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**

COMPARATIVE STUDY AMONG OUTLYING DETECTION METHODS AND TWO TRANSFORMATIONS IN EXPERIMENTAL DATA

Paulo M. Oliveira, Roberto H. Marks

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

Instituto de Pesquisas Energéticas e Nucleares, IPEN - CNEN/SP
Av. Professor Lineu Prestes, 2242
05508-000 São Paulo, SP
hazenfratz@gmail.com

ABSTRACT

Identifying outliers is an important step prior to multivariate analyzes which are sensitive to them. This paper presents the comparative study between the log base-10 and compositional transformation for the determination of outliers for one data set with 89 samples of ceramic fragments, analyzed by neutron activation analysis.. Five procedures were employed: Mahalanobis distance, cluster analysis, principal component analysis, factor analysis and standardized residuals. The results showed that although cluster analysis is one of the procedures widely employed for outliers identification, it can fail by not identifying outliers that are identified by other methods

1. INTRODUCTION

The interest in the detection of atypical data points by the statistical community has increased since the middle of the last century. Some of the methods developed were Mahalanobis distance [1], mask [2], ellipsoid minimum volume [3] and decisive of the covariance matrix [3]. In general, the authors have concluded that it is not possible to determine, with precision, the outliers in a data set [1, 2, 3] for log base-10 and that, for compositional data, it is necessary to use adequate distributions [4].

Some sources of outlying results are: uncontrolled processes, inadequate analytical technique, contamination during sample preparation, error in measurements, transcription mistakes and others. In general, identification of outliers is rather subjective, although different statistical methods exist [5].

As to log base-10 or compositional transformation, 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 in estimates. Therefore, their detection is very important and should be done before data analysis [6, 7, 8]. A comparative study between different methods of detection, for this purpose, is necessary in the experimental results.

The statistical analysis of compositional data is based on determining the appropriate transformation from the simplex to real space. Possible transformations and outliers strongly

interact: parameters of transformations may be influenced particularly by outliers and the result of good-on-fit tests will reflect their presence [8].

In this work, a comparative study of outlier identification was performed by 5 methods: Mahalanobis distance, cluster analysis (average linkage with Euclidean distance), principal component analysis, factor analysis and standardized residuals. The data set comprised the elemental concentration vectors of 13 elements (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Tb and U) for 89 samples measured by instrumental neutron activation analysis.

With this purpose, elementary concentrations transformed by log base 10, normalized and compositional, standardized by the median [9], also serve to compensate the magnitude differences of the elements that are in percentage and at trace level [10]. The elemental concentrations were transformed to base-10 logarithms and to compositional data standardized by the median [9]. One of the main objectives is to compensate the magnitude differences of macro and trace elements [10].

2. DEVELOPMENT

2.1. Motivation

The elemental concentrations were measured by instrumental neutron activation analysis (INAA). This technique is regarded as very sensitive and appropriate for qualitative and quantitative analysis of macro and trace elements [11]. The samples were irradiated in the IEA-R1 research nuclear reactor at the Research Reactor Center of the Nuclear and Energy Research Institute (IPEN).

In this work, this study was carried out using 5 methods along with log base-10 and compositional for transformations of concentration data: Mahalanobis distance, cluster analysis (Euclidean distance and average linkage), principal components, factorial analysis and standardized residual.

2.2. Mahalanobis Distance

The Mahalanobis distance is an important measure of dissimilarity in statistics, introduced by Mahalanobis [12, 13]. It is suggested by many authors as an appropriate metric to detect outliers in multivariate data. The Mahalanobis distance (D_i) from each sample to the centroid is calculated by the expression:

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

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

where, $S = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)' \left(x_i - \bar{x}\right)$ is the sample covariance matrix. In this work, the Mahalanobis distance of each sample is compared to the critical value calculated by the Wilks lambda criterion [8, 10], defined by:

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

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, α = 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 [14].

2.3 Cluster analysis

It is a graphical visualization method, where one can identify outliers by the inspection of dendrograms. It was employed two methods of cluster analysis in this work: single linkage and Ward, with the Euclidean and squared Euclidean distance as the dissimilarity measure [1, 15].

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.

2.4. Principal components analysis

The PCA was first described by Karl Pearson [16]. He believed it was the right solution for some of the problems of interest for biometrics at the time. One practical description of computational methods came late by Hoteling [17]. The basic idea is that the latent variables represent a linear combination of the original variables which can be correlated [18].

