Quim. Nova, Vol. 30, No. 2, 481-487, 2007

Educação

RECONHECIMENTO DE PADRÕES POR MÉTODOS NÃO SUPERVISIONADOS: EXPLORANDO PROCEDIMENTOS QUIMIOMÉTRICOS PARA TRATAMENTO DE DADOS ANALÍTICOS

Paulo R. M. Correia*
Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, Av. Arlindo Bettio, 1000, 03828-000 São Paulo - SP, Brasil
Márcia M. C. Ferreira
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13084-971 Campinas - SP, Brasil

Recebido em 27/10/05; aceito em 23/6/06; publicado na web em 19/1/07

NON-SUPERVISED PATTERN RECOGNITION METHODS: EXPLORING CHEMOMETRICAL PROCEDURES FOR EVALUATING ANALYTICAL DATA. An activity for introducing hierarchical cluster analysis (HCA) and principal component analysis (PCA) during the Instrumental Analytical Chemistry course is presented. The posed problem involves the discrimination of mineral water samples according to their geographical origin. Thirty-seven samples of 9 different brands were considered and the results from the determination of Na, K, Mg, Ca, Sr and Ba were taken into account. Non-supervised methods for pattern recognition were explored to construct a dendrogram, score and loading plots. The devised activity can be adopted for introducing Chemometrics devoted to data handling, stressing its importance in the context of modern Analytical Chemistry.

Keywords: analytical chemistry; chemometrics; pattern recognition.

INTRODUÇÃO

Além do planejamento experimental, a estatística multivariada aplicada à química é freqüentemente utilizada no tratamento de dados analíticos[1-6]. Essa área da quimiometria desenvolve ferramentas computacionais que permitem explorar os resultados obtidos por meio de análises químicas, a fim de verificar a existência de similaridades entre as amostras que, por sua vez, correspondem às semelhanças na composição química. O reconhecimento de padrões, uma das principais vertentes do uso da estatística multivariada em química analítica[2-4,6], viabiliza a obtenção de mais informações quando comparado com os procedimentos univariados que são usualmente adotados.

O número de parâmetros analisados (variáveis) nos estudos de reconhecimento de padrões é elevado, e a representação gráfica de todo o conjunto de dados facilita a interpretação dos resultados. Alguns algoritmos foram desenvolvidos para elaborar gráficos que representem a maior quantidade possível das informações contidas em um conjunto de dados analíticos. Entre eles, destacam-se a análise por agrupamento hierárquico (HCA) e a análise de componentes principais (PCA)[1,4,6].

HCA e PCA permitem a visualização gráfica de todo o conjunto de dados, mesmo quando o número de amostras e variáveis é elevado. O uso desses algoritmos tem como objetivo principal aumentar a compreensão do conjunto de dados, examinando a presença ou a ausência de agrupamentos naturais entre as amostras. Ambos são classificados como exploratórios ou não supervisionados, visto que nenhuma informação com relação à identidade das amostras é levada em consideração[6]. A HCA busca agrupar as amostras em classes, baseando-se na similaridade dos participantes de uma mesma classe e nas diferenças entre os membros de classes diferentes. A representação gráfica obtida é chamada de dendrograma, um gráfico bidimensional independentemente do número de variáveis do conjunto de dados[4,6]. A utilização da PCA visa reduzir a dimensionalidade do conjunto de dados original, preservando a maior quantidade de informação (variância) possível. Essa redução é obtida por meio do estabelecimento de novas variáveis ortogonais entre si, denominadas componentes principais (PCs). Organizadas em ordem decrescente de importância, as PCs são combinações lineares das variáveis originais. Os gráficos obtidos representam as amostras em um sistema cartesiano onde os eixos são as PCs[7]. Tanto HCA quanto PCA permitem a interpretação multivariada de conjuntos de dados grandes e complexos por meio de gráficos bi ou tridimensionais. Estes gráficos apresentam informações que expressam as inter-relações que podem existir entre as variáveis, facilitando a interpretação multivariada do comportamento das amostras[4,6,7].

A apresentação dos resultados experimentais na forma de gráficos facilita a interpretação dos dados, visto que o ser humano é dotado de um potente sistema visual de reconhecimento de padrões. Por esse motivo, a identificação de grupos de amostras com características parecidas é quase imediata quando se utiliza HCA e PCA. Além disso, é possível verificar quais dos parâmetros analisados (variáveis) são os principais responsáveis pela formação dos grupos de amostras. A avaliação das PCs pode auxiliar no estabelecimento de uma assinatura química particular para cada grupo de amostras segregado após a PCA. Esse é o objetivo principal dos estudos de reconhecimento de padrões, que busca encontrar uma maneira de relacionar a identidade de uma amostra com suas características químicas.

O uso da composição química para verificar a similaridade entre amostras empregando métodos de reconhecimento de padrões é amplamente explorado, com a finalidade de garantir a autenticidade de produtos agroindustriais[8-11]. Os trabalhos pioneiros envolvendo o reconhecimento de padrões foram propostos no final da década de 70[12,13]. Amostras de vinho da uva Pinot Noir, provenientes da França e dos Estados Unidos, foram discriminadas por meio da composição química elementar[12] e da análise de algumas substâncias orgânicas[13]. Para isso, os resultados analíticos foram avaliados utilizando-se ferramentas quimiométricas para reconhecimento de padrões, confirmando que a combinação entre a química analítica e a quimiometria viabiliza a identificação da origem geográfica das amostras de vinho[12,13].

A autenticação de alimentos é um nicho de pesquisa estabelecido, que busca desenvolver procedimentos para controlar e assegu-

*e-mail: prmc@usp.br

rar a qualidade dos produtos agroindustriais, a partir das informações sobre composição química[8-13]. Recentemente, as questões relacionadas com a autenticação de alimentos têm despertado grande interesse devido aos problemas de adulteração, contaminação e utilização indevida de organismos geneticamente modificados. Adicionalmente, a necessidade de indicar informações confiáveis a respeito da composição química nos rótulos dos produtos industrializados, atestando sua qualidade, também intensificou os estudos dedicados à autenticação de bebidas e alimentos[14-18]. Alguns trabalhos encontrados na literatura exploram produtos tipicamente brasileiros, tais como sucos de frutas[15], cachaça[17,18] e café[17]. Nesse contexto, o desenvolvimento de novos procedimentos analíticos e o aprimoramento de ferramentas estatísticas para tratamento de quantidades crescentes de dados favorecem a atuação do químico na área de reconhecimento de padrões.

