Avaliação de análise de agrupamentos pelo coeficiente de correlação cofenética: estudo preliminar

Priscilla R. Carvalho*, Casimiro S. Munita,

Instituto de Pesquisas Energéticas e Nucleares, IPEN – CNEN/SP

*e-mail: prii.ramos@usp.br

Introdução

Tendo em vista a facilidade de se armazenar dados e o crescente avanço das técnicas analíticas em estudos de diversas áreas do conhecimento, a quantidade de resultados gerados tem aumentado significativamente. Para a interpretação desses resultados, faz-se necessário o uso de métodos estatísticos cada vez mais sofisticados, tais como as técnicas multivariadas. Nas diversas áreas do conhecimento uma das técnicas mais utilizadas é análise de agrupamentos [1]. O seu emprego em áreas tais como agricultura, medicina, marketing, administração, arqueometria, entre outras, vem aumentando nos últimos anos.

A análise de agrupamentos, também conhecida como *cluster analysis*, tem por finalidade primária agrupar as amostras com base na similaridade ou dissimilaridade [2]. Os grupos são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles.

As distâncias são as medidas de dissimilaridade mais utilizadas no estudo de banco de dados com variáveis quantitativas. Um grande número de medidas de dissimilaridade tem sido proposto e utilizado em análise de agrupamentos e vários são os tipos de métodos de agrupamentos encontrados na literatura [3, 4, 5], tendo o pesquisador que tomar a decisão de qual é o mais adequado ao seu propósito, uma vez que, diversos métodos com base em diferentes medidas de dissimilaridade podem levar a distintos padrões de agrupamento [1,3].

Com base nesse cenário, o propósito deste trabalho é realizar um estudo comparativo do uso de diferentes medidas de distância no método de Ward, utilizando para essa comparação o dendrograma e o coeficiente de correlação cofenética (CCC). O estudo foi realizado usando uma base de dados do Grupo de Estudos Arqueométricos do IPEN-CNEN/SP, na qual foram analisadas 45 amostras de fragmentos cerâmicos de três sítios arqueológicos nos que foram determinados, por análise de ativação com nêutrons instrumental (INAA), 13 elementos (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, U).

Análise de Agrupamentos

A análise de agrupamentos é uma técnica estatística multivariada que foi originalmente desenvolvida para classificação biológica [6]. Uma importante contribuição ao desenvolvimento das técnicas de agrupamento foi feita a partir do livro “Principles of Numerical Taxonomy” de Sokal e Sneath [7]. Sokal e Sneath reuniram as principais publicações em áreas biológicas e lançaram os fundamentos para um estudo em bases numéricas, demonstrando que a análise de agrupamento pode ser utilizada de forma eficiente para a classificação de um conjunto de dados [6].

Análise de agrupamentos é uma denominação genérica para um grande grupo de técnicas que podem ser utilizadas para criar uma classificação. Para Hair et al [8], a análise de agrupamentos agrupa indivíduos ou objetos em grupos de modo que objetos em um mesmo grupo são mais parecidos entre si do que em relação a outros grupos. É nesse sentido que o principal objetivo da análise de agrupamentos é agrupar casos a partir de determinadas características que os tornam similares. Para tanto, esta técnica procura não só minimizar a variância dentro do grupo, mas também maximizar a variância entre os grupos, ou seja, maximizar a homogeneidade dentro dos grupos e a heterogeneidade entre eles.

Para isso, as amostras são inicialmente tratadas de maneira individual e, em seguida, são analisadas em uma matriz de correlação, ou matriz de similaridade/dissimilaridade das amostras, onde são calculadas distâncias amostra-amostra, amostra-grupo e grupo-grupo, sucessivamente, até a formação de um único grupo. De uma forma geral, quanto menor for a distância entre as amostras, maiores são suas similaridades.

Sendo assim, pode-se dizer que processo de agrupamento envolve basicamente duas etapas: a primeira relaciona-se com a estimação de uma medida de similaridade (ou dissimilaridade) entre as unidades amostrais; e a segunda, com a adoção de uma técnica de agrupamento para a formação dos grupos.

