2011 International Nuclear Atlantic Conference - INAC 2011
Belo Horizonte,MG, Brazil, October 24-28, 2011
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
***ISBN:* 978-85-99141-04-5**

VARIABLE SELECTION USING PROCRUSTES ANALYSIS FOR AN INAA COMPOSITIONAL STUDY IN THE CENTRAL AMAZON

Roberto Hazenfratz[1], Paulo M. S. Oliveira[1], Casimiro S. Munita[1], Eduardo G. Neves[2],

[1] Instituto de Pesquisas Energéticas e Nucleares (IPEN – CNEN/SP)
Av. Professor Lineu Prestes 2242
05508-000 São Paulo, SP, Brasil
robertohm@usp.br

[2] Museu de Arqueologia e Etnologia, Universidade de São Paulo (MAE-USP)
Av. Prof. Almeida Prado, 1466
05508-900 - Cidade Universitária - São Paulo, SP, Brasil

ABSTRACT

A set of 140 archaeological ceramic samples from Osvaldo and Lago Grande archaeological sites was chosen to perform a variable selection study. The elemental concentrations of 24 elements (As, K, La, Lu, Na, Nd, Sb, Sm, U, Yb, Ba, Ce, Co, Cr, Cs, Eu, Fe, Hf, Rb, Sc, Ta, Tb, Th and Zn) were measured by INAA. Firstly, the quantity of missing values was analyzed for each variable. Then, the precision of results was addressed and the accuracy of experimental results was analyzed by z-score tests with the elemental concentrations of IAEA *Soil 7* reference material. The variables eliminated in those steps were K, Nd, Sb, Ba, Tb, Zn, Rb and Sm. Finally, a Procrustes analysis was performed in the remaining elements. It was concluded that the diagnostic elements for the present work were Lu, Na, Yb, Cr, Eu, Hf and Sc, because they are sufficient to represent the multivariate structure of data. The PCA with the 16 elements selected prior to Procrustes analysis, and the same analysis for the 7 elements selected by the method were compared. No significant information was lost regarding the multivariate compositional patterns of data. It confirmed the variable selection efficiency by Procrustes analysis. The authors argue that it is useful to perform variable selection studies before further statistical analyzes are performed, instead of believing that the higher the quantity of variables, the better will be the results of multivariate analysis.

1. INTRODUCTION

Many archaeometric compositional studies employ instrumental neutron activation analysis (INAA) as the analytical technique to measure elemental concentrations due to many advantages. Some of them are: high precision, accuracy and sensitivity for many elements; ease of preparation of small samples and the simultaneous measurement of 30 or more elements [1].

The multivariate data produced are normally analyzed by multivariate statistical methods, as cluster, discriminant and principal component analysis. However, their use can be problematic and misinterpreted with too many variables. The addition of a variable that is not informative about the data structure can mislead the perception of real patterns, and generate computational and sample size problems. Furthermore, discarding variables may save time, money and fewer variables may be measured [2]. To accomplish such an objective, it is possible to select a subset of variables that retains most of the multivariate data structure, with no damages for the future multivariate analyzes [3]. Although variable selection may appear to be counter to the commonly held view “the higher the quantity of measured variables the better”, it is not obvious that this view is appropriate for elemental

***ISBN:* 978-85-99141-04-5**

fingerprinting studies. The inclusion of variables that are not informative about the multivariate structure may mislead the perception of the real data patterns [3, 4, 5].

There are many methods listed in the literature to perform variable selection analysis, as methods based on cluster analysis; methods based on multiple correlation and methods based on principal component analysis, as Procrustes [2, 6]. The reasons for choosing Procrustes analysis is its good performance and because many compositional studies focus on data grouping and representation in principal components.

Hurley and Cattell were the first researchers to suggest the name "Procrustes Analysis" for the techniques which compare different sets of principal components [4]. The name is based on the Greek myth of Damastes (or Procrustes), a rogue smith from Attica, son of Poseidon. He forced each of his visitors to lie and fit his iron bed, by stretching them or cutting off their legs [7].