The PCA technique performs a linear transformation on a set of *p* variables to yield a smaller set of *k* non-correlated new variables, which explain a large portion of the data covariance structure [10]. The p transformed variables (Y_1, Y_2, ... , Y_p) calculated from the original set are called principal components. They are ordered so that the first component Y_1 explains the largest portion of the data variability, the second component Y_2 explains the second largest portion, and so on.

PCA is useful in archaeometric studies because it makes possible to represent the original high dimensional system with a reduced number of principal components. This dimension reduction is important in the case of elemental concentrations determined by INAA, where one may have more than 20 variables in the data set. Many researchers affirm that it is usual to have more than 70% of the total system variance explained with three principal components [19, 20].

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

2.5. Factor Analysis

The factor analysis´ basic idea is to describe one set of p variables $X_1, X_2, \ldots, X_p$ in terms of smaller number of indices or factors, and in the process gain a better understanding of the relationship of these variables. The knowledge of the factor analysis is the result of the work of Charles Spelmann. While studying many correlations observed could be contained in a simple model [21]

The factor analysis is used to describe the covariance structure among the original variables as a function of few random quantities. In other words, it describes the dependence structure of a set of variables by the creation of factors which are supposed to measure common aspects.

One advantage of this technique when compared to PCA is that the latter is not a statistical technique, but a base change in the space of the original variables. The factor analysis, on the other hand, is a statistical method with an explicit purpose of explaining the covariance structure. The matrix product of data rotational factors is called *factor score matrix*, which represents the contribution of several factors to each original observation. It can be used to group samples.

In this work, the score dispersion diagram for the first and second score components was used in a configuration that considers the principal components and varimax rotation [22].

2.6 Standardized residual

The residuals represent the amount of information not explained by the regression equation, possibly due to the effect of omitted explaining variables and the natural variability in the data. On the other hand, the standardized residual is the residual divided by the square root of the quadratic error mean, and has the advantage of comparison possibility [23].

2.7 Compositional Data Analysis

Compositional data have particular and important numerical properties that have major consequences for any statistical analysis. These have been elucidated and discussed by a number of authors since Karl Pearson from 1897 [9].

We shall call an n x p data matrix fully-compositional if the rows sum to a constant, and sub compositional if the variables are a subset of a fully-compositional data set. Such data occur widely in archaeometry, where it is common to determine the chemical composition of ceramic, glass, metal or other materials using techniques such as neutron activation analysis and X-ray fluorescence analysis (XRF), among others. Interest often centers on whether there are distinct chemical groups within the data or should, for example, these groups be associated with different origins or manufacturing technologies [11].

The sample space of compositional data is, thus, simplex space. It is a $D - 1$ dimensional subset R^D. Standard statistical methods can lead to misleading results if they are directly applied to original closed data. For this reason, centred logratio (clr) was introduced. The clr

transformation is a transformation from $\boldsymbol{S^D}$ to $\boldsymbol{R^D}$, and the results for an observation $\boldsymbol{x \in R^D}$ are the transformed data $\boldsymbol{y \in R^D}$ with

$$y = (y_1, \ldots, y_D)' = \left(\log \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \cdots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right) \qquad (3)$$

Compositional data are those which contain relative information. They are parts of some whole information. In most cases, they are record as closed data, i.e., data summing to a constant, such as 100%, archaeological data being classic examples. Compositional data have important and particular properties that prelude the application of standard statistical techniques on such data in raw form [9]**.**

3. RESULTS AND DISCUSSION

The methods presented in this work were applied to thirteen elemental concentrations of the elements As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Tb e U in 89 ceramic fragments samples excavated at an archaeological site. The data used is presented in [24], Table 1. The analytical procedure for the measurements was already published in [25].

Table 1 shows the Mahalanobis distance (*D*) and in the last row the lambda Wilks critical values for $\log_{10}$ and compositional. For $\log_{10}$, 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 sample 6 is an outlier. So the sample 6 was eliminated from the data and the Mahalanobis distance was calculated again. In this case, the value of critical *D* was 31.5 and the sample that has a larger *D* than the critical value was eliminated. In the example, sample 42 is an outlier. The procedure continues until *D* found is lower than the critical value. The study showed that the samples 6, 11, 12, 13, 42, 44 and 61 are outliers. The same procedure for compositional data was used and the samples 5, 6, 11. 12, 13, 42, 44, 48 and 61 were found outliers.

Following, the data ($\log_{10}$ and compositional) were studied by means of cluster analysis method using single linkage and Euclidean distance. The dendrogram is showed in Figures 1a and 1b. As can be seen in the figures for $\log_{10}$ transformation the sample 48 is an outlier and for compositional data the sample 5.

