2017 International Nuclear Atlantic Conference - INAC 2017
Belo Horizonte, MG, Brazil, October 22-27, 2017
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR – ABEN

Agência Brasileira do ISBN
ISBN 978-85-99141-07-6
9 788599 141076

VALIDITY STUDIES AMONG HIERARCHICAL METHODS OF CLUSTER ANALYSIS USING COPHENETIC CORRELATION COEFFICIENT

Priscilla R. Carvalho[1], Casimiro S. Munita[1] and André L. Lapolli[1]

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

ABSTRACT

The literature presents many methods for partitioning of data base, and is difficult choose which is the most suitable, since the various combinations of methods based on different measures of dissimilarity can lead to different patterns of grouping and false interpretations. Nevertheless, little effort has been expended in evaluating these methods empirically using an archaeological data base. In this way, the objective of this work is make a comparative study of the different cluster analysis methods and identify which is the most appropriate. For this, the study was carried out using a data base of the Archaeometric Studies Group from IPEN-CNEN/SP, in which 45 samples of ceramic fragments from three archaeological sites were analyzed by instrumental neutron activation analysis (INAA) which were determinated the mass fraction of 13 elements (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, U). The methods used for this study were: single linkage, complete linkage, average linkage, centroid and Ward. The validation was done using the cophenetic correlation coefficient and comparing these values the average linkage method obtained better results. A script of the statistical program R with some functions was created to obtain the cophenetic correlation. By means of these values was possible to choose the most appropriate method to be used in the data base.

1. INTRODUCTION

In the last years cluster analysis have increasing emphasis in multivariate data analysis. However, clustering techniques are tools where the application and interpretation are subjective, depending on the experience and perspicacity of the user [1]. Different clustering methods produce different results when applied to the same data [2]. Nevertheless, little effort has been expended in evaluating these methods empirically using an archaeological data base.

In archaeological studies several analytical techniques are used to study the chemical and mineralogical composition of many materials of archaeological origin, generating a large data base. Thus, the multivariate statistical methods become indispensable for the interpretation of the results.

These multivariate techniques, unsupervised and supervised, are accompanied by modern computational programs, which provide visualization and interpretation. Several methods have been used, such as cluster analysis, discriminant analysis, principal component analysis, among others. However, the most used is cluster analysis [3]. The purpose of cluster analysis

is to group the samples based on similarity or dissimilarity [4]. The groups are determined in order to obtain homogeneity within the groups and heterogeneity between them [5].

The literature presents many methods for partitioning of data base [2, 5, 6, 7, 8] and to choose which is the most suitable is difficult, since the various combinations of methods based on different measures of dissimilarity can lead to different patterns of grouping and false interpretations [2].

In this way, the objective of this work is make a validation study of the different methods of cluster analysis and to identify which is the most appropriate in archaeological data base. This study was accomplished using a data base of the Archaeometric Studies Group from IPEN-CNEN/SP, of 45 ceramic fragment samples from three archaeological sites, named A, B and C, in which they were analyzed by Instrumental Neutron Activation Analysis (INAA) to determine the mass fractions of 13 chemical elements: As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th and U.

The methods used for this study were: Single Linkage, Complete Linkage, Average Linkage, Centroid and Ward. The validation was done using the cophenetic correlation coefficient, whose purpose is to analyze the quality of the grouping generated by the hierarchical methods of cluster analysis, as well as a criterion for evaluate the efficiency of the various grouping techniques [9].

In addition, taking into account the existence of several statistical programs and even the complexity of certain programs, a script of the statistical program R with some functions was created to obtain the cophenetic correlation coefficient. Thus, the identification of the most appropriate method to be used in the study is faster.

2. DEVELOPMENT

2.1. Cluster Analysis

Cluster analysis is a statistical technique of interdependence whose primary purpose is to group the samples based on similarity or dissimilarity [4] from predetermined variables. The groups are formed so that each sample is similar to the others in the grouping, thus seeking to minimize the variance within the group and to maximize the variance between the groups, that is, to maximize the homogeneity within the groups and the heterogeneity among them [5]. Thus, if the classification is successful, the objects within the groupings will be close together when represented graphically and different groupings will be distant.