A água mineral de uma determinada fonte pode ser caracterizada por meio da sua composição inorgânica. Determinações multielementares por espectrometria de emissão ótica com fonte de plasma acoplado indutivamente (ICP-OES) são convenientes nessas situações, por permitirem determinar dezenas de elementos em poucos segundos[16,19]. Nesse caso, a obtenção de diferentes assinaturas químicas para amostras provenientes de diferentes localidades é possível devido às características minerais do solo e das rochas encontradas nas regiões próximas a cada fonte. A assinatura química permite verificar a ocorrência de adulteração de amostras de água mineral.

Frente à importância de realizar o tratamento de dados analíticos por meio de ferramentas quimiométricas, o presente trabalho propõe uma atividade para introduzir os métodos não supervisionados para o reconhecimento de padrões durante a disciplina de Química Analítica Instrumental. Para isso, elementos alcalinos (Na e K) e alcalino-terrosos (Mg, Ca, Ba e Sr) são determinados por técnicas instrumentais, a fim de verificar se é possível discriminar as amostras de água mineral em função de sua origem geográfica.

PROCEDIMENTOS

Amostragem

Amostras de água mineral (*I*=37) de 9 marcas e lotes diferentes foram adquiridas nos supermercados da região metropolitana de São Paulo. A origem geográfica declarada no rótulo foi considerada como critério de discriminação das amostras (Tabela 1). A existência de uma maior quantidade de amostras para as marcas A-D permitiu avaliar se é possível obter uma assinatura química para cada uma delas, a fim de discriminá-las a partir das determinações de alguns metais alcalinos e alcalino-terrosos. As amostras das marcas E-I foram incorporadas no estudo somente para verificar se elas são ou não similares às amostras A-D.

Tabela 1. Informações relativas às amostras de água mineral que foram utilizadas na avaliação quimiométrica por métodos não supervisionados de reconhecimento de padrões

Marca	Nº de amostras	Cidade/Estado	Latitude	Longitude
A	10	Mogi das Cruzes/SP	46º11'18"W	23º31'22"S
B	7	Campos do Jordão/SP	45º35'29"W	22º44'22"S
C	4	Águas da Prata/SP	46º43'00"W	21º56'12"S
D	9	São Lourenço/MG	45º03'16"W	22º06'59"S
E*	1	Campo Largo/PR	49º31'42"W	25º27'31"S
F*	2	Itu/SP	47º17'57"W	23º15'51"S
G*	1	Petrópolis/RJ	43º10'43"W	22º30'18"S
H*	1	Petrópolis/RJ	43º10'43"W	22º30'18"S
I*	2	São Paulo/SP	46º38'10"W	23º32'51"S

* As amostras E-I foram incluídas no estudo somente para verificar se existe ou não similaridade química com as amostras A-D.

Determinação instrumental dos metais alcalinos e alcalino-terrosos

A parte experimental dessa atividade pode ser executada durante as aulas referentes à espectrometria atômica do curso de Química Analítica Instrumental. As determinações de Na, K e Ca por espectrometria de emissão atômica com chama (FAES) e de Mg, Sr e Ba por espectrometria de absorção atômica com chama (FAAS) podem ser desenvolvidas pelos alunos. Alternativamente, alguns resultados podem ser fornecidos para os alunos (K, Ca, Sr e Ba), restringindo a parte experimental às determinações de Na por FAES e de Mg por FAAS. Desta forma, respeita-se a limitação de tempo que freqüentemente é imposta pelo calendário escolar.

Os dados apresentados no presente trabalho foram obtidos simultaneamente para os 6 elementos de interesse por ICP-OES. Os princípios teóricos e os procedimentos para a realização das determinações de K, Na, Mg, Ca, Sr e Ba podem ser encontrados em livros didáticos de Química Analítica Quantitativa[20,21].

Avaliação multivariada dos resultados analíticos

O tratamento dos dados analíticos referentes às amostras de água mineral foi realizado em duas etapas. Inicialmente, foram consideradas as concentrações de Na, K e Mg, a fim de introduzir a abordagem multivariada de análise de dados por meio de gráficos simples. Como foram utilizadas somente 3 variáveis, toda a informação contida no conjunto de dados foi representada em um sistema cartesiano de 3 eixos. Qualquer pacote computacional para elaboração de gráficos pode ser utilizado para essa finalidade, sendo que no presente trabalho optou-se pelo software Microcal Origin versão 5.0 (Microcal Software Inc., Northamptomn, MA, EUA). Posteriormente, foram utilizados todos os resultados analíticos obtidos para Na, K, Mg, Ca, Sr e Ba e, conseqüentemente, métodos não supervisionados de reconhecimento de padrões foram empregados para avaliar de maneira multivariada o conjunto de dados completo de 6 dimensões. Essa tarefa pode ser realizada com pacotes computacionais dedicados à quimiometria, bem como por meio de programação em ambiente MatLab (The Mathworks, Natick, MA, EUA). Os resultados apresentados a seguir foram obtidos por meio do software Pirouette versão 3.11 (Infometrix, Bothell, WA, EUA).

RESULTADOS E DISCUSSÃO

Pré-processamento dos dados

A primeira etapa da avaliação dos dados analíticos por meio de HCA ou PCA consiste na montagem de uma matriz **X** contendo todas as informações x_{ij} relacionadas com as amostras (dispostas em linhas) e com as variáveis (dispostas em colunas). O formato da matriz obtida, a partir da determinação de K, Na, Mg, Ca, Sr e Ba (j = 1, 2 ..., 6) nas amostras de água mineral (i = 1, 2 ..., 37), é **X** (37x6). Posteriormente, seleciona-se o método mais adequado para realizar o pré-processamento dos dados originais das análises químicas contidos na matriz. Essa etapa de preparação do conjunto de dados é crítica para obter sucesso no tratamento multivariado[6]. Considerando que no presente estudo todas as variáveis (K, Na, Mg,