After that, for the method of principal component analysis, the dispersion diagram was made as shown in Figure 2, for $\log_{10}$ and compositional data. The first principal component explained 41.6% of the variance and the second principal component explained 17.5% of the variance. The ellipses represent a confidence level of 95%. The outliers samples were 6, 11, 12, 13, 42 and 44 when was used $\log_{10}$ and 6, 13, and 48 for compositional data.

Figure 2, using PCA, it can be seen that sample 61 is not an outlier; however, this sample is at the ellipse limit for the confidence level of 95%.

Table 1. Results for Mahalanobis Distance.

log base 10								Compositional data								
Sample	D_1^2	D_2^2	D_3^2	D_4^2	D_5^2	D_6^2	D_7^2	sample	D_1^2	D_2^2	D_3^2	D_4^2	D_5^2	D_6^2	D_7^2	D_8^2
1	10.8	10.8	10.8	10.7	10.6	10.5	10.3	1	7.9	7.9	8.1	8.1	8.0	7.9	7.8	7.7
2	9.7	9.6	9.5	9.4	9.4	9.5	9.2	2	7.4	7.4	7.2	7.2	7.1	7.0	6.9	6.9
3	9.8	9.8	10.4	10.7	10.6	10.6	10.4	3	10.3	10.2	11.6	12.4	12.5	12.4	12.2	12.0
4	16.5	16.5	16.3	16.1	16	15.9	16.2	4	17.5	17.3	16.9	16.8	16.8	16.6	16.7	17.1
5	13.8	14.5	14.7	14.6	14.4	14.8	15	**5**	**82.3**							
6	**34.6**							**6**	**33.4**							
7	11.8	12.3	12.2	12.1	11.9	12	12.1	7	19.0	19.4	22.2	22.1	21.9	21.9	21.9	21.8
8	12.4	12.8	13.6	14.1	14	14.4	18.7	8	8.4	9.5	10.3	10.3	10.5	11.1	12.0	14.2
9	21.2	20.9	21.1	22	22.7	22.4	22.4	9	15.2	21.8	21.8	23.2	23.3	23.0	23.3	23.0
10	10.4	10.3	10.3	11	11.3	11.5	14.5	10	14.2	14.3	15.1	15.0	15.4	15.6	16.2	19.0
11	**25.1**	**24.9**	**24.7**	**25.6**	**33.4**			**11**	**23.0**	**24.2**	**24.1**	**25.6**	**32.9**			
12	**26.9**	**26.6**	**26.5**	**26.2**	**29.8**	**35**		**12**	**27.6**	**27.1**	**26.8**	**26.5**	**29.9**	**44.1**		
13	**28.6**	**28.3**	**28.5**	**33.8**				**13**	**30.1**	**29.5**	**29.6**	**32.4**				
14	6	6.4	6.5	6.5	6.5	6.5	6.4	14	7.8	8.5	9.7	9.6	9.5	9.3	9.3	9.1
15	11.8	11.7	11.6	11.5	11.4	11.2	11.1	15	11.6	11.5	11.3	11.4	11.3	11.1	11.0	11.2
16	7.7	8	7.9	7.9	7.8	7.7	8.2	16	5.0	5.3	5.3	5.2	5.1	5.1	5.4	5.6
17	12.2	12	12.6	13.9	14	13.9	14.1	17	11.0	10.8	11.3	12.2	12.1	11.9	11.8	11.7
18	10.9	11.5	11.5	11.5	11.7	11.7	11.6	18	9.3	11.3	11.2	11.3	11.6	11.5	11.4	11.3
19	13.1	13.2	13.7	14.7	14.7	14.5	15.9	19	9.7	9.9	10.0	10.4	10.3	10.1	10.0	10.5
20	11	11	10.9	11.7	11.7	11.6	11.5	20	11.3	11.1	11.3	11.9	12.0	11.8	11.7	11.8
21	14.5	15.1	15.2	15.3	15.2	15.1	15.8	21	10.7	18.8	19.5	19.8	19.6	19.4	19.1	20.4
22	12.3	12.3	12.3	12.2	12.1	12.3	15.3	22	6.6	8.6	9.2	9.2	9.3	10.0	11.9	12.0
23	4	4.4	4.5	4.5	4.5	4.5	4.8	23	3.0	4.7	4.8	4.8	4.7	4.8	5.1	5.0
24	10.2	10.1	10	10.6	11.1	11	10.8	24	11.0	10.9	10.7	11.5	11.9	11.9	11.7	11.6
25	9.1	9.1	9	8.8	8.7	8.6	11.7	25	7.4	7.5	7.9	7.9	7.8	7.8	7.9	10.5
26	4.5	4.4	4.4	4.4	4.4	4.3	4.4	26	3.3	3.2	3.4	3.4	3.4	3.3	3.3	3.3
27	11.3	11.2	11.1	11.3	11.3	11.2	10.9	27	10.2	10.3	10.1	10.5	10.7	10.7	10.6	10.6
28	8.1	9.2	9.5	10.1	9.9	9.8	10.1	28	12.9	13.7	16.5	17.3	17.7	17.4	17.2	18.0
29	5.3	5.9	5.9	5.9	5.9	6	6.3	29	5.0	7.3	8.0	7.9	8.1	8.2	8.3	8.3
30	21.1	20.8	20.7	21.4	21.3	21.2	20.7	30	23.9	23.8	23.8	24.9	24.6	24.3	24.0	23.7
31	24.2	24.3	24.2	24	27.6	27.7	27.1	31	21.4	24.8	24.6	24.4	27.2	27.2	26.9	26.8
32	13.6	13.7	13.5	13.9	13.7	13.6	13.7	32	14.5	14.6	14.3	14.4	14.3	14.3	14.3	14.1
33	5.8	6.7	6.7	6.6	6.7	6.8	7.1	33	5.1	6.1	6.0	6.0	6.0	6.2	6.1	6.6
34	5.1	5.2	5.3	5.2	5.1	5.1	5.1	34	4.8	5.1	5.2	5.1	5.1	5.1	5.1	5.0
35	10.4	10.2	10.2	10.3	10.4	10.3	12	35	6.6	9.6	9.7	9.9	10.0	9.9	10.0	11.4
36	12.2	14	15.9	17.7	17.5	17.8	20.4	36	14.7	16.3	17.6	19.7	19.6	19.8	20.8	22.1
37	4.2	4.2	4.2	4.1	4.1	4	3.9	37	5.0	4.9	5.9	5.8	5.8	5.7	5.6	5.6
38	15.2	15.3	15.3	15.7	15.9	16.4	16.4	38	12.9	14.1	14.8	15.0	15.2	15.7	15.6	15.5
39	8.5	8.5	8.4	8.3	8.5	8.5	8.5	39	8.5	8.4	8.4	8.3	8.6	8.5	8.4	8.8
40	14	13.8	13.9	14.9	14.8	14.8	14.5	40	12.4	13.0	13.1	13.9	13.7	13.7	13.6	13.5
41	13	12.9	13.4	13.7	13.7	13.7	15.5	41	12.0	11.9	12.4	12.6	12.4	12.3	13.0	15.3
42	**28.5**	**31.7**						**42**	**27.5**	**31.9**						
43	8.5	9	9	9.8	10.1	10	9.7	43	8.8	9.3	9.2	9.6	9.8	9.7	9.6	9.6
44	**30.6**	**31.1**	**42.9**					**44**	**29.8**	**30.8**	**41.1**					

