PROGRAMA

1 -INTRODUÇÃO

2 -CLASSIFICAÇÃO AUTOMÁTICA

- 2.1 - Introdução
- 2.2 - Classificação não Hierárquica
- 2.3 - Classificação Hierárquica
- 2.4 - Medidas de Proximidade

3 -ANÁLISE DISCRIMINANTE

- 3.1 - Introdução
- 3.2 - Análise Discriminante(Caso Bigrupal)
- 3.3 - Análise Discriminante (Caso Multigrupal)
- 3.4 - Discriminação Baricêntrica

4 -ANÁLISE GEOESTATÍTICA DE DADOS

- 4.1 - Introdução
- 4.2 - Variogramas e Covariâncias Cruzadas
- 4.3 - Análise Factorial de Corregionalizações
- 4.4 - Krigagem Factorial

5 -REGRESSÃO

- 5.1 - Introdução
- 5.2 - Regressão Linear
  - 5.2.1 - Regressão Linear Simples
  - 5.2.2 - Regressão Linear Múltipla
- 5.3 - Regressão não Linear

6 -MODELAÇÃO EM RECURSOS NATURAIS

- 6.1 - Introdução
- 6.2 - Introdução aos modelos lineares

BIBLIOGRAFIA

Davis, J. C. (1986) - *Statistics and data analysis in Geology*. Wiley, 646 pp.

L. Lebart, A. Morineau & K. M. Warwick - *Multivariate Descriptive Statistical Analysis*. Wiley, New York, 1984.

Jambu, M. (1978) - C*lassification Automatique pour Analyse des Données.* Dunod, 310 pp..

Jambu, M. (1989) - *Exploration Informatique et Statistique des Données*. Dunod, 505 pp..

Pereira, H.G. (1985) - *Métodos de Classificação*. LMPM/IST, 10p..

Reis, E. (1997) – *Estatística Multivariada Aplicada*. Edições Sílabo, 343 pp..

Sharma, S. (1996) - *Applied Multivariate Techniques*. Wiley, New York, 493 pp..

Sneath, P.H.A., Sokal, R.R. (1973) - *Numerical Taxonomy*. W.H. Freeman, S. Francisco, 573 pp..

Sousa, A. J. (1989) - *Geoestatística Multivariada*, LMPM/IST, Lisboa, 15 pp.

Sousa, A. J. (2000) - *Análise Geoestatística de Dados*, CVRM / Centro de Geo-sistemas do IST, Lisboa, 17 pp.

CLASSIFICAÇÃO AUTOMÁTICA
(Taxonomia Numérica)

OBJECTIVO

Construção automática de grupos de amostras (indivíduos, objectos) ou variáveis (propriedades), no interior dos quais existe elevada *proximidade* (de acordo com um critério definido *a priori*). A proximidade entre os elementos de cada grupo deve ser maior do que em relação a qualquer elemento exterior.

