J Radioanal Nucl Chem (2010) 283:433–437
DOI 10.1007/s10967-009-0382-2

Comparative study between three methods of outlying detection on experimental results

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

Received: 24 July 2009 / Published online: 28 October 2009

Abstract This paper describes experimental results through multivariate statistical methods that might reveal outliers that are rarely taken into account by analysts. The results were submitted to three procedures to detect outliers: Mahalanobis distance, MD, cluster analysis, CA, and principal component analysis, PCA. The results showed that although CA is one of the procedures most often used to identify outliers, it can fail by not showing the samples that are easily identified as outliers by other methods, like MD. Mahalanobis distance proved to be the simpler application, with sensitive procedures to identify outliers in multivariate datasets.

Keywords Outliers · Mahalanobis distance · Cluster analysis · Principal component analysis · Archaeometry

Introduction

In studies where samples are grouped in clusters, like in ceramic archaeological studies, the presence of outliers is critical. This is because frequently, two or more groups of samples are found where clusters can represent different chemical compositional groupings [1], and the effect of an outlier on the means, variances and on the correlations between the variables may need to be taken into account.

Outliers are atypical results that can happen due to uncontrolled process, wrong analytical technique, contamination during the preparation of the sample, sample inhomogeneity, measurements with a high systematic error, and so on. In some cases it is easy to see the outlier, especially in the univariate dataset, when the result is clearly different from the other values [2, 3]. When the analyst uses an analytical technique that determines several elements simultaneously, like INAA or another one, the value can be hidden in the other values. This is especially true when the numerical value is of the same order as the others. So, the presence of outliers needs to be the primary concern of the analyst, because he can arrive to a misinterpretation and distortion in the results [4, 5].

In the literature, several papers were published using many methods to detect outliers, however, almost all of them were directed at a single univariate case [6]. However, much less work has been done on multivariate outliers and the methods employed are the MD [1, 3], PCA [7], CA [8], mask [9], ellipsoid minimum volume [10], decisive minimum of the covariance matrix [11], and so on.

In this paper, a comparative study was made on a real dataset obtained via instrumental neutron activation analysis, INAA, using three procedures used very frequently in archaeometric studies to detect the outliers: MD, CA and PCA.

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. After that, these materials were dried in an oven at 105 °C for 24 h, and stored in a desiccator.

Constituent Elements in Coal Fly Ash (NIST-SRM-1633b) were used as standards, and IAEA-Soil-7, Trace

P. M. S. Oliveira · C. S. Munita (✉) · R. Hazenfratz
Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP,
Av. Prof. Lineu Prestes, 2242, Cidade Universitária, São Paulo,
SP 05508-000, Brazil
e-mail: camunita@ipen.br

Elements in Soil, were used to check samples in every analysis. These materials were dried in an oven at 105 °C for 24 h.

About 100 mg of different ceramic samples, NIST-SRM-1633b and IAEA-Soil-7 were weighed in polyethylene bags and wrapped in aluminum foil. Groups of 8 samples, and one of each reference material were packed and irradiated in the research reactor pool, IEA-R1, from the IPEN-CNEN/SP, at a thermal neutron flux of about 5×10^{12} n cm^{-2} s^{-1} for 8 h.

Two measurement series were carried out using Ge (hyperpure) detector, model GX 1925 from Canberra, resolution of 1.90 keV at the 1332.49 keV gamma peak of ^{60}Co, with S-100 MCA of Canberra with 8192 channels. K, La, Lu, Na, Nd, U, and Yb were measured after 7 days cooling time, and Ba, Ce, Co, Cr, Cs, Eu, Fe, Hf, Rb, Sb, Sc, Sm, Ta, Tb, Th, and Zn after 25–30 days. Gamma ray spectra analysis and the concentrations were carried out using the Genie-2000 Neutron Activation Analysis Processing Procedure from Canberra. A detailed description of the standard sample preparation, and the procedure were published elsewhere [12].

Statistical method

Mahalanobis distance

The MD 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 one of the n samples and p variables, the MD (D_i) from the sample to the centroid is calculated by means of the expression $D_i = \sqrt{(x_i - \bar{x})^t S^{-1} (x_i - \bar{x})}$ for $i = 1,\ldots,n$ where ^{t} is the transpose matrix, $\bar{x}$is the arithmetic mean of the concentrations and S is the variance-covariance sampling.