For this, the samples are initially treated individually and then analyzed in a correlation matrix, or similarity/dissimilarity matrix of the samples, where sample-sample, sample-group and group-group distances are calculated successively, until the formation of a single group. In general, the smaller the distance between the samples, the greater their similarities.

Thus, it can be said that the clustering process basically involves two stages: the first relates to the estimation of a measure of similarity (or dissimilarity) between the sample units; and the second, with the adoption of a grouping technique for group formation.

The distances are the measures of dissimilarity most used in the study of data base with quantitative variables. A large number of measures of dissimilarity have been proposed and used in cluster analysis [2, 7]. Among these, those chosen to perform the work were the distances: Euclidean, Squared Euclidean, Manhattan (or City-Block) and Mahalanobis. Once the metric is chosen, the second step is to choose which clustering algorithm will be used to form the groups.

In the literature, several methods of grouping are found [2, 5, 6, 7, 8], and the researcher has to decide which is most suitable for its purpose. Most methods can be classified into two large families of methods: hierarchical and non-hierarchical. In this work, will be studied the hierarchical algomerative methods (Single Linkage, Complete Linkage, Average Linkage, Centroid and Ward).

2.1.1 Single linkage method

The single linkage method is one of the oldest methods, its origins being traced to polish researchers in the 1950s [10]. It was first described by Florek *et al.* [11] and later by Sneath [12] and Johnson [13]. The defining feature of the method is that the distance between groups is defined as that of the closest pair of samples, where only pairs consisting of one sample from each group are considered [8].

2.1.2 Complete linkage method

The complete linkage method is similar to the single linkage method except that the distance between two clusters is now defined as the largest distance between pairs of samples in each cluster, rather than the smallest [14].

2.1.3 Average linkage method

In average linkage – also known as the unweighted pair-group method using the average approach (UPGMA) – the distance between two clusters is the average of the distance between all pairs of samples that are made up of one sample from each group [8]

All these three methods (single, complete and average) use a proximity matrix as input, and the inter-cluster distances they use are each illustrated graphically in Fig. 1.

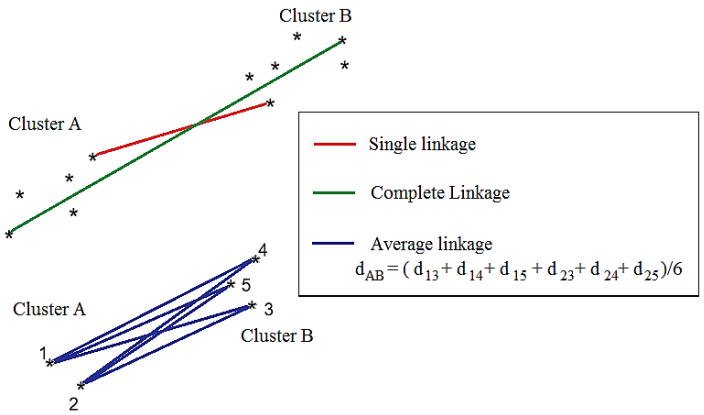

Figure 1: Examples of three inter-cluster distance measures: single, complete and average [8].

2.1.4 Centroid's method

In centroid's method the dissimilarity of two clusters is expressed as the distance of centroids of these clusters. Each cluster is represented by the average of its samples, which is called the centroid. The distance between clusters is determined by the Lance-William correlation:

$$d(C_1, C_2 \cup C_3) = \frac{n_2}{n_2 + n_3} d(C_1, C_2) + \frac{n_3}{n_2 + n_3} d(C_1, C_3) - \frac{n_2 n_3}{n_2 + n_3} d(C_2, C_3) \quad (1)$$

where n_1, n_2 and n_3 are the number of samples in clusters C_1, C_2 and C_3 [4, 10, 15].

2.1.5 Ward's method

Ward's Method was proposed by Ward in 1963 [16] and is also called "Minimum Variance" [2]. In this method, the fusion of two clusters is based on the size of an error sum-of-squares criterion [8], in order to maximize the internal homogeneity of the groups [4]. The distance between clusters is determined by the Lance-William correlation:

$$d(C_1, C_2 \cup C_3) = \frac{n_1 + n_2}{n_1 + n_2 + n_3} d(C_1, C_2) + \frac{n_1 + n_3}{n_1 + n_2 + n_3} d(C_1, C_3) - \frac{n_1}{n_1 + n_2 + n_3} d(C_2, C_3) \quad (2)$$

where n_1, n_2 and n_3 are the number of samples in clusters C_1, C_2 and C_3 [4, 10, 15].

