*2013 International Nuclear Atlantic Conference - INAC 2013*
*Recife, PE, Brazil, November 24-29, 2013*
*ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN*
***ISBN:*** **978-85-99141-05-2**

# OUTLIER DETECTION BY ROBUST MAHALANOBIS DISTANCE IN GEOLOGICAL DATA OBTAINED BY INAA TO PROVENANCE STUDIES

**José O. dos Santos[1], Casimiro S. Munita[2] and Emílio A. A. Soares[3]**

[1] Coordenação de Física - Instituto Federal de Educação, Ciência e Tecnologia de Sergipe
Av. Lourival Baptista, s/n
79000-000 Lagarto, SE
osmansantos@ig.com.br

[2] Instituto de Pesquisas Energéticas e Nucleares (IPEN / CNEN - SP)
Av. Professor Lineu Prestes 2242
05508-000 São Paulo, SP
camunita@ipen.br

[3]Departamento de Geociências – Universidade Federal do Amazonas (UFAM)
Av. Gal. Rodrigo O. J. Ramos, 300
69077-000 Manaus, AM
easores@ufam.edu.br

## ABSTRACT

The detection of outlier in geochemical studies is one of the main difficulties in the interpretation of dataset because they can disturb the statistical method. The search for outliers in geochemical studies is usually based in the Mahalanobis distance (MD), since points in multivariate space that are a distance larger the some predetermined values from center of the data are considered outliers. However, the MD is very sensitive to the presence of discrepant samples. Many robust estimators for location and covariance have been introduced in the literature, such as Minimum Covariance Determinant (MCD) estimator. When MCD estimators are used to calculate the MD leads to the so-called Robust Mahalanobis Distance (RD). In this context, in this work RD was used to detect outliers in geological study of samples collected from confluence of Negro and Solimões rivers. The purpose of this study was to study the contributions of the sediments deposited by the Solimões and Negro rivers in the filling of the tectonic depressions at Paraná do Ariaú. For that 113 samples were analyzed by Instrumental Neutron Activation Analysis (INAA) in which were determined the concentration of As, Ba, Ce, Co, Cr, Cs, Eu, Fe, Hf, K, La, Lu, Na, Nd, Rb, Sb, Sc, Sm, U, Yb, Ta, Tb, Th and Zn. In the dataset was possible to construct the ellipse corresponding to robust Mahalanobis distance for each group of samples. The samples found outside of the tolerance ellipse were considered an outlier. The results showed that Robust Mahalanobis Distance was more appropriate for the identification of the outliers, once it is a more restrictive method.

**Keywords:** INAA, Multivariate Outliers, Sediments, Solimões-Amazon River, Robust Mahalanobis Distance.

## 1. INTRODUCTION

Geochemistry datasets often contain outliers which represent the influence of extraneous and exotic processes such as those correlated to singular rock types, mineral deposits or anthropogenic phenomena. Generally, in geochemistry studies, outliers are observations

resulting from secondary processes and are almost always multivariate. In addition, atypical samples should not be simply ignored because contain information about data quality and rare phenomenon in the area of interest [1, 2].

The outlier detection in multivariate space is one important task in statistical interpretation of geochemical data, since outlying samples can contain a lot of valuable information about data set and its presence can interfere in statistical interpretation of data, especially in provenance studies [3]. Generally, atypical samples in a geological data set does not need to be especially high (or low) in relation to all values of a variable, and thus attempts to identify these samples with classical univariate methods commonly fail. In order to resolve this difficulty, several methods to identify outlier on multidimensional data set have been propose based on covariance matrix and mean vector, such as the methods which use Mahalanobis distance [4]. This statistical method often uses classical Mahalanobis Distance (MD) to detect if an observation is far from the centre of the data distribution and if MD is larger than critical value the sample is candidate to be outlier.

The classical Mahalanobis distance estimator (MD) between the points $\overline{X}_i$ and $\overline{X}$ is defined by eq.(1), where $\overline{X}_i$ is the ith observation from a p-dimensional data set, with p equal to number of variable, S is a covariance matrix and is a $\overline{X}$ mean vector.

$$MD_i = \sqrt{\left(X_i - \overline{\mathrm{X}}\right)^T S^{-1}\left(X_i - \overline{\mathrm{X}}\right)} \tag{1}$$

To a data set with normal distribution, MD has a $\chi^2_p$ distribution and observations with a large distance, in generally $MD^2 > \chi^2_{p,0.975}$, are indicated as outliers [5]. However, masking effects decrease the MD of an outlier, on the other hand, swamping effects increase the MD of non-outlying observation [6]. This way, the MD needs to be estimated by a robust procedure in order to provide reliable measurements because the Mahalanobis distance is very sensitive to the presence of outliers.

