Variable selection study using Procrustes analysis

Casimiro Sepúlveda Munita,[1]
Lúcia Pereira Barroso,[2]
Paulo M.S. Oliveira[1]

[1]Instituto de Pesquisas Energéticas e Nucleares, Universidade de São Paulo, São Paulo; [2]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil

Abstract

Several analytical techniques are often used in archaeometric studies, and when used in combination, these techniques can be used to assess 30 or more elements. Multivariate statistical methods are frequently used to interpret archaeometric data, but their applications can be problematic or difficult to interpret due to the large number of variables. In general, the analyst first measures several variables, many of which may be found to be uninformative, this is naturally very time consuming and expensive. In subsequent studies the analyst may wish to measure fewer variables while attempting to minimize the loss of essential information. Such multidimensional data sets must be closely examined to draw useful information. This paper aims to describe and illustrate a stopping rule for the identification of redundant variables, and the selection of variables subsets, preserving multivariate data structure using Procrustes analysis, selecting those variables that are in some senses adequate for discrimination purposes. We provide an illustrative example of the procedure using a data set of 40 samples in which were determined the concentration of As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U obtained via instrumental neutron activation analysis (INAA) on archaeological ceramic samples. The results showed that for this data set, only eight variables (As, Cr, Fe, Hf, La, Nd, Sm, and Th) are required to interpret the data without substantial loss information.

Introduction

Ceramics are one of the main types of artifacts used by archaeologists, because these materials are a product of human activity and are recognizable when raw materials are displaced from their natural settings. A scientific account of this type of record requires a description of the kind and quantities of raw materials that were displaced, along with the distance and direction of movement. Archaeologists commonly refer of this type of study as artifact sourcing or provenance determination. Provenance studies permit archaeologists to investigate such diverse topics as mobility patterns, prehistoric migrations and commerce, and they are essential to understanding cultural development. For many years, these studies have investigated the provenance or other aspects of ceramic fragments and have utilized a number of techniques to classify these materials into a particular group (Tite, 2008). One method that has been used in such studies involves classifying samples according to their physical characteristics, such as color, texture, decoration, and style. An essential problem is this type of analysis is that ceramics manufactured in different places can appear to be identical based on visual inspection alone. Another frequently used method is a form of chemical *fingerprinting* of the ceramic fragments, in which their elemental composition is determined by chemical analysis (Glascock, 1992; Baxter *et al.*, 2008). It is becoming more common to analyze samples using more than one analytical technique, such as instrumental neutron activation analysis (INAA) and X-ray fluorescence (XRF), among others. In such cases, the number of the variables assessed is approximately 30 or more. To study these data sets, it is necessary to use multivariate statistical methods, such as cluster, principal components and/or discriminant analyses. However, difficulties and problems can arise when the number of variables increases without an increase in the number of samples (Baxter and Jackson, 2001). When multivariate statistical methods are used, the requirement is that the number of samples in a group exceeds the number of variables, preferably by a factor at least three (Baxter and Jackson, 2001). When this condition is not satisfied, it is necessary to reduce the number of variables used in the analysis, which can be accomplished through variable selection.

The purpose of this study is to reduce the number of variables (elements) used. This goal differs from the widely accepted practice in ceramic studies which assumes that using the largest possible number of measured variables is better (Harbottle, 1982). However, there is a distinction between the number of variables measured and the number that should be used in the study. Normally, the analyst measures a large number of variables, many of which may not be very informative. The variables used in the analysis need to show different concentrations in different types of ceramics and should also show small variations in ceramics of the same type. Among the various techniques, INAA employing g-ray spectrometry seems to be the most suitable analytical technique because it does not require mineralization of samples and allows the determination of numerous elements simultaneously with high sensitivity, accuracy and precision (Bishop *et al.*, 1990).

Thus, the aim of this paper is to identify the most relevant subset of variables, and to remove the variables with the least amount of relevant information, while preserving multivariate data structure and minimising the loss of essential information. In other words, we seek to select those variables that are in some sense adequate for discrimination purposes. We used Procrustes analysis in conjunction with a stopping rule. This procedure seems to perform well and is especially useful in archaeometric studies when the initial structure of the data set is unknown and when a principal components analysis (PCA) is used. To verify the reliability of the procedure, this technique was applied using the results of ceramic samples from one archaeological site.

