IDENTIFICAÇÃO DE VALORES DISCREPANTES POR MEIO DA DISTÂNCIA MAHALANOBIS

PAULO TADEU MEIRA E SILVA DE OLIVEIRA
Depto de Estatística, Universidade de São Paulo, e-mail: poliveir@ime.usp.br
JOSE OSMAN DOS SANTOS
Supervisão de Radioquímica, *IPEN – CNEN /SP, e-mail: josantos@ipen.br*
CASIMIRO SEPÚLVEDA MUNITA
Supervisão de Radioquímica, IPEN – CNEN /SP, e-mail: camunita@net.ipen.br

1. INTRODUÇÃO

Conforme a similaridade química elementar entre os fragmentos cerâmicos, a detecção de amostras discrepantes em estudos composicionais de cerâmicas constitui uma das principais tarefas antes da aplicação das técnicas de classificação das amostras. A razão para esta busca, reside no fato que as amostras discrepantes podem alterar os grupos composicionais ou distorcer os critérios de otimização em que podem alterar a separação entre os grupos (Hardin et al, 2004).

Tendo em vista a natureza geoquímica dos dados obtidos em estudos composicionais de cerâmicas arqueológicas, tem sido citado na literatura que as amostras discrepantes são, geralmente, observações resultantes de processos secundários, e não de valores extremos a partir da distribuição natural dos dados (Reimann, 2000). Os resultados analíticos discrepantes são gerados por processos fora de controle, técnicas analíticas erradas, contaminação durante a preparação das amostras, medida com alto erro, etc.

Os valores analíticos derivados de um processo secundário não precisam ser, necessariamente, observações de valores altos ou baixos em relação a todas as variáveis na base de dados, logo a tentativa de identificar amostras discrepantes com métodos univariados clássicos são pouco confiáveis. Desta forma, vários métodos para identificação das amostras discrepantes no espaço multidimensional têm sido propostos (Hadi, 1992). Entre os métodos, os que são determinados por meio da matriz de covariância tem sido aplicados com sucesso. Com esse propósito, a distância de Mahalanobis tem sido aplicada para identificação das amostras discrepantes.

A observação de um valor discrepante por meio da distância de Mahalanobis, de uma dada amostra ao centróide dos dados, é realizada, de uma forma geral, comparando-se a distância com um valor crítico. Como apresentado em trabalho anterior, a escolha do valor crítico é fundamental para identificação das amostras discrepantes, uma vez que, influencia no julgamento se uma amostra deve ser considerada discrepante (Oliveira et al., 2003). Tão importante quanto a escolha do valor crítico, os estimadores de localização e dispersão, usados para obtenção da distância de Mahalanobis, tem influência no processo de identificação das amostras discrepantes, uma vez que estes estimadores não são robustos em relação à própria presença das amostras discrepantes (Rousseeuw, 1990).

Para contornar a questão da falta de robustez das medidas de localização e dispersão dos dados na obtenção da distância de Mahalanobis, tem sido discutida na literatura a estimativa destes parâmetros por

meio do subconjunto dos dados que minimizam o determinante da matriz de covariância (Minimum Covariance Determinant – MCD). A distância obtida utilizando-se os estimadores baseado no MCD é denominada distância de Mahalanobis robusta; visto que a distância de Mahalanobis robusta distribui-se, aproximadamente, segundo uma χ_p^2, pode ser utilizada, com eficácia, para detecção de amostra discrepante.

Dentro deste contexto, neste trabalho, para identificar amostras discrepantes em uma base de dados oriunda de estudos arqueométricos de cerâmicas, aplicou-se a distância de Mahalanobis clássica, usando como valor crítico o critério lambda Wilks (Oliveira et al., 2003), e a distância Mahalanobis robusta, usando como valor crítico o valor da distribuição χ_p^2, em 13 elementos químicos obtidos por meio do método de análise por ativação com nêutrons em 41 amostras.

2. DESENVOLVIMENTO

Inicialmente, os resultados das concentrações elementares foram normalizados mediante transformação logaritma com base 10 para compensar as diferenças em magnitude de elementos que são determinados em porcentagem dos que estão ao nível de traços Sayre (1975). Vários autores sugerem a distância Mahalanobis, D_i^2, como método para detecção de resultados discrepantes quando são determinadas várias variáveis Penny (1987).

Neste trabalho, a distância Mahalanobis ao quadrado foi calculada para cada uma das 41 amostras e os resultados estão apresentados nas três últimas colunas da Tabela 1.

Freqüentemente é sugerido que a distribuição *F,* para calcular o valor crítico é mais adequada que a distribuição χ^2, especialmente quando o número de amostras é pequeno.

