2009 International Nuclear Atlantic Conference - INAC 2009
Rio de Janeiro,RJ, Brazil, September27 to October 2, 2009
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
ISBN: **978-85-99141-03-8**

COMPARATIVE STUDY ON METHODS OF OUTLYING DATA DETECTION IN EXPERIMENTAL RESULTS

P.M.S. Oliveira, C.S. Munita, R. Hazenfratz

Instituto de Pesquisas Energéticas e Nucleares, IPEN - CNEN/SP
Av. Professor Lineu Prestes, 2242
05508-000 São Paulo, SP
ptoliveira@ipen.br
camunita@ipen.br
robertohm@usp.br

ABSTRACT

The interpretation of experimental results through multivariate statistical methods might reveal the outliers existence, which is rarely taken into account by the analysts. However, their presence can influence the results interpretation, generating false conclusions. This paper shows the importance of the outliers determination for one data base of 89 samples of ceramic fragments, analyzed by neutron activation analysis. The results were submitted to five procedures to detect outliers: Mahalanobis distance, cluster analysis, principal component analysis, factor analysis, and standardized residual. The results showed that although cluster analysis is one of the procedures most used to identify outliers, it can fail by not showing the samples that are easily identified as outliers by other methods. In general, the statistical procedures for the identification of the outliers are little known by the analysts.

1. INTRODUCTION

Since the middle of the 20th century, it has increased the concern by the statisticians to detect and to treat atypical experimental results. Among the employed methods there are Mahalanobis distance [1], mask [2], ellipsoid minimum volume [3] and decisive minimum of the covariance matrix [3]. In general, the authors concluded that it is not possible to determine, with precision, the outliers in a data set [1, 2, 3].

Outlying results can happen due to any of the following problems: uncontrolled process; wrong analytical technique; contamination during the preparation of the sample; measurements with high error; transcription mistake; mistake when considering a sample that does not belong to the group of interest, and other factors. In general, the identification of the outlying values is subjective, although different statistical methods exist [4].

In the literature, not considering the publications on statistics, few works have been published about the identification of outlying values in samples that involve more than one variable. Most of the proposed methods are graphical and subjective. The presence of outliers can bring distortions in the results of the models and estimates. Therefore, their detection is very important and should be done before data analysis [5, 6]. A comparative study between different methods of detection, for this purpose, is necessary in the experimental results.

In this work, a comparative study of the effect of the outlying values was performed using 5 methods: Mahalanobis distance, cluster analysis (Euclidean distance and average linkage), principal component analysis, factor analysis and standardized residual. These studies were accomplished using a data base of 89 samples whose variables were the elemental concentration of As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Tb and U, obtained by neutron activation analysis. Initially, it was assumed that the results obey a normal distribution. For this, the transformation to a 10 logarithm base normalized the elementary concentrations, serving, also, to compensate the differences in the magnitude of the elements that are in percentage, and those that are at trace level [7].

2. DEVELOPMENT

2.1 Mahalanobis Distance

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

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

for $i = 1,...,n$.

where, $S = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)' \left(x_i - \bar{x}\right)$ is a variance-covariance sampling matrix; and, $\left(x_i - \bar{x}\right)$ is the vector of difference between the concentrations measured in a group and the concentrations measured in the other group. Each of these values is compared to the critical value that can be calculated through the Wilks lambda criterion [8, 9], defined by:

$$\frac{p(n-1)F_{p,n-1,\alpha/n}}{n(n-p-1+p)F_{p,n-1,\alpha/n}} \tag{2}$$

where,
p, is a number of variables;
n, is a number of samples;
F, is the *F* statistics value for *p* degrees of freedom in the numerator and *n-1,* degrees of freedom in the denominator under a significance level of α/n, $\alpha = 5\%$.

When the value found by the expression (1) is larger than the critical value by the expression (2), the sample is considered an outlier [10].

2.2 Cluster analysis

It is a method of graphical visualization, usually through the dendrogram, for outliers detection. There are two methods of cluster analysis: single linkage and Wards, together with the measurement of dissimilarity as Euclidean and Euclidean squared distance, applied to the variables transformed by base log 10 [1, 11].

These methods for cluster analysis already exist, implemented in several computational packages as: SAS, Minitab, SPSS, R, Statistica and another detection method, consisting of verifying the dendrogram samples, which are isolated in a single group, or with the measurement of dissimilarity distance.

In this work, the methods of single linkage and Euclidean distance were used, because the objective of this analysis is the detection of possible outlying samples.

2.3. Principal components analysis

The principal components analysis is a technique that transforms, linearly, one set of *p* variables, observed in a smaller set of k non-correlated variables, and that explain a substantial portion of the data covariance structure [7]. The *p* transformed variables (Y_1, Y_2,..., Y_p) calculated from the original variables are denominated principal components. The principal components are ordered so that the first component (Y_1) explains the largest portion of the variability, the second component (Y_2) explains the second largest portion, and so on.

In archaeometric studies of ceramics, the technique of principal components is extremely useful, because with the modern analytical techniques it is possible to determine a great number of variables, which are frequently correlated. The composition of each original species can be converted into the principal scores, becoming more easily interpreted. Several researchers highlight that, in archaeometric studies of archeological ceramics, about 70% or more of the total data variance is explained in terms of the first three principal components [12, 13].