Correspondence: Casimiro Sepúlveda Munita, Instituto de Pesquisas Energéticas e Nucleares, Universidade de São Paulo, Avenida Lineu Prestes 2242, 05508-000 São Paulo, Brazil.
Tel. +55.11.3133.9984 - Fax: +55.11.3133.9960.
E-mail: camunita@ipen.br

Key words: Procrustes analysis, archaeometry, variable selection, archaeological ceramic.

Citation: Munita et al., 2013. Variable selection study using Procrustes analysis. In: RH Tykot (ed.), Proceedings of the 38th International Symposium on Archeometry – May 10th-14th 2010, Tampa, Florida. Open Journal of Archaeometry 1:XXX.

Acknowledgments: the authors wish to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), contracts 06/57343-3 and 08/54867-7, for financial support.

Conflict of interests: the author declares no potential conflict of interests.

Presented at the 38th International Symposium on Archaeometry – May 10th-14th 2010, Tampa, Florida.

Open Journal of Archaeometry 2013; 1:e7
doi:10.4081/arc.2013.e7

Materials and Methods

Sample preparation

The ceramic powder samples were obtained by cleaning the outer surface of the ceramic and drilling with a tungsten carbide rotary file attached to the end of a variable-speed drill with a flexible shaft. Five holes were drilled as deep into the core of the ceramic material as possible without drilling through the walls. Forty ceramic samples were analyzed. After that, the materials were dried in an oven at 105°C for 24 h (Santos *et al.*, 2009).

Constituent elements in coal fly ash (NIST-SRM-1633b) were used as standards. IAEA-Soil-7, trace elements in soil, was used to check samples in each analysis. These materials were also dried in an oven at 105°C for 4 h (Santos *et al.*, 2009).

Approximately 100 mg of samples, along with NIST-SRM-1633b and IAEA Soil-7 were irradiated in the research reactor pool, IEA-R1, from the IPEN-CNEN/SP, at a thermal neutron flux of approximately 5×10^{12} cm^{-2} s^{-1} for 8 h.

Two measurement series were carried out using a Ge (hyperpure) detector, model GX 1925 from Canberra with a resolution of 1.90 keV at a gamma peak of ^{60}Co 1332.49 keV and S-100 MCA with 8192 channels. As, La, Na, Sm, and U were measured after a 7-days cooling period and Ce, Cr, Eu, Fe, Hf, Nd, Sc, and Th were measured after 25-30 days. The gamma ray spectra analysis and the concentrations were carried out using the Genie-2000 neutron activation analysis processing procedure from Canberra (Santos *et al.*, 2009).

Procrustes analysis

The idea of the Procrustes analysis is to select a subset of variables that preserve the structure revealed by PCA from the full data set. To illustrate this procedure, we will consider a data base of 40 samples of the ceramic fragments for which levels of As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U had been determined via INAA. This data will be represented in matrix X with *n* samples analyzed and *p* variables (elements) measured. This resulted in an *n* x *p* data matrix or a 40x13 matrix. If a PCA is applied on the *n* x *p* data matrix X, the scores of the *n* samples on the first *k* principal components are retained in the matrix *Z*, forming a new matrix *n* x *k*. So, the resulting matrix *Z* contains the scores of the first *k* principal components of data matrix X. When transformed the scores of the matrix *Z*(*n* x *k*) represent the best approximation to the original data configuration of X(*n* x *p*). If the first principal components are 2 or 3, *i.e.* *k*=2 or 3, plots based on the samples, *n*, of matrix *Z* can be used to identify patterns in the data.

