ANÁLISE HIERÁRQUICA DE AGRUPAMENTOS PARA DISTRIBUIÇÕES MULTINOMIAIS

ANTONIO JOSÉ MANZATO

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA A OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

ÁREA DE CONCENTRAÇÃO: ESTATÍSTICA
ORIENTADOR: *Prof. Dr. Wilton de Oliveira Bussab*

- SÃO PAULO, OUTUBRO DE 1983 -

ÍNDICE

C A P Í T U L O I

INTRODUÇÃO

Na área das ciências observacionais, existe um crescente interesse em identificar grupos de objetos, que empiricamente melhor represente certas medidas de similaridade ou dissimilaridade. Frequentemente, grandes conjuntos de dados são coletados, mas a ausência de teorias compatíveis para analisá-los sugere a procura de estruturas naturais. Assim, Johnson (1967) diz então, que o problema é a descoberta da existência de uma estrutura natural inerente aos dados, isto é, um arranjo dos objetos em grupos homogêneos. Acontece porém que raramente o pesquisador dispõe de um critério teórico para definir estes grupos, tornando necessário então, o uso de técnicas numéricas para descobrir tal estrutura.

O método numérico que visa agrupar n objetos de uma amostra em grupos homogêneos internamente é mais conhecido como Análise de Conglomerados (Cluster Analysis). Existe uma confusão natural entre Classificação e Análise de Conglomerados, e Kendall e Stuart (1966) procuram caracterizar a diferença existente. Classificação tem por objetivo alocar novos objetos em classes já conhecidas. A informação disponível sobre o número de classes e as características de cada uma delas é que possibilita classificar cada um dos novos objetos em suas respectivas categorias. Já a análise de conglomerados tem por objetivo descobrir as classes e caracterizá-las. Pode-se assim dizer que a análise de conglomerados antecede a classificação.

Esta técnica se desenvolveu em áreas bastante diversas como: Biologia, Psicologia, Geologia, etc., o que explica a variedade de nomes pelos quais é chamada, tais como Tipologia, Ball (1971), Taxionomia Numérica [Sokal, 1963], etc.

Estaremos neste trabalho usando indistintamente os nomes Análise de Conglomerados e Análise de Agrupamento para a técnica em si, bem como as palavras Grupos ou Conglomerados.

Na Seção 1 apresentamos um resumo do desenvolvimento das técnicas de agrupamento.

Na Seção 2 tratamos de alguns problemas mais comuns em análise de conglomerados.

Na Seção 3 encontramos a proposta deste trabalho.

1.1 - DESENVOLVIMENTO DAS TÉCNICAS DE AGRUPAMENTO

Classificar pessoas em tipos é um passatempo antigo. Os hindus usaram sexo, características físicas e comportamentais para classificar pessoas em seis tipos, que eles designaram por nomes de animais. Gregos e romanos desenvolveram várias tipologias baseados em variações das características físicas do caule das plantas. No século XVIII, Linneu trabalhou com classificação no reino animal e vegetal (Genera Plantarum, primeira publicação 1737).

Segundo Everitt (1974), muitos dos conhecimentos reais que possuímos dependem do método pelo qual distinguimos o similar do dissimilar. Ao maior número de distinções naturais compreendida pelo método, mais clara se tem a idéia a respeito dos objetos. A medida que os objetos exigem mais nossa atenção, é necessário criarmos critérios mais rigorosos de distinção, por consequência mais difícil se torna o estabelecimento do método e cada vez mais ele se faz necessário.

A classificação de plantas e animais consistia mais em arte do que um método científico. Talvez a imprecisão dos métodos até então utilizados, tenha impelido os pes

quisadores a procurar técnicas mais objetivas e, os métodos numéricos foram surgindo naturalmente. A descrição inicial do que hoje é conhecida como análise de conglomerados foi formulada por Tryon (1939). Zubin (1938) e Thorndike (1953) tentaram usar os métodos de agrupamento em outras áreas além das ciências naturais, mas não obtiveram sucesso pelo fato do trabalho operacional ser muito grande. Somente a partir da última década com o advento dos computadores eletrônicos é que as técnicas numéricas de classificação tiveram seu uso difundido em outras áreas. Jardine e Sibson (1971) deram uma abordagem axiomática aos métodos relacionados com a taxionomia biológica.

Podemos então dizer que com os computadores, não só o uso das técnicas de análise de conglomerados foram difundidas, como também propiciaram o desenvolvimento de inúmeros algorítmos para agrupar objetos semelhantes.

Do ponto de vista estocástico, somente a partir de 1970 é que a análise de agrupamento vem recebendo uma abordagem mais rigorosa. Para isto se considera uma distribuição de probabilidade subjacente aos dados, envolvendo consequentemente conceitos de estimação de parâmetros, testes de significância, etc. Sob este ponto de vista, um conglomerado é definido em termos das características da função densidade da população da qual os dados foram amostrados, enquanto que nos procedimentos heurísticos as próprias observa

ções é que sugerem os conglomerados.

A análise de conglomerados cada vez mais se firma dentro da Estatística, principalmente como uma técnica descritiva de dados multivariados.