2.2. Cophenetic Correlation Coefficient

After applying the method chosen for the formation of groups the cophenetic correlation coefficient (CCC) has been used to verify the quality of the grouping. Since its introduction by Sokal and Rohlf [17], the CCC (3) has been widely used in studies, both as a measure of the degree of fit of a classification of a data base and as a criterion for evaluating the efficiency of various clustering techniques [9].

$$CCC = \frac{\sum_{i=1}^{n-1} \sum_{k=i+1}^{n} (c_{ik} - \bar{c})(d_{ik} - \bar{d})}{\sqrt{\sum_{i=1}^{n-1} \sum_{k=i+1}^{n} (c_{ik} - \bar{c})^2} \sqrt{\sum_{i=1}^{n-1} \sum_{k=i+1}^{n} (d_{ik} - \bar{d})^2}} \quad (3)$$

Where:

c_{ik} = dissimilarity value between samples i and k, obtained from the cophenetic matrix;

d_{ik} = dissimilarity value between samples i and k, obtained from the dissimilarity matrix.

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} c_{ik} \quad (4)$$

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} d_{ik} \quad (5)$$

The cophenetic correlation coefficient consists in comparing the observed distances between the samples and the distances predicted from a grouping process [6, 15], by measuring the degree of fit between the original dissimilarity matrix, matrix D, and the matrix resulting from the simplification provided by the clustering method, matrix C [15].

In this work, the cophenetic correlation coefficient was used to validate the methods and find the most suitable for the data base studied.

2.3. Script

The statistical study was performed using the statistical program R. The R is a programming environment with an integrated set of software tools for data manipulation, calculations and graphical presentation [18]. The structure is open source and the software is public and free, so R has been widely accepted by researchers around the world. However, by using programming language, the R, requires the user a brief programming knowledge.

In this way, a script with functions of the statistical program R was developed to calculate the cophenetic correlation coefficient for identification of the hierarchical method of cluster analysis more suitable to be in a data base. The purpose of this guide is to facilitate the study of researchers who are not from the statistical area or are not familiar with the program.

3. RESULTS AND DISCUSSION

The study was made using a data base of 45 ceramic fragment samples which were determined As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U by INAA. Table 1 shows the values of the mass fractions.

Initially, the results were transformed to $\log_{10}$. This transformation before applying multivariate statistical techniques is a usual procedure in archaeometric studies and there are two reasons for this: the first is explained by the fact that a normal logarithmical distribution of the elements exists. The other is the difference magnitude between elements, which it was found in percentage and trace level [19].

After this, the detection of the outliers was done by means of Mahalanobis distance using the lambda Wilks criterion as critical value [20]. In this outlier detection method, when the calculated value for the Mahalanobis distance is greater than the critical value, the sample is considered outlier. For this data base, no outliers were detected.

Table 1: Results of the elementary concentrations in mg/kg of samples of ceramic fragments.