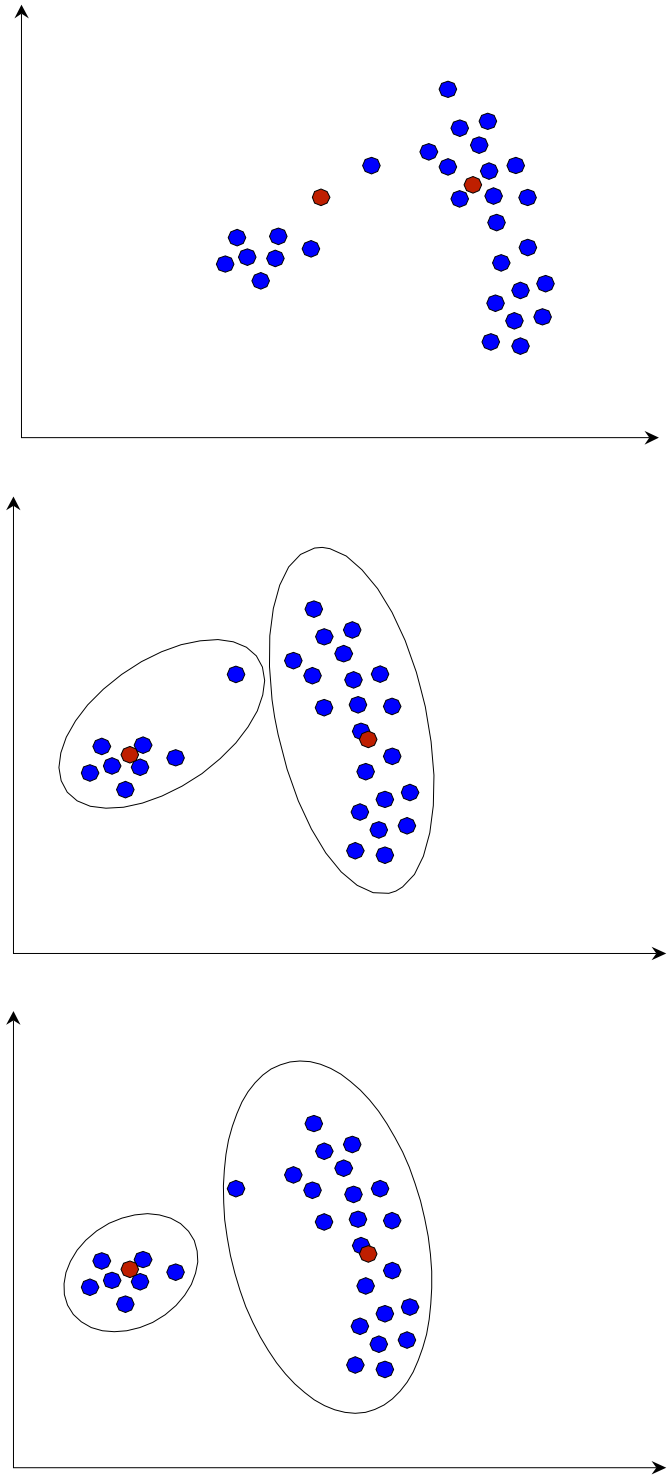
TIPOS DE MÉTODOS

- Não hierárquicos - Os grupos resultam de uma partição da matriz inicial em classes, conduzindo a uma estrutura em rede. Os elementos dos grupos são realocados dinamicamente em cada fase.

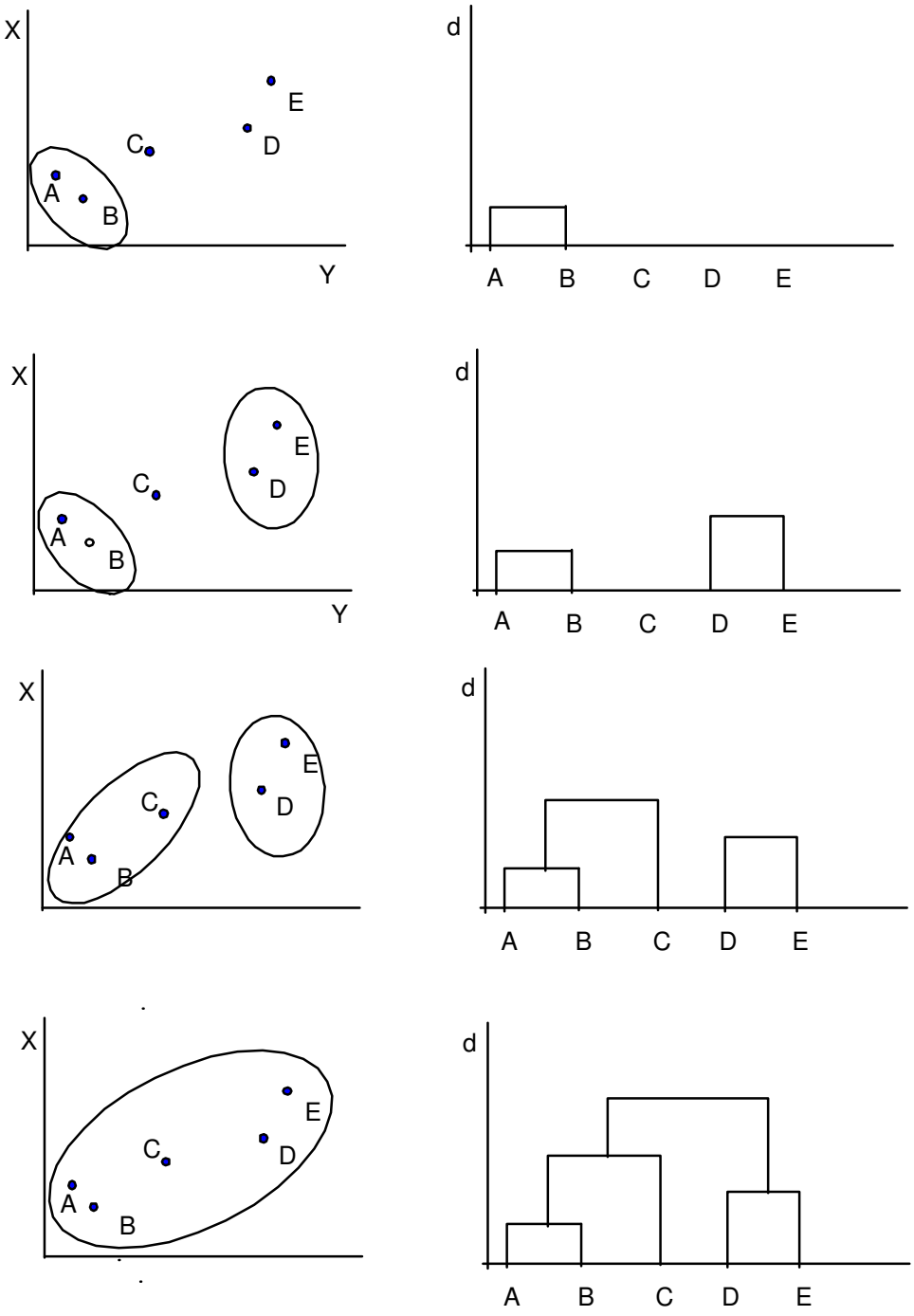
- Hierárquicos - Os grupos formados em cada fase vão sendo sucessivamente imbricados uns nos outros, conduzindo a uma estrutura em árvore. Os grupos formados em cada fase do algoritmo nunca mais se desfazem.