Conversely, suppose that we select *q* variables from the original *p*, so that the selection represents the same structure as the original variables. For this to be true, *q* needs to be less than *p* and greater than or equal to *k*. Therefore, *q*<*p* and *q*≥*k*. In this case, suppose that $\tilde{X}$ is the matrix *n* x *q* which retains only *q* selected variables, and *Z* is the *n* x *k* matrix of the PC scores of these reduced data. $\tilde{Z}$ is therefore the best *k*-dimensional approximation to the *q*-dimensional configuration defined by the subset data. The concept of Procrustes uses the measured distance, M^2, between the two *k*-dimensional configurations *Z*(*n* x *k*) and $\tilde{Z}$ (*n* x *k*), and deletes the *p*-*q* variables in order to keep this distance as small as possible. The diagram shows the steps of the procedure:

$$\begin{tikzcd}
X(n \text{ x } p) \arrow[r, "\text{select}"] \arrow[d, "\text{PCA}"'] & \tilde{X}(n \text{ x } q) \arrow[r, phantom, "\text{PCA}"'] & q<p \arrow[d] & \text{(eq. 1)} \\
Z(n \text{ x } k) \arrow[r, leftrightarrow, "\text{Procrustes}"'] & \tilde{Z}(n \text{ x } k) & q\geq k &
\end{tikzcd}$$

The residual produced by the lost information through the deletion of some variables is the sum of squared differences between the two configurations, *Z* and *Z*, and is given by the expression:

$$M^2 = trace\ \{ZZ' + \tilde{Z}\tilde{Z}' - 2\tilde{Z}Q'Z'\} \qquad \text{(eq. 2)}$$

where *trace* is the sum of the diagonal elements of the matri is the transpose matrix, Q is given by multiplying two of the matrices of the singular value decomposition of the *k* x *k* square matrix $\tilde{Z}$'Z.

The value of M^2 is determined for each variable and the resulting value indicates the effect in the configuration and identifies the variable that has the lowest effect when eliminated. A practical backward elimination procedure is then used to find the minimum M^2, to delete the variable, and to repeat the process. The stopping rule for determining an appropriate value for the variable was discussed by Krzanowski (1987, 1996) who showed that if the variable is important for explaining the structure of the data, the sum of residues (M^2) will be higher than the critical value (*cv*). The critical value is approximately, $(1+c^2)\sigma^2$ times a chi-squared distribution on *nk*-1/2 *k* (*k*+1) degrees of freedom if the deleted variables do not influence the structure of the data, where $c=\sqrt{(p-i-k)/(p-k)}$. If some of the deleted variables influence the structure of the data, then the residual sum of squares will be greater. A suitable confidence level of the chi-squared distribution times $(1+c^2)\sigma^2$ will provide a stopping rule for the process until that of the calculated M^2 exceeds the critical value. However, s^2 is unknown, and it is necessary to replace this parameter with an estimator. More details of the procedure can be found elsewhere (Baxter *et al.*, 2008; Krzanowski, 1987, 1996).

Results and Discussion

The study was made using a data set of 40 ceramic samples which were assessed in terms of their As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U content by INAA. Table 1 shows the values of the elemental concentrations for the 40 samples. All of the elements used had precision of less than 10% and were tested using 25 independent determinations of the reference material IAEA Soil-7. The observed results were compared to the certified values. The precision level used in this study is in agreement with the criteria recommended for archaeometric studies (Bishop *et al.*, 1990).

In geochemistry, concentration data are often assumed to follow a lognormal distribution after being $\log_{10}$ – transformed, as suggested by Ahrens (1954); however, in geochemistry, this assumption rarely holds true. For the majority of the variables, a log base 10 transformation does not result in a normal distribution (Reimann and Filzmoser, 2000). This may have serious consequences for the further statistical treatment of data sets because the vast majority of advanced statistical methods require not only that each variable shows a normal distribution, but also that the variables show a multivariate normal distribution. In addition, although the data set does not present the total composition of the samples (*i.e.* the variables measured are <100%) this type of data frequently displays a curvature and linear techniques, such as principal component analysis cannot be used. The present study used the transformation proposed by Aitchison (1983), which transforms each sample x_{ij} (*i*=1,….*n* and *j*=1,…., *p*) in y_{ij} by taking the natural log transformation and subtracting the mean of the transformed variables, *i.e.*:

$$y_{ij} = \ln x_{ij}, \quad \bar{y}_i = \frac{1}{p}\sum_{j=1}^{p} y_{ij} \quad z_{ij} = y_{ij} - \bar{y}_i \quad \text{(eq. 3)}$$

In addition, the data were standardized to compensate for the large difference in magnitude between the measured elements at the trace level and the larger elements (Templ *et al.*, 2008). The method used was the z-transformation, in which the median is subtracted from the raw data and then divided by the median absolute deviation (MAD) as follows (Templ *et al.*, 2008):

$$\text{z-transformation} = \frac{z_{ij} - median\left(z_i\right)}{MAD\left(z_i\right)} \qquad \text{(eq. 4)}$$

After being transformed, the data set was submitted to outlying tests using the *Mahalanobis* distance (Oliveira *et al.*, 2010).