As distâncias são as medidas de dissimilaridade mais utilizadas no estudo de banco de dados com variáveis quantitativas. Um grande número de medidas de dissimilaridade tem sido proposto e utilizado em análise de agrupamentos [4, 5]. Entre essas, as escolhidas para realizar o trabalho foram as distâncias: Euclidiana, Euclidiana quadrática, *Manhattan* (ou *City-Block*) e *Mahalanobis*, cujas métricas estão descritas na Tabela 1.

Tabela 1. Métricas utilizadas para calcular as distâncias

Distância	Métrica
Euclidiana	$d_{ik} = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{kj})^2}$
Euclidiana quadrática	$d_{ik} = \sum_{j=1}^{p}(x_{ij} - x_{kj})^2$
Manhattan (ou *City-Block*)	$d_{ik} = \sum_{j=1}^{p}\lvert x_{ij} - x_{kj}\rvert$
Mahalanobis	$d_{ik} = \sqrt{(X_i - X_k)'S^{-1}(X_i - X_k)}$

Onde:

d_{ik} é a distância entre a amostra i e a amostra k, com i, k = 1, 2, 3, ..., n;

x_{ij} e x_{kj} são os valores observados da variável j, j = 1, 2, 3, ..., p, para as amostras i e k;

X_i é o vetor de médias do i'ésimo grupo;

X_k é o vetor de médias do k'ésimo grupo;

S é a matriz de covariância.

Dado um conjunto de *n* amostras e *p* variáveis, a estimação das medidas de similaridade ou dissimilaridade consiste na conversão da matriz das amostras *n* x *p*, em uma matriz quadrada e simétrica de ordem *n*, de similaridades ou dissimilaridades individuais, que são medidas de distância entre pares de amostras. Na posição (i, k) dessa matriz encontra-se a distância entre a i-ésima e a k-ésima amostra.

Escolhida a métrica, as distâncias são calculadas para todos os elementos e armazenadas em uma matriz, chamada de matriz de distâncias ou de dissimilaridades, que é simétrica e com zeros na diagonal principal. Agora o segundo passo é escolher qual algoritmo de agrupamento será utilizado para formação dos grupos.

Na literatura, são encontrados vários métodos de agrupamentos [3, 4, 5], tendo o pesquisador que tomar a decisão de qual é o mais adequado ao seu propósito. A maioria dos métodos pode ser classificada em duas grandes famílias de métodos: hierárquicos e de partição. Neste trabalho será utilizado o método de Ward, que é um dos métodos hierárquicos aglomerativos. Entre os métodos de análise de agrupamento, o método de Ward é um dos mais utilizados por basear-se numa medida com forte apelo estatístico e por gerar grupos que possuem uma alta homogeneidade interna [2].

O método de Ward foi proposto por Ward em 1963 [9] e é também chamado de “Mínima Variância” [5]. Nesse método a formação dos grupos se dá pela maximização da homogeneidade dentro dos grupos. A soma de quadrados dentro dos grupos é usada como medida de homogeneidade. Isto é, o método de Ward tenta minimizar a soma de quadrados dentro do grupo. Os grupos formados em cada passo são resultantes de grupo solução com a menor soma de quadrados.

Assim como nos outros métodos hierárquicos aglomerativos, os resultados do método de Ward são resumidos até que seja estabelecido um dendrograma, sendo este um diagrama bidimensional em forma de árvore ilustrando as fusões efetuadas em cada nível sucessivo, no qual o eixo das abscissas representa as amostras e o eixo das ordenadas as distâncias obtidas após a utilização de uma metodologia de agrupamento.

Os ramos da árvore fornecem a ordem das (n-1) ligações, em que o primeiro nível representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos formem um grupo.