In this study, the scores of the first two principal components were considered for the outliers determination.

2.4. Factor Analysis

The factor analysis has the purpose of describing the covariance structure among the original variables, in function of few random measures. In other words, it describes the dependence structure of a set of variables, through the creation of factors that, supposedly, measure common aspects.

An advantage of the factor analysis in relation to the technique of principal components is that the latter does not constitute a statistical technique, but a single base change in the space of the original variables. The factor analysis is a statistical method that seeks to explain the data covariance structure. The product of the rotational factors matrix for the data is denominated "factor scores" matrix, representing the contribution estimates of the several factors, to each original observation. They are used to group of samples.

In this work, the score dispersion diagram for the first and second score components was used, with the configuration that considers rotation for principal components and varimax rotation [14].

2.5. Standardized residual

It is known that the residual represents the amount that the regression equation does not explain. Possibly, it is due to the effect of omitted explanatory variables and to the natural variability among the samples. On the other hand, the standardized residual is the residual divided by the square root of the medium quadratic error, which guarantees, as an advantage, the comparison possibility [15].

3. RESULTS AND DISCUSSION

The methods previously presented were applied to the results for thirteen variables (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Tb e U), regarding 89 samples of ceramic fragments collected at an archeological site. In Table 1, the data of the elementary concentrations are shown. The details on the sample preparation and the analytical method were published in another work [16].

Table 1 shows the Mahalanobis distance (D) and the Wilks critical value, in the last row. In the first stage, the Mahalanobis distance for sample 6 was 34.6, which is larger than the critical value (31.6). It implicates that the sample 6 is an outlier. To proceed, the sample 6 is eliminated from the data and the Mahalanobis distance is calculated again. In this case, the value of critical D was 31.5 and the sample that has a larger D than the critic is eliminated. In the example, the sample 42 is an outlier. The procedure continues until no D, higher than the critical value, is found. The study showed that the samples 6, 11, 12, 13, 42, 44 and 61 are outliers.

In the case of the cluster method, the Euclidean distance and the method of single linkage were used, as shown in Figure 1. The sample outlier is that presenting the largest distance value and it corresponds to sample 48. In Figure 2, the samples 7, 10, 21, 28 and 48 are outliers, by the method of the standardized residual. For the method of principal component analysis, the dispersion diagram was made, as shown in Figure 3.

In Figure 3, the first principal component explained 41.6% of the total variance and the second principal component explained 17.5% of the total variance. The samples outliers are those out of the ellipse, limiting the 95% confidence region: samples 6, 11, 12, 13, 42 and 44 are outliers.

Finally, in the factor analysis, the rotation varimax, the extraction by principal components and the dispersion diagram that represents the scores of first and second factor were used. The results are presented in Figure 4. Again, the samples 6, 11, 12, 13, 42 and 44 are located out of the ellipse, corresponding to the outliers.

Using the Mahalanobis distance, 7 outliers were found (6, 11, 12, 13, 42, 44 and 61): they are the same samples that were found by the methods of principal components analysis and factor analysis. In Figures 3 and 4, using PCA and FA, it can be seen that the sample 61 is not an outlier, however, this sample is at the ellipse limit for the confidence level of 95%. The good result obtained by the Mahalanobis distance was due to the number of samples, which was higher than the critical value obtained by the expression (2). Then, the main limitation to use the Mahalanobis distance is the necessity that the number of samples, n, be three times larger than the number of variables, and, preferentially $n > 3\,p$, for the effect of the variance

Table 1. Results of the elementary concentrations in ppm, except when suitable, and values for the Mahalanobis distance.

