UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

Orientador: Prof. Dr. Roy E. Bruns

ANÁLISES ESPECTROMÉTRICAS DE ÁGUAS MINERAIS E SUAS CLASSIFICAÇÕES POR MEIO DE RECONHECIMENTO PADRÃO

Ieda Spacino Scarminio

Tese de Mestrado

CAMPINAS - 1981

AGRADECIMENTOS

- Ao Prof. Dr. Roy E. Bruns, pela orientação, pelas discussões e pela amizade durante todo este trabalho.
- Ao Prof. Dr. Bruce R. Kowalski, pelas discussões estimulantes sobre reconhecimento de padrão e pela ajuda na análise dos dados.
- Ao Prof. Dr. José W. Martins e Paulo R. Queiroz, pela ajuda na operação do espectrofotômetro de absorção atômica.
- Ao Prof. Dr. Oswaldo E.S. Godinho e Prof. Dr. Luis M. Aleixo, pela utilização do laboratório.
- Ao Elias Zagatto e Antonio O. Jacintho, pela orientação nas análises das amostras.
- Ao meu esposo Jair, pelos excelentes desenhos, pelas discussões, ajuda na coleta das amostras, pela compreensão e incentivo durante todo este trabalho.
- Ao Sr. e Sra. Paulo Labatte, pelas amostras provenientes da Fonte Nossa Senhora Aparecida (Distribuidor).
- Aos responsáveis pelas engarrafadoras Lindóya, Nossa Senhora Aparecida e Fonte Mécia.
- À Diretoria do Instituto de Química da Universidade Estadual de Campinas, pelas facilidades concedidas para a realização deste trabalho.
- Ao Centro de Computação do Instituto de Matemática da UNICAMP.
- Ao Sr. William Kalaf, pelo excelente trabalho datilográfico e pela ajuda nas correções.

- À Janaina, Juliana e Bruns, pela ajuda na coleta das amostras.

- A todos aqueles que, de forma direta ou indireta, contribuíram para este trabalho.

- À Coordenação de Aperfeiçoamento de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelas bolsas de estudo.

- E, em especial, aos amigos Mozart, Eliana e Lucia.

Aos meus pais

e Jair.

ÍNDICE

Página

FIGURAS

Figura

Página

Figura

TABELAS

Tabela

Página

Tabela

Página

Tabela

RESUMO

Métodos estatísticos e de reconhecimento de padrão contidos no programa ARTHUR foram utilizados no problema de identificação de fontes de águas minerais e na verificação da autenticidade das águas comercializadas. Os dados foram obtidos pela determinação dos traços dos elementos naturais das mesmas, utilizando-se um espectrômetro de emissão atômica de plasma e um espectrofotômetro de absorção atômica. As amostras foram coletadas nas fontes naturais de Serra Negra, Lindóia e Valinhos. Os elementos com concentrações mais significantes foram: potássio, sódio, silício, cálcio, fósforo e magnésio e somente os quatro primeiros foram importantes na separação destas fontes. O método KNN mostrou 94,8% de confiança na discriminação das categorias de Serra Negra, Lindóia e Valinhos. Águas comercializadas foram testadas para as três categorias.

ABSTRACT

Statistical and pattern recognition methods contained in the ARTHUR program were used to identify mineral water sources and the authenticity of commercial mineral water samples based on their chemical analysis. The data consisted of trace element concentrations measured using an emission spectrometer with an inductively coupled plasma source and an atomic absorption spectrophotometer. The samples were collected at mineral water sources in Serra Negra, Lindóia and Valinhos of São Paulo State, Brazil. Elements with significant concentrations in these samples were potassium, sodium, silicon, calcium, phosphorus and magnesium. Only the first four of these elements are important in discriminating the different sources into the geographical categories, Serra Negra, Lindóia and Valinhos. Commercial samples from these three categories were also tested for their autheticity.

TERMINOLOGIA USADA

OBJETO - Uma amostra para qual medidas químicas, físicas e biológicas podem ser obtidas.

MEDIDAS - Qualquer variável que pode ser quantitativamente determinada para cada objeto.

CARACTERÍSTICAS - Podem ser variáveis ou medidas simples, algumas combinações de medidas ou transformações matemáticas destas medidas.

CATEGORIA - Objetos tendo uma mesma propriedade.

CONJUNTO DE TREINAMENTO - É um subconjunto do conjunto de dados tendo propriedades e categorias conhecidas usadas para desenvolver regras de classificações.

CONJUNTO TESTE - É um subconjunto dos dados tendo propriedades e categorias conhecidas usadas para testar a habilidade das regras de classificações desenvolvidas sobre o conjunto de treinamento.

CAPÍTULO I

INTRODUÇÃO

A análise de dados experimentais em química normalmente assume duas formas: primeiro um modelo teórico pode ser construido para descrever o problema usando quantidades fundamentais como, massa e carga eletrônica, massas dos núcleos, geometria molecular, etc. Segundo, o problema pode ser explicado usando analogias com um modelo pré-existente. Por exemplo, o conceito de orbitais moleculares, para descrever propriedades moleculares, foi baseado nos orbitais atômicos que sairam naturalmente da aplicação da mecânica quântica aos átomos. Contudo em muitas situações nenhuma destas formas são viáveis, e então, precisamos recorrer a métodos matemáticos e estatísticos empíricos, cujo processamento geralmente envolve técnicas computacionais. Convém observar que estes últimos métodos abrangem uma grande quantidade de problemas de análises em química experimental. Por exemplo, muitas vezes o químico é obrigado a trabalhar com misturas químicas, produtos naturais, sistemas bioquímicos, etc. Estes sistemas são bastante complexos para serem tratados utilmente com modelos que funcionam bem para sistemas mais simples. Tais sistemas dependem normalmente de um grande número de variáveis, necessitando de uma análise de dados multivariados.

Às vezes o problema pode ser solucionado utilizando-se métodos estatísticos razoavelmente simples, mas o processo de análise pode também necessitar de métodos de matemática aplicada que são raramente usados para resolver problemas químicos. Este tipo de análise é especialmente importante, por exemplo, em métodos espectroscópicos modernos, que geram uma enorme quantidade de dados multivariados dos quais somente uma pequena porção é interpretada.

Reconhecimento de Padrão (RP), que é um subcampo da inteligência artificial (1,2) é um tipo de tratamento que pode analisar dados multivariados. Somente nos últimos cinco anos a aplicação de RP em problemas químicos tem recebido bastante ênfase, apesar destes métodos serem comumente aplicados nas últimas três décadas em engenharia elétrica, economia, biologia, psicologia, etc.

Kowalski e Bender (3) descreveram o tipo de problema químico que o RP pode resolver. Dado um conjunto de objetos de diferentes categorias (objetos tendo uma mesma propriedade) e uma série de medidas feitas nestes objetos, é possível encontrar e/ou predizer propriedades dos objetos que não sejam diretamente medidas, mas que estão relacionadas com estas medidas via algumas relações não conhecidas. Os objetos com categorias conhecidas pertencem ao conjunto chamado treinamento ou referência, do

qual informações são extraídas para tornar possível a classificação de novos objetos (conjunto teste), com base nas mesmas medidas feitas para estes objetos (4). Em algumas aplicações, estas informações são usadas para classificar um conjunto de objetos adicionais, onde se conhecem as medidas, mas as categorias não são conhecidas, figura 1.

A representação gráfica mostra uma idéia central dos métodos de RP. Os objetos são considerados pontos no espaço M dimensional, onde M é o número de medidas feitas sobre estes objetos. As medidas são usadas para gerar o que chamamos de características, podendo ser medidas simples, algumas combinações lineares destas ou transformações matemáticas mais complexas. As medidas podem ser grandezas e unidades químicas ou físicas tais como: pH, concentração, densidade, medidas espectroscópicas, etc.

A distância euclidiana entre dois pontos i e j de coordenadas y_{ki} e y_{kj} é dado por:

$$d_{ij} = \left[\sum_{k=1}^{M} (y_{ki} - y_{kj})^2 \right]^{\frac{1}{2}} \qquad |1|$$

onde a soma é feita sobre as M medidas. Esta distância é muitas vezes uma excelente medida de semelhança; onde d_{ij} é uma medida de similaridade recíproca. Considera-se que quanto menor a distância entre estes, melhor a similaridade. Uma medida de

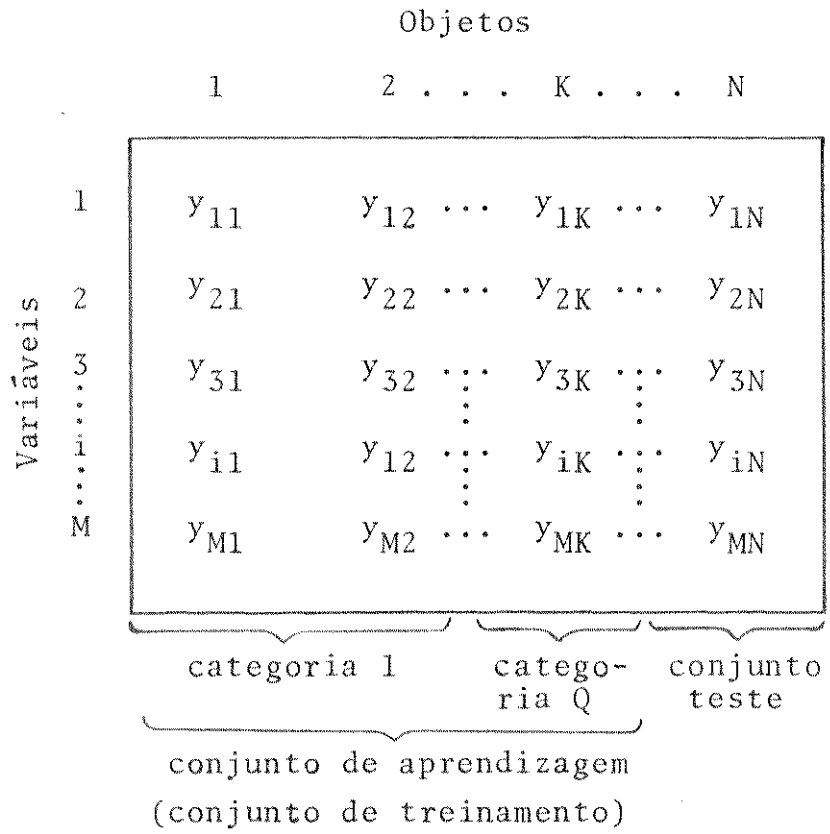


Figura 1. Dados (y_{ik} da i-ésima medida e do K-ésimo objetos) usadas em problemas de RP formando uma matriz M x N.

similaridade é convenientemente definida como:

$$S_{ij} = 1 - d_{ij}/(d_{ij})_{máx.} \qquad |2|$$

onde $(d_{ij})_{máx.}$ é a maior distância interponto no espaço M dimensional (3). Para objetos idênticos, $S_{ij} = 1$, enquanto que $S_{ij} = 0$ descreve dois pontos que são separados pela maior distância interponto para o conjunto sob estudo.

A medida de similaridade constitui uma das bases mais fundamentais para os métodos de RP e é usada de uma maneira in-

tuitiva pelos químicos. Por exemplo, normalmente substâncias similares são miscíveis em uma outra, ao passo que aquelas diferentes não o são. Espera-se que categorias de solventes miscíveis e não miscíveis poderão ser definidas com base nas similaridades (ou falta), determinadas à partir de suas outras propriedades (físicas ou químicas).

Os dados que são normalmente tratados por RP podem ser divididos em dois grupos: 1) paramétricos, onde se conhece a distribuição estatística dos dados ou esta pode ser estimada através de um algoritmo matemático e 2) não paramétricos, onde a distribuição estatística não é conhecida. Este tipo de problema é mais frequente em aplicações químicas e será discutido no restante deste trabalho.

Normalmente os métodos de RP são mais eficientes se os dados estão sujeitos a um tratamento chamado pré-processamento (5). Estes tratamentos podem modificar as variáveis originais em novas variáveis que terão propriedades matemáticas mais convenientes para análise, sem perder a informação inerente aos dados. Vários métodos de pré-processamento foram usados neste trabalho e serão discutidos mais tarde.

Os métodos de RP podem ser aplicados de dois modos:

1) Aprendizagem supervisionada: As amostras no espaço M dimensional são de classificação conhecida (categoria), cons-

tituindo o conjunto de treinamento. O objetivo é desenvolver um método de aprendizagem baseado nas informações contidas nos dados das amostras, para que estas sejam classificadas corretamente e então aplicar esta regra para amostras não conhecidas visando sua classificação. Este tipo de aprendizagem foi usado com mais frequência neste trabalho.

2) Análise não supervisionada: Na ausência do conjunto de treinamento, agrupamentos naturais poderão ser detectados e o objetivo é encontrar regiões no espaço com altas densidades de pontos ou agrupamentos de amostras no espaço M. Às vezes nenhuma informação sobre as categorias e objetos é possível de se obter, então estes agrupamentos substituirão esta informação.

Os problemas tratados por RP podem ser divididos em quatro níveis (6). O nível |1| supõe que todos objetos nos conjuntos de treinamento e teste pertencem a uma das categorias que já foram definidas inicialmente. O nível |2| além dos aspectos do nível |1| considera a possibilidade que alguns objetos no conjunto de referência ou treinamento pertençam a uma categoria que não foi definida pelo investigador. Neste nível, por exemplo, cada categoria é contida em um hipervolume no espaço M (4). Estas categorias são construidas de tal forma que um objeto dentro de um hipervolume pertence a uma determinada categoria, mas para objetos fora de qualquer um destes, admite-se a possibili-

dade que não pertençam a nenhuma categoria definida.

Nos níveis |3| e |4|, através das variáveis medidas, podemos prever quantitativamente uma (nível 3) ou mais (nível 4) propriedades externas, ou seja, outras variáveis que não as medidas para os objetos (6).

A maior parte das aplicações das técnicas de RP em química até agora está incluída nos níveis |1| e |2|. Neste trabalho usaremos exclusivamente estes níveis. Podemos citar como exemplo: a classificação de vinhos (7), uísques (8), artefatos arqueológicos (9), petróleo (10), estruturas moleculares (4) e doenças baseadas em dados químicos (5). Como as técnicas de RP são bastante eficientes para problemas de classificação, decidimos aplicá-las ao problema de identificação de fontes de águas minerais.

A água mineral difere das outras potáveis em sua composição química ou concentração. Os íons mais comumente encontrados naquelas são: sódio (Na), potássio (K), cálcio (Ca), magnésio (Mg), manganês (Mn), ferro (Fe), cloro (Cl), íons sulfatos (SO_4^{2-}) e carbonatos (CO_3^{2-}), geralmente provenientes de sais dissociados (11). No estado de São Paulo existem diversas fontes, mas por conveniência escolhemos as de Serra Negra, Lindóia e Valinhos.

Os traços dos elementos contidos nas amostras foram me-

didos por um espectrômetro de emissão atômica com uma fonte de plasma e no espectrofotômetro de absorção atômica no CENA (Centro de Energia Nuclear na Agricultura) em Piracicaba. Os resultados mostraram que a discriminação das diversas fontes de água mineral não poderia ser feita por observação direta dos dados e por isto a classificação destas fontes foi escolhida como a primeira aplicação das técnicas de Reconhecimento de Padrão em problemas químicos a ser desenvolvido pelo grupo de Quimiometria do Instituto de Química da UNICAMP.

OBJETIVOS

Os objetivos deste trabalho são:

1) Estabelecer se a aplicação dos métodos de RP podem ser bem sucedidos na classificação das três fontes de água mineral baseada na análise química de seus elementos traços, porque regras de classificação simples não são suficientes para fazer isso.

2) Estabelecer um conjunto de dados conhecidos e precisos que poderão ser usados para desenvolver métodos mais poderosos de RP.

3) Testar a autenticidade das amostras de água mineral comercializadas nos supermercados e distribuidores.

No capítulo II descreveremos os equipamentos e resulta-

dos das análises experimentais. No capítulo III serão descritos os métodos estatísticos e de RP contidos no programa ARTHUR (12). No capítulo IV serão apresentados os resultados da aplicação do programa ARTHUR aos dados experimentais e discussões. Finalmente no capítulo V serão feitas as conclusões dos resultados acima.

CAPÍTULO II

DESCRIÇÃO EXPERIMENTAL

Equipamentos

A determinação dos traços dos elementos das amostras de água mineral foi feita através do "Inductively Coupled Argon Plasma Emission Spectrometer", Jarrel-ash, modelo 96-750 (espectrômetro de leitura direta), locado no CENA. Este equipamento permite a análise quantitativa de até 48 elementos, embora atualmente o mesmo é feito para 18 elementos. A análise é rápida e os resultados são impressos por uma impressora acoplada a um minicomputador.

A figura 2 mostra o esquema do espectrômetro (13). Um gerador de rádio frequência (RF) fornece energia para a formação do plasma. A amostra é atomizada através de um nebulizador e transportada até o plasma, por meio de um gás de arraste (no caso argônio), onde os elementos da amostra são excitados, emitindo fótons de energia radiante que serão transformados em sinais elétricos por meio de uma fotomultiplicadora. Um computador acoplado ao espectrômetro converte os sinais em unidades de concentração (%, ppm e ppb) que podem ser lidas diretamente na impressora.

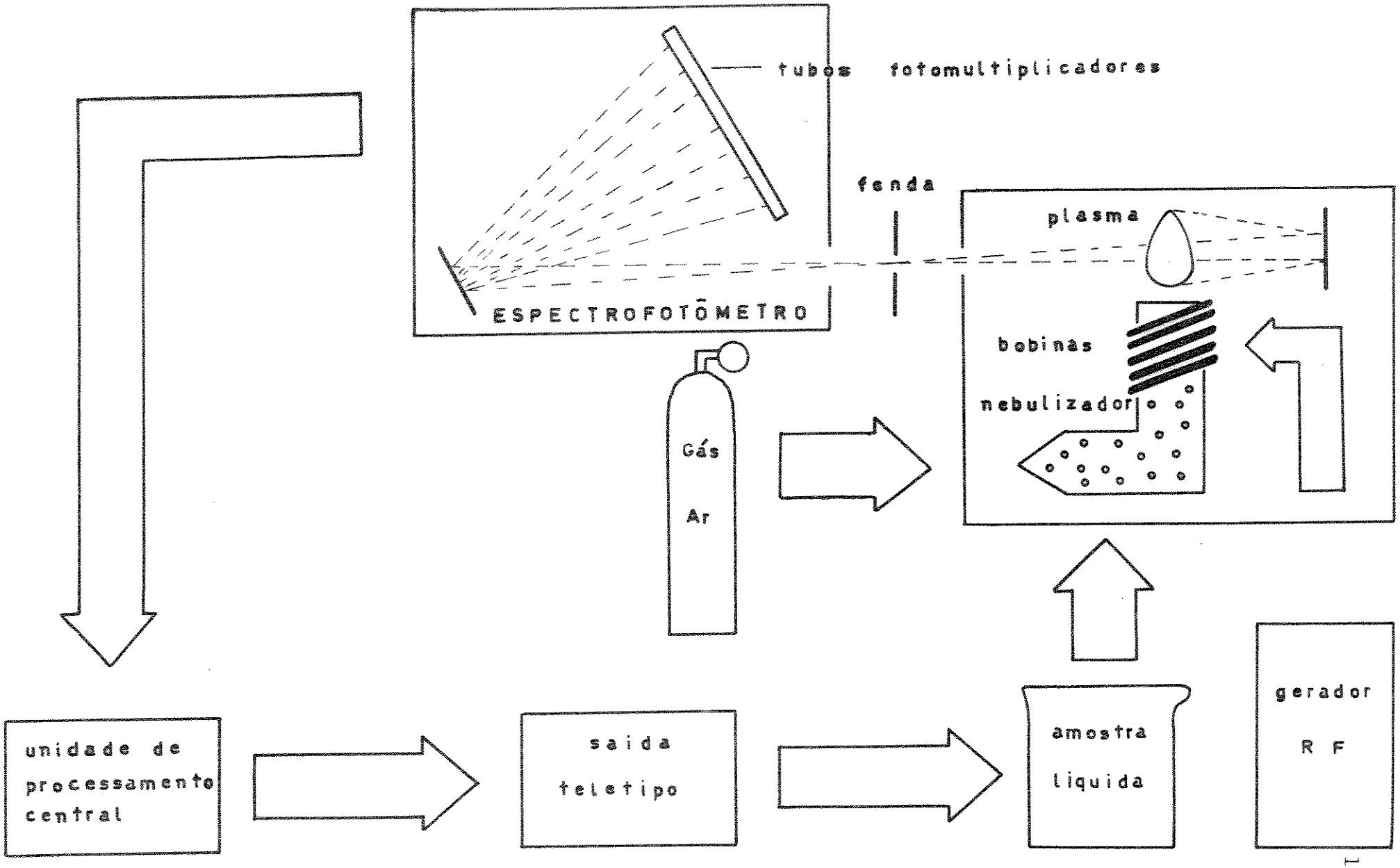


Figura 2. ESQUEMA DO ICAP

Daremos a seguir alguns dados técnicos mais relevantes do espectrômetro de emissão atômica utilizado. Os 18 elementos analisados são: prata (Ag), alumínio (Al), arsênio (As), boro (B), bário (Ba), cálcio (Ca), cádmio (Cd), cromo (Cr), cobre (Cu), ferro (Fe), magnésio (Mg), manganês (Mn), sódio (Na), níquel (Ni), chumbo (Pb), fósforo (P), silício (Si) e zinco (Zn). O limite teórico de detecção dos elementos em unidades ppm é dado na tabela 1. A tabela 2 especifica o intervalo de concentração de confiabilidade em ppm.

