Journal of Archaeological Science: Reports 19 (2018) 100–114

Contents lists available at ScienceDirect

Journal of Archaeological Science: Reports

journal homepage: www.elsevier.com/locate/jasrep

Statistical processing of compositional data. The case of ceramic samples from the archaeological site of Xalasco, Tlaxcala, Mexico

Pedro López-García[a,*], Denisse Argote-Espino[b], Kamila Fačevicová[c]

[a] Posgrado de Arqueología, Escuela Nacional de Antropología e Historia, Periférico Sur esq. Calle Zapote, Col. Isidro Fabela, Deleg. Tlapan, Mexico City, Mexico
[b] Dirección de Estudios Arqueológicos, Instituto Nacional de Antropología e Historia, Lic. Primo Verdad no. 2, Col. Centro, Deleg. Cuauhtémoc, Mexico City, Mexico
[c] Department of Mathematics, Faculty of Science, Palacký University, 17, Listopadu 12, CZ-77146 Olomouc, Czech Republic

ARTICLE INFO

Keywords:
Archaeometry
Compositional data analysis
Statistical processing
Ceramics
Xalasco
Tlaxcala
Mexico

ABSTRACT

This paper focuses on the implementation of a methodology for identifying chemical groups, the search for data patterns and the determination of outliers in a collection of ceramic samples after compositional analysis. In order to obtain the chemical composition, pottery samples from the archaeological site of Xalasco were analyzed by means of a portable XRF technique. The discrimination of groups was based on the methodology proposed by Aitchison in 1986 for the analysis of compositional data. First, we applied a stepwise method for the selection of variables with the use of biplots. The detection of outliers was managed by using the method known as FAST-MCD. In a second stage, the detected groups and their components were analyzed in a supervised manner using the PLS-DA method. Once the groups were validated, we obtained the balances and the CoDa-dendrogram with the isometric transformation to appreciate the differences in their chemical composition. The results showed a high degree of accuracy in the discrimination between groups, which can be interpreted as different manufacturing processes. The chemical differences among the samples can be understood in terms of the cultural influences reflected in the preparation of the ceramics pastes through time, the availability of raw materials and the organization of pottery production.

1. Introduction

In recent decades, significant efforts have been made in experimental research by using different instrumental techniques for the study of archaeological materials, which provide qualitative and quantitative information of the analyzed material. Nowadays, these techniques have proven to be fundamental tools for the characterization and classification of a wide variety of archaeological artifacts. The new perspective in specialized archaeometric studies focuses on the identification of groups of artifacts with similar chemical compositions, looking for patterns through quantitative methods. Although several studies have adopted the concept of compositional data in the analysis of archaeological materials, few observe the theoretical assumptions of these methods.

In archaeology, several applications have searched for the establishment of chemical groups, using multivariate methods. When analyzing compositional data, several issues have to be considered. The first problem is the lack of unified criteria for the analysis of this type of data, especially in the pre-processing stage. On the other hand, there is the problem of the outlier identification, which can greatly influence the results of the analysis. Lastly, due to the high amount of existing multivariate methods (i.e. PCA, LDA, PLS, CA, PLS-DA, among others), the selection of the most adequate technique for processing the data has to be taken in account. These three criteria are intimately related with the solution of the problem of compositional data.

In the case of data pre-processing, the main debate is focused in the type of transformation to be used; in other words, the available transformations to "open" closed data. Among these transformations, we can mention standardization, logarithmic transformation ($\log_{10}$), the use of log-ratios and standardization of the data after the use of the logarithmic or log-ratio transformations. This issue was approached by Baxter (1995), Baxter et al. (2005), Baxter and Freestone (2006) and Beardah et al. (2003), who tested the behavior of the data by using these transformations in several databases of archaeological materials and applying multivariate techniques to them, evaluating the performance of the algorithms to reveal the existence of patterns with an archaeological significance.

Aitchison et al. (2002) emphasize data analysis through log-ratios as a meaningful and interpretable methodology for archaeometry, highlighting that the underlying and necessary principles of the analysis of

* Corresponding author.
E-mail address: dplopez14@gmail.com (P. López-García).

https://doi.org/10.1016/j.jasrep.2018.02.023
Received 7 August 2017; Received in revised form 15 February 2018; Accepted 18 February 2018

compositional data should be taken into account. For this type of data, it is typical that the relevant information be contained in the ratios between parts, rather than in its absolute values, and that they follow Aitchison geometry, respecting its relative structure (Aitchison et al., 2002; Pawlowsky-Glahn et al., 2015). Due to the relative character of the geochemical data, application of standard statistical methods, which mostly rely on Euclidean geometry, may conduct to misleading results owed by their impossibility to properly interpret the covariance and the correlation coefficients due to the restriction of constant sum.

Beardah et al. (2003) suggest that the structure of certain kinds of archaeometric data is such that common log-ratio analysis will sometimes produce results with no substantive meaning, even when substantively interpretable structure exists. They also mention that the standardization of the data after the centered log-ratio transformation and the merge of some components work well, but the produced results are similar to the standardized data analysis. On the other hand, Glascock (1992) proposes the $\log_{10}$ transformation, assuming a log-normal distribution of the data. He argues that compositional data, mainly trace elements, mostly follow a normal distribution and that the logarithmic transformation compensates for the differences in the magnitudes between the major elements (i.e. Al, K and Fe) and the trace elements (i.e. REE). Baxter (1995) debates the use of the $\log_{10}$ transformation, pointing out that the manufacturing of archaeological artifacts involves the use of diverse raw materials, which can produce multi-normal statistic distributions and not necessarily log-normal. However, he mentions that, when comparing normal standardized data with $\log_{10}$ standardized data, the differences are minimal as long as the data is clean from outliers.

