Aplicação de método de imputação para substituição de dados discrepantes univariados obtidos em resultados experimentais

PAULO TADEU MEIRA E SILVA OLIVEIRA[1], CASIMIRO JAIME ALFREDO SEPÚLVEDA MUNITA[*, 1].

[1] *Instituto de Pesquisas Energéticas Nucleares, IPEN – CNEN / SP, Brasil*

ptoliveira@ipen.br

1 - Introdução

Na interpretação dos resultados em estudos arqueométricos de cerâmicas arqueológicas, para classificar, estudar a similaridade/dissimilaridade, a proveniência das amostras e a tecnologia de produção são utilizados métodos estatísticos multivariados, tais como: análise de conglomerados (do inglês, *cluster analysis*), análise de componentes principais (do inglês, *principals components*), análise discriminantes do inglês, *discriminant analysis*), entre outros. Contudo, para que seja viável a utilização destas técnicas estatísticas se faz necessário que a matriz das amostras não possua valores discrepantes (do inglês, *outliers*) e que esteja completa, isto é, inexistência de valores faltantes (do inglês, *missing values*), para uma posterior análise dos dados completos (Stanimirova and Walczak, 2008).

Resultados discrepantes podem ser gerados por processo fora de controle, técnica analítica errada, contaminação durante a preparação da amostra, medida com alto erro, etc (Oliveira and Camunita, 2003). Estes resultados podem ser classificados em univariado (relativos a uma única variável) e multivariados (relativo a duas ou mais variáveis conjuntamente).

Uma alternativa, neste caso, para resolver o problema de dados discrepantes univariados em resultados experimentais são tratá-los como valores perdidos, aplicando método de imputação para estimação de valores plausíveis para substituir estes dados e continuar a análise (Stanimirova and Walczak, 2008).

A ocorrência de dados faltantes pode ocorrer devido a problemas como medição da concentração elementar após o elemento completar o tempo de meia vida, troca inesperada das condições experimentais, pequena quantidade da concentração do elemento na amostra a ser analisada entre outras, provocando falhas na obtenção das medidas de concentração de um ou mais elementos existentes naquela amostra.

Os objetivos deste trabalho são detectar os dados discrepantes univariadose tratá-los como faltantes utilizandoum método de imputação para obtenção de valores plausíveis para sua substituição e avaliar em termos comparativos a quantidade de dados discrepantes univariados e multivariados antes e após a aplicação deste método de imputação.

Para este trabalho; foram utilizados dados de concentrações elementares de As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th e U determinadas por análise por ativação com neutrons instrumental em amostras de fragmentos cerâmicos de um sítio arqueológico; foi feito um estudo de detecção de dados discrepantes univariados aplicando o método de Box-Plot e um método de análise discrepantes multivariados utilizando o método da distancia de Mahalanobis antes e após aplicação do método de imputação por decomposição do valor singular.

2 – Método

2.1 - Motivação

Para este estudo, foram considerados para análise os dados de 31 amostras de concentrações elementares de As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th e U determinadas por análise por ativação instrumental, tida como uma técnica bastante sensível, utilizada nas análises qualitativas e quantitativas de elementos presentes numa ampla faixa de concentrações, da ordem de percentagens a níveis de traço(Aguiar, 2001), em amostras de fragmentos cerâmicos coletadas de sítios arqueológicos no Reator Nuclear IEA-R1 no Centro do Reator de Pesquisa no Instituto de Pesquisas Energéticas Nucleares (IPEN).

Dados discrepantes são caracterizados por observações que se destacam quanto a uma ou mais variáveis do estudo e que podem influenciar na modelagem estatística dos dados e, portanto, a análise do comportamento dos mesmos (Baxter, 1994).

No caso em que se observa apenas uma variável para cada elemento químico do conjunto de amostras, esses dados discrepantes são univariados. Já, quando se trata de um conjunto de dados com duas ou mais variáveis sendo analisadas, os dados discrepantes, neste caso, são multivariados. (Giroldo, 2008)

Para este trabalho, os valores discrepantes univariados para cada variável foram obtidos aplicando o método de Box-plot.

Os dados discrepantes univariados foram substituídos por valores obtidos pelo método da decomposição do valor singular que se baseia na decomposição da matriz dos valores das variáveis dos scores das componentes principais seguindo o algoritmo proposto por Krzanovsky (1987).

2.2 – Método de Box - plot

Trata-se de um gráficoque mostra caracteristicas dos dados ordenados como Q_1 (primeiro quartil), que

* ptoliveira@ipen.br

corresponde a posição dos primeiros 25% destes dados; Md (mediana), que corresponde a posição dos primeiros 50% destes mesmos dados, e por fim; Q_3 (terceiro quartil), que corresponde a posição dos primeiros 75% dos dados já ordenados.

