Variable selection using Procrustes analysis with stopping rule in archaeometric studies

C.S. Munita, L.P. Barroso[1] and P.M.S. Oliveira

Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP, Av. Prof. Lineu Prestes 2242, Cidade Universitária, USP, CEP 05508-000 São Paulo, SP, Brazil, e-mail: camunita@ipen.br

[1]Instituto de Matemática e Estatística, Cidade Universitária, USP, C.P. 66281, CEP 05315-970, São Paulo, SP, Brazil

Abstract Several analytical techniques are used in archaeometric studies, and in combination, this can result in 30 or more elements being determined. Frequently multivariate statistical methods are used to interpret such data set, but their applications can be problematic or difficult to interpret with too many variables. In this paper, the application of Procrustes analysis with a stopping rule for the identification of redundant variables is presented. One illustrative example of the procedure being done with a data set obtained via instrumental neutron activation analysis, INAA, on archaeological ceramic samples is provided.

Keywords: Procrustes analysis, principal components analysis, archaeometry, variable selection, INAA, arcaheological ceramic

Introduction

Ceramics is one of the main categories of artefacts used by archaeologists because is a product of human activity and is recognizable by displacement of raw material from their natural settings. A scientific account of this record necessitates a description of the kind and amounts of raw materials that were displaced, along with the distance and direction of movement. Archaeologits commonly refer of this type of study as artefact 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. Then, for many years they have been interested in the provenance or of another kind of study of ceramic fragments and have utilized a number of techniques to classify these materials into a particular group [1]. One method used for establishing such studies has been to

classify samples according to their physical characteristics, such as color, texture, decoration, style and so on. An essential problem is that ceramics manufactured in different places can appear to be identical on the basis of visual inspection only. Another method has been to utilize a form of chemical "fingerprinting" of the ceramic fragments by determining their elemental composition [2, 3].

It is becoming more frequent to have the same sample analyzed by more than one analytical technique, such as instrumental neutron activation analysis, INAA, X-ray fluorescence, XRF, among others. In such a case the number of the variable determined is about 30 or more. In order to study the data set, it is necessary to use multivariate statistical methods like cluster, principal components and/or discriminant analysis. However, difficulties and problems can appear when the number of determined variables has increase without an increase to the number of samples [4]. It is known that when multivariate statistical methods are used, the requirement exist that the number of samples in a group exceeds the number of variables, preferably by a factor at least three [4]. When this condition is not satisfied, it is necessary for some form of variable reduction, which can include variable selection.

The purpose in this work in to reduce the number of variables (elements) used that may apparently to seem to be against what is accepted widely in ceramic studies, in that the greater the number of variables measured is better [5]. However, is not always recognized that there is a distinction between the number of variables measured and the number needed to be used in the study. Normally the analyst measures a large number of variables, many of which may not be very informative. In order to do that a measured variable has to be included in ceramic studies and have certain. This provides a good base for archaeological interpretation. It additionally needs to show different concentrations in ceramics of different types and small variations in ceramic of the same type. This is all while covering a wide range of chemical properties. This is in order to be determined with analytical precision of less than 10% [6].

So, the purpose of this paper is to identify a subset of the variables that are truly the most relevant. Also, it is to remove the less productive information and preserving multivariate data structure without losing essential information. In other words, by selecting those variables which are in some sense adequate for discrimination purposes. The procedure used was the Procrustes analysis in conjunction with a stopping rule. This procedure seems to be adequate. Especially in archaeometric studies when the initial structure in the data set is unknown. It is also when the principal components

analysis, PCA, is used. This study was based using the results of ceramic samples from one archaeological site.

Experimental

Sample preparation and description of the method

The ceramic powder samples were obtained by cleaning the outer surface and drilling, using 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, these materials were dried in an oven at 105^{o}C for 24 h [7].

Constituent Elements in Coal Fly Ash (NIST-SRM-1633b) were used as standards. Then IAEA-Soil-7, Trace Elements in Soil, was used to check samples in every analysis. These materials were also dried in an oven at 105^{o}C for 4 h [7].

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

Two measurement series were carried out using a Ge (hyperpure) detector. It was a model GX 1925 from Canberra with a resolution of 1.90 keV. It had a gamma peak of ^{60}Co 1332.49 keV. It had a Canberra S-100 MCA with 8192 channels. As, La, Na, Sm, and U were measured after 7 days cooling time and Ce, Cr, Eu, Fe, Hf, Nd, Sc, and Th, 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 [7].