Based on the asymptotic distribution of the MD, the critical value, cv, should be chosen. For smaller samples, several authors showed that the Lambda Wilks criteria to calculate the cv can be an appropriate choice [13]. In this paper, this criteria was adopted using the expression $cv = p(n-1)^2\ F_{p,\ n-p-1;\ \alpha/n}/n(n-p-1+pF_{p,\ n-p-1,\ \alpha/n})$ where p is a number of variables; n is a number of samples and F is the F test also called Fisher distribution ($F = s_1^2/s_2^2$ where s_1^2 and s_2^2 are sample variances) with p degrees of freedom at a significance level of α/n, $\alpha = 0.05$.

Cluster analysis

It is a graphical visualization method, and seems to be the common method used for identifying multivariate outliers. The interpretation of CA is usually based on dendrograms (trees) and is subjective. There are two methods of cluster analysis that are mostly used in archaeometric studies: average linkage and Wards, using Euclidean or squared-mean Euclidean distance as a dissimilarity measure applied to standardized or logarithmically transformed data [1, 8]. Single linkage is usually avoided as a means of group definition because of the chaining phenomenon, which can link distinct groups [1].

The identification of the sample outliers is by means of the leaves of the tree that are distinct from the main dataset, and are separated out at a high level of dissimilar distance. The samples that are isolated in a single group or with a measure of dissimilar distance much larger than the others are outliers.

Principal component analysis

PCA is a technique that transforms linearly one set of p variables observed in a smaller set of k non-correlates variables, and that explains 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 PCs 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.

Several studies of archaeological ceramics show that more than 70% of the total variance is explained in the first two PC scores.

Results and discussion

To evaluate the analytical process, and to establish the chemical elements which can be used to compare the performance of MD, CA and PCA in identifying outliers, the elemental concentrations of the 18 samples of the IAEA Soil 7 reference material were statistically compared with the certified values. The elements with an accurate and relative standard deviation (RSD) of less than 10% were Ce, Cr, Cs, Eu, Fe, Hf, La, Lu, Na, Sc, Tb, U and Yb [12]. Afterwards, these elements were used in the subsequent data analyses.

Table 1 shows the values of the elemental concentrations for 31 samples of ceramic fragments. Initially, the results were transformed to $\log_{10}$ 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 [12]. Then, throughout the work, it was assumed that the dataset was log-normally distributed. After logarithmic transformation, the dataset was submitted to outlying tests.

Table 1 Results for the concentration data for ceramic samples in μg/g, unless otherwise indicated, and Mahalanobis distance