In this context, the Minimum Covariance Determinant estimator (MCD) is frequently used as robust estimator for location and covariance, where MCD is determined by a subset of observation of size k, generally $k \approx 0{,}75n$ (n is a number of samples), which minimizes the determinant of the sample covariance matrix [5]. The choice of k determines the robustness of the estimator. Using robust estimators of location and scatter in eq. (1) results in so-called Robust Distances (RDs) and if $RD^2 > \chi^2_{p,0.975}$ the sample can be declared a candidate outlier [2].

In this paper, a geochemical data set of samples from the confluence of Negro and Solimões rivers has been utilized to study outliers in multivariate space by means of Robust Mahalanobis Distance (RD). The Amazon system is the largest and most complex terrestrial ecosystem, and its formation is correlated by dynamics of the Amazonas River system [7-8]. The main changes in the landscape of the Amazon region, which have been shown on the pattern of sedimentation, relief and in the distribution of current biodiversity, have contributions from Andean tectonics and climate change that occurred in the Cenozoic. These transitions in the landscape can be observed through of the analysis of the sedimentary

deposits of the Amazon basin, since they are results of migration and overlapping of different river systems from the Cretaceous [9].

In order to develop an evolutionary sedimentary-tectono model to Pleistocene period to the confluence of the rivers Negro and Solimões, it is running a multidisciplinary project to stratigraphic-sedimentological characterization of the Pleistocene succession [10]. In particular, the analysis of trace elements of sediments samples to geochemical study of sediments, it can contributed mainly to determine the contribution of sediment load deposited from Negro and Solimões rivers in filling of tectonic depressions and changes in the weathering degree between younger and older units, and thus have additional subsidies to build the geological evolution of the area. The geochemistry of trace elements has been extensively investigated during last years because of their importance to study of provenance and petrogenical and geochemical problems. Trace elements, in particular rare earth elements, form a coherent group with similar chemical composition, which is important to reveal chemical processes in geological systems, may also provide a characteristic fingerprint of different mineral [11,12].

## 2. METHODOLOGY

### 2.1. Sample Collection

The confluence of Negro and Solimões rivers results in the formation of the Amazon River downstream from the city of Manaus, Brazil (Figure 1). The Negro river, with basin area of 686,810 $km^2$, drains the western slopes of the Guyana shield and joins Solimões River. The Solimões River has a drainage area of 2,150,000 $km^2$, delivers sediment-rich waters, dammed from Andes located in the west of basin [13].

Evidence of the Quaternary tectonic and sedimentary history of Solimões-Amazon river system is registered in fluvial deposits found in the confluence zone of Negro and Solimões Rivers [9]. Extensional tectonics originated tectonic depression that have controlled the Pleistocene sedimentation influenced by Solimões (Paraná do Ariaú – GPA and Lago do Miriti – GLM grabens) and Negro river dynamics. The Pleistocene units occur in three levels of terraces. The Negro river always exhibited a straight style, confined in basement Cretaceous and Miocene rocks and with restricted alluvial plain, whereas the fluvial pattern of Solimões River has changed during the Holocene.

For chemical compositional characterization of sediments from confluence zone of Negro (Cachoeira do Castanho Grabens – GCC and Cacau do Pirera Grabens – GCP), Paraná do Ariaú Grabens (GPA) and Solimões rivers, samples of sediments deposited by fluvial system were collected (113 samples) at high points of terrace (Figure 1), stored into plastic bags, where the water contents are about 2-3% and where they did not remain immersed during long periods of time.