CLASSIFICAÇÃO AUTOMÁTICA

CLASSIFICAÇÃO NÃO HIERÁRQUICA



Classificação Ascendente Hierárquica



CLASSIFICAÇÃO AUTOMÁTICA

CLASSIFICAÇÃO ASCENDENTE HIERÁRQUICA

1. Calcular a matriz de distâncias (ou similitudes) entre todos os pares de indivíduos.

2. Seleccionar o par de indivíduos mais próximos (distância mínima ou similitude máxima).

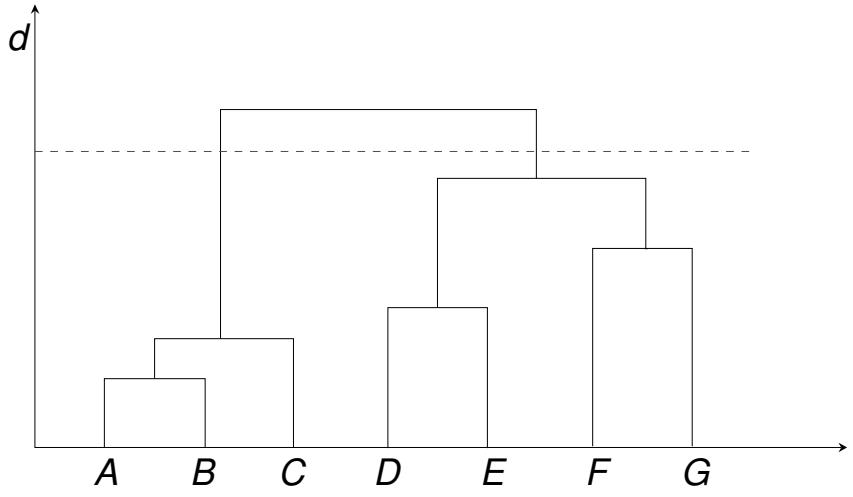
3. Calcular a distância (ou similitude) deste grupo a todos os restantes indivíduos e grupos já formados.

4. Reconstruir a matriz de distâncias (ou similitudes).

5. Reiterar o processo até todos os indivíduos estarem agrupados.

6. Construir o dendrograma.

7. Interpretar os resultados e dividir em grupos.



CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE PROXIMIDADE

As medidas de proximidade pretendem medir o grau de semelhança (similitude) ou de diferença (dissimilitude) entre duas amostras (indivíduos) ou variáveis (propriedades).

MEDIDAS DE SIMILITUDE

As medidas de similitude $c(a, b)$, que crescem com o aumento da proximidade, satisfazem às seguintes condições:

1. $0 \leq |c(a,b)| \leq 1$
2. $c(a,b) = 1$ se e só se a e b forem idênticos
3. $c(a,b) = c(b,a)$

Habitualmente esta medidas são conhecidas por medidas de correlação.

MEDIDAS DE DISSIMILITUDE

A medidas de dissimilitude $d(a, b)$, que decrescem com o aumento da proximidade, satisfazem às seguintes condições:

1. $d(a,b) \geq 0$
2. $d(a,b) = 0$ se e só se a e b forem idênticos
3. $d(a,b) = d(b,a)$

Habitualmente esta medidas são conhecidas por medidas de distância.

CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE SIMILITUDE

Em geral, as medidas de similitude podem ser transformadas em medidas de distância:

$$d(a,b) = 1 - c(a,b) \qquad \text{ou} \qquad d(a,b) = \sqrt{1 - c(a,b)}$$

MEDIDAS DE SIMILITUDE (VARIÁVEIS CONTÍNUAS)

Cos θ

$$d(a,b) = \sqrt{1 - \sum_j \frac{x(a,j) - m(j)}{s(j)} \; \frac{x(b,j) - m(j)}{s(j)}}$$

onde a e b representam duas amostras (indivíduos) e $m(j)$ e $s(j)$ são, respectivamente, a média e o desvio padrão da variável j.