Considere um retângulo com base determinadas por Q_1 e Q_3, conforme ilustra a Figura 1. Marcado com um seguimento a posição da mediana. Considere os limites liminf = $Q_1 - 1{,}5(Q_3 - Q_1)$ e limsup =$Q_3 + 1{,}5(Q_3 - Q_1)$ onde liminf e limsup são os limites inferior e superior, respectivamente, para as concentrações de cada variável isoladamente. As observações que estiverem acima do limite superior (limsup) ou abaixo do limite inferior (liminf)estabelecidos são observações destoantes das demais, são chamados de valores discrepantes univariados (Baxter, 1994; Bussab and Moretin, 2009).

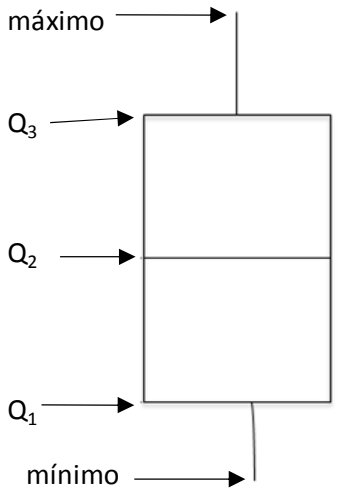

Figura 1. Box plot (desenho esquemático)

A figura 1 mostra a porção dos dados (50%) entre o primeiro e o terceiro quartil e a posição da mediana.O Box plot dá uma idéia da posição, assimetria, caudas e dados discrepantes univariados. A posição central é dada pela mediana e a dispersão é dada por $Q_3 - Q_1$. As posições relativas de Q_1, Q_2 e Q_3traz uma noção da assimetria da distribuição.

2.3 – Distância de Mahalanobis

Para cada uma das *n* amostras no conjunto de *p* variáveis, a distância D_i^2 é calculada. Se $\bar{x}$ é a média do vetor e *S* é a matriz de covariância amostral, então

$$S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \Big/ (n-1) \quad (1)$$

e

$$D_i = \sqrt{\{(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})\}} \quad \text{para i =1....n} \quad (2)$$

onde $(x_i - \bar{x})$ é o vetor da diferença entre os valores da medida em um grupo e a média dos valores do outro grupo (Oliveira and Camunita, 2003).

O valor desta distância foi comparado ao valor crítico obtido pela expressão

$$\frac{p(n-1)^2 F_{p,n-p-1;\alpha/n}}{n\left(n - p - 1 + pF_{p,n-p-1;\alpha/n}\right)} \quad (3)$$

em que *p* representa o número de variáveis, *n* o número de amostras, *F* é o teste *F* também chamado de distribuição de Fisher ($F = s_1^2 / s_2^2$ onde s_1^2 e s_2^2 são variâncias amostrais) com *p* graus de liberdade a um nível de significância α/n, α = 0,05 (Oliveira et all., 2010).

2.4 – Decomposição de valor singular

A principal característica da decomposição por valor singular está na capacidade do método em determinar o "rank" ou o grau de singularidade de uma matriz. Quando trata de dados de concentração esta singularidade reflete-se no grau de correlação entre os traços de concentração. Quanto menor for o "rank" maior será a correlação entre o traço com a informação da matriz de dados de contração representada por *X* concentrada em poucas componentes principais, que estarão associadas a grandes valores singulares.

Para a elaboração do método da decomposição do valor singular,considere a matriz de dados de concentração representada por $X_{n \times p}$ na forma

$$X = UDV' \quad (4)$$

em que $U'U = I_p$, $V'V = VV' = I_p$ e $D = diag(d_1, \ldots, d_p)$ com $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$. As matrizes $X'X$ e XX' tem os mesmos autovalores, e os elementos d_i são as raizes quadradas destes autovalores; a i-ésima coluna $v_i = (v_{i1}, \ldots, v_{ip})'$ da matriz $V_{(n \times p)}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $X'X$; enquanto a j-ésima coluna $u_i = (u_{i1}, \ldots, u_{ni})$ da matriz $U_{(n \times p)}$é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de XX';. A decomposição (4) possui a seguinte representação elementar:

$$x_{ij} = \sum_{t=1}^{p} u_{it} d_t v_{tj} \quad (5)$$

Esta representação é utilizada como uma base para determinar a dimensionalidade do conjunto de dados multivariados, se a estrutura dos dados é m-dimensional, então a variação na dimensão resultante $(p - m)$pode ser tratada como ruído aleatório. As características principais dos dados estarão no espaço