As especificações mais importantes são: temperatura do plasma - 8.000 a 9.000 K, frequência da bobina de RF - 27 MHz, e gás de arraste de alta pureza, consumo 5 ℓ/min.

Dos resultados obtidos neste equipamento escolhemos os que apresentaram maiores concentrações nas amostras, são eles: cálcio, sódio, silício, fósforo e magnésio, sendo que os outros estão, em geral, abaixo do limite de determinação.

Como a linha de emissão máxima para o potássio não dá resultados de muita confiança para baixas concentrações, optou-se por outro método para sua determinação.

Determinação do Potássio

O método escolhido foi espectrofotometria de absorção atômica, onde o potássio pode ser facilmente determinado. Foi

Tabela 1. Limite teórico de detecção em ppm.

P	Ag	Al	As	B	Ba
0.03	0.001	0.01	0.015	0.002	0.005
Ca	Cd	Cr	Cu	Fe	Mn
0.004	0.001	0.002	0.002	0.0001	0.0005
Mg	Na	Pb	Si	Zn	Ni
0.02	0.2	0.015	1.0	0.001	0.004

Tabela 2. Intervalo de concentração em ppm.

P	Ag	Al	As	B	Ba
<0.1- >1000	<0.001- >10	<0.01- >100	<0.1- >1000	<0.01- >100	<0.001- >10
Ca	Cd	Cr	Cu	Fe	Mn
<0.01- >100	<0.001- >10	<0.01- >100	<0.01- >100	<0.001- >10	<0.001- >10
Mg	Na	Pb	Si	Zn	Ni
<0.1- >1000	<0.1- >1000	<0.1- >1000	<100- >5000	<0.001 >10	<0.01- >100

construída uma curva de calibração a partir de uma série de soluções de KCl em 10^{-3} M de HCl, respectivamente apresentando teores de 0,5, 1, 2, 3, 4, 5, 6 e 7 ppm K.

A partir da curva foi possível determiná-lo numa faixa de concentração inferior a 5 ppm, onde a curva é linear.

O equipamento usado foi o espectrofotômetro de absorção atômica, modelo Perkin Elmer-306.

As especificações mais importantes foram:

comprimento de onda - 765,5 nm

fenda - 2,0 nm

corrente na lâmpada - 15 mA

gás suporte - Ar, 30 psi

combustível - acetileno, 8 psi.

RESULTADO DAS ANÁLISES

As amostras foram coletadas em diferentes fontes públicas das cidades de Serra Negra, Valinhos e Lindóia mostradas na tabela 3. Usamos frascos de polietileno com capacidade de 500 ml para obtenção das amostras. Estes foram lavados primeiro com ácido clorídrico 0,01 N e depois água destilada de acordo com o padrão do CENA. Antes da coleta os frascos foram novamente lavados com a água mineral a ser coletada. Uma parte das amostras foi obtida nos supermercados e distribuidores. Estas fo-

Tabela 3. Fontes de água mineral usadas para obter o conjunto de dados.

Cidade	Fonte
Serra Negra	São Carlos
" "	Sto. Agostinho F-1
" "	Sto. Agostinho F-2
" "	Sto. Agostinho F-3
" "	Sto. Agostinho F-4
" "	Turismo Macaquinhos F-1
" "	Turismo Macaquinhos F-2
" "	Turismo Macaquinhos F-3
" "	Turismo Macaquinhos F-4
" "	Turismo Macaquinhos F-5
" "	Turismo Macaquinhos F-6
" "	N. Sra. Aparecida F-1
" "	N. Sra. Aparecida F-2
" "	Camping Serra Negra
" "	Dr. Jovino Silveira
Lindóia	São Benedito, São José
"	São Jorge
"	São Bernardo, São Francisco
"	Água Azul
"	Engarrafadora Mantovani
"	BC16A
Valinhos	Santa Tereza F-1
"	Santa Tereza F-2
"	Sônia

ram analisadas diretamente sem tratamento prévio.

Como durante o tempo de armazenamento da água pode ocorrer significantes perdas nos teores dos elementos constituintes (14,15), é aconselhável que as amostras sejam acidificadas para que as referidas perdas sejam minimizadas. Normalmente os ácidos usados são: ácido sulfúrico concentrado, ácido acético, ácido nítrico concentrado ou ácido clorídrico, mantendo o pH numa faixa aproximada de 1 a 3,5 (14). Geralmente esta acidificação reduz consideravelmente a razão de adsorção dos metais (15,16), mas tem havido pouco ou nenhum estudo sistemático da validade desta hipótese.

Nossas amostras foram acidificadas com ácido sulfúrico concentrado (PA), industrializado pela Merck S.A. Usamos o volume de 0,5 ml de ácido em 500 ml de água de acordo com (16). Como teste, usamos também ácido nítrico concentrado (PA) industrializado pela Carlo Erba, com o volume de 0,25 ml de ácido em 500 ml de água (15). Através da tabela 4, pode-se observar que os resultados usando os dois ácidos são equivalentes dentro do erro experimental com exceção do fósforo que mostra uma grande diferença percentual. Isto pode ser devido ao fato de sua concentração estar muito próxima ao limite de detecção.

A reprodutibilidade dos dados foi considerada adequada tendo em vista os resultados indicados na tabela 5, que se refe

re a cinco análises consecutivas de uma mesma amostra. O sódio foi menos reprodutivo com média de 6,17 ± 0,57, mostrando um êrro de 9%. Apesar desta variação não ser satisfatória para muitos estudos, podemos esperar que este êrro não prejudique seriamente as classificações usando os métodos de RP.

Além desta analisamos duas amostras da mesma fonte coletadas simultaneamente, usando o mesmo ácido, tabela 6. Aqui podemos observar que com exceção do fósforo analisado por ICAP e do potássio analisado por espectrofotometria de absorção atômica, os desvios dos traços dos elementos estão dentro da reprodutibilidade dos dados.

Outros testes foram feitos para definir o melhor tratamento das amostras. Primeiro, coletamos várias amostras da mesma fonte simultaneamente, para serem analisadas em diferentes épocas para observar a variação das concentrações dos traços dos elementos entre a coleta e a análise. Usamos como agentes preservantes os ácidos sulfúrico e nítrico concentrados, tabelas (7a e 7b). O comportamento dos elementos usando os diferentes ácidos parecem ser semelhantes, mostrando uma tendência no aumento das concentrações à medida que aumenta o tempo de armazenamento.

Como prova em branco, testamos água deionizada com os ácidos sulfúrico e nítrico, com o mesmo volume usado nas amos-

Tabela 4. Resultado das análises de uma mesma amostra, acidificadas com diferentes ácidos.
B - ácido nítrico
C - ácido sulfúrico

Amostra	P	Ca	K*	Mg	Na	Si
87B	0.08	6.33	1.95	1.19	6.36	19.22
87C	0.14	6.12	2.05	1.16	6.12	18.96

Tabela 5. Reprodutibilidade: Análises consecutivas de uma mesma amostra.

Amostra	P	Ca	K*	Mg	Na	Si
A	0.15	7.81	3.36	3.54	5.60	16.83
B	0.12	7.82	3.40	3.52	5.75	16.79
C	0.16	7.92	3.36	3.48	6.27	16.21
D	0.13	8.08	3.38	3.52	6.46	16.73
E	0.15	8.04	3.40	3.47	6.78	16.06

Tabela 6. Resultado das análises de duas amostras coletadas simultaneamente, (mesma acidificação).

Amostra	P	Ca	K*	Mg	Na	Si
68A	0.14	5.29	3.20	0.94	7.41	18.65
68B	0.05	5.50	2.80	0.99	7.98	18.48

* Analisado por espectrofotometria de absorção atômica.

Tabela 7.a. Variação da concentração com o tempo de armazenamento, acidificação com ácido sulfúrico (data da coleta: 17/06).

Amostra	Data	P	Ca	K*	Mg	Na	Si
59A	24/06	0.05	4.46	1.40	2.06	5.07	15.53
59B	17/07	0.07	4.90	1.50	2.33	5.35	14.94
59C	29/07	0.11	5.14	1.50	2.36	5.64	15.23

Tabela 7.b. Variação da concentração com o tempo de armazenamento, acidificação com ácido nítrico (data da coleta: 09/09).

Amostra	Data	P	Ca	K*	Mg	Na	Si
87B	25/09	0.08	6.33	1.95	1.19	6.36	19.22
87A	19/10	0.14	6.80	3.35	1.28	6.49	21.79

* Analisado por espectrofotometria de absorção atômica.

tras de água mineral, para verificar o teor de pureza do material empregado. Com este teste, foi possível observar que não houve detecção de traços, isto é, a concentração dos elementos estava abaixo do limite de determinação.

Uma outra análise foi feita usando duas amostras coletadas simultaneamente, acidificando apenas uma para observar se haveria variação nos traços dos elementos, tabela 8. Verificou-se que, quando o tempo de armazenamento é pequeno a acidificação parece reduzir as concentrações dos elementos. Nas tabelas 8a e 8b esta variação é pequena, mostrando estar contida na reprodutibilidade dos dados, enquanto que na tabela 8c esta variação é um pouco maior.

Para verificar se os métodos de RP seriam eficientes na classificação, mesmo havendo variação nas concentrações das amostras devido causas naturais, determinamos um intervalo de 10 meses para coleta das amostras. As tabelas 9, 10, 11, 12 e 13 mostram esta variação.

Tabela 8. Resultado das análises de três pares de amostra 8a), 8b) e 8c) de uma mesma fonte quando acidificadas (A) e não acidificadas (B).

Amostra	P	Ca	K*	Na	Mg	Si
96A	0.05	7.51	3.36	3.58	4.89	17.11
96B	0.08	7.54	3.35	3.65	5.21	17.61

Tabela 8a

Amostra	P	Ca	K*	Na	Mg	Si
97A	0.15	7.81	3.35	3.54	5.60	16.83
97B	0.05	7.55	3.36	3.64	5.06	17.78

Tabela 8b

Amostra	P	Ca	K*	Na	Mg	Si
101A	0.10	11.68	2.25	3.14	8.86	19.01
101B	0.05	13.10	2.40	3.47	9.91	19.76

Tabela 8c

* Analisado por espectrofotometria de absorção atômica.

Tabela 9. Resultado da variação dos traços de elementos da mesma fonte durante oito meses (Valinhos), Fonte Santa Tereza F-1.

Amostra	Mês da coleta	Ca	K*	Mg	Na	Si	P
15	05-03-80	15.33	4.62	5.55	14.13	16.39	0.13
48	12-04-80	7.40	3.39	3.35	6.55	15.49	0.27
54	24-05-80	7.74	3.42	3.47	6.22	15.99	0.21
66	21-06-80	13.16	3.70	4.77	13.53	14.16	0.10
72	23-07-80	10.26	4.42	3.87	5.76	15.56	0.09
83	06-08-80	15.30	4.40	5.26	10.66	14.12	0.04
96	14-10-80	7.51	3.36	3.58	4.89	17.11	0.05

Tabela 10. Resultado da variação dos traços de elementos da mesma fonte durante oito meses (Valinhos), Fonte Santa Tereza F-2.

Amostra	Mês da coleta	Ca	K*	Mg	Na	Si	P
16	05-03-80	6.37	3.50	3.17	7.53	14.75	0.05
49	12-04-80	6.84	3.65	3.22	7.50	14.74	0.01
53	24-05-80	7.69	3.37	3.36	6.40	16.04	0.01
67	21-06-80	6.82	3.80	3.04	6.90	15.70	0.10
73	23-07-80	6.11	3.30	2.75	5.85	15.90	0.15
82	06-08-80	6.69	3.30	3.17	4.78	16.08	0.07
97	14-10-80	7.81	3.35	3.54	5.60	16.83	0.15

* Analisado por espectrofotometria de absorção atômica.

Tabela 11. Resultado da variação dos traços de elementos da mesma fonte durante nove meses (Serra Negra), Fonte Santo Agostinho F-1.

Amostra	Mês da coleta	Ca	K*	Mg	Na	Si	P
2	23-02-80	5.11	1.48	2.49	5.33	16.61	0.11
28	23-03-80	4.75	1.50	2.35	4.52	15.85	0.15
59	17-06-80	4.47	1.40	2.06	5.07	14.53	0.05
79	02-08-80	3.66	1.30	1.78	3.62	15.91	0.08
106A	17-10-80	6.00	1.60	2.49	4.20	16.12	0.12

Tabela 12. Resultado da variação dos traços de elementos da mesma fonte durante nove meses (Serra Negra), Fonte Santo Agostinho F-2.

Amostra	Mês da coleta	Ca	K*	Mg	Na	Si	P
5	23-02-80	3.47	1.00	1.18	4.28	13.89	0.05
42	23-03-80	3.20	1.00	1.14	3.33	13.62	0.12
56	17-06-80	3.22	1.00	1.06	4.48	12.68	0.01
80A	02-08-80	2.44	2.27	0.80	3.08	13.19	0.05
107A	17-10-80	5.32	1.25	1.38	3.84	13.88	0.02

Tabela 13. Resultado da variação dos traços de elementos da mesma fonte durante oito meses (Lindóia), Fonte São Jorge.

Amostra	Mês da coleta	Ca	K*	Mg	Na	Si	P
23	23-03-80	2.01	2.21	0.89	3.73	13.90	0.05
64	17-06-80	2.07	2.20	0.82	5.38	13.30	0.01
77	02-08-80	1.82	2.11	0.81	3.64	13.81	0.01
103A	17-10-80	2.37	2.00	1.07	5.03	13.07	0.09

* Analisado por espectrofotometria de absorção atômica.

CAPÍTULO III

MÉTODOS ESTATÍSTICOS E DE RECONHECIMENTO DE PADRÃO

Para a análise de nossos dados usamos o programa de computação ARTHUR (12), que contém métodos de RP e estatísticos (escrito em FORTRAM IV) os quais discutiremos a seguir.

MÉTODOS DE PRÉ-PROCESSAMENTO

Se as medidas dos objetos são representadas em forma de uma matriz $\underset{\sim}{Y}$ com elementos y_{ik} (i-ésima variável e k-ésimo objeto), o processo de operação em $\underset{\sim}{Y}$ para mudar a estrutura dos dados no espaço M é conhecido como pré-processamento. Muitas vezes estes métodos são aplicados para reduzir o número de variáveis (5). Discutiremos a seguir alguns métodos de pré-processamento usados neste trabalho.

ESCALAMENTO - É chamado de "auto-escalamento" quando as características são transformadas tal que a média é zero e a variança igual a um (3,17).

$$\bar{y}_i = \sum_{k=1}^{N} y_{ik}/N \qquad |3|$$

e

$$\sigma_i = \left\{ \sum_{k=1}^{N} (y_{ik} - \bar{y}_i)^2 \right\}^{\frac{1}{2}} / (N-1)^{\frac{1}{2}} \qquad |4|$$

onde: $\bar{y}_i$ é a média para a variável i, N é o número de amostras no conjunto de treinamento e σ_i é a variança.

As equações transformadas podem ser escritas como:

$$y'_{ik} = (y_{ik} - \bar{y}_i)/(N)^{\frac{1}{2}} \cdot \left\{ \sum_{k=1}^{N} (y_{ik} - y_i)^2/N \right\}^{\frac{1}{2}}$$

$$\sigma'_i = \left\{ \sum_{k=1}^{N} (y'_{ik} - \bar{y}'_i)^2 \right\}^{\frac{1}{2}} / (N-1)^{\frac{1}{2}} \qquad |5|$$

$$\bar{y}'_i = \sum_{k=1}^{N} y'_{ik}/N$$

onde σ'_i é a variança dos dados transformados ($\sigma'_i = 1$) e $\bar{y}'_i$ = é a média dos dados transformados ($\bar{y}_i = 0$).

O outro método de escalamento é chamado de "intervalo de escalamento", onde os dados são transformados tal que o valor mínimo da característica é zero e o máximo é um.

$$y'_{ik} = \frac{y_{ik} - y_{ik}\,(\text{min})}{R_i} \qquad |6|$$

Onde $R_i = y_{ik}\,(\text{máx}) - y_{ik}\,(\text{min})$ e representa o intervalo varrido pelos valores da i-ésima variável.

O segundo método geral de pré-processamento comumente usado é o peso das variáveis, onde é avaliado a importância individual de cada característica ou variável para separar categorias. Existem vários métodos de avaliação, mas, os mais importantes são o peso de Fisher e o peso de variança.

Normalmente as várias medidas são autoescaladas para permitirem comparações na mesma escala; por outro lado os métodos de peso determinam quais destas características são úteis na separação das categorias, devendo ser enfatizadas e quais poderão ser eliminadas.

O PESO DE FISHER: estima quantitativamente a utilidade de uma característica para separar duas categorias

$$W(F)_{j,m,n} = \frac{\left(\sum_{m=1}^{N_m} y_{jm}/N_m - \sum_{n=1}^{N_n} y_{jn}/N_n \right)^2}{(S_{jm}^2 + S_{jn}^2)} \qquad |7|$$

onde $W(F)_{j,m,n}$ é a medida da utilidade da característica, j, para separar a categoria m da categoria n e S_{jm}^2 é o quadrado dos desvios padrões da categoria m para característica j:

$$S_{jm}^2 = \frac{\left(\sum_{k=1}^{N_m} y_{jk}^2 \right) N_m - \left(\sum_{k=1}^{N_m} y_{jk} \right)^2}{(N_m - 1)}$$

onde N_m é o número de amostra na categoria m e y_{jk} é a j-ésima característica do k-ésimo padrão na categoria m.

A equação |7| dá uma idéia qualitativa deste peso, ou seja, se a diferença das médias para duas categorias for grande e a soma de suas variânças for pequena o peso será grande. Neste caso a variável j fornece a base para boa separação das categorias m e n. Então o valor da característica j é pesado pelo fator multiplicativo $W(F)_{j,m,n}$ para enfatizar esta separação. O

peso de Fisher global é a média aritmética de $W(F)_{j,m,n}$ para todos os pares de categorias:

$$W(F)_j = \{ \sum_{m=1}^{NCAT-1} \sum_{n=m+1}^{NCAT} W(F)_{j,m,n} \} / \{ (NCAT)(NCAT-1)/2 \} \quad |8|$$

onde NCAT é o número de categorias.

O PESO DE VARIANÇA é a razão da variança interclasse à variança intraclasse (3). Para as categorias m e n e a variável j, o peso de variança é dado por:

$$W(V)_{j,m,n} = \frac{(\sum_{m=1}^{N_m} y_{jm}^2)/N_m + (\sum_{n=1}^{N_n} y_{jn}^2)/N_n - 2\,(\sum_{m=1}^{N_m} y_{jm})/N_m\,(\sum_{n=1}^{N_n} y_{jn})/N}{S_{jm}^2 + S_{jn}^2} \quad |9|$$

$$S_{jm}^2 = \frac{(\sum_{k=1}^{N_m} y_{jk}^2)N_m - (\sum_{k=1}^{N_m} y_{ik})^2}{(N_m)^2} .$$

O peso de variança global para todos os pares de categorias é a média geométrica dos pesos de variança para a separação de cada par de categoria.

$$W(V)_j = \sum_{m=1}^{NCAT} \sum_{n=m+1}^{NCAT} W(V)_{j,m,n} / (NCAT)(NCAT-1) \quad |10|$$

O terceiro método de pré-processamento usado neste trabalho é a TRANSFORMAÇÃO DE KARHUNEN-LOEVE (18). Este gera novas variáveis as quais são combinações lineares das variáveis originais. A primeira nova variável contém a maior variança e

assim sucessivamente, até um ponto onde as últimas variáveis tem variâncas que são zero ou perto de zero (19). Esta transformação tem a propriedade de conservar a variança total dos dados. Muitas vezes é também usado como um método de redução de dimensionalidade de um conjunto de dados, isto é, projeta M dimensões num espaço de menor dimensão (normalmente duas). Aquelas variáveis com variança perto de zero são truncadas. Esta transformação pode ser descrita como segue; uma matriz de covariança $\underline{C}$ é gerada:

$$(\underline{C})_{ij} = \sum_{k=1}^{N} (y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j)/(N-1)$$

onde $\bar{y}_i = \sum_{k=1}^{N} y_{ik}/N$.