Depois de aplicar o método escolhido para a formação de grupos tem-se utilizado o coeficiente de correlação cofenética (CCC) para verificar a qualidade do agrupamento. Desde a sua introdução por Sokal e Rohlf [10], o CCC (eq. 1.1) tem sido amplamente utilizado em estudos, tanto como uma medida do grau de adequação de uma classificação de um conjunto de dados, como um critério para avaliar a eficiência das várias técnicas de agrupamento [6].

$$CCC = \frac{\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(c_{ik}-\bar{c})(d_{ik}-\bar{d})}{\sqrt{\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(c_{ik}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(d_{ik}-\bar{d})^2}} \qquad (1.1)$$

Onde:

c_{ik} = valor de dissimilaridade entre as amostras i e k, obtidos a partir da matriz cofenética (1.2);

d_{ik} = valor de dissimilaridade entre as amostras i e k, obtidos a partir da matriz de dissimilaridade (1.3);

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} c_{ik} \quad (1.2)$$

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} d_{ik} \quad (1.3)$$

Material e Métodos

Os dados utilizados no presente estudo foram fornecidos pelo Grupo de Estudos Arqueométricos do IPEN-CNEN/SP. A base de dados contém 45 amostras de fragmentos cerâmicos de três sítios arqueológicos, as que foram analisadas por análise de ativação com nêutrons instrumental (INAA), onde foram determinadas as concentrações elementares de As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, U (Tabela 2).

Inicialmente as concentrações elementares foram transformadas em $\log_{10}$ para compensar a diferença de magnitude entre elementos que são determinados em porcentagem e em nível de traços. Em seguida, foram estudadas as amostras discrepantes (*outliers*) por meio da distância de *Mahalanobis* usando com valor crítico o critério lambda *Wilks* [11].

Posteriormente, foi realizada uma classificação por meio da análise de agrupamentos usando o método de Ward e quatro medidas de distância: Euclidiana, Euclidiana quadrática, *Manhattan* e *Mahalanobis.*

Para comparação dos resultados utilizou-se, além do dendrograma, o CCC que mede o grau de ajuste entre a matriz de dissimilaridade original e a matriz resultante da simplificação proporcionada pelo método de agrupamento. Assim quanto mais próximo de 1 for o CCC, melhor será a qualidade do agrupamento [3, 4]. Na Tabela 2 apresentam-se os resultados das 45 amostras.

Resultados & discussão

Inicialmente foi estudada a presença de *outliers*, na base de dados das amostras de fragmentos cerâmicos dos três sítios arqueológicos, por meio da distância de *Mahalanobis* usando com valor crítico o critério lambda *Wilks*. Neste método de detecção de *outlier*, quando o valor calculado para a distância de *Mahalanobis* for superior ao valor crítico de Wilks, a amostra é considerada *outlier*. Para essa base de dados, não foi detectada a presença de *outliers*. Após isso, foram calculadas as distâncias e posteriormente a aplica-

Tabela 2. Resultados das concentrações elementares das amostras de fragmentos cerâmicos em mg/kg.