Procrustes analysis [3]

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 explain the procedure we consider a data base of 40 samples of the ceramic fragments which were determined As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U via INAA. To facilitate the explanation the procedure will be presented by a data matrix **X** with *n* samples analyzed and *p* variables (elements) measured. Then there is a *n* x *p* data matrix or 40 x 13 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 obtained matrix ***Z*** has the scores of the first *k* principal components of data matrix **X**. The scores of the matrix ***Z***(*n* x *k*) transformed have 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 patterning in the data.

On the other hand, suppose that there are selected q variables from the original p, to that the selection does take place and it recovers the same structure that with the original variables, q needs to be less than p and higher than or equal to k. Therefore, $q < p$ and $q \geq k$. In this case, suppose that $\tilde{X}$ is the matrix n x q which retains only q selected variables, and $\tilde{Z}$ is the n x k matrix of the PC scores of these reduced data, this later is therefore the best k-dimensional approximation to the q-dimensional configuration defined by the subset data. The Procrustes idea is to measure the distance, M^2, between the two k-dimensional configurations ***Z***(n x k) and $\tilde{Z}$ (n x k), and to delete the p - q variables that keep this distance as small as possible. The diagram shows the steps of the procedure:

$$
\begin{array}{lcll}
X\,(n \text{ x } p) & \xrightarrow{\text{select}} & \tilde{X}\,(n \text{ x } q) & q < p \\
\text{PCA} \downarrow & & \text{PCA} \downarrow & \\
\boldsymbol{Z}\,(n \text{ x } k) & \xleftrightarrow{\text{Procrustes}} & \tilde{Z}\,(n \text{ x } k) & q \geq k
\end{array}
$$

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 $\tilde{Z}$, and is given by the expression

$$M^2 = \text{trace}\left\{\text{ZZ}' + \tilde{\text{Z}}\tilde{\text{Z}}' - 2\tilde{\text{Z}}\text{Q}'\text{Z}'\right\}$$

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

The value of M^2 is determined for each variable and the value found indicate the effect in the configuration and identifies the variable whose elimination has the least affects. Then, a practical backward elimination procedure is 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 [9] who showed that if

the variable is important to explain the data structure, the sum of residues will be higher than the critical value (*cv*). The critical value has, 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 are not structuring-carrying, where $c = \sqrt{(p-i-k)/(p-k)}$. If some of the deleted variables are structure-carrying, then the residual sum of squares will clearly 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, σ^2 is unknown, and it is necessary to replace it by its estimator. More details of the procedure is possible to find elsewhere [8, 9].

Results and discussion

The study was made using a data set of 40 samples of ceramic samples which were determined As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U by INAA. Table 1 shows the values of the elemental concentrations for 40 samples. All the elements used had precision less than 10% that was tested using 25 independent determinations of the reference material IAEA Soil-7. For that the results found were statistically compared with the certified values. The precision chosen it is in agreement with the criteria recommended for archaeometric studies [6].

Initially the results were transformed to $\log_{10}$, in order to compensate for the large magnitude differences between the measured elements at the trace level and the larger ones. Another reason for this is the belief that within manufactured raw materials, elements have a natural log normal distribution, and that data normalization is desirable [2]. So, throughout the work, it was assumed that the data sets were log-normally distributed.

After logarithmic transformation, the data set was submitted to outlying tests, using the *Mahalanobi*s distance [10].

The *Mahalanobis* distance is an important measure in statistic and it is suggested by many authors as the method to detect outliers in multivariate data. For each one of the *n* samples and *p* variables, the *Mahalanobis* distance (D_i) from the sample to the centroid is calculated by the expression [11]

$$D_i = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})} \quad (1)$$

Table 1. Results for ceramic samples in μg/g, 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	8.75
2	1.58	138.04	186.21	1.29	1.74	10.96	38.90	0.06	44.67	26.92	8.13	19.50	4.68	7.43
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	16.06
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	3.65
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	9.28
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	11.21
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	10.47
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	22.41
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	11.87
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	4.97
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	5.77
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	6.80
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	4.25
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	14.59
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	12.91
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	14.83
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	8.19
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	16.04
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	14.35
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	17.08
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	21.73
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	18.05
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	22.64
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	9.71
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	10.71
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	4.77
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	9.80
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	8.33
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	3.55
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	22.16
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	12.33
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	13.97
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	10.68
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	6.06
35	1.60	109.9	260.02	3.80	2.60	12.30	48.31	0.02	59.02	44.06	13.24	5.79	0.90	13.55
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	19.3
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	5.03
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	14.53
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	13.35
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	6.83