Na Tabela 1 mostra-se os valores da distância de Mashalanobis calculadas para nível de confiança 95% usando o critério lambda Wilks. Na primeira coluna, as amostras 10, 26 e 37 são consideradas discrepantes, recalculando novamente o valor da distância de Mahalanobis, na segunda coluna, a amostra 21 é discrepante em relação às demais.

Calculando a distância Mahalanobis robusta, por meio da expressão $MCD = (\overline{X}_J^*, \overline{S}_J^*)$

onde $J = \left\{\text{conjunto de h pontos} : \left|S_J^*\right| \le \left|S_K^*\right| \forall \text{conjunto K} \# |K| = h\right\}$ e $\#|\omega|$ define o número de elementos no conjunto ω

$$\overline{X}_J^* = \frac{1}{h}\sum_{i \in J} x_i \qquad (1)$$

$$S_J^* = \frac{1}{h}\sum (x_i - \overline{X}_J^*)(x_i - \overline{X}_J^*)^t \qquad (2)$$

A seleção do tamanho da amostra h foi determinada por um compromisso entre a robustez e eficiência dos estimadores obtidos do subconjunto das observações, a qual tem determinante mínimo. Desta forma, um valor de $h \approx 0{,}75n$ (n é o tamanho total da amostra) foi empregado para obtenção do

estimadores baseados no MCD. Substituindo-se os estimadores de localização e dispersão robustos na equação (2) são obtidas as distâncias de Mahalanobis robustas (RD). De acordo com (Rousseeuw, 1984) ,se a RD quadrática para uma observação é maior que $\chi^2_{p;0,98}$, esta amostra pode ser declarada discrepante.

Tabela 1. Resultados das concentrações elementares, em ppm, exceto quando indicado e valor da distância Mahalanobis.