1.2 - ALGUNS PROBLEMAS COMUNS

O conceito geral de que um conglomerado é a reunião de objetos semelhantes e, o fato da análise de agrupamento ser uma técnica aplicável às diferentes áreas de pesquisa, levaram ao aparecimento de um número muito grande de técnicas para agrupar dados. A maioria delas consiste em definir medidas de similaridade ou dissimilaridade, que aqui denominamos medidas de proximidade, para então, de acordo com estas medidas, agrupar os objetos mais similares entre si.

O problema com essas técnicas, é a determinação da medida mais conveniente, uma vez que as técnicas baseadas em diferentes medidas de proximidade nem sempre levam aos mesmos resultados. A maioria das técnicas de agrupamento começam com o cálculo da matriz de proximidade entre os objetos observados, e sem dúvida, as técnicas de agrupamento podem ser vistas como forma de resumir as informações sobre o relacionamento entre os objetos que compõem a matriz de proximidade.

As dificuldades com as técnicas de agrupamento são claras:

- Quais medidas de proximidade serão usadas; desde que diferentes medidas podem levar-nos a diferentes resultados?

A escolha da medida adequada seria muito mais simples se tivéssemos conhecimento a priori da estrutura dos dados, o que em geral não temos.

- Um outro problema comum a todas as técnicas de agrupamento reside na dificuldade em se decidir sobre o número de grupos presente aos dados.

Com técnicas hierárquicas de agrupamento vários autores têm proposto diferentes regras de parada para a determinação do número de grupos.

Ball (1971) lista sete possíveis usos das técnicas de agrupamento aos interessados:

i) Encontrar uma verdadeira tipologia

ii) Ajustar modelos

iii) Predição baseado nos grupos

iv) Testar hipóteses

v) Exploração de dados

vi) Geração de hipóteses

vii) Redução de dados.

1.3 - OBJETIVOS

Pretendemos aqui apresentar algumas medidas de proximidade juntamente com as técnicas de análise de agrupamento, aplicados a dados que podem ser considerados como amostras de distribuições multinomiais com r categorias de resposta, com o intuito de fornecer ao leitor interessado um guia prático de aplicação de análise de conglomerados.

No Capítulo II apresentamos algumas medidas de proximidade, descrevendo suas principais propriedades.

No Capítulo III apresentamos as técnicas hierárquicas de agrupamento, se restringindo a três métodos de técnicas hierárquicas aglomerativas:

i) Método do encadeamento completo (Complete Linkage)

ii) Método do vizinho mais próximo (Single Linkage)

iii) Método do encadeamento médio (Average Linkage)

Para finalizar o capítulo apresentamos dois métodos de avaliação dos resultados obtidos pela aplicação dos métodos (i), (ii) e (iii), anteriormente citados.

No Capítulo IV abordamos a questão da determinação do número de grupos, onde são apresentadas duas possíveis regras de parada.

No Capítulo V comparamos através de simulações os desempenhos dos processos de agrupamento citados, das medidas de proximidade e das regras de parada sugeridas.

Em apêndice apresentamos:

Apêndice I - Resultados obtidos através de simulação, sendo estes representados em tabelas.

Apêndice II - Uma medida de similaridade sensível à contribuição das espécies raras, proposta por Grassle e Smith (1976).

Apêndice III - Alguns pacotes contendo programas prontos de análise de agrupamento.

C A P Í T U L O I I

ALGUMAS MEDIDAS DE PROXIMIDADE PARA COMPARAÇÃO ENTRE DOIS OBJETOS

2.1 - INTRODUÇÃO

Uma das maiores dificuldades encontrada pelo usuário das técnicas de agrupamento é o problema de se saber qual medida deva ser usada para comparação dos objetos em estudo. Ao recorrermos à bibliografia existente, encontramos um número variado de tais medidas, por exemplo veja Hartigan (1967).

Não vamos aqui fazer um trabalho exaustivo a respeito de tais medidas, mas sim apresentaremos algumas delas, por serem frequentemente utilizadas por pesquisadores de Biologia, Geologia, Ecologia, Química, Pesquisa de Mercado, etc.

Por exemplo, a distância Sanghvi (1968) com aplicações em genética, a medida de Morisita (1959) e Horn (1966) em estudos ecológicos, a medida de Kullback em problemas de energia, e a distância Euclidiana por ser uma medida cujas propriedades matemáticas são bem conhecidas dos pesquisadores, e por ser esta encontrada na maioria dos programas prontos de análise de agrupamento.

Para isso passemos à caracterização da nossa massa de dados. Consideremos s amostras extraídas de populações desconhecidas, cada uma de tamanho $n_1, n_2, \ldots, n_s$ respectivamente, com r categorias distintas de resposta. Estes resultados podem ser resumidos numa tabela de frequência cruzada $s x r$ conforme Tabela 2.1.

TABELA 2.1 - TABELA DE DADOS

	CATEGORIA DE RESPOSTA					
AMOSTRAS (OBJETOS)	1	2	..... j	.....	r	TOTAL
1	n_{11}	n_{12}	.... n_{1j}	....	n_{1r}	n_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	.... n_{ij}	....	n_{ir}	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
s	n_{s1}	n_{s2}	n_{sj}		n_{sr}	n_s

n_{ij} denota a frequência de resposta da *j-ésima* categoria para a *i-ésima* amostra.

Neste trabalho ao invés de usarmos a frequência

absoluta n_{ij}, usaremos a frequência relativa $p_{ij} = \frac{n_{ij}}{n_i}$ $i=$ $=1,\ldots,s;\ j=1,\ldots,r$ com $\sum_{j=1}^{r} p_{ij} = 1$.