Os autovalores e autovetores λ_j e μ_j da matriz $\underline{C}$ são calculados por diagonalização e a soma de todos os λ_j deve corresponder a 100% de variança total. Como o autovalor é a variança ao longo de seu autovetor correspondente, a percentagem da variança contida nos primeiros ℓ autovetores é dada por:

$$\% V = (\sum_{i}^{\ell} \lambda_i).100/ \sum_{i=1}^{M} \lambda_i \ . \qquad |11|$$

O espaço M é reduzido usando em lugar das variáveis originais os autovetores correspondentes aos autovalores com valo-

res de variança significativa. Esta é especialmente útil quando os dois ou três primeiros autovetores contêm uma fração de variança significativa. Neste caso o espaço M pode ser projetado usando esta transformação para um espaço de dimensão mais baixa de fácil visualização para o pesquisador.

Um outro método de pré-processamento que diminui o número de variáveis é o de SELEÇÃO. Este gera características ortogonais (ou linearmente independentes) que são importantes para classificação e retém a identidade das medidas originais (20). Basicamente a seleção é uma ortogonalização de Schmidt comumente usada na ortogonalização dos orbitais moleculares em química quântica.

As M medidas são escaladas e normalizadas tal que para as i-ésimas características:

$$< y_i / y_i > \equiv \sum_{k=1}^{M} y_{ik}^2 = 1 \qquad |12|$$

onde a soma é sobre todas as amostras. A seguir é selecionada a medida mais importante usando por exemplo, o peso de Fisher ou de variança. A característica com mais alto peso é selecionada como primeira. O próximo passo é remover a correlação da primeira característica das outras N-1 variáveis. Estas são decorrelacionadas da primeira (y_i) por combinações lineares:

$$x_j = a y_i + b_{yj} \qquad |13|$$

com a condição que:

$$< x_j / x_j > = 1 \qquad |14|$$

e que:

$$< y_i / x_j > \equiv \sum_{k=1}^{M} y_{ik} \, x_{jk} = 0$$

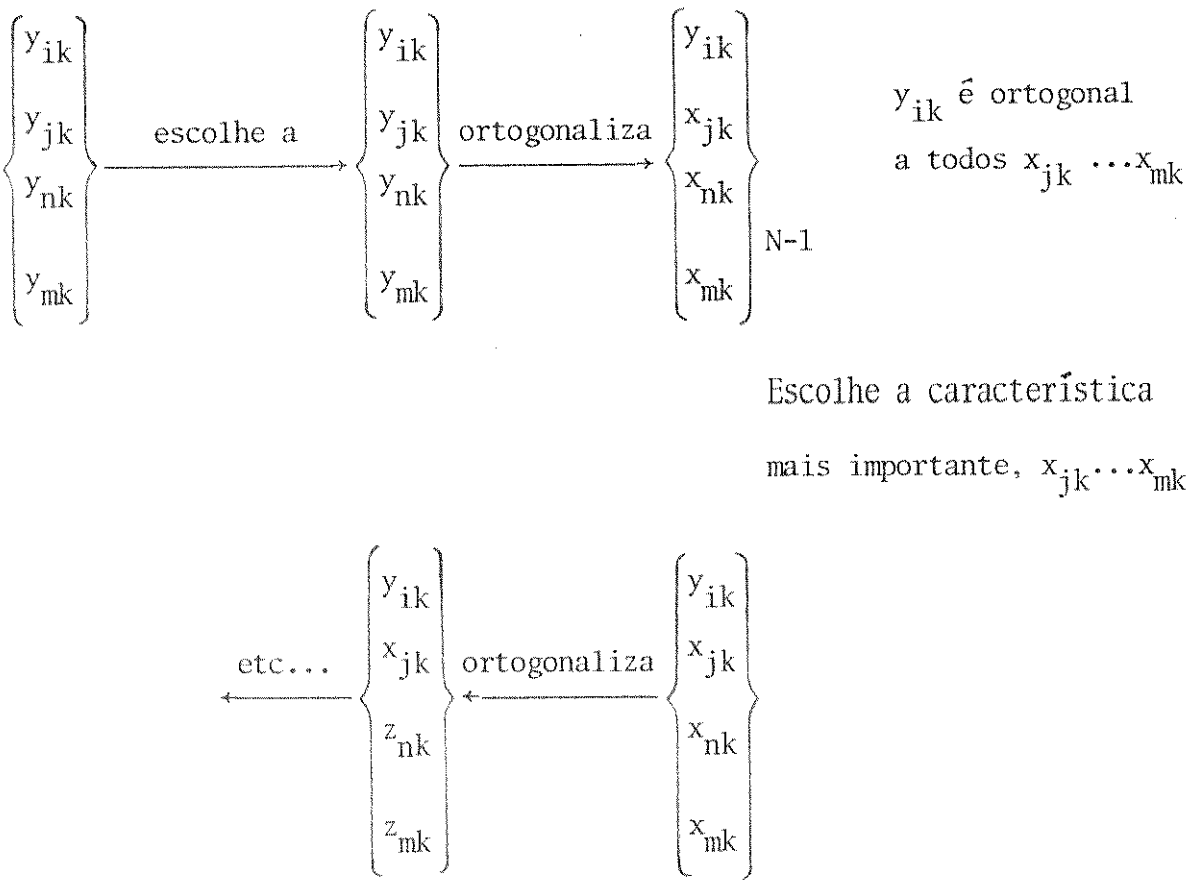
onde as características y_i e x_j são ortogonais.

O cálculo de a e b, equação |13|, dá uma nova característica x_j a qual não será correlacionada com a característica escolhida y_i, mas vai conter pelo menos parte da sua identidade original.

Depois de removida a correlação da primeira característica com as N-1 variáveis restantes, será selecionada a mais importante desta N-1 variáveis e sua correlação com as N-2 características restantes será removida de uma maneira análoga a primeira. Assim, cada característica selecionada retém sua identidade menos as correlações das características previamente selecionadas, como mostra o esquema da página seguinte.

Existem outros métodos de pré-processamento que não serão discutidos neste trabalho, tais como GRAB (17) e mapeamento não linear (NLM) (19,21). A investigação de novos métodos de

pré-processamento é também uma das ênfases mais importantes no campo de análise de dados multivariados.



Além dos métodos de pré-processamento usamos também alguns métodos de reconhecimento e conhecimento de padrão, os quais discutiremos a seguir.

MÉTODOS DE RECONHECIMENTO DE PADRÃO

REGRA DO VIZINHO MAIS PRÓXIMO (KNN) - Os métodos de classificação que são estritamente usados em aprendizagem supervisionada são chamados de métodos de reconhecimento de padrão. Um método conceitualmente simples e bastante usado é a regra do vi

zinho mais próximo, o qual necessita de um conjunto de treina - mento (22).

Neste método uma amostra com categoria não conhecida é classificada de acordo com as categorias de seus K vizinhos mais próximos, podendo ser K = 1, 2, 3, ... 10 (4). Normalmente estes vizinhos são definidos pela distância euclidiana. Em geral uma amostra é classificada na classe a qual a maioria dos seus vizinhos mais próximos pertencem, figura 3.

O conjunto de treinamento é usado para determinar o valor de K, que dá uma melhor predição dos vizinhos mais próximos no conjunto de treinamento. Assim uma amostra não conhecida que pertence ao conjunto teste é classificada de acordo com o número de vizinhos mais próximos, K, determinado pelo conjunto de treinamento.

MÁQUINA DE APRENDIZAGEM LINEAR (PLANE) - Considere um conjunto de treinamento finito Y de N padrões, onde Y é dividido em dois subconjuntos de treinamento Y_1 e Y_2. O subconjunto Y_1 contém os padrões os quais pertencem à categoria 1 e o Y_2 contém os padrões pertencentes à categoria 2 (23,24). Vamos supor que estes subconjuntos são separáveis linearmente, como mostra a figura 4. Um hiperplano pode então dividir o espaço padrão em dois subespaços onde um destes é chamado R_1 e o outro R_2, figura 4.

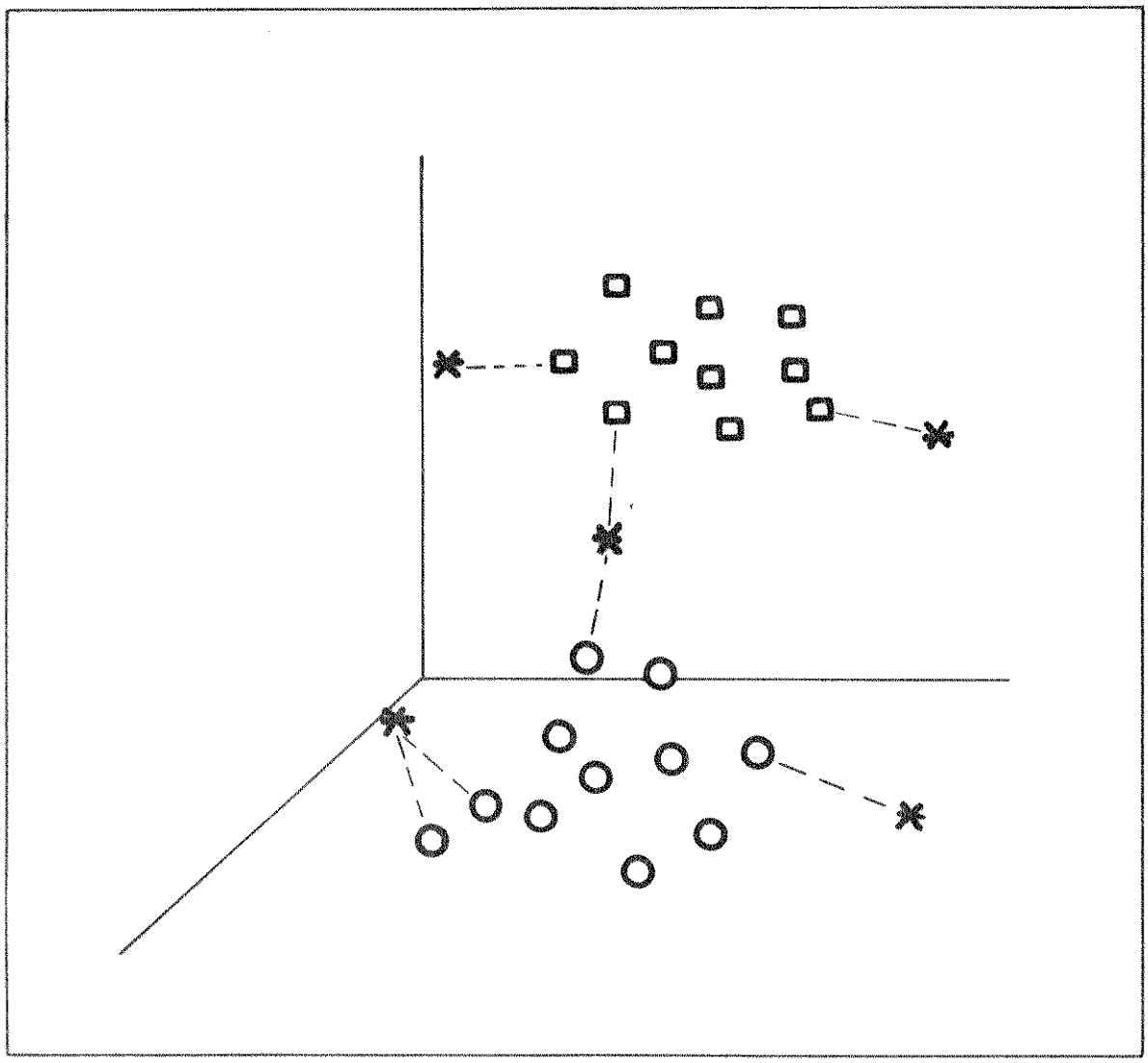


Figura 3. O método KNN classifica um objeto de acordo com a maioria dos seus vizinhos mais próximos (* objetos não classificados).

Os hiperplanos são descritos matemáticamente usando a equação generalizada de um plano:

$$\sum_{i=1}^{n} W_i y_i + W_n = 0 \qquad |15|$$

onde os valores W_i são chamados pesos. Como os y_i para as amostras são conhecidos, a máquina de aprendizagem linear necessita de uma solução para os pesos W_i, para fazer a classificação. Esta poderá ser feita mais convenientemente no espaço de peso (onde os W_i definem os eixos) do que no espaço de configuração, figura 5.

Para fazer isto o vetor padrão original Y será aumentado por um componente (d + 1) i-ésimo cujo valor é sempre igual a um. Este valor aumentado será simbolizado por X, onde seus componentes aumentados serão dados por x_1, x_2, ... x_D, onde D = d + 1 e $y_i = x_i$ para i = 1, ..., D-1, e $y_{d+1} = x_D = +1$. A função discriminante linear de Y pode então ser escrita em termos de X.

$$g(Y) = \vec{X}.\vec{W} \qquad |16|$$

Qualquer padrão aumentado X, é representado por um hiperplano no espaço peso, como mostra a figura 4, que é o locus de todos os pontos peso para qual:

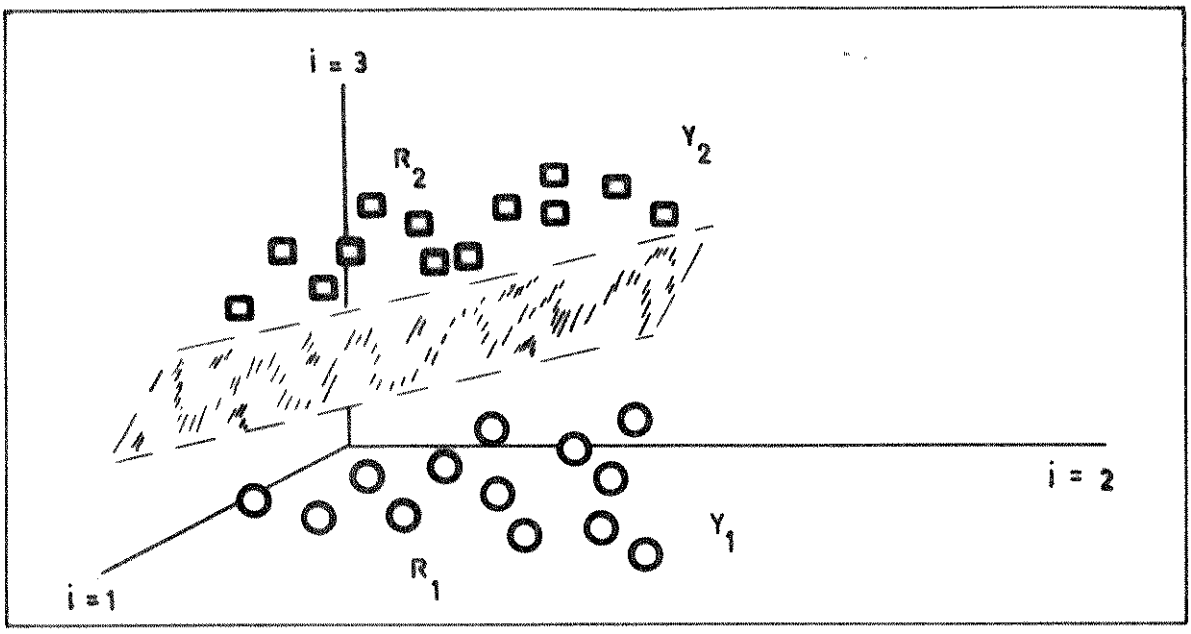


Figura 4. A máquina de aprendizagem linear separa as classes por meio de um hiperplano M-1 dimensional.

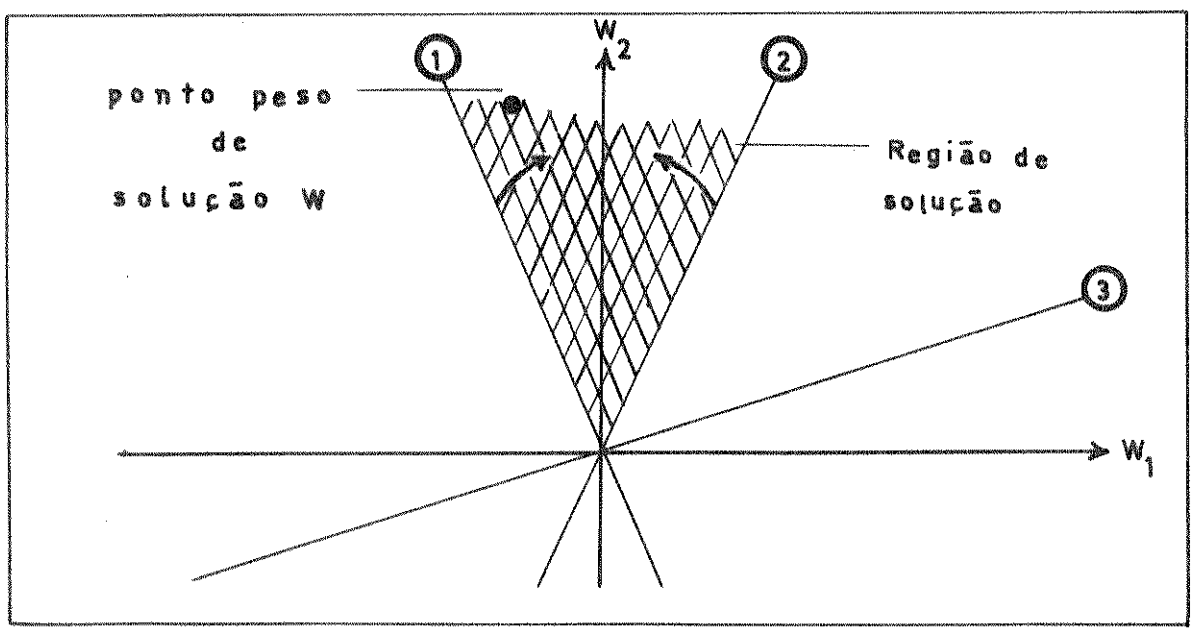


Figura 5. Espaço de peso em duas dimensões com três hiperplanos padrão. As flechas indicam o lado positivo (definido arbitrariamente) de cada hiperplano e os números no círculo indicam os vários padrões ou amostras.

$$\vec{W}.\vec{X} = 0 \ . \qquad |17|$$

Aqueles pontos pesos, $\vec{W}$, com $\vec{W}.\vec{X} > 0$ estão em um lado do hipoplano padrão. Aqueles com o valor de $\vec{W}$ tal que $\vec{W}.\vec{X} < 0$ estão do outro lado deste hiperplano ($\vec{W}.\vec{X} = 0$ onde X está no hiperplano). Cada elemento Y_1 e Y_2 será obtido por padrões aumentados X_1 e X_2 respectivamente, onde:

$$\left.\begin{array}{l} Y_1 \rightarrow X_1 \\ Y_2 \rightarrow X_2 \end{array}\right\} \text{ subconjuntos de treinamento aumentados}$$

$$\left.\begin{array}{l} Y \rightarrow X \end{array}\right\} \quad X = X_1 U X_2$$

Se X_1 e X_2 são linearmente separáveis existe pelo menos um vetor peso $\vec{W}$, chamado vetor peso de solução, tal que:

$$\vec{X}.\vec{W} > 0 \quad \text{para cada X em } X_1$$

e $$\vec{X}.\vec{W} < 0 \quad \text{para cada X em } X_2 \ .$$

A figura 5 ilustra esta situação para três hiperplanos padrão correspondendo às amostras em X_1. A região achuriada representa todos os vetores peso $\vec{W}$, satisfazendo $\vec{X}.\vec{W} > 0$. Esta região de solução varre todo o espaço peso situado no lado positivo (como indica a flecha nos três hiperplanos).