Procrustes technique is a very efficient method to select variables [4, 8]. Firstly, the principal components are extracted from the data matrix **X**. Then, it is determined a significant number of principal components *k* to represent the data variability in a satisfactory way. The quantity *k* may be decided either formally, using available methods, or informally [9]. The coordinates of the samples on these components are considered as the standard configuration, which we denote here by **Y**(n × k). For each variable removed in turn, the first *k* principal components are recalculated for the $n \times (p-1)$ data matrix and a matrix of coordinates $\mathbf{Z}_{(j)}$ (n × k) is generated for each j^{th} variable removed. From the comparison of $\mathbf{Z}_{(j)}$ with **Y**, the algorithm removes the variable that yields the lowest discrepancy value $M^2_{(j)}$ [4]. The discrepancy value is given by [9]:

$$M^2 = Trace\left(\mathbf{Y}\mathbf{Y}' + \mathbf{Z}_{(j)}\mathbf{Z}'_{(j)} - 2\mathbf{\Sigma}\right) \quad (1)$$

where, from the singular value decomposition for the matrix $Z'_{(j)}Y$ (k × k):

$$\mathbf{\Sigma} = diag\left(\sigma_1, \sigma_2, ..., \sigma_k\right) \quad (2)$$

The quantity M^2 given by Equation (1) represents the sum of the squared differences between the configuration $\mathbf{Z}_{(j)}$ and **Y.** It is a measure of the loss of data structure information when *q* variables are used instead of the p original variables.

The Procrustes procedure explained briefly here is repeated up to the point in which the discrepancy measure is higher than a critical value. Mathematically, the stopping rule adopted in this work for *i* variables removed is [10]:

$$M^2 \geq \left(1 + c^2\right)\hat{\sigma}^2 \cdot \chi^2\left(0.95, nk - 0.5k\left(k+1\right)\right) \quad (3)$$

where,

***ISBN:* 978-85-99141-04-5**

$$c = \sqrt{\frac{(p-i-k)}{(p-k)}} \quad (4)$$

and $\hat{\sigma}^2$ represents a suitable estimate for the per-observation error variance in **X**.

Prior to Procrustes analysis, it is necessary to perform a quality control test in order to assess the precision and accuracy of results. In this work, the quantity adopted as a measure of agreement between the experimental values and the certified ones were the z-score. For elemental concentrations measured by INAA, the statistic may be calculated by [11]:

$$z_i = \frac{C_i - C_{\text{ref,i}}}{\sqrt{\sigma_i^2 + \sigma_{\text{ref,i}}^2}} \quad (5)$$

where:

C_i is the experimental concentration of element *i*
$C_{ref,i}$ is the certified concentration of element *i*
σ_i is the experimental uncertainty in the concentration of element *i*
$\sigma_{ref,i}$ is the uncertainty in the certified concentration of element *i*

The samples used in this work are archaeological ceramic fragments from Osvaldo and Lago Grande archaeological sites, in the central Amazon, Brazil. The ceramic samples were provided by the Museum of Archaeology and Ethnology of the University of São Paulo (MAE-USP), excavated during the Central Amazon Program (CAP), an effort to understand the pre-colonial occupation of the central Amazon region. Those sites were chosen due to the intensive excavation work performed and because they represent a microcosm of the studied region [12]. This work is part of a bigger project regarding the comparison of both sites by archaeometric studies.

2. MATERIALS AND METHODS

One hundred and forty ceramic fragments from Osvaldo and Lago Grande archaeological sites were analyzed by INAA in order to characterize their elemental composition. The ceramic fragments selected represent all the stratigraphic levels of both sites. The experimental procedure is further described in [13]. The data was logarithmically ($\log_{10}$) transformed. The measured elements were As, K, La, Lu, Na, Nd, Sb, Sm, U, Yb, Ba, Ce, Co, Cr, Cs, Eu, Fe, Hf, Rb, Sc, Ta, Tb, Th and Zn.