Table 1. Continued

sample	D_1^2	D_2^2	D_3^2	D_4^2	D_5^2	D_6^2	D_7^2	sample	D_1^2	D_2^2	D_3^2	D_4^2	D_5^2	D_6^2	D_7^2	D_8^2
45	10.4	10.5	10.4	10.4	10.4	10.5	11	45	10.6	10.8	10.6	10.6	10.7	10.8	10.9	11.1
46	8.3	8.8	9.7	12.9	14.2	14.3	14.3	46	8.1	9.0	10.2	12.6	13.6	13.7	13.5	13.4
47	11.4	11.3	11.6	12.2	12.1	12	12.8	47	10.9	10.8	10.9	11.5	11.5	11.4	11.7	11.8
48	9.9	10.1	10.1	10.1	10.1	10.3	11.7	**48**	**31.3**	**32.5**						
49	16.1	15.9	15.8	15.6	16.9	16.7	18.6	49	13.2	13.6	14.1	13.9	14.1	13.9	14.5	14.9
50	8.5	8.8	9.2	9.2	9.3	9.9	11.8	50	6.0	6.6	6.8	6.7	6.6	7.2	8.3	9.2
51	16.8	16.7	16.8	17.2	17.2	18.4	19.6	51	13.0	17.2	18.1	18.8	18.6	20.2	21.3	21.3
52	11.1	11.4	11.5	12.4	12.2	12.1	12.2	52	12.2	12.8	12.8	13.5	13.4	13.3	13.1	13.3
53	5.8	6.4	6.8	7.8	7.8	7.9	7.8	53	6.2	7.0	7.3	8.4	8.4	8.4	8.3	8.3
54	7.8	7.7	7.8	7.7	7.6	8.2	9	54	8.0	8.5	8.7	8.7	8.7	9.2	9.8	10.0
55	6.9	6.8	6.8	6.7	6.8	6.8	7.3	55	8.1	8.0	8.2	8.1	8.1	8.0	8.0	8.8
56	8.1	8.1	8.2	8.3	8.3	8.3	8.2	56	7.0	8.3	9.0	9.2	9.1	9.1	8.9	9.2
57	5.9	5.9	5.8	6.1	6	6.2	11.9	57	6.7	6.5	6.4	6.9	6.8	6.8	8.7	12.9
58	14.8	14.7	14.5	14.5	14.5	14.5	14.4	58	8.2	12.2	11.9	11.7	11.6	12.0	12.3	12.4
59	16.1	15.9	15.9	15.7	16.6	16.9	20.8	59	10.2	10.6	10.6	10.8	11.4	11.3	12.5	17.9
60	18.5	21.6	21.6	22.2	21.9	21.7	22.7	60	14.4	18.0	18.8	19.9	19.9	19.6	19.7	19.7
61	**29.4**	**29.3**	**29.1**	**29.4**	**30.7**	**34.1**		**61**	**20.9**	**24.1**	**24.3**	**24.0**	**26.3**	**30.4**	**50.7**	
62	6.6	6.6	6.6	6.5	6.6	6.5	6.4	62	6.4	6.4	6.6	6.5	6.6	6.5	6.5	6.4
63	6.4	6.7	6.6	6.6	6.5	6.5	6.3	63	6.3	7.0	7.4	7.3	7.2	7.2	7.2	7.2
64	12	12.4	12.3	12.2	12	11.9	12	64	11.0	11.8	11.9	12.0	11.9	11.7	11.8	12.0
65	6.1	6.1	6.3	6.7	6.9	6.9	7	65	7.8	7.8	8.0	8.2	8.3	8.4	8.5	8.5
66	12.4	12.5	12.5	12.4	12.6	12.5	12.7	66	12.9	13.6	14.6	14.5	14.8	14.7	14.5	14.5