Coeficiente de correlação

$$d(a,b) = \sqrt{1 - r}$$

onde r é o coeficiente de correlação entre as variáveis (propriedades, colunas) a e b.

CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE DISSIMILITUDE

Distância euclidiana

$$d(a,b) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left[x(a,j) - x(b,j)\right]^2}$$

Distância euclidiana reduzida

$$d(a,b) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left[\frac{x(a,j) - m(j)}{s(j)} - \frac{x(b,j) - m(j)}{s(j)}\right]^2}$$

onde $m(j)$ e $s(j)$ são, respectivamente, a média e o desvio padrão da variável j.

Distância *Chebychev*

$$d(a,b) = \text{Máximo}\left|x(a,j) - x(b,j)\right|$$

CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE DISSIMILITUDE

Distância *Manhatan* (*city block*)

$$d(a,b) = \sum_{j=1}^{p} \left| x(a,j) - x(b,j) \right|$$

Distância de *Gower*

$$d(a,b) = \frac{1}{p} \sum_{j=1}^{p} \frac{\left| x(a,j) - x(b,j) \right|}{R(j)}$$

onde $R(j)$ =Max[$x(j)$] – Min[$x(j)$] é a amplitude da propriedade j

Distância do χ^2

$$d(a,b) = \sqrt{\frac{1}{x(j)} \sum_{j=1}^{p} \left[\frac{x(a,j)}{x(a)} - \frac{x(b,j)}{x(b)} \right]^2}$$

onde $x(j)$ é a soma da coluna j *e* $x(a)$ e $x(b)$ são as somas das linhas correspondentes aos indivíduos a e b.

CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE SIMILITUDE (VARIÁVEIS ORDINAIS)

Coeficiente de correlação de Spearman

$$d(a,b) = \sqrt{1-r}$$

onde r é o coeficiente de correlação de Spearman entre as variáveis (propriedades, colunas) a e b.

CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE SIMILITUDE (VARIÁVEIS ORDINAIS)

Coeficiente de correlação de Spearman

$$d(a,b) = \sqrt{1-r}$$

onde r é o coeficiente de correlação de Spearman entre as variáveis (propriedades, colunas) a e b.

CLASSIFICAÇÃO AUTOMÁTICA

MEDIDAS DE SIMILITUDE (VARIÁVEIS BINÁRIAS)

SMC (*Simple Matching Coefficient*)

$$SMC = \frac{n(1,1) + n(0,0)}{n(1,1) + n(1,0) + n(0,1) + n(0,0)}$$

em que:

$n(1,1)$ - Nº de coincidências de 1 nos indivíduos a e b
$n(0,0)$ - Nº de coincidências de 0 nos indivíduos a e b
$n(1,0)$ - Nº de diferenças com 1 no indivíduo a e 0 no indivíduo b
$n(0,1)$ - Nº de diferenças com 0 no indivíduo a e 1 no indivíduo b

Índice de Jaccard

$$S_J = \frac{n(1,1)}{n(1,1) + n(1,0) + n(0,1)}$$

Índice de Russel & Rao

$$S_{RR} = \frac{n(1,1) + n(0,0)}{q}$$

em que:

q - Nº de variáveis qualitativas

CLASSIFICAÇÃO AUTOMÁTICA

ESTRATÉGIAS (CRITÉRIOS) DE AGLOMERAÇÃO

Salto mínimo (*Single Linkage*) - a distância entre dois grupos é a mínima distância entre os elementos de cada um dos grupos.

Salto máximo (*Complete linkage*) - a distância entre dois grupos é a máxima distância entre os elementos de cada um dos grupos.

Distância média (*UPGMA*) - a distância entre dois grupos é a média das distâncias entre os elementos de cada um dos grupos.

Distância média pesada (*WPGMA*) - a distância entre dois grupos é a média das distâncias entre os elementos de cada um dos grupos, pesadas pelos efectivos de cada grupo.

Centróides - a distância entre dois grupos é a distância entre os centros de gravidade (dados pela média dos valores das variáveis) de cada um dos grupos.

Método de Ward - em cada iteração, o novo grupo formado é aquele que minimiza a soma das variâncias inter grupos (*within-group variances*).