Firstly, the quantity of missing values was checked for each variable, in order to remove the elements which presented more than 10% of missing values. Then, it was presented a univariate analysis of the elemental concentrations in order to check for precision, by analyzing the relative uncertainties for each element. The accuracy of the experimental data was addressed by quality control tests using IAEA *Soil 7* as the reference material in INAA. The z-score was adopted as the measure of agreement between experimental and certified values for the elemental concentrations. The most problematic elements for the univariate analysis were eliminated. Finally, the elemental data set comprising the concentration of the

***ISBN:* 978-85-99141-04-5**

remaining elements was analyzed by Procrustes technique on the first four principal components with the objective of determining a variable set with the best discriminant properties possible.

3. RESULTS AND DISCUSSION

Figure 1 shows the percentage of missing values for the elements analyzed by INAA.

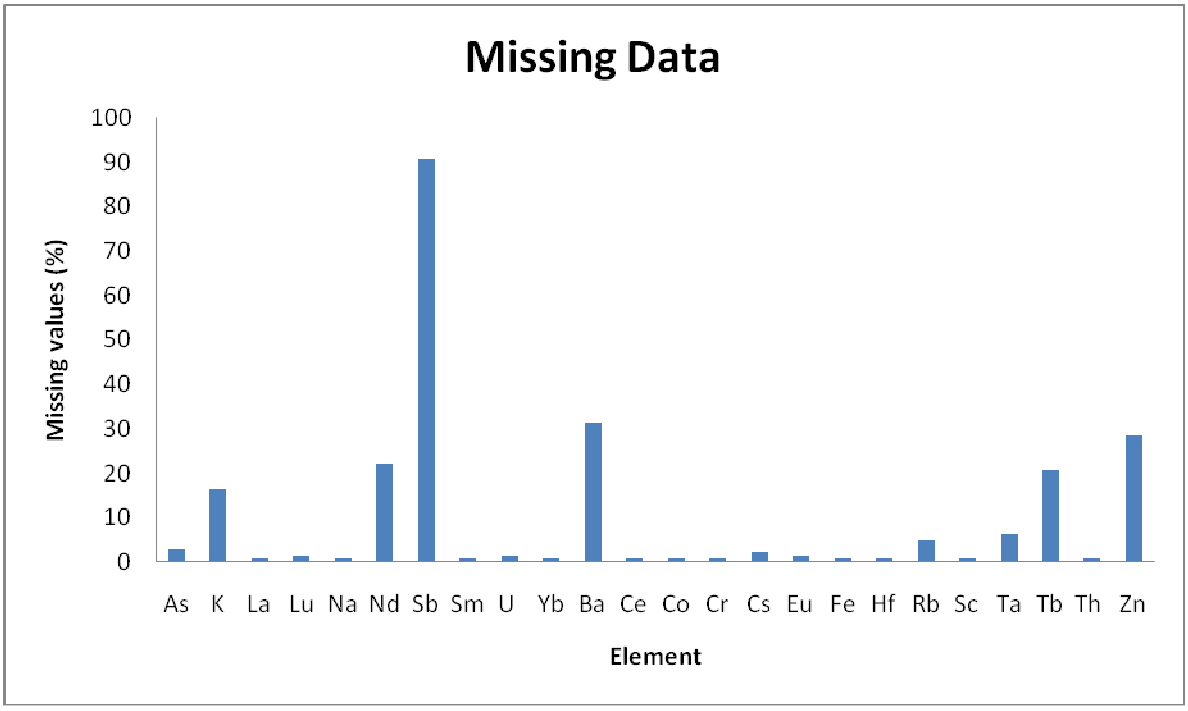

Figure 1: Missing data analysis for the elements determined by INAA

The elements which presented more than 10% of missing values were eliminated, in order to avoid excluding a high number of samples from analysis. Then, the elements removed were: K, Nd, Sb, Ba, Tb and Zn.

The uncertainties regarding the gamma spectrometry were addressed for each element. Figure 2 presents the percentage of uncertainties higher than 20% for each element, where it was possible to verify that U, Rb and Ta were the elements with the highest frequency (> 10%) of uncertainties above 20%.