Amostras	Sítio	As	Ce	Cr	Eu	Fe	Hf	La	Na	Nd	Sc	Sm	Th	U
A01	Sítio A	1.8	117.5	175.0	1.0	17300.0	10.0	38.5	786.0	57.0	26.7	7.8	19.2	4.5
A02	Sítio A	1.6	137.2	186.0	1.3	17200.0	11.0	38.9	727.0	45.0	27.0	8.1	19.5	4.7
A03	Sítio A	2.5	113.4	123.0	1.5	38100.0	8.8	31.5	302.0	35.0	31.5	7.7	17.8	4.6
A04	Sítio A	1.8	105.4	142.0	1.2	26600.0	9.3	27.2	543.0	26.0	27.9	6.4	16.4	3.3
A05	Sítio A	1.8	108.2	157.0	1.3	30700.0	9.2	29.3	552.0	36.0	31.4	6.8	17.9	6.3
A06	Sítio A	1.8	117.6	156.0	1.4	29800.0	8.8	33.0	590.0	32.0	30.2	7.4	18.7	3.5
A07	Sítio A	1.4	120.9	152.0	1.4	29600.0	9.0	33.5	621.0	39.0	30.4	7.8	18.5	5.4
A08	Sítio A	1.8	113.5	170.0	1.3	29900.0	9.5	30.0	635.0	27.0	31.3	7.0	17.2	4.3
A09	Sítio A	1.4	102.9	114.0	1.4	36100.0	8.7	40.4	644.0	38.0	27.6	7.8	17.0	4.3
A10	Sítio A	1.2	113.2	138.0	1.3	28000.0	8.5	31.4	557.0	29.0	28.6	7.0	15.8	4.8
A11	Sítio A	1.5	104.0	136.0	1.3	26300.0	8.4	29.3	579.0	38.0	27.6	6.8	16.0	3.5
A12	Sítio A	1.6	115.4	124.0	1.7	38400.0	8.4	30.4	328.0	43.0	32.5	7.4	17.7	3.9
A13	Sítio A	1.7	120.3	115.0	1.7	36000.0	9.0	32.6	377.0	40.0	30.7	8.1	16.6	4.9
A14	Sítio A	2.1	121.0	121.0	1.6	37300.0	9.1	33.5	493.0	34.0	31.8	6.6	17.6	5.2
A15	Sítio A	1.8	131.0	140.0	1.6	26500.0	8.9	35.3	593.0	46.0	29.1	6.5	16.5	5.0
B01	Sítio B	1.5	108.3	134.2	2.5	32000.0	7.8	64.1	1961.0	63.0	12.9	8.9	9.8	1.3
B02	Sítio B	2.7	122.3	133.0	2.6	38600.0	6.3	83.4	1487.0	64.0	15.2	10.1	12.6	1.0
B03	Sítio B	2.0	111.9	138.0	2.3	37800.0	8.4	62.7	2254.0	49.0	12.6	8.4	12.1	0.9
B04	Sítio B	1.2	125.6	150.0	2.7	34400.0	9.3	83.4	1617.0	51.0	17.2	11.3	13.5	1.3
B05	Sítio B	3.9	123.8	175.0	2.7	43900.0	9.1	72.5	2254.0	63.0	16.8	10.2	15.0	1.3
B06	Sítio B	2.5	160.3	183.0	3.8	38800.0	7.6	96.8	2613.0	68.0	18.0	13.1	14.2	1.2
B07	Sítio B	0.2	114.5	110.0	2.1	28500.0	7.2	66.1	1023.0	50.0	12.5	9.1	11.4	1.3
B08	Sítio B	3.3	123.4	151.0	2.6	40800.0	7.8	66.8	1702.0	54.0	16.3	9.0	14.0	1.0
B09	Sítio B	1.5	104.6	135.0	2.1	24500.0	9.2	60.7	1015.0	46.0	14.9	8.2	13.7	1.3
B10	Sítio B	1.0	127.4	171.0	2.1	18500.0	6.5	57.9	841.0	54.0	17.2	8.1	12.5	1.2
B11	Sítio B	1.6	104.5	150.0	2.4	30900.0	7.7	61.8	2437.0	47.0	12.8	8.7	11.0	1.3
B12	Sítio B	1.9	85.5	147.0	2.3	28800.0	10.4	61.5	1480.0	44.0	14.0	9.3	11.7	1.6
B13	Sítio B	1.8	121.6	160.0	2.6	29300.0	8.6	72.4	1712.0	63.0	16.4	9.9	11.1	1.2
B14	Sítio B	1.8	138.5	192.0	2.7	32100.0	9.3	78.2	2183.0	57.0	19.7	10.5	15.5	1.7
B15	Sítio B	3.0	127.3	166.0	2.6	41000.0	9.9	80.9	2223.0	72.0	17.0	11.2	14.0	1.2
C01	Sítio C	1.7	75.8	205.0	2.9	85500.0	12.5	31.8	121.0	45.0	41.8	9.0	6.9	1.6
C02	Sítio C	1.6	56.4	183.0	2.4	81600.0	10.8	28.0	120.0	35.0	43.4	7.5	6.4	1.5
C03	Sítio C	2.2	62.5	195.0	2.8	91300.0	11.3	29.3	92.0	46.0	42.5	9.2	7.1	1.3
C04	Sítio C	1.5	90.8	303.0	3.2	121200.0	11.0	39.5	266.0	52.0	41.7	10.2	5.6	1.1
C05	Sítio C	1.8	101.5	230.0	3.4	139600.0	11.7	45.5	144.0	51.0	45.0	11.4	7.7	1.3
C06	Sítio C	1.2	63.4	183.0	2.9	98300.0	10.5	33.9	130.0	44.0	40.7	9.6	6.7	1.7
C07	Sítio C	2.7	67.8	236.0	3.0	110000.0	11.0	33.8	139.0	55.0	41.2	10.0	6.3	1.4
C08	Sítio C	1.9	109.7	218.0	3.3	75800.0	11.7	37.8	181.0	60.0	39.4	10.3	5.2	1.1
C09	Sítio C	1.6	78.9	230.0	3.2	86000.0	10.9	41.1	189.0	69.0	40.0	11.3	5.1	1.1
C10	Sítio C	2.5	54.5	203.0	3.0	125900.0	10.9	34.1	138.0	44.0	44.7	9.6	6.8	1.2
C11	Sítio C	1.8	129.0	95.0	3.8	129700.0	14.1	54.1	1339.0	62.0	40.5	12.3	7.0	1.1
C12	Sítio C	3.1	104.0	99.0	3.9	156900.0	11.1	48.9	583.0	65.0	37.8	12.6	6.2	0.8
C13	Sítio C	3.0	134.0	70.0	4.4	132100.0	14.4	56.8	360.0	67.0	45.1	9.2	7.7	1.1
C14	Sítio C	2.4	123.2	224.0	4.3	91600.0	12.8	51.5	176.0	58.0	47.8	14.0	7.4	1.6
C15	Sítio C	1.8	92.7	253.0	3.6	149400.0	12.8	44.2	125.0	63.0	48.3	11.7	6.4	1.2