$D_{critical\ value\ at\ significance\ level\ of\ 0.05}$ 25.38

where ´ is the transpose matrix; $S = \sum_{i=1}^{n} (x_i - \bar{x})' (x_i - \bar{x})$ is a variance-covariance sampling matrix; and, $(x_i - \bar{x})$ is the vector of difference between the concentrations measured in a group and the concentrations measured in the other group. Each one of these values is compared to the critical value that can be calculated through the lambda Wilks criteria [11], *cv*, calculated by

$$\frac{p(n-1)^2 F_{p,n-p-1;\alpha/n}}{n(n-p-1+pF_{p,n-p-1,\alpha/n})} \quad (2)$$

where *p* is a number of variables; *n* is a number of samples and F , is the *F* test called 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 the expression (1) is larger than the critical value by the expression (2), the sample is considered an outlier [11]. So, the *Mahalanobis* distance for each sample was calculated and the critical value. In the last column of Table 1 are the *Mahalanobis* distance values of each sample, and the end for the critical value, calculated using the lambda Wilks criteria [11]. The stopping rule is when the *Mahalanobis* distance calculated in the samples does not exceed the critical value. In accordance with the *Mahalanobis* distance, in the Table 1 any one sample is outlier.

With the purpose of verifying the possibility of the reduction of data dimensionality in the compositional analysis or, in other words, to eliminate variables without altering data structure, the data was studied through the Procrustes analysis.

Applying PCA in the log normalized data sets showed that the variance explained in the first fourth PCs was 47.0, 36.7, 6.5 and 3.3%, respectively, being 93,5% the total variance. Then using $k = 2$ seems to be adequate because the first two PC explain 83.7% of the total variance. Table 2 shows the results of the selection procedure, including the sequence of elimination.

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

Variable	Na	Eu	La	Th	Hf	Nd, Ce, Cr, U, As, Fe, Sc, Sm
M^2	2.2	4.5	11.2	18.4	30.2	
cv	35.5	33.4	31.3	29.2	27.1	

In the Table 2, the variable Na is the first element for elimination because the value of M^2 is 2.2 which measures the distance of the scores of the PC of the two matrices, while using all variables, and represents the loss of information caused by the elimination of the variable. To know if the scores of the two configurations are significantly different, it was calculated the critical value (*cv*) as obtained using Krzanowski stopping rule at 5% of the significance level [9]. As can be seen in Table 2 the critical value for Na was 35.5, which is higher than 2.2 (M^2). This shows that the elimination of Na does not affect significantly 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 a point in which the elimination of the variable will affect the associated configuration. This point is reached when M^2 is greater than the critical value, and this point is reached when Hf is deleted. This procedure suggests the application of the stopping rule at the point in which $M^2 \geq cv$. This suggests that Hf, Nd, Ce, Cr, U, As, Fe, Sc and Sm must be retained without loss information in the configuration. To confirm this assumption, the same data set were submitted to PCA. The plot is useful for visually displaying group separation. A bivariate plot of two first principal components using all the elements is presented in Figure 1. As can be seen, the results show that the samples form three

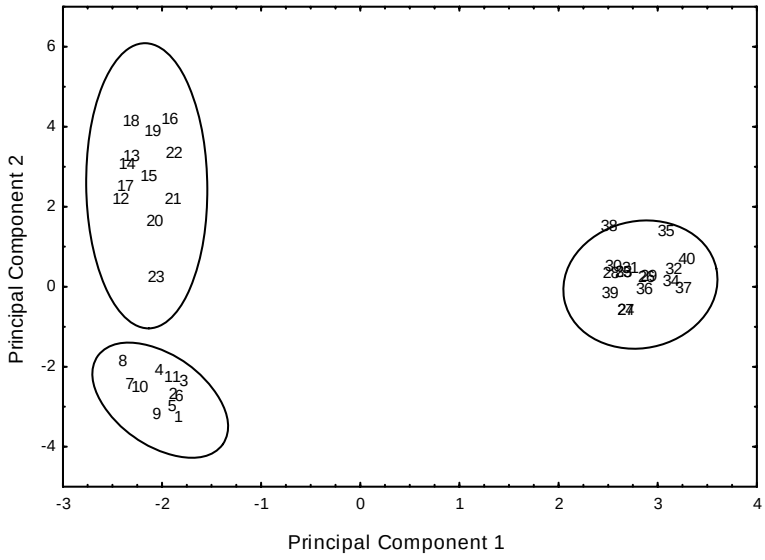

Figure 1. Plot of the first two principal components for all variables, As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U, n=40.

clusters with chemically homogeneous groups, showing a high degree of chemical similarity among them. Figure 2 shows the plot for the two first principal components using the selected variables.