It was performed a z-score test for each variable with the elemental concentrations of IAEA *Soil 7* as the reference material. It was chosen the critical value 2.58 for the standardized normal distribution, with level of significance 0.01. Figure 3 presents the frequency of

***ISBN:* 978-85-99141-04-5**

rejection of the null hypothesis for the equivalence between the measured values and the certified ones for the concentrations.

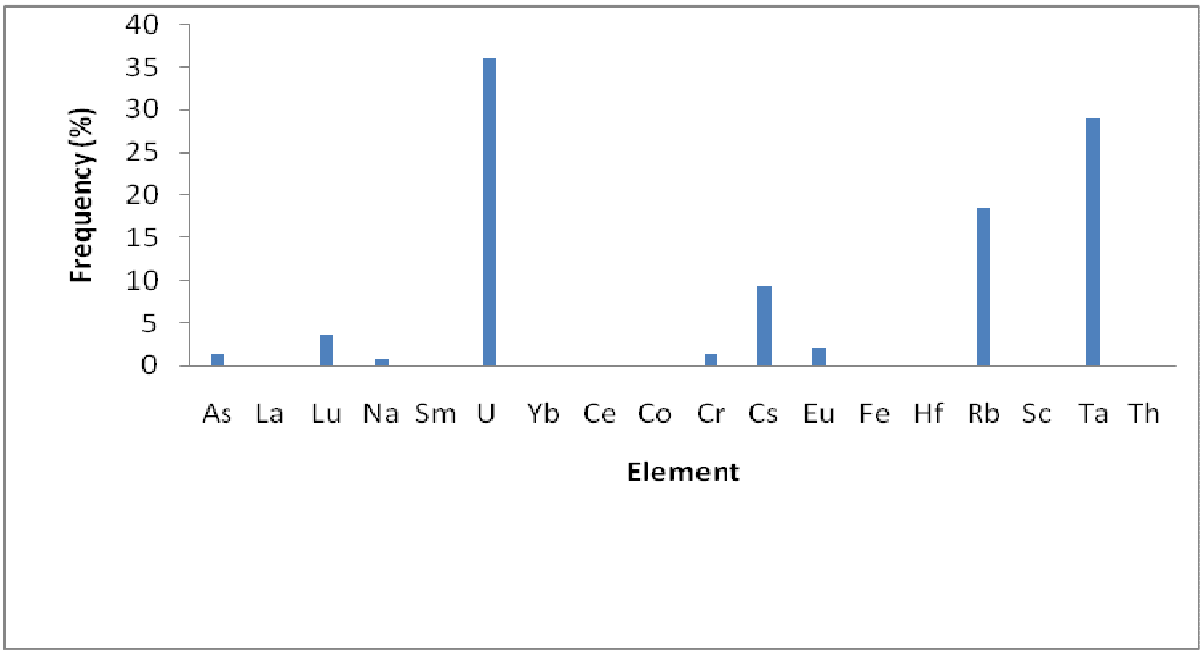

Figure 2: Frequency distribution for gamma detection uncertainties higher than 20% for elements analyzed by INAA