dos *m* primeiros componentes principais. A correspondência entre as quantidadesno lado direito de (5) e os erros principais da configuração dos dados sugerem o modelo correspondente

$$x_{ij} = \sum_{t=1}^{m} u_{it} d_t v_{tj} + \varepsilon_{tj} \qquad (6)$$

Agora suponha o modelo (6) para um específico valor de *m*, mas que uma única observação x_{ij}, foi perdida na matriz de dados. Então x_{ij} é estimado por

$$\hat{x}_{ij}^{(m)} = \sum_{t=1}^{m} u_{it} d_t v_{tj} \qquad (7)$$

em que u_{ti}, d_t e v_{tj}devem ser estimados com o restante dos dados. Simbolizado por $X_{(-i)}$ a matriz dos dados obtida retirando-se a i-ésima linha de *X,* e por, $X_{(-j)}$ a matriz dos dados obtida retirando-se a j-ésima coluna de *X*. Tornando a decomposição da matriz singular dessas matrizes como:

$$X_{(-i)} = \overline{UDV} \text{ com } \overline{U} = (\overline{u}_{st}), \overline{V} = (\overline{v}_{st}) \text{e } \overline{D} = diag(\overline{d}_1, \cdots \overline{d}_p) \quad (8)$$

e

$$X_{(-j)} = \tilde{U}\tilde{D}\tilde{V} \text{ com } \tilde{U} = (\tilde{u}_{st}),\ \overline{V} = (\overline{v}_{st}) \text{e } \tilde{D} = diag(\tilde{d}_1, \cdots \tilde{d}_{p-1}) \quad (9)$$

A estimativa de u_{it} e v_{tj} em (7) com o máximo dos dados, é u_{st} e v_{st}, enquanto $d_t$pode ser estimado por $\overline{d}_t$ e $\tilde{d}_t$ou por alguma combinação dos dois. Um meio termo adequado parece ser $\sqrt{\overline{d}_t}$ e $\sqrt{\tilde{d}_t}$, em que uma estimativa do valor predito x_{ij} (Krzanowski, 1987) é dada por

$$\hat{x}_{ij}^{(m)} = \sum_{t=1}^{m} \left(\tilde{u}_{it} \sqrt{\tilde{d}_t} \right) \left(\overline{v}_{tj} \sqrt{\overline{d}_t} \right) \qquad (10)$$

Por fim, seguindo o preceito do máximo de dados, utiliza-se o valor mais elevado de *m* possível. De (6) este é *p – 1*, de f41orma que o valor atribuído x_{ij} será

$$\hat{x}_{ij} = \sum_{t=1}^{m} \left(\tilde{u}_{it} \sqrt{\tilde{d}_t} \right) \left(\overline{v}_{tj} \sqrt{\overline{d}_t} \right) \qquad (11)$$

Para os casos em que ocorreram mais de um valor perdido nos dados, foi implementado um programa iterativo, em que as estimativas iniciais foram atribuídas para todos os valores perdidos, e cada um deles foi recalculado utilizando a expressão (11). Cada uma dessas estimativas requer duas decomposições de valores singulares, isto é, de $X_{(-i)}$ e $X_{(-j)}$para *i* e *j* necessários (usando a estimativa do valor perdido para "completar" X). O processo é iterativo até ser alcançada a estabilidade nos valores atribuídos (Bérgamo, 2007).

3 – Resultados e Discussão

A análise por INAA (Análise por ativação por neutrons instrumental) tem se destacado na determinação da combinação química em amostras de cerâmicas por apresentar vantagens como alta sensibilidade, precisão e exatidão (Toyota, 2009).

A Tabela 1 mostra os valores das concentrações elementares para 31 amostras de fragmentos cerâmicos e o valor da Distância de Mahalanobis. Inicialmente, os resultados foram transformados por log10 para compensar a diferença de grande magnitude entre os valores dos elementos medidos no nível traço e maiores que um (Oliveira and Camunita, 2010). Outra razão para isto é o fato de acreditar que em materiais manufaturados, medidas de concentrações elementares têm distribuição lognormal e a normalização dos dadosé desejável (Santos et. all , 2007).Na Tabela 1, observa-se que as concentrações marcadas em negrito são os casos detectados como discrepantes univariados após aplicar o método de Box-plot, cujo resumo dos resultados é apresentado na Tabela 2 para cada elemento que foi detectado pelo menos um resultado discrepante e que as amostras e valores de distância de Mahalanobis marcadas em negrito são as amostras consideradas como discrepantes multivariadas (amostras 5, 11 e 31) após aplicar o método da distância de Mahalanobis.