Ca, Sr e Ba) têm igual importância para auxiliar na discriminação das amostras, optou-se pelo autoescalamento dos dados: os resultados obtidos para uma mesma variável (coluna) foram subtraídos do valor médio $\bar{x}_j$ e divididos pelo desvio-padrão do conjunto de resultados obtidos para essa variável (s_j). Esse cálculo é realizado para cada uma das amostras contidas no conjunto de dados original. A Equação 1 indica o cálculo matemático que foi realizado para cada um dos elementos de interesse (variáveis) da matriz de dados. A matriz de dados autoescalados para as amostras de água mineral está apresentada na Tabela 2. Esses são os valores utilizados para iniciar a análise dos dados experimentais.

$$\text{Autoescalamento } (as) \quad x_{ij(as)} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (1)$$

Tabela 2. Matriz de dados autoescalados que foi obtida para as amostras de água A-I

Amostra	K	Na	Mg	Ca	Sr	Ba
A1	-0,491	0,980	-1,129	1,323	-0,519	-0,189
A2	-0,476	0,753	-1,118	1,414	-0,311	0,684
A3	-0,487	0,900	-1,125	1,373	-0,229	0,492
A4	-0,488	0,946	-1,124	1,353	-0,216	0,690
A5	-0,480	0,919	-1,123	1,387	-0,181	0,671
A6	-0,491	0,890	-1,128	1,330	-0,291	0,582
A7	-0,470	1,019	-1,109	1,583	-0,119	0,332
A8	-0,423	0,761	-1,072	1,815	-0,096	0,212
A9	-0,475	1,001	-1,112	1,522	-0,282	0,919
A10	-0,476	1,014	-1,112	1,545	-0,037	0,548
B1	-0,631	-1,282	0,752	-0,270	-0,721	-0,687
B2	-0,631	-1,282	0,736	-0,339	-0,446	-0,791
B3	-0,631	-1,282	0,727	-0,325	-0,234	-0,788
B4	-0,632	-1,284	0,729	-0,342	-0,307	-0,806
B5	-0,632	-1,284	0,721	-0,360	-0,283	-0,775
B6	-0,631	-1,282	0,712	-0,348	-0,295	-0,860
B7	-0,632	-1,283	0,720	-0,342	-0,315	-0,823
C1	-0,502	-1,178	-0,217	-0,201	-1,395	-0,906
C2	-0,518	-1,207	-0,405	-0,380	-1,043	-0,950
C3	-0,514	-1,197	-0,335	-0,322	-0,671	-0,871
C4	-0,502	-1,180	-0,218	-0,183	-0,727	-0,963
D1	1,818	0,672	0,917	-0,776	-0,130	1,125
D2	1,593	0,663	0,820	-0,845	0,221	2,357
D3	1,963	0,787	1,045	-0,740	0,474	1,404
D4	1,805	0,680	0,924	-0,763	0,853	0,826
D5	2,020	0,874	1,115	-0,693	0,169	1,179
D6	1,597	0,615	0,805	-0,822	0,387	2,154
D7	1,617	0,543	0,766	-0,770	0,357	0,850
D8	1,594	0,555	0,738	-0,779	0,123	0,799
D9	1,563	0,518	0,768	-0,777	0,209	0,439
E1	-0,680	-1,283	2,093	-1,204	-0,383	-0,560
F1	-0,581	0,628	1,041	0,742	2,950	-0,522
F2	-0,582	0,602	1,033	0,732	3,188	-0,570
G1	-0,676	-1,258	-1,295	-1,328	-1,631	-1,343
H1	-0,677	-1,257	-1,297	-1,324	-1,333	-1,610
I1	-0,581	0,638	-1,127	-0,939	1,678	-1,127
I2	-0,577	0,583	-1,115	-0,946	1,586	-1,124

Avaliação multivariada dos resultados obtidos para Na, K e Mg

A primeira avaliação multivariada será feita com apenas 3 variáveis, permitindo que 100% da informação contida no conjunto de dados seja representada por meio de gráficos simples. Nesse caso, a matriz de dados passa a ter o formato 37x3, pois as colunas com resultados de Ca, Sr e Ba não serão consideradas. Gráficos bidimensionais combinando as 3 variáveis (MgxNa, Figura 1a; KxNa, Figura 1b e KxMg, Figura 1c) foram preparados a partir dos valores da matriz autoescalada. Cada um desses gráficos correlaciona 2 das 3 variáveis consideradas para a discriminação das amostras. Os eixos desses gráficos apresentam valores que estão entre -1,5 e 2,5 devido ao autoescalamento dos dados. O teor de Na foi o primeiro critério observado para analisar os gráficos (Figura 1), visto que existem amostras com elevado teor de Na (A, D, F e I) e com baixo teor de Na (B, C, E, G e H).

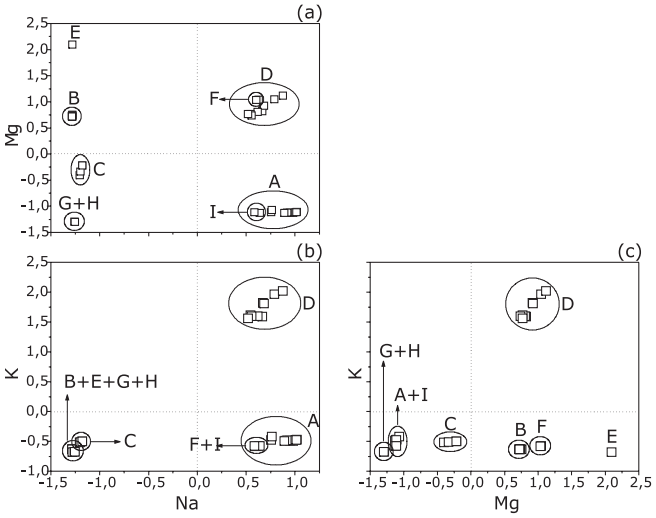

Figura 1. *Gráficos de correlação entre as variáveis (a) MgxNa, (b) KxNa e (c) KxMg para as amostras de água mineral (n=37)*

As amostras A, D, F e I aparecem destacadamente à direita nas Figuras 1a e 1b, pois possuem um teor mais elevado de Na que as demais. Por outro lado, as amostras D e F possuem maior teor de Mg que as amostras A e I, aparecendo na parte de cima do gráfico da Figura 1a. Na Figura 1b, as amostras D aparecem isoladas na parte superior do gráfico, indicando que possuem o maior teor de K entre todas as marcas consideradas no estudo. As amostras A, F e I apresentam teores baixos de K, aparecendo na parte inferior desse gráfico (Figura 1b). Na Figura 1c, verifica-se a confirmação das informações relacionadas com Mg e K, visto que as amostras D aparecem na parte superior à direita (altos teores de Mg e K), as amostras F aparecem na parte central e inferior (alto teor de Mg e baixo teor de K), e as amostras A e I aparecem na parte inferior à esquerda (baixos teores de Mg e K).