sample	As	Ce	Cr	Eu	Fe(%)	Hf	La	Na(%)	Nd	Sc	Sm	Tb	U	D_1^2	D_2^2	D_3^2	D_4^2	D_5^2	D_6^2	D_7^2
1	1.5	108.3	134.2	2.5	3.2	7.8	64.1	2.0	63.0	12.9	8.9	1.1	1.3	10.8	10.8	10.8	10.7	10.6	10.5	10.3
2	4.4	133.5	148.0	2.8	4.6	8.4	86.6	2.5	66.0	14.9	11.7	1.2	1.9	9.7	9.6	9.5	9.4	9.4	9.5	9.2
3	1.4	110.6	156.0	2.3	3.4	9.6	71.4	2.9	60.0	14.5	9.8	0.9	2.0	9.8	9.8	10.4	10.7	10.6	10.6	10.4
4	4.6	124.4	141.7	2.7	5.6	7.2	79.6	2.0	82.0	15.4	9.7	1.2	1.2	16.5	16.5	16.3	16.1	16.0	15.9	16.2
5	2.4	98.8	128.0	2.0	4.0	7.5	56.3	1.1	52.0	13.1	7.8	0.9	1.0	13.8	14.5	14.7	14.6	14.4	14.8	15.0
6	**1.5**	**180.6**	**275.0**	**3.5**	**2.3**	**8.2**	**91.4**	**0.7**	**80.0**	**21.9**	**14.4**	**1.5**	**1.8**	**34.6**						
7	0.5	138.8	150.0	3.4	3.7	7.2	96.0	2.9	66.0	15.6	12.1	1.3	1.4	11.8	12.3	12.2	12.1	11.9	12.0	12.1
8	2.2	84.9	125.0	1.9	3.1	6.8	59.6	3.2	49.0	10.2	7.5	0.7	1.4	12.4	12.8	13.6	14.1	14.0	14.4	18.7
9	1.5	113.2	202.0	3.0	4.2	8.4	78.4	3.1	85.0	19.8	10.3	1.4	1.3	21.2	20.9	21.1	22.0	22.7	22.4	22.4
10	4.6	102.8	107.4	2.4	4.4	6.8	77.0	1.6	48.0	13.3	9.8	1.1	1.0	10.4	10.3	10.3	11.0	11.3	11.5	14.5
11	**2.4**	**100.1**	**83.0**	**1.8**	**3.5**	**11.1**	**47.6**	**0.3**	**44.0**	**20.5**	**7.1**	**0.8**	**0.9**	**25.1**	**24.9**	**24.7**	**25.6**	**33.4**		
12	**1.8**	**99.2**	**82.0**	**1.9**	**3.5**	**11.9**	**56.2**	**0.2**	**48.0**	**20.0**	**7.7**	**0.7**	**1.2**	**26.9**	**26.6**	**26.5**	**26.2**	**29.8**	**35.0**	
13	**2.4**	**96.5**	**98.0**	**1.8**	**2.6**	**9.8**	**43.6**	**0.3**	**37.0**	**23.3**	**6.3**	**0.9**	**1.3**	**28.6**	**28.3**	**28.5**	**33.8**			
14	3.6	99.9	139.0	2.2	3.5	9.3	66.8	2.5	47.0	12.8	8.9	0.9	1.3	6.0	6.4	6.5	6.5	6.5	6.5	6.4
15	2.7	122.3	133.0	2.6	3.9	6.3	83.4	1.5	64.0	15.2	10.1	1.1	1.0	11.8	11.7	11.6	11.5	11.4	11.2	11.1
16	2.5	133.6	182.0	2.3	3.3	9.8	70.7	1.7	57.0	18.2	9.8	1.0	1.6	7.7	8.0	7.9	7.9	7.8	7.7	8.2
17	1.8	102.6	118.0	2.2	3.5	5.7	72.9	2.3	46.0	12.6	9.2	0.9	1.2	12.2	12.0	12.6	13.9	14.0	13.9	14.1
18	2.0	111.9	138.0	2.3	3.8	8.4	62.7	2.3	49.0	12.6	8.4	0.9	0.9	10.9	11.5	11.5	11.5	11.7	11.7	11.6
19	2.0	107.4	122.0	2.6	3.4	5.4	76.0	2.1	63.0	13.6	10.1	1.0	1.1	13.1	13.2	13.7	14.7	14.7	14.5	15.9
20	1.1	137.5	172.6	2.6	2.6	9.6	73.7	1.6	80.0	17.2	10.0	1.2	1.3	11.0	11.0	10.9	11.7	11.7	11.6	11.5
21	0.7	105.7	184.0	2.5	3.0	8.9	73.9	1.5	60.0	19.7	9.8	1.0	1.2	14.5	15.1	15.2	15.3	15.2	15.1	15.8
22	1.6	99.2	148.0	2.3	3.5	7.8	67.6	2.0	52.0	12.6	8.9	1.2	1.2	12.3	12.3	12.3	12.2	12.1	12.3	15.3
23	2.2	127.1	143.0	2.4	3.4	8.0	71.4	2.1	62.0	14.8	9.3	0.9	1.2	4.0	4.4	4.5	4.5	4.5	4.5	4.8
24	2.2	116.4	130.0	2.2	2.8	9.0	70.1	1.7	60.0	12.2	9.7	1.1	1.5	10.2	10.1	10.0	10.6	11.1	11.0	10.8
