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

Oliveira, Paulo
Nuclear Energy Research, IPEN - CNEN/SP, CRPQ
Lineu Prestes, 2242
São Paulo (05508-000), Brazil
ptoliveira@ipen.br

Munita, C. S.
Nuclear Energy Research, 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 ceramic artifacts, associated with archeological and historical research has allowed the reconstitution of the cultural customs and lifestyles of ancient communities. This study aimed to study chemical composition and mineralogy of archaeological ceramics collected from three different archaeological sites located in Brazil. Through analysis by instrumental neutron activation (AANl) and allowed to set ceramic compositional groups according to the chemical similarity of ceramic pastes, which reflects the composition of the raw material used in its manufacture by prehistoric man, and infer atmosphere and sintering temperature of ceramics. Aberrant specimens were identified by means of Mahalanobis distances. The results were interpreted by principal component Analysis (PCA) and Biplot (Santos, 2007).

The data of elemental concentrations were standardized by log base 10 and also were standardized from its compositional form standardized by the median to determine which method best reduces the differences in magnitudes of the concentrations recorded (Santos, 2007).

2. Methods

2.1. Motivation

For this study, were considered for data analysis of the three sites with 34, 89 and 42 samples of elemental concentrations of As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th and U determined by instrumental activation analysis, taken as a very sensitive technique, used in qualitative and quantitative analysis of elements present in a broad range of concentrations in the range of percentages to trace levels (Aguilar, 2001), in samples of ceramic fragments collected from archaeological sites in the IEA-R1 Nuclear Reactor the Research Reactor Center at the Institute of Nuclear Energy Research (IPEN).

Biplot technique allows viewing on a plane relations and interrelations between the elements and cases of the matrix after being graphed. This graph demonstrates the existence of clusters of cases (Souza, 2010).

The Multidimensional Scaling (MDS) method checks the similarity / dissimilarity of the data as a set of distances between points in a geometric space that may be inter-correlated. This method allows for an easier way to visually of the data structure.

This paper proposes the use of MDS and Biplot methodologies to determine associations between variables and cases on experimental data and make comparison between MDS and Biplot and establishing the conditions to check which method has satisfactory performance and its specificities.

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 set. The p original variables ($X_1,...,X_p$) are transformed into variables p ($Y_1,..., Y_p$), so that Y_1 is one that explains the largest portion of total data variability, Y_2 explains the second largest plot and so on.

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

The analysis is performed in order to summarize the pattern of correlation between variables and in some cases it is possible to reach set of variables that are not correlated, thus leading to a grouping of them. Algebraically, principal components are linear combinations of original variables. Geometrically, the principal components are the coordinates of sampling points in an axis system obtained by rotating the original system of axes in the direction of maximum variability of the data.

The PCA depends only on the covariance (Σ) or the correlation matrix (ρ) of $X_1,...,X_p$. It 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 mathematicians, designed to describe and interpret the data that comes 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 graph a data matrix, such that this representation allows in a plan view of the relations and interrelations between the rows and columns of this matrix. Factoring the matrix of original data by Singular Value Decomposition (SVD) as the sum of products of matrices that contains the markers of rows and columns that are elements for the graphical representation, can be a visual assessment of the structure of data matrix (Gower, 1996).

$\mathbf{Y}_{\mathbf{n\ x\ p}}$ is a data matrix, where the n rows correspond to individuals (samples) and p columns correspond to the measures elemental concentrations on the samples. The Biplot of the matrix $\mathbf{Y}$ is a graphical representation made by vectors called markers a_1, a_2,...,a_n to the rows of $\mathbf{Y}$ and markers of b_1, b_2,...,b_p for the columns of $\mathbf{Y}$, 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 $\mathbf{Y}$.

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

The structure of the matrix $\mathbf{Y}$ is displayed, representing the Euclidean space in two or three dimensions. The decomposition in general is not unique. There are several ways of decomposition of a matrix. The best known method of approximation to a matrix of lower rank is the SVD, according to Gabriel (1971) and Greenacre (1984).

2.4. Compositional Data Analysis

We shall call an n x p data matrix fully-compositional if the rows sum to a constant, and sub compositional 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 ceramic, glass, metal or other artefacts using techniques such as neutron activation analysis, X-ray fluorescence analysis (XRF) among others. Interest often centres on whether there are distinct chemical groups within the data and whether, for example, these can be associated with dofferent origins or manufacturing technologies (Baxter, 2003).

The sample space of compositional data is thus simplex space is a $\boldsymbol{D} - 1$ dimensional subset $\boldsymbol{R}^{\boldsymbol{D}}$.

Standard statistical methods can lead to misleading results if they are directly applied to original closed data. For this reason, centred logratio (clr) was introduced.

The clr transformation is a transformation from $\boldsymbol{S}^{\boldsymbol{D}}$ to $\boldsymbol{R}^{\boldsymbol{D}}$, and the result for an observation $\boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{D}}$ are the transformed data $\boldsymbol{y} \in \boldsymbol{R}^{\boldsymbol{D}}$ 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 comes to represent the measures of closeness (Similarities or dissimilarities) between pairs of objects as distances in a space multidimensional in short supply, thus allowing visual inspection of the structure of data.

The MDS approach is defined as, consider the dissimilarity matrix $\Delta = [\delta_{ij}] \in \Re$ order $(n \times n)$, where δ_{ij} represents a measure of proximity between the ith and jth object. A reduction algorithm obtains 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, ie $(n > q)$, 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 between data can be given by MDS. This technique can be metric (space Euclidean, two-dimensional) or non-metric (Minkowski).