As amostras B, C, E, G e H são notadamente diferentes das amostras A, D, F e I, pois apresentam baixo teor de Na e aparecem à esquerda na Figura 1a. As amostras B e E estão dispostas na parte de cima do gráfico porque possuem maiores teores de Mg em comparação com as amostras C, G e H. Essa situação não é verificada no gráfico da Figura 1b, visto que as amostras B, C, E, G e H aparecem muito próximas na região inferior à esquerda. Além de não favorecer a discriminação, isso indica que os teores de K e Na nessas amostras são baixos e semelhantes. Essa situação melhora considerando-se as diferenças nos teores de Mg (Figura 1c), sendo que a amostra E apresenta os maiores teores, seguida pelas amostras B, C, G e H. Em outras palavras, a discriminação das amostras das marcas B, C, E, G e H só é possível se a variável Mg for considerada.

As principais informações extraídas dos gráficos bidimensionais da Figura 1 podem ser combinadas por meio de um gráfico

tridimensional, onde as variáveis Na, Mg e K são representadas nos eixos cartesianos (Figura 2). A avaliação desse gráfico mostra que é possível diferenciar as amostras de água mineral das marcas consideradas no presente estudo, excetuando-se os casos das marcas A/I e G/H. A limitação do número de variáveis consideradas nessa primeira análise multivariada dos dados não permitiu obter uma assinatura química única para cada uma das marcas de água mineral.

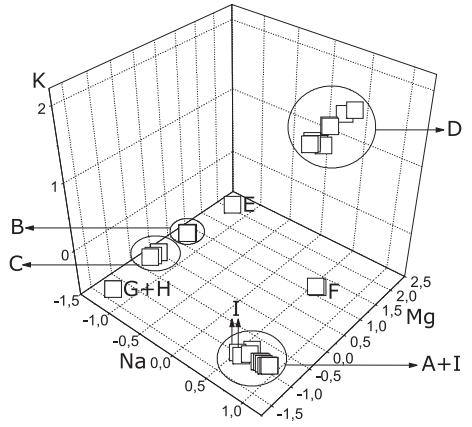

Figura 2. *Gráfico tridimensional de correlação entre as variáveis Na, K e Mg para as amostras de água mineral (n=37)*

Uma outra maneira de visualizar os agrupamentos naturais existentes entre as amostras é por meio de um dendrograma. Para sua construção, considera-se a distância entre as amostras no espaço amostral (Figura 2). Como foram considerados os valores obtidos para K, Na e Mg, pode-se calcular a distância euclidiana entre duas amostras quaisquer ("*a*" e "*b*") utilizando a Equação 2, onde *x*, *y* e *z* representam as coordenadas de uma amostra qualquer, para as 3 variáveis em questão. Esse cálculo foi realizado para as amostras utilizando os valores autoescalados para K, Na e Mg. Os valores de distância obtidos para as amostras 1 e 2 das marcas A-D são apresentados na Tabela 3.

Distância euclidiana $$d_{ab} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2} \qquad (2)$$

As distâncias calculadas entre amostras de uma mesma marca são bem menores (0,016 a 0,245) que os valores obtidos para amostras de marcas diferentes (0,967 a 3,276). Como o gráfico da Figura 2 agrupa amostras com teores semelhantes de Na, K e Mg, é possível constatar que distâncias pequenas implicam em amostras parecidas, geralmente de uma única marca. Desta forma, as distâncias calculadas (Tabela 3) podem auxiliar na busca por similaridade entre as amostras, sendo fácil verificar que a amostra B1 é muito parecida com a amostra B2 (d=0,016) e muito diferente da amostra D1 (d=3,137). Um dendrograma (Figura 3) foi obtido organizando-se as amostras no eixo y e o índice de similaridade no eixo x, sendo que as amostras são incluídas em função da sua proximidade: inicia-se incluindo as amostras mais próximas (similares), terminando com as amostras mais distantes (diferentes). O cálculo do índice de similaridade segue a Equação 3 e é feito depois que todas as amostras foram agrupadas, sendo d_{ab} a distância calculada entre duas amostras quaisquer ("*a*" e "*b*") e d_{max} a maior distância calculada entre as amostras. A vantagem de utilizar o índice de similaridade como escala ao invés da distância é que ele sempre varia entre 0 (se $d_{ab} = d_{max}$) e 1 (quando as amostras são idênticas)[6].

Tabela 3. Cálculos de distância euclidiana para algumas amostras do conjunto de dados, considerando 3 variáveis (Na, K e Mg)

	A1	A2	B1	B2	C1	C2	D1	D2
A1	0	0,228	2,945	2,935	2,343	2,304	3,100	2,871
A2	-	0	2,768	2,757	2,131	2,086	3,068	2,836
B1	-	-	0	0,016	0,983	1,165	3,137	2,955
B2	-	-	-	0	0,967	1,149	3,138	2,956
C1	-	-	-	-	0	0,191	3,177	2,976
C2	-	-	-	-	-	0	3,276	3,075
D1	-	-	-	-	-	-	0	0,245
D2	-	-	-	-	-	-	-	0

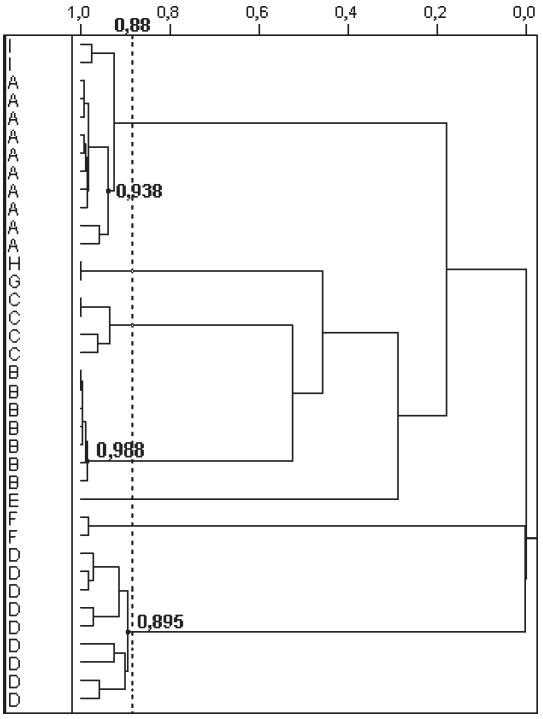

Figura 3. *Dendrograma obtido para as amostras de água mineral (n=37) por HCA, a partir das variáveis Na, K e Mg*