ção do método. Os resultados obtidos estão apresentados nos dendrogramas da Figura 1, A, B, C e D.

A partir da análise dos dendrogramas, Figuras 1A, 1B e 1C, percebe-se a formação de três grupos bem definidos. Para as distâncias Euclidiana, Euclidiana quadrática e *Manhattan* os grupos formados são os mesmos e constituídos de amostras do mesmo sítio arqueológico:

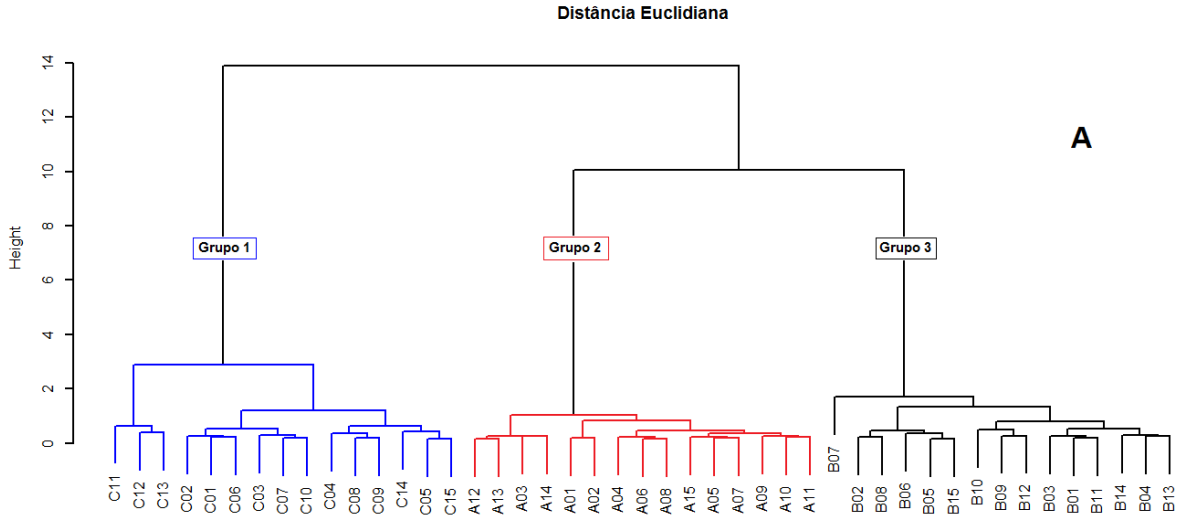
- Amostras do Grupo 1 – C01, C02, C03, C04, C05, C06, C07, C08, C09, C10, C11, C12, C13, C14 e C15;
- Amostras do Grupo 2 – A01, A02, A03, A04, A05, A06, A07, A08, A09, A10, A11, A12, A13, A14 e A15;
- Amostras do Grupo 3 – B01, B02, B03, B04, B05, B06, B07, B08, B09, B10, B11, B12, B13, B14 e B15.

