essential when outliers are indeed present in the data, various modifications of Mahalanobis distance as a tool for outlier detection are proposed in the literature. Modifications are mainly consisted of using other, robust estimates, of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ instead of $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ (Rousseeuw and Van Zomeren, 1990).

Once outliers are identified we proceed with the cluster analysis by keeping those special observations aside. A further examination of the set of outliers with respect to their compositions will possibly reveal the reason why they diverse from the remaining data set. Outliers may be quite important observations and meaningful in archaeological terms, but as far as the clustering problem is concerned they do not cluster with any of the existing groups and they will only confound any clustering technique.

5. Classification

Classification is a supervised multivariate technique and refers to the problem of classifying a new observation to one of the known identified groups in the data. It is an interesting problem in Archaeometry when a clustering at previous stage has been conducted and a number of groups corresponding to different origin have been established. If new samples, artifacts in this case, are available from another study classification will be the appropriate method to assign the new artifact or artifacts to one of the existing groups and consequently make conclusions about its origin.

As in clustering, classification is a multivariate statistical problem and a number of approaches and techniques can be applied in this context. One approach is to consider the

problem at a probabilistic manner where either a prediction model is used to predict the classification category for the new observation or the probability for the unclassified observation to belong to each one of the existing groups is calculated and assignment is based on the maximum probability. Within this framework belong the methodologies of linear regression, logistic regression, multiple logistic regression methods, Bayes classifier and Naive classifier.

Discriminant Analysis (DA) is another widely used technique for classification problem where the task is to construct a classification rule which discriminates the groups. The rule can be linear or quadratic leading to linear discriminant and quadratic discriminant analysis respectively. The resulting rule may also be useful to visualize the data in a lower dimension space than the original dimension of the data. This visualization can be considered as an alternative to PCA projection. PCA and DA utilize a different criterion according to which the components are constructed. In particular, DA takes into account the information of existed groups and DA components result as the projection of the original data which maximizes the between groups variation. Other classification techniques are the K nearest neighbor and classification trees. James et al. (2013) is a reference book with emphasis the applications of the classification.

We present here only an illustration of classification trees technique, because this is the less mathematical, no assumptions about the population distributions are required, it can accommodate both continuous and discrete data and most importantly the results are easy to interpret. A software is necessary, but the implementation is straight forward.

Example 2. For this experiment we assume that the dataset consists of a 73x24 data matrix with ten variables to represent continuous variables of chemical measurements and 14 variables as categorical, binary or ordinal. Three groups have been identified and labels ‘gr1’, ‘gr2’ and ‘gr3’ have assigned to those groups.

Applying a classification tree method on the data using the group label results to the tree plotted in Figure 9. From the 24 variables, only three variables corresponding to equal number of nodes are sufficient in order to construct a discriminant rule. The rule is apparent and written on each branch of the tree. For example is the new subject has measurement feldspar variable 2,4 or 6, the right branch is suggested where another test will follow with respect the value of the new subject in Na2O and more specifically if it less than 0.925 or not.