2.2 - MEDIDA DE DISSIMILARIDADE

Sejam P e Q dois pontos, podendo os mesmos representarem medidas sobre dois objetos ou indivíduos. Uma função de valor real $d(P,Q)$ é uma função distância se possui as seguintes propriedades:

i) Simetria $d(P,Q) = d(Q,P)$;

ii) Não negatividade $d(P,Q) \geqq 0$ e

iii) $d(P,P) = 0$

para muitas funções distância as demais propriedades são asseguradas:

iv) $d(P,Q) = 0$ se e somente se $P=Q$

v) $d(P,Q) \leqq d(P,R) + d(R,Q)$ (desigualdade triangular)

se as cinco propriedades são válidas então d é chamada de métrica.

A razão do termo "dissimilaridade" apareceu em função de que a medida que $d(P,Q)$ cresce, diz-se que a divergência entre P e Q aumenta, ou seja, tornam-se cada vez mais dissimilares.

De acordo com Mardia, Kent e Bibby (1979); mesmo

que as propriedades (iv) e (v) não estejam satisfeitas $d(P,Q)$ é chamado de coeficiente de dissimilaridade entre P e Q.

2.2.1 - DISTÂNCIA MINKOWSKI (1911), DEFINIDA PELA NORMA ℓ^p

Para dois objetos i e j, define-se $d(i,j)$ como sendo:

$$d(i,j) = \left[\sum_{k=1}^{r} |p_{ik}-p_{jk}|^p\right]^{1/p}, \quad i,j=1,\ldots,s \qquad [2.1]$$

que é uma função distância se $1 \leq p < \infty$, onde p_{mk}, $m=i,j$; $k=1,\ldots,r$ representa a proporção da *m-ésima* amostra para a *k-ésima* categoria.

Quando $p=2$ temos a função distância mais comum de uso prático, conhecida como distância Euclidiana

$$d_1(i,j) = \left[\sum_{k=1}^{r} (p_{ik}-p_{jk})^2\right]^{1/2} \qquad [2.2]$$

Embora esta medida possa não ser a melhor, ou seja, mais adequada para a comparação de dois objetos, é porém a de mais fácil acesso em programas prontos de análise de agrupamento.

2.2.2 - DISTÂNCIA DE BHATTACHARYYA (1946)

A distância devido a Bhattacharyya é dada por

$$d_2(i,j) = \left[\sum_{k=1}^{r} (\sqrt{p_{ik}} - \sqrt{p_{jk}})^2 \right]^{1/2} \qquad [2.3]$$

A origem desta medida provém do fato que ao trabalharmos com proporções, a variância amostral depende da média, e com o intuito de se eliminar esta dependência, transformações nos dados originais são propostas. Neste caso a transformação sugerida é $Arc\ sen\sqrt{p_{ik}}$.

Podemos também procurar uma interpretação geométrica para esta medida:

$\theta_{ik} = arc\ sen\sqrt{p_{ik}} \iff p_{ik} = sen^2(\theta_{ik})$, sujeito a restrição de que

$\sum_{k=1}^{r} p_{ik} = 1$. Assim

$$\sum_{k=1}^{r} p_{ik} = \sum_{k=1}^{r} sen^2(\theta_{ik}) = 1;$$

o que significa dizer que as observações amostrais pertencem a uma esfera no espaço r-dimensional com centro na origem. A distância entre dois pontos i e j poderá então ser dada pela corda que conectará os dois pontos, ou algo diretamente proporcional a esta corda. Calculando-se a corda entre os dois pontos encontramos:

$$d(i,j) = (2-2cos(\psi_{ij}))^{1/2},$$

onde ψ_{ij} é o ângulo entre os vetores i e j com centro na origem.

Mas $cos(\psi_{ij})$ é dado por

$$cos(\psi_{ij}) = \sum_{k=1}^{r} sen(\psi_{ik})sen(\psi_{jk}) = \sum^{r} \sqrt{p_{ik}}\sqrt{p_{jk}}$$

Portanto:

$$d(i,j) = (2-2cos(\psi_{ij}))^{1/2} = (2-2\sum_{k=1}^{r} \sqrt{p_{ik}}\sqrt{p_{jk}})^{1/2}$$

$$= (\sum_{k=1}^{r} p_{ik} + \sum_{k=1}^{r} p_{jk} - 2\sum_{k=1}^{r} \sqrt{p_{ik}}\sqrt{p_{jk}})^{1/2}$$

$$= \left[\sum_{k=1}^{r} (\sqrt{p_{ik}} - \sqrt{p_{jk}})^2\right]^{1/2} = d_2(i,j)$$

a distância $d(i,j)=(2-2cos(\psi_{ij}))^{1/2}$ é conhecida como medida de Edward e Cavalli-Sforza (1964).