Amostra	As	Ce	Cr	Eu	Fe, %	Hf	La	Na	Nd	Sc	Sm	Th	U	D_1^2	D_2^2	D_3^2
1	2,6	67,8	212,0	2,9	1,3	10,8	31,8	132,0	41,0	39,9	9,4	6,4	1,3	6,3	6,1	6,6
2	1,7	75,8	205,0	2,9	0,9	12,5	31,8	121,0	45,0	41,8	9,0	6,9	1,6	5,9	8,5	8,7
3	1,9	61,1	215,0	2,9	1,0	10,9	30,8	176,0	45,0	46,2	9,1	7,3	1,5	3,5	5,6	5,6
4	1,6	56,4	183,0	2,4	0,8	10,8	28,0	120,0	35,0	43,4	7,5	6,4	1,5	14,1	16,5	19,0
5	2,8	68,6	215,0	3,0	1,2	11,8	34,0	145,0	48,0	45,0	9,3	6,3	1,4	2,2	3,3	3,5
6	2,0	61,7	212,0	3,0	0,9	10,8	34,0	125,0	48,0	47,4	9,2	6,9	1,7	6,1	7,8	7,6
7	2,2	62,5	195,0	2,8	0,9	11,3	29,3	92,0	46,0	42,5	9,2	7,1	1,3	8,9	8,2	8,0
8	1,6	82,0	187,0	3,2	1,1	10,8	37,2	260,0	47,0	37,2	9,8	4,8	1,2	10,5	11,8	11,8
9	1,5	90,8	303,0	3,2	1,2	11,0	39,5	266,0	52,0	41,7	10,2	5,6	1,1	14,0	14,1	13,7
10	1,3	20,2	57,7	0,8	0,2	2,5	39,0	244,0	45,0	9,5	10,1	1,2	0,9	38,0		
11	2,8	64,6	216,0	2,8	1,2	10,9	31,2	271,0	40,0	44,5	9,7	7,4	1,5	7,3	14,1	15,4
12	2,4	85,2	214,0	3,3	1,6	10,8	37,6	155,0	53,0	43,9	10,8	5,2	1,2	9,3	9,6	9,5
13	1,9	135,0	150,0	4,6	0,7	12,3	54,5	160,0	67,0	50,9	14,1	5,7	1,2	19,4	21,2	20,6
14	1,8	101,5	230,0	3,4	1,4	11,7	45,5	144,0	51,0	45,0	11,4	7,7	1,3	10,8	11,2	14,0
15	0,5	104,5	214,0	3,5	1,4	11,8	46,6	144,0	59,0	48,1	12,2	6,5	1,5	6,7	7,5	7,5
16	1,4	95,2	245,0	3,5	1,2	12,1	44,0	187,0	57,0	43,0	11,3	5,8	1,4	6,3	5,9	6,2
17	2,0	67,5	205,0	2,8	1,2	10,6	36,6	93,0	42,0	43,5	10,1	7,0	1,9	9,7	14,8	14,5
18	1,2	63,4	183,0	2,9	1,0	10,5	33,9	130,0	44,0	40,7	9,6	6,7	1,7	6,0	6,4	6,3
19	3,0	65,3	212,0	2,9	1,3	10,5	33,5	138,0	50,0	42,6	9,7	6,8	1,6	3,1	3,4	3,4
20	2,7	67,8	236,0	3,0	1,1	11,0	33,8	139,0	55,0	41,2	10,0	6,3	1,4	4,1	4,0	3,9
21	1,5	83,5	82,0	2,5	0,4	6,6	29,2	976,0	45,0	34,2	8,6	5,2	1,6	24,6	27,2	
22	1,9	52,5	195,0	2,7	0,9	11,6	26,2	136,0	43,0	43,2	8,5	7,3	1,4	7,6	9,4	9,1
23	1,9	109,7	218,0	3,3	0,8	11,7	37,8	181,0	60,0	39,4	10,3	5,2	1,1	10,0	10,4	10,3
24	1,7	87,8	241,0	3,3	1,2	10,9	40,8	200,0	71,0	45,6	11,0	7,0	1,3	9,5	10,7	10,4
25	1,6	78,9	230,0	3,2	0,9	10,9	41,1	189,0	69,0	40,0	11,3	5,1	1,1	7,8	15,4	17,3
26	1,9	112,9	48,0	2,9	0,8	11,5	47,7	1583,0	68,0	35,1	10,7	10,0	2,0	26,7		
27	1,2	68,9	204,0	2,9	0,8	11,4	32,8	191,0	51,0	44,3	10,2	6,8	1,6	3,4	6,4	9,9
28	2,5	54,5	203,0	3,0	1,3	10,9	34,1	138,0	44,0	44,7	9,6	6,8	1,2	7,4	8,5	8,8
29	1,1	245,0	162,0	4,7	0,7	11,6	53,5	172,0	72,0	48,4	15,6	5,3	1,0	23,5	23,8	23,8
30	1,4	70,9	192,0	3,0	0,8	11,9	36,1	117,0	61,0	46,1	10,3	7,4	1,5	8,0	9,9	9,9
31	1,4	93,2	243,0	3,4	1,3	12,8	40,9	189,0	54,0	45,8	11,4	6,1	1,2	2,3	3,3	3,5
32	1,8	129,0	95,0	3,8	1,3	14,1	54,1	1339,0	62,0	40,5	12,3	7,0	1,1	12,0	20,6	21,2
33	1,6	110,0	260,0	3,8	1,3	12,3	48,3	159,0	59,0	44,1	13,2	5,8	0,9	7,6	11,2	12,2
34	1,7	95,2	204,0	3,4	1,4	12,5	43,5	192,0	48,0	50,1	11,1	6,8	1,2	9,5	9,1	9,2
35	3,1	104,0	99,0	3,9	1,6	11,1	48,9	583,0	65,0	37,8	12,6	6,2	0,8	15,6	16,0	15,7
36	3,0	137,9	94,0	3,7	1,3	11,0	51,7	535,0	54,0	37,3	12,7	5,6	0,8	12,1	14,7	16,1
37	3,0	134,0	70,0	4,4	1,3	14,4	56,8	360,0	67,0	45,1	9,2	7,7	1,1	30,0		
38	1,3	89,2	249,0	3,4	1,5	12,3	39,5	165,0	62,0	48,9	11,1	5,7	1,4	12,3	12,4	12,5
39	2,4	123,2	224,0	4,3	9,2	1,3	51,5	176,0	58,0	47,8	14,0	7,4	1,6	20,0	19,3	18,8
40	1,8	97,5	238,0	3,3	8,0	1,2	38,0	167,0	52,0	42,3	10,4	6,2	1,8	15,8	14,8	16,6
41	1,8	92,7	253,0	3,6	14,9	1,3	44,2	125,0	63,0	48,3	11,7	6,4	1,2	21,7	20,8	20,9

$\frac{p(n-1)^2 F_{p,n-p-1;\alpha/n}}{n(n-p-1+pF_{p,n-p-1;\alpha/n})}$	25,6 ($n = 41$)	24,9 ($n = 38$)	24,6 ($n = 37$)

A elipse de tolerância pode ser construída por meio dos estimadores robustos para obtenção da distância de Mahalanobis igual $\chi^2_{p;0,98}$, No hiperespaço das variáveis medidas, as amostras que estiveram fora desta elipse são consideradas discordantes da estrutura geral de dados (Filzmoser et al, 2005).

Na figura 1 as composições químicas elementares das cerâmicas são apresentadas no espaço formado pelos dois componentes principais, os que explicam 52,5 % da variabilidade total dos dados. Neste gráfico, também, são apresentados 4 elipses de tolerância correspondentes aos quantis 0,25, 0,5, 0,75 e 0,98 da $\chi^2_{p;0,98}$. As amostras discrepantes, considerando-se a distância de Mahalanobis robusta, estão foram da elipse de tolerância que corresponde à $\chi^2_{p;0,98}$.

