Comparatives studies of the Biplot and Multidimensional Scaling Analysis in experimental data

Oliveira, Paulo
IPEN - CNEN/SP, CRPQ
Lineu Prestes, 2242
São Paulo (01319-001), Brazil
ptoliveira@ipen.br

Munita, C. S.
IPEN - CNEN/SP, CRPQ
Lineu Prestes, 2242
São Paulo (05508-000), Brazil
camunita@ipen.br

1. Introduction

The detailed study of the physical and chemical properties of ceramics artifacts, associated with archaeological and historical research, has contributed to the reconstitution of the cultural habits and lifestyles of ancient communities. The present work aimed to study the elemental composition and mineralogy of archaeological ceramics collected from three different archaeological sites located in Brazil. By using instrumental neutron activation analysis (INAA) combined with multivariate statistical techniques, it is possible to define groups of ceramics in terms of elemental concentrations, which reflects the raw material composition used in its manufacture (Jones, 2004). Discrepant values were identified by Mahalanobis distance method. The results were analyzed by principal component analysis (PCA) and Biplot (Santos, 2007).

The elemental concentration data were normalized by log base-10 transformation (Santos, 2007) and also standardized by the compositional transformation combined with the median, in order to assess the method that best reduces the differences in magnitudes of the concentrations recorded.

2. Methods

2.1. Motivation

For this study, three archaeological sites were considered with, respectively, 34, 89 and 42 samples. The elements As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th and U were measured by instrumental neutron activation analysis (INAA). It is considered a very sensitive technique, used in the qualitative and quantitative analysis of macro and trace level elements (Aguilar, 2001). The samples were irradiated in the IEA-R1 nuclear reactor at the Research Reactor Center of the Nuclear and Energy Research Institute (IPEN).

Biplot technique allows viewing the relation and interrelations among the elements and samples on a bidimensional graph. The graphs also shows the existence of case clusters (Souza, 2010).

The Multidimensional Scaling (MDS) method checks the similarity/dissimilarity of data by representing them as points in a geometric space, which may be intercorrelated. This method offers an easy manner of visualizing the data structure.

This paper proposes the use of MDS and Biplot methodologies for do comparative study between log base-10 transformation and compositional transformation and, characterization of the Biplot (as the smallest loss of information) and MDS (as the smallest stress measure).

2.2. Principal Component Analysis

PCA is a statistical technique that linearly transforms a set of p variables in a set with a smaller number (k) of uncorrelated variables that explain a substantial portion of the information from the original data set. The p original variables ($X_1,...,X_p$) are transformed into p variables ($Y_1,...,Y_p$), so that Y_1 is the one that explains the largest amount of total data variability, Y_2 explains the second largest amount and so on.

The main objectives of the principal component analysis are: reducing the data dimensionality, obtaining interpretable combinations of variables, and finally, the discrimination and understanding of the correlation structure of variables.

Algebraically, the principal components are linear combinations of the original variables. Geometrically, the principal components are the samples coordinates in an axis system obtained by rotating the original system to the direction of maximum data variability.

The PCA depends only on the covariance (Σ) or the correlation matrix (ρ) of $X_1,...,X_p$, and requires no assumption about the form of multivariate distribution of these variables.

2.3. Biplot

Multivariate statistical analysis involves a set of statistical methods and mathematics designed to describe and interpret the data that arises from the observation of several variables together and some correlation structures (Johnson; Wichern, 2006).

Biplot is a multivariate technique proposed by Gabriel (1971), with the objective of graphing a data matrix, in such a way that its representation can show the relations and interrelations among the rows and columns of that matrix. Factoring the original data matrix by Singular Value Decomposition (SVD), as the sum of products of matrices that contain the marker of rows and columns, which are elements for graphical representation, can be considered a visual assessment of the data matrix structure (Gower, 1966).

$\mathbf{Y}_{\mathbf{n\ x\ p}}$ is a data matrix, where the n rows correspond to individuals (samples) and p columns correspond to the measured elemental concentrations of the samples. The Biplot of the matrix **Y** is a graphical representation made by vectors called markers a_1, a_2,...,a_n for the rows of **Y** and markers b_1, b_2,...,b_p for the columns of **Y**, so that the intern product of, for $i = 1$,..., n **e** $j = 1$,..., p, is equal or close to the elements $\mathbf{Y}_{\mathbf{ij}}$ of the original matrix **Y.**

If we consider the markers a_1, a_2,...,a_n as rows of matrix **A** and markers b_1, b_2,...,b_p as rows of matrix **B,** the decomposition of the matrix **Y** is given by: $Y \approx AB^T$

The structure of the matrix Y is displayed in the Euclidean space in two or three dimensions. The decomposition, in general, is not unique. There are several methods of matrix decomposition, which the best is considered the approximation of matrix with lower rank by the SVD, according to Gabriel (1971) and Greenacre (1984).