Índice de similaridade $$S_{ab} = 1{,}0 - \frac{d_{ab}}{d_{max}} \qquad (3)$$

Uma avaliação dos agrupamentos existentes no dendrograma obtido (Figura 3), considerando 0,88 como valor limite para o índice de similaridade, mostra que todas as marcas são adequadamente separadas, com exceção das amostras das marcas A/I e G/H. Esse fato novamente indica que, a partir dos resultados obtidos para as determinações de Na, K e Mg, não é possível discriminar as amostras dessas marcas. O alto índice de similaridade selecionado (próximo de 1) mostra que as amostras das marcas A/I e G/H são muito parecidas. Uma possível alternativa para melhorar o poder de discriminação das amostras de água mineral é a inclusão de mais variáveis no tratamento multivariado, que podem ser os teores de Ca, Sr e Ba.

Outros aspectos podem ser observados a partir de uma avaliação mais pormenorizada dos gráficos obtidos. Os agrupamentos das amostras B são menos dispersos do que aqueles verificados para as demais amostras (Figuras 1 a 3). Apesar de parecer somente 1 amostra, o agrupamento B contém 7 amostras diferentes, que são extremamente homogêneas e graficamente sobrepostas no gráfico tridimensional (Figura 2). Como conseqüência, o índice de similaridade para as amostras B é alto no dendrograma da Figura 3 (0,988).

A situação oposta pode ser verificada com as amostras D, que apresentam menor homogeneidade entre si e maior dispersão (Figura 2), com índice de similaridade igual a 0,895 no dendrograma. Além disso, é importante destacar que as amostras G e H sempre apareceram sobrepostas em qualquer um dos gráficos apresentados (Figuras 1 e 2), e com índice de similaridade igual a 1,0 no dendrograma (Figura 3). Apesar de serem de marcas distintas, elas apresentam um elevado grau de semelhança considerando-se os teores de Na, K e Mg. A explicação para isso reside no fato dessas marcas serem produzidas utilizando água mineral da mesma fonte, ou seja, a origem geográfica dessas marcas é a mesma (Tabela 1).

Avaliação multivariada dos resultados obtidos para Na, K, Mg, Ca, Sr e Ba

A utilização de um número de variáveis maior que 3 impede a representação gráfica direta dos dados, como empregada anteriormente. Nesse segundo momento, há 6 variáveis e o espaço amostral agora possui 6 dimensões, ou seja, a matriz de dados a ser considerada apresenta o formato 37x6. A representação gráfica do conjunto de dados, que auxilia na identificação de agrupamentos naturais de amostras, deve ser realizada por meio da HCA ou da PCA.

A PCA é utilizada para transformar dados complexos, visando explicitar as informações mais importantes para facilitar sua interpretação. Caso existam correlações significativas entre as 6 variáveis consideradas (Na, K, Mg, Ca, Sr e Ba), é possível encontrar novas variáveis (PCs), em número menor que as 6 iniciais, que sejam capazes de descrever, aproximadamente, toda a informação contida nos dados originais. Esta redução do número de variáveis é denominada compressão dos dados e é obtida através da combinação linear das variáveis originais, que busca agrupar aquelas que fornecem informações semelhantes[2,4].

A Tabela 4 apresenta as correlações entre as 6 variáveis iniciais consideradas nesse estudo (Na, K, Mg, Ca, Sr e Ba). Essa informação será utilizada para ilustrar como as PCs serão formadas. As correlações mais altas ocorrem entre as variáveis Na, Ca, Mg e Ba. Portanto, é possível considerar que essas 4 variáveis possam ser combinadas para formar uma única PC. Por outro lado, K apresenta correlação com Ba e ambos podem ser combinados para formar uma outra PC. Já o Sr não apresenta correlação significativa com nenhuma das outras variáveis iniciais, sendo possível utilizá-lo para formar uma terceira PC. A partir das correlações entre as variáveis, os 6 elementos podem ser divididos em 3 grupos diferentes: grupo 1 (Na, Ca, Mg e Ba), grupo 2 (K e Ba) e grupo 3 (Sr). Essa é uma indicação de que apenas 3 novas variáveis (PCs) serão suficientes para descrever grande parte da informação original dos conjunto de dados, havendo uma compressão do espaço amostral de 6 dimensões (Na, K, Mg, Ca, Sr e Ba), para um novo espaço de 3 dimensões (PC1, PC2 e PC3).

A primeira PC (PC1) é definida pela direção que descreve a máxima variância dos dados originais. A segunda PC (PC2) tem a direção de máxima variância dos dados no subespaço ortogonal à PC1, e as PCs subseqüentes são ortogonais às anteriores e orientadas de tal maneira que descrevam sempre a máxima variância restante. Pela própria maneira como estas novas variáveis são definidas, é possível descrever quase toda a informação contida nos dados originais utilizando poucas PCs. Isso permite representar as amostras usando um espaço cuja dimensão *A* é bem reduzida se comparada à dimensão do espaço que descreve os dados originais. No presente caso, há uma grande chance de *A* ser igual a 3. Cabe ressaltar que as relações entre as amostras não são alteradas por esta transformação de eixos.

Tabela 4. Valores de correlação calculados para as variáveis originais

	K	Na	Mg	Ca	Sr	Ba
K	1,0	0,423	-0,267	0,013	0,144	0,490*
Na	-	1,0	0,471*	0,627*	-0,123	-0,392
Mg	-	-	1,0	0,932*	0,238	-0,525*
Ca	-	-	-	1,0	0,434	-0,426
Sr	-	-	-	-	1,0	0,191
Ba	-	-	-	-	-	1,0

* Maiores correlações entre as variáveis originais, que ajudam a definir a composição das PCs.

Uma vez definidas as PCs, os dados originais são projetados neste novo sistema de eixos. Por isto, o método PCA é conhecido como um método de projeção, pois as amostras são projetadas em um espaço de dimensão menor.