CLASSIFICAÇÃO AUTOMÁTICA

ESTRATÉGIAS (CRITÉRIOS) DE AGLOMERAÇÃO

Agrupamento flexível - se dois grupos *a* e *b* se fundem num grupo *k*, a distância entre este novo grupo e qualquer outro *h* é dada por:

$$d(k,h) = \alpha_1\, d(h,a) + \alpha_2\, d(h,b) + \beta\, d(a,b) + \gamma \left| d(h,a) - d(h,b) \right|$$

Estratégia	α_1	α_2	β	γ
Salto mínimo	0.5	0.5	0	0.5
Salto máximo	0.5	0.5	0	-0.5
Distância média	$\frac{n_a}{n_a + n_b}$	$\frac{n_b}{n_a + n_b}$	0	0
Método de Ward	$\frac{n_a}{n_k + n_h}$	$\frac{n_b}{n_k + n_h}$	$\frac{-n_h}{n_k + n_h}$	0

CLASSIFICAÇÃO AUTOMÁTICA

DADOS (RESULTADOS ELEITORAIS – 1987)

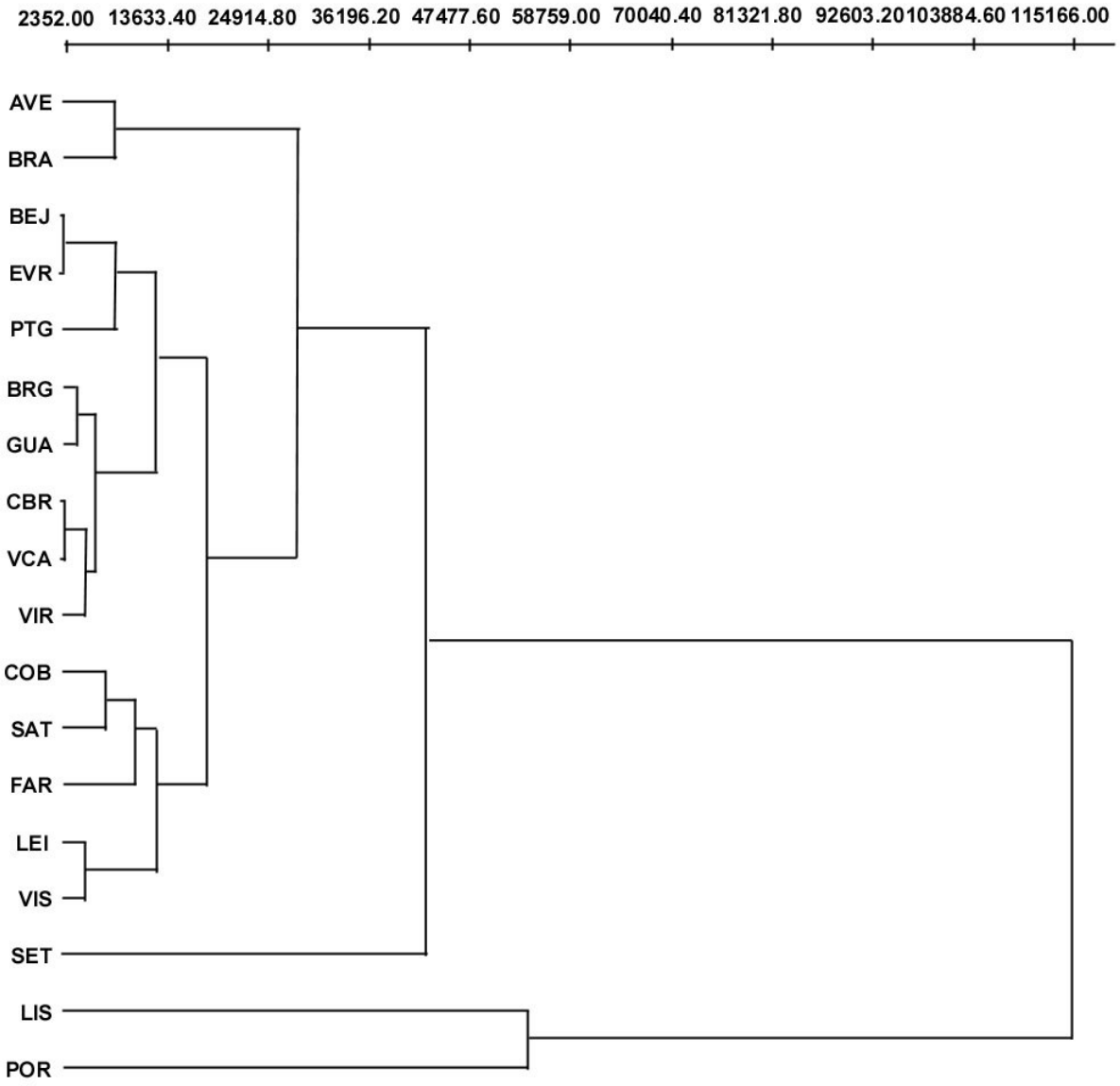
	PCP	PS	PSD	CDS	PRD
AVE	15832	82230	215623	18981	9517
BEJ	39581	20730	25098	2037	5869
BRA	24600	103935	214141	23737	13407
BRG	3026	17757	56413	7003	1188
CBR	9757	30848	71794	6425	8189
COB	17367	69962	121640	10988	8412
EVR	39750	17002	35294	2314	8474
FAR	20734	47401	8676	5901	11887
GUA	3879	25493	70069	7603	2366
LEI	14311	45270	146831	14600	7569
LIS	202917	261079	563845	45419	84509
PTG	18052	21911	32520	2657	5515
POR	87341	249451	475591	36999	37570
SAT	33740	57947	127870	9572	19592
SET	128973	69406	128334	7411	34132
VCA	8737	28339	76107	10751	6751
VIR	5545	27542	85303	6759	1838
VIS	6538	40705	144148	15662	3921