Em geral, existem N hiperplanos para N objetos, a região de solução situada no lado positivo destes hiperplanos padrão pertencem a X_1 e no lado negativo aqueles que pertencem a X_2.

Quando N aumenta a região de solução torna-se melhor definida. A máquina de aprendizagem linear, por um processo interativo usa todos os hiperplanos padrão, para calcular o valor do vetor peso de solução contido na região de solução (figura 5). Este vetor peso quando projetado no espaço de configuração divide este em duas regiões (como na figura 4). Um contendo os pontos em Y_1 e o outro contendo os pontos em Y_2.

A máquina de aprendizagem linear pode funcionar com mais de duas categorias. Neste caso a função discriminante pode ser representada como o produto do vetor peso com um vetor padrão aumentado:

$$g_i(Y) = \vec{W}(i) . \vec{X} \quad \text{para } i = 1, \ldots, R$$

onde R categorias estão sendo estudadas. Análogo a máquina de aprendizagem linear com duas categorias, este divide o espaço padrão em um número de regiões igual ao número de categorias.

STATISTICAL ISOLINEAR MULTIPLE COMPONENT ANALYSIS (SIMCA)

O método SIMCA é baseado no fato em que os dados y^q_{ik} (figura 1) observados num grupo de objetos similares de categoria q podem ser descritos pelo modelo de componentes principais (25, 26).

$$y_{ik}^{q} = \alpha_i^{(q)} + \sum_{a=1}^{A_q} \beta_{ia}^{(q)} \theta_{ia}^{(q)} + \varepsilon_{iq}^{(q)} \qquad |18|$$

onde A é o número de componentes e $\varepsilon_{iq}^{(q)}$ são os desvios devido aos erros aleatórios nas medidas. O valor de $\alpha_i^{(q)}$ é o valor médio da i-ésima variável para a q-ésima classe. Os valores $\beta_{ia}^{(q)}$ e $\theta_{ak}^{(q)}$ são obtidos por métodos de diagonalização, similar àqueles usados na transformação de Karhunen-Loeve.

Há formação da matriz $\underset{\sim}{Z} = \underset{\sim}{Y} - \alpha_i$, para um componente principal. Esta matriz é diagonalizada como $ZZ^t \approx \underset{\sim}{\beta}\ \underset{\sim}{\theta}\ \underset{\sim}{\theta}^t\ \underset{\sim}{\beta}^t$ e no final temos, $\underset{\sim}{\Lambda} = \theta\ \theta^t = \beta^t\ Z\ Z^t\ \underset{\sim}{\beta}$ onde $\underset{\sim}{\theta} \approx \underset{\sim}{\beta}^t\ \underset{\sim}{Z}$ ou $\underset{\sim}{Z} \approx \underset{\sim}{\beta}\ \underset{\sim}{\theta}$. Este argumento é similar para A_q maior que 1.

Este método de RP é baseado nos seguintes passos:

1) Para os dados $y_{ik}^{(q)}$ dos objetos no conjunto de treinamento, os parâmetros (A_q, β_{ia}^{q} e $\theta_{ia}^{(q)}$) são determinados para minimizar a soma dos desvios quadrados (ver equação |18|).

2) A classificação dos objetos com categorias não conhecidas, é então feita pelo ajustamento dos dados de cada destes tais objetos (denotados por x_{ij}) para cada modelo da classe Q (índice q) com parâmetros $\alpha_i^{(q)}$ e $\beta_{ia}^{(q)}$ fixados com os valores obtidos no passo 1. Este corresponde a uma regressão linear para cada objeto e cada classe, isto é, a determinação dos coeficientes $t_{aj}^{(q)}$ para minimizar os $e_{ij}^{(q)}$ residuais por interpretação dos mínimos quadrados:

$$x_{ij} - \alpha_i^{(q)} = \sum_{a=1}^{A} t_{aj}^{(q)} \beta_{ia}^{(q)} + e_{ij}^{(q)} \qquad |19|$$

O objeto é então atribuído à classe que contém a menor variança residual $|S_j^{(q)}|^2$:

$$|S_j^{(q)}|^2 = \sum_{i=1}^{M} \{e_{ij}^{(q)}\}^2 / (M-A) \ . \qquad |20|$$

Se a variança do conjunto teste é maior que o desvio padrão do conjunto de referência, o objeto não pertence a classe Q.

Diferente dos outros métodos de RP discutidos até agora, o SIMCA não desconsidera a possibilidade de um objeto não pertencer a nenhuma classe ou a possibilidade de pertencer a mais que uma.

Este método opera no segundo nível de RP e leva vantagem do KNN e PLANE, porque pode indicar que as amostras não pertencem a nenhuma categoria definida previamente, enquanto que KNN e PLANE classificam de acordo com as categorias definidas inicialmente.

No SIMCA as categorias são contidas num hipervolume no espaço M dimensional, como mostra o esquema da figura 6.

MÉTODOS DE CONHECIMENTO DE PADRÃO

Estes métodos podem ser usados para dados que não pos-

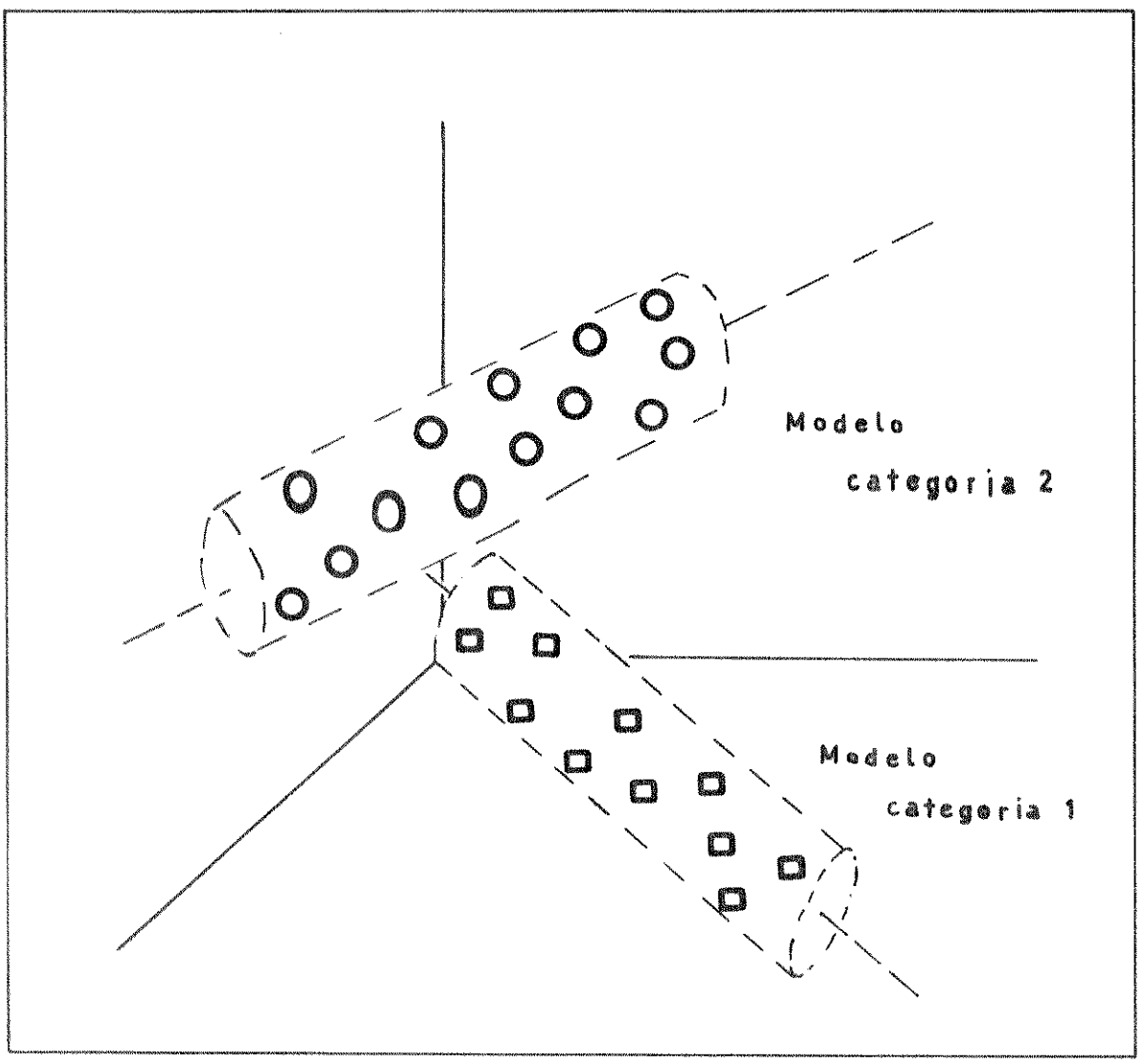


Figura 6. Os pontos dentro de um hipervolume pertencem a uma determinada categoria e os que ficam fora, não pertencem a nenhuma categoria.

suem conjunto de treinamento. Procuram altas densidades ou agrupamentos de pontos no espaço M dimensional. Estes métodos serão usados neste trabalho para investigar a estrutura geral dos dados em lugar de ser usado para finalidades de classificação.

AGRUPAMENTOS POR HIERARQUIA (HIER) - Encontra agrupamentos nos dados multidimensionais usando as medidas de similaridade, equação |2|, na forma de matriz. A matriz é percorrida para o maior valor e os dois pontos produzindo este valor são somados. Daí em diante os dois pontos são considerados como único (formando um centro de gravidade) para calcular uma nova e menor matriz de similaridade. Este processo continua até que todos os pontos estejam contidos em um agrupamento (3).

ÁRVORE DE VARREDURA MÍNIMA (TREE) - Esta técnica pode ser muitas vezes usada para investigar agrupamentos. Pode ser definida como sendo a árvore (conexão de todos os pontos no conjunto para seus vizinhos mais próximos), tal que o tamanho dos galhos da árvore é mínimo. Apesar da árvore resultante ser mínima neste sentido, os pontos equidistantes poderão ser rearranjados. Usando várias combinações das distâncias vizinhas e densidade dos pontos ligados, será possível separar os agrupamentos com sucesso (27).

4 - PROGRAMA

Faremos um breve resumo dos subprogramas dos métodos es

tatísticos e RP do programa ARTHUR usados neste trabalho que são úteis para análise de dados em geral.

SCALE - Escalamento

Pode ser autoescalamento ou intervalo de escalamento. O primeiro transforma os dados tal que a média é zero e variança igual a um, e o segundo transforma os dados tal que o valor máximo é um e o mínimo zero.

WEIGHT - Peso

Avalia a importância individual de cada característica para separar duas categorias. Avalia o peso de Fisher e o peso de variança.

KARLOV - Transformação de Karhunen-Loeve

Método de pré-processamento baseado na conservação da variança da amostra. Muitas vezes é usado como método de redução de dimensionalidade de um conjunto de dados.

SELECT - Seleção

Gera características ortogonais que são importantes para classificação, mas que também retém a identidade das medidas originais.

KNN - Regra do vizinho mais próximo

É usado para classificar uma amostra na categoria igual àquela em que a maioria dos seus vizinhos mais próximos perten-

cem.

PLANE e MULTI - Máquina de aprendizagem linear

Procuram os hiperplanos usando a máquina de aprendizagem linear no espaço M, tal que ele separa linearmente este espaço em regiões contendo só os pontos de uma categoria específica.

SIMCA - Statistical Isolinear Multiple Components Analysis

Análise de componentes principais que pode ser usado para classificar objetos no 2º nível de RP.

Existem outros métodos usados para classificação que são mais convenientes com dados contínuos, do que dados de categorias distintas considerados até agora, entre eles podemos citar:

PNN - Propriedade contínua do vizinho mais próximo

Este prediz a propriedade para cada padrão, sendo a média dos valores característicos dos seus vizinhos mais próximos (17).

STEP - Regressão múltipla linear por passos

Método sistemático para determinar as variáveis mais importantes na regressão múltipla linear (17).

PIECE - Regressão múltipla por pedaço

Faz regressões múltiplas pelo método dos mínimos quadra

dos naqueles pontos mais próximos ao ponto de interesse (17).

Além dos métodos citados o programa ARTHUR pode fazer gráficos em duas dimensões usando a impressora. Como o homem é um reconhecedor de padrão mais eficiente que o computador, estes gráficos darão uma idéia do comportamento geral dos dados. Para aplicações envolvendo M medidas (espaço M), existem 1/2 M (M-1) diferentes gráficos que serão examinados, não contendo erros de projeção e também não havendo perda de informação nos resultados (3). Por outro lado é difícil para o pesquisador analisar estes gráficos simultaneamente. Por isso os métodos de redução de espaço são importantes (reduzem o número de gráficos que precisam ser analisados). Geralmente os métodos usados são o KARLOV, SELEC, WEIGHT. Além do programa funcionar com os dados crus, existem outros métodos os quais podem transformar automaticamente as variáveis y_{ij} para y_{ij}^{-1}, y_{ij}^{-2}, y_{ij}^{2}, $\frac{y_{ij}}{y_{kj}}$, $\frac{1}{y_{ij}\ x_{kj}}$, $\log y_{ij}$, etc.

CAPÍTULO IV

RESULTADOS E DISCUSSÕES

Os dados analíticos, reproduzidos no apêndice A, foram analisados através do programa ARTHUR referido no capítulo anterior. Foram usadas as medidas das concentrações (características) de 6 traços de elementos. O conjunto de treinamento consta de 114 amostras e o conjunto teste de 11 amostras. O comportamento geral dos dados para estudar a melhor maneira de separar as diversas fontes de Serra Negra, Lindóia e Valinhos (categorias SN, L e V respectivamente), foi primeiramente investigado, usando estas características para "plotar" os $\frac{1}{2}$ M (M-1) diferentes gráficos (M é o número de características) bidimensionais possíveis. Através deste método vimos que o gráfico potássio x sódio era o que apresentava melhor capacidade de separação para as fontes de Serra Negra e Valinhos, enquanto que a separação das fontes de Serra Negra e Lindóia não era possível por uma maneira tão simples. Na figura 7, podemos ver que há uma sobreposição dos pontos para estas duas categorias. Ainda podemos observar que, como existe um plano que separa as amostras de Serra Negra e Valinhos, podemos dizer que estas duas características (K e Na) são suficientes para a discriminação destas.

Após este resultado testamos a habilidade das 6 caracte

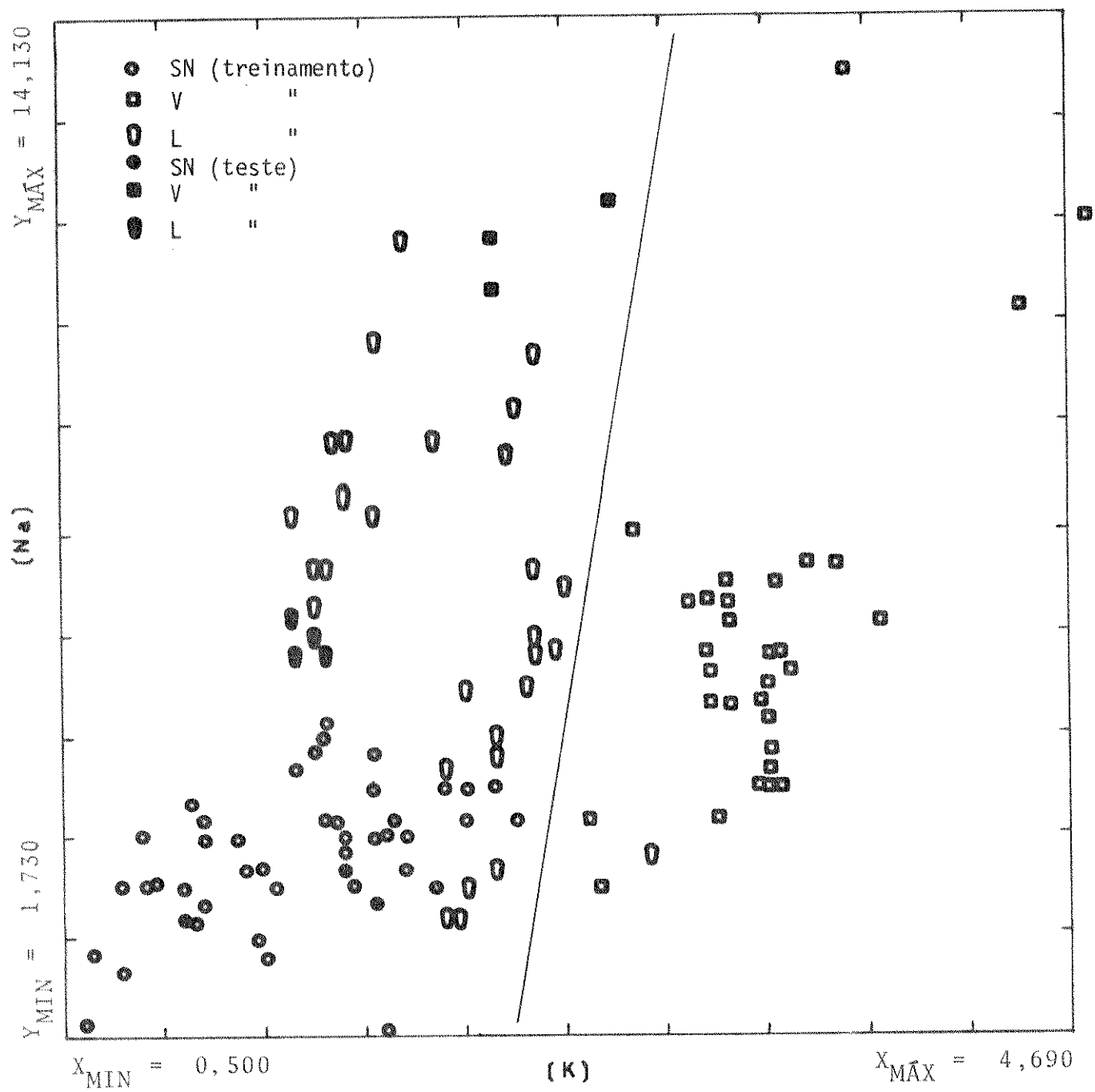


Figura 7. Gráfico da concentração de Na x K para as três categorias (SN, V e L).

rísticas usando os pesos de variança e Fisher para a classificação das fontes de água mineral, apenas com o conjunto de treinamento. As medidas foram pré-processadas por auto-escalamento para que fiquem com média zero e variança igual a 1. A tabela 14 mostra o resultado desta análise para o peso de variança (intervalo de $1,0 < W_V < \infty$), onde o menor valor do peso significa menor capacidade de separação. Podemos observar que o elemento potássio é um ótimo discriminador pois tem alto peso para separar as categorias SN x V, L x V, ou seja, é bom para separar a categoria V das categorias SN e L; esta separação pode ser vista na figura 7. Para as categorias SN e L este elemento não torna possível esta separação, talvez fosse, se houvesse uma combinação com as outras características. Da mesma maneira o sódio separa a categoria SN das categorias L e V. O silício é bom para separar as categorias SN e L da categoria V e cálcio separa um pouco SN das categorias L e V. Usando este mesmo argumento para o magnésio e o fósforo, vimos que estes não eram importantes para estas classificações.

Ainda para selecionar as melhores características, usamos um outro método para comparar com o anterior. O resultado da transformação de Karhunen-Loeve é mostrado na tabela 15, onde são apresentados os valores para os autovalores, a percentagem de variança correspondendo à cada autovetor, a percentagem

Tabela 14. Peso de variança* para separação das 3 categorias.

SN x L		SN x V		L x V	
Característica	Peso	Característica	Peso	Característica	Peso
Na	2,589	K	6,971	K	3,925
Si	2,389	Na	2,261	Si	1,509
K	2,131	Ca	1,534	Mg	1,087
Ca	1,399	Si	1,178	P	1,054
Mg	1,244	Mg	1,086	Ca	1,007
P	1,007	P	1,030	Na	1,005

* As características estão em ordem decrescente.

Tabela 15. Informações obtidas pela transformação de Karhunen-Loeve para as 3 categorias (SN, V e L).