67	8.2	8.2	8.2	8.1	8	7.9	10.8	67	6.6	6.4	6.5	6.4	6.5	6.6	7.2	8.8
68	6.8	7.7	7.9	8.1	8.3	8.1	8.1	68	7.1	8.8	10.0	9.9	10.1	10.0	10.1	10.1
69	13.7	13.6	13.5	14.6	15	15.5	16.4	69	14.0	13.8	13.7	14.7	14.8	15.3	15.9	16.6
70	9.6	10.2	10.1	10.6	10.8	11	10.7	70	9.8	10.8	12.0	12.1	12.4	12.9	12.7	12.5
71	6.4	6.7	6.9	6.9	6.8	6.8	6.7	71	5.0	6.7	7.2	7.4	7.3	7.3	7.2	7.2
72	21.2	22.1	22	21.8	21.6	21.3	21.6	72	13.3	20.9	20.5	20.3	20.1	19.9	19.6	20.5
73	13.8	13.8	13.9	13.8	13.7	13.9	13.9	73	12.8	13.2	13.7	13.6	13.6	13.6	13.6	13.4
74	9.2	9.8	9.7	9.6	9.8	9.7	9.7	74	8.0	8.8	10.1	10.0	10.7	10.6	10.5	10.4
75	5.7	6.9	7.1	7	7.1	7.2	7	75	5.9	7.0	8.3	8.2	8.2	8.2	8.1	8.0
76	14.5	15.1	14.9	15.4	16	16.4	17.5	76	13.0	15.2	15.5	16.2	16.6	17.0	18.2	18.1
77	6.9	7.5	7.4	7.4	7.3	7.3	7.1	77	4.5	5.8	5.8	6.0	6.0	5.9	5.9	5.8
78	1.8	1.8	1.7	1.7	1.7	1.7	1.7	78	1.7	1.8	1.8	1.9	1.9	1.9	1.8	1.9
79	12.1	12.1	12	12	12	11.9	12	79	10.8	10.7	10.6	10.6	10.7	10.6	10.4	10.7
80	11.6	12.2	12	13.4	13.7	13.8	14.4	80	8.7	12.0	11.8	13.1	13.4	13.3	13.3	14.3
81	11.6	11.6	11.4	11.9	11.8	11.8	12	81	12.0	12.2	11.9	12.5	12.4	12.4	12.3	12.5
82	8.9	8.8	8.8	8.7	8.6	8.5	8.9	82	4.3	4.2	4.3	4.5	4.7	4.6	4.7	4.7
83	3.8	3.8	4.2	4.7	4.6	4.8	8.6	83	4.4	4.4	5.0	5.2	5.2	5.3	6.0	9.1
84	8.2	8.5	10.5	12.3	12.2	12.1	15.2	84	8.6	9.2	11.3	13.3	13.4	13.3	13.2	16.1
85	10.4	10.6	10.6	10.4	10.3	10.6	12.2	85	7.3	8.6	8.5	8.4	8.4	8.4	8.3	9.3
86	4.4	4.5	4.5	4.6	4.6	4.6	4.6	86	3.1	3.5	3.5	3.6	3.5	3.5	3.5	3.5
87	7.5	7.4	7.4	7.3	7.4	7.3	7.6	87	6.8	7.4	7.3	7.2	7.3	7.2	7.1	7.2
88	8.4	9.1	9.1	10.3	10.2	10.2	10.2	88	5.8	7.7	7.8	9.1	9.2	9.2	9.1	9.4
89	3.4	3.3	3.5	4.1	4.2	4.3	4.2	89	3.1	3.1	3.8	4.4	4.6	4.7	4.7	4.8
$D_{critical}$	31.6	31.5	31.4	31.4	31.3	31.2	31	$D_{critical}$	31.6	31.4	31.3	31.2	31.1	31.0	31.0	30.9