Sample	Site	As	Ce	Cr	Eu	Fe	Hf	La	Na	Nd	Sc	Sm	Th	U
1	A	1.80	117.50	175.00	1.01	1730.00	10.00	38.50	786.00	57.00	26.69	7.75	19.20	4.50
2	A	1.60	137.20	186.00	1.28	1720.00	11.00	38.90	727.00	45.00	26.96	8.07	19.50	4.70
3	A	2.50	113.40	123.00	1.51	3810.00	8.80	31.50	302.00	35.00	31.51	7.74	17.80	4.60
4	A	1.80	105.40	142.00	1.16	2660.00	9.30	27.20	543.00	26.00	27.91	6.35	16.40	3.30
5	A	1.80	108.20	157.00	1.26	3070.00	9.20	29.30	552.00	36.00	31.40	6.75	17.90	6.30
6	A	1.80	117.60	156.00	1.40	2980.00	8.80	33.00	590.00	32.00	30.16	7.43	18.70	3.50
7	A	1.40	120.90	152.00	1.42	2960.00	9.00	33.50	621.00	39.00	30.37	7.76	18.50	5.40
8	A	1.80	113.50	170.00	1.27	2990.00	9.50	30.00	635.00	27.00	31.29	7.00	17.20	4.30
9	A	1.40	102.90	114.00	1.36	3610.00	8.70	40.40	644.00	38.00	27.64	7.84	17.00	4.30
10	A	1.20	113.20	138.00	1.33	2800.00	8.50	31.40	557.00	29.00	28.62	7.02	15.80	4.80
11	A	1.46	104.00	136.00	1.30	2630.00	8.40	29.33	579.00	38.00	27.63	6.83	16.00	3.50
12	A	1.60	115.40	124.00	1.68	3840.00	8.40	30.40	328.00	43.00	32.48	7.43	17.70	3.90
13	A	1.70	120.30	115.00	1.70	3600.00	9.00	32.60	377.00	40.00	30.72	8.09	16.60	4.90
14	A	2.10	121.00	121.00	1.61	3730.00	9.10	33.50	493.00	34.00	31.80	6.63	17.60	5.20
15	A	1.80	131.00	140.00	1.64	2650.00	8.90	35.30	593.00	46.00	29.07	6.50	16.50	5.00
16	B	1.50	108.30	134.20	2.52	3200.00	7.82	64.10	196.10	63.00	12.87	8.89	9.81	1.30
17	B	2.70	122.30	133.00	2.57	3860.00	6.30	83.40	148.70	64.00	15.23	10.14	12.60	0.99
18	B	2.00	111.90	138.00	2.31	3780.00	8.40	62.70	225.40	49.00	12.60	8.43	12.10	0.90
19	B	1.20	125.60	150.00	2.67	3440.00	9.30	83.40	161.70	51.00	17.24	11.34	13.50	1.30
20	B	3.90	123.80	175.00	2.65	4390.00	9.10	72.50	225.40	63.00	16.78	10.17	15.00	1.30
21	B	2.50	160.30	183.00	3.79	3880.00	7.60	96.80	261.30	68.00	18.04	13.10	14.20	1.20
22	B	3.30	123.40	151.00	2.61	4080.00	7.80	66.80	170.20	54.00	16.26	9.04	14.00	0.99
23	B	1.50	104.60	135.00	2.12	2450.00	9.20	60.70	101.50	46.00	14.87	8.16	13.70	1.30
24	B	2.30	105.10	142.50	2.09	2230.00	8.50	62.50	125.00	61.00	14.44	8.83	15.00	1.60
25	B	1.60	104.50	150.00	2.42	3090.00	7.70	61.80	243.70	47.00	12.82	8.73	11.00	1.28
26	B	1.90	85.50	147.00	2.33	2880.00	10.40	61.50	148.00	44.00	14.02	9.28	11.70	1.60
27	B	1.80	121.60	160.00	2.55	2930.00	8.60	72.40	171.20	63.00	16.41	9.88	11.10	1.20
28	B	1.80	138.50	192.00	2.67	3210.00	9.30	78.20	218.30	57.00	19.71	10.54	15.50	1.70
29	B	2.00	131.90	169.00	2.98	3490.00	9.30	77.60	103.70	60.00	17.77	10.34	14.40	1.70
30	B	3.00	127.30	166.00	2.63	4100.00	9.90	80.90	222.30	72.00	16.99	11.16	14.00	1.20
31	C	2.60	67.80	212.00	2.94	11270.00	10.80	31.80	132.00	41.00	39.90	9.43	6.40	1.30
32	C	1.70	75.80	205.00	2.94	8550.00	12.50	31.80	121.00	45.00	41.75	8.98	6.90	1.60
33	C	1.60	56.40	183.00	2.39	8160.00	10.80	28.00	120.00	35.00	43.40	7.45	6.40	1.50
34	C	2.20	62.50	195.00	2.82	9130.00	11.30	29.30	92.00	46.00	42.46	9.21	7.10	1.30
35	C	1.50	90.80	303.00	3.20	12120.00	11.00	39.50	266.00	52.00	41.72	10.21	5.60	1.10
36	C	1.80	101.50	230.00	3.40	13960.00	11.70	45.50	144.00	51.00	45.00	11.43	7.70	1.30
37	C	1.20	63.40	183.00	2.85	9830.00	10.50	33.90	130.00	44.00	40.71	9.57	6.70	1.70
38	C	2.70	67.80	236.00	3.02	11000.00	11.00	33.80	139.00	55.00	41.16	9.99	6.30	1.40
39	C	1.90	109.70	218.00	3.29	7580.00	11.70	37.80	181.00	60.00	39.36	10.31	5.20	1.10
40	C	1.60	78.90	230.00	3.20	8600.00	10.90	41.10	189.00	69.00	40.01	11.33	5.10	1.10
41	C	2.50	54.50	203.00	2.95	12590.00	10.90	34.10	138.00	44.00	44.70	9.61	6.79	1.20
42	C	1.40	70.90	192.00	3.00	8320.00	11.90	36.10	117.00	61.00	46.10	10.31	7.40	1.50
43	C	2.40	123.20	224.00	4.31	9160.00	12.80	51.50	176.00	58.00	47.80	14.04	7.40	1.60
44	C	1.80	97.50	238.00	3.27	8030.00	11.90	38.00	167.00	52.00	42.30	10.36	6.20	1.80
45	C	1.80	92.70	253.00	3.60	14940.00	12.80	44.20	125.00	63.00	48.30	11.70	6.40	1.20

