Bol.IG-USP, Sér.Esp., 13:45-63, 1992.

MÉTODOS ESTATÍSTICOS PARA AGRUPAMENTO NÃO HIERÁRQUICOS DE DADOS

E.Acarini[1]
G.Amaral[2]

RESUMO

Os métodos estatísticos de análise de agrupamento (Cluster Analysis) possibilitam estudar um conjunto de dados subdividindo-o em diversos grupos, com características internas semelhantes, e distintas entre si. Os métodos K-médias e Isodata procuram estabelecer tais relações sem qualquer base hierárquica influindo no processo de formação de grupos. Isto possibilita uma grande mobilidade para as unidades de dados analisadas e a conseqüente formação de grupos com relações internas mais próximas. O método K-médias utiliza-se de distâncias euclideanas entre as médias das variáveis de um grupo, em relação a cada unidade do conjunto de dados, procurando estabelecer K grupos com as menores distâncias internas possíveis. O método Isodata opera com valores paramétricos, permitindo a individualização de um número qualquer de grupos.

O presente trabalho procura apresentar dois programas desenvolvidos em FORTRAN 77, Centro de Processamento de Dados do Instituto de Geociências da Universidade de São Paulo, e uma breve análise dos resultados obtidos por programas quando aplicados a um exemplo real.

ABSTRACT

[1]Bolsista de Iniciação Científica (FAPESP), Instituto de Geociências/USP, São Paulo.
[2]Departamento de Paleontologia e Estratigrafia, Instituto de Geociências/USP, São Paulo.

A cluster is usually thought to be a subset of data points which are highly similar or associated and relatively unassociated with data points outside the subset. Cluster Analysis makes it possible to study data sets and subdivide them into several distinct clusters with similar internal characteristics. The K-means and ISODATA algorithms can establish this relationship, without any hierarchical base influencing in the cluster formation process. This gives great mobility to the data units analyzed and facilitates the consequent formation of clusters with the closest internal relationships. The K-means algorithm uses Euclidean distances between the means of the variables of a cluster with respect to each unit of the data set and estabilishes K clusters for the smallest interval distance possible. In the ISODATA technique (which operates in the same way) each pattern is put into cluster for which the squared distance between it and the cluster mean is smallest. The new clusters means are then computed and the whole procedure repeated.

INTRODUÇÃO

Os métodos estatísticos de análise de agrupamento (Cluster Analysis) surgiram da necessidade de classificar, reunir e separar os objetos em estudo de acordo com suas características mensuráveis.

Uma aplicação, que se tem mostrado bastante útil em geologia, é a subdivisão em grupos de amostras analisadas quimicamente. As informações obtidas a partir deste agrupamento podem permitir relacionar cada grupo a unidades litológicas com características geoquímicas próprias.

FUNDAMENTOS TEÓRICOS

Inicialmente serão explicados os conceitos básicos envolvidos nos programas.

Uma população é um conjunto de objetos ou indivíduos, finitos ou infinitos, que constitui o conjunto universo. Para se analisar uma população não são necessários dados de toda população, mas apenas de uma pequena parte, chamada amostra.

Os métodos estatísticos permitem o estudo de uma população (ou amostra), cuja distribuição apresenta certas características, conhecidas como parâmetros, tais como: o ponto médio, as medidas de simetria da distribuição e as medidas que indicam o valor da dispersão.

Os programas utilizam como um dos parâmetros a média aritmética, que é definida como a soma de todas as observações dividida pelo número de observações.

$$\bar{X} = \frac{\sum_{i=1} X_i}{N}$$

Um outro parâmetro é a variância, que mede a dispersão dos elementos de uma amostra em relação a média

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$$

X = média

N = números de elementos

e o desvio padrão, que é a raiz quadrada da variância.

Na natureza é comum serem encontrados conjuntos de populações com distribuição estatística normal (ou distribuição de Gauss) definida pela equação

$$Y = \frac{1}{\sigma \sqrt{2\Pi}} e^{-1/2(x-\mu)^2/\sigma^2}$$

porém outros tipos de distribuição (binominal, Poisson, etc) também podem ser encontradas.

As populações podem, no entanto, apresentar dificuldades quando se procura separá-las. Para tal, utiliza-se a análise de variância, que proporciona a caracterização das diferenças existentes entre as populações do conjunto.

A Figura 1 assume que a variável analisada possui distribuição estatística normal, com a mesma variância em cada população. Nota-se que todas as populações são centradas em pontos diferentes e possuem distribuição contínua.

Novas técnicas, reunidas sob a designação de análises multivariantes, procuram estudar grupos diferentes de populações em um espaço multidimensional. A diferença entre a análise de variância e a análise multivariante é que aquela possui uma única variável dependente e esta possui

um vetor de variáveis. A população agora não está mais centrada em torno de um grupo, mas em torno de um vetor de variáveis (Fig. 2).