25	1.2	125.6	150.0	2.7	3.4	9.3	83.4	1.6	51.0	17.2	11.3	1.2	1.3	9.1	9.1	9.0	8.8	8.7	8.6	11.7
26	2.0	142.0	166.0	3.1	4.1	8.3	86.4	2.4	72.0	16.9	11.6	1.5	1.4	4.5	4.4	4.4	4.4	4.4	4.3	4.4
27	2.4	132.5	148.0	3.1	3.8	8.1	80.9	3.0	64.0	15.3	11.7	1.4	1.7	11.3	11.2	11.1	11.3	11.3	11.2	10.9
28	4.8	138.8	222.0	2.5	2.5	8.7	64.7	1.5	57.0	19.9	9.0	0.9	1.7	8.1	9.2	9.5	10.1	9.9	9.8	10.1
29	2.2	110.9	155.0	2.7	4.5	7.9	75.7	1.9	69.0	14.7	10.3	1.2	1.3	5.3	5.9	5.9	5.9	5.9	6.0	6.3
30	2.4	143.5	147.1	3.8	3.2	7.7	100.2	1.8	102.0	16.4	13.5	1.7	1.4	21.1	20.8	20.7	21.4	21.3	21.2	20.7
31	0.9	137.9	195.0	2.2	2.1	8.1	54.0	1.9	39.0	18.3	7.3	1.0	1.6	24.2	24.3	24.2	24.0	27.6	27.7	27.1
32	4.4	120.1	159.0	2.4	4.2	9.0	70.2	2.5	58.0	15.2	10.3	1.5	1.2	13.6	13.7	13.5	13.9	13.7	13.6	13.7
33	2.1	123.9	141.8	2.6	3.9	8.3	73.0	2.4	66.0	14.8	9.7	1.3	1.2	5.8	6.7	6.7	6.6	6.7	6.8	7.1
34	3.9	123.8	175.0	2.7	4.4	9.1	72.5	2.3	63.0	16.8	10.2	1.3	1.3	5.1	5.2	5.3	5.2	5.1	5.1	5.1
35	1.8	151.5	150.0	2.6	2.9	7.6	79.3	2.2	59.0	14.6	11.1	1.3	1.1	10.4	10.2	10.2	10.3	10.4	10.3	12.0
36	2.4	158.8	215.0	2.6	2.1	9.1	72.5	1.4	55.0	18.2	10.5	1.4	2.0	12.2	14.0	15.9	17.7	17.5	17.8	20.4
37	3.6	111.0	156.0	2.3	3.5	8.4	64.9	2.1	48.0	14.7	8.6	0.8	1.3	4.2	4.2	4.2	4.1	4.1	4.0	3.9
38	1.9	118.9	185.0	2.8	2.6	8.8	66.8	1.9	64.0	20.3	10.0	1.1	1.0	15.2	15.3	15.3	15.7	15.9	16.4	16.4
39	2.2	100.0	145.0	2.1	3.7	9.9	58.2	1.9	54.0	13.2	8.1	0.9	1.2	8.5	8.5	8.4	8.3	8.5	8.5	8.5
40	2.4	116.5	154.0	2.0	3.5	8.2	61.0	1.5	49.0	13.6	7.5	0.7	1.8	14.0	13.8	13.9	14.9	14.8	14.8	14.5
41	2.5	160.3	183.0	3.8	3.9	7.6	96.8	2.6	68.0	18.0	13.1	1.5	1.2	13.0	12.9	13.4	13.7	13.7	13.7	15.5
42	**1.2**	**221.0**	**271.0**	**3.9**	**2.6**	**9.7**	**102.2**	**0.9**	**78.0**	**16.8**	**13.3**	**1.6**	**1.7**	**28.5**	**31.7**					
43	2.1	98.2	130.0	2.1	3.3	7.8	61.1	2.5	37.0	12.4	8.1	0.9	1.4	8.5	9.0	9.0	9.8	10.1	10.0	9.7
44	**1.7**	**227.0**	**275.0**	**3.8**	**2.3**	**9.8**	**116.6**	**1.2**	**93.0**	**16.2**	**13.8**	**1.3**	**1.6**	**30.6**	**31.1**	**42.9**				
45	2.1	126.5	157.0	2.8	3.5	8.5	77.3	1.5	62.0	17.7	10.8	1.1	0.8	10.4	10.5	10.4	10.4	10.4	10.5	11.0
46	2.2	141.7	160.0	3.2	4.6	8.3	95.8	1.3	80.0	16.7	12.3	1.3	1.1	8.3	8.8	9.7	12.9	14.2	14.3	14.3
47	2.2	97.8	130.0	2.0	3.9	7.3	67.2	2.8	41.0	12.0	8.7	0.9	1.1	11.4	11.3	11.6	12.2	12.1	12.0	12.8
48	0.2	114.5	110.0	2.1	2.9	7.2	66.1	1.0	50.0	12.5	9.1	1.1	1.3	9.9	10.1	10.1	10.1	10.1	10.3	11.7
49	0.9	133.0	188.0	2.1	2.0	7.2	57.8	1.7	49.0	16.9	7.9	1.1	1.2	16.1	15.9	15.8	15.6	16.9	16.7	18.6
50	1.5	114.0	145.0	2.1	2.6	8.0	58.4	1.6	57.0	15.8	7.9	0.8	1.0	8.5	8.8	9.2	9.2	9.3	9.9	11.8
51	1.6	148.7	142.0	2.3	2.7	7.8	70.3	1.9	57.0	17.0	10.5	1.2	1.1	16.8	16.7	16.8	17.2	17.2	18.4	19.6