	Autovalor	% Variança Total	% Variança Acumulada	Coeficientes para os autovetores					
				Ca	K	Mg	Na	Si	P
1	17,0400	68,3	68,3	-0,2003	-0,9196	-0,1104	-0,2911	-0,1274	-0,0319
2	4,720	18,9	87,2	-0,3701	-0,3562	-0,3502	-0,4819	-0,6094	-0,0997
3	1,089	4,4	91,6	-0,3004	-0,1049	-0,4694	-0,7707	-0,1816	-0,2269
4	1,022	4,1	95,7	0,2519	-0,0889	0.3655	-0,0053	-0,2718	-0,8492
5	0,9528	3,8	99,5	0,3469	-0,0804	0,2884	-0,2681	-0,7109	-0,4612
6	0,1163	0,5	100,0	0,7409	-0,0453	-0,6543	-0,1307	0,0128	0,0603

de variança acumulada e os coeficientes das concentrações dos elementos. Por exemplo nota-se que os 2 primeiros autovetores contêm 87,2% da variança total. A soma dos autovalores devem corresponder a 100% de variança total. Os autovetores são funções das concentrações multiplicadas por uma constante. Podemos então escrever:

$$KARL_1 = -0,2003\ [Ca] - 0,9196\ [K] - 0,1104\ [Mg] - 0,2911\ [Na] - 0,1274\ [Si] - 0,0319\ [P]$$

$$KARL_2 = -0,3701\ [Ca] + 0,3562\ [K] - 0,3502\ [Mg] - 0,4819\ [Na] - 0,6094\ [Si] + 0,0997\ [P] .$$

O $KARL_1$ tem 68,3% da variança total onde a única característica importante é o elemento potássio. A variança do $KARL_2$ é de 18,9% com o sódio e silício importantes. Este resultado vem confirmar o anterior, onde os elementos fósforo e magnésio podem ser desprezados. Este método tem também a finalidade de reduzir a dimensão do espaço de 6 para 2, desprezando os autovetores de pequena significância, isto é, desprezando $KARL_3$ até $KARL_6$ com uma perda de somente 12,8% de variança (ou informação). Na figura 8 o gráfico dos dois autovetores dá uma idéia da estrutura geral dos dados. Através desta figura podemos ver que as amostras de Serra Negra estão situadas em uma certa região, as amostras de Valinhos em outra, enquanto que asde Lindóia pa-

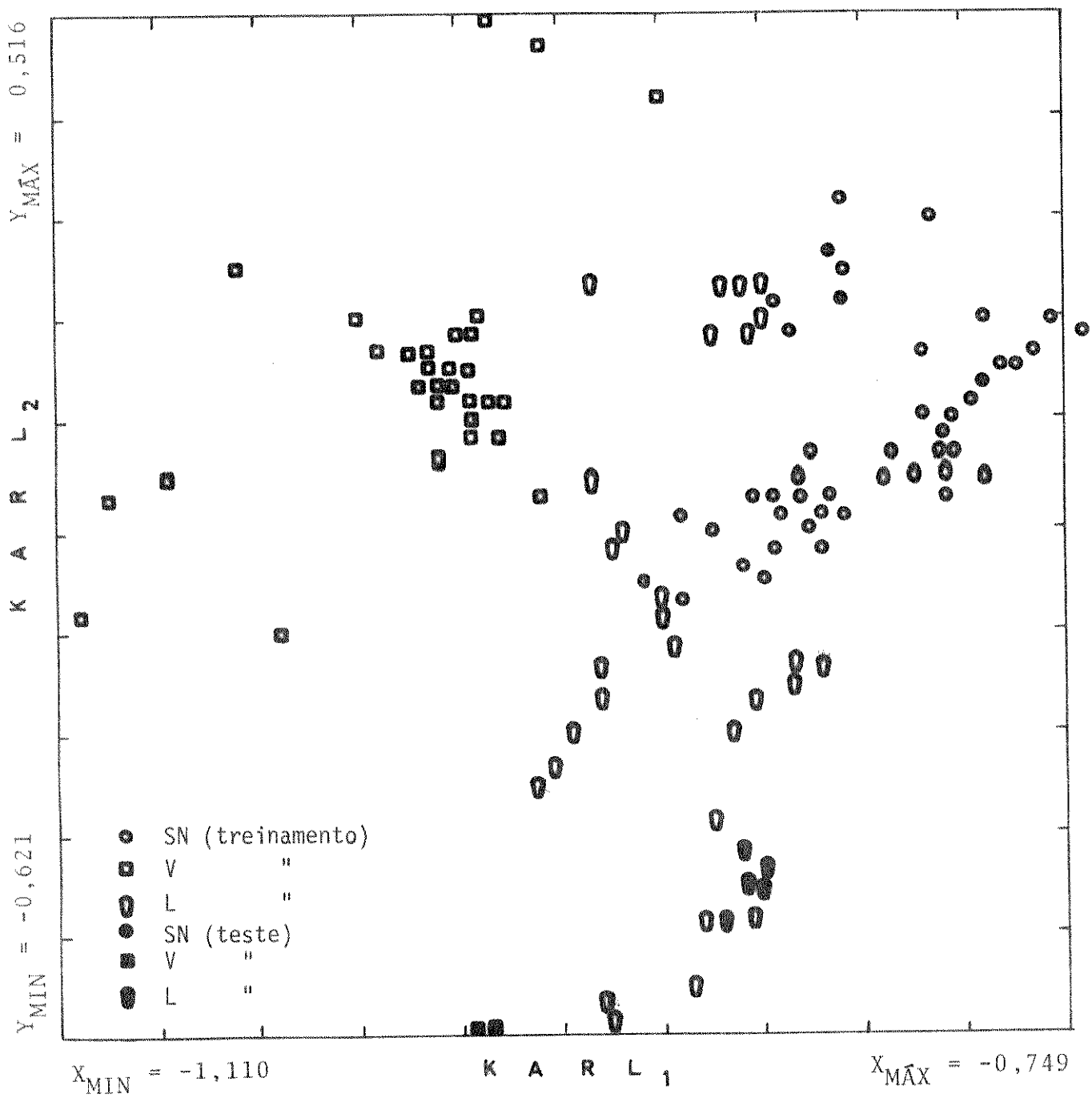


Figura 8. Projeção de Karhunen-Loeve em duas dimensões para as três categorias (SN, L, V).

recem estar divididas em 2 regiões, isto para o conjunto de treinamento. No conjunto teste as amostras da fonte Mécia de Valinhos estão mais próximas da região de Lindóia, enquanto que as amostras de Serra Negra e Valinhos estão contidas em suas respectivas regiões.

A classificação foi feita através da regra do vizinho mais próximo, tabela 16a. Para as amostras de Serra Negra, a maioria dos pontos foi classificada como categoria L. Para as amostras de Lindóia a maioria foi classificada como categoria V e a classificação para categoria V foi dada como categorias L e SN. O melhor resultado foi obtido com K = 1, onde a percentagem de classificação correta foi de 94,8% e com os demais vizinhos uma media de 90%. Apesar do melhor resultado ter sido obtido por K = 1 devemos usar K = 3, 5, 7 ou 9 visto que K = 1 é uma situação artificial, pode-se ver porém que independente dos 3, 5, 7 ou 9 vizinhos os resultados são mais ou menos iguais.

Para o conjunto teste este resultado é mostrado na tabela 16b, onde as amostras de Lindóia e Serra Negra foram corretamente classificadas, enquanto que as amostras da fonte Mécia incorretamente. Isto pode ser entendido, pois somente uma amostra desta fonte foi usada no conjunto de treinamento. Como pode ser visto na figura 8, os pontos desta fonte ficam longe das outras de Valinhos e mais próximas das fontes de Lindóia.

Tabela 16. Resultado da regra do vizinho mais próximo para as amostras do conjunto de treinamento (16a) e conjunto teste (16b).

Tabela 16a

Categoria	Nº de amostras	Nº de pontos classificados incorretamente				
		1 NN	3 NN	5 NN	7 NN	9 NN
SN	46	4	4	3	3	3
L	31	2	3	4	3	3
V	39	0	5	4	6	6
Total:	116	6	12	11	12	12
% de informação correta		94,8	89,7	90,5	89,7	89,7

Tabela 16b

Fonte	Embalagem	Classificação obtida				
		1 NN	3 NN	5 NN	7 NN	9 NN
N.Sra. Aparecida	garrafa plástica	SN	SN	SN	SN	SN
" " "	" "	SN	SN	SN	SN	SN
" " "	" "	SN	SN	SN	L	SN
São José	" "	L	L	L	L	L
" "	" "	L	L	L	L	L
" "	copo	L	L	L	L	L
" "	"	L	L	L	L	L
São Sebastião	garrafa vidro	L	L	L	L	L
Mécia	" "	L	L	L	L	L
"	" "	L	L	L	L	L
"	" "	L	L	L	L	L

O resultado do SIMCA é mostrado na tabela 17a. Este método forma 3 hipervolumes no espaço, sendo um para cada categoria. Os pontos que ficam dentro de um hipervolume são classificados de acordo com esta categòria (classificação correta) e os que ficam fora deste hipervolume podem ser classificados numa outra categoria ou podem ser classificados como amostras que não pertencem a nenhuma das categorias. Através desta tabela podemos ver que para as amostras de Serra Negra, 25 pontos foram classificados incorretamente, isto pode ser devido ao espalhamento geral dos pontos da categoria L (ver figura 8) pois como esta forma um hipervolume muito grande, as amostras de Serra Negra foram então classificadas como pertencentes a categoria L. As amostras de Lindóia foram classificadas corretamente e as amostras de Valinhos, talvez pelo mesmo motivo das de Serra Negra, tiveram sua classificação como categoria L.

Para o conjunto teste, tabela 17b, o resultado é semelhante ao KNN, onde a melhor classificação acontece com as amostras de Lindóia. As amostras de Serra Negra 2 foram classificadas como categoria L, como no conjunto de treinamento, e as amostras da fonte Mécia de Valinhos foram classificadas como não pertencentes a nenhuma categoria definida inicialmente.

Após estes resultados, investigamos novamente estes métodos para cada categoria separadamente, assim poderíamos ter u

Tabela 17. Resultado do método SIMCA para classificação das amostras do conjunto de treinamento (17a) e conjunto teste (17b).

Tabela 17a

Categoria	Nº de amostras	Pontos classificados incorretamente
SN	46	25
L	31	0
V	39	3

Tabela 17b

Categoria	Nº de amostras	Pontos classificados corretamente
SN	3	1
L	5	5
V	3	0

ma melhor idéia da estrutura dos dados. A primeira categoria estudada foi SN. Pelo diagrama de hierarquia, figura 9, pudemos observar, usando o valor de similaridade (ver equação |2|) de 0,65, quatro diferentes grupos dentro da mesma categoria. Este valor parece ser uma boa escolha porque para um menor, teríamos poucos grupos e um maior, muitos grupos, dificultando a análise.

A tabela 18 mostra o resultado da Transformação de Karhunen-Loeve. No primeiro autovetor,

$$KARL_1 = -0,4773\ [Ca] - 0,4401\ [K] - 0,4837\ [Mg] + \\ + 0,3660\ [Na] - 0,4587\ [Si] - 0,0161\ [P]$$

com exceção do fósforo, todos os outros coeficientes são importantes. Neste 65,6% de variança está contido. O segundo autovetor mais significante, contém 18,4% de variança. Este consiste quase que somente da concentração de fósforo,

$$KARL_2 = 0,0484\ [Ca] - 0,2481\ [K] - 0,0592\ [Mg] + \\ + 0,2971\ [Na] + 0,0454\ [Si] - 0,9177\ [P]$$

enquanto que o terceiro autovetor é determinado principalmente por sódio. Fazendo o gráfico dos dois autovetores mais importantes, figura 10, podemos representar simbólicamente os quatro grupos determinados pelo diagrama de hierarquia com S = 0,65. Infelizmente estes grupos aparentemente não tem pontos de ori-

Figura 9. Diagrama de hierarquia para as fontes da categoria SN.

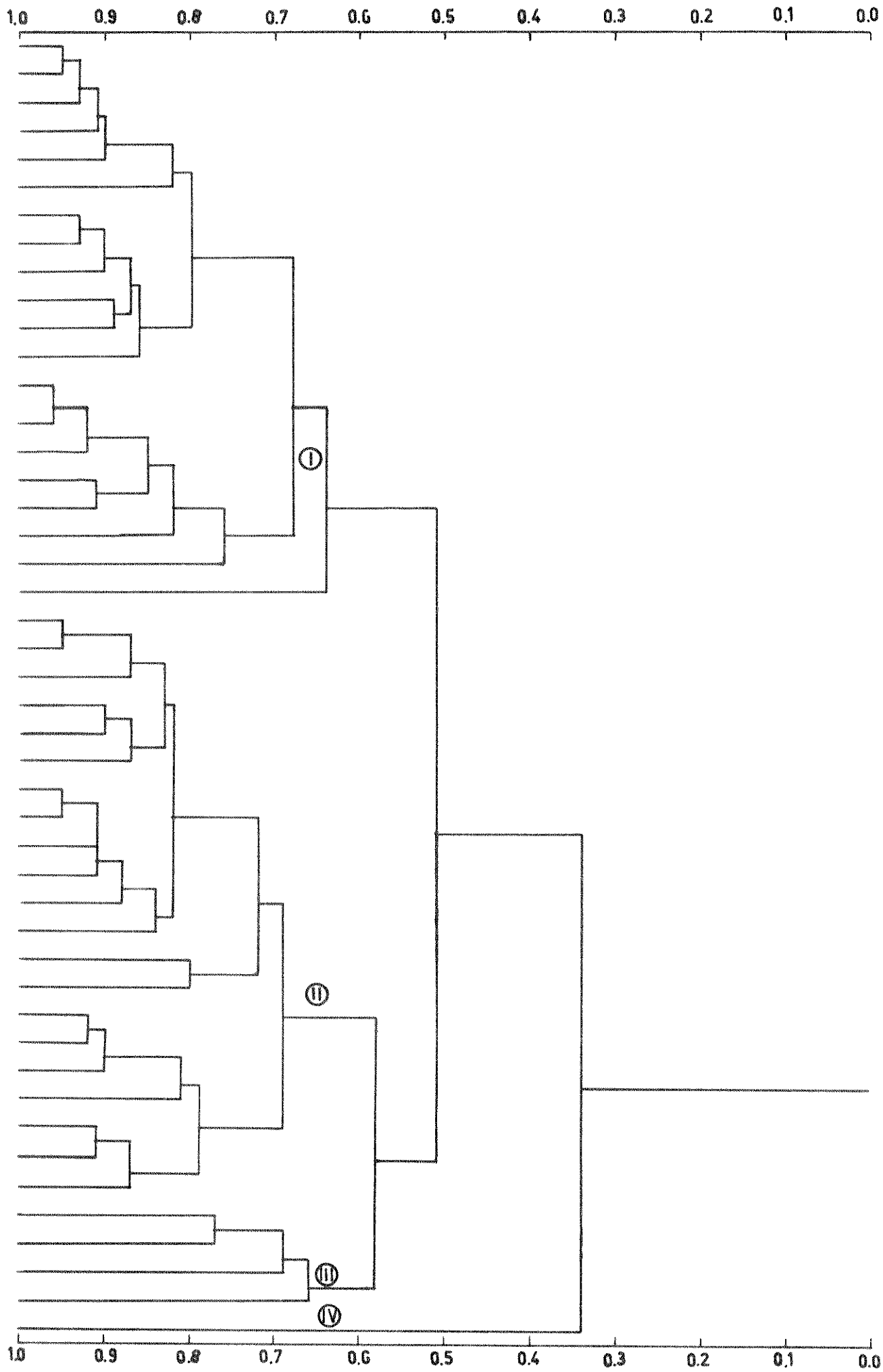


Tabela 18. Informações obtidas pela transformação de Karhunen-Loeve para a categoria SN.

	Autovalor	% Variança Total	% Variança Acumulada	Coeficientes para os autovetores					
				Ca	K	Mg	Na	Si	P
1	3,9350	65,6	65,6	-0,4773	-0,4401	-0,4837	-0,3660	-0,4587	-0,0161
2	1,1050	18,4	84,0	0,0484	-0,2481	-0,0592	0,2971	0,0454	-0,9177
3	0,5308	8,8	92,8	-0,2749	-0,2447	-0,2679	0,8107	0,1446	0,3386
4	0,2391	4,0	96,8	0,1164	-0,6346	0,0383	-0,3082	0,6894	0,1096
5	0,1556	2,6	99,4	-0,5297	0,5153	-0,3313	-0,1584	0,5385	-0,1705
6	0,0348	0,6	100,0	0,6326	0,1288	-0,7613	-0,0224	0,0373	0,0423

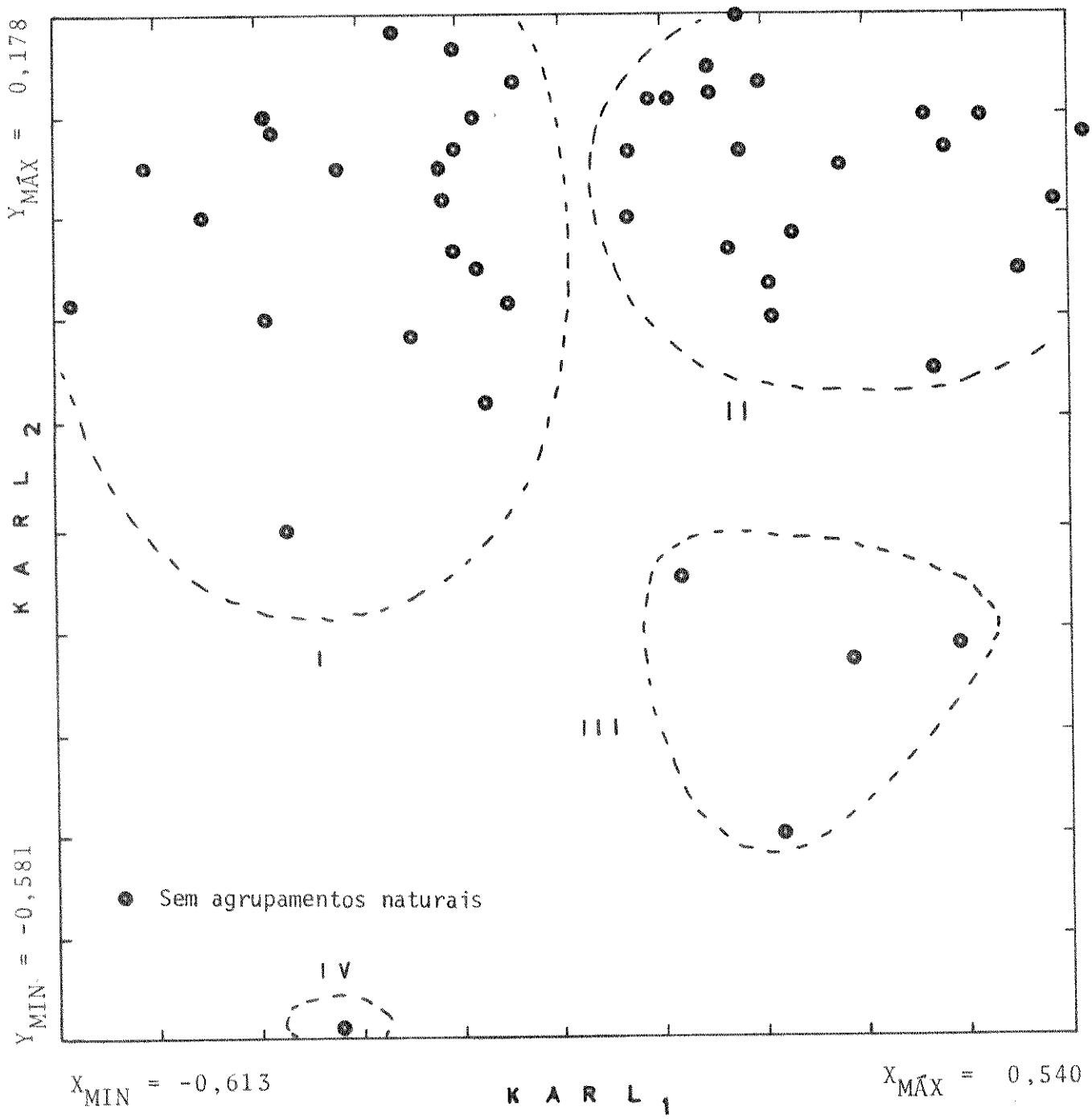


Figura 10. Projeção de Karhunen-Loeve para as fontes da categoria SN.

gem comum, por exemplo, fontes, época de escolha das amostras, etc.