Posteriorly the detection of the outliers, the 45 ceramic samples results were submitted to cluster analysis using the methods: Single Linkage, Complete Linkage, Average Linkage, Centroid and Ward with distances: Euclidean, Squared Euclidean, Manhattan and Mahalanobis.

The results of the hierarchical methods are summarized until a dendrogram is established, this being a two-dimensional diagram in the form of a tree illustrating the fusions performed at each successive level, in which the abscissa axis represents the samples and the axis of the ordinates the distances obtained after the use of a grouping methodology.

In general, the dendrograms generated by the different methods formed three well-defined groups. For the Euclidean, Squared Euclidean and Manhattan distances, the groups formed are the same and consist of samples from the same archaeological site. The same does not happen when clustering methods were associated with distance Mahalanobis. At this distance the formed groups end up mixing samples from different sites, which leads to false interpretations. To illustrate this fact, were chosen two dendrograms represented in Fig. 2 and Fig. 3.

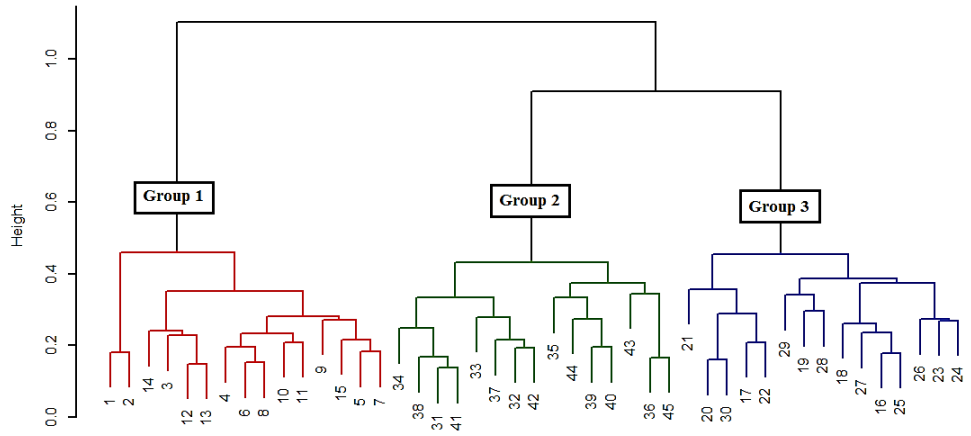

Figure 2: Dendrogram of the ceramics sample using Euclidean distance and Average Linkage method.