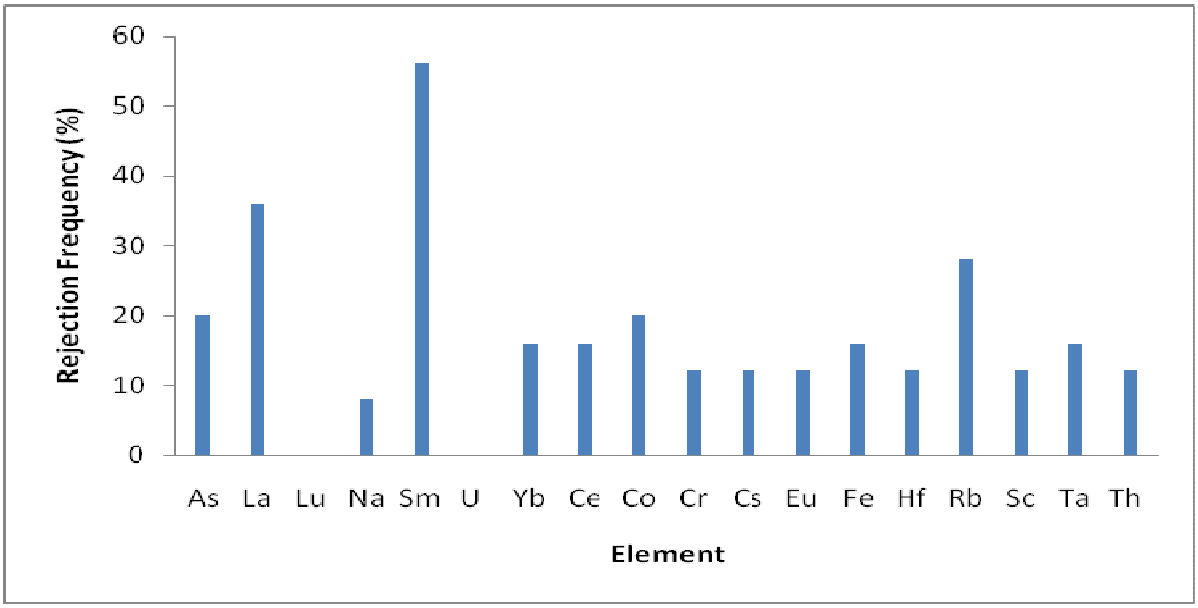

Figure 3: Results of the z-score test for the elemental concentrations of IAEA *Soil 7* reference material

***ISBN:* 978-85-99141-04-5**

By Figure 3, the most problematic elements for the control quality analysis were La, Sm and Rb.

As the element Rubidium was an element that presented one of the highest number of uncertainties above 20% (19%) and one the highest percentages of rejection in the control quality analysis (28%), it was eliminated. Samarium was also excluded because it presented more than 50% of rejection frequency in the control quality analysis.

The remaining elements for Procrustes analysis were: As, La, Lu, Na, U, Yb, Ce, Co, Cr, Cs, Eu, Fe, Hf, Sc, Ta e Th. Thirteen samples were excluded from the primary data set with 140 samples due to remaining missing values.

Table 1 summarizes the results of the Procrustes algorithm implementation in the software *R,* performed in our laboratory. The discrepancy value M^2 and the critical values are plotted for each step representing the elimination of one variable. The analysis was based on the first four principal components, which represented 76% of the total system variance for the complete data set with 16 elements.

Table 1: Results of Procrustes analysis for a data set with 16 elements. Level of significance: 5%

Element	Ta	Th	La	Fe	U	Co	Ce	As	Cs	Na	Lu	Cr	Yb, Eu, Hf, Sc
M^2	10.3	22.3	41.5	65.1	93.0	128.2	168.6	210.5	277.6	316.6	369.1	444.0	–
cv	465.6	445.3	425.1	404.9	384.6	364.4	344.1	323.9	303.6	283.4	263.2	242.9	–

From Table 1, it is possible to observe that the discrepancy value M^2 remains lower than the critical value c.v. up to the element Cesium. From the next step on, representing the elimination of Sodium, the discrepancy value turned higher than the critical value, and one may conclude that alterations in the multivariate structure of data becomes statistically significant from that point on. The algorithm finishes at the point with 4 variables remaining (Yb, Eu, Hf and Sc), because the number of principal components chosen were 4.

Therefore, the remaining elements that were not eliminated by Procrustes analysis were: Lu, Na, Yb, Cr, Eu, Hf and Sc. By that approach, it is the elemental data set formed by the diagnostic elements which represent the multivariate structure of interest in this work. It does not mean that all the removed elements could not carry information about other aspects of interest for an archaeometric study. The reasons of their elimination will be analyzed in the future, attempting to correlate them to geochemical, analytical and contamination causes.