CLASSIFICAÇÃO AUTOMÁTICA

DISTÂNCIA EUCLIDIANA

$$d(a,b) = \sqrt{\frac{1}{p}\sum_j [x(a,j) - x(b,j)]^2}$$

Dendrograma

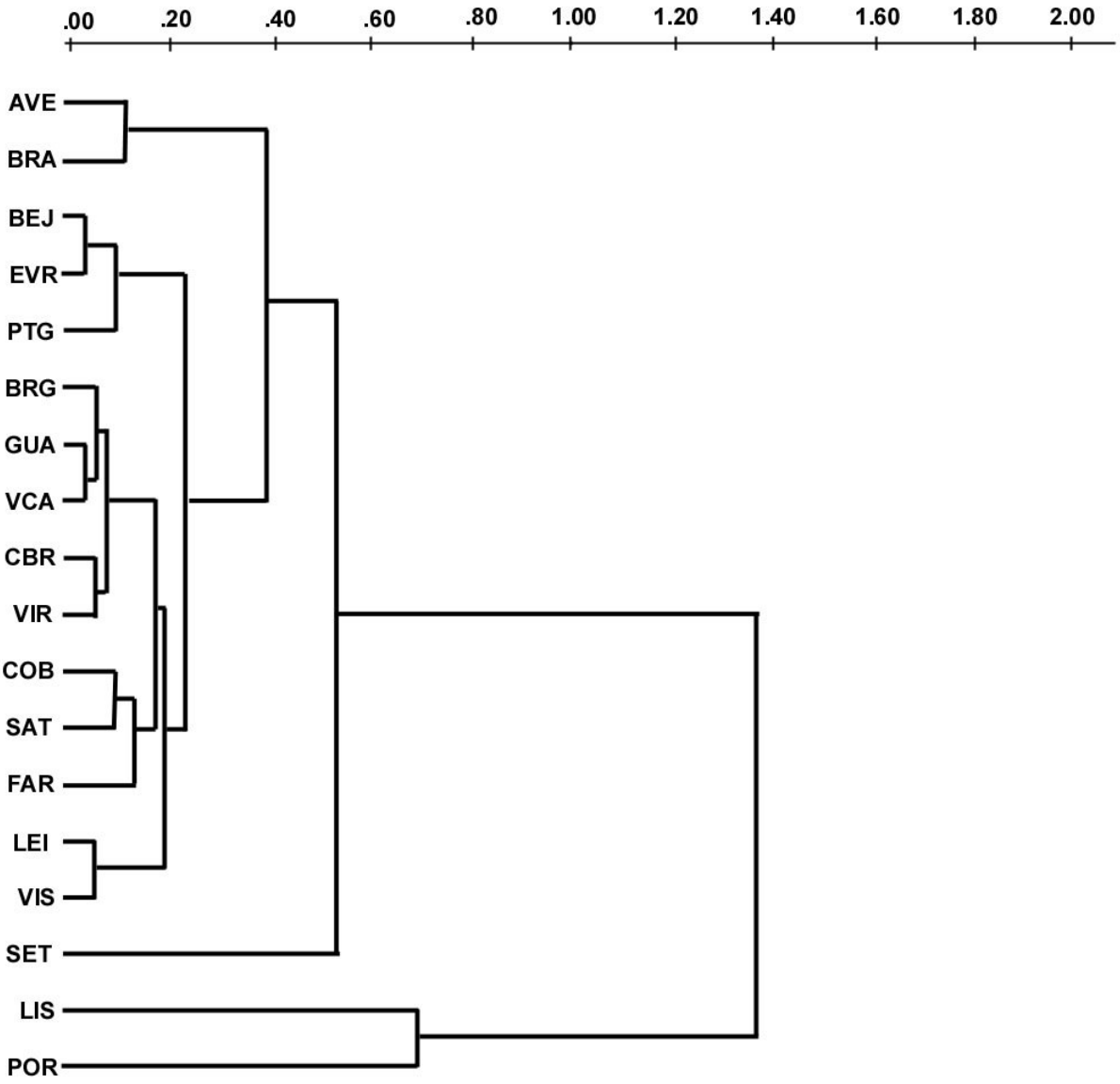


CLASSIFICAÇÃO AUTOMÁTICA

DISTÂNCIA EUCLIDIANA REDUZIDA

$$d(a,b) = \sqrt{\frac{1}{p}\sum_j \left[\frac{x(a,j) - m(a)}{s(a)} - \frac{x(b,j) - m(b)}{s(b)} \right]^2}$$