Sample	Na (%)	Lu	U	Yb	La	Th	Cr	Cs	Sc	Fe(%)	Eu	Ce	Hf	Tb	D_1	D_2	D_3
1	0.05	0.34	3.02	2.45	36.31	14.13	57.54	7.76	13.80	3.09	1.02	74.13	8.91	0.60	5.1	7.3	7.0
2	0.07	0.40	3.39	2.09	35.48	13.18	63.10	7.08	14.45	3.09	0.79	66.07	6.76	0.79	7.8	8.4	8.0
3	0.05	0.30	3.09	2.29	33.11	12.88	66.07	9.12	13.18	2.51	0.79	72.44	7.94	1.10	15.3	15.0	14.6
4	0.05	0.30	3.72	1.82	27.54	13.80	67.61	7.94	15.85	3.80	0.60	48.98	6.17	0.79	15.0	18.3	17.6
5	**0.07**	**0.30**	**3.72**	**2.40**	**28.84**	**15.14**	**52.48**	**14.13**	**12.59**	**2.29**	**0.79**	**63.10**	**6.03**	**0.79**	**21.8**	**23.0**	
6	0.10	0.40	3.24	2.69	38.02	12.30	50.12	5.25	11.22	2.82	0.89	74.13	12.02	1.02	13.6	15.3	14.8
7	0.09	0.34	3.24	2.57	22.91	10.47	60.26	11.75	13.80	2.88	0.69	44.67	4.68	0.41	16.6	19.9	19.2
8	0.10	0.30	2.88	1.82	26.92	10.96	47.86	8.91	11.48	3.02	0.71	53.70	7.59	0.40	14.7	14.2	13.7
9	0.15	0.33	3.31	2.51	34.67	15.14	64.57	10.23	15.14	3.39	0.95	63.10	7.41	0.65	7.1	7.8	8.5
10	0.18	0.33	3.24	2.34	22.91	14.79	63.10	7.76	13.80	3.31	0.74	45.71	7.59	0.37	7.4	8.0	9.2
11	**0.37**	**0.28**	**2.51**	**2.14**	**25.12**	**10.72**	**48.98**	**13.49**	**12.59**	**3.02**	**0.09**	**53.70**	**4.79**	**0.65**	**28.3**		
12	0.20	0.60	5.25	2.88	39.81	14.45	66.07	10.00	15.49	3.47	1.29	77.62	6.92	0.71	16.7	16.2	16.4
13	0.20	0.50	3.47	3.02	40.74	14.79	64.57	8.32	15.49	3.47	1.20	79.43	8.13	0.60	3.4	4.0	5.1
14	0.20	0.50	4.17	3.47	43.65	14.45	66.07	11.22	16.22	3.47	1.41	79.43	7.24	0.79	6.4	6.2	5.9
15	0.20	0.40	3.98	3.02	39.81	14.45	60.26	10.72	15.14	3.47	1.29	77.62	6.92	0.89	4.7	4.7	12.0
16	0.22	0.45	4.17	3.16	39.81	14.45	64.57	9.55	15.85	3.98	1.35	72.44	8.71	0.81	5.9	7.3	11.8
17	0.20	0.60	3.63	3.24	39.81	15.14	64.57	8.91	15.85	3.98	1.41	77.62	8.32	0.89	11.4	15.0	15.0
18	0.20	0.40	4.37	3.09	43.65	15.14	58.88	12.02	16.22	3.31	1.41	104.71	9.12	1.00	12.7	12.3	13.0
19	0.20	0.40	3.47	2.82	35.48	12.30	56.23	10.72	14.45	3.02	1.20	85.11	7.08	1.10	8.7	12.0	12.5
20	0.10	0.34	3.89	2.40	36.31	12.02	64.57	22.91	14.45	2.63	1.15	112.20	6.03	0.69	21.1	20.0	19.5
21	0.20	0.50	4.17	3.72	67.61	14.79	63.10	6.17	15.14	3.72	1.91	138.04	8.51	1.82	15.7	15.4	15.3
22	0.20	0.60	4.68	4.47	56.23	16.98	69.18	9.12	17.38	3.80	1.70	123.03	9.55	1.20	6.6	11.6	11.3
23	0.20	0.48	5.13	3.80	44.67	17.38	77.62	11.48	19.95	4.47	1.66	95.50	7.94	1.20	8.6	8.9	9.6
24	0.20	0.50	2.69	2.69	36.31	13.49	60.26	5.75	14.79	3.02	1.02	70.79	6.61	0.50	13.6	14.6	14.1
25	0.19	0.37	3.55	2.75	36.31	14.13	63.10	4.79	15.85	3.98	1.23	75.86	6.92	0.76	7.0	8.6	8.3
26	0.10	0.51	3.89	3.89	47.86	18.20	79.43	12.30	19.95	4.90	1.66	120.23	6.76	0.71	13.6	14.9	14.6
26	0.30	0.44	4.27	2.82	37.15	13.18	58.88	4.68	14.79	3.31	1.17	72.44	8.13	0.83	11.2	11.4	11.2
27	0.35	0.35	3.89	2.57	31.62	13.80	61.66	2.88	14.45	5.01	0.95	60.26	7.24	0.65	17.6	17.0	16.7
28	0.07	0.55	5.37	3.80	35.48	27.54	100.00	9.33	17.78	5.13	1.17	89.13	14.45	0.89	18.9	18.0	17.3
30	0.07	0.47	3.98	3.63	32.36	19.95	77.62	8.91	15.85	4.07	1.15	83.18	11.48	0.81	9.6	9.1	8.9
31	**0.06**	**0.42**	**3.31**	**2.95**	**67.61**	**24.55**	**66.07**	**6.61**	**13.80**	**2.00**	**1.32**	**141.25**	**11.48**	**0.68**	**24.2**		
$D_{\text{critical value at significance level of 0.05}}$															23.6	22.8	22.3

In the last three columns of Table 1 are the MD values of each sample, and the end for the critical value, calculated using the lambda Wilks criteria. The MD for each sample was calculated, and the samples with *D* being higher than the critical value, was excluded in the dataset, and *D* was recalculated for the reduced dataset. The stopping rule is when *D* calculated in the samples does not exceed the critical value. In accordance with the MD, in the Table 1, samples 5, 11 and 31 are outliers.

After that, the data were submitted to CA. It is well known that different clustering methods produce different results when applied to the same data, and a particular method may not reflect the true structure of the data. In this paper Ward's method was used and a similarity between the samples was measured by a squared-mean Euclidean distance, because it is the most widely applied method in archaeology, due to giving greater emphasis on larger differences between variables [1]. On the other hand, forming groups by estimating each stage of the clustering process, and knowing the combination of groups will produce the minimum increase in the error sum of squares, as measured by the total sum of squared deviations, from every member of the cluster, and from the mean of that cluster. Ward's method has a bias toward the discovery of relatively dense, well separated, hyperspherical clusters, and will create them if necessary. It also tends to find clusters of equal size, and emphasizes the creation of homogeneous groups [2, 8].