2.2.3 - DISTÂNCIA DE KULLBACK (1952)

Dadas duas populações π_i e π_j, caracterizadas pela distribuição multinomial, Kullback definiu a informação média para discriminação entre π_i e π_j através de π_i como sendo:

$$I(i:j) = \sum_{k=1}^{r} p_{ik} log \frac{p_{ik}}{p_{jk}}; \quad p_{jk} \neq 0; \quad 0 < p_{mk} < 1; \quad m=i,j.$$

Analogamente a informação média para discriminação entre π_i e π_j através de π_j como sendo:

$$I(j:i) = \sum_{k=1}^{r} p_{jk} \log \frac{p_{jk}}{p_{ik}}; \quad p_{ik} \neq 0; \quad 0 < p_{mk} < 1; \quad m = i, j.$$

A partir de $I(i:j)$ e $I(j:i)$, Kullback definiu como medida de divergência entre π_i e π_j, $d(i,j)$ dada por:

$$d(i,j) = I(i:j) + I(j:i)$$

$$= \sum_{k=1}^{r} (p_{ik} - p_{jk}) \log \frac{p_{ik}}{p_{jk}}; \quad p_{jk} \neq 0; \quad 0 < p_{mk} < 1; \quad m = i, j.$$

A medida $d(i,j)$ não satisfaz a (v) propriedade de distância, ou seja a desigualdade triangular.

Para medida de comparação entre dois objetos i e j baseado na teoria da informação, adotamos então a medida de divergência definida por Kullback

$$d_3(i,j) = \sum_{k=1}^{r} (p_{ik} - p_{jk}) \log \frac{p_{ik}}{p_{jk}}; \quad p_{jk} \neq 0; \quad 0 < p_{mk} < 1; \quad m = i, j \qquad [2.4]$$

OBSERVAÇÃO:

Em situações práticas poderá ocorrer casos onde $p_{jk} = 0$. Nestes casos, devemos optar por algum critério para atribuir um valor bem pequeno para p_{jk}, a fim de que possamos realizar os cálculos. A respeito da base do logarítmo, a maioria das vezes estaremos trabalhando com o logarítmo na base natural e.

2.2.4 - DISTÂNCIA DE SANGHVI E BALAKRISHNAN (1968)

O desenvolvimento de um índice para medir a distância entre duas populações com dados caracterizados por atributos, foi primeiramente tratado por Sanghvi [1952; 1953], que propôs uma medida análoga ao quiquadrado estatístico. Esta medida foi definida por:

$$d^2(i,j) = \frac{100}{\text{nº de graus de liberdade}} \sum_{k=1}^{r} \left[\frac{(p_{ik}-p_k)^2}{p_k} + \frac{(p_{jk}-p_k)^2}{p_k} \right]$$

onde $p_k = \frac{1}{2}(p_{ik}+p_{jk})$ e p_{ik}, p_{jk} são as proporções da *k-ésima* categoria.

Neste índice a multiplicação por um fator constante tal como 100, não é essencial. Desde que somente valores relativos são de interesse. A divisão pelo número de graus de liberdade foi proposta para situações onde algumas das populações não tinham respostas para todas as categorias.

Removendo-se os fatores não essenciais, um índice padrão pode ser dado por:

$$d_4(i,j) = 2 \sum_{k=1}^{r} \frac{(p_{ik}-p_{jk})^2}{p_{ik}+p_{jk}} \quad \text{onde} \quad p_{ik}+p_{jk} \neq 0; \quad 0 \leq p_{mk} \leq 1; \quad m=i,j \qquad [2.5]$$

OBSERVAÇÃO:

M. Hills (1967) observou que usando a aproximação:

$$\frac{1}{2}\log\frac{p_{ik}}{p_{jk}} \approx \frac{p_{ik}-p_{jk}}{p_{ik}+p_{jk}}; \quad 0<p_{mk}<1; \quad m=i,j$$

Teremos então que:

$$\sum_{k=1}^{r}(p_{ik}-p_{jk})\log\frac{p_{ik}}{p_{jk}} \approx \sum_{k=1}^{r}(p_{ik}-p_{jk})\frac{2(p_{ik}-p_{jk})}{p_{ik}+p_{jk}} = \sum_{k=1}^{r}\frac{2(p_{ik}-p_{jk})^2}{p_{ik}+p_{jk}}$$

portanto, [2.4] e [2.5] se equivalem em determinadas situações.

O erro de aproximação cresce a medida que p_{ik} e p_{jk} se distanciam um do outro, e a aproximação se torna melhor quando as proporções não se distanciam muito umas das outras, para as respectivas categorias de resposta. Para ilustrarmos a situação vejamos um exemplo.

EXEMPLO 2.1 - Sejam as amostras A_1, A_2 e A_3 caracterizadas pelas distribuições:

$A_1 = [0,26; \quad 0,44; \quad 0,30]$

$A_2 = [0,20; \quad 0,38; \quad 0,42]$

$A_3 = [0,05; \quad 0,62; \quad 0,33]$

Usando [2.4] e [2.5] para cálculo da matriz de distância, e usando a base natural para o logarítmo encontramos D e D' respectivamente

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \end{array}\begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{bmatrix} 0,00 & & \\ 0,06 & 0,00 & \\ 0,40 & 0,33 & 0,00 \end{bmatrix} \end{array} \qquad D' = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \end{array}\begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{bmatrix} 0,00 & & \\ 0,06 & 0,00 & \\ 0,89 & 0,29 & 0,00 \end{bmatrix} \end{array}$$

Para termos idéia do comportamento das medidas até aqui apresentadas consideremos um exemplo.