By Table 1, it can be seen that the remaining elements which were the most problematic for uncertainty and control quality analysis (U, Ta and La) were eliminated by Procrustes algorithm. It will be investigated if their elimination were due to the lack of precision and accuracy in the results, or due to other geochemical reasons, as correlation with other

***ISBN:* 978-85-99141-04-5**

variables. It is also interesting to observe that among the elements selected by Procrustes Analysis, there are 3 rare earth elements (Lu, Yb and Eu) and also 3 transition metals (Cr, Hf and Sc). It reflects a well know fact that those types of trace elements, rare earths and transition metals, has a distinct geochemical behavior which is characteristic of the geological environment.

For a graphical visualization of the Procrustes algorithm, a plot of the statistic M^2 based on the experimental results, along with the critical values adopted as the stopping rule is showed in Figure 4. From the increasing statistic M^2, one can observe that it is positively dependent on the step number, which represents the variable being extracted. It is according to what is expected from the algorithm: with the successive extraction of elements, the configuration $\mathbf{Z}_{(j)}$ turns progressively more discrepant from configuration **Y** (see explanation on page 2).

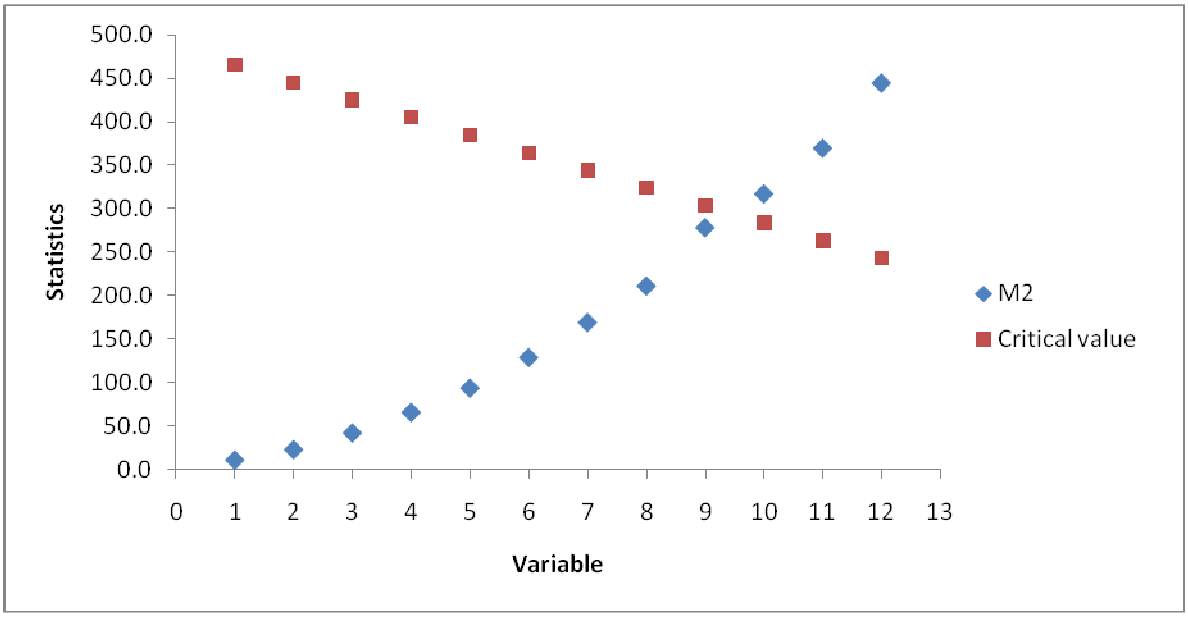

Figure 4: Discrepancy (M^2) and critical value behavior during Procrustes Analysis

Figure 5 shows the principal component analysis (PCA) for the 16 elements data set and the same analysis for the variables selected by the Procrustes algorithm. It may be seen from that figure that although the relative position of groups change, the analysis remains practically the same in terms of group composition, which was also checked by comparing group assignment from both Ward clusters for 16 elements and for 7 elements. The intersection between the ellipses representing the 95% confidence region for groups 1 and 2 is minimized with 7 elements. Based on the eigenvalues of the covariance matrix, it was calculated that total system variance represented by the two first principal components increased from 56%, for 16 elements, to 86% with the 7 elements selected. Thus, we may conclude that the PCA

***ISBN:* 978-85-99141-04-5**

analysis were more efficient after the variable selection procedure. Figure 5 shows that no significant information was lost regarding the multivariate compositional patterns of data, and confirms the efficiency of the variable selection by Procrustes analysis.