Os parâmetros utilizados pela análise multivariante são extrapolações dos parâmetros da análise de variância (ANOVA).

Os métodos K-médias e ISODATA constroem grupos com técnicas da análise multivariante. Estes dois métodos são hierárquicos de agrupamentos, e são assim denominados porque agrupam unidades de dados em vários grupos, que podem ser especificados a priori (K-médias), ou determinados como parte do método de agrupamento (ISODATA). Diferem dos métodos hierárquicos por não fornecerem uma "árvore" de classificação.

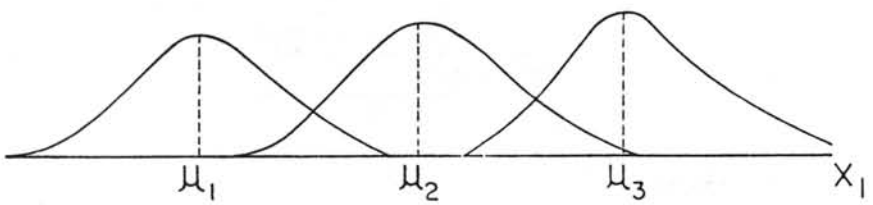

Figura 1 - Populações diferenciadas pela análise de variância.

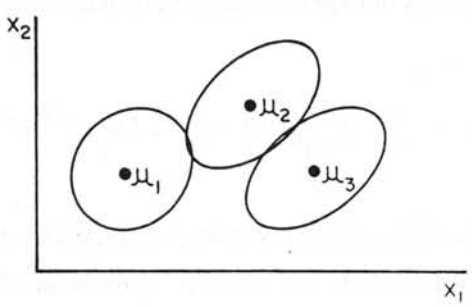

Figura 2 - Populações distinguidas pela análise multivariante (apenas duas variáveis).

ESTRUTURAS DOS DADOS

A estrutura dos dados é constituída de um número de casos (ou amostras), e cada caso é constituído de uma série de variáveis (por exemplo: as porcentagens em óxidos de uma amostra de rocha).

NORMALIZAÇÃO E PADRONIZAÇÃO

Muitas vezes é necessário destacar variáveis que apresentem valores extremamente pequenos e compará-los a variáveis como valores mais expressivos. Um exemplo ilustrativo é apresentado na Tabela 1. Trata-se de uma série de análises químicas de rochas carbonáticas do Permiano da Bacia do Maranhão (COIMBRA, 1983). Os resultados analíticos estão em **ppm** e evidenciam uma grande diferença entre as variáveis. Destacam-se de um lado os elementos traços, como o Ni, Zr, Sr e de outro os elementos maiores, como o Ca, Al, Mg.

Para comparar os resultados destas análises seria necessário normalizar, criar um critério para equilibrar o peso das variáveis. Um modo de se fazer isto é estabelecer o valor 100 para o maior valor da variável e o valor 0 (zero) para o menor valor, e interpolar os demais valores dentro deste intervalo. Outro modo é o da padronização estatística, onde cada valor é subtraído de sua média e dividido pelo desvio padrão, resultando numa nova variável com média zero e variância unitária.

O MÉTODO K-MÉDIAS

O objetivo deste algoritmo é pesquisar a melhor partição de dados, em que cada grupo formado abrangerá os casos mais próximos em relação a sua média. Este método foi desenvolvido inicialmente por MacQUEEM (1967) e modificado posteriomente por ANDERBERG (1973).

Considerações iniciais

O I-ésimo caso da J-ésima variável possui o valor A (I,J) {I = 1...M, J = 1...N}. Vamos supor que exista preliminarmente uma partição dos M casos em K agrupamentos, P (M,K). Cada grupo é representado pela média das variáveis dos casos que o compõe, MÉDIA (L,J). Local (I) designa o grupo ao qual pertence o caso I. O número de casos no grupo L é N (L) e a distância euclideana entre o Lésimo grupo e o Iésimo caso é dada por:

$$D(I,K)=\sqrt{\sum_{j=1}^{N} (A(I,J) - MÉDIA(L,J))^2}$$

O erro cometido por um determinado particionamento do conjunto de dados é dado por:

$$e = \sum_{I=1}^{M} D(I, LOCAL(I))^2$$

PROCEDIMENTOS

1 - calcular a média dos grupos e o erro da partição inicial;

2 - calcula-se, para cada caso I, a distância entre este caso e os grupos formados nesta partição;

3 - caso I, pertencente a grupo L(I), esteja mais próximo do grupo L, mude I do grupo L(I) para o grupo L. Atualize a média e corrija o erro;

4 - repita o passo 2 até I=M;

5 - se não ocorrer mais nenhum movimento de casos de um grupo para o outro, pare; caso contrário, volte ao passo 2.