Do ponto de vista matemático, a matriz dos dados originais ou pré-processados, $\mathbf{X}(I\text{x}J)$ é inicialmente decomposta em dois vetores, um de escores $\mathbf{t}_1$ e um de pesos ("loadings") $\mathbf{l}_1$ como mostrado na Equação 4.

$$\begin{bmatrix} \mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1 \end{bmatrix} \begin{bmatrix} \mathbf{l}_1^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} \mathbf{E}_1 \end{bmatrix} \tag{4}$$

O vetor $\mathbf{t}_1$ é formado pelas coordenadas de cada amostra na *primeira* nova variável (PC1) enquanto a coluna $\mathbf{l}_1$ contém a informação do quanto cada variável original contribuiu (seu peso) na formação da primeira PC (PC1). Os pesos podem variar entre +1 e –1 e são os co-senos dos ângulos entre PC1 e os eixos das variáveis originais. Valores elevados para os pesos indicam altas correlações, sendo que o ângulo entre PC1 e a variável original é pequeno. $\mathbf{E}_1$ é a matriz de resíduos que contém toda a informação original que não foi descrita por PC1. Esta matriz ($\mathbf{E}_1$) será utilizada para calcular a segunda PC (PC2), conforme mostrado na Equação 5. Ao utilizar a matriz de resíduos para cálculo da próxima PC, fica evidente uma propriedade importante das PCs: elas são completamente não-correlacionadas e ortogonais entre si.

$$\begin{bmatrix} \mathbf{E}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{t}_2 \end{bmatrix} \begin{bmatrix} \mathbf{l}_2^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} \mathbf{E}_2 \end{bmatrix} \tag{5}$$

Como mencionado anteriormente, *A* é o número de PCs necessário para descrever as informações relevantes do conjunto de dados e $\mathbf{E}_A$ é a matriz de resíduos final. O número máximo de PCs que podem ser calculadas é igual a 6, visto que esse é o menor entre os seguintes valores que dimensionam a matriz **X**: *I*=36 e *J*=6. A Equação 6 representa matematicamente a decomposição da matriz **X** em *A* PCs.

$$\mathbf{X} = \mathbf{T} * \mathbf{L}^{\mathrm{T}} = \mathbf{T}_A * \mathbf{L}_A^{\mathrm{T}} + \mathbf{E}_A \tag{6}$$

onde, $\mathbf{T}_A = [\mathbf{t}_1 \quad \cdots \quad \mathbf{t}_A]$ e $\mathbf{L}_A = [\mathbf{l}_1 \quad \cdots \quad \mathbf{l}_A]$

T é a matriz de escores e **L** é a matriz de pesos. Estas matrizes podem ser obtidas utilizando-se tanto o algoritmo NIPALS[22], quanto o método de decomposição de valores singulares (SVD)[4].

Outro ponto relevante na PCA diz respeito à quantidade de informação dos dados originais que cada uma dessas novas variáveis é capaz de descrever. Esta informação está contida nos escores. O produto $\mathbf{t}^{\mathrm{T}}_1 * \mathbf{t}_1$ é igual à variância dos dados originais, λ_1, descrita pela primeira PC. Portanto, a quantidade de informação contida nesta PC é dada, na Equação 7, pela porcentagem de variância

explicada, $(\%Var_1)$

$$\%Var_1 = \frac{\lambda_1}{\sum_{k=1}^{6}\lambda_k} \times 100 = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5+\lambda_6} \times 100 \quad \text{onde } \lambda_k \text{ é } \mathbf{t}_k^{T} * \mathbf{t}_k \quad (7)$$

A determinação do número de PCs (*A*) que devem ser utilizadas para se ter uma boa descrição do conjunto de dados, sem perder informações relevantes por um lado, nem incluir resíduos por outro, é muito importante e há várias maneiras de fazê-lo[2,4].

Os resultados da PCA são visualizados na forma de gráficos, facilitando a identificação de estruturas e agrupamentos existentes no conjunto de dados. As Figuras 4 e 5 foram obtidas por meio da PCA, a partir dos dados autoescalados. Os pesos encontrados para PC1, PC2 e PC3 foram, respectivamente

$$\mathbf{l}_1 = \begin{matrix} \text{K} \\ \text{Na} \\ \text{Mg} \\ \text{Ca} \\ \text{Sr} \\ \text{Ba} \end{matrix}\begin{bmatrix} -0{,}082 \\ 0{,}412 \\ 0{,}561 \\ 0{,}572 \\ 0{,}137 \\ -0{,}403 \end{bmatrix}, \mathbf{l}_2 = \begin{bmatrix} 0{,}723 \\ 0{,}289 \\ -0{,}059 \\ 0{,}185 \\ 0{,}383 \\ 0{,}458 \end{bmatrix}, \mathbf{l}_3 = \begin{bmatrix} -0{,}315 \\ -0{,}544 \\ 0{,}164 \\ 0{,}154 \\ 0{,}718 \\ 0{,}197 \end{bmatrix}$$

Deve-se notar que os altos pesos correspondem exatamente às variáveis indicadas como correlacionadas na Tabela 4. Utilizando estes pesos podemos calcular os escores de cada amostra usando a expressão $\mathbf{T}=\mathbf{X}*\mathbf{L}$ ($\mathbf{L}$ é uma matriz ortogonal, *i. e.*, $\mathbf{L}^{T}=\mathbf{L}^{-1}$). Uma vez obtidos os escores é possível calcular a quantidade de informação (variância explicada) contida em cada componente principal utilizando a Equação 7. No presente estudo, verifica-se que a porcentagem de variância explicada pela PC1 é

$$\%Var_1 = \frac{2{,}77}{2{,}77+1{,}58+1{,}23+0{,}34+0{,}077+0.0015} \times 100 = 46{,}2\%$$

Para PC2 e PC3 os resultados obtidos são 26,4% e 10,5%, respectivamente. Desta forma, a utilização de 3 novas variáveis (*A* = 3) permite representar 83% das informações originais do conjunto de dados (*A* = 6).

Os gráficos das Figuras 4 e 5 apresentam nos seus eixos as 2 PCs mais importantes, condensando graficamente mais de 70% da informação multivariada que pode ser extraída a partir dos dados analíticos. As informações relacionadas com as amostras de água mineral são apresentadas no gráfico de escores (Figura 4). Já a avaliação das variáveis (K, Na, Mg, Ca, Sr e Ba) pode ser feita no gráfico de pesos (Figura 5).