Outliers can have a considerable influence on PCs because they can disturb homogeneous groups.

The *Mahalanobis* distance is an important measure in statistic and has been suggested by many authors to be the method that should be used to detect outliers in multivariate data. For each of the n samples and p variables, the *Mahalanobis* distance (D_i) taken from the sample to the centroid is calculated by the expression (Penny, 1996):

$$D_i = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})} \qquad \text{(eq. 5)}$$

where ' is the transpose matrix; $S = \sum_{i=1}^{n} (x_i - \bar{x})' (x_i - \bar{x})$ is the variance-covariance sampling matrix; and $(x_i - \bar{x})$ is the vector of difference between the concentrations measured in one group and the concentrations measured in the other group. Each one of these values is compared with the critical value, *cv*, that can be calculated through the

Table 1. Results for ceramic samples in g g^{-1}, unless otherwise indicated (n=40).

Sample	As	Ce	Cr	Eu	Fe, %	Hf	La	Na, %	Nd	Sc	Sm	Th	U	D
1	1.82	104.71	141.25	1.15	2.63	9.33	26.92	0.05	25.70	28.18	6.31	16.22	3.31	9.21
2	1.45	104.71	134.90	1.29	2.63	8.32	29.51	0.06	38.02	27.54	6.76	15.85	3.47	7.75
3	1.29	109.65	138.04	1.35	2.88	8.32	30.90	0.05	33.88	29.51	6.76	17.78	1.82	16.78
4	1.82	117.49	154.88	1.41	2.95	8.71	33.11	0.06	32.36	30.20	7.41	18.62	3.47	3.84
5	1.58	112.20	162.18	1.35	3.02	9.12	29.51	0.05	25.70	31.62	6.76	17.78	4.07	9.48
6	1.82	112.20	169.82	1.26	3.02	9.55	30.20	0.06	26.92	31.62	7.08	17.38	4.27	10.87
7	1.58	120.23	151.36	1.38	2.82	8.13	33.11	0.06	33.88	28.84	7.41	18.20	5.01	10.45
8	1.51	107.15	109.65	1.48	3.24	9.55	40.74	0.08	39.81	26.30	7.76	18.20	4.27	20.71
9	1.55	109.65	114.82	1.38	2.14	7.59	28.18	0.05	28.84	30.20	6.76	16.22	3.31	9.25
10	1.55	117.49	144.54	1.41	2.82	8.32	32.36	0.07	30.90	29.51	7.24	17.38	4.17	4.94
11	1.41	112.20	141.25	1.35	2.82	8.32	31.62	0.06	36.31	28.18	7.08	16.98	2.57	6.40
12	2.20	127.06	142.89	2.39	3.44	8.00	71.45	0.21	61.94	14.79	9.29	12.39	1.20	7.64
13	2.00	141.91	165.96	3.07	4.13	8.30	86.50	0.24	71.94	16.87	11.61	13.90	1.40	3.65
14	2.40	132.43	147.91	3.06	3.78	8.09	80.91	0.30	63.97	15.28	11.72	10.50	1.70	12.61
15	2.20	110.92	154.88	2.70	4.45	7.91	75.68	0.19	69.02	14.72	10.26	10.89	1.30	11.33
16	2.40	143.55	147.23	3.79	3.22	7.66	100.23	0.18	102.09	16.41	13.49	12.62	1.40	14.65
17	2.10	123.88	141.91	2.62	3.88	8.30	72.95	0.24	66.07	14.79	9.66	12.05	1.20	8.47
18	2.50	160.32	182.81	3.79	3.88	7.60	96.83	0.26	68.08	18.03	13.09	14.19	1.20	15.48