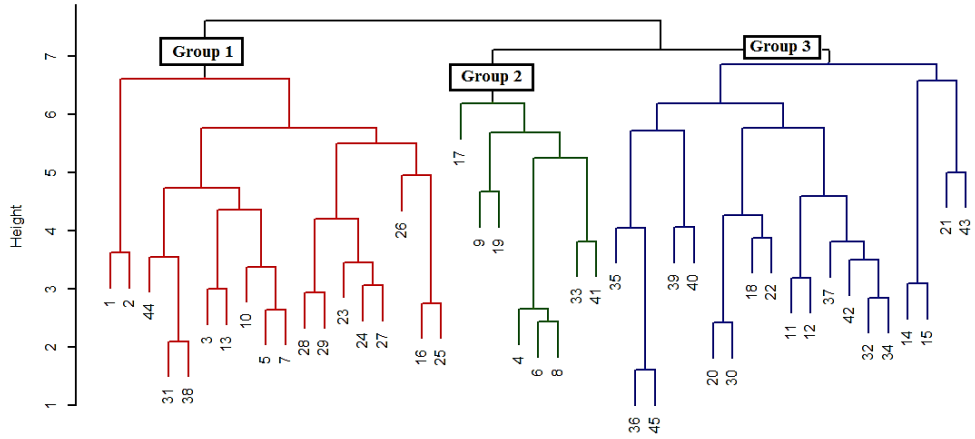

Figure 3: Dendrogram of the ceramics sample using Mahalanobis distance and Complete Linkage method.

To validate and compare the clustering methods, we estimated the cophenetic correlation coefficient (CCC), which measures the degree of fit between the original dissimilarity matrix and the resulting matrix of simplification provided by the clustering method. Thus, the closer to 1 is the CCC, the better the grouping quality [6, 7]. According to Rohlf [21], in practice dendrograms with CCC less than 0.7 would indicate the inadequacy of the grouping method to summarize the data base information. These values are represented in Table 2.

Table 2: The cophenetic correlation coefficient values

Distance measure	Clustering method	Cophenetic correlation coefficient
Euclidean	Single	0.914515
	Complete	0.928495
	Average	**0.940344**
	Centroid	0.934583
	Ward	0.930007
Squared Euclidean	Single	0.839967
	Complete	0.873087
	Average	**0.885749**
	Centroid	0.883074
	Ward	0.881126
Manhattan	Single	0.915449
	Complete	0.924147
	Average	**0.929580**
	Centroid	0.923110
	Ward	0.918241
Mahalanobis	Single	0.628300
	Complete	0.397273
	Average	**0.692614**
	Centroid	0.665084
	Ward	0.404547

Thus, since the CCC value for the dendrogram of Fig. 3 is 0.397273, this explain the fact that it generated false groupings. Even comparing the CCC values, it can be observed that regardless of the distance metric used, the Average Linkage method obtained better results, which corroborates with the literature [9, 22, 23].

Finally, to facilitate the statistical study of researchers who do not have much familiarity with statistical programs, the script developed becomes very useful, since it is enough to just insert the data base in the statistical program R and to execute it thus obtaining a table with all the cophenetic correlation values. This way, the researcher can easily check which method and distance is most appropriate for your data base. The Fig. 4 shows the table generated by the script developed in this work.

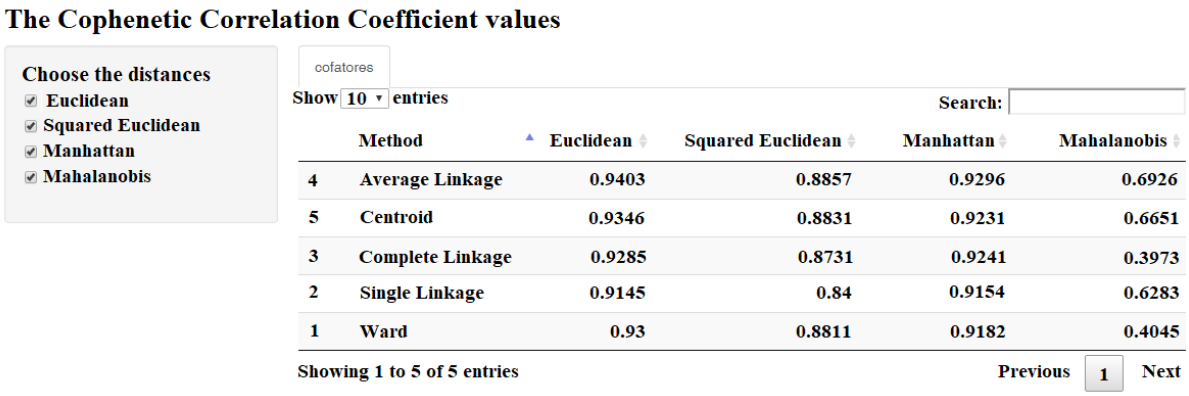

Figure 4: Table generated by the script developed.