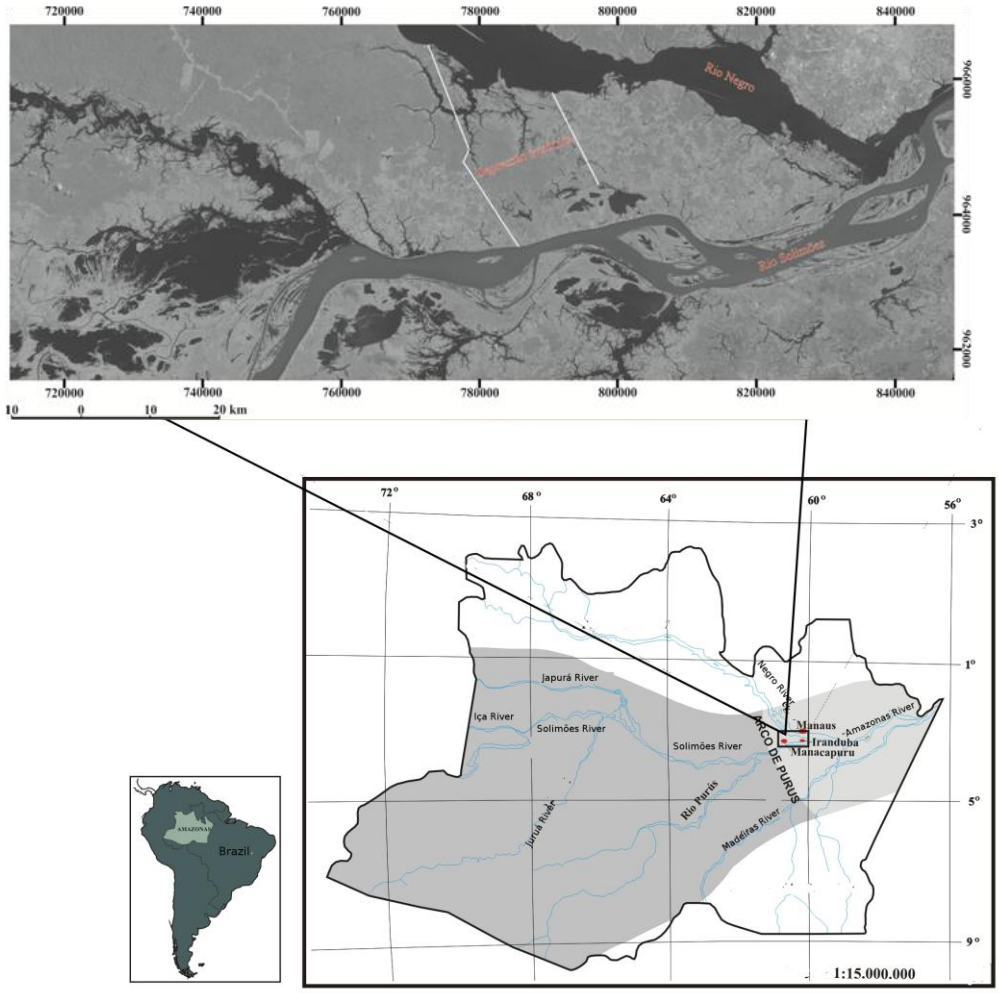


**Figure 1. Location of studies sites.**

## 2.2. Sample Preparation

Sediment samples were ground in an agate mortar until a granulometry of 100-200 mesh was achieved and quarted for the chemical determination by INAA. Finally, the powered samples were dried in an oven at 105$^{o}$C for 24 hrs and stored in desiccators. Different authors have considered that there is no significant volatilization of the elements studied in this work when heated at this temperature [14, 15].

Constituent Elements in Coal Fly Ash - NIST-SRM-1633b, was used as standard in all analyses. The standard reference material Brick Clay - NIST-SRM-679 was used to check the analytical quality of the results. The standards and the samples were dried in an oven at 105$^{o}$C, the standards for 4 hrs and samples for 24 hrs and stored in desiccators until weighing.

## 2.3. Irradiation

About 100 mg of sediments samples, and NIST-SRM-1633b were weighed in polyethylene bags and wrapped in aluminum foil. The groups of eight sediments samples and two

reference materials were packed in aluminum foil and irradiated in the swimming pool research reactor, IEA-R1 (IPEN/CNEN – SP) at a thermal neutron flux at about 5 x $10^{12}$ $n \cdot cm^{-2} \cdot s^{-1}$ for 8hrs.

### 2.4. Gamma Spectrometry

Two measurement series were carried out using Ge (hyperpure) detector, model GX 1925 from Canberra, with a resolution of 1.90 keV at the 1332.49 keV gamma peaks of $^{60}Co$, coupled to a S-100 MCA of Canberra with 8192 channels. As, K, La, Lu, Na, Nd, Sm, U, and Yb were measured after 7 days cooling time and Ba, Ce, Co, Cr, Cs, Eu, Fe, Hf, Rb, Sb, Sc, Ta, Tb, Th, and Zn after 25-30 days. Gamma ray spectrum analysis was carried out using the software Genie 2000 NAA Procedure from Canberra.

### 2.5. Statistical Interpretation

The first stage in statistical interpretation of geochemical data is to describe the ranges of concentrations of the elements and to get on identification of the presence of outliers, which are observations that appear to be inconsistent with the rest of the data. In this work, multiple outliers have been detected by means of using RD based in MCD, according eq. (1), where 75% of samples were utilized to calculate the MCD as a compromise between robustness and efficiency [2].