O MÉTODO ISODATA

Dado um conjunto inicial de dados, o Isodata reparte este conjunto em grupos distintos, com base em valores paramétricos, buscando a melhor partição através da separação, combinação e eliminação dos grupos. Este método foi desenvolvido por BALL & HALL (1965).

Considerações iniciais

Inicialmente deve-se fornecer os parâmetros, que servirão de base para a separação, combinação e eliminação dos grupos. Estes parâmetros são: o desvio padrão máximo, DPMAX, que é o parâmetro acima do qual um grupo se separará; a distância mínima, DEMIN, abaixo da qual grupos distintos devem combinar-se; o número mínimo de pontos que um grupo deve conter, NMIN, abaixo do qual o grupo é eliminado. Também devem ser fornecidos o número de iterações a ser realizado, NITER, e a seqüência de passos a serem executados.

Tabela 1 - Análises químicas das rochas carbonáticas da Bacia do Maranhão (com algumas descrições litológicas).

AMOST	P	Al	B	Ba	Ca	Cr	Fe	K
	Mg	Mn	Na	Pb	Zn	Ni	Zr	Sr
	Ti	V			**********LITOLOGIA**********			
62C1	242.	5522.	283.	646.	190580.	270.	6202.	4250.
	102220.	2708.	1189.	92.	1111.	29.	45.	193.
	241.	96.4		Doloesparito				
13	3314.	33400.	80.	1480.	134880.	63	20980.	16429.
	5604.	550.	513.	156.	171.	0.050	110.	139.
	1228.	66.6		Arenito esparítico				
88	815.	33960.	359.	556.	154520.	362.2	19700.	21111.
	99260.	277.	277.	109.	305.	62.	51.	534.
	974.	116.0		*SDP				
* SDP sem descrição petrográfica								
28C	237.	21890.	122.	264.	110420.	176.6	10332.	17097.
	61560.	522.	277.	85.	167.	9.	55.	251.
	917.	56.2		SDP				
22C	172.	39500.	114.	355.	61340.	126.5	18219.	25556.
	29260.	756.	686.	07.	448.	6.	131.	199.
				Siltito silicoso dolomicroesparítico				
8G	302.	24620.	139.	184.	113400.	152.8	9346.	15161.
	68220.	674.	387.	116.	235.	12.	78.	57.2
	1242.	54.2		Doloesparito síltico silicoso				
64	0.05	26260.	54.	454.	56560.	122.8	10738.	18387.
	35880.	295.	1927.	69.	824.	9.	70.	173.
	1024.	72.6		Arenito silicoso doloesparítico				
37A	111.	50360.	384.	754.	8192.	76.	13124.	22222.
	18699.	204.	0.05	114.	156.	16.	150.	60.
	2426.	57.2		Siltito micáceo silicoso				
37B	234.	40500.	114.	601.	33320.	99.2	9354.	18571.
	24220.	476.	89.	58.	341.	0.05	138.	77.
	2009.	48.9		SDP				
31A	639.	23560.	131.	1127.	98780.	172.8	8984.	15645.
	67940.	868.	246.	106.	249.	22.	73.	187.
	1101.	62.8		SDP				