O aumento na quantidade de informações relacionadas com a composição química auxiliou o processo de discriminação das amostras

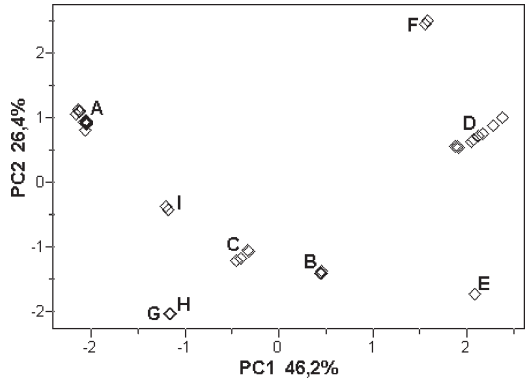

Figura 4. *Gráfico de escores das amostras obtido por PCA*

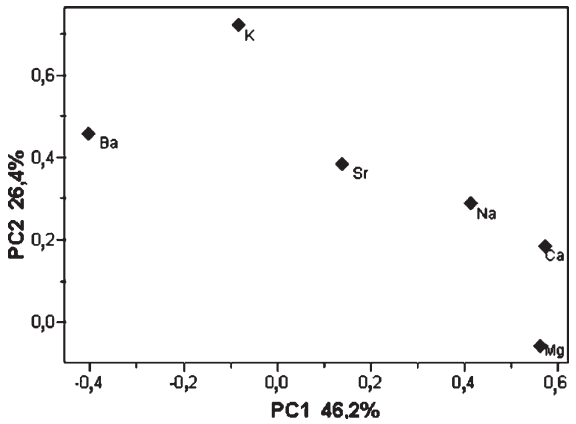

Figura 5. *Gráfico de pesos das variáveis obtido por PCA. Valores dos pesos para PC1 (Ca=0,572, Mg=0,561, Na=0,412, Ba=-0,403, Sr=0,137 e K=-0,082) e para PC2 (K=0,723, Ba=0,458, Sr=0,383, Na=0,289, Ca=0,185 e Mg=-0,059)*

de água mineral. A partir dos resultados para os 6 diferentes elementos, foi possível observar agrupamentos isolados para a maioria das marcas consideradas (Figura 4), em contraste com o verificado anteriormente (Figura 1). As amostras A e I foram convenientemente separadas, permitindo a diferenciação entre elas. Por outro lado, as amostras G e H, que são provenientes da mesma fonte hidromineral, continuaram muito próximas e não foram discriminadas.

O posicionamento das amostras no gráfico de escores (Figura 4) pode ser interpretado a partir do gráfico de pesos (Figura 5), que está relacionado com os elementos químicos considerados no estudo. O posicionamento dos agrupamentos de amostras é determinado pela PC1 (sentido horizontal) e pela PC2 (sentido vertical). As variáveis mais importantes para determinar o posicionamento horizontal das amostras no gráfico de escores (Figura 4) são aquelas que apresentam maiores pesos para PC1 (Figura 5): Ca (0,572), Mg (0,561), Na (0,412) e Ba (-0,403). As amostras localizadas à direita no gráfico dos escores tendem a apresentar maiores teores de Ca e Mg (variáveis localizadas à direita no gráfico dos pesos) e menores teores de Ba (variável localizada à esquerda no gráfico dos pesos). Esse é o caso das amostras D, E e F, que aparecem à direita (Figura 4). O raciocínio inverso é igualmente válido e, por esse motivo, é possível afirmar que as amostras A, G, H e I que aparecem à esquerda (Figura 4) possuem maiores teores de Ba e menores teores de Ca, Mg e Na. As demais amostras estão enquadradas em situações intermediárias. As variáveis mais importantes para determinar o posicionamento vertical das amostras no gráfico de escores (Figura 4) são aquelas que apresentam maiores pesos para PC2 (Figura 5): K (0,723) e Ba (0,458). Os teores desses elementos para as amostras B, C, E, G e H tendem a ser menores que aqueles verificados para as marcas A, D e F. A avaliação conjunta dos resultados analíticos para os 6 elementos permite identificar as variáveis Ca/Mg (PC1) e K/Ba (PC2) como as principais responsáveis pela discriminação das amostras A e I, que não ocorreu quando foram considerados somente os resultados de Na, K e Mg (Figura 2). A composição das amostras A apresenta maiores teores de Na e Ba, enquanto as amostras I possuem menores teores de Ba, e teores ligeiramente menores de Ca e Na. Assim, verifica-se que a inclusão dos resultados obtidos especialmente para Ba foi decisiva para separar essas 2 marcas em agrupamentos distintos.

Cabe ainda ressaltar que a terceira PC tem maior contribuição dos elementos Sr e Na (gráficos não incluídos). Esta PC, que apresenta 10,5% de variância explicada, discrimina as amostras F, que são as únicas que têm um alto teor de Sr, e as amostras D, que têm altos teores de Na.

A HCA considerando as 6 variáveis produziu um novo dendrograma (Figura 6), que pode ser comparado com aquele previamente obtido (Figura 3). Estabelecendo um índice de similaridade igual a 0,9, é possível perceber que as amostras das marcas A e I aparecem em agrupamentos distintos (Figura 6). Já as amostras G e H formam um único agrupamento em ambos os casos, pois o índice de similaridade delas é igual a 1 (Figuras 3 e 6).