For outlier detection by RD analysis the "*mvoutlier"* and "*chemometrics"* packages running under R environment were used [16-18]. It was constructed graphics with classical Mahalanobis distance of data against the robust Mahalanobis distance, graphics called dd-plot, which can be used to detect outlier [19]. In addition, the first principal components against the second principal component was plotted from data set using different symbols according to RD, and four ellipses were drawn, on which Mahalanobis distances are constant corresponding to 25%, 50%, 75% and adjusted quantile of the chi-square distribution. In this last plot, samples that have a RD higher than criteria were defined outliers. Finally, three graphics, one for each group (Negro, Solimões and Paraná do Ariaú) were generated, showing MD and RD versus the observation numbers [19], where samples which distance was higher than cutoff are considered as a potential outlier.

After outlier detection, linear discriminant analysis (LDA) was applied to INAA results to determine the level of chemical variability between the characterized samples and differentiate between the sedimentary deposits. LDA is a multivariate method which produces a data reduction that enabling the consideration of extensive variables that result from multi-element chemical characterization such as INAA. In order that the LDA is a technique that can be used for pattern recognition under supervision, so requires a priori classification. Here, it was assumed that samples collected from each sedimentary deposit was originated at that site collection (Paraná do Ariaú – GPA; Rio Negro; Rio Solimões), which established a priori compositional groups.

## 3. RESULTS AND DISCUSSION

It was analyzed by INAA the elemental composition of 113 samples from Amazon system, 46 from of the influence of Solimões River, 35 from Negro River and 32 from Paraná do Ariaú Graben, and 24 chemical elements were analyzed (Fe, K, Na, Ba, As, Cs, Rb, Sb, Sc, Ta, Zn, Co, Cr, Th, U, Hf, La, Ce, Nd, Sm, Eu, Tb, Yb and Lu). To evaluate the analytical process and to establish the chemical elements which can be used in the data interpretation, the elemental concentrations for reference material Brick Clay - NIST-SRM-679 were statistically compared with the data found in our laboratory. The precision of several elements (La, Th, Sc, Fe, Eu, Ce, Zn, Hf, and Co) was better than 5%. Some elements presented a RSD (Relative Standard Deviation) less than 10% (Nd, Rb, Sm, Ba, Sb, Ta, and Tb) and are similar to those from the literature [20].

Elements that have low precision can reduce the discriminating effects of other well measured elements. In this study all the elements with precision of less than 10% were considered for interpretation of the results (Na, Lu, Yb, La, Th, Cr, Cs, Sc, Ce, Fe, Eu, Zn, Co, Ta, U and Hf). The Zn presented RSD better than 10% but was excluded from the data set because its determination suffers strong gamma ray interferences of $^{46}$Sc and $^{182}$Ta. The K and Sb were better than 10%, however they were excluded because they presented 15% of missing values.

The results were transformed to $\log_{10}$ to compensate for the large magnitude difference between the measured elements at the trace level and the larger ones. One reason for this is the belief that elements from geochemical studies have a natural lognormal distribution, and that data normalization is desirable [21].

Using the RD estimator, based in minimum covariance determinant (k=0.75n), we can find 14 points that are unusually far away from location and call those points outlying in our data set. The dd-plot which is shown in Figure 2, where we can identify five outliers from Rio Negro (RN), four samples in Paraná do Ariaú (PA) and five in Solimões area (S). In all graphics in Figure 2 both axes we have the cutoff value $\sqrt{\chi^2_{24;0.975}}$ indicated, if the data set was not contaminated then all points would lie near the cutoff lines. However, in Figures 2, we can see that five samples in RN, 4 samples in PA and five samples in S are outliers, where RDs to each area is above at cutoff line.