Table 1. Continued

Sample	As	Ce	Cr	Eu	Fe(%)	Hf	La	Na(%)	Nd	Sc	Sm	Tb	U	D_1^2	D_2^2	D_3^2	D_4^2	D_5^2	D_6^2	D_7^2
52	1.6	113.8	180.0	2.0	3.0	10.0	54.9	1.6	47.0	15.3	7.0	0.7	1.4	11.1	11.4	11.5	12.4	12.2	12.1	12.2
53	1.8	142.7	168.0	2.5	2.7	8.2	78.7	1.3	53.0	16.7	10.6	1.1	1.3	5.8	6.4	6.8	7.8	7.8	7.9	7.8
54	3.3	123.4	151.0	2.6	4.1	7.8	66.8	1.7	54.0	16.3	9.0	0.9	1.0	7.8	7.7	7.8	7.7	7.6	8.2	9.0
55	2.7	115.2	145.0	2.5	3.2	7.4	70.0	2.2	61.0	13.9	9.4	0.8	1.5	6.9	6.8	6.8	6.7	6.8	6.8	7.3
56	1.2	137.2	144.0	2.6	2.8	8.4	72.6	1.7	59.0	15.0	10.1	1.4	1.1	8.1	8.1	8.2	8.3	8.3	8.3	8.2
57	1.5	104.6	135.0	2.1	2.5	9.2	60.7	1.0	46.0	14.9	8.2	0.7	1.3	5.9	5.9	5.8	6.1	6.0	6.2	11.9
58	4.5	148.2	173.0	2.4	4.5	9.0	68.0	2.4	66.0	16.6	9.4	1.0	1.2	14.8	14.7	14.5	14.5	14.5	14.5	14.4
59	2.1	146.3	242.0	3.1	3.8	10.2	84.7	4.1	81.0	20.4	12.1	1.2	1.3	16.1	15.9	15.9	15.7	16.6	16.9	20.8
60	1.0	127.4	171.0	2.1	1.9	6.5	57.9	0.8	54.0	17.2	8.1	0.7	1.2	18.5	21.6	21.6	22.2	21.9	21.7	22.7
61	**1.2**	**108.9**	**122.0**	**2.5**	**3.9**	**15.5**	**75.4**	**0.4**	**54.0**	**22.3**	**9.8**	**1.2**	**1.6**	**29.4**	**29.3**	**29.1**	**29.4**	**30.7**	**34.1**	
62	2.0	116.8	183.0	2.3	2.7	8.1	61.4	1.7	59.0	17.7	8.2	0.8	1.4	6.6	6.6	6.6	6.5	6.6	6.5	6.4
63	5.8	124.8	177.0	2.7	4.8	8.8	74.2	2.2	68.0	17.4	10.3	1.1	1.3	6.4	6.7	6.6	6.6	6.5	6.5	6.3
64	2.3	105.1	142.5	2.1	2.2	8.5	62.5	1.3	61.0	14.4	8.8	0.9	1.6	12.0	12.4	12.3	12.2	12.0	11.9	12.0
65	1.1	119.5	184.0	2.6	2.7	9.6	68.4	1.8	50.0	16.5	9.5	0.9	1.4	6.1	6.1	6.3	6.7	6.9	6.9	7.0
66	3.3	109.1	127.0	2.1	5.0	8.0	64.5	2.3	55.0	10.9	8.5	1.0	1.5	12.4	12.5	12.5	12.4	12.6	12.5	12.7
67	1.6	104.5	150.0	2.4	3.1	7.7	61.8	2.4	47.0	12.8	8.7	0.9	1.3	8.2	8.2	8.2	8.1	8.0	7.9	10.8
68	2.3	104.7	161.0	2.2	2.9	9.0	63.0	2.5	50.0	15.0	8.2	1.0	1.2	6.8	7.7	7.9	8.1	8.3	8.1	8.1
69	2.7	104.0	129.2	2.4	4.0	8.6	60.7	2.3	60.0	13.4	9.1	1.1	1.8	13.7	13.6	13.5	14.6	15.0	15.5	16.4
70	1.0	120.9	141.0	2.8	3.3	7.0	87.1	1.4	59.0	14.9	11.2	1.0	1.5	9.6	10.2	10.1	10.6	10.8	11.0	10.7
71	2.7	115.1	155.0	3.0	3.6	7.6	79.2	1.7	62.0	15.7	10.7	1.1	1.3	6.4	6.7	6.9	6.9	6.8	6.8	6.7
72	1.9	85.5	147.0	2.3	2.9	10.4	61.5	1.5	44.0	14.0	9.3	1.0	1.6	21.2	22.1	22.0	21.8	21.6	21.3	21.6
73	2.7	117.3	187.0	2.2	2.7	10.5	67.3	2.6	57.0	16.1	9.1	1.0	2.4	13.8	13.8	13.9	13.8	13.7	13.9	13.9
74	2.7	123.1	186.0	2.7	3.3	8.6	71.6	2.4	59.0	17.6	9.0	1.0	1.5	9.2	9.8	9.7	9.6	9.8	9.7	9.7
75	2.1	126.8	166.0	2.5	3.6	8.2	65.6	1.7	59.0	16.3	9.6	1.2	1.5	5.7	6.9	7.1	7.0	7.1	7.2	7.0
76	2.4	120.1	141.0	2.2	3.3	7.3	59.9	1.7	52.0	15.0	8.8	0.9	2.0	14.5	15.1	14.9	15.4	16.0	16.4	17.5
77	2.0	117.5	184.0	2.5	3.4	9.2	69.5	2.2	57.0	17.0	9.8	1.0	2.0	6.9	7.5	7.4	7.4	7.3	7.3	7.1
78	1.8	121.6	160.0	2.6	2.9	8.6	72.4	1.7	63.0	16.4	9.9	1.1	1.2	1.8	1.8	1.7	1.7	1.7	1.7	1.7
79	3.1	96.0	145.0	2.2	4.6	7.8	61.2	2.6	49.0	13.0	8.0	0.7	1.1	12.1	12.1	12.0	12.0	12.0	11.9	12.0
80	1.3	152.1	158.0	2.6	2.5	7.4	80.7	2.0	68.0	15.3	10.1	1.0	1.1	11.6	12.2	12.0	13.4	13.7	13.8	14.4
81	1.1	125.5	182.0	2.2	1.8	9.8	68.9	1.3	50.0	17.5	9.3	1.0	1.5	11.6	11.6	11.4	11.9	11.8	11.8	12.0
82	1.8	138.5	192.0	2.7	3.2	9.3	78.2	2.2	57.0	19.7	10.5	1.0	1.7	8.9	8.8	8.8	8.7	8.6	8.5	8.9
83	1.2	125.2	158.0	2.8	3.0	9.2	71.3	1.2	58.0	17.7	9.9	1.1	1.4	3.8	3.8	4.2	4.7	4.6	4.8	8.6
84	2.0	131.9	169.0	3.0	3.5	9.3	77.6	1.0	60.0	17.8	10.3	1.3	1.7	8.2	8.5	10.5	12.3	12.2	12.1	15.2
85	2.0	121.2	152.0	2.7	4.1	8.7	89.0	2.1	64.0	15.8	10.8	1.2	1.7	10.4	10.6	10.6	10.4	10.3	10.6	12.2
86	1.4	115.1	147.0	2.3	2.8	7.7	65.3	2.2	47.0	14.6	8.9	0.9	1.2	4.4	4.5	4.5	4.6	4.6	4.6	4.6
87	3.0	127.3	166.0	2.6	4.1	9.9	80.9	2.2	72.0	17.0	11.2	1.3	1.2	7.5	7.4	7.4	7.3	7.4	7.3	7.6
88	1.1	116.3	130.0	2.1	2.6	7.8	66.5	1.4	44.0	12.7	8.2	0.8	1.2	8.4	9.1	9.1	10.3	10.2	10.2	10.2
89	1.4	112.7	137.0	2.3	3.2	8.1	69.4	1.5	50.0	13.1	9.0	0.9	1.5	3.4	3.3	3.5	4.1	4.2	4.3	4.2
$D_{critical}$														31.6	31.5	31.4	31.4	31.3	31.2	31.0