EXEMPLO 2.2 - Uma amostra de 1806 famílias de origem rural que se encontravam de passagem pelo Departamento de Imigração e Colonização do Estado de São Paulo no período de setembro de 1969 a agosto de 1970 foi utilizada.

Diversos foram os testes de laboratório feitos sobre o sangue coletado. Um deles foi a tipagem do grupo sanguíneo ABO, cuja distribuição fenotípica de ABO entre regiões é dada pela Tabela 2.2.

TABELA 2.2 - DISTRIBUIÇÃO FENOTÍPICA DE ABO POR REGIÕES

REGIÃO	O	A_1	A_2	B	A_1B	A_2B	TOTAL
1	685	377	119	167	41	9	1398
2	446	220	60	121	27	8	882
3	136	77	21	34	3	4	275
4	287	171	51	87	6	3	605
5	184	101	26	45	12	0	368
TOTAL	1738	946	277	454	89	24	3528

Região *1:* é compreendida por SP, RJ, ES e PR;

Região *2:* é compreendida por MG;

Região *3:* é compreendida por BA;

Região *4:* é compreendida por SE, AL e PE;

Região 5: é compreendida por CE, RN, PI, MA, PB e PA.

Fonte: Pedro Hernan Cabello Acero (1976).

Para o cálculo das distâncias entre as amostras R_1, R_2, R_3, R_4 e R_5, ao invés de usarmos a Tabela 2.2, vamos usar a tabela de proporções dada pela Tabela 2.3.

TABELA 2.3 - DISTRIBUIÇÃO DA PROPORÇÃO FENOTÍPICA ABO POR REGIÕES

REGIÃO	O	A_1	A_2	B	A_1B	A_2B
1	0,490	0,270	0,085	0,120	0,029	0,006
2	0,505	0,250	0,068	0,137	0,030	0,010
3	0,494	0,280	0,076	0,124	0,011	0,015
4	0,474	0,283	0,084	0,144	0,010	0,005
5	0,500	0,274	0,071	0,122	0,033	0,000

Usando [2.2], [2.3], [2.4] e [2.5] obtemos respectivamente as matrizes de distância D_1, D_2, D_3 e D_4.

$$D_1 = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,015 & 1,000 & & & \\ 0,025 & 0,040 & 1,000 & & \\ 0,037 & 0,053 & 0,031 & 1,000 & \\ 0,019 & 0,030 & 0,028 & 0,044 & 1,000 \end{bmatrix} \end{array}$$

$$D_2 = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,050 & 1,000 & & & \\ 0,084 & 0,079 & 1,000 & & \\ 0,079 & 0,093 & 0,066 & 1,0000 & \\ 0,082 & 0,105 & 0,144 & 0,115 & 1,000 \end{bmatrix} \end{array}$$

$$D_3 = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,010 & 1,000 & & & \\ 0,027 & 0,027 & 1,000 & & \\ 0,026 & 0,035 & 0,015 & 1,000 & \\ 0,027 & 0,050 & 0,099 & 0,054 & 1,000 \end{bmatrix} \end{array}$$

$$D_4 = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,010 & 1,000 & & & \\ 0,025 & 0,025 & 1,000 & & \\ 0,024 & 0,033 & 0,015 & 1,000 & \\ 0,015 & 0,024 & 0,052 & 0,042 & 1,000 \end{bmatrix} \end{array}$$

Ao tentarmos fazer comparações entre D_1, D_2, D_3 e D_4, simplesmente olhando para os números, teremos dificuldades em dizer quais das matrizes estariam produzindo resultado similar. Embora ainda não tenhamos visto as técnicas de

agrupamento, mencionadas no Capítulo III, estaremos aqui apresentando o dendrograma resultante do método de agrupamento, vizinho mais próximo, aplicado a cada uma das matrizes, para podermos visualizar graficamente o que cada medida produziu em termos do agrupamento. Os resultados da aplicação do método de agrupamento sobre D_1, D_2, D_3 e D_4 estão nas Figuras 1, 2, 3 e 4.