Dendrograma

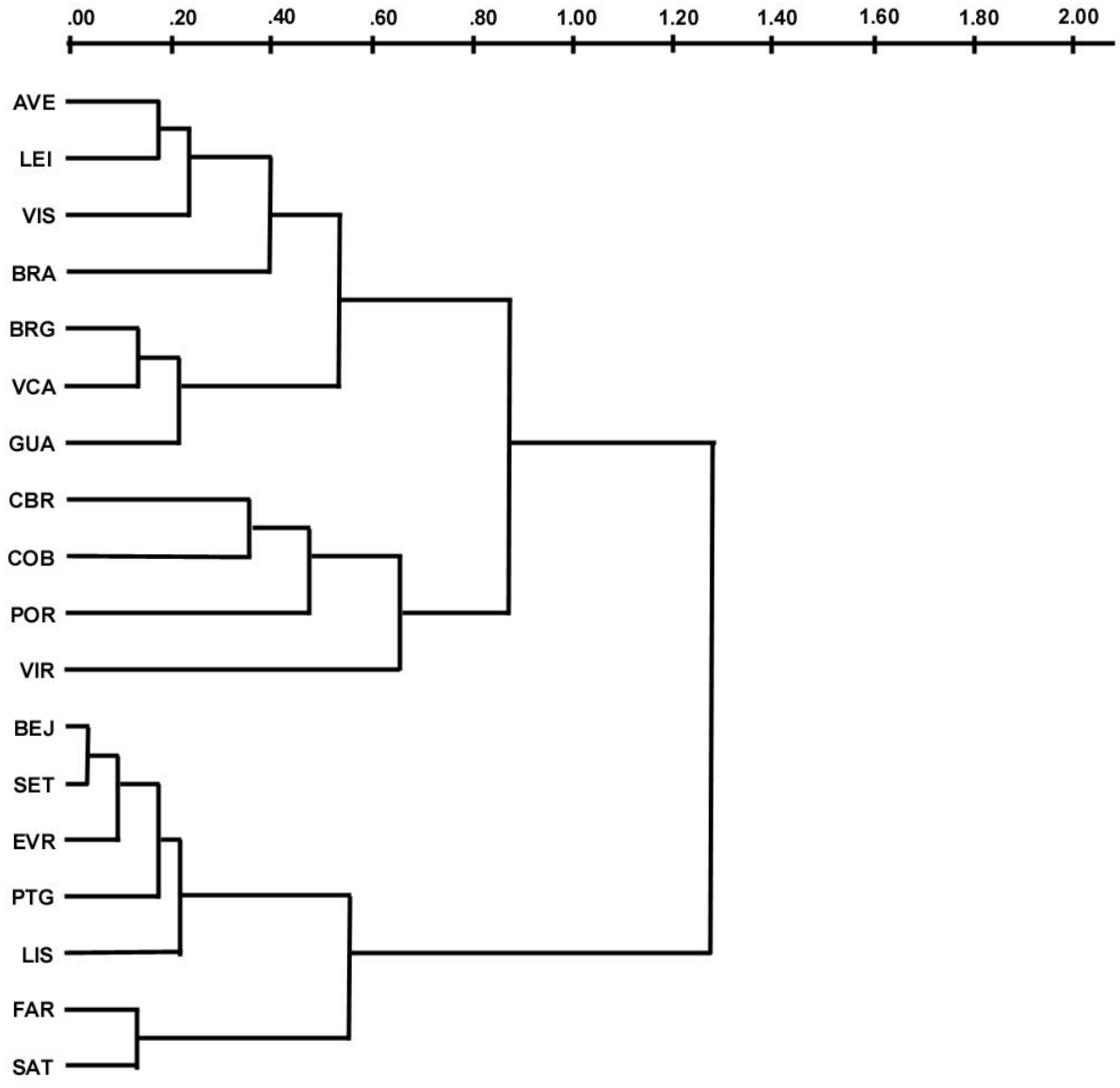


CLASSIFICAÇÃO AUTOMÁTICA

DISTÂNCIA COS θ

$$d(a,b) = \sqrt{1 - \sum_j \frac{x(a,j) - m(a)}{s(a)} \frac{x(b,j) - m(b)}{s(b)}}$$