19	2.20	141.58	159.96	3.23	4.57	8.30	95.72	0.13	79.98	16.71	12.25	13.49	1.10	13.95
20	0.99	120.78	140.93	2.84	3.26	7.00	87.10	0.14	59.02	14.86	11.19	12.19	1.50	18.34
21	2.70	123.03	186.21	2.72	3.32	8.59	71.61	0.24	59.02	17.58	8.95	13.00	1.50	22.55
22	3.00	127.35	165.96	2.63	4.10	9.91	80.91	0.22	71.94	16.98	11.17	14.00	1.20	16.78
23	1.10	116.41	130.02	2.13	2.60	7.80	66.53	0.14	43.95	12.68	8.15	11.19	1.20	21.03
24	1.60	82.04	187.07	3.20	1.87	10.79	37.24	0.03	46.99	37.15	9.79	4.80	1.20	7.77
25	1.50	90.78	302.69	3.20	3.03	10.99	39.54	0.03	52.00	41.69	10.21	5.60	1.10	10.80
26	2.40	85.11	213.80	3.30	2.14	10.79	37.58	0.02	52.97	43.85	10.74	5.20	1.20	9.35
27	1.60	82.00	187.00	3.20	1.87	10.80	37.20	0.03	47.00	37.17	9.80	4.80	1.20	7.87
28	1.80	101.39	230.14	3.40	2.30	11.69	45.50	0.01	51.05	44.98	11.43	7.69	1.30	13.64
29	1.40	95.28	244.91	3.50	2.45	12.11	43.95	0.02	57.02	42.95	11.35	5.79	1.40	3.80
30	1.90	109.65	217.77	3.29	2.18	11.69	37.76	0.02	59.98	39.36	10.30	5.20	1.10	22.89
31	1.70	87.70	240.99	3.30	2.41	10.89	40.83	0.02	70.96	45.60	11.02	7.00	1.30	12.50
32	1.60	78.89	230.14	3.20	2.30	10.89	41.11	0.02	69.02	39.99	11.32	5.11	1.10	12.63
33	1.50	90.80	303.00	3.20	3.03	11.00	39.50	0.03	52.00	41.72	10.21	5.60	1.10	10.78
34	1.40	93.11	243.22	3.44	2.43	12.79	40.93	0.02	53.95	45.81	11.40	6.10	1.20	9.47
35	1.60	109.90	260.02	3.80	2.60	12.30	48.31	0.02	59.02	44.06	13.24	5.79	0.90	13.46
36	1.70	95.28	204.17	3.42	2.04	12.50	43.45	0.02	47.97	50.12	11.04	6.75	1.20	18.05
37	1.30	89.13	248.89	3.40	2.49	12.30	39.54	0.02	61.94	48.87	11.09	5.70	1.40	6.22
38	2.40	123.31	223.87	4.31	2.24	12.79	51.52	0.02	57.94	47.75	14.03	7.40	1.60	11.67
39	1.80	97.50	238.23	3.27	2.38	11.91	38.02	0.02	52.00	42.27	10.35	6.19	1.80	10.97
40	1.80	92.68	252.93	3.60	2.53	12.79	44.16	0.01	62.95	48.31	11.69	6.40	1.20	9.97
Total														25.38

As, arsenic; Ce, cerium; Cr, chromium; Eu, europium; Fe, iron; Hf, halfnium; La, lanthanum; Na, sodium; Nd, neodymium; Sc, scandium; Sm, samarium; Th, thorium; U, uranium; D, critical value at significance level of 0.05.

lambda Wilks criteria (Penny, 1996), calculated by:

$$\frac{p(n-1)^2 F_{p,n-p-1;\alpha/n}}{n(n-p-1+pF_{p,n-p-1,\alpha/n})} \qquad \text{(eq. 6)}$$

where p is the number of variables; n is the number of samples and , is the F test called the Fisher distribution ($F=s_1^2/s_2^2$where s_1^2 and s_2^2 are the sample variances) with p degrees of freedom at a significance level of α/n, $\alpha=0.05$.