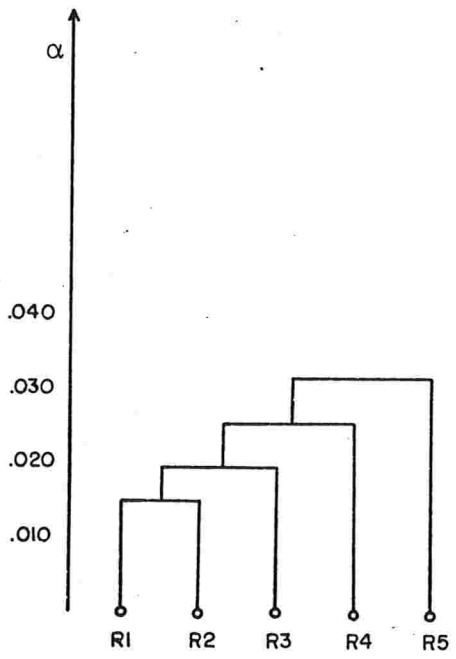


FIGURA 1 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO V.P. SOBRE D_1

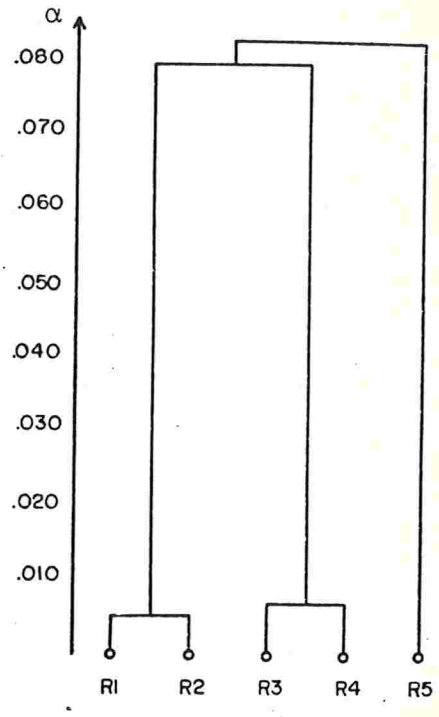


FIGURA 2 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO V.P. SOBRE D_2

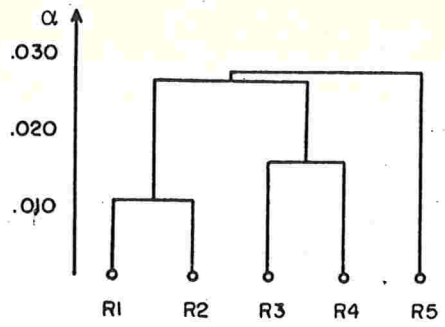


FIGURA 3 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO V.P. SOBRE D_3

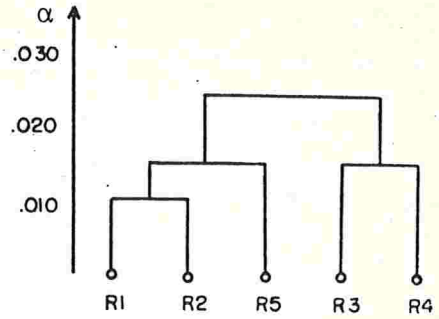


FIGURA 4 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO V.P. SOBRE D_4

Ao observarmos as Figuras 1, 2, 3 e 4 notamos que as medidas D_2 e D_3 produzem os mesmos resultados de agrupamento, enquanto que D_1 e D_4 não. Mas ao observarmos atentamente, verificamos que não são tão diferentes. As Figuras 2, 3 e 4 nos mostra que a distribuição fenotípica de SP, RJ, ES e PR é bastante parecida com a de MG, enquanto que a distribuição fenotípica de BA, SE, AL e PE também são bem similares. O resultado produzido por D_1 é diferente dos demais.

Conforme podemos observar, a adoção da medida pode levar a resultados bem distintos, e a escolha necessita o conhecimento e participação do pesquisador específico da área.

2.3 - MEDIDA DE SIMILARIDADE

Medidas de similaridade frequentemente são também chamadas de coeficiente de similaridade e às vezes não necessariamente estão definidas no intervalo [0,1].

Segundo Mardia, Kent e Bibby (1979), uma medida de similaridade entre dois objetos P e Q, denotado por $s(P,Q)$, deve satisfazer as seguintes propriedades:

i) $s(P,Q) = s(Q,P)$;

ii) $s(P,Q) \geq 0$;

iii) $s(P,Q)$ cresce a medida que a semelhança entre P e Q cresce.

Com esta definição, vemos que o comportamento da medida de similaridade e o da medida de dissimilaridade caminham em sentidos opostos, pois a similaridade cresce a medida em que os dois objetos em comparação são cada vez mais semelhantes, e a dissimilaridade diminui cada vez mais. Assim dada uma medida de similaridade $s(P,Q)$ entre P e Q, podemos facilmente derivar uma correspondente medida de dissimilaridade d tal como:

$$d(P,Q) = constante - s(P,Q).$$

Aqui também, como se pode notar, existe uma variedade muito grande de coeficientes de similaridade, mas iremos apresentar somente dois deles.

2.3.1 - COEFICIENTE DE SIMILARIDADE DE MORISITA

Consideremos duas populações, ambas com as mesmas r categorias de resposta. De cada uma delas, amostras características de tamanhos n_i e n_j respectivamente são consideradas, satisfazendo a condição:

$$\sum_{k=1}^{r} n_{ik} = n_i \; ; \qquad \sum_{k=1}^{r} n_{jk} = n_j$$

n_{ik} representa o número de observações da *k-ésima* categoria para a *i-ésima* população.

Ao selecionarmos aleatoriamente, dois elementos, um de cada uma das populações, a probabilidade que ambos pertençam a mesma categoria de resposta é estimada por:

$$\hat{p}_{ij} = \sum_{k=1}^{r} \frac{n_{ik} n_{jk}}{n_i n_j} \qquad [2.6]$$

Agora ao selecionarmos aleatoriamente, sem reposição dois elementos da população i, a probabilidade que ambos pertençam a mesma categoria de resposta é estimada por:

$$\lambda_i = \sum_{k=1}^{r} \frac{n_{ik}(n_{ik}-1)}{n_i(n_i-1)} \qquad [2.7]$$

analogamente se obtém λ_j.