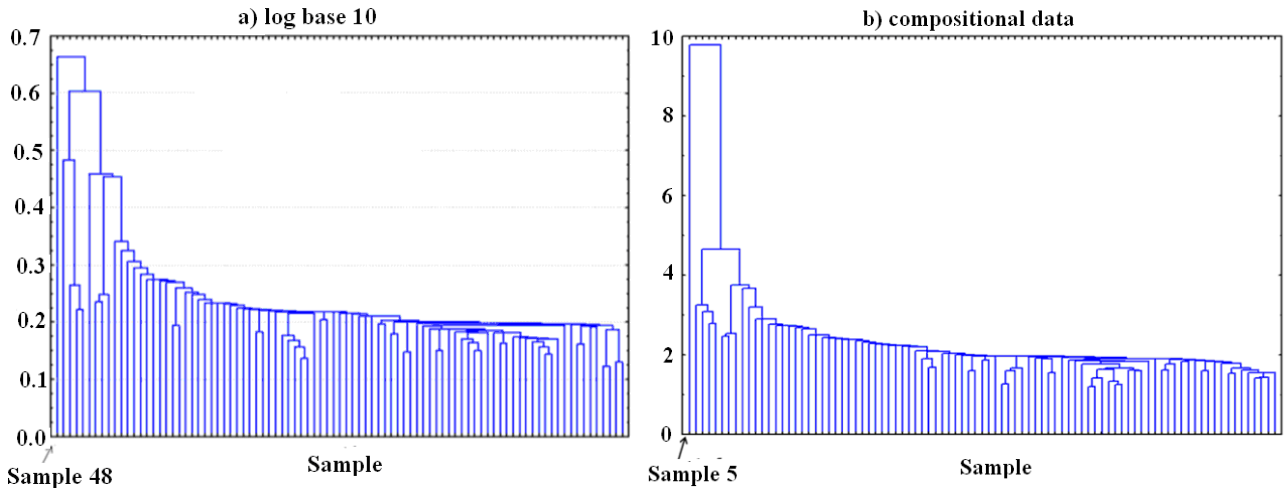

Figure 1. Cluster analysis dendrogram, by single linkage method for the data regarding 89 samples: a) log base 10 and b) compositional.

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 Figures 3a and 3b, respectively. In the Figure 3a six samples are outliers: 6, 11, 12, 13, 42 and 44 and in Figure 3b 11, 12, 13, 48 and 61.

For the method of the standardized residual, Figure 4, with $\log_{10}$, 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 other methods (Mahalanobis distance, PCA, FA), except for sample 48, which was, also, considered an outlier by cluster analysis. For the compositional data, samples 54 and 68 were considered outliers, but they were not found by the other methods

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 covariance sampling matrix. On the other hand, when the transformation to $\log_{10}$ was used plus compositional transformation 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.

Using the cluster analysis method did not show to be efficient to determine outliers, because the sample 48 is not an outlier, in accordance with other methods, such as Mahalanobis distance, principal component analysis and factor analysis for $\log_{10}$ transformation, while the sample 5 is an outlier for Mahalanobis distance only for compositional data.

Figure 4 shows that samples 7, 10, 21, 28 and 48 are outliers, by the method of the standardized residual for log base 10 (a) and samples 54 and 68 for compositional data (b).