89A	3398.	7796.	122.	219.	141720	196.8	5998.	4786.
	79780.	1728.	1550.	96.	147.	0.05	41.	883.
	245.	61.8	Doloesparico arenoso					
31D	307.	25140.	177.	300.	125620.	230.2	9910.	4143.
	80180.	855.	214.	115.	126.	42.	83.	74.
	1151.	63.6	Doloesparito siltico silicoso					
62C2	0.05	1595.	88.	256.	78240.	73.2	3374.	9565.
	24000.	340.	2432.	43.	232.	0.05	31.	221.
	346.	60.4	Silexito arenoso doloesparítico					
30B	163.	10088.	351.	107.	120500.	80.8	4738.	6521.
	20640.	329.	0.05	70.	170.	0.05	37.	97.
	244.	29.8	Silexito arenoso esparítico magnesiano					
28E	0.05	19966.	55.	157.	84760.	133.2	9730.	11935.
	55400.	518.	0.05	100.	84.	0.05	56.	60.
	990.	49.2	SDP					
41	117.	2024.	46.	950.	366600.	64.2	1046.	33333.
	2700.	675.	0.05	30.	50.	0.05	5.	170.
	97.	3.4	Esparito					
58C	88.	33440.	63.	421.	45040.	117.8	4732.	1000.
	26420.	235.	497.	54.	57.	19.	108.	130.
	1373.	116.4	SDP					
28B	77.	25380.	207.	159.	94460.	176.8	12894.	18871.
	55720.	499.	0.05	105.	108.	0.05	53.	428.
	1034.	57.4	SDP					
58A	183.	22240.	88.	283.	93100.	166.6	8206.	13870.
	57700.	464.	717.	77.	81.	17.	65.	191.
	1011.	42.4	Arenito doloesparítico silicoso					
62F	88.	15558.	131.	361.	141100.	171.8	5056.	9673.
	79160.	1169.	2542.	94.	180.	0.05	57.	133.
	649.	65.6	Doloesparito arenosiltoso					
90	292.	16528.	182.	280.	154720.	228.	17064.	10645.
	89060.	2048.	57.	91.	301.	9.	33.	106.
	735.	84.2	Doloesparito silicoso					
22B	2380.	49760.	207.	464.	46760.	78.	18622.	24444.
	19384.	50.	183.	121.	123.	17.	130.	120.
	2396.	79.6	Siltito silicoso dolomicroesparítico					
89B	314.	6484.	105.	201.	160280.	174.4	7024.	4143.
	92000.	1780.	937.	55.	102	9.	38.	846.
	221.	46.2	Doloesparito					
38B	0.05	54760.	46.	302.	9798.	35.8	20380.	1000.
	15686.	301.	450.	136.	408.	3.	138.	76.
	1016.	71.4	SDP					

30A	0.05	21380.	46.	102.	42560.	25.8	10122.	15161
	30160.	262.	0.05	0.05	52.	0.05	34.	138.
	950.	0.05	SDP					
28D	75.	14264.	129.	129.	100840.	131.0	8580.	11451.
	39800.	531.	4916.	168.	563.	6.	85.	410.
	950.	99.	SDP					
35C	0.05	11172.	144.	443.	284800.	95.	7186.	5434.
	41580.	490.	3752.	83.	127.	16.	64.	146.
	572.	106.4	Esparito magnesiano					
38A	13.	211980.	122.	301.	69520.	92.	9928.	12741.
	7490.	375.	3958.	146.	122.	26.	114.	558.
	1459.	80.2	SDP					
31B	0.05	11172.	114.	230.	80180.	144.	6100.	6413.
	52620.	490.	3028.	85.	116.	9.	63.	106.
	621.	74.4	Silexito arenoso doloesparítico					
62D	138.	22160.	80.	295.	47820.	97.	9118.	18871.
	26580.	223.	6112.	126.	85.	13.	91.	136.
	1156.	92.6	Arenito doloesparítico silicoso					
35D	28.	18000.	207.	353.	210800.	42.	14716.	12742.
	6598.	1311.	3140.	134.	105.	36.	67.	559.
	1219.	102.8	Microesparito siltico					
58H	822.	12516.	156.	258.	103460.	157.	5900.	9022.
	67400.	606.	2212.	134.	71.	36.	68.	153.
	583.	68.6	Silexito doloesparítico					
58B	12.	5750.	359.	193.	123960.	218.0	2582.	1714.
	62640.	327.	1944.	74.	53.	26.	52.	274.
	344.	456.2	Biodoloesparito silicoso					
28F	22.	15030.	46.	84.	71980.	108.	7526.	8281.
	43420.	402.	2054.	116.	68.	29.	56.	51.
	648.	64.	SDP					
35A	0.05	11904.	241.	440.	274000.	27.	6880.	6630.
	5714.	1318.	2188.	48.	87.	0.05	63.	594.
	592.	89.80	Esparito					
22C	194.	28100.	377.	706.	65380.	137.	15504.	27778.
	27900.	1091.	2242.	161.	170.	23.	170.	252.
	2158.	194.4	Siltito arenoso dolomicroesparítico silicoso					
79	135.	15992.	750.	2084.	98540.	178.0	7716.	17903.
	59880.	449.	4668.	219.	226.	13.	70.	154.
	792.	76.2	Doloesparito arenoso					
35B	109.	2398.	106.	266.	28700.	41.	9110.	9565.
	9722.	1185.	2526.	78.	110.	18.	48.	149.
	1087.	62.2	Micrito					

32A	359.	19836.	55.	171.	178400.	270.	10164.	11451.
	104600.	1347.	749.	66.	77.	32.	39.	413.
	741.	79.	Dolodismicrito					
36B	0.05	8316.	21.	223.	136180.	147.	4322.	3607.
	53660.	1041.	0.05	25.	91.	18.	21.	149.
	341.	0.05	SDP					
52A	119.	7746.	87.	767.	205800.	232.	3272.	4035.
	98360.	4778.	1299.	76.	373.	33.	30.	109.
	275.	88.4	Dolomicrito					
58G	227.	6020.	301.	145.	184360.	210.	2682.	5217.
	76540.	943.	984.	84.	61.	0.05	30.	190.
	288.	48.2	Esparito dolomítico silicoso					
28A	145.	311220.	72.	178.	52460.	212.	13370.	21111.
	34560.	330.	1676.	66.	60.	29.	90.	77.
	1473.	75.2						