A utilização do método de Box plot para determinação de dados discrepantes univariados foi motivado pelo fato de ser um método bastante utilizado, fácil de usar e de maior precisão para a detecção de observações verdadeiramente atípicas.

Tabela 2. Relação de amostras discrepantes por elemento.

Elementos	Amostras discrepantes
La	21, 22 e 31
Th	27, 29, 30 e 31
Cr	8, 23, 27, 29 e 30
Cs	20
Sc	23 e 27
Ce	21 e 31
Hf	6 e 29
Tb	21

Verifica-se na Tabela 2 que o elemento que apresenta maior quantidade de amostras discrepantes é o Cr (amostras 8, 23, 27, 29 e 30) e em todos os elementos pesquisados foram encontrados 20 casos de dados discrepantes após aplicar o método de Box-plot. Considerando que as causas que levam a ocorrência de dados discrepantes já mencionadas anteriormente (medição da concentração elementar após completar a meia vida do elemento, troca inesperada das condições experimentais, pequena quantidade da concentração do elemento na amostra a ser analisada entre

outras)podem ser consideradas como faltantes. Para este trabalho foi aplicado um método de imputação para dados faltantes que foi o método de decomposição singular.

O método de decomposição singular, diferente de outros métodos de imputação como imputação pela média, imputação pela normal univariada e imputação normal multivariada dentre outros; trata-se de um método de atribuição perfeitamente geral e livre de qualquer restrição quanto a distribuição de valores (Bergamo, 2007).

Tabela 1. Resultados para os dados de concentração para amostras de cerâmica em μg/g, exceto caso contrário indicado e valores da distância de Mahalanobis.

Amostras	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	0.30	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	0.10	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
28	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
29	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 *critical value at significance level of 0.05*															23.6	22.8	22.3

A Tabela 3 mostra os valores das concentrações elementares e em negrito mostra os valores dos dados discrepantes calculados após aplicar o método de decomposição do valor singular que substituíram os valores discrepantes existentes na Tabela 1.

Pode observar que os valores de concentração em negritos foram os valores discrepantes univariados obtidos para os dados após ser feita a substituição destes dadospelos valores estimados pelo método de imputação de decomposição por valor singular,cujo resumo é mostrado na Tabela 4 e os número de amostras e distancia de Mahalanobis estão em negrito na Tabela 3 é a amostra (amostra 11) detectada como discrepantes multivariada após aplicar a distância de Mahalanobis.

Tabela 4. Relação de amostras discrepantes univariadas por elemento após ter sido aplicado o método de imputação por decomposição do valor singular.

Elementos	Amostras discrepantes
Th	31
Cr	31
Ce	22, 28 e 29
Hf	23, 30 e 31

É possível verificar na Tabela 4, que ocorreu uma diminuição na quantidade de dados discrepantes univariados (de 20 na Tabela 2 para 8 na Tabela 4) e a maior quantidade foi detectada para os elementos Ce (amostras 22, 28 e 29) e Hf (23, 30 e 31) com tres amostras em cada uma.

Tabela 3. Resultados para os dados de concentração para amostras de cerâmica em μg/g, exceto caso contrário indicado, após substituição dos dados discrepantes por valores plausíveis obtidos pelo método de decomposição singular e valores da distância de Mahalanobis.