When the value found by expression (5) is larger than the critical value produced by expression (6), the sample is considered to be an outlier (Penny, 1996). Thus, the *Mahalanobis* distance for each sample was calculated and compared to the critical value. The last column of Table 1 shows the *Mahalanobis* distance values for each sample, as well as the end for the critical value, calculated using the lambda Wilks criteria. The stopping rule is applied when the *Mahalanobis* distance calculated in the samples does not exceed the critical value. In accordance with the *Mahalanobis* distance rule, any sample in the Table 1 could be an outlier.

To verify the reduction of data dimensionality in the compositional analysis (in other words, to eliminate variables without altering data structure), the data were studied using a Procrustes analysis.

Applying the robust PCA to the natural log-transformed and standardized data sets indicated that the variance explained in the first and fourth robust PCs was 51.1, 28.8, 8.2 and 5.7%, respectively, representing 93.8% of the total variance. Thus, using k=2 seems to be adequate because the first two robust PCs explain 79.9% of the total variance. Table 2 shows the results of the selection procedure, including the sequence of elimination.

In Table 2, the variable Eu is the first element for elimination because the value of M^2 is 2.3. This parameter represents the distance of the scores of the robust PC of the two matrices while using all variables and represents the loss of information caused by the elimination of the variable. To determine whether the robust scores of the two configurations are significantly different, the critical value (*cv*) was calculated using the Krzanowski stopping rule at 5% of the significance level (Krzanowski, 1996). As shown in Table 2, the critical value for Eu was 48.8, which is higher than 2.3 (the value of M^2). This shows that the elimination of Eu does not significantly affect the scores in the configuration of the PCs. When the variables are eliminated, the distance of the PC scores, M^2, increases and the critical value that depends on the number of variables, decreases, until the elimination of the variable affects the associated configuration. This point is reached when M^2 is greater than the critical value, and when Sm is deleted. This procedure suggests that the stopping rule be applied at the point at which $M^2 \geq cv$. This suggests that Sm, As, Cr, Fe, Hf, La, Nd, and Th must be retained without the loss of information in the configuration. To confirm this assumption, the same data set were submitted to a robust PCA. The PCA plot is useful for visually displaying group separation. A bivariate plot of first two robust principal components using all the elements is presented in Figure 1. The results show that the samples form three clusters with chemically homogeneous groups with a high degree of chemical similarity among the groups. Figure 2 shows the plot for the first two robust principal components using the selected variables. The variance explained for the first two robust principal components was 76.6 and for the fourth PC was 96.1%. In both plots the ellipses represent a confidence level of 90%. Comparison of Figures 1 and 2 shows that a PCA performed using only on eight variables produced similar results to a PCA produced using all variables. In other words, these results indicate that for this data set, only eight variables are required to interpret the data without a substantial loss of information because the plots for both configurations are similar. In addition, in order for Procrustes

Table 2. Results of the deletion procedure for the data set (n=40).

Variable	Eu	Sc	Ce	Na	U	Sm	As, Cr, Fe, Hf, La, Nd, Th
M2	2.3	6.0	11.6	19.2	30.5	45.5	-
cv	48.8	46.0	43.1	40.2	37.4	34.5	-

Eu, europium; Sc, scandium; Ce, cerium; Na, sodium; U, uranium; Sm, samarium; As, arsenic; Cr, chromium; Fe, iron; Hf, halfnium; La, lanthanum; Nd, neodymium; Th, thorium; M^2, measured distance; *cv,* critical value.

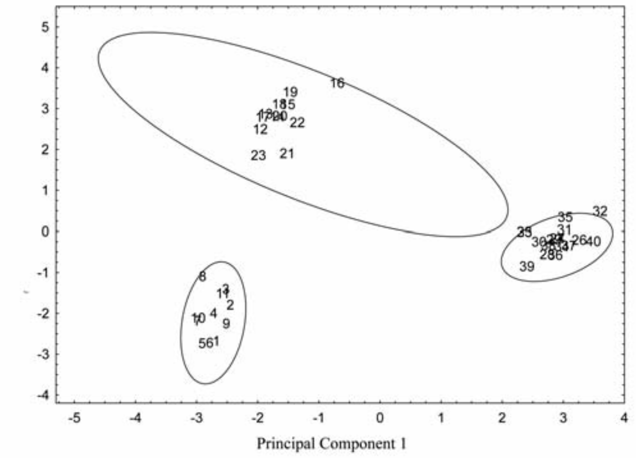

Figure 1. Plot of the first two robust principal components for all variables (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U). The ellipses represent a confidence level of 90%.