Filzmoser et al. (2009) also compares the behavior of univariate compositional data by using several transformations like log(x), log (x)/(1 − x) y *ilr*(x), where *ilr* refers to the isometric transformation. They discuss the employment of first and second order statistics, as well as graphic tools, to examine the effect of the different transformations in the results. In their analyses, they found that the original data as well as the data with a $\log_{10}$ transformation presented skewed distributions with the occurrence of outliers, while the *ilr* transformation displayed a greater tendency to normality. In the case of central trend and dispersion statistics, the computations performed with the original and the $\log_{10}$ transformed data are inadequate because the geometry in which the calculations are made is inconsistent with the Aitchison geometry. Only the *ilr* transformation allows the right geometric representation of compositional data in the Euclidean space, where standard statistical methods can be applied.

Having mentioned some basic aspects about the type of transformations employed in compositional data, it is worth mentioning that the analyses performed by Baxter (2008), Baxter and Freestone (2006) and Beardah et al. (2003) are based in the additive log-ratio (*alr*) and the centered log-ratio (*clr*) transformations, which present certain inconvenient for the multivariate analysis. The *alr* transformation is not recommended because it does not result in an orthogonal basis system, which is necessary for diagnostic tools following outlier detection (Filzmoser et al., 2012). On the other hand, the *clr* transformation preserves distances but leads to a singular covariance matrix (Pawlowsky-Glahn et al., 2015). In this work, we focus on the *ilr* transformation, which allows us to analyze relative information avoiding the issue of singularity of the *clr*. The most important property of the *ilr* transformation is its isometry, meaning that it relates the geometry on the simplex directly to the Euclidean geometry (Egozcue et al., 2003). By using this transformation along with robust methods, better results are obtained than with the other approaches.

Other debate arises around the decision of applying unsupervised methods versus supervised methods; the selection depends mostly on the purpose of the analysis. Exploratory methods employ unsupervised techniques, while classification methods use supervised techniques. One of the most common non-supervised strategies utilized for the exploration of archaeological data are the principal component analysis (PCA) and the cluster analysis (CA). The PCA explains the structure of the covariance of the data with only few components, but is very sensitive to outliers since it is based in the classic location and scatter estimators and the concept of variance. Outliers can greatly influence the scores and loadings, sometimes even to the extent that they will dominate the first PCs (Wehrens, 2011). Consequently, the first components are often attracted toward outlying points, and may not capture the variation of the regular observations. Therefore, data reduction based on classical PCA becomes unreliable if outliers are present in the data (Hubert et al., 2005). On the other hand, the selection of the first two components for the interpretation of the data is not necessarily adequate, since they may not display all the relevant information. It is therefore advisable to perform cross validation and bootstrap techniques for a statistically based estimation of the optimum number of PCA components (Varmuza and Filzmoser, 2009).

The CA defines groups based on the estimation of the similitudes within the samples through different metrics, like the Euclidean distance, which are of heuristic nature. Moreover, assessing the quality of the clustering is tricky since the “real” clustering is by definition unknown (Wehrens, 2011). Other problems are the selection of the optimal number of groups, the election of the proper metric, the detection of outliers and the use of validation techniques. The statistical properties of this method are generally unknown, precluding the possibility of formal inference (Fraley and Raftery, 2002). Due to the aforementioned issues, the PCA and the CA should be used in the first stage of the exploration of the data; performing a posterior supervised analysis for their validation is recommended.

From the chemometrical point of view, classification can be defined as finding a mathematical model capable of recognizing the membership of each object to its proper class (Ballabio, 2006). Some of the classification techniques commonly used in archaeology are the Mahalanobis distance (MD) and the Discriminant Analysis (DA). Nowadays, there are other statistical tools for the prediction of memberships of groups, like the Partial Least Squares Discriminant Analysis (PLS-DA). This method combines the properties of the PLS regression with the DA, optimizing the separation between the different groups present in the samples (Kalivodová et al., 2014). Unlike non-supervised methods, the PLS-DA can evaluate the overall quality of the model, providing a better description of the data; in other words, the better discrimination of the samples into classes or groups. When comparing the performance of the PCA versus the PLS-DA, the second one turns out to be more effective in the reduction of dimensionality when the objective is the separability of classes. PCA is only capable of identifying gross variability and is not capable of distinguishing ‘among-groups’ and ‘within-groups’ variability, as is the explicit goal of the simple linear DA paradigm (Barker and Rayens, 2003).

This research works on a methodological proposal applied to the study of archaeological materials, focusing on the characterization of the chemical formula of pottery through an elemental analysis technique and an adequate statistical processing based on the compositional data approach. The methodology, which is introduced in detail in further sections, is applied to pottery datasets from the archaeological site of Xalasco (Tlaxcala, Mexico), in order to demonstrate advances of the log-ratio approach compared to standard statistical techniques. Chemical information extracted from archaeological pottery can provide relevant information about aspects related to the technology used, its origin, function and social meaning. Through this methodology, it is possible to infer different technological characteristics of pottery production.

2. Archaeological setting

We applied the proposed statistical processing adapted for compositional data (described in further detail in the following section) to a data set of pottery samples proceeding from the archaeological site of Xalasco (Fig. 1). This site is located in the northeast portion of Tlaxcala