PROCEDIMENTOS

0 - inicializa (lê uma matriz de dados e os parâmetros, calcula a média e os desvios padrões iniciais);

1 - calcula a distância euclideana entre os grupos;

2 - muda a estrutura dos grupos;

2.1 - elimina - destrói grupos;

2.2 - separa se for uma iteração do tipo (1) e vai para 3, senão continua;

2.3 - combina

3 - Se é a última iteração numa seqüência Separa(1)-combina(0), senão vai para a próxima iteração, para 1.

EXEMPLO

Como exemplo de aplicação foram utilizados os dados de análises químicas de rochas carbonáticas do Permiano da Bacia do Maranhão (COIMBRA, 1983). Sobre estes dados foram

aplicados os dois métodos de agrupamento discutidos anteriormente, com as variáveis normalizadas e não normalizadas. Em seguida procurou-se indentificar os principais fatores e variáveis que influenciaram no resultado.

Apresentam-se, na Tabela 1, os dados que representam resultados de análises químicas de rochas da Formação Pedra de Fogo e Motuca.

ANÁLISE DOS RESULTADOS

Variáveis não padronizadas

K-médias

Ao aplicarmos o K-médias sobre estes dados foram realizadas várias sugestões de subdivisão de grupos (2, 3, 4, 5, 6, 9, 11). Todos os resultados mostraram grupos com relações internas em maior ou menor grau, dependendo unicamente do nível de detalhe exigido. Quando é sugerida uma subdivisão em apenas dois grupos, consegue-se separar as rochas carbonáticas com maior contribuição de terrígenos. Detalhando-se melhor esta subdivisão obtém-se :

AMOSTRAS

GRUPO 1: 22A, 22B, 22Ç 28A, 30A, 37A, 37B, 38B, 58Ç 62D, 64

GRUPO 2: 38A

GRUPO 3:28B, 28Ç 28D, 28E, 28F, 28G, 30B, 31A, 31B, 31D, 36B, 58A, 58B, 58H, 62C2, 79

GRUPO 4: 32A, 52A, 58G, 62C1, 62F, 88, 89A, 89B, 90

GRUPO 5: 13, 35D

GRUPO 6: 35A, 35B, 35Ç 41

Estes 6 grupos foram obtidos com as variáveis não padronizadas. Isto implica que as variáveis com valores maiores influenciaram mais o resultado, ou seja, os valores de Ca, Mg, Al, Fe, e K foram os fatores determinantes na subdivisão dos grupos.

A Tabela 2 procura definir os teores, em termos relativos, e as características de cada grupo, em termos dos 5 elementos que mais influenciaram na subdivisão.

Do resultado (Tabela 2) e da descrição litológica das amostras (COIMBRA, 1983), podemos caracterizar cada grupo como segue :

GRUPO 1: siltitos, arenitos e margas dolomíticas da Formação Pedra de Fogo.

GRUPO 2: argilas com muito pouco carbonato da Formação Pedra de Fogo.

GRUPO 3: dolomitos, dolomitos arenosos e arenitos dolomíticos da Formação Pedra de Fogo.

GRUPO 4: calcáreos dolomíticos da Formação Pedra de Fogo e Motuca.

GRUPO5: calcáreos calcíticos síltico da Formação Motuca e Formação Pedra de Fogo.

GRUPO 6 : calcáreos calcíticos da Formação Motuca.

Isodata

O Isodata propicia resultados em função da adoção de três parâmetros: desvio padrão máximo que uma variável pode assumir dentro de um grupo, a distância mínima entre dois grupos e o número mínimo de pontos que um grupo deve conter. Para as variáveis não padronizadas é fácil notar que estas possuem uma ampla distribuição, o que implica em desvios padrões e distâncias com valores elevados.