The objective this analysis is to rearrange the distribution of objects (or variables) in order to study detect smaller significant in explaining similarities or dissimilarities (distances) between 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, therefore, when given the increasing or decreasing order of similarity measures, the algorithm that determines which is the graphic that best fits ***t***he experimental values. Thus, 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.

Since the metric type is characterized by the need to use 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 sets of ceramic fragments and were obtained the following results, as can be seen in Table 1, *n PC's* is the number of components needed for the percentage of variance explained is greater than 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	explication (%)	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 sets of data were transformed log base 10 explanation of the percentage of total variance slightly larger, lower rate of stress and higher values of $\boldsymbol{R}^2$, which means that in this case, the adjustments made to the data transformations by log base 10 were better than for compositional data.

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

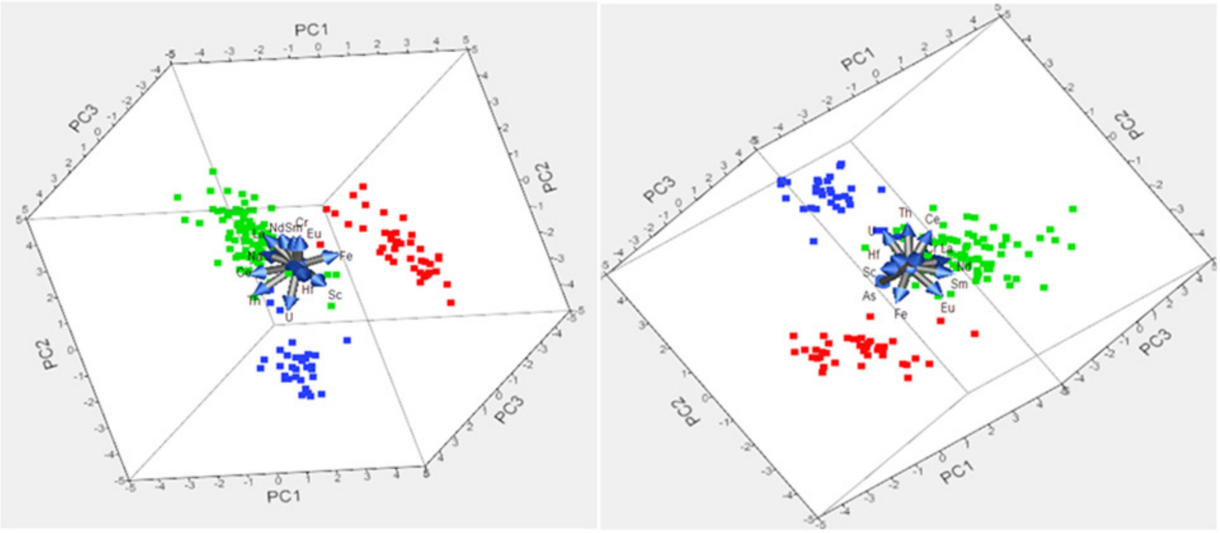

Figure 1. Graphics Biplots for three first principal components for a) Log base 10 transformation and b) compositional transformation

It is observed from Figure 1 that the graphics Biplots compositional data for the first three principal components are more spread out compared to the same chart that consider data transformed by log base 10.

The percentage of variance explained for the first three components principals from data transformed by log base 10 is 78%, while the same variance explained for the compositional data was 77.26%. This means 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 data sets with all 165 samples by considering the transformations log base 10 (a) and compositional (b) obtained from 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).

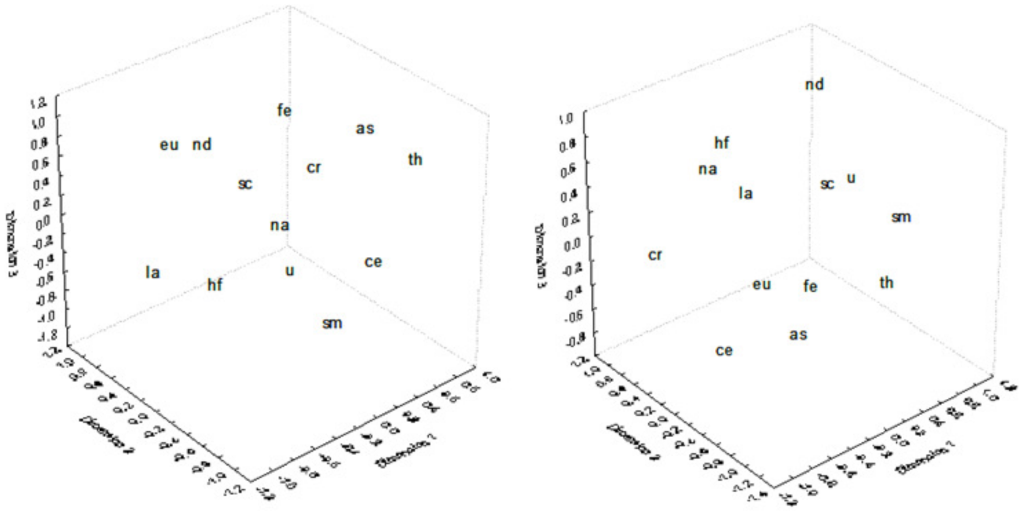

Figure 2. Graphics of principal coordinates a) Log base 10 transformation and b) compositional data

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) in section 3 -1.2 dimension to the data transformed by log base 10 (a), while For compositional data (b) the data are all above the scale of -0.8 in the dimension 3.

The stress ratio obtained for the data transformed by log base 10 was 0.1658, while for the index of compositional data stress obtained was 0.2159. This means that data transformed by log base 10 had a better fit than the compositional data, since, had a 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.

REFERENCES (RÉFERENCES)

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.

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.