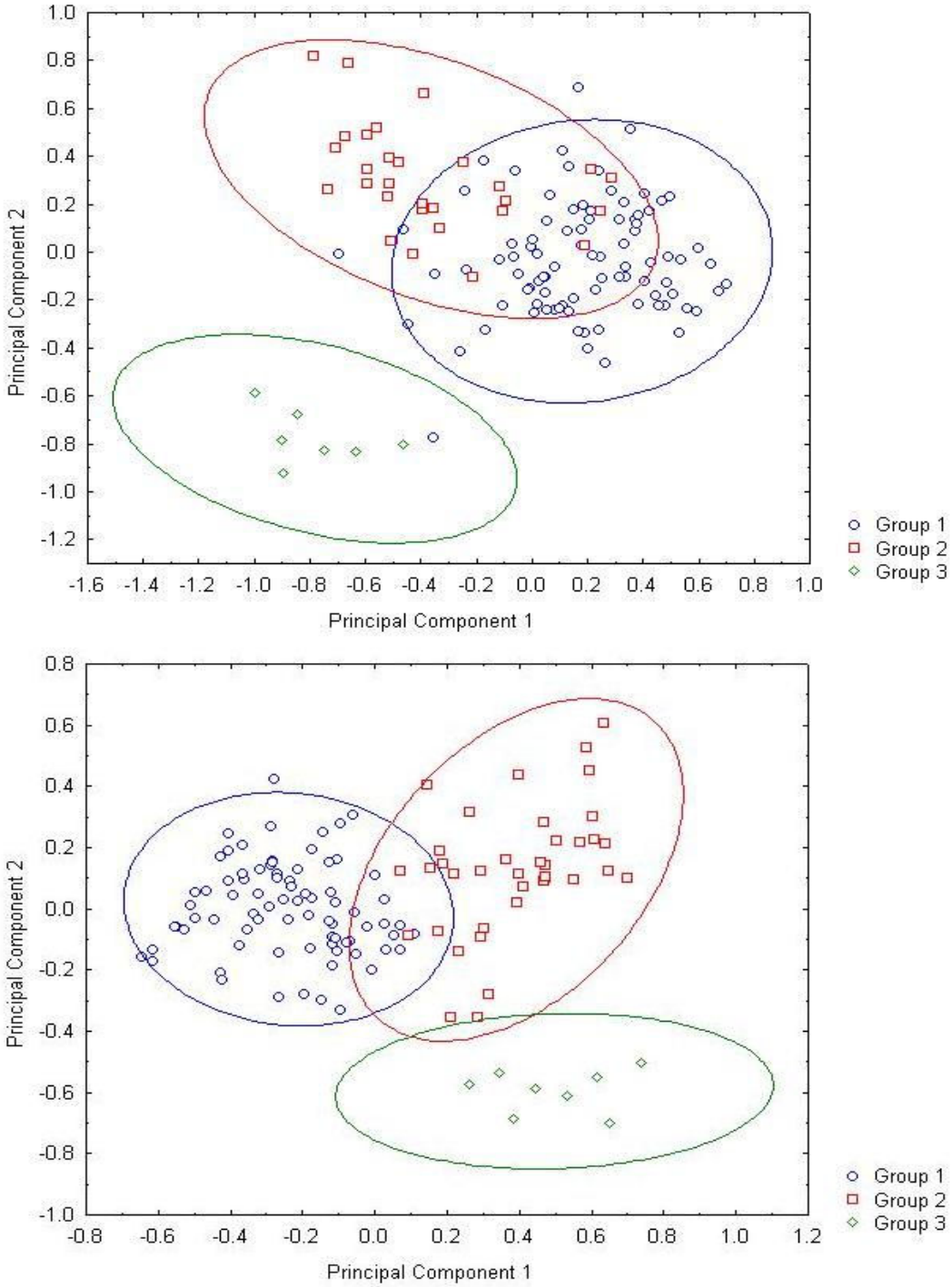

Figure 5: Comparison of PCA for 16 elements (up) and for the 7 elements selected by Procrustes analysis (down). The ellipses represent the 95% confidence region.