A estrutura dos dados para categoria L é mostrada também através do diagrama de hierarquia, onde o valor de 0,55, aproximadamente, mostra 5 grupos, I, II, III, IV e V (figura 11). O resultado da transformação de Karhunen-Loeve é mostrado na tabela 19. O $KARL_1$ possui 48,6% de variança total:

$$KARL_1 = -0,4793\ [Ca] + 0,1861\ [K] - 0,5054\ [Mg] - 0,4817\ [Na] - 0,4795\ [Si] - 0,1351\ [P].$$

Neste, com exceção do fósforo e potássio, os outros coeficientes são importantes. O segundo autovetor com 23,6% de variança tem como maior coeficiente o elemento potássio:

$$KARL_2 = 0,3994\ [Ca] + 0,7290\ [K] + 0,2479\ [Mg] - 0,1547\ [Na] - 0,3202\ [Si] + 0,3481\ [P].$$

A figura 12, mostra o gráfico dos dois autovetores mais importantes. Através deste gráfico podemos ver os 5 grupos apontados pelo diagrama de hierarquia, onde todas as amostras dos grupos III e IV pertencem às fontes BC16A e São Jorge respectivamente. Quanto aos grupos I, II e V, não está havendo agrupamentos naturais, podemos com certeza dizer que não é devido à época da coleta.

Figura 11. Diagrama de hierarquia para as fontes da categoria L.

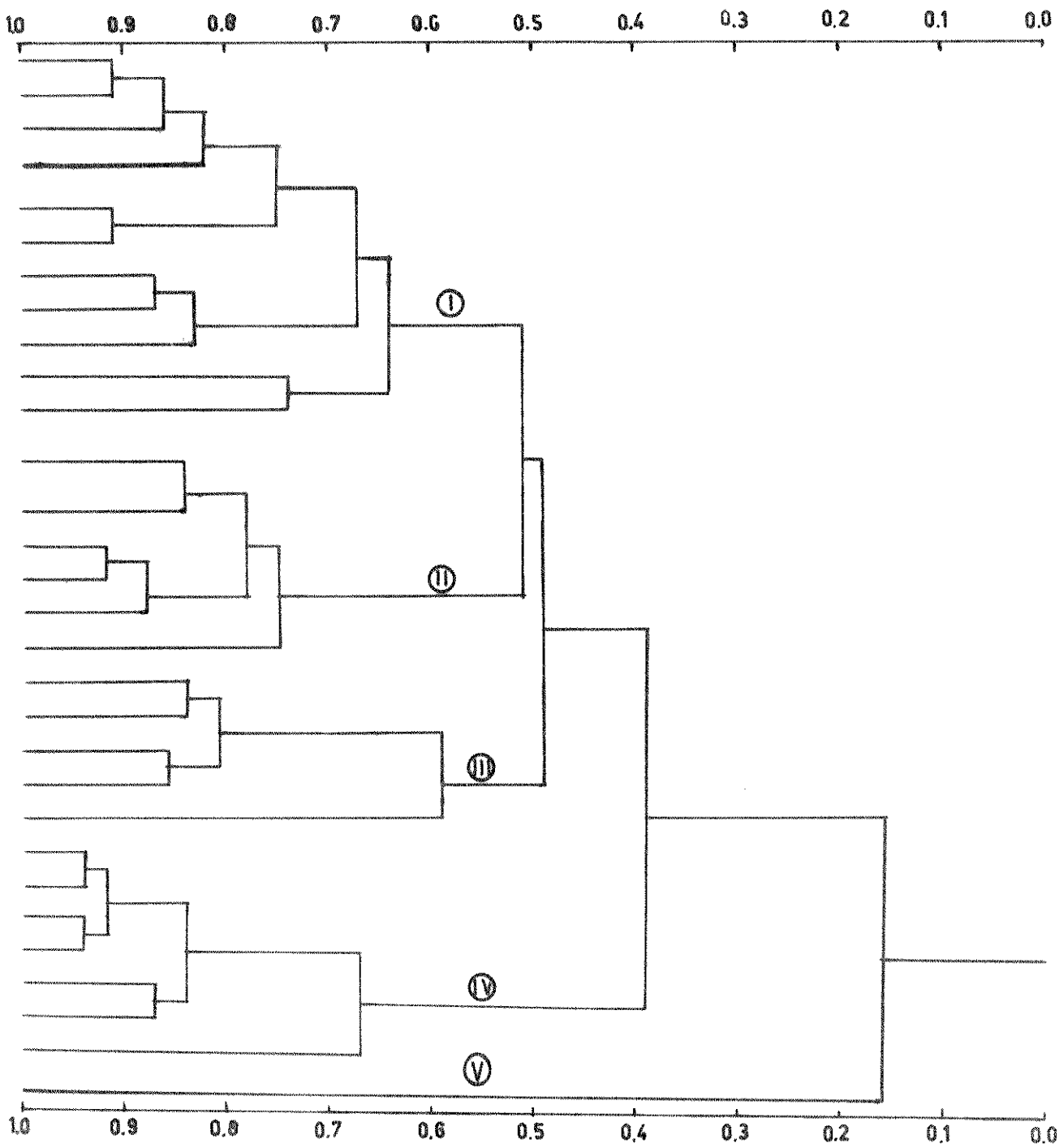


Tabela 19. Informações obtidas pela transformação de Karhunen-Loeve para a categoria L.

	Autovalor	% Variança Total	% Variança Acumulada	Coeficientes para os autovetores					
				Ca	K	Mg	Na	Si	P
1	2,9160	48,6	48,6	-0,4793	0,1861	-0,5054	-0,4817	-0,4795	-0,1351
2	1,4140	23,6	72,2	0,3994	0,7290	0,2479	-0,1547	-0,3202	0,3481
3	0,9208	15,3	87,5	0,2980	0,1614	0,1860	-0,1256	-0,0479	-0,9124
4	0,4373	7,3	94,8	0,0281	-0,4317	0,5138	-0,7260	-0,0472	0,1400
5	0,2735	4,6	99,4	0,1017	-0,3791	0,1629	-0,3954	-0,8142	-0,0122
6	0,0383	0,6	100,0	0,7147	-0,2789	-0,5984	-0,2120	-0,0048	-0,0915

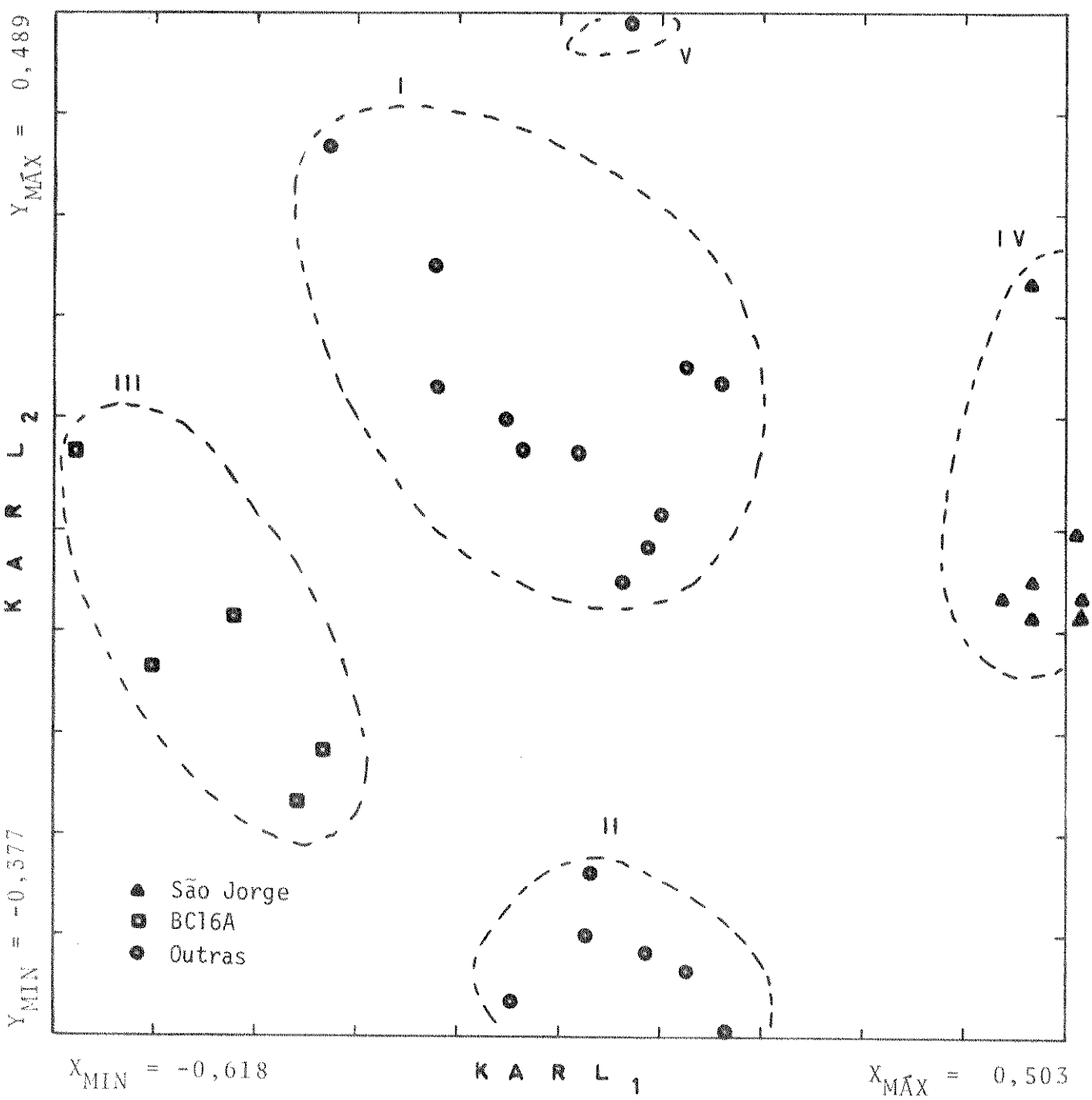


Figura 12. Projeção de Karhunen-Loeve para as fontes da categoria L.

Por último, para a categoria V o diagrama de hierarquia é mostrado na figura 13. Para o valor de similaridade de 0,65 temos 5 grupos, enquanto que para o valor de 0,70 temos 6 grupos.

A tabela 20 mostra os coeficientes mais importantes para os autovetores. O primeiro autovetor contém 49,7% de variança total,

$$KARL_1 = -0,5609\ [Ca] - 0,4199\ [K] - 0,4425\ [Mg] - 0,4698\ [Na] - 0,1912\ [Si] + 0,2367\ [P]$$

onde os coeficientes dos elementos Ca, K, Mg e Na são importantes. O segundo autovetor com 21,5% de variança, mostra que os coeficientes mais importantes,

$$KARL_2 = -0,0264\ [Ca] - 0,4239\ [K] - 0,0347\ [Mg] - 0,1344\ [Na] + 0,6165\ [Si] - 0,6483\ [P]$$

são dos elementos fósforo, silício e potássio e para o terceiro autovetor com 14,5% de variança, estes coeficientes são magnésio, sódio, silício e fósforo.

O gráfico $KARL_1$ x $KARL_2$, aparentemente mostra 4 grupos bem separados (figura 14). Segundo o diagrama de hierarquia com similaridade 0,7 temos 6 grupos. O maior grupo na figura 14 é dividido em 2 grupos, por este diagrama (ver figura 15). O gru

Figura 13. Diagrama de Hierarquia para as fontes da categoria V.

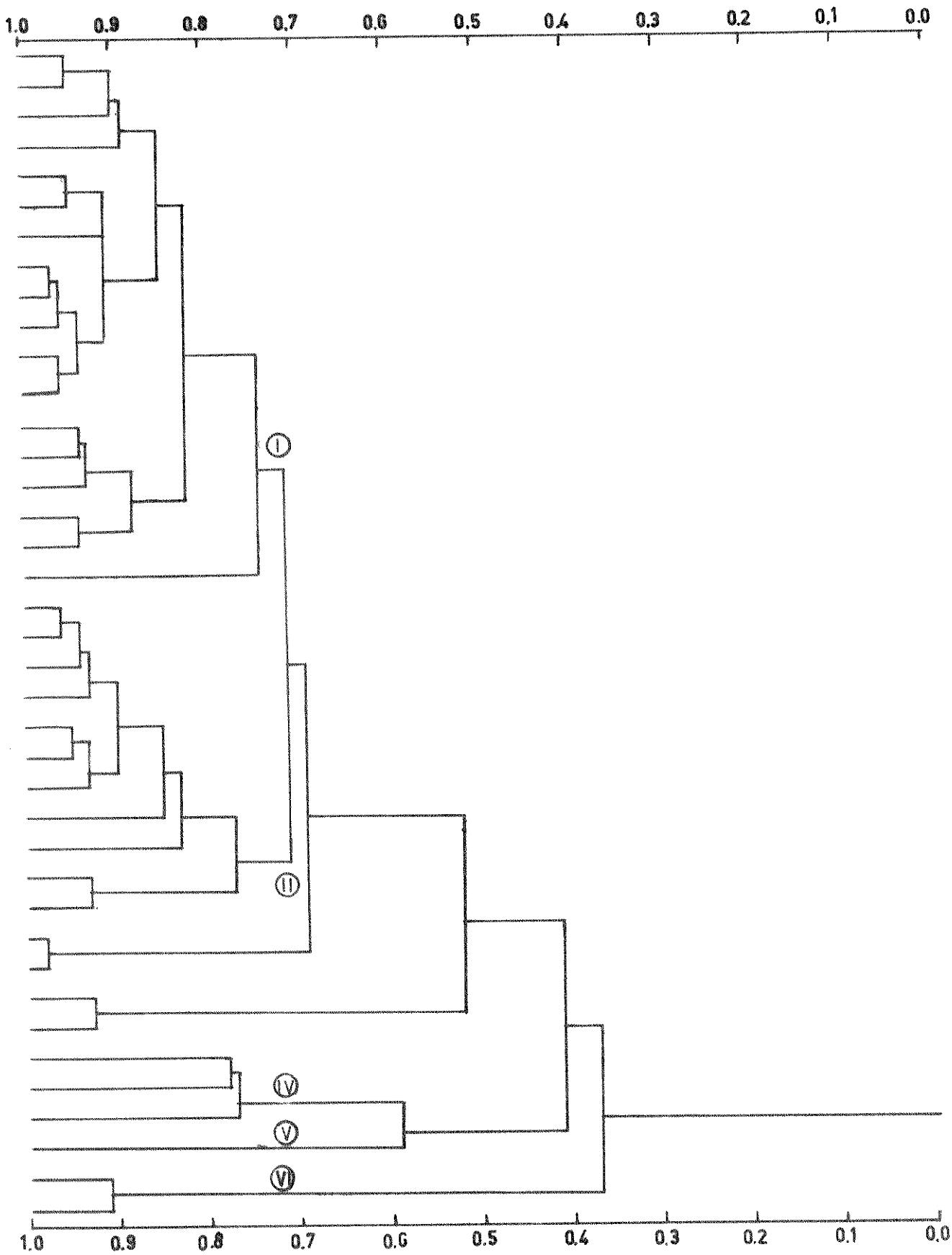


Tabela 20. Informações obtidas pela transformação de Karhunen-Loeve para a categoria V.

	Autovalor	% Variança Total	% Variança Acumulada	Coeficientes para os autovetores					
				Ca	K	Mg	Na	Si	P
1	2,9800	49,7	49,7	-0,5609	-0,4199	-0,4425	-0,4698	-0,1912	0,2367
2	1,8900	21,5	71,2	-0,0264	-0,4239	-0,0347	-0,1344	0,6165	-0,6483
3	0,8686	14,5	85,7	-0,0158	-0,1304	0,5564	-0,3649	-0,5938	-0,4329
4	0,4340	7,2	92,9	-0,0462	0,3283	0,4045	-0,6376	0,4574	0,3327
5	0,3557	5,9	98,8	-0,2036	0,7120	-0,4261	-0,1964	-0,0840	-0,4737
6	0,0710	1,2	100,0	0,8005	-0,1107	-0,3851	-0,4277	-0,1204	0,0346

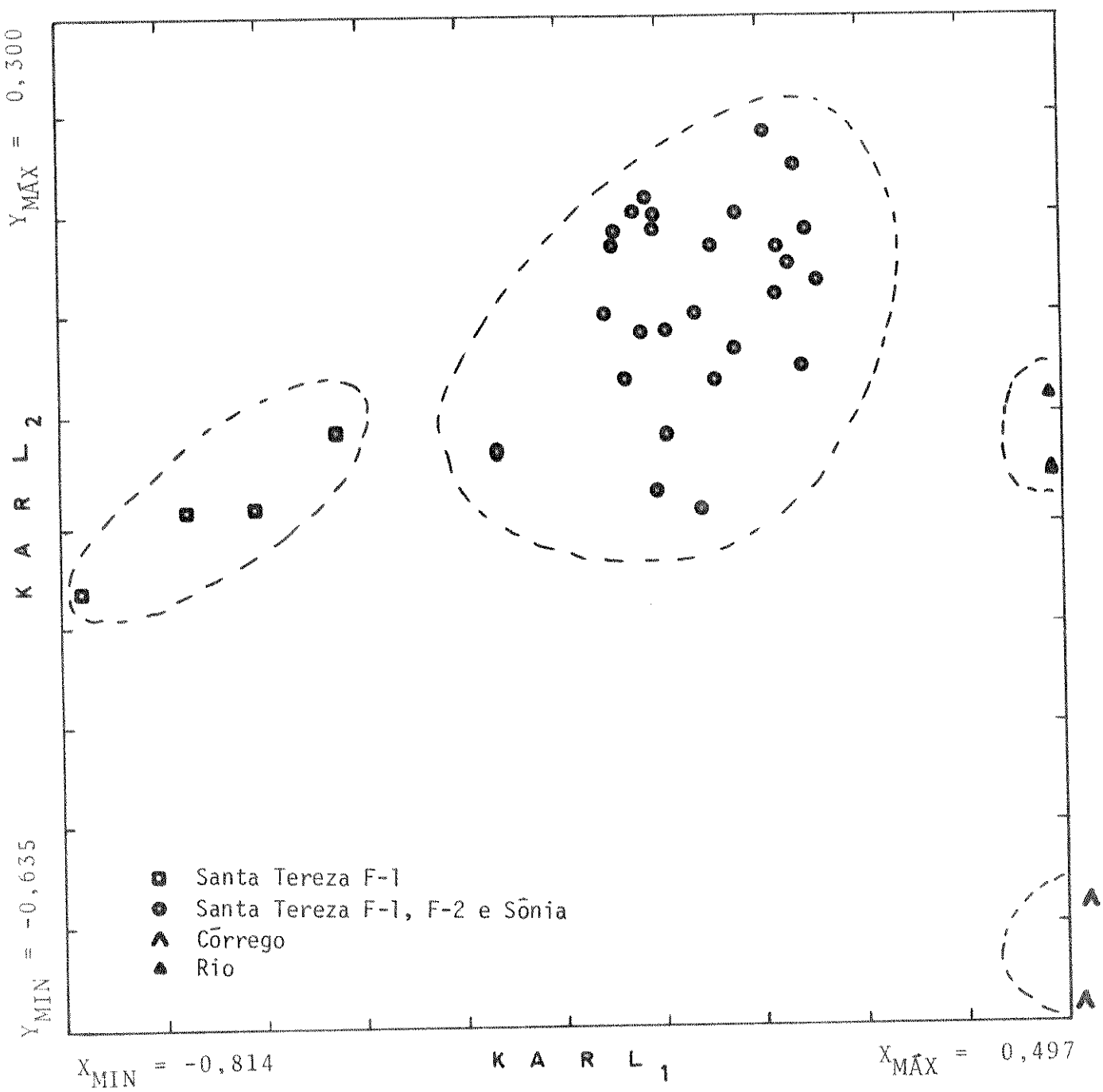


Figura 14. Projeção de Karhunen-Loeve para as fontes da categoria V.