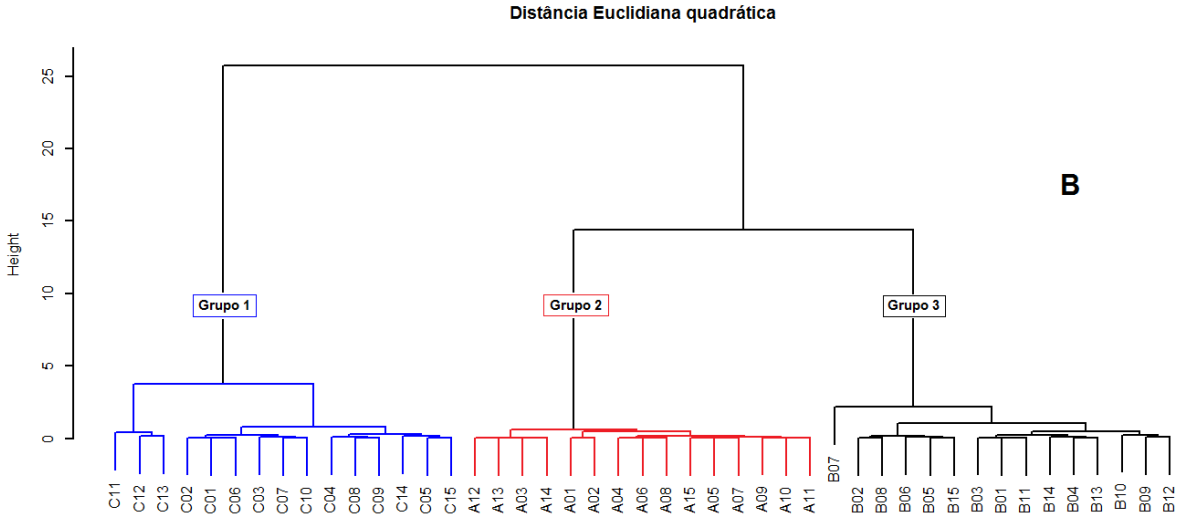
Em relação ao dendrograma obtido pelo método de Ward com base na distância de *Mahalanobis*, não se tem a formação de grupos bem definidos. Então, foi feito um corte no dendrograma, de modo a obter 3 grupos para se comparar com os grupos formados com base nas outras medidas de dissimilaridade. Pode-se observar pela Figura 1D que os grupos formados não são os mesmos, e nestes tem se amostras de sítios diferentes.