2.4. Compositional Data Analysis

We shall call an n **x** p data matrix as fully-compositional if the rows sum to a constant, and subcompositional if the variables are a subset of a fully-compositional data set. Such data occur widely in archaeometry, where it is common to determine the chemical composition of ceramics, glass, metal or other artifacts using techniques such as neutron activation analysis, X-ray fluorescence analysis (XRF), among others. The interest is often centered on whether there are distinct compositional groups within the data and whether, for example, they can be associated with different origins and manufacturing technologies (Baxter, 2003).

The sample space of compositional data is thus a simplex space in a $\boldsymbol{D-1}$ dimensional subset $\boldsymbol{R^D}$**.** Standard statistical methods can lead to misleading results if they are directly applied to original closed data. For this reason, the centered logratio (clr) transformation was introduced.

The clr transformation is a transformation from $\boldsymbol{S^D}$ to $\boldsymbol{R^D}$, and the result for an observation $\boldsymbol{x \in R^D}$ is the transformed data point $\boldsymbol{y \in R^D}$ with

$$y = (y_1,\dots,y_D)' = \left(\log\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}},\cdots,\log\frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}}\right)$$

2.5. Multidimensional Scaling

MDS or Proximity Analysis is a method that represents the measures of closeness (similarities and dissimilarities) between pairs of objects as distances in a multidimensional space in short supply, thus allowing the visual inspection of the data structure.

The MDS approach is defined considering the dissimilarity *n x n* matrix $\Delta = [\delta_{ij}] \in \Re$ where δ_{ij} represents a measure of proximity between the i^th and j^th objects. A reduction algorithm obtains a dimensional configuration of points (vectors of coordinates) called by $\mathbf{x}_i = (x_{i1},\cdots,x_{iq})$ of order $(n \times q)$ on a smaller scale, i.e. $(n > q)$, also must verify that the matrix of Euclidean distance $D = [d_{ij}]$ of order $(n \times n)$, being that $d_{ij} = \|x_i - x_j\|$, where $i = (1,\cdots,n)$ and $j = (1,\cdots,q)$, obtained from this if t of points, approaching the maximum of the original dissimilarity matrix, ie $D \approx \Delta$ (Souza, 2010). The determination of the relationship among data can be given by MDS. This technique can be metric (space Euclidean, two-dimensional) or non-metric (Minkowski).

The objective of this analysis is to rearrange the distribution of objects (or variables) in order to study and detect most significant in explaining similarities or dissimilarities (distances) among them.

MDS is a technique that allows testing with certain criteria the differences between objects interest that are mirrored in the corresponding empirical differences these objects and it is a statistical technique that can provide a spatial representation of a set of measures elemental concentrations from measurements of similarities between them.

The option uses the non-metric ordination of measures (Minkoeski). It yields and increasing or decreasing order of similarity measures, so that the algorithm determines which is the graphic that best fits the experimental results. This adjustment is such that the order of the distances between points on the graphical configuration is as close as possible to the order of similarities.

The metric type is characterized by the need of using values in the adjustment process and consists of a method for constructing the configuration from the Euclidean distances between points, using a method highly related to PCA.

3. Results and Discussion

PCA and MDS were applied in the three data sets of ceramic elemental concentrations and the results obtained can be seen in Table 1, where n is the number of components needed for the percentage of variance explained to be greater or equal to 70, stress is the stress ratio and RSQ is the index of corrected R^2.

Table 1. Results of the PCA and MDS for each site and for each transformation

site	samples	transformation	n PC's	explanation (%)	stress	RSQ
1	34	log base 10	4	75.1	0.0764	0.9999
1	34	compositional	4	74.7	0.2318	0.7508
2	89	log base 10	3	72	0.0695	0.9999
2	89	compositional	3	70.2	0.2137	0.8034
3	42	log base 10	3	80.4	0.0147	0.9994
3	42	compositional	3	78.3	0.1924	0.7869

In Table 1, one can observe that the base-10 log transformations yielded slightly larger percentages of the total variance explanation, lower rates of stress and higher values R^2, which means that, in the case, the base-10 log transformation was better than compositional transformation.

Figure 1 shows Biplot and PCA for the first three principal components considering all the 165 samples after applying base-10 log (a) and compositional (b) transformations.