***ISBN:* 978-85-99141-04-5**

4. CONCLUSION

The elemental data set formed by the elements Lu, Na, Yb, Cr, Eu, Hf and Sc selected by Procrustes analysis represents adequately the multivariate structure of the data by PCA. It is an evidence that it is useful to inspect what are the diagnostic variables before further statistical analyzes are performed. It is not a good practice believing that the more the quantity of variables is increased, the better will be the results of multivariate analysis. The reasons of the elimination of As, K, La, Nd, Sb, Sm, U, Ba, Ce, Co, Cs, Fe, Rb, Sc, Ta, Tb, Th and Zn will be analyzed in the future, attempting to correlate them to geochemical, analytical and contamination causes.

ACKNOWLEDGMENTS

The present work was realized with the support from "São Paulo Research Foundation" – FAPESP – Brazil. Process Number: 2010/07659-0.

REFERENCES

1. M. D. Glascock, H. Neff, K. J. Vaughn, "Instrumental Neutron Activation Analysis and Multivariate Statistics for Pottery Provenance", *Hyperfine Interactions*, **154**, pp. 95-105 (2004).

2. I. T. Jolliffe, "Discarding Variables in a Principal Component Analysis. I: Artificial Data", *Journal of the Royal Statistical Society*, **21**, pp.160-173 (1972).

3. M. J. Baxter, C. M. Jackson, "Variable Selection in Artefact Compositional Studies", *Archaeometry*, **43**, pp. 253-268 (2001).

4. J. M. Andrade, M. P. Gómez-Carracedo, W. Krzanowski, M. Kubista, "Procrustes rotation in analytical chemistry, a tutorial", *Chemometrics and Intelligente Laboratory Systems*, **72**, pp. 123-132 (2004).

5. K. Michelaki, R. G. V. Hancock, "Chemistry versus Data Dispersion: Is There a Better Way to Assess and Interpret Archaeometric Data?", *Archaeometry*, doi: 10.1111/j.1475-4754.2011.00590.x (2011)

6. M. M. Sithole, *Variable Selection in Principal Component Analysis: Using measures of Multivariate Association*, Thesis for Master of Science in Mathematics, School of Mathematics and Statistics, Curtin University of Technology (1992).

7. "Procrustes.", *Encyclopædia Britannica. Encyclopædia Britannica Online*. Encyclopædia Britannica, http://www.britannica.com/EBchecked/topic/477822/Procrustes (2011).

8. C. S. Munita, L. P. Barroso, P. M. S. Oliveira, "Variable selection using Procrustes analysis with stopping rule in archaeometric studies", *Proceeding of the 38th International Symposium on Archaeometry*, Tampa, Florida, May 10th-15st (2010).

***ISBN:* 978-85-99141-04-5**

9. W. J. Krzanowski, "Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components", *Journal of the Royal Statistical Society*, **36(1)**, pp. 22-33 (1987).

10. W. J. Krzanowski, "A stopping rule for structure-preserving variable selection", *Statistics and Computing*, **6**, pp. 51-56 (1996).

11. P. Bode, *Instrumental and Organizational Aspects of a Neutron Activation Analysis Laboratory*, Delft University of Technology, The Netherlands (1996).

12. R.C. Portocarrero, *A variabilidade espacial no sítio Osvaldo. Estudo de um assentamento da tradição barrancóide na Amazônia central*, Master´s Dissertation, Museum of Archeology and Ethnology, University of São Paulo (2006).

13. C. S Munita, R. P. Paiva, M. A. Alves, P. M. S. Oliveira, E. F. Momose, "Provenance study of archaeological ceramic", *Journal of Trace and Microprobe Techniques*, **21(4)**, pp. 697-706 (2003).