covariance sampling matrix. On the other hand, when the transformation to base log 10 isused, to normalize the data, this may, also, produce outliers, when working with results next to zero; but, obviously, to work with null values cannot be done.

For the method of cluster analysis, a single sample, sample 48, is the one that presented the largest distance among the samples in the group, being considered, therefore, an outlier (Figure 1). The method of cluster analysis did not show to be efficient to determine outliers, because sample 48 is not an outlier, in accordance with other methods, such as Mahalanobis distance, principal component analysis and factor analysis.

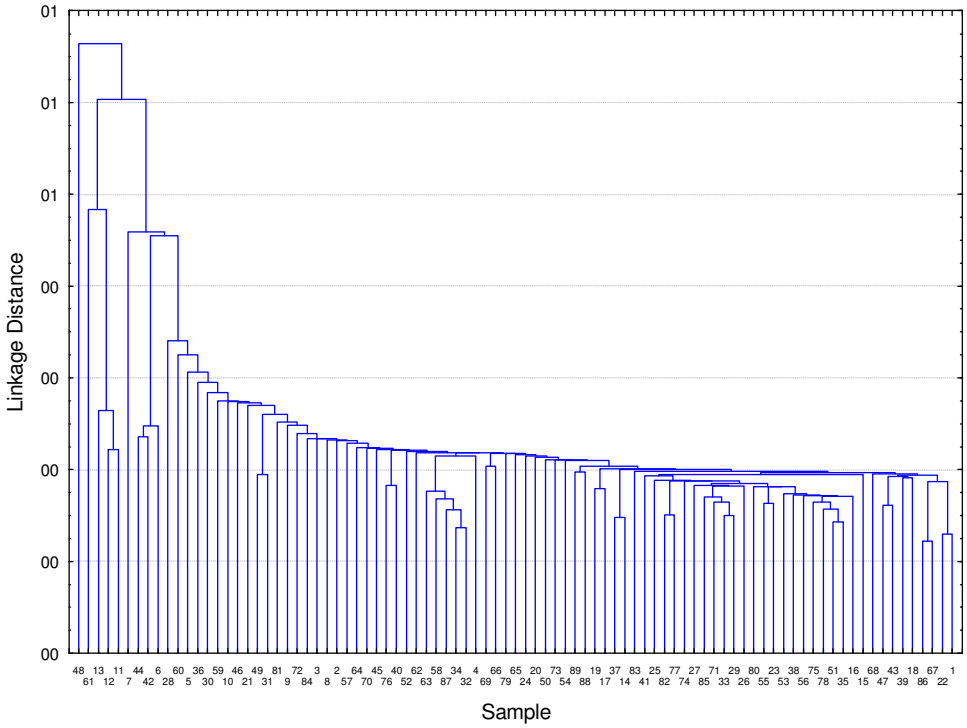

Figure 1. The cluster analysis dendrogram, by the Single Linkage method, for the data regarding 89 samples, from one archaeological site.