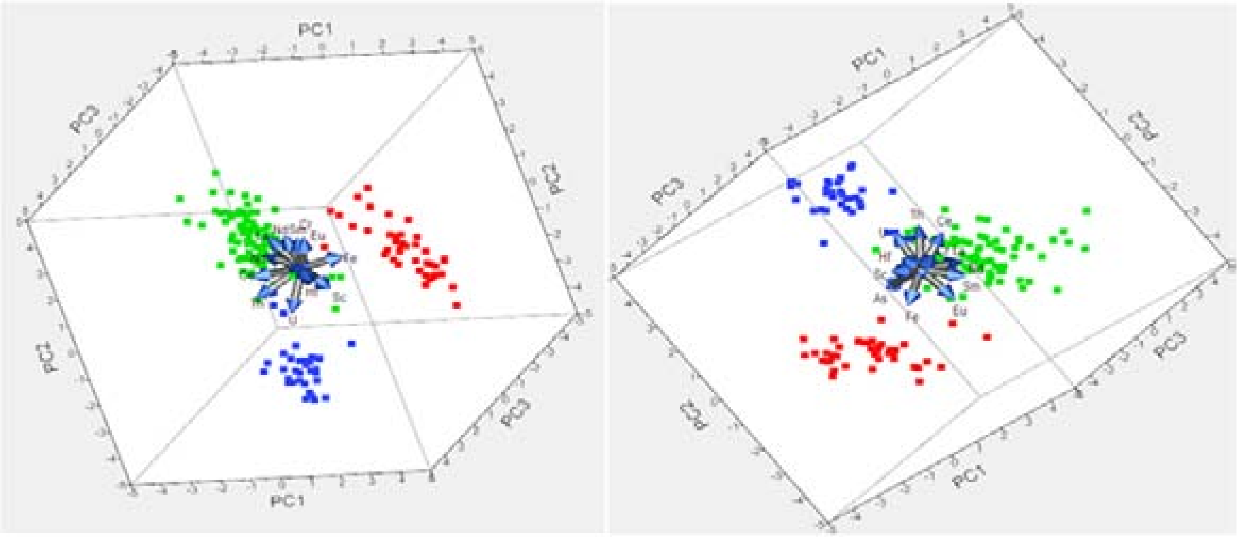

Figure 1. Graphics Biplots for three first principal components for a) Log base 10 transformation and b) compositional transformation

It is observed from Figure 1 that the Biplot tridimensional graphs for compositional data presents more spread data, when compared with the graph for base-10 log transformations.

The percentage of variance explained by the three first principal components is 78% for base-10 log transformation and 77.26% for the compositional data. This mean that the information loss of all data processed was lower than the same loss in compositional data.

Figure 2 shows the graphs of principal coordinates for all the data sets, composed by 165 samples, considering the base-10 log (a) and compositional (b) transformations, obtained by the method of MDS the matrix Euclidean distance, which in this case can be understood as a general case of PCA when the dissimilarity is measured by Euclidean distance (Souza, 2010).

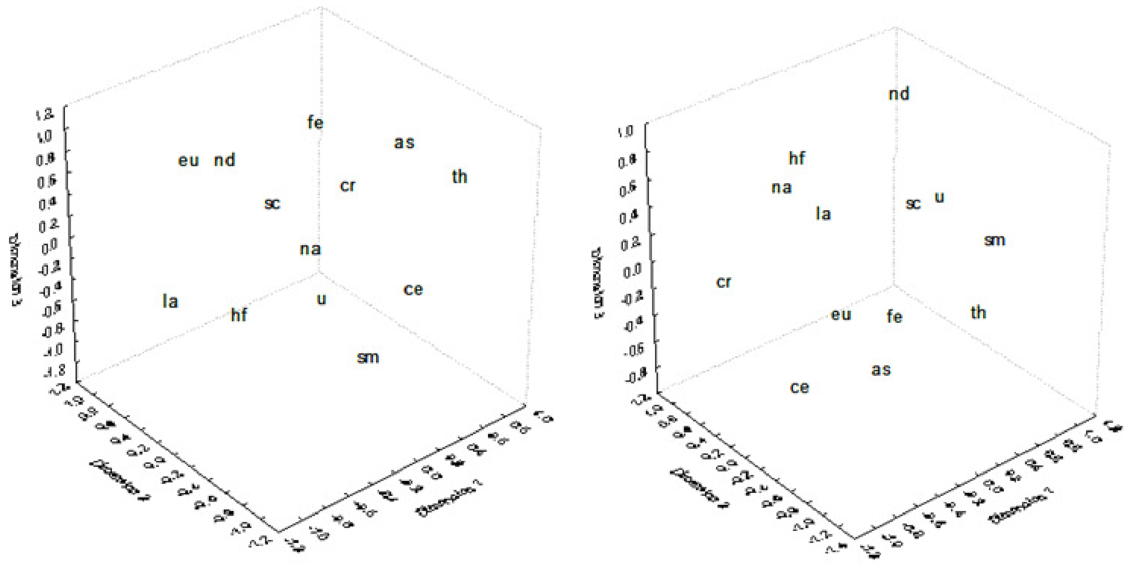

Figure 2. Graphics of principal coordinates a) Log base 10 transformation and b) compositional data

In Figure 2 we observed the three-dimensional positions of the variables that are all above the plane parallel to the (dimension 1 versus dimension 2) and verify that shortest distance was between Hf and Na for compositional transformation.