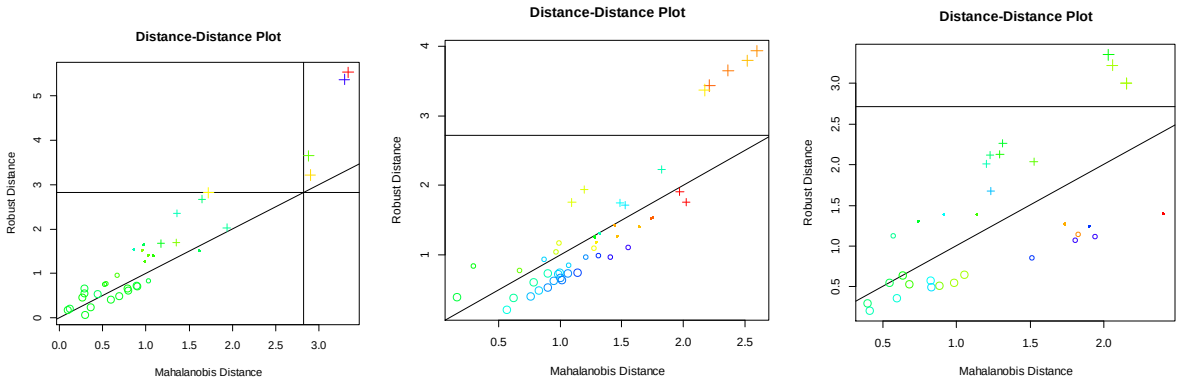


**Figure 2. dd-plot to RN (left), S (center) and PA (right) samples. Samples far from cutoff line can be considered an outlier.**

The RDs were computed and four inner ellipses – Figure 3 to RN, Figure 4 to PA and Figure 5 to S – are shown for 0.25, 0.5, 0.75 and adjusted quantiles of $\sqrt{\chi^2_{24;0.975}}$, where observations between 0.25 and 0.5 tolerance ellipses are shown by a larger dot and the most distance non-outlier are plotted as a small plus. Finally, multivariate outliers that are outside the outer tolerance ellipse are represented by a large plus [16]. Considering this criterion, five samples are showed in RN (Figure 3), four in AP (Figure 4) and five in S (Figure 5).

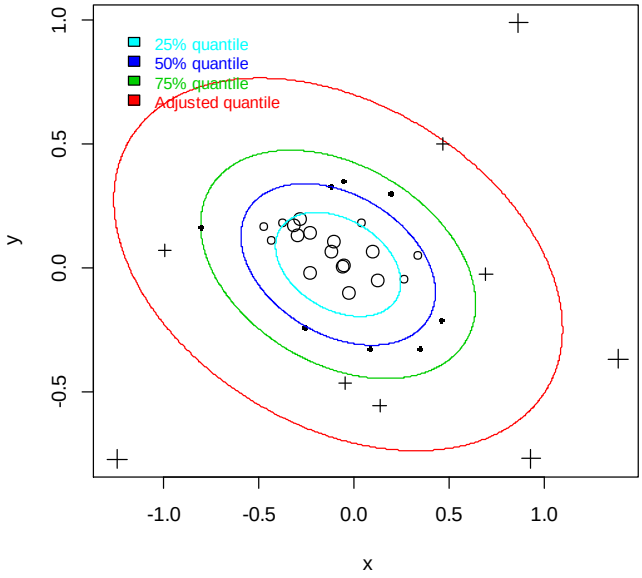


**Figure 3. The groups are defined by tolerance defined for chi-squared quantiles 0.25, 0.5, 0.75 and adjustable quantile to Rio Negro samples. x represents first principal component and y second principal component.**

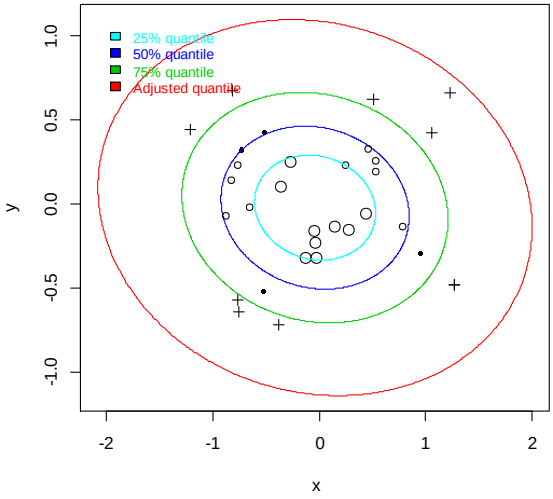


**Figure 4. The groups are defined by tolerance defined for chi-squared quantiles 0.25, 0.5, 0.75 and adjustable quantile to Paraná do Ariaú samples. x represents first principal component and y second principal component.**