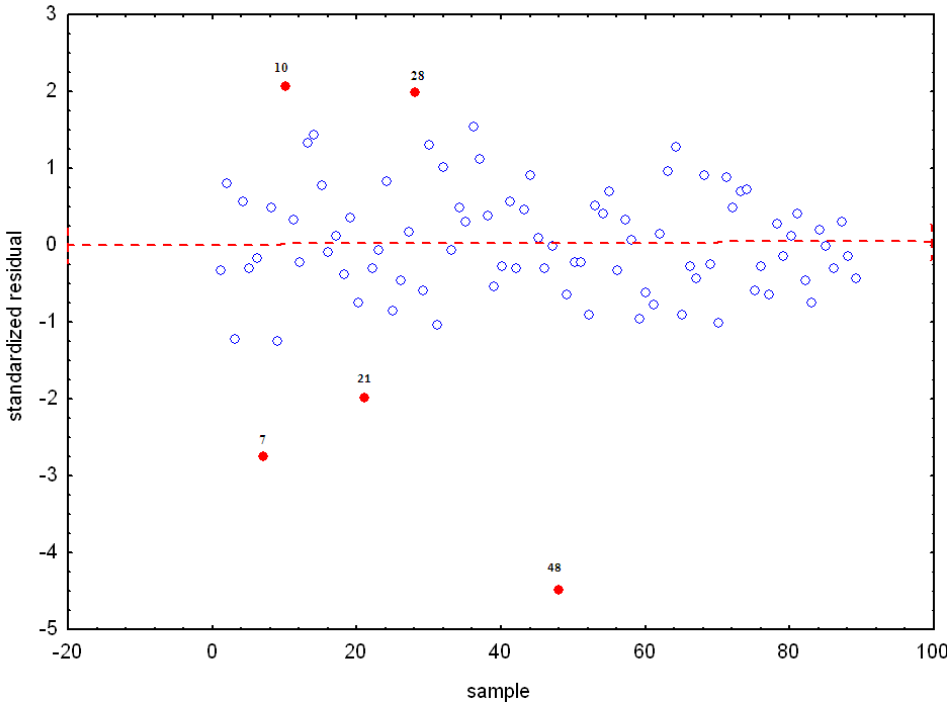

Figure 2. Identification of the samples versus standardized residual.

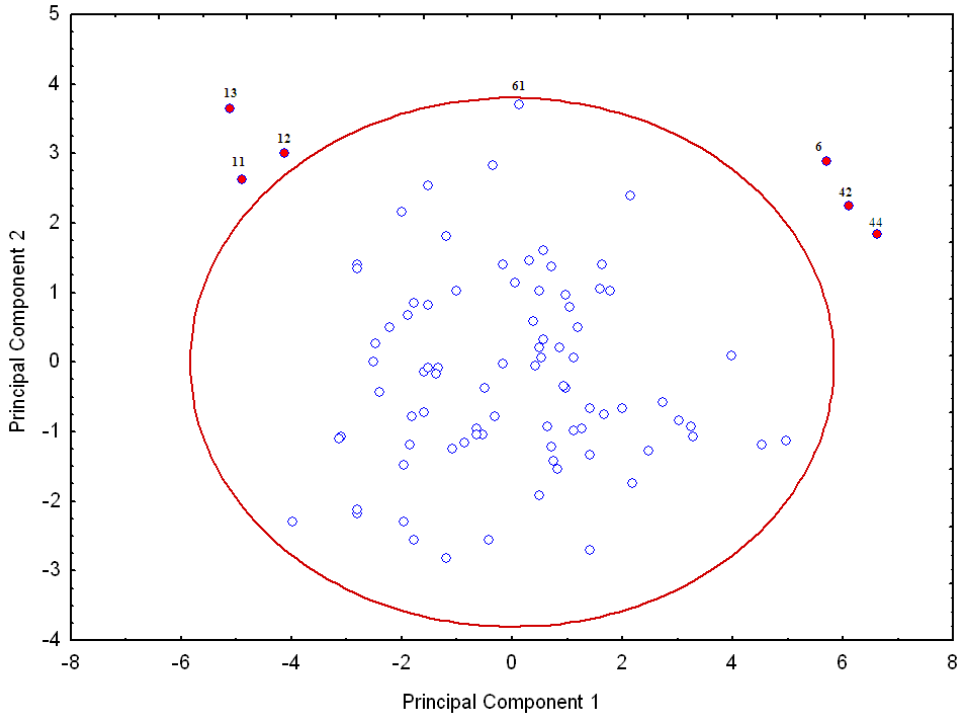

Figure 3. Dispersion diagram for the scores of the first principal component, versus the score of the second principal component. The ellipse represents the confidence level of 95%.