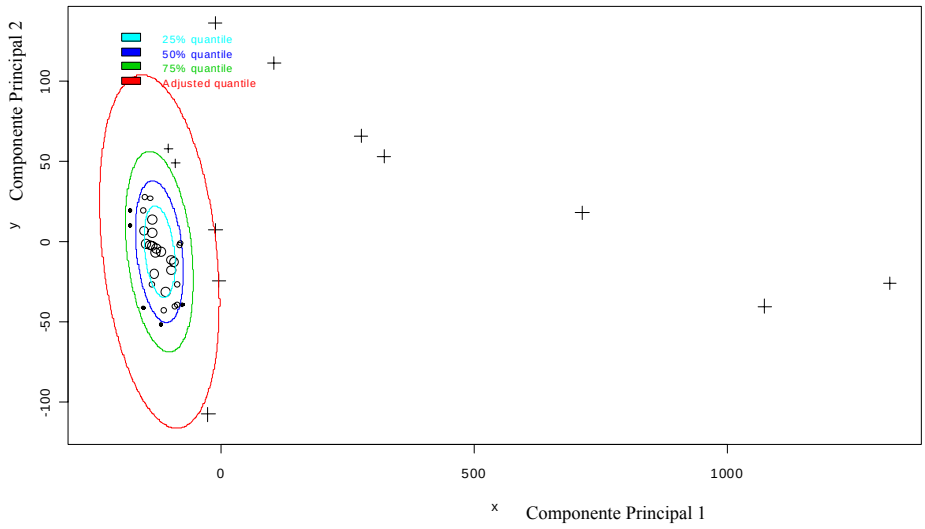

Figura 1. Componente Principal 1 versus Componente Principal 2

As elipses correspondem aos quantis 0,25, 050, 075 e 0,98 da distribuição χ^2_p.

De acordo com o procedimento baseado na distância de Mahalanobis robusta é possível a identificação de 10 amostras discrepantes: 8, 9, 10, 11, 21, 26, 32, 35, 36 e 37. Este resultado mostra que a identificação das amostras discrepantes por meio deste método é mais restritivo, visto que além das quatro amostras (10, 21, 26 e 37) consideradas discrepantes pelo primeiro método mais seis amostras forma identificadas como discrepantes. A diferenças observadas nos resultados pode ser explicado pelo fato que no primeiro método a presença de amostras discrepantes provoca redução nas distâncias de Mahalanobis em relação ao centro do grupo, uma vez que estas distâncias são ponderadas pela variabilidade dos dados. Também, em virtude disso, aumenta-se a probabilidade das amostras terem distância de Mahalanobis menor que o valor crítico. Portanto, o método de detecção de amostras discrepantes baseadas no MCD foram mais eficazes na detecção das amostras discrepantes na base de dados apresentada neste trabalho.

3. CONCLUSÕES

A aplicação de dois métodos para identificação de amostras discrepantes no espaço multidimensional por meio da distância de Mahalanobis clássicas e robustas permitiu a identificação de 4 e 10 amostras, respectivamente. O resultado mostrou que a do método de distância de Mahalanobis Robusta é mais eficiente para obtenção destas amostras discordantes.

4. REFERENCIAS BIBLIOGRÁFICAS

FILZMOSER, P., GARRET, R. G., REIMANN, C. Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31, 579-587, 2005.

HADI, A. Identifying multiple outliers in multivariate data. Journal Royal Statistics Society , 54, 761-771, 1992.

HARDIN, J., ROCKE, D. M. Outlier detection in multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis, 44, 625-638, 2004.

OLIVEIRA, P. T. M. MUNITA, C. S. Influência do valor crítico na detecção de valores discrepantes em arqueometria, 48 Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, 7-11 de julho de 2003. Lavras – MG.

PENNY, KAY I., Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. In: Applied Statistics, 35, Royal Statistical Society, UK, 1987. p.153-162.

REIMANN, C., BANKS, D., KASHULINA, G. Processes influencing the chemical composition of the O – horizon of podzols along a 500 km north-south profile from coast of the Barents Sea the Arctic Circle. Geoderma 95, 113-139, 2000.

ROUSSEEUW, P. J., VAN ZOMEREN, B. C. Unmasking multivariate outliers and leverage points. Journal of American Statistical Association, 85 (411), 633-651, 1990.

ROUSSEEUW, P. Least median of squares regression. Journal of American Association, 79, 871-880, 1984.

SAYRE, E. V. Brookhaven Procedures for statistical analyses of multivariate archaelmetric data. Brookhaven National Laboratory Report BNL-21693, New York, 1975.