Dendrograma

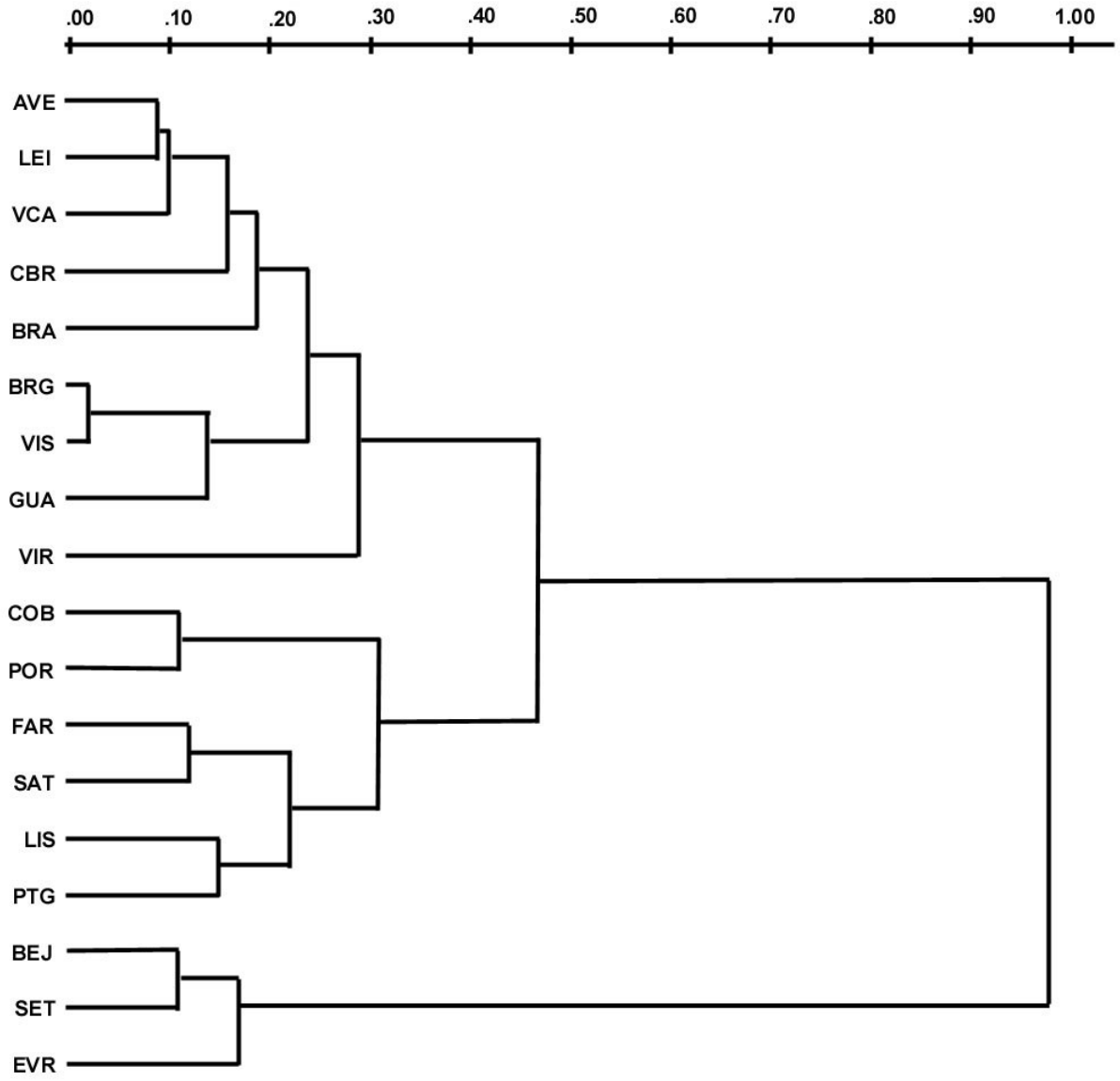


CLASSIFICAÇÃO AUTOMÁTICA

DISTÂNCIA DO χ^2

$$d(a,b) = \sqrt{\sum_j \frac{1}{x(j)} \left[\frac{x(a,j)}{x(a)} - \frac{x(b,j)}{x(b)} \right]^2}$$

Dendrograma



CLASSIFICAÇÃO AUTOMÁTICA

DISTÂNCIA DE GOWER

$$d(a,b) = \frac{1}{p}\sum_{j}\frac{|x(a,j) - x(b,j)|}{R(j)}$$

Dendrograma