O resultado obtido pelo Isodata, adotando-se um desvio padrão de 1100, uma distância de 460 e um número mínimo de 2 unidades de dados para cada grupo, é:

Amostras ignoradas: 38A, 41, 58C

AMOSTRAS

GRUPO 1: 22A, 22B, 22Ç 28A, 28B, 28Ç 28D, 28E, 28F, 28G, 30A, 31A, 31B, 36B, 37B, 58A, 58B, 58H, 62C2, 62D, 64, 79

GRUPO 2: 35A, 35B, 35C

GRUPO 3: 13, 30B, 31D, 62F, 89A

GRUPO 4: 88, 89B, 90

GRUPO 5: 32A, 35D, 52A, 58G, 62C1

GRUPO 6: 37A, 38B

Como podemos observar pela Tabela 3, o Isodata deixa-se influenciar marcantemente pelas variáveis que possuem os maiores valores e em decorrência os maiores desvios padrões. Observa-se que a subdivisão dos grupos sofreu forte influência das variáveis Ca e Mg e, secundariamente, das variáveis Al, Fe e K.

Tabela 2 - Valores Teores relativos à mediana para os grupos obtidos pelo programa K-médias.

ELEM.	GRUPO 1	GRUPO 2	GRUPO 3	GRUPO 4	GRUPO 5	GRUPO 6
Ca	teores mais baixos	teor baixo	teores médios	teores médios	teores altos	teores mais altos
Mg	teores médios	teor baixo	teores médios a médios altos	teores mais altos	teores baixos	teores muito baixos a baixos
Al	teor alto	teor altíssimo	teores médios	teores baixos a médio-baixos	teores médios a altos	teores baixos a médios
Fe	teores médios	teor médio-alto	teores médios a baixos	teores baixos a altos	teores altos	teores baixos
K	teores médios a altos	teor médio	teores médios e baixos	teores baixos	teores médios a altos	teores médios a altos

Tabela 3 - Teores relativos à mediana para os grupos obtidos pelo programa Isodata.

ELEM	GRUPO 1	GRUPO 2	GRUPO 3	GRUPO 4	GRUPO 5	GRUPO 6
Ca	teores baixos	teores mais altos	teores médios	teores médios	teores médios altos	teores mais baixos
Mg	teores médios	teores baixos	teores variáveis	teores altos	teores variáveis	teores médios baixos
Al	teores variáveis	teores baixos e médios baixos	teores variáveis	teores variáveis	teores baixos a médios	teores altos
Fe	teores variáveis	teores baixos a médios-baixos	teores médios baixos a altos	teores médios a altos	teores baixos a médios	teores altos e muito altos
K	teores variáveis	teores baixos a médios baixos	teores baixos	teores variáveis	teores baixos médios baixos	teores baixos a médios

Comparação dos resultados

O grupo 1, obtido pelo Isodata, é a soma entre os grupos 1 e 3 obtidos pelo K-médias, ou seja, é a união entre os dolomitos síltico e arenosos da Formação Pedra de Fogo.

O grupo 2, obtido pelo Isodata, é o grupo 6 do K-médias, representam os calcáreos calcíticos da Formação Motuca.

O grupo 3, obtido pelo Isodata, é constituído de doloesparitos e esparitos areno siltosos da Formação Pedra de Fogo.

Os grupos 4 e 5, obtidos pelo Isodata, compõem o grupo 4 do K-médias, e são constituídos por calcáreos dolomíticos das Formações Pedra de Fogo e Motuca.

O grupo 6, obtido pelo Isodata, é constituido por siltitos micáceos silicosos da Formação Pedra de Fogo. Destaca-se como característica principal os mais baixos teores em Ca.

Conclusão

Os resultados obtidos pelo K-médias e pelo Isodata mostraram-se muito similares, porém com características distintas. O Isodata separa os grupos ressaltando mais as variáveis com valores e desvios padrões maiores, enquanto no K-médias tal efeito é amenizado.

Variáveis padronizadas

Foram realizadas várias tentativas de subdividir em grupos os dados da Tabela 1. Os resultados são apresentados na forma de tabelas, que permitem a distinção das variáveis mais importantes na caracterização de cada grupo.

K-médias (fazendo K = 10)

AMOSTRAS

GRUPO 1: 30A, 30B, 36B, 62C2

GRUPO 2: 41

GRUPO 3: 28F, 31B, 58C 58G, 58H, 62F

GRUPO 4: 35A, 35B, 35Ç 35D

GRUPO 5: 22A, 28A, 28D, 37B, 38B, 62D, 64

GRUPO 6: 28B, 28C 28E, 28G, 31A, 31D, 58A, 90

GRUPO 7: 13, 22B, 38A

GRUPO 8: 32A, 58B, 89A, 89B

GRUPO 9: 22C 37A, 79

GRUPO 10: 52A, 88, 62C1

Isodata

Adotando-se desvio padrão máximo 26, distância mínima 23 e número mínimo por grupo igual a 2, obtém-se:

Amostras eliminadas: 13, 31A, 31B, 58B, 64

GRUPO 1: 28B, 28C, 28E, 28G, 58A

GRUPO 2: 31D, 38B, 62F, 90

GRUPO 3: 28D, 35C 62D, 79

GRUPO 4: 35A, 35B, 35D, 38A, 41

GRUPO 5: 32A, 52A, 62C1, 88

GRUPO 6: 89A, 89B

GRUPO 7: 22A, 22B, 30A, 37A, 37B

GRUPO 8: 22C 28A, 36B, 58C

GRUPO 9: 28F, 58H

GRUPO 10: 30B, 58G, 62C2

Comparação e conclusão

Dos resultados obtidos pelos dois métodos aplicados a variáveis normalizadas conclui-se: o grupo 1 obtido pelo Isodata apresenta todas as unidades de dados dentro do grupo 6 do K-médias. Observa-se nas Tabelas 4 e 5 que estes dois grupos apresentam um comportamento bastante similar para os diversos elementos químicos. A maioria destas amostras é descrita petrograficamente.

O grupo 1 obtido pelo K-médias agrupa as variáveis 30B e 62C1, que também estão agrupadas no grupo 10 obtido pelo Isodata. Estas amostras estão descritas petrograficamente como silexitos arenosos esparíticos magnesianos e silexitos arenosos doloesparíticos, respectivamente. Como pode ser observado na Tabela 1, os teores estão bastante próximos para a maioria dos elementos analisados (P, B, Ca, Fe, K, Mg, Mn, Zn, Zr, Sr, Ti).

As amostras 31D e 90 do grupo 2 do Isodata também estão contidas dentro do grupo 6 do K-médias e vários elementos possuem comportamento similar (P, Al, B, Ba, V, Pb, Ca, Cr, Ti, Fe, entre outros). Estas amostras são descritas petrograficamente como doloesparitos síltico

Tabela 4 - Tabela de teores relativos à mediana das variáveis (elementos químicos), dentro dos grupos obtidos pelo programa K-médias.

Grupo / Teor		1	2	3	4	5	6	7	8	9	10
BAIXO		Zr,Pb, Ni,Cr, Ti,V, P	Al,Ti, Pb,Zr, Cr,Mg, Ni,B,Zn, Fe,V		P,Al			Mg		Ca	
BAIXO	MÉDIO	Mg,Ca, Al,Sr		Na,Pb, Al,Fe, K	Cr, Mg	P,B,Ca, Cr,Mg	Na	Ca,Cr	Zr,Al Fe,K,Ti		Zr,Al, Ti
MÉDIO		Fe	P	Cr,Mg, Mn,Zn, Sr	Zr,Ba, K,Pb, Zn,Ti, V	Mn,Pb, V	Al,Ca, Pb,V,B K,Mn,Zr	Mn,V, Zn	Ba,Ca, Pb,Zn	P,Cr,Fe	Na,P
MÉDIO A	ALTO			B,Ba V	B,Fe,Mn, Sr	Al,Fe,K, Zr,Ti	Cr,Fe, Mg,Ti	Ba,Fe, K,Pb	Mg,Cr	Mg,Pb, Zr,Sr, Ti,K	P,B
ALTO			Ca, K		Ca,Na			Al,Zr, Ti	Na,Sr	B,Ba	Ba,Ca V,Cr,Mg Zn,Ni
VARIÔ	VEL	Mn, B, Na		P,Zr, Ca,Ti, Ni	Ni	Ba,Na, Zn,Ni, Sr	P,Ba,Sr, Ni,Sr	P,B,Na, Ni,Sr	P,B,Ni, V	Al,Mn, Na,Zn, V	Fe,K,Mn Sr

Tabela 5 - Tabela de teores relativos à mediana das variáveis (elementos químicos), dentro de cada grupo obtido pelo programa Isodata.

Grupo / teor	1	2	3	4	5	6	7	8	9	10
BAIXO				Cr,Mg	Zr	Al,Ti, K	Mg,Na, Ca	Al,Fe,	Al,Ti, Zn,Ti,	Ni,Fe K,Zr
BAIXO A MÉDIO	Ca,V, Na,Ba	K,Sr	Cr,Mg, Mn,P, Al,Fe	Pb,P, Zn	Pb,Al, Ti	Pb,Zr, Fe	Cr,Mn, Pb,Zn, P	Ca,Mg, Pb,Zn, P		Pb,Na
MÉDIO	Al,Cr, Mg,Mn, Pb,Zr	Ba,V	K,Ti		Na	B,Ba, Cr,Na, Zn,V		Cr,Fe,	B,Ba, Ca,Cr, Mg,Mn, Pb,Zr, Sr,V	Sr
MÉDIO A ALTO	Fe,K,Ti	Pb,Ti, Ca,Zn, Fe	Ni,Zr, B,Na, Pb,Sr, V	Ba,Ca,K, Mn,Sr,	P,Ba,Mn, Zn,Sr, V	Ca,Mg, Ni	Fe,Zn, Ti	Zr,V, Ti	Na	Ca,Cr, P
ALTO				Na	Ca,Cr, Mg,Ni	P,Mn, Sr	Al,K, Zr	Ni	Ni	B
VARIÓ VEL	P,B,Zn, Ni,Sr	P,Al,B, Cr,Na, Mn,Ni, Zr	Ba,Ca, Zn	Al,B,Fe, Ni,Zr, Ti,V	B,Fe,K		P,B,Ba, Ni	Al,B,Ba, K,Mn, Na,Sr	P	Ba,Mg, Mn,Zn, V