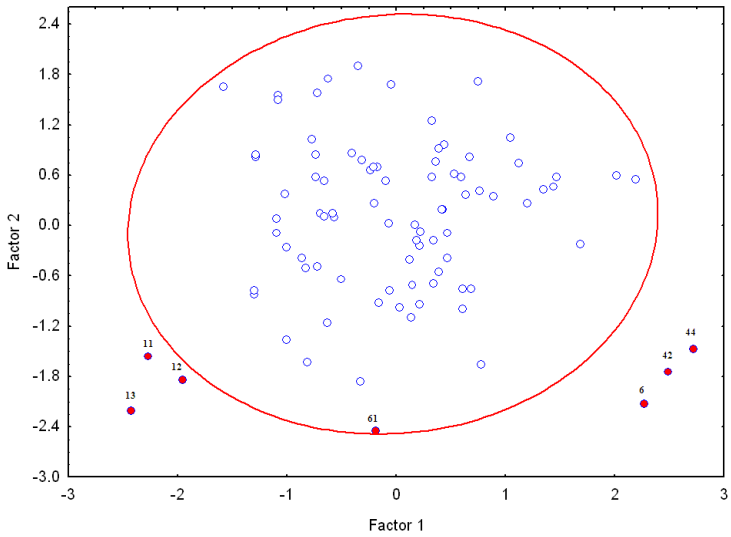

Figure 4. Dispersion diagram for the first and second factor scores. The ellipse represents a confidence level of 95%.

For the method of the standardized residual, the samples 7, 10, 21, 28 and 48 were considered outliers because they are those with the largest residues. The outliers found by this procedure were different from those found by the other methods (Mahalanobis distance, PCA, FA), except for the sample 48, which was, also, considered an outlier by the cluster analysis. The

different samples found as outliers, by the standardized residual, it was due to the fact that residue takes into account the part not explained by the adjustment of the multiple regression, which considers the first variable as dependent, and the others as independent ones.

4. CONCLUSION

The outliers detection in a data base is a technical problem that depends on the scientific work and on the questions wished to be answered. However, researchers, usually, do not take into consideration the identification and elimination of the outliers at the end of the analysis. Among the studied statistical methods (Mahalanobis distance, cluster analysis, principal component, factor analysis, standardized residual) to determine outliers in a data base, the results showed that the Mahalanobis distance, using the lambda Wilks criterion to determine the critical value, is the method that showed to be the most convenient and accurate. The other two methods (PCA and FA), also, showed to be convenient to identify outlying values in a data base. On the other hand, this study showed that the cluster analysis and the standardized residual methods are not appropriate to identify outliers, in the present case.

ACKNOWLEDGMENTS

Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, Process number 2008/54867-7, for financial support.

REFERENCES

1. E. V. Sayre, "Brookhaven Procedures for Statistical Analyses of Multivariate Archaeometric Data," *Brookhaven National Laboratory Report BNL*, New York, 21693 (1975).
2. W. J. Egan; S. L. Morgan, "Detection in Multivariate Analytical Chemical Data" *Ana1. Chem.* **70,** pp. 2372—2379 (1998)
3. J. Papageorgiou; M.J. Baxter, "Model-based cluster analysis of artefact compositional data," *Archaeometry,* **43(4)**, pp. 571—588 (2001).
4. I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, USA (2002).
5. P.T.M.S**.** Oliveira; C. S. Munita; A. Nascimento; S. Luna; R. P. Paiva; M. A. Alves; E. F. Momose, "Aplicação de métodos estatísticos multivariados em estudos arqueométricos," *VIII Escola de modelos de regressão,* Conservatória, RJ, 23 a 26 de fevereiro, (2003).
6. J. Bacon-Shone; W.K. Fung, "A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data," *Applied Statistics*, *Royal Statistical Society,* **35**, pp.153-162 (1987).
7. P. T. M. S. Oliveira; C.S. Munita, "Influência do valor crítico na detecção de valores discrepantes em arqueometria", *48ª Região Brasileira da Sociedade Internacional de Biometria – RBRAS, 9º Simpósio de Estatística Aplicada à Experimentação Agronômica – SEAGRO*, Lavras, MG, 7 a 11 de julho, (2003).
8. S. S. Wilks, Multivariate statistical outliers. *Sankhya,* **25**, pp, 407–426(1963).
9. Ali S. Hadi, "Identifying Multiple Outliers in Multivariate Data. " *J. R. Statisc. Soc. B*, **54(3),** pp. 761—771 (1992).
10. V. Barnett; T. Lewis, *Outliers in Statistical Data*, Wiley & Sons, New York , USA (1994).

11. M. J. Baxter, "Detecting multivariate outliers in artefact compositional data, *Archaeometry*," **41,** pp.321-338 (1999).
12. N. R. Draper; H. Smith, *Applied Regression Analysis*. Wiley & Sons, New York, USA (1998).
13. R. A. Jonhson; D. W. Wichern, *Applied Multivariate Analysis*, Prentice Hall, New Jersey, USA (1998).
14. L. P. Barroso; R. Artes, "Análise multivariada,", *48ª Região Brasileira da Sociedade Internacional de Biometria – RBRAS, 9º Simpósio de Estatística Aplicada à Experimentação Agronômica – SEAGRO*, Lavras, MG, 7 a 11 de julho, (2003).
15. F. R. S. Giroldo, *Alguns Métodos Robustos Para Detectar Outliers Multivariados*, Dissertação, IME-USP (2008).
16. C. S. Munita; M. A. Alves; R. P. Paiva; P. M. S. Oliveira; E. F. Momose, "Contribution of Neutron Activation Analysis to Archaeological Studies, "*J. Trace and Microprobe Techniques*, **18(3)**, pp. 381-387 (2000).