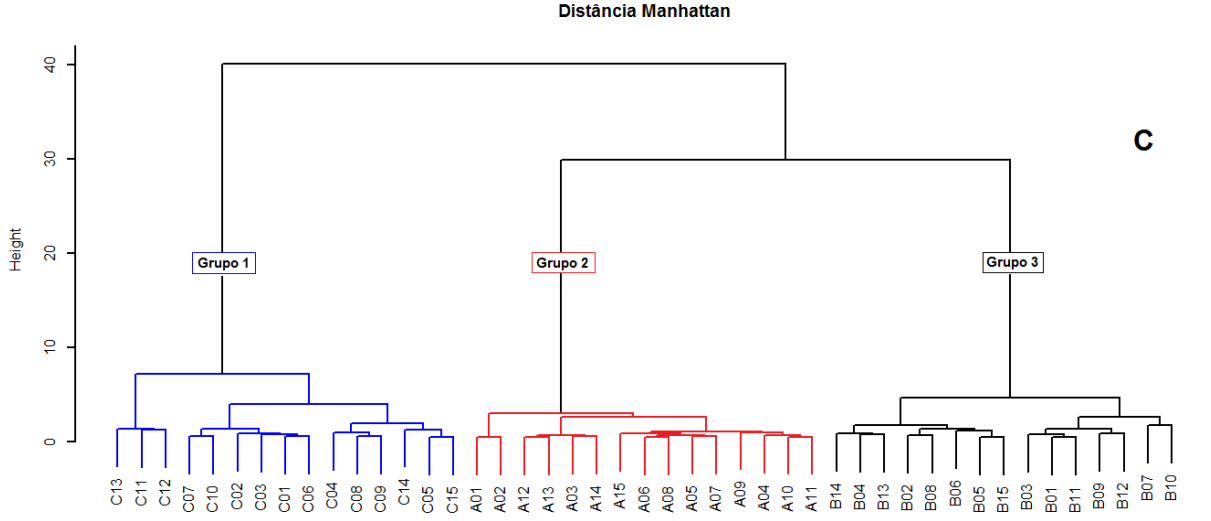
Em relação à avaliação do grau de ajuste entre as matrizes de dissimilaridade e as matrizes resultantes dos agrupamentos, para a formação dos dendrogramas, foi estimado o coeficiente de correlação cofenética (CCC). Segundo Rohlf [12], na prática, dendrogramas com CCC menor que 0.7 indicariam a inadequação do método de agrupamento para resumir a informação do conjunto de dados. Sendo assim, os resultados da Tabela 3 revelam a maior consistência dos agrupamentos formados com base nas distâncias Euclidiana, Euclidiana quadrática e *Mahanttan* em relação à *Mahalanobis.*

Tabela 3. Coeficientes de correlação cofenética (CCC)

Distância	**CCC**
Euclidiana	0,892
Euclidiana quadrática	0,853
Manhattan	0,923
Mahalanobis	0,321


Distância Euclidiana
A
Height
0
2
4
6
8
10
12
14
Grupo 1
Grupo 2
Grupo 3
C11 C12 C13 C02 C01 C06 C03 C07 C10 C04 C08 C09 C14 C05 C15
A12 A13 A03 A14 A01 A02 A04 A06 A08 A15 A05 A07 A09 A10 A11
B07 B02 B08 B06 B05 B15 B10 B09 B12 B03 B01 B11 B14 B04 B13


Distância Euclidiana quadrática
B
Height
0
5
10
15
20
25
Grupo 1
Grupo 2
Grupo 3
C11 C12 C13 C02 C01 C06 C03 C07 C10 C04 C08 C09 C14 C05 C15
A12 A13 A03 A14 A01 A02 A04 A06 A08 A15 A05 A07 A09 A10 A11
B07 B02 B08 B06 B05 B15 B03 B01 B11 B14 B04 B13 B10 B09 B12


Distância Manhattan
C
Height
0
10
20
30
40
Grupo 1
Grupo 2
Grupo 3
C13 C11 C12 C07 C10 C02 C03 C01 C06 C04 C08 C09 C14 C05 C15
A01 A02 A12 A13 A03 A14 A15 A06 A08 A05 A07 A09 A04 A10 A11
B14 B04 B13 B02 B08 B06 B05 B15 B03 B01 B11 B09 B12 B07 B10