[2.7] é conhecido como Índice de Simpson (1949). Morisita (1959b) para medida de similaridade entre duas amostras definiu:

$$s_1(i,j) = \frac{\hat{p}_{ij}}{(\lambda_i+\lambda_j)/2} \qquad [2.8]$$

$$s_1(i,j) = \frac{2\sum_{k=1}^{r} n_{ik}n_{jk}}{(\lambda_i+\lambda_j)n_i n_j}$$

$s_1(i,j)$ dada por [2.8] varia no intervalo $[0;1]$. Assume valor zero quando as amostras não apresentam categoria de resposta comum, isto é, observamos uma frequência n_{ik} não nula para a *k-ésima* categoria da *i-ésima* amostra e observamos frequência nula para *k-ésima* categoria da *j-ésima* amostra, ou observamos frequência nula para ambas as amostras i e j na *k-ésima* categoria. Consideremos um exemplo.

EXEMPLO 2.3 - Sejam A_1, A_2, A_3, A_4 e A_5 amostras caracterizadas pela seguinte distribuição de frequência:

$A_1 = [\ 0;\quad 0;\quad 0;\quad 3;\quad 97]$

$A_2 = [44;\quad 10;\quad 1;\quad 12;\quad 33]$

$A_3 = [\ 0;\quad 9;\quad 85;\quad 6;\quad 0]$

$A_4 = [20;\quad 19;\quad 26;\quad 18;\quad 17]$

$A_5 = [98;\quad 2;\quad 0;\quad 0;\quad 0]$

Usando [2.8] para cálculo da similaridade entre as amostras obtemos a matriz de similaridade S_M dada por:

$$S_M = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \end{array} \begin{array}{c} \begin{array}{ccccc} A_1 & A_2 & A_3 & A_4 & A_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,513 & 1,000 & & & \\ 0,002 & 0,047 & 1,000 & & \\ 0,299 & 0,724 & 0,536 & 1,000 & \\ 0,000 & 0,676 & 0,002 & 0,345 & 1,000 \end{bmatrix} \end{array}$$

Quando os n_{ik} são grandes podemos usar um esquema amostral com reposição. Dessa forma λ_i será dado por:

$$\lambda_i = \sum_{k=1}^{r} \frac{n_{ik}^2}{n_i^2} \quad \text{e,}$$

$$s_1(i,j) = \frac{2 \sum_{k=1}^{r} n_{ik} n_{jk}}{\left[\sum_{k=1}^{r} \frac{n_{ik}^2}{n_i^2} + \sum_{k=1}^{r} \frac{n_{jk}^2}{n_j^2} \; n_i n_j \right]} \qquad [2.9]$$

Se $n_i = n_j$ então $s_1(i,j)$ é dada por:

$$s_1(i,j) = \frac{2 \sum_{k=1}^{r} n_{ik} n_{jk}}{\sum_{k=1}^{r} n_{ik}^2 + \sum_{k=1}^{r} n_{jk}^2} \qquad [2.10]$$

Se ao invés de observarmos as frequências n_{ik} para as categorias de resposta, tivermos as proporções p_{ik} então $s_1(i,j)$ será dada por:

$$s_1(i,j) = \frac{2\sum_{k=1}^{r} p_{ik}p_{jk}}{\sum_{k=1}^{r} p_{ik}^2 + \sum_{k=1}^{r} p_{jk}^2} \qquad [2.11]$$

onde $\lambda_i = \sum_{k=1}^{r} p_{ik}^2$ e $\sum_{k=1}^{r} p_{ik} = 1$.

Usando [2.11] para os dados do Exemplo 2.3 encontramos uma nova matriz de similaridade dada por S'_M.

$$S'_M = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \end{array} \begin{array}{c} \begin{array}{ccccc} A_1 & A_2 & A_3 & A_4 & A_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,510 & 1,000 & & & \\ 0,002 & 0,046 & 1,000 & & \\ 0,297 & 0,704 & 0,530 & 1,000 & \\ 0,000 & 0,672 & 0,002 & 0,342 & 1,000 \end{bmatrix} \end{array}$$

2.3.2 - COEFICIENTE DE SIMILARIDADE DE HORN

Henry S. Horn (1966) usando medida de informação de Shannon-Wiener, descreve um coeficiente de similaridade para duas amostras i e j.

Para a *i-ésima* amostra a medida de informação dada por Shannon-Wiener é conhecida como entropia e é dada por:

$$H(i) = -\sum_{k=1}^{r} \frac{n_{ik}}{n_i} \log\left[\frac{n_{ik}}{n_i}\right] \qquad [2.12]$$

A informação conjunta para as amostras i e j é definida como:

$$H(i+j) = -\sum_{k=1}^{r}\left[\frac{n_{ik}+n_{jk}}{n_i+n_j}\right]\log\left[\frac{n_{ik}+n_{jk}}{n_i+n_j}\right] \qquad [2.13]$$

RESULTADOS

1) Se as amostras i e j possuirem as mesmas proporções para as respectivas *k-ésimas* categorias de resposta, então

$H(i) = H(j)$ e $H(i+j) = H(j)$

De fato, se:

$$\frac{n_{ik}}{n_i} = \frac{n_{jk}}{n_j} \iff n_{ik}n_j = n_{jk}n_i$$

$$H(i+j) = -\sum_{k=1}^{r}\left[\frac{n_{ik}+n_{jk}}{n_i+n_j}\right]\log\left[\frac{n_{ik}+n_{jk}}{n_i+n_j}\right] =$$

$$= -\sum_{k=1}^{r}\left[\frac{n_i n_{jk}+n_j n_{jk}}{(n_i+n_j)n_j}\right]\log\left[\frac{n_i n_{jk}+n_j n_{jk}}{(n_i+n_j)n_j}\right]$$

$$= -\sum_{k=1}^{r}\frac{n_{jk}}{n_j}\log\left[\frac{n_{jk}}{n_j}\right] = H(j).$$

2) $H(i+j)$ será máximo quando as amostras i e j possuirem observações em categorias totalmente distintas. Isto é, ao observarmos n_{ik} elementos na *k-ésima* categoria de resposta da *i-ésima* amostra, não observamos elemento algum para a *k-ésima* categoria de resposta da *j-ésima* amostra.