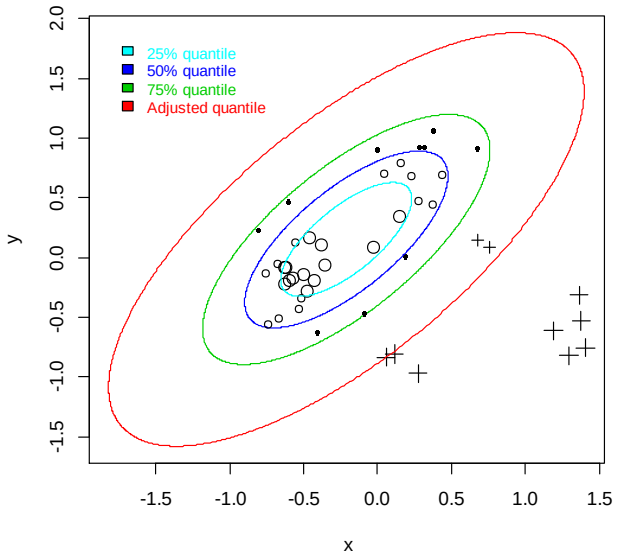


**Figure 5. The groups are defined by tolerance defined for chi-squared quantiles 0.25, 0.5, 0.75 and adjustable quantile to Solimões samples. x represents first principal component and y second principal component.**

To confirm the samples which considered outlier suspects we showed in Figures 6 – 8, Classical Mahalanobis Distance (MD) and Robust Mahalanobis Distance (RD) for each group of the samples, as well a line corresponding a cutoff value. The horizontal lines correspond to the cutoff value $\sqrt{\chi^2_{24;0.975}}$. We can see in Figures 6 – 8 that using RDs estimator several outliers have been identified which we were not considered discrepant by means of MDs.

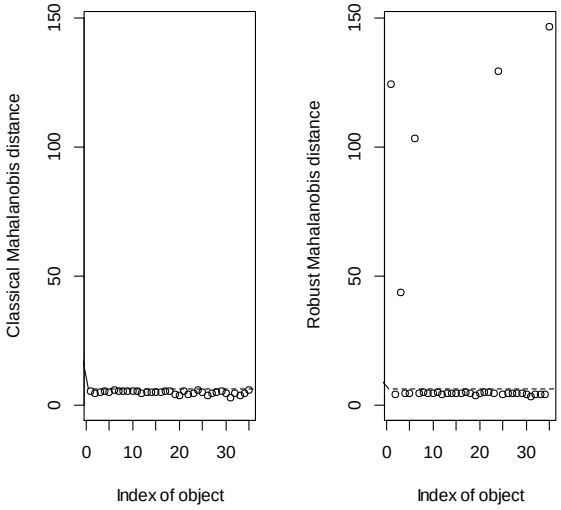


**Figure 6. Plots show the MDs (left) and RDs (right) to Negro river samples versus the object number. The horizontal line represents cutoff line (5 samples above cutoff line to RDs).**

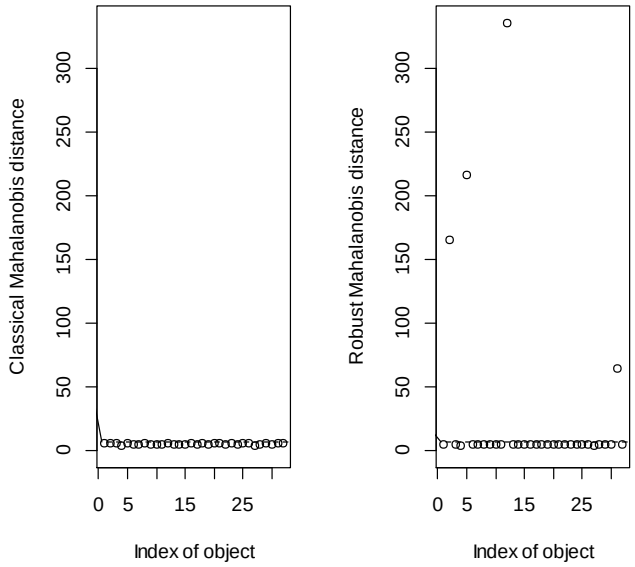


**Figure 7. Plots show the MDs (left) and RDs (right) to Paraná do Ariaú samples versus the object number. The horizontal line represents cutoff line (4 samples above cutoff line to RDs).**

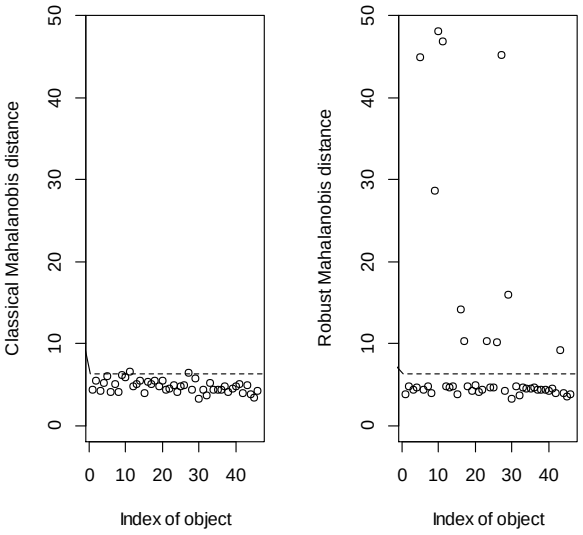


**Figure 8. Plots show the MDs (left) and RDs (right) to Solimões river samples versus the object number. The horizontal line represents cutoff line (5 samples above cutoff line to RDs).**