Amostras	Na (%)	Lu	U	Yb	La	Th	Cr	Cs	Sc	Fe (%)	Eu	Ce	Hf	Tb	D_1	D_2
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	7.51	7.70
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	7.55	8.21
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	16.58	16.25
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	15.39	15.10
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	19.75	19.10
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	5.83	1.02	17.30	18.88
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	21.18	22.23
8	0.10	0.30	2.88	1.82	26.92	10.96	54.21	8.91	11.48	3.02	0.71	53.70	7.59	0.40	18.84	18.20
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	7.59	7.30
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	10.27	10.87
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	**27.90**	
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	16.49	16.91
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	4.27	4.82
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	7.97	7.85
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	3.42	3.67
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	6.63	6.55
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	11.46	14.54
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	12.18	13.31
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	14.98	14.80
20	0.10	0.34	3.89	2.40	36.31	12.02	64.57	12.55	14.45	2.63	1.15	112.20	6.03	0.69	20.09	19.83
21	0.20	0.50	4.17	3.72	50.45	14.79	63.10	6.17	15.14	3.72	1.91	**120.76**	8.51	1.09	6.68	8.15
22	0.20	0.60	4.68	4.47	48.04	16.98	69.18	9.12	17.38	3.80	1.70	**123.03**	9.55	1.20	6.30	7.77
23	0.20	0.48	5.13	3.80	44.67	17.38	67.03	11.48	17.39	4.47	1.66	95.50	7.94	1.20	6.53	7.86
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	13.84	14.04
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	7.19	8.64
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	11.22	11.43
26	0.10	0.51	3.89	3.89	47.86	16.18	68.73	12.30	18.51	4.90	1.66	**120.23**	6.76	0.71	13.65	13.34
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	15.42	14.97
28	0.07	0.55	5.37	3.80	35.48	**19.83**	**78.42**	9.33	17.78	5.13	1.17	89.13	**13.24**	0.89	14.56	15.61
30	0.07	0.47	3.98	3.63	32.36	17.71	68.82	8.91	15.85	4.07	1.15	83.18	**11.48**	0.81	9.75	9.53
31	0.06	0.42	3.31	2.95	34.46	15.05	66.07	6.61	13.80	2.00	1.32	73.27	**11.48**	0.68	17.52	19.52
D *critical value at significance level of 0.05*															23.56	23.17

Ao aplicar a distancia de Mahalanobis nos dados da Tabela 1 nota-se a presença de três amostras discrepantes (5, 11 e 31), enquanto que para os dados da Tabela 3, após substituir os valores discrepantes univariados pelo método da decomposição singular, nota-se que apenas uma amostra foi considerada discrepante (amostra 11) e essa diminuição deve ter sido motivada pelo fato de que ao substituir as amostras discrepantes univariadas pelos valores plausíveis obtidos pelo método de imputação mostrou-se que muita delas deixaram de ser discrepantes ocorrendo uma melhorana distribuição dos dados da Tabela 3 ao serem comparados com os dados da Tabela 1.

4 – Conclusões

Ao fazer a substituição dos dados discrepantes univariados aplicando o método de imputação por decomposição singular foi possível concluir que ocorreu uma diminuição da quantidade de dados discrepantes por variável e menor quantidade de amostras discrepantes multivariadodapós aplicar o método de distância de Mahalanobis.

5 – Referências

[1]AGUIAR, A. M. (2001). *Aplicação do método de análise por ativação com nêutrons à determinação de elementos traços em unhas humanas*. Dissertação de mestrado, IPEN-CNEN/SP.

[2]BAXTER, M.J. (1994) *Exploratory multivariate analysis in archaeology*. Edimburgh. University Press Ltd Edimburgh, Great Britain.

[3]BÉRGAMO, G. C. (2007). *Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação*. Tese de doutorado, ESALQ-USP

[4]BUSSAB, W. O.; MORETTIN, P. A. (2009). *Estatística Básica*. Editora Sarauva, São Paulo-SP.

[5]GIROLDO, F. R. S. (2008). *Alguns métodos robustos para detectar outliers multivariados*. Dissertação de mestrado, IME-USP.

[6]KRZANOWSKY, W. J. (1987). Cross-validation in principal component analysis. *Biometrics,* **43:**575—584.

[7]OLIVEIRA, P.M.S.; MUNITA, C. S. (2003) Influência do valor crítico na detecção de valores discrepantes em arqueometria. 48ª Reunião Anual da Sociedade Internacional de Biometria – RBRAS e 9º Simpósio de Estatística Aplicada à Experimentação Agronômica – SEAGRO, Lavras – MG, 7 a 11 de julho.

[8]OLIVEIRA, P.M.S.; MUNITA, C. S.; HAZENFRATZ, R. (2010). Comparative study between three methods of outlying detection on experimental results.*Journal Radioanalytical Nuclear Chemistry*, **283(2):**433—437.

[9]SANTOS, J. O.; MUNITA, C. S.; VERGUE, C.; OLIVEIRA, P.M.S. (2007) Normalização e padronização por meio da transformação logarítmica em estudos arqueométricos de cerâmicas. 52ª Reunião Anual da Sociedade Internacional de Biometria – RBRAS e 12º Simpósio de Estatística Aplicada à Experimentação Agronômica – SEAGRO, Santa Maria – RS , 23 a 27 de julho.

[10]TANIMIROVA, I.; WALCZAK, B. (2008).Classification of data with missing elements and outliers.*Talanta*, **76:**279—289.

[11]TOYOTA, R. G. (2009) *Caracterização química da cerâmica Marajoara*. Dissertação de mestrado, IPEN-CNEN/SP.