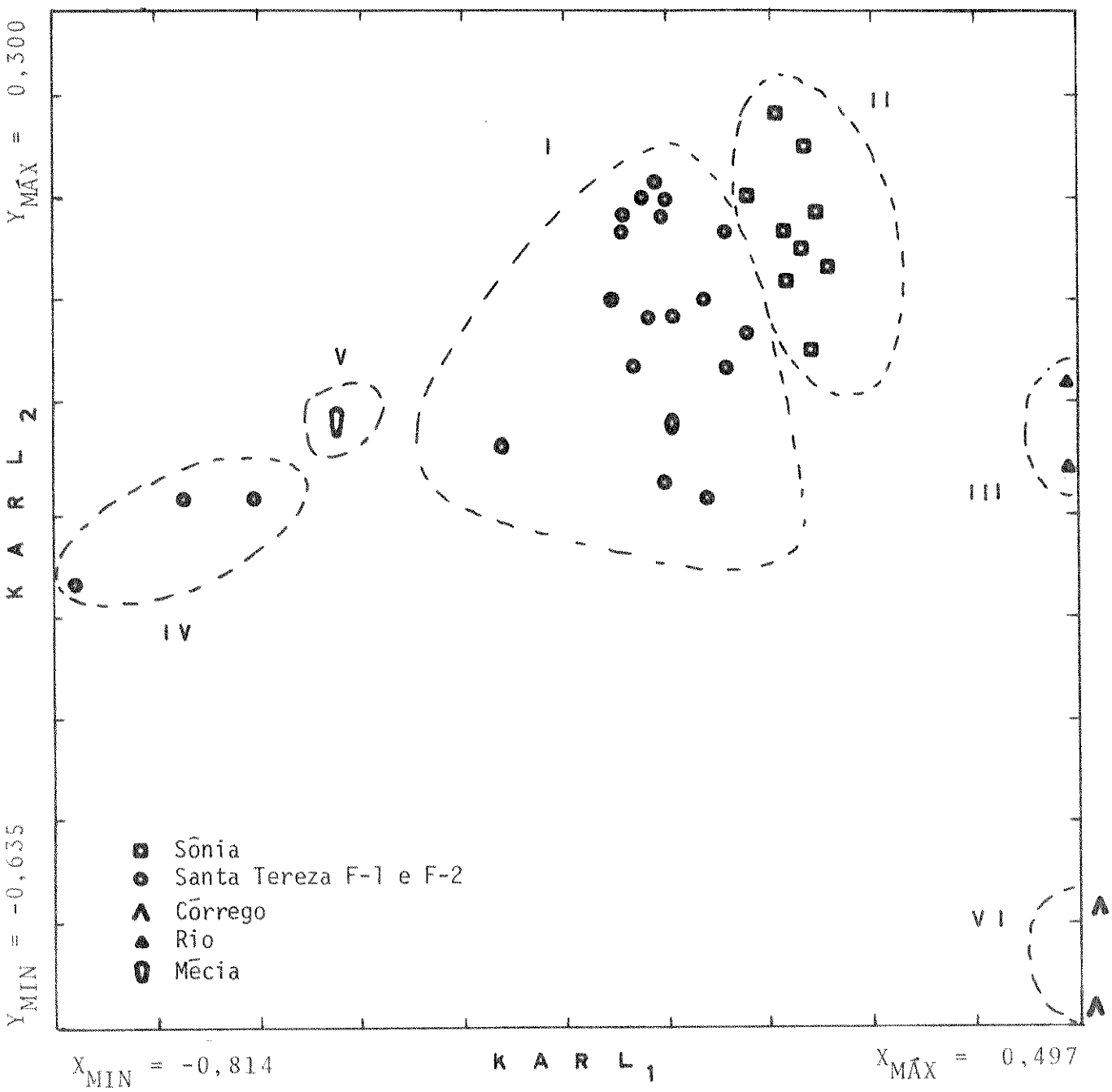


Figura 15. Projeção de Karhunen-Loeve para as fontes da categoria V.

po II nesta figura contém todos os pontos da fonte Sônia e os grupos III e VI contêm todos os pontos da água de um rio e de um córrego respectivamente, os quais passam dentro do local onde esta fonte está situada. No grupo I temos as amostras da fonte Santa Tereza F_1 e F_2 enquanto que no grupo IV temos pontos que pertencem somente à fonte Santa Tereza F_1 e o que chamamos de grupo V é a amostra (única) da fonte Mécia, que pertence ao conjunto de treinamento.

Além do gráfico dos dois primeiros autovetores, viu-se também que esta mesma separação era possível com $KARL_3$ x $KARL_1$, como mostra a figura 16. Isto implica que $KARL_1$ e $KARL_3$ são importantes para separar estas categorias, enquanto que $KARL_2$ não é útil nesta separação.

Apesar de se observar pelo diagrama de hierarquia e também pela transformação de Karhunen-Loeve que era possível separar as amostras das fontes Santa Tereza e Sônia, verificamos também se isto seria possível por meio da máquina de aprendizagem linear. O resultado foi excelente com 100% de classificação correta, onde as amostras da fonte Sônia ficaram em um lado do hiperplano e as amostras das fontes Santa Tereza do outro lado deste.

Por meio dos estudos anteriores pudemos observar que os elementos magnésio e fósforo, não eram tão importantes para a

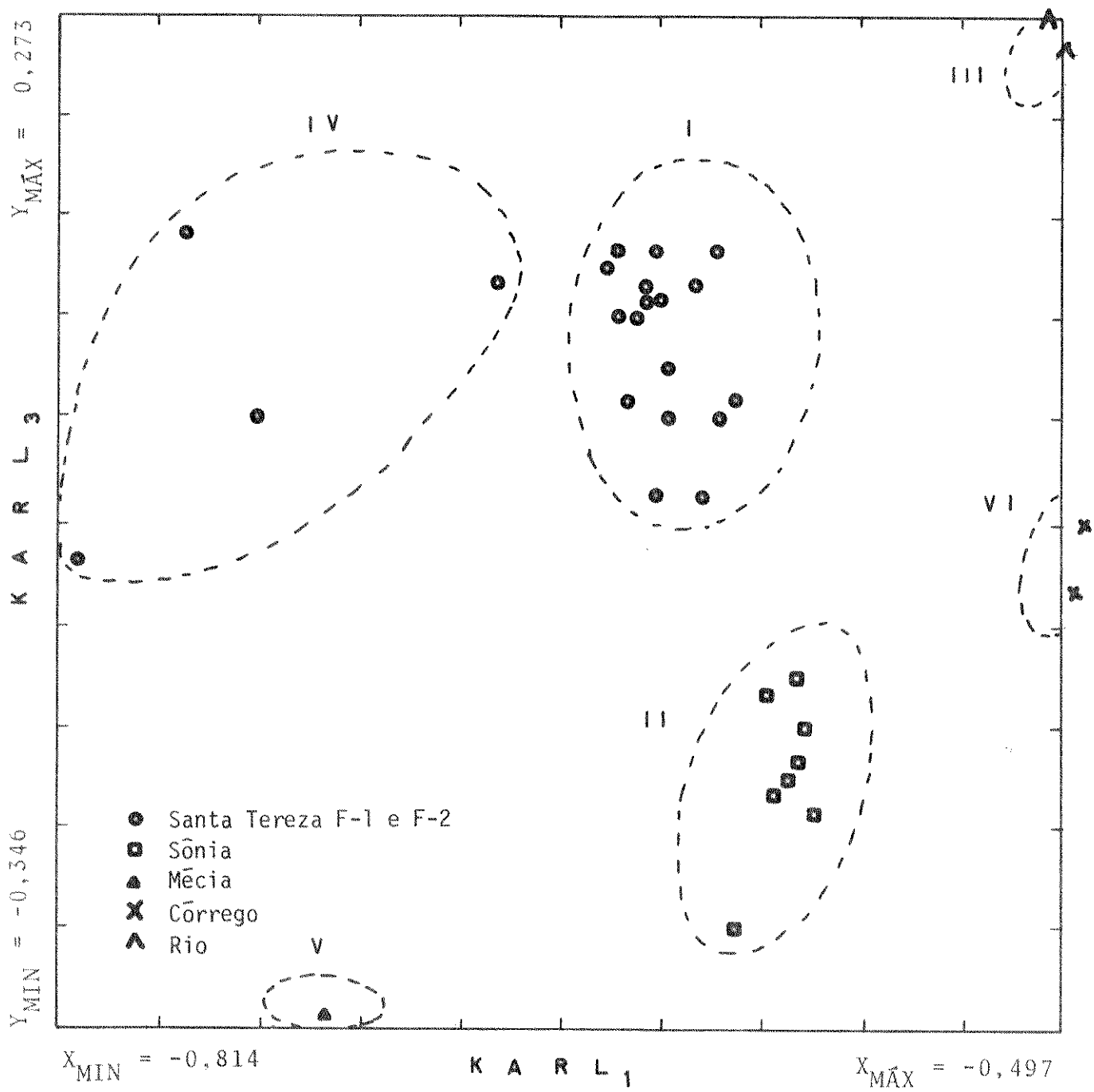


Figura 16. Projeção de Karhunen-Loeve para as fontes da categoria V.

discriminação das fontes de água mineral usadas neste trabalho, como mostram as tabelas 14 e 15. Fizemos então outro estudo, desprezando estes dois elementos, passando a usar apenas 4 características: sódio (Na), potássio (K), silício (Si) e cálcio (Ca).

Um outro fato que chamou a atenção foi que havia dois grupos de Lindóia bem separados entre si, figura 8. Pode-se verificar que o grupo menor só possui pontos pertencentes à fonte São Jorge, enquanto que os pontos correspondentes as outras fontes estão situados no outro grupo. Para ver se melhorava a classificação, resolvemos chamar este grupo como uma nova categoria, passando então a ter 4 categorias: Serra Negra, Lindóia, São Jorge e Valinhos (SN, L, SJ, V, respectivamente).

A habilidade das 4 características para discriminação destas fontes foram testadas usando o peso de variança. As medidas foram primeiramente autoescaladas.

A tabela 21 mostra o peso dos elementos para as 4 categorias. Seguindo o mesmo critério da tabela 14, de que os elementos com baixo peso não são bons discriminadores, podemos notar que o potássio continua separando bem as fontes de Lindóia e São Jorge das fontes de Valinhos e também dá uma separação razoável das fontes de Serra Negra e da fonte São Jorge. O sódio separa as fontes de Serra Negra das fontes de Lindóia e Vali-

Tabela 21. Peso de Variança* para discriminação das 4 categorias.

SN x L		SN x SJ		SN x V		L x SJ		L x V		SJ x V	
Caracte-rística	Peso	Caracte-rística	Peso	Caracte-rística	Peso	Caracte-rística	Peso	Caracte-rística	Peso	Caracte-rística	Peso
Na	5,278	K	3,355	Na	2,254	Na	5,487	K	6,487	K	5,837
Si	4,808	Ca	1,491	Ca	1,491	Ca	3,820	Si	2,648	Ca	3,415
Ca	2,031	Si	1,003	Si	1,141	K	1,247	Ca	1,151	Na	2,283
K	1,927	Na	1,000	K	1,044	Si	1,030	Na	1,117	Si	1,267

* As características estão em ordem decrescente.

nhos e também produz uma boa separação entre as fontes de Lindóia e São Jorge. O elemento silício separa as fontes de Serra Negra e Lindóia das de Valinhos, e por último o cálcio separa as fontes de Lindóia e São Jorge das fontes de Valinhos. Estes resultados estão de acordo com o apresentado anteriormente com as 6 características. A figura 17 continua praticamente igual, porque leva em consideração apenas as concentrações dos elementos sódio e potássio, a diferença consiste em um outro grupo (categoria) das águas da fonte São Jorge. Com este grupo separado, podemos ver por meio desta figura que há uma separação visível de Serra Negra e Lindóia.

Usando a Transformação de Karhunen-Loeve para também selecionar as melhores características, podemos observar pela tabela 22 que para o maior autovalor a que pertence a maior parte de variança, o maior coeficiente é do elemento potássio. O autovetor é dado por:

$$KARL_1 = -0,3461\ [Ca] - 0,7973\ [K] - 0,4162\ [Na] - 0,2670\ [Si].$$

No $KARL_2$, o segundo autovalor mais significativo, os maiores coeficientes são dos elementos silício e potássio. O autovetor é dado por:

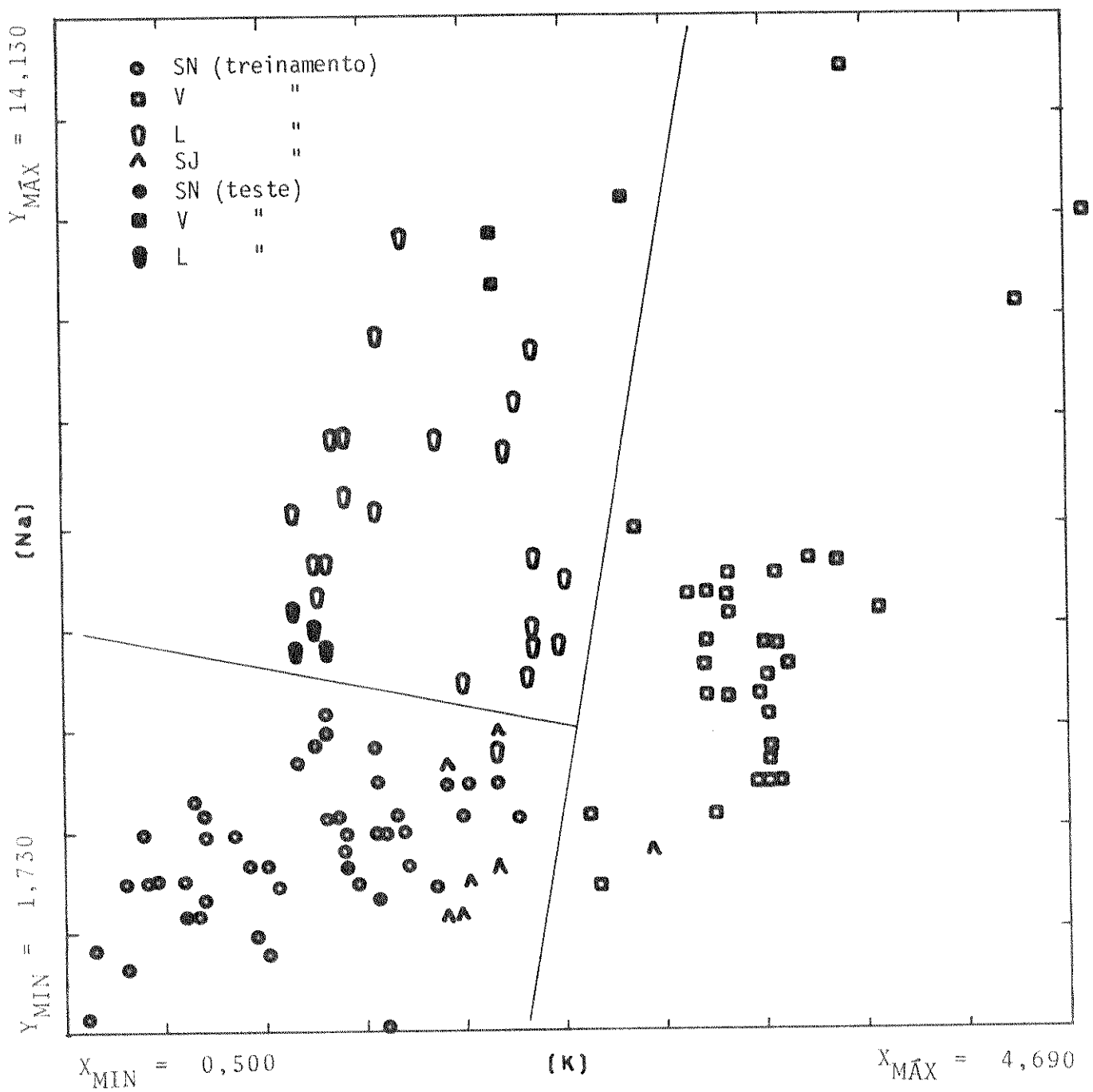


Figura 17. Gráfico da concentração de Na x K para as categorias SN, V, SJ e L.

Tabela 22. Informações obtidas pela transformação de Karhunen-Loeve para as 4 categorias (SN, L, V e SJ).

	Autovalor	% Variança Total	% Variança Acumulada	Coeficientes para os autovetores			
				Ca	K	Na	Si
1	19,390	64,5	64,5	-0,3461	-0,7973	-0,4162	-0,2670
2	7,487	24,9	89,4	-0,3155	-0,5516	-0,3543	-0,6861
3	1,901	6,3	95,7	-0,2077	0,2372	-0,6957	0,6455
4	1,297	4,3	100,0	0,8588	-0,0613	-0,4662	-0,2035

$$KARL_2 = -0,3155\ [Ca] + 0,5516\ [K] - 0,3543\ [Na] - 0,6861\ [Si].$$

E o $KARL_3$ onde 6,3% de variança podemos escrever,

$$KARL_3 = -0,2077\ [Ca] + 0,2372\ [K] - 0,6957\ [Na] + 0,6455\ [Si]$$

onde os coeficientes mais significativos são sódio e silício. Este resultado está de acordo com o peso de variança, onde os elementos com maiores pesos são: potássio, silício e sódio.

Ainda usando o mesmo método para reduzir o espaço, agora de 4 para 2 dimensões, (figura 18), pudemos notar que o resultado foi semelhante ao da figura 8. Isto comprova que os elementos magnésio e fósforo não são importantes para este caso.

Com a regra do vizinho mais próximo, o resultado foi o esperado, houve uma melhora na classificação passando de 94,8% com K = 1, para 97,3% e para os demais vizinhos uma media de 95,5% (tabela 23a). Podemos ver através desta que o resultado melhorou para todas as categorias. No conjunto teste o resultado não foi alterado (tabela 23b), para as amostras das fontes de Lindóia e Valinhos, apenas uma amostra de Serra Negra, que havia sido classificada como categoria L (ver tabela 16b), com K = 7, agora passou a ter classificação correta.

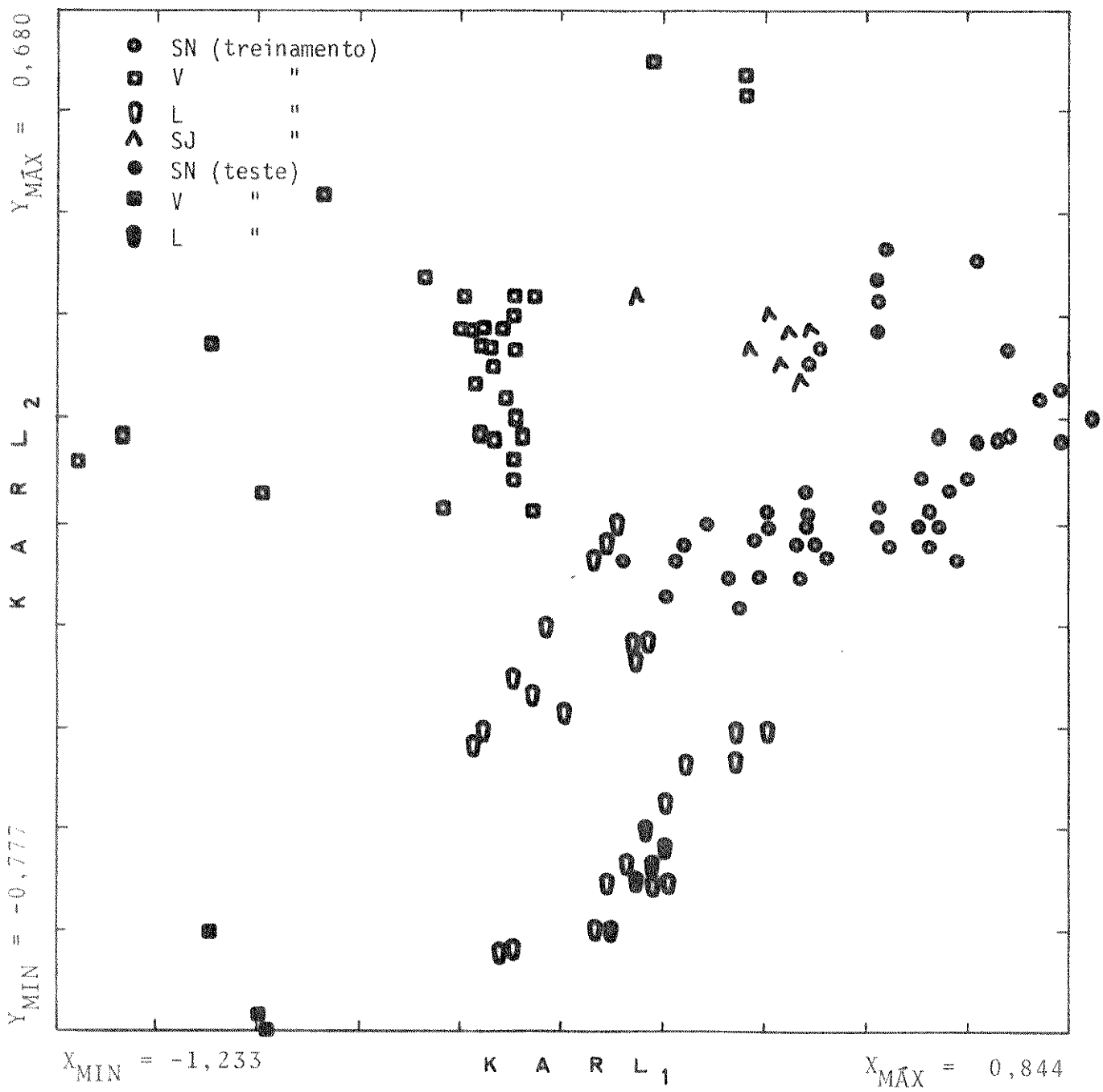


Figura 18. Projeção de Karhunen-Loeve para as categorias SN, V, L e SJ.