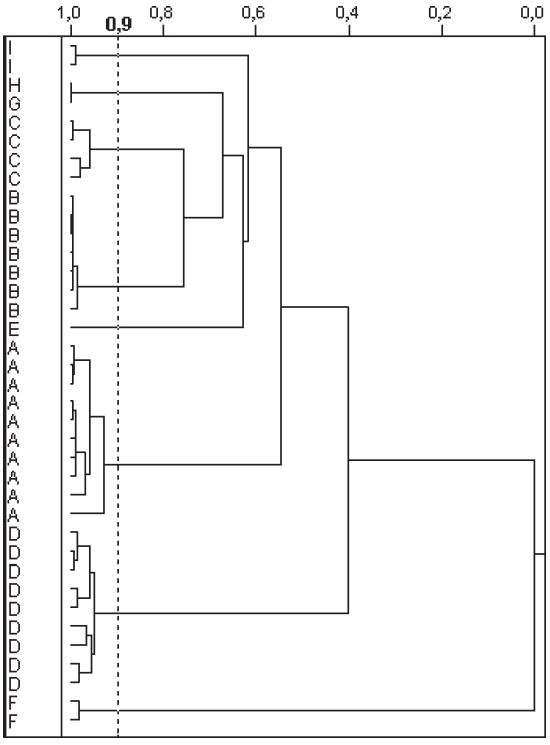

Figura 6. *Dendrograma obtido para as amostras de água mineral (n=37) por HCA, a partir das variáveis Na, K, Mg, Ca, Sr e Ba*

CONSIDERAÇÕES FINAIS

A atividade proposta introduz procedimentos multivariados visando o tratamento de dados analíticos, por meio de métodos não supervisionados de reconhecimento de padrões. A análise de agrupamentos hierárquicos (HCA) e a análise de componentes principais (PCA) são ferramentas quimiométricas amplamente utilizadas para essa finalidade, tornando explícitas as inter-relações entre as variáveis e permitindo a visualização das informações latentes que não seriam observadas através de um tratamento univariado.

A análise dos gráficos obtidos por HCA (dendrograma) e PCA (escores e pesos) é complementar e, por esse motivo, recomenda-se a utilização dessas 2 ferramentas nos estudos não supervisionados de reconhecimento de padrões. Além de aumentar a consistência das interpretações, o entendimento do conjunto de dados é facilitado e as inter-relações entre as variáveis ficam mais evidentes.

A possibilidade de processar uma grande quantidade de dados com o auxílio de recursos computacionais tem despertado o interesse das indústrias para essas ferramentas quimiométricas, que são utilizadas no controle de qualidade de produtos e no controle do processo industrial. A difusão do uso da quimiometria no setor produtivo e na área acadêmica justifica sua apresentação durante os cursos de graduação, por meio de atividades didáticas que motivem os estudantes a explorarem os cálculos computacionais de maneira crítica, evitando o uso automático e mecânico que transforma as ferramentas quimiométricas em uma "caixa preta".

A atividade proposta encaixa-se nesse contexto, permitindo adaptações às realidades de cada professor. A intenção foi oferecer uma seqüência de atividades que podem ser desenvolvidas pelos estudantes, valorizando aspectos conceituais. A mesma estrutura pode ser empregada utilizando-se amostras diferentes e resultados analíticos obtidos por meio de outras técnicas instrumentais.

AGRADECIMENTOS

P. R. M. Correia agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico pela concessão da bolsa de pós-doutoramento (CNPq 150325/2004-5). Os autores também agradecem aos Profs. Drs. P. V. Oliveira (IQ/USP) e E. Oliveira (IQ/USP) pela permissão em utilizar o instrumento para obtenção dos dados experimentais apresentados nesse estudo.

ACRÔNIMOS

FAAS: Flame Atomic Absorption Spectrometry
FAES: Flame Atomic Emission Spectrometry
HCA: Hierarchical cluster analysis
ICP-OES: Inductively coupled plasma optical emission spectrometry
NIPALS: Non-linear Iterative Partial Least Squares
PC: Principal component
PCA: Principal component analysis
SVD: Singular value decomposition

REFERÊNCIAS

1. Neto, J. M.; Moita, G. C.; *Quim. Nova* **1998**, *21*, 467.
2. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.
3. Hopke, P. K.; *Anal. Chim. Acta* **2003**, *500*, 365.
4. Beebe K. R.; Pell, R. J.; Seasholtz, M. B.; *Chemometrics: a practical guide*, John Wiley & Sons: New York, 1997.
5. Teófilo, R. F.; Ferreira, M. M. C.; *Quim. Nova* **2006**, *29*, 338.
6. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R.; *Chemometrics*, John Wiley & Sons: New York, 1986.
7. Christie, O. H. J.; *Chemometr. Intell. Lab.* **1995**, *29*, 177.
8. Dennis, M. J.; *Analyst* **1998**, *123*, 151R.
9. Tzouros, N. E.; Arvanitoyannis, I. S.; *Crit. Rev. Food Sci. Nutr.* **2001**, *41*, 287.
10. Cordella, C.; Moussa, I.; Martel, A. C.; Sbirrazzuoli, N.; Lizzani-Cuvelier, L.; *J. Agric. Food Chem.* **2002**, *50*, 1751.
11. Lees, M; *Food authenticity and traceability*, Woodhead Publishing: Cambridge, 2003.
12. Kwan, W.; Kowalski, B. R.; Skogerboe, R. K.; *J. Agric. Food Chem.* **1979**, *27*, 1321.
13. Kwan, W.; Kowalski, B. R.; Skogerboe, R. K.; *J. Agric. Food Chem.* **1980**, *28*, 356.
14. Sloan, A. E.; *Food Tech.* **2003**, *57*, 26.
15. Ferreira, E. C.; Rodrigues, S. H. B. G.; Ferreira, M. M. C.; Nóbrega, J. A.; Nogueira, A. R. A.; *Eclet. Quim.* **2002**, *27*, 77.
16. Silva, F. V.; Kamogawa, M. Y.; Ferreira, M. M. C.; Nóbrega, J. A.; Nogueira, A. R. A.; *Eclet. Quim.* **2002**, *27*, 91.
17. Cardoso, D. R.; Andrade-Sobrinho, L. G.; Leite-Neto, A. F.; Reche, R. V.; Isique, W. D.; Ferreira, M. M. C.; Lima-Neto, B. S.; Franco, D. W.; *J. Agric. Food Chem.* **2004**, *52*, 3429.
18. Fernandes, A. P.; Santos, M. C.; Lemos, S. G.; Ferreira, M. M. C.; Nogueira, A. R. A.; Nóbrega, J. A.; *Spectrochim. Acta* **2005**, *60B*, 717.
19. Yabe, M. J. S.; de Oliveira, E.; *Quim. Nova* **1998**, *21*, 551.
20. Harris, D. C.; *Análise química quantitativa*, LTC: Rio de Janeiro, 2001.
21. Skoog, D. A.; Holler, F. J.; Nieman, T. A.; *Princípios de análise instrumental*, Bookman: Porto Alegre, 2002.
22. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.