The stress ratio obtained for the log-transformed data was 0.1658, and 0.2159 for the compositional data. One can conclude that the base-10 log transformed data yielded a better fit than the compositional data, due to its lower incidence of stress.

4. Conclusion

Techniques like Biplot of PCA allow a better assessment regarding the loss of information, while MDS technique that better assesses the quality of data fitting.

The base- log transformation was less loss of information and lowest stress measure that compositional transformation.

REFERENCES

Aguiar, A.M. (2001) *Aplicação do método de análise por ativação com nêutrons à determinação de elementos traços em unhas humanas.* Dissertation, Nuclear Energy Resourch, IPEN – CNEN / SP, São Paulo, Brazil.

Baxter, M.J. (2003). *Compositional data analysis in archaeometry.* Universitat of Girona.

Ferreira, D. F. (2010) *Estatística Multivariada*,. Editora UFLA: Lavras, Brazil.

Gabriel, KR.(1971) The Biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58(3):**453—467.

Greenacre, M.J.(1984) *Theory and application of correspondence analysis*. London Academic Press.

Gower, J.C. & Hand, D.J. (1996) *Biplots.* New York: Chapman & Hall.

Hair Jr., J.F.; Blach, W.C; Babin, B.J.; Anderson, R.G & Tathan, R.L. (2006) *Multivariate Data Analysis*, Sixth edition. Prentice-Hall, New-Jersey, USA.

Johnson, R.A. & Wichern, D.W. (2007) *Applied Multivariate Statistical Analysis*, Sixth Edition. Prentice-Hall, New Jersey, USA.

JONES, A. (2004) Archaeometry and materiality: materials based analysis in theory and practice. *Archaeometry*, 46(3):327—338.

Manly, B.F. (2008*) Métodos Estatísticos Multivariados*, Third Edition. Bookman: Porto Alegre, Brazil.

Santos. J.O. (2007) *Estudos arqueométricos de sitios arqueológicos do baixo São Francisco*. Thesis of Doctor of Science, Nuclear Energy Resourch, IPEN – CNEN / SP, SãoPaulo, Brazil.

Souza, E.C. (2010) *The Biplot methods and multidimensional scaling in experimental design*. Thesis of Doctor of Science, São Paulo University, Piracicaba, Brazil.

ABSTRACT

Archaeometry is an established area since the decade of 1960. Some of its techniques employ nuclear methods in the characterization of art, archaeological and cultural heritage objects in general. For the interpretation of results of ceramics elemental concentration data, many multivariate statistical methods, such as principal component analysis and biplot, are used extensively.

Multivariate statistical analysis involves a set of statistical and mathematical methods designed to describe and interpret the data, which predicts the observation of several variables together and some correlation structures (Hair et all., 2006).

The principal component analysis is related to the explanation of variance and covariance structure by linear combinations of the original data (Ferreira, 2008). The components depend only on the sample covariance matrix and do not require multivariate normality of the data (Manly, 2008).

The application of principal component analysis has the following objectives: reducing the dimensionality of data, obtaining interpretable combinations of original variables, description and understanding of the correlation structure of variables.

Biplots are graphs that are plotted in the scores of the two most important principal components, along with the values of the corresponding eigenvectors. In this graphical representation, it is possible to observe associations among the sites and the variables, indicating which ones are responsible for the explanation of each point (Jonhson and Wichrn, 2006; Souza, 2010). The compositional data is the ***n*** x ***p*** matrix, where the sum of the values of all the lines results in a same constant.

Multidimensional scaling is a multivariate method that checks the similarity of data on a set of objects and can be inter-related (correlated) to model multidimensional data such as the distance among points in a geometric space (Souza, 2010). This technique is based solely on the distance or similarity matrix between n samples in a ***p***-dimensional space. It obtains a representation of the corresponding observations, yielding a similarity or distance matrix and finds the principal coordinates (unknown dimensions for the multivariate samples).

In this work, a comparative study of the applications of biplot and multidimensional analysis is performed. The data sets are formed by elemental concentrations of As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th and U in samples from three archaeological sites, which are transformed to base-10 logarithms and normalized.