silicosos e doloesparitos silicosos, respectivamente.

O grupo 4 do Isodata também apresenta grande similaridade no comportamento dos teores relativos (Tabelas 4 e 5), com o grupo 4 do K-médias, ambos contendo as amostras 35A, 35B e 35D. No grupo 9 do Isodata e no grupo 3 do K-médias encontram-se as amostras 28F e 58H, que nas diversas partições obtidas sempre se apresentaram juntas. Isto geralmente ocorre entre amostras muito próximas. Já o grupo 5 do Isodata apresenta as amostras 52A, 62C1 e 88, que formam o grupo 10 do K-médias. Estes grupos apresentam teores relativos bastante similares para quase todos os elementos analisados.

Outras relações podem ser estabelecidas entre o conteúdo dos diversos grupos, porém não nos demoraremos mais estabelecendo estas relações. O leitor pode observar que quando as variáveis são padronizadas, o comportamento dos grupos não sofre mais a influência das variáveis (elementos) com valores mais expressivos, mas de todas as variáveis igualmente.

Até agora foi observado uma série de similaridades entre os grupos formados pelo K-médias e pelo Isodata. No entanto, uma breve análise mostra que existem algumas divergências entre os grupos obtidos por um método em relação aos grupos do outro. O que causa então estas divergências?

Um estudo mais pormenorizado entre os procedimentos destes dois métodos esclarece este comportamento. A diferença básica ocorre porque o Isodata elimina grupos com um pequeno número de unidades de dados. Com isto elimina as grandes distorções que afetam a distribuição interna dos grupos normalmente ocasionadas por um número pequeno de unidades de dados, cujas variáveis possuem valores distantes da maioria dos grupos formados durante o processo. O método K-médias não elimina grupos, apenas incorpora estas unidades de dados nos grupos existentes. Isto força o deslocamento dos centros dos grupos (dado pelo vetor médio) da região central, para pontos mais periféricos, permitindo a incorporação de novos elementos e ocasionando distorções no comportamento das distribuições internas de cada grupo.

AGRADECIMENTOS

Agradecemos ao Professor Dr. Armando Márcio Coimbra pela ajuda prestada durante a fase de interpretação dos resultados, ao Professor Dr. Thomas Fairchild pela revisão do abstract e à FAPESP pela bolsa de iniciação científica, que permitiu a elaboração deste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERBERG, M.R. (1973) **Cluster analysis for applications**. New York, Academyi. Press. 359p.

BALL G.H. & HALL D.J. (1965) ISODATA, novel method of data analysis and pattern classification. Stanford Res. Inst., Menlo Park, California, Tech. Rep.

BALL, G.H. & HALL, D.J. (1967) A clustering technique for summarizing multivariate data. **Behav. Sci., 12**:153-155.

COIMBRA, A.M. (1983) **Estudo sedimentológico e qeoquímico do Permo-Triássico da Bacia do Maranhão**. São Paulo. 2v. (Tese de Doutorado - Instituto de Geociências/USP).

COOLEY, W.W. & LOHNES, P.R. (1971) **Multivariate data analysis**. New York, John Willey. 364p.

DAVIS, J.C. (1973) **Statistics and data analysis in geology**. New York, John Willey. 550p.

DIXON, W.J. (1981) **BMDP Statistical Software 1981.** Berkeley, University of California Press. 727p.

EVERITT, B., (1974) **Cluster analysis.** London, Heinemann. 122p.

HARALICK, R.M. & DINSTEIN, I. (1971) **An iterative Clustering Procedure**. Systems, Man, and Cybernetics, july 1971, v.smc-1, number 3. IEEE Transaction on. New York, N.Y.

MacQUEEN, J.B. (1967) **Some methods for classification and analysis of multivariate observations.** Proc. Symp. Math. Statist and Probability, 5Th, 1., Berkeley. University of California Press, Berkeley. p.281-297.

TRYON, R.C. & BAILEY, D.E. (1970) **Cluster analysis.** New York, McGraw-Hill. 347p.