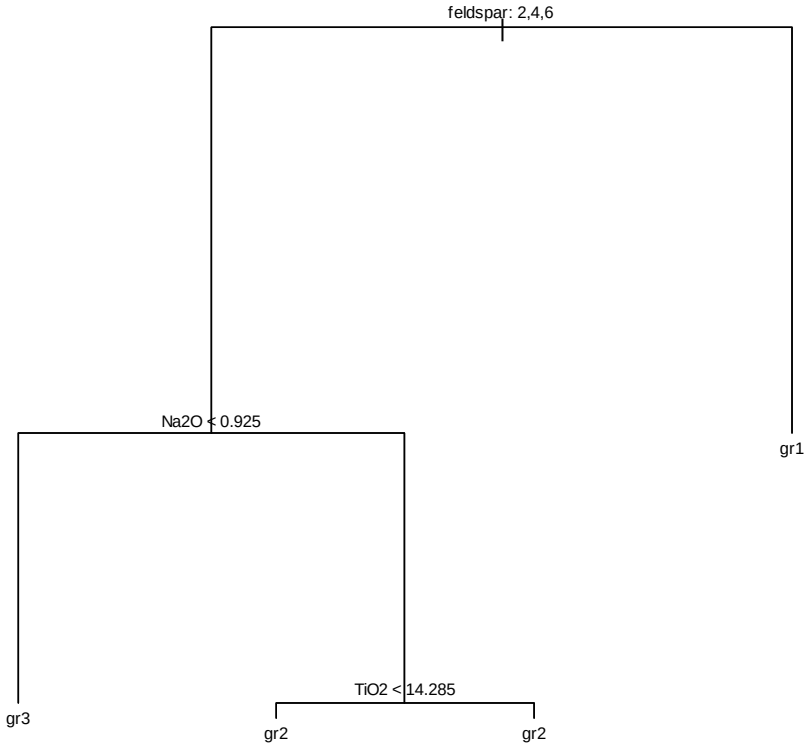

Figure 9. Classification tree.

6. Conclusions

Summarizing, we list the structure of the statistical analysis in steps.

Step 1. Run a preliminary analysis of the data, without any transformation, including univariate and bivariate study. Identify departure from normality, long tail for variables, outliers (univariate or multivariate using Mahalanobis distance) and possible structure in the data.

Step 2 Data transformation.

Step 3. Implement PCA to transformed data in order to (i) identify or confirm outliers located at step 1 and (ii) check if data (using all measured variables) suggest the existence of different groups.

Step 4 If PCA gives grounds to a non-homogeneous dataset, implement cluster analysis by using various methods and ways of data transformations. Verify outlier indicated by PCA

or Mahalanobis distance and confirm with data inspection. Remove those and repeat cluster analysis. Identify clear groups, inference about these groups and justify the result. Set aside the distinct groups and repeat cluster analysis.

Step 5. In the process of cluster analysis, comparison and assessment of the clustering results as proposed from various methods is performed.

Step 6. For the compact groups of the analysis proceed with their characterization and determine the discriminating factors among groups.

Step 7. If discrete data from microscopy study are also available and have a quantitative nature, a cluster analysis for this type of data can also be implemented and results with respect to clusters suggested can be seen in comparison with the clustering suggested from the continuous data.

Step 8. Mixed-mode cluster analysis is suggested if quantitative discrete data are also available.

Step 9. A classification analysis is appropriate when the aim is to classify a new subject, an artifact, into one of the existing identified groups.

Closing this article, I would like to borrow a paragraph from Whallon (1984) that Baxter (2015) also uses as an introduction to chapter 11. The paragraph is "Archaeologists are ill-trained to, the rigorous and logical thought necessary form an informed use of quantitative methods, while the rare statisticians who have tried their hands at archaeology, typically have understood the nature of archaeological data, questions, and models only partially, vaguely, or incorrectly, so that their efforts are usually no better than the archaeologist's own". As mentioned in the Introduction section, Archaeologists now-days are more educated in using quantitative tools. However, interaction and collaboration between the archaeologist and the statistician are key components that cannot be substituted from technology and training. To my opinion any result is safe only when it can be confirmed from both sciences.

Bibliography

Aitchison, J (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

Binford, L. R. (1964) A Consideration of Archaeological Research Design, *American Antiquity*, **29**, 4, 425-441.

Banfield, J.D. and Raftery, A.E. (1993) Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, Vol. 49, 803-821.

Baxter, M.J. (2015) Exploratory Multivariate Analysis in Archaeology (Foundations of Archaeology). 2nd Edition. Eliot Werner Publications/Percheron Press.

Baxter, M.J. (2015) Spatial k-means clustering in archaeology – variations on a theme. Working paper – November 2015 (accessed in Academia.edu).

Baxter, M.J., Beardah, C.C., Papageorgiou, I., Cau, P.M., Day, P.M. And Kilikoglou, V. (2008). On Statistical Approaches To The Study Of Ceramic Artefacts Using Geochemical And Petrographic Data. *Archaeometry*, **50**, 142–157. Https://Doi.Org/10.1111/J.1475-4754.2007.00359.X

Baxter, M. J. (2006) A review of supervised and unsupervised pattern recognition in archaeometry. *Archaeometry*, **48**, 4, 671–694.

Baxter, M.J. (2001) Statistical modelling of artefact compositional data. *Archaeometry* **43**, 1, 131-147.