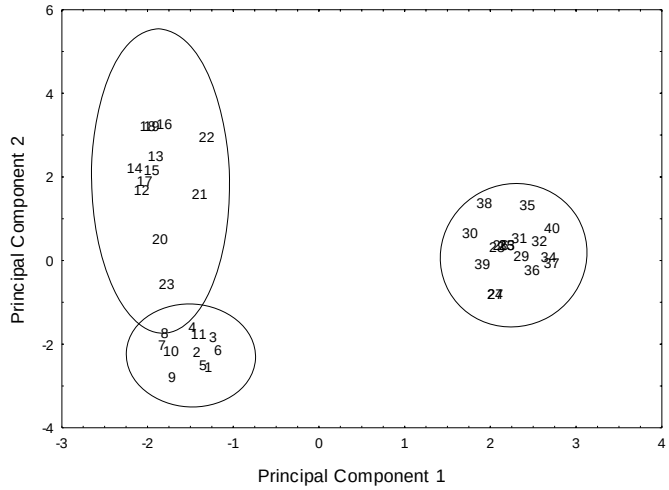

Figure 2. Plot of the first two principal components for selected variables, Hf, Nd, Ce, Cr, U, As, Fe, Sc and Sm for dataset.

The variance explained for the first two principal component was 79.3% and for the fourth PC was 92.6%. In both plots the ellipses represent a confidence level of 95%. When comparing the Figures 1 and 2, it is possible to see that when PCA was applied based only on the 9 variables, there were similar results to a PCA using all variables. In other words, it proves that for this data set only nine variables are sufficient to do the interpretation without loss information because the plots for both configurations are similar. In addition, it is important to have account that to use Procrustes analysis and to obtain good results in the configurations, it is necessary that the variance explained by the first two components need to be high.

Conclusion

In this paper, it was shown, with one illustrative example, that in a data matrix it is possible to determine a subset of variables using the Procrustes analysis without loss

information in the data set. The study was confirmed by using a principal component analysis based on the best nine variables. This produces similar results to a PCA using all the variables. This paper have provided important contribution to archaeometric

studies using compositional data set because was demonstrated that it is possible to use a subset of variable obtained via Procrustes analysis without loss information.

Acknowledgements

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

References

1. Tite MS (2008) Ceramic production, provenance and use – a review. Archaeometry 50(2):216-231.

2. Glascock MD (1982) Characterization of ceramics at MURR by NAA and multivariate statistics. In: Neff H (ed) Chemical characterization of ceramic paste in archaeology, Monographs in World Archaeology. Prehistoriy Press, NewYork, section 1, pp 11-26.

3. Baxter MJ, Beardah, CC, Papageorgiou I, Cau MA, Day, PM (2008) On statistical approaches to the study of ceramic artifacts using geochemical and petrographic data. Archaeometry 50(1):142-157.

4. Baxter MJ, Jackson CM (2001) Variable selection in artifact compositional studies. Archaeometry 43(2):253-268.

5. Harbottle G (1982) Chemical characterization in archaeology. In: Jonathon EE and Earle TK (eds) Contexts for prehistoric exchange, Academic Press, NewYork, section 2, pp 13-51.

6. Bishop RL, Canouts V, Grown PL, Attas M, De Atley SP (1990) Sensitivity, precision, and accuracy: their roles in ceramic compositional databases. Amer Antiquity 55(3):537-546.

7. Santos J.O., Munita C.S., Toyota R.G., Vergne C., Silva R.S., Oliveira P.M.S (2009). The archaeometry study of the chemical and mineral composition of pottery from Brazil´s Northeast. J Radioanal Nucl Chem 281:189-192.

8 Krzanowski WJ (1987) Selection of variables to preserve multivariate data structure, using principal components. Applied Statist 36(1):22-33.

9. Krzanowski WJ (1996) A stopping rule for structure-preserving variable selection. Statistics and Computing 6:51-56.

10. Oliveira PMS, Munita CS, Hazenfratz R (2010) Comparative study between three methods of outlying detection on experimental results. J Radioanal Nucl Chem 283(2): 433-437.
11. Penny KI (1996) Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. Appl Statist 45(1):73-81.