Tabela 23. Resultado da regra do vizinho mais próximo para as amostras do conjunto de treinamento (23a) e conjunto teste (23b).

Tabela 23a

Categoria	Nº de amostras	Nº de pontos classificados incorretamente				
		1 NN	3 NN	5 NN	7 NN	9 NN
SN	46	2	3	3	2	2
L	24	0	0	0	1	2
SJ	7	1	2	1	1	1
V	39	0	0	0	0	2
Total:	116	3	5	4	4	7
% de informação correta		97,3	95,5	96,4	96,4	93,8

Tabela 23b

Fonte	Embalagem	Classificação obtida				
		1 NN	3 NN	5 NN	7 NN	9 NN
N.Sra. Aparecida	garrafa	SN	SN	SN	SN	SN
" " "	"	SN	SN	SN	SN	SN
" " "	"	SN	SN	SN	SN	SN
São José	garrafa plástica	L	L	L	L	L
" "	" "	L	L	L	L	L
" "	copo	L	L	L	L	L
" "	"	L	L	L	L	L
Mécia	garrafa vidro	L	L	L	L	L
"	" "	L	L	L	L	L
"	" "	L	L	L	L	L

O resultado do SIMCA é mostrado na tabela 24a. Para a categoria SN houve uma melhora na classificação (ver tabela 18), passando de 25 pontos classificados incorretamente para 16 pontos, onde 14 destes foram classificados como categoria SJ e os outros 2 como categorias V e L. Este resultado poderia ser esperado pela figura 17, pois os pontos da categoria SN estão misturados com a categoria SJ. Para categoria L este resultado não foi tão bom, pois usando apenas três categorias não houve nenhum ponto classificado incorretamente, enquanto que agora possui 5, onde 4 foram classificados como categoria SN e apenas um como categoria V. Para categoria SJ todas as amostras foram classificadas corretamente e para a categoria V os 3 pontos foram classificados como categoria L, como na tabela 17a. No conjunto teste (tabela 24b) não houve alteração, ou seja, continuou igual o resultado da tabela 17b.

Tabela 24. Resultado do método SIMCA para classificação das amostras do conjunto de treinamento (24a) e conjunto teste (24b).

Tabela 24a

Categoria	Nº de amostras	Pontos classificados incorretamente
SN	46	16
L	24	5
SJ	7	0
V	39	3

Tabela 24b

Categoria	Nº de amostras	Pontos classificados corretamente
SN	3	1
L	5	5
SJ	-	-
V	3	0

CAPÍTULO V

CONCLUSÃO

No inicio deste trabalho usamos seis características para discriminar as fontes de água mineral, sendo elas determinadas pelos traços dos elementos cálcio, potássio, silício, sódio, magnésio e fósforo. Verificou-se porém que para discriminar as fontes das categorias SN e V, apenas os traços dos elementos sódio e potássio são suficientes (figura A). Vimos também que para a discriminação das fontes de Serra Negra, Lindóia, Valinhos e fonte São Jorge são necessários apenas quatro traços destes elementos: cálcio, potássio, silício e sódio. Isto porque os resultados obtidos, desprezando-se os traços de fósforo e magnésio, praticamente não foram alterados (ver figuras 8 e 18).

Observou-se ainda, que o KNN é um método bom para discriminar fontes de água mineral. O resultado com o mesmo usando as seis características, foi bom porque com K = 1 obtivemos aproximadamente 95% de classificação correta e com K = 3, 5, 7 e 9 uma media de aproximadamente 90%.

A classificação com o método SIMCA foi ruim, visto estar havendo espalhamento dos pontos da categoria L, onde algumas amostras das categorias SN e V foram classificadas como pertencentes àquela categoria.

Observou-se também que a classificação com o KNN poderia ser melhorada, se as amostras da fonte São Jorge (as quais estavam bem separadas das outras da categoria L) fossem definidas como uma nova categoria. Além disso usamos apenas as quatro características mais importantes. O resultado passou então de 95% com K = 1 para 97,3% de classificação correta e para os demais vizinhos uma média de aproximadamente 95,5%.

Quanto ao método SIMCA, a classificação continuou ruim, só que agora o hipervolume da categoria L ficou menor por ter removido as amostras da fonte São Jorge. Então houve uma pequena alteração na classificação, onde, para as amostras de Serra Negra verificou-se uma sensível melhora, enquanto para as amostras de Lindóia o resultado foi pior. Como as de Valinhos estavam longe destas duas categorias, o resultado não foi alterado. Portanto o SIMCA não foi um bom método para discriminação das fontes de água mineral com nossos dados, enquanto que KNN sim.

Para melhorar o método de classificação achamos que as amostras não deveriam ser acidificadas, já que aquelas do conjunto teste não foram. Um outro ponto importante é talvez quanto à análise das amostras. A melhor maneira seria analisá-las no mesmo dia, independente do tempo de coleta, se foram ou não acidificadas, etc. Assim o sistema estaria sob as mesmas condições de calibração para todas as amostras e haveria menor possi

bilidade de contaminação por outras amostras "sujas", como solos, plantas, etc. Pelas tabelas 7a e 7b pode-se verificar que fazendo as análises em dias diferentes pode-se obter resultados com desvios padrões significantes.

Quanto a análise de dados para serem usados na separação das fontes, estes podem ser obtidos pela espectrofotometria de absorção atômica, já que é preciso apenas quatro elementos.

Em relação aos objetivos do trabalho podemos enfatizar os seguintes ítens: a regra do vizinho mais próximo, peso de variança e a transformação de Karhunen-Loeve deram informações importantes para discriminação das fontes de águas minerais. Agora também temos um conjunto de dados que poderão ser usados para testar outros métodos de RP.

Para as amostras dos supermercados e distribuidores a melhor classificação foi com as de Lindóia e depois Serra Negra. Com as da fonte Mécia de Valinhos o resultado não foi bom, porque só havia uma amostra desta fonte no conjunto de treinamento.

APÊNDICE A

CONJUNTO DE DADOS USADOS

Amostra			Ca	K	Mg	Na	Si	P
1	0001S	1.	8.74	2.09	4.80	4.90	18.80	0.05
1	0002S	1.	5.11	1.48	2.49	5.33	16.61	0.11
1	0003S	1.	6.63	1.71	2.78	4.93	15.42	0.17
1	0004S	1.	7.64	1.71	3.84	4.19	17.19	0.05
1	0005S	1.	3.47	1.00	1.18	4.28	13.89	0.05
1	0006S	1.	2.01	0.77	0.64	4.27	14.36	0.15
1	0007S	1.	1.28	0.80	0.34	3.69	11.09	0.05
1	0008S	1.	1.09	0.55	0.33	2.75	10.71	0.11
1	0009S	1.	1.05	0.93	0.71	3.52	8.99	0.05
1	0010S	1.	0.50	0.68	0.29	3.55	9.91	0.05
1	0011S	1.	2.46	0.95	1.19	4.60	14.37	0.05
1	0013S	1.	2.66	1.62	1.29	3.56	9.99	0.11
1	0014S	1.	2.33	1.14	1.41	4.17	12.94	0.12
1	0027S	1.	8.21	2.10	4.67	4.34	18.16	0.06
1	0028S	1.	4.75	1.50	2.35	4.52	15.85	0.15
1	0029S	1.	6.21	1.60	2.73	4.02	14.87	0.18
1	0030S	1.	7.13	1.79	3.76	4.34	16.54	0.55
1	0031S	1.	0.66	0.50	0.21	2.02	10.28	0.01
1	0032S	1.	0.70	0.65	0.26	2.48	9.70	0.06
1	0033S	1.	1.07	1.25	0.72	2.71	8.39	0.28
1	0034S	1.	1.93	1.95	0.63	3.68	14.04	0.25
1	0035S	1.	0.93	0.76	0.30	3.52	11.23	0.19
1	0036S	1.	2.35	1.00	1.21	3.45	14.69	0.01
1	0038S	1.	2.52	1.71	1.31	1.76	10.42	0.23
1	0039S	1.	2.20	1.20	1.48	2.92	13.14	0.41
1	0040S	1.	3.31	1.85	2.02	3.82	12.46	0.01
1	0041S	1.	4.49	2.20	3.06	4.86	19.98	0.25
1	0042S	1.	3.20	1.00	1.14	3.33	13.62	0.12
3	0021L	3.	7.36	2.49	3.11	6.41	17.41	0.71
3	0022L	3.	15.04	2.50	6.90	7.30	21.49	0.19
3	0023L	3.	2.01	2.21	0.89	3.73	13.90	0.05
3	0024L	3.	3.43	1.58	1.40	9.02	21.12	0.23
3	0025L	3.	10.18	2.35	4.73	6.14	22.69	0.02
3	0026L	3.	12.26	1.85	7.90	11.41	23.13	0.33
2	0015V	2.	15.33	4.62	5.55	14.13	16.39	0.13
2	0016V	2.	6.37	3.50	3.17	7.53	14.75	0.05
2	0045V	2.	2.53	2.60	0.66	4.42	6.73	0.09
2	0046V	2.	5.74	3.10	1.26	7.16	17.40	0.23
2	0047V	2.	2.97	3.40	0.95	4.80	7.61	0.58
2	0048V	2.	7.40	3.39	3.35	6.55	15.49	0.27
2	0050V	2.	2.58	2.65	0.67	3.57	7.77	0.05
2	0051V	2.	5.86	3.03	1.25	7.04	18.79	0.18
2	0052V	2.	3.06	3.15	0.98	4.43	8.19	0.51
2	0053V	2.	7.69	3.37	3.36	6.40	16.04	0.01
2	0054V	2.	7.74	3.42	3.47	6.22	15.99	0.21
1	0055S	1.	7.57	2.00	4.01	4.92	16.54	0.14
1	0056S	1.	3.22	1.00	1.06	4.48	12.68	0.01
1	0057S	1.	6.62	1.60	3.20	4.21	14.60	0.10
1	0058S	1.	7.20	1.70	2.87	5.22	14.34	0.01

1	0A59S	1.	4.46	1.40	2.06	5.07	15.53	0.05
3	0060L	3.	11.13	1.40	6.55	8.23	22.08	0.01
3	0061L	3.	8.58	2.10	3.58	6.05	20.58	0.01
3	0062L	3.	4.12	1.70	1.58	10.19	23.95	0.01
3	0063L	3.	15.21	2.40	5.74	7.48	18.68	0.01
3	0064L	3.	2.07	2.20	0.82	5.38	13.30	0.01
3	0065L	3.	10.51	2.30	3.01	9.35	16.86	0.01
2	0066V	2.	13.16	3.70	4.77	13.53	14.16	0.10
2	0A67V	2.	6.54	3.20	2.81	6.81	15.37	0.16
2	0B67V	2.	6.82	3.80	3.04	6.90	15.70	0.10
2	0A68V	2.	5.29	3.20	0.94	7.41	18.65	0.14
2	0B68V	2.	5.50	2.80	0.99	7.92	18.48	0.05
2	0071AV	2.	5.86	3.20	1.09	7.11	18.32	0.19
2	0071BV	2.	5.44	3.10	1.02	6.48	18.67	0.14
2	0072V	2.	15.30	4.40	5.26	10.66	14.12	0.04
2	0073AV	2.	6.11	3.30	2.75	5.85	15.90	0.15
2	0073BV	2.	7.61	3.40	3.24	7.41	15.21	0.23
3	0074AL	3.	8.26	2.39	2.61	6.74	18.14	0.16
3	0074BL	3.	8.36	2.39	2.61	6.57	17.45	0.09
3	0075AL	3.	7.78	2.20	3.42	5.29	21.62	0.05
3	0075BL	3.	7.41	2.20	3.28	4.85	21.62	0.12
3	0076AL	3.	3.22	1.51	1.28	7.45	20.80	0.05
3	0076BL	3.	2.89	1.40	1.14	6.46	21.01	0.01
3	0077AL	3.	1.82	2.11	0.80	3.64	13.81	0.01
3	0077BL	3.	1.64	2.05	0.73	3.25	14.10	0.05
3	0078AL	3.	10.43	1.45	6.40	7.12	23.08	0.12
3	0078BL	3.	13.67	1.55	8.14	9.00	22.86	0.05
1	0079S	1.	3.66	1.30	1.78	3.62	15.91	0.08
1	0080AS	1.	2.44	0.95	0.84	3.08	13.19	0.05
1	0080BS	1.	3.25	1.00	1.09	4.13	13.21	0.08
2	0082V	2.	6.69	3.30	3.17	4.78	16.08	0.07
2	0082AV	2.	6.99	3.35	3.28	5.34	15.76	0.11
2	0083V	2.	10.26	4.42	3.87	5.76	15.56	0.09
2	0084AV	2.	5.67	3.10	1.16	6.31	19.12	0.22
2	0084BV	2.	5.48	3.10	1.11	5.94	19.09	0.10
2	0049V	2.	6.84	3.65	3.22	7.50	14.74	0.01
1	0059BS	1.	4.90	1.50	2.33	5.35	14.94	0.07
1	0059CS	1.	5.15	1.50	2.36	5.63	15.23	0.11
2	0086CV	2.	7.24	2.00	3.34	5.33	15.88	0.04
2	0086DV	2.	7.36	2.01	3.37	5.70	15.98	0.05
2	0087BV	2.	6.33	1.95	1.19	6.36	19.22	0.08
2	0087CV	2.	6.12	2.05	1.16	6.12	18.96	0.14
2	0096A	2.	7.51	3.36	3.58	4.89	17.11	0.05
2	0096B	2.	7.54	3.35	3.65	5.21	17.61	0.08
2	0097A	2.	7.81	3.35	3.54	5.60	16.83	0.15
2	0097B	2.	7.55	3.36	3.64	5.06	17.78	0.05
3	00101A	3.	11.68	2.25	3.14	8.86	19.01	0.10
3	00101B	3.	13.10	2.40	3.47	9.91	19.76	0.05
3	00102A	3.	3.85	1.45	1.31	7.42	21.25	0.11
3	00102B	3.	4.00	1.70	1.52	8.21	23.40	0.07
3	00103A	3.	2.37	2.00	1.07	5.03	13.07	0.09
3	00103B	3.	4.32	2.85	1.26	4.06	15.25	0.01
3	00103C	3.	2.45	2.00	1.03	3.27	15.01	0.05
3	00104A	3.	13.09	1.60	7.11	8.34	22.23	0.14

3	00104B	3.	14.87	1.95	8.73	8.89	25.18	0.04
1	00105	1.	11.82	2.30	5.26	4.43	17.84	0.10
1	00106A	1.	6.00	1.60	2.49	4.20	16.12	0.12
1	00106B	1.	6.02	1.55	2.53	4.52	16.78	0.09
1	00107A	1.	5.32	1.25	1.38	3.84	13.88	0.02
1	00107B	1.	4.80	1.15	1.29	3.78	14.62	0.03
1	00108A	1.	8.47	1.85	3.96	4.24	17.91	0.01
1	00108B	1.	8.86	1.75	3.96	4.27	18.05	0.03
2	0086A	2.	8.11	3.20	3.71	5.81	17.77	0.08
2	0086B	2.	8.19	3.35	3.65	6.11	17.17	0.08
2	0087A	2.	6.80	3.35	1.28	6.49	21.79	0.20
1	0059DS	1.	5.14	1.50	2.33	5.64	15.14	0.05
9999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999								
4	0069	4.	2.68	1.70	1.22	3.42	9.63	0.07
4	0085	4.	2.92	1.60	1.32	3.71	10.64	0.06
4	0088	4.	16.26	1.50	6.27	6.56	19.99	0.10
4	0089	4.	22.93	1.50	5.34	5.63	21.75	0.09
4	0090	4.	16.82	1.45	6.58	6.68	20.85	0.11
4	0092	4.	15.19	1.40	5.89	6.87	21.04	0.08
4	0093	4.	16.11	1.39	6.09	6.55	20.10	0.10
4	0094	4.	1.15	0.85	0.52	1.73	4.44	0.01
4	0091	4.	3.10	0.90	1.38	3.12	11.20	0.03
4	0098	4.	40.98	2.21	2.06	10.92	17.04	0.09
4	0099	4.	36.00	2.75	1.94	11.95	17.11	0.05
4	00100	4.	40.87	2.21	1.99	11.42	16.35	0.05
9999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999								

APÊNDICE B

REVISÕES GERAIS DAS APLICAÇÕES DE RP EM QUÍMICA

1. B.R. Kowalski, Anal. Chem., 52 (1980) 112R.

2. P.S. Shoenfeld e J.R. DeVoi, Anal. Chem., 48 (1976) 403R.

3. K. Varmuza, Anal. Chim. Acta, 112 (1980) 227.

4. P.C. Jurs e T.L. Isenhour, Chemical Applications of Pattern Recognition, Wiley-Interscience, New York, 1975.

5. B.R. Kowalski, Chemometrics: Theory and Application, Am. Chem. Soc., Washington, 1977.

REFERÊNCIAS

1. W.S. Meisel, Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.

2. H.C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition, Wiley Interscience, New York, 1972.

3. B.R. Kowalski e C.F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.

4. M. SjÖström e B.R. Kowalski, Anal. Chim. Acta, 112 (1979) 11.

5. B.R. Kowalski, Anal. Chem., 47 (1975) 1152 A.

6. C. Albano, W. Dum III, V. Edlund, E. Johanson, B. Nordém, M. SjÖström e S. Wold, Anal. Chim. Acta, 103 (1978) 1320.

7. W.U. Kawn e B.R. Kowalski, J. Food Science, 43 (1978) 1320.

8. B.E.H. Saxberg, D.L. Duewer, J.L. Booker e B.R. Kowalski, Anal. Chim. Acta, 103 (1978) 201.

9. B.R. Kowalski, T.F. Schatzki, F.H. Stross, Anal. Chem., 44 (1972) 2176.

10. D.L. Duewer e B.R. Kowalski, Anal. Chem., 47 (1975) 1573.

11. Perfil Analítico de Águas Minerais, Boletim 49, vol. II (1978), Ministério das Minas e Energia, Depto. Nacional da Produção Mineral.

12. D.L. Duewer, J.R. Koskinen e B.R. Kowalski, ARTHUR, dispo-

nível por B.R. Kowalski, Department of Chemistry, BG-10, University of Washington, Seattle, WA 98195.

13. Manual do "Inductively-Coupled Argon Plasma (ICPA) Source"; Jarrel-ash, modelo 96-750, Fisher Scientific Company.

14. A.L. Wilson, The Chemical Analysis of Water, The Society for Analytical Chemistry, London, 1974.

15. A.D. Shendrikar, V. Dharmarajan, H.W. Merrick e P.W. West, Anal. Chim. Acta, 84 (1976) 408.

16. Standard Methods for Examination of Water and Waste Water, APHA AWWA WPCF, 13a. ed., 1971.

17. D.L. Duewer, J.R. Koskinen e B.R. Kowalski, Documentation for ARTHUR, Version 1-8-75 (7), Chemometrics Society Report Nº 2.

18. K. Fukunaga e W.L.G. Koontz, IEEE Trans. Comp., C-19 (1970) 311.

19. B.R. Kowalski e C.F. Bender, J. Am. Chem. Soc., 95 (1973) 686.

20. B.R. Kowalski e C.F. Bender, Pattern Recognition, 8 (1976)1.

21. J.W. Sammon, Jr. IEEE Trans Comp., C-18 (1969) 401.

22. T.M. Cover e P.E. Hart, IEEE Trans. Comp., IT-13 (1967) 21.

23. N.J. Nilsson, Learning Machines, MacGraw Hill, New York, 1965.

24. P.C. Jurs, B.R. Kowalski e T.L. Isenhour, Anal. Chem., 41

(1969) 21.

25. O. Strouf e S. Wold, Acta Chem. Scand. 31A (1977) 391.

26. S. Wold, Pattern Recognition, 8 (1976) 127.

27. H.C. Andrews, Introduction to Mathematical Thecniques in Pattern Recognition, Wiley Interscience, New York, 1972.