Baxter, M.J. (2008) Mathematics, Statistics and Archaeometry – The last 50 years or so. *Archaeometry* **50**, 6, 968–982.

Baxter, M.J. (1995) Standardization and Transformation In Principal Component Analysis, with Applications to Archaeometry. *Applied Statist.* (1995) ***44,*** *No.4, pp. 513-527.*

Beardah, C. C., Baxter, M. J., Cool, H. E. M., and Jackson, C. M. (2003) Compositional data analysis of archaeo-logical glass: problems and possible solutions, in CoDaWork'03: Compositional Data Analysis Workshop, Girona, Spain; available at http://ima.udg.es/Activitats/CoDaWork03/paper_baxter_Beardah2.pdf

Bieber, A.M. JR., Brooks, D.W., Harbottle, G. and Sayre, E.V. (1976) Application of Multivariate Techniques to Analytical Data on Aegean Ceramics. *Archaeometry* 18, 59-74.

Buxeda i Garrigós, J. (1999) Alteration and Contamination of Archaeological Ceramics: The Perturbation Problem. *Journal of Archaeological Science*, 26, 295-313.

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.

Everitt, B.S., Landau, S. and Leese, M. (2011) *Cluster Analysis*, 5th Edition. John Wiley & Sons.

Everitt, B.S. and Dunn, G. (2010) *Applied Multivariate Data Analysis*. 2nd Edition. John Wiley & Sons.

Filzmoser, P., Garrett, R.G. and Reimann, R. (2005) Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, **31**, 579-587.

Gordon A.D. (1998) Cluster Validation. In: Hayashi C., Yajima K., Bock HH., Ohsumi N., Tanaka Y., Baba Y. (eds) Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Tokyo.

Hubert, L. and Arabie, P. (1985) Comparing Partitions, *Journal of the Classification*, 2, 193-218.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Kassambara, A. (2017) Practical Guide To Cluster Analysis in R. Unsupervised Machine Learning: Volume 1 (Multivariate Analysis). STHDA Publishing.

Martín-Fernández,J.A., Buxeda i Garrigós, J. and Pawlowsky-Glahn, V. (2015) Logratio Analysis in Archeometry: Principles and Methods. In Barcelo,J.A. and Bogdanovic, I. (Eds) Mathematics and Archaeology. CPC press, Boca Raton FL, 178-189.

Palarea-Albaladejo, J. and Martin-Fernandez, J.A. (2015) zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems. **143**, 85-96.

Papageorgiou, Ioulia. (2018) Cluster Analysis. The SAS Encyclopedia of Archaeological Sciences, John Wiley & Sons. https://doi.org/10.1002/9781119188230.saseas0099

Papageorgiou, I., Baxter, M.J. and Cau, M.A (2001) Model-Based Cluster Analysis of Artefact Compositional Data. *Archaeometry*, **43**, 571-588.

Papageorgiou, I. and Moustaki, I. (2005) Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics & Data Analysis*, **48**, 659 – 675.

Pawlowsky-Glahn, V. & Buccianti, A. (2002) International Journal of Earth Sciences, **91**, 357–368. https://doi.org/10.1007/s005310100222

Pawlowsky-Glahn, V. & Buccianti, A. (2011) *Compositional Data Analysis: Theory and Applications*. Wiley.

Rogers, S. and Girolami, M. (2016) A First Course in machine Learning. 2nd Edition. Chapman and Hall.

Rousseeuw, P. and Van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633-639.

Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Saraçli, S., Doğan, N. & Doğan, İ. (2013) Comparison of hierarchical cluster analysis methods by cophenetic correlation. *J Inequal Appl* 2013, 203. doi:10.1186/1029-242X-2013-203

Shotwell, M.S. (2013). profdpm: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, 53, 1-18. URL http://www.jstatsoft.org/v53/i08/.

Steinley, D. (2006) K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59, 1–34.

Whallon, R. (1984) Unconstrained Clustering for the Analysis of Spatial Distributions in Archaeology. In *Intrasite Spatial Analysis In Archaeology*, edited by H. J. Hietala, pp. 242-277. Cambridge University Press, New York.

Xu, R. and Wunsch-II, D.C. (2008). *Clustering*. Wiley, John & Sons, Inc.