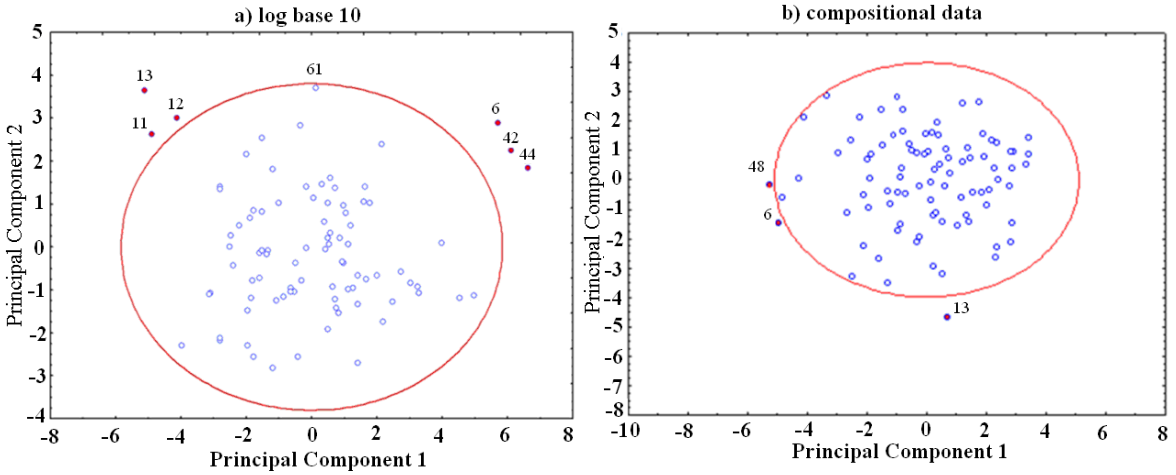

Figure 2. Dispersion diagram for the scores of the first principal component, versus the score of the second principal component. The ellipse represents the confidence level of 95% with: a) log base 10 and b) compositional

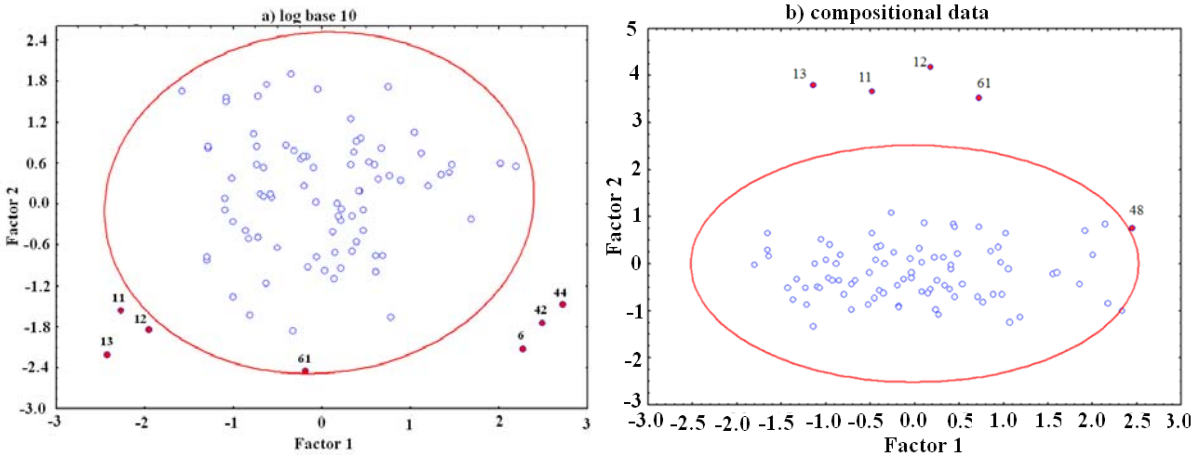

Figure 3. Dispersion diagram by the first score factor versus score of the second factor. The ellipse represents a confidence level of 95% with: a) log base 10 and b) compositional data.