4. CONCLUSIONS

Several types of clustering methods are found in the literature, with the researcher deciding which is most suitable for their purpose, since the various combinations of methods based on different measures of dissimilarity can lead to different patterns of grouping. This work aimed to compare the methods and point the most appropriate to the data base used. The results show that the method Average linkage was the one which has the best cophenetic correlation coefficient result. In addition, the script developed will help researchers to find the most appropriate grouping method for their data base.

ACKNOWLEDGMENTS

The author thanks the CAPES/PROEX for the financial support.

REFERENCES

1. L. P. Fávero; P. Belfiore; F. L. Silva; B. L. Chan, *Análise de dados: modelagem multivariada para tomada de decisões*, Elsevier, Rio de Janeiro, Brazil (2009).
2. S. A. Mingoti, *Análise de dados através de métodos estatísticos multivariada: uma abordagem aplicada*, Editora UFMG, Belo Horizonte, Brazil (2005).
3. J. Papageorgiou; M. J. Baxter, "Model-based cluster analysis of artefact compositional data", *Archaeometry*, **43(4)**, p. 571-588 (2001).
4. P. Trebuna; J. Halcinová, "Mathematical tools of cluster analysis", *Applied Mathematics*, **4**, pp.814-816 (2013).
5. J. F. Jr. Hair; R. E. Anderson; R. L. Tatham; C. Black, *Análise multivariada de dados*, Bookman, Porto Alegre, Brazil (2005).
6. 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).
7. W. O. Bussab; E. S. Miazaki; D. F. Andrade, *Introdução à análise de agrupamentos*, ABE, São Paulo, Brazil (1990).
8. B. S. Everitt; S. Landau; M. Leese; D. Stahl, *Cluster analysis*, Edward, London (2011).

9. S. Saraçli; N. Dogan; I. Dogan, “Comparison of hierarchical cluster analysis methods by cophenetic correlation”, *J. Inequalities and Applications*, **203** (2013).
10. F. Murtagh; P. Contreras, “Methods of Hierarchical Clustering”, Data Mining and Knowledge Discovery, Wiley-Interscience, **2(1)**, pp.86-97 (2012).
11. K. Florek; L. Lukaszewiez; L. Perkal et al, “Sur la liaison et la division des points d’un ensemble fini, Colloquium Mathematicum”, **2**, pp.282-285 (1951).
12. P. H. A. Sneath, “The application of computers to taxonomy”, *J. General Microbiology*, **17**, pp.201-226 (1957).
13. S. C. Johnson, “Hierarchical clustering schemes”, *Psychometrika*, **32**, pp.241–254 (1967).
14. K. V. Mardia; J. T. Kent; J. M. Bibby, Multivariate Analysis, Academic Press, London (1989).
15. M. A. Albuquerque, *Estabilidade em análise de agrupamento (Cluster Analysis)*, Dissertação, UFRPE (2005).
16. J. H. Ward, “Hierarchical grouping to optimize an objective function”, *J. Applied Statistics*, **58**, pp.236-244 (1963).
17. R. R. Sokal; F. J. Rohlf, “The comparison of dendrograms by objective methods”, *Taxon*, **11**, pp.33-40 (1962).
18. W. N. Venables; D. M. Smith; The R Core Team, “An introduction to R”, https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf (2017).
19. P. M. S. Oliveira, C. S. Munita, “Influência do Valor Crítico na Detecção de Valores Discrepantes em Arqueometria”, *48ª Reunião Anual Região Brasileira da Sociedade Internacional de Biometria*, Lavras, MG, Brazil, 07-11 de julho (2003).
20. P. M. S. Oliveira; C. S. Munita; R. Hazenfratz, “Comparative study between three methods of outlying detection on experimental results”, *J. Radioanalytical and Nuclear Chemistry*, **283**, pp.433-437 (2010).
21. F. J. Rohlf, “Adaptative hierarquical clustering schemes”, *Systematic Zoology*, **19(1)**, pp.58-82 (1970).
22. F. K. Kuiper; L. A. Fisher, “A Monte Carlo comparison of six clustering procedures”, *Biometrics*, **31**, pp.777-783 (1975).
23. G. W. Milligan; M. C. Cooper, “A study of standardization of variables in cluster analysis”, *J. Classif.*, **5**, pp.181-204 (1988).