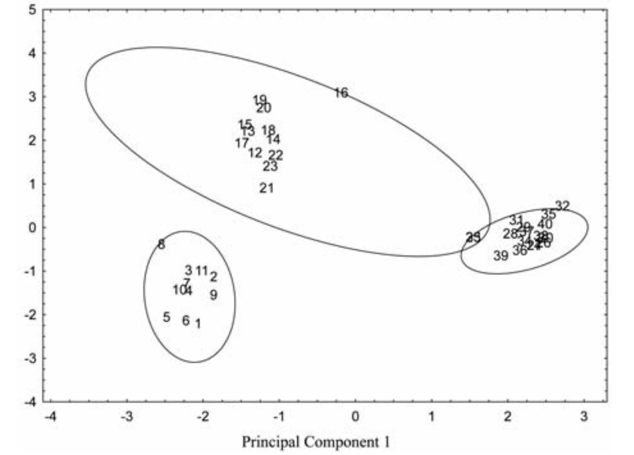

Figure 2. Plot of the first two robust principal components for selected variables (Sm, As, Cr, Fe, Hf, La, Nd and Th). The ellipses represent a confidence level of 90%.

analysis to be used to obtain good results in the configurations, the variance explained by the first two components should be high.

Conclusions

This paper demonstrates with one illustrative example, that it is possible to determine a subset of variables from a data matrix using the Procrustes analysis without losing information in the data set. This finding was confirmed by a robust principal component analysis based on the best eight variables, because the subsets captured all of the information. The PCA produced using eight variables gave results similar to those of the PCA produced using all of the variables. This paper provides an important contribution to archaeometric studies using a compositional data set by demonstrating that it is possible to use a subset of variables obtained through a Procrustes analysis without losing information.

References

Ahrens LH, 1954. The lognormal distribution of the elements (a fundamental law of geochemistry and its subsidiary). Geochim Cosmochim Ac 5:49-73.

Aitchison J, 1983. Principal component analysis of compositional data. Biometrika 70:57-65.

Baxter MJ, Beardah CC, Papageorgiou I, Cau MA, Day PM. On statistical approaches to the study of ceramic artifacts using geochemical and petrographic data. Archaeometry 2008;50:142-57.

Baxter MJ, Jackson CM, 2001. Variable selection in artefact composition studies. Archaeometry 43:253-68.

Bishop RL, Canouts V, Grown PL, Attas M, De Atley SP, 1990. Sensitivity, precision, and accuracy: their roles in ceramic compositional databases. Am Antiquity 55:537-46.

Glascock MD, 1992. Characterization of ceramics at MURR by NAA and multivariate statistics. In: Neff H (ed.), Chemical characterization of ceramic pastes in archaeology. Prehistory Press, New York, NY; USA, pp 11-26.

Harbottle G, 1982. Chemical characterization in archaeology. In: Ericson JE, Earle TK (ed.), Contexts for prehistoric exchange. Prehistory Press, New York, NY; USA, pp 13-51.

Krzanowski WJ, 1987. Selection of variables to preserve multivariate data structure, using principal components. Appl Statist 36:22-33.

Krzanowski WJ, 1996. A stopping rule for structure-preserving variable selection. Stat Comput 6:51-6.

Oliveira PMS, Munita CS, Hazenfratz R, 2010. Comparative study between three methods of outlying detection on experimental results. J Radioanal Nucl Ch 283:433-7.

Penny KI, 1996. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. Appl Statist 45:73-81.

Reimann, C, Filzmoser P, 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ Geol 39:1001-14.

Santos JO, Munita CS, Toyota RG, Vergne C, Silva RS, Oliveira PMS, 2009. The archaeometry study of the chemical and mineral composition of pottery from Brazil's Northeast. J Radioanal Nucl Ch 281:189-92.

Templ M, Filzmoser P, Reimann C, 2008. Cluster analysis applied to regional geochemical data. Problems and possibilities. Appl Geochem 23:2198-213.

Tite MS, 2008. Ceramic production, provenance and use: a review. Archaeometry 50:216-31.