Figure 1 shows the dendrogram, presenting the results of CA of 31 samples, and the distance measure used was squared-mean Euclidean. Visual inspection of the dendrogram is useful as a method of identifying preliminary groups and outliers. The dendrogram suggests that there are two groups of samples: one group formed by samples 1–10 and the other group, by the samples 12–31. The sample 11 is clearly an outlier.

The third studied procedure, to determine the number of outliers in the results presented in Table 1 was the PCA technique. PCA is a powerful technique, and essential when compositional data are highly correlated. In PCA, a transformation of the dataset, based on eigenvector methods, is preformed to determine the direction and magnitude of maximum variance. Then PCA begins with the *p* correlated variables, and the procedure transforms these to an uncorrelated set of *p* new variables. So PCA provides a means for reducing the dimensionality of the dataset, with the minimum loss of information [7]. Figure 2 shows the plot of PC1 vs PC2, and the variance explained in the first two PCs was 64%. As can be seen in Fig. 2, sample 31 is an outlier and was confirmed plotting the scores of PC2 in a box-plot, as shown in Fig. 3. Looking at the plot (Fig. 2), it is not possible confirm that sample 11 is an outlier. However, when only the sample 31 is excluded in the dataset and plotting PC1 vs PC2 there are two groups with some samples overlapping. On the other hand, when are excluded in the dataset both (sample 11 and 31) the plot showed two groups and any overlap. We think that sample 11 is missing, because in the plot, only the first two PCs are presented which explained 64% of the total variance, and the variables are highly inter-correlated.

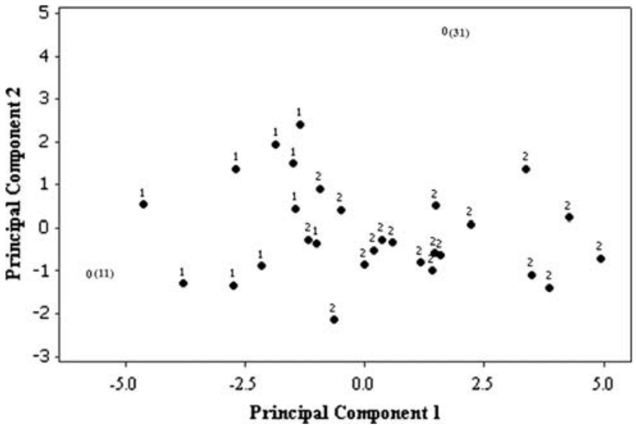


Fig. 2 First and second principal component biplot for 31 ceramic samples

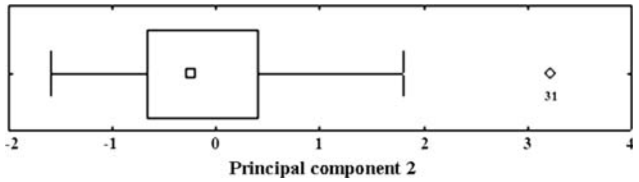
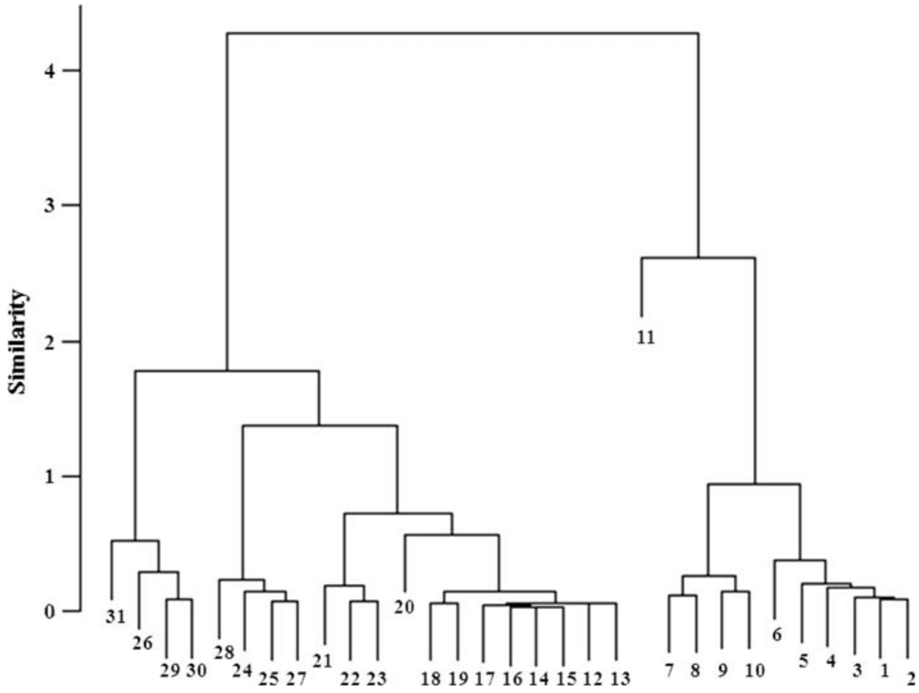