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

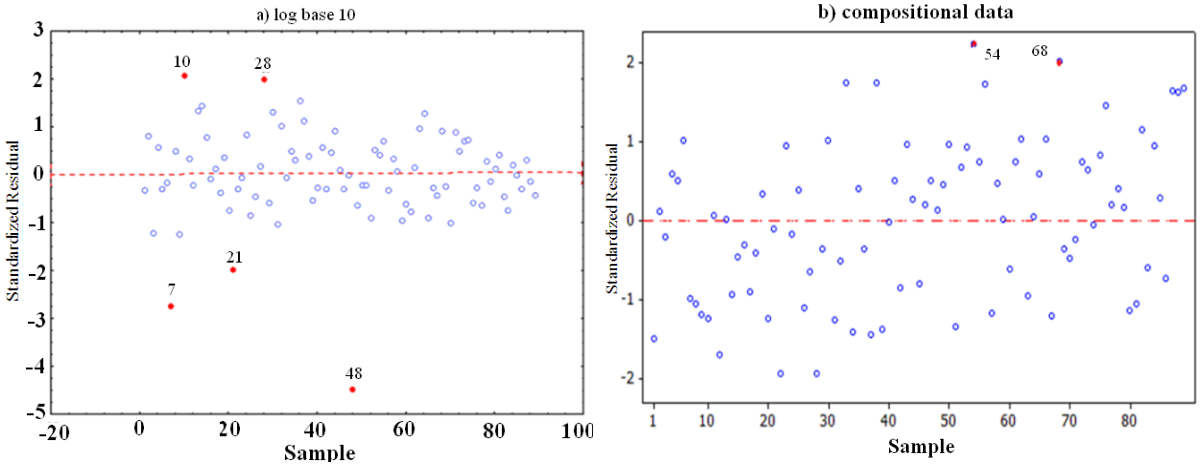

Figure 4. Identification of the samples versus standardized residual with: a) log_{10} and b) compositional data.

4. CONCLUSION

The detection of outliers in a data base is a technical problem that depends on scientific work and on the questions required 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, was 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.

Increased sensitivity between different transformations for the detection of outliers varies according to the method that was used. While the compositional transformation was more sensitive for detection of outliers for Mahalanobis distance, the transformation to log_{10} was more sensitive for detection of the outliers for the methods: principal components analysis, factorial analysis and standardized residuals

ACKNOWLEDGMENTS

Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, process number 2008/54867-7, for the 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 Artifact Compositional Data," *Archaeometry,* **43(4)**, pp. 571-588 (2001).
4. J. Aitchison, *The Statistical Analysis of Compositional Data*, Chapman and Hall, London, United Kingdon (1986).
5. I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, USA (2002).
6. 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).
7. 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).
8. C. Barceló; V. Pawlowsky; E. Grunky. "Some Aspect Transformations of Compositional Data and the Identification of Outliers," *Mathematical Geology*, **28(4)**, pp. 501-518 (2005).
9. A. Buccianti, G. Mateu-Figueras; V. Pawlowsky-Glahn. *Compositional Data Analysis in the Geosciences from Theory to Practice*, Geological Society Special 264 (2006).
10. 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).
11. A. M. Aguiar. *Aplicação do Método de Análise por Ativação à Determinação de Elemento Traços em Unhas Humanas.* Dissertation, Nuclear Energy Research, IPEN – CNEN / SP (2001).
12. P. C. Mahalanobis, "On the Generalized Distance in Statistics," *Proceeding of the National Institute of Sciences of India*, **12**, pp. 49-55, (1936).
13. P. C. Mahalanobis, Historic Note on the D^2 Statistic, *Sanklya*, **9**, pp.237, (1948).
14. S. S. Wilks, "Multivariate Statistical Outliers," *Sanklya,* **25**, pp, 407-426(1963).
15. Ali S. Hadi, "Identifying Multiple Outliers in Multivariate Data," *J. R. Statisc. Soc. B*, **54(3),** pp. 761—771 (1992).
16. K. Pearson, "On Lines and Planes of Closest fit to a System of Points in pace," *Philos. Mag*, **2**, pp. 557—572 (1901).
17. H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educational Psychol.*, **24** pp. 498-520 (1933).
18. B. J. F. Manly, *Métodos Estatísticos Multivariados uma Introdução*, Artmed & Bookmann, Porto Alegre-RS, Brasil (2008)
19. V. Barnett; T. Lewis, *Outliers in Statistical Data*, Wiley & Sons, New York, USA (1994).
20. M. J. Baxter, "Detecting Multivariate Outliers in Artifact Compositional Data," *Archaeometry*, **41,** pp.321-338 (1999).
21. C. Spearman, "General Intelligence, Objectively Determined and Measured," *Am. J. Psychol*., **15**, pp. 201-293 (1904).
22. R. A. Jonhson; D. W. Wichern, *Applied Multivariate Analysis*, Prentice Hall, New Jersey, USA (1998).
23. N. R. Draper; H. Smith, *Applied Regression Analysis*. Wiley & Sons, New York, USA (1998).

24. P. M. S. Oliveira; C. S. Munita; R. Hazenfratz, Comparative Study on Methods of Outlying Data Detection in Experimental Results, International Nuclear Atlantic Conference – INAC 2009, September 27 to October 2, (2009).
25. 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).