In order to confirm compositional groups, LDA was used to sample groups previously defined according to the individualized Pleistocene stratigraphic units in the confluence zone, related floodplains of the Negro and Solimões rivers and tectonic depressions. Figure 9 presents a bivariate plot of discriminant functions showing some overlap between samples

from Solimões River and Paraná do Ariaú Graben (GPA), suggesting similarities of these samples according to chemical composition. However, in Figure 9 can be seen a separation between samples from Negro river and GPA. Thus, these results suggest an indicative of the strong influence of sediments from Solimões River in the filling in GPA, since its formation, while the separation between groups RN and GPA can be interpreted that the sedimentary input from Negro in GPA is insignificant.

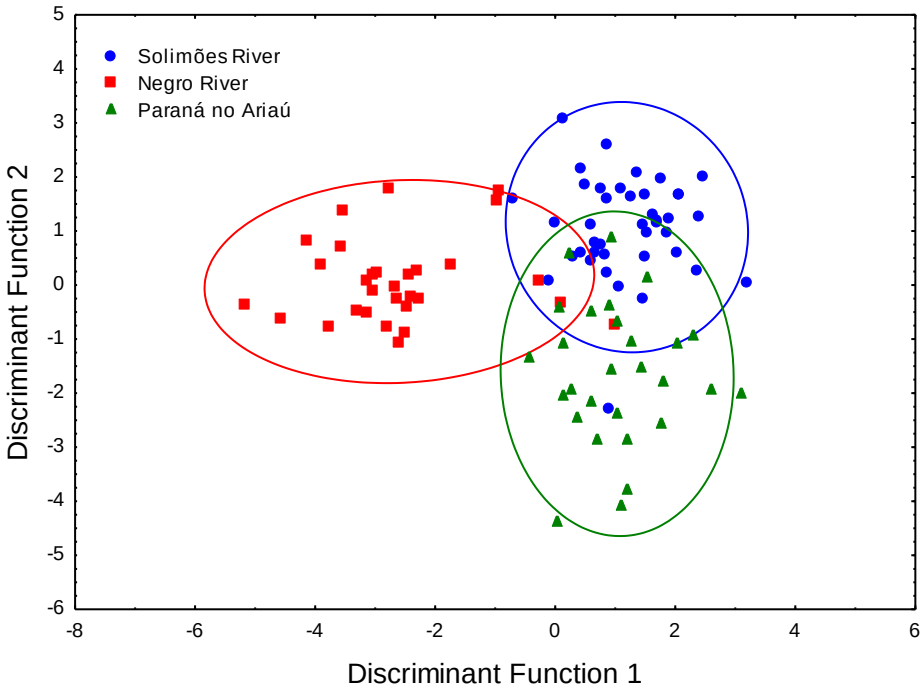


**Figure 9. Linear discriminant analysis of sediments samples from GPA and Negro. Ellipses represent a 95% confidence level**

## 4. CONCLUSIONS

A method to identify outliers in multivariate space by means of MCD was applied to data set obtained from INAA of samples collected from Amazon. The results showed that to multivariate outlier detection it is necessary consider the shape and structure of data set, since the Robust Mahalanobis distance estimator by Minimum Covariance Determinant was able to detect atypical samples that the Classical Mahalanobis distance has not been able to identify. According to statistical interpretation to INAA results it was possible to infer that the elemental chemical composition of samples from Solimões River and GPA are not significantly different and samples from Negro river and GPA are distinct, which indicate the strong influence of sediment supply from Solimões River in the filling of GPA. The results provided information about the fluvial dynamic of confluence zone of Negro and Solimões Rivers contributing with subsidies to reconstruction of the geological evolution history of Amazon basin.

## ACKNOWLEDGMENTS

We thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for parcial sponsoring this research.