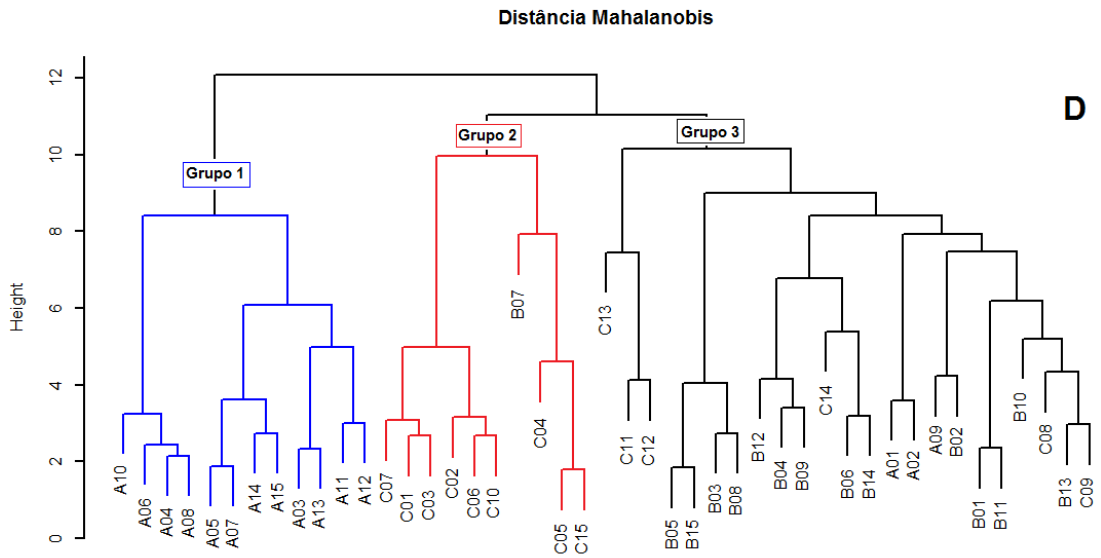


Figura 1. Dendrogramas obtidos pelo método de Ward usando como medida de dissimilaridade a distância Euclidiana (a), Euclidiana quadrática (b), *Manhattan* (c) e *Mahalanobis* (d).

Conclusões

Com base nos resultados obtidos, pode-se verificar que a utilização das medidas de distância Euclidiana, Euclidiana quadrática e *Manhattan*, para essa base de dados, gerou agrupamentos similares e de boa qualidade, fato decorrente dos valores de CCC serem altos. Além disso, pode-se dizer que a qualidade do agrupamento para essas distâncias é melhor do que o agrupamento gerado pelo uso da distância de *Mahalanobis*. Sendo assim, levando em consideração que os resultados correspondem a três sítios arqueológicos, a distância de *Mahalanobis* levaria a falsas interpretações uma vez que estaria indicando que foi usada a mesma matéria-prima na fabricação dessas peças.

Agradecimentos

Agradecemos ao apoio financeiro da CAPES/PROEX.

Referências

[1] Papageorgiou, J.; Baxter, M.J. Model-based cluster analysis of artefact compositional data. Archaeometry, 43(4), 571-588, 2001.

[2] Trebuna, P.; Halcinová, J. Mathematical tools of cluster analysis. Applied Mathematics, 4, 814-816, 2013.

[3] Barroso, L. P.; Artes, R. 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.

[4] Bussab, W. O.; Miazaki, E. S.; Andrade, D. F. Introdução à análise de agrupamentos. São Paulo: ABE, 1990.

[5] Mingoti, S. A. Análise de dados através de métodos estatísticos multivariados: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

[6] Saraçli, S.; Dogan, N.; Dogan, I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. J. Inequalities and Applications, 203, 2013.

[7] Sneath, H. A.; Sokal, R. R. Numerical Taxonomy: The principles and practices of numerical classification, p. 573. Freeman, San Francisco, 1973.

[8] Hair, J. F., Anderson, R. E., Tatham, R. L., Black, C. Análise multivariada de dados. Bookman, Porto Alegre, 2005.

[9] Ward, J. H. Hierarchical grouping to optimize an objective function. Journal of Applied Statistics, 58, 236-244, 1963.

[10] Sokal, R. R.; Rohlf, F. J. The comparison of dendrograms by objective methods. Taxon 11, 33-40, 1962.

[11] Oliveira, P. M. S; Munita, C. S.; Hazenfratz, R. Comparative study between three methods of outlying detection on experimental results. J. Radioanalytical and Nuclear Chemistry, 283, 433-437, 2010.

[12] Rohlf, F. J. Adaptative hierarquical clustering schemes. Systematic Zoology, V.19, n.1, 58-82, 1970.