Fig. 3 Box-plot for the second principal component for 31 samples

We can quickly say that Ward's method of CA using squared-means Euclidean distance procedure is used very much in archaeometric studies as a method to identify outliers, has been shown to be a procedure that is not very objective. In practice, it has limitations, because some

Fig. 1 Dendrogram for 31 ceramic samples using Ward's method and squared-mean Euclidean distance



cases depend on the ability of the analyst to look at the sample inside of the leaves of the tree (dendrogram), which is different from the group of samples. When there are too many samples, it is practically impossible to find it. With frequency, the outliers are mixed and are forming part of the group. On our case, using a small dataset, CA was sufficiently clear to identify only one outlier, sample 11, however, in the dataset there are three (samples 5, 11 and 31), as suggested by the MD. When the data were tested using single linkage CA, the same outliers (samples 11 and 31) were found, and with an average linkage samples 11 and 27 were found. Other author [1], using different forms of CA, also found that CA can fail to reveal outliers clearly identified by other methods. The problem using single linkage is that the partition of the group it is not very clear.

A very important point is that the presence of the outliers in a dataset simply does not need to be excluded from the results, after applying some procedures to identify, it is necessary to carry out the data base, to reduce the impact on the interpretation. After that, the outliers outside the dataset will need a more detailed study.

Conclusions

Only through visual inspection of the data is probably the most common method used by the analyst in identifying outliers, but in the largest datasets, visual inspection of the data may be quite impossible. Thus, it becomes necessary to apply some type of objective criterion at the outset. Keeping this in mind, throughout this paper, were presented three methods: MD, Ward's CA method using squared-means Euclidean distance, and PCA, used to identify outliers. As was shown, a sample can be outlier by one statistical method and not to be detected by other procedure because the outliers can inflate estimates of the variances without affecting the correlations very much point. For example, an outlier which mainly affected a subset of variables with low or highly inter-correlations would not be detectable by examining the projection of the points onto the first few principal components. By the results obtained, there is not specific method that can be recommended for outlier detection. However, in the dataset the more convenient procedure was MD using lambda Wilks as critical value. The MD has the advantage that is easy to apply, to identify and is very sensitive to the presence of outliers.

Acknowledgement We thank to Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, Process 08/54867-7 for financial support.

References

1. Baxter MJ (1999) Detecting multivariate outliers in artefact compositional data. Archaeometry 41: 321–338
2. Beckman RJ, Cook RD (1983) Outliers. Technometrics 25: 119–149
3. Willems G, Joe H, Zamar R (2009) Diagnosing multivariate outliers detected by robust estimator. J Comput Graph Stat 18: 73–91
4. Andrews DF, Pregibon D (1978) Finding the outliers that matter. J R Stat Soc B 40: 85–93
5. Alqallaf F, van Aelst S, Yohai VJ, Zamar RH (2009) Propagation of outliers in multivariate data. Ann Stat 37: 311–331
6. Cerioli A, Riani M (1999) The ordering of spatial data and the detection of multiple outliers. J Comput Graph Stat 8: 239–258
7. Serneels T, Verdock T (2008) Principal component analysis for data containing outliers and missing elements. J Comput Stat Data Anal 52: 1712–1727
8. Papageorgiou J, Baxter MJ (2001) Model-based cluster analysis of artefact compositional data. Archaeometry 43: 571–588
9. Zhang J, Wang X (1998) Unmasking test for multiple upper or lower outliers in normal samples. J Appl Stat 25: 257–261
10. Rousseeuw PJ, van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85: 633–639
11. Hardin J, Rocke DM (2004) Outlier detection in multiple cluster setting using the minimum covariance determinant estimator. J Comput Stat Data Anal 44: 625–638
12. Santos JO, Munita CS, Valério MEG, Vergne C, Oliveira PMS (2008) Correlations between chemical composition and provenance of Justino site ceramics by INAA. J Radioanal Nucl Chem 278: 185–190
13. Penny KI (1996) Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. Appl Stat 45: 73–81