## REFERENCES

1. C. Reimann, P. Filzmoser, "Normal and lognormal data distribution in geochemistry: death of myth. Consequences for the statistical treatment of geochemical and environmental data", *Earth and Environmental Science*, **39 (9)**, pp.1101-1014 (1999).
2. P. Filzmoser, R. G. Garret, R., C. Reimann. "Multivariate outlier detection in exploration geochemistry", *Computers & Geosciences*, **31**, pp. 579-587, 2005.
3. J. O.Santos, C.S Munita, R.G. Toyota, C. Vergne, R.S. Silva, P. M. S. Oliveira, "The archaeometry study of the chemical and mineral composition of pottery from Brazil Northeast", *Journal of Radionalytical and Nuclear Chemistry*, **281**, p. 189-192, 2009.
4. A. Hadi. "Identifying multiple outliers in multivariate data", *Journal Royal Statistics Society*, **54**, p. 761-771, 1992.
5. P. J. Rousseeuw, B. C. Van Zomeren. "Unmasking multivariate outliers and leverage points", *Journal of American Statistical Association*, **85 (411)**, pp. 633-651, 1990.
6. I. T. Jolliffe, B. Jones, J. T. Morgan. "Identifying influential observations in hierarchical cluster analysis", *Journal of Applied Statistics*, **22 (1)**, 1995.
7. J. L. Guyot, J. M. Jouanneau, L. Soares, G. R. Boaventura, N. Maillet, C. Lagane, "Clay mineral composition of river sediments in the Amazon Basin", *Catena*, **71**, pp.340-356 (2007).
8. J. L. Gaillardet, B. Dupré, C. J. Allègre, P. Négrel, "Chemical and physical denudation in the Amazon River Basin, *Chemical Geology*, **142**, pp.141-173 (1997).
9. E. A. A. Soares, S. H. Tatumi, C. Riccomini, "OSL age determinations of Pleistocene fluvial deposits in Central Amazonia", *Anais da Academia Brasileira de Ciências*, **82(3)**, pp.691-699 (2010).
10. E. A. A., Soares. Depósitos pleistocenos da região de confluência dos rios Negro e Solimões, porção oeste da Bacia do Amazonas, PhD Thesis – Instituto de Geociências, Universidade de São Paulo, São Paulo. 2007.
11. R. Ravisankar, E. Manikandan, M. Dheenathayalu, B. Rao, N. P. Seshadreesan, K. G. M. Nair, "Determination and distribution of rare earth elements in beach rock samples using instrumental neutron activation analysis", *Nuclear Instruments and Methods in Physics Research B*, **251**, pp.496-500 (2006).
12. M. A. Phedorin, V. A. Bobrov, E. L. Goldberg, J. Navez, K. V. Zolotaryov, M. A. Grachev, "SR-XRA as method of choice in the search of signals of changing palaeoclimates in the sediments of Lake BaiKal, compared to INAA and ICP-MS", *Nuclear Instruments and Methods in Physics Research A*, **448**, pp.394-399 (2000).
13. A. Laraque, J. L. Guyot, N. Filizola, "Mixing processes in the Amazon River at confluences of the Negro and Solimões Rivers, Encontro das Águas, Manaus, Brazil", *Hidrological Processes*, **23**, pp.3131-3140 (2009).
14. J. A. Cogswell, H. Neff, M. D. Glascock, "The effect of firing temperature on the elemental characterization of pottery", *Journal of Archaeological Science*, **23**, pp. 283-287 (1996).
15. A. Schwedt, H. Mommsen, " On the influence of drying and firing of clay on the formation of trace element concentration profiles within pottery", *Archaeometry*¸**49**, pp 495-509 (2007).

16. K. Varmuza,vP. Filzmoser. “Introduction to Multivariate Statistical Analysis in Chemometrics”, CRC Press, Boca Raton, FL, 2009.
17. J. Hardin, D. M. Rocke. “Outlier detection in multiple cluster setting using the minimum covariance determinant estimator”, *Computational Statistics & Data Analysis*, **44**, pp. 625-638, 2004.
18. R development core team, 2013, R: A language and environment for statistical computing: Vienna, http://www.r-project.org.
19. P. Filzmoser, K. Hron, and C. Reimann. “Interpretation of multivariate outliers for compositional data”, *Computers & Geosciences*, **39**, pp. 77-85, 2012.
20. C. S. Munita, R. P. Paiva, M. A. Alves, P. M. S. Oliveira, E. F. Momose, “Contribution of neutron activation analysis to archaeological studies”, *Journal of Microprobe Techniques*, **18**, pp.381-387 (2000).
21. C. Reimann, P. Filzmoser, “Normal and lognormal data distribution in geochemistry: death of myth. Consequences for the statistical treatment of geochemical and environmental data”, *Earth and Environmental Science*, **39 (9)**, pp.1101-1014 (1999).