De fato: sejam i^* e j^* duas amostras de tamanhos n_i e n_j respectivamente, com as características anteriormente enunciadas, a menos da *s-ésima* categoria de resposta onde ocorrem observações para as duas amostras.

Então $H(i^*+j^*)$ é dado por:

$$H(i^*+j^*) = \sum_{\substack{k=1 \\ k\neq s}}^{r} \frac{n_{ik}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{ik}}\right] + \sum_{\substack{k=2 \\ k\neq s}}^{r} \frac{n_{jk}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{jk}}\right] +$$

$$+ \frac{n_{is}+n_{js}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{is}+n_{js}}\right]$$

Para todo i e j, amostras com as características anteriormente enunciadas, temos:

$$H(i+j) = \sum_{k=1}^{r} \frac{n_{ik}+n_{jk}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{ik}+n_{jk}}\right]$$

$$= \sum_{k=1}^{r} \frac{n_{ik}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{ik}}\right] + \sum_{\substack{m=2 \\ m\neq k}}^{r} \frac{n_{jm}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{jm}}\right]$$

Vamos supor que a *s-ésima* categoria de resposta pertença ao primeiro somatório de $H(i+j)$. Então o segundo somatório de $H(i^*+j^*)$ e de $H(i+j)$ são iguais, e o primeiro somatório de $H(i^*+j^*)$ e $H(i+j)$ diferem somente pela *s-ésima* categoria de resposta. Mas,

$$\frac{n_{is}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{is}}\right] \geq \frac{n_{is}+n_{js}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{is}+n_{js}}\right]$$

então

$H(i+j) \geqq H(i^*+j^*)$.

De forma análoga, se tivermos duas categorias de resposta, s e p, com observações em ambas as amostras i^* e j^*, $H(i+j) \geqq H(i^*+j^*)$, $\forall$ i e j amostras satisfazendo a condição anteriormente enunciada.

Assim quando tivermos observações em todas as categorias de resposta para ambas as amostras i^* e j^*, $H(i+j) \geqq H(i^*+j^*)$, $\forall$ i e j amostras satisfazendo a condição anteriormente enunciada.

Dessa forma $H(i+j)$ será máximo. Portanto,

$$H_{max}(i+j) = \sum_{k=1}^{r} \frac{n_{ik}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{ik}}\right] + \frac{n_{jk}}{n_i+n_j} \log\left[\frac{n_i+n_j}{n_{jk}}\right] \qquad [2.14]$$

3) Em geral o que ocorre na prática com maior frequência não são as situações apresentadas pelos resultados 1 e 2 anteriormente citados, mas sim situações em que diferentes proporções ocorrem para as amostras i e j, nas respectivas *k-ésimas* categorias de resposta. Em geral $H(i) \neq H(j)$, e para $H(i+j)$ usamos como estimativa o seu valor mínimo dado por:

$$H_{min}(i+j) = \frac{n_i}{n_i+n_j} H(i) + \frac{n_j}{n_i+n_j} H(j) \qquad [2.15]$$

Com os resultados obtidos Horn apresenta um índice de similaridade entre as amostras i e j, que é dado por:

$$s_2(i,j) = \frac{H_{max}(i+j)-H_{obs}(i+j)}{H_{max}(i+j)-H_{min}(i+j)} \qquad [2.16]$$

$H_{obs}(i+j)$ é a informação conjunta observada para as amostras i e j. Desenvolvendo [2.16] encontramos:

$$s_2(i,j) = \frac{\sum_{k=1}^{r}(n_{ik}+n_{jk})\log(n_{ik}+n_{jk}) - \sum_{k=1}^{r} n_{ik}\log(n_{ik}) - \sum_{k=1}^{r} n_{jk}\log(n_{jk})}{(n_i+n_j)\log(n_i+n_j) - n_i\log(n_i) - n_j\log(n_j)} \qquad [2.17]$$

Se ao invés de trabalharmos com as frequências n_{ik}, n_{jk} estivermos trabalhando com as proporções p_{ik}, p_{jk}, então $s_2(i,j)$ é dado por:

$$s_2(i,j) = \frac{\sum_{k=1}^{r}(p_{ik}+p_{jk})\log(p_{ik}+p_{jk}) - \sum_{k=1}^{r} p_{ik}\log(p_{ik}) - \sum_{k=1}^{r} p_{jk}\log(p_{jk})}{2\log 2} \qquad [2.18]$$

Para obtenção de [2.18] consideramos:

i) $H(i) = -\sum_{k=1}^{r} \frac{n_{ik}}{n_i}\log\left(\frac{n_{ik}}{n_i}\right) = -\sum_{k=1}^{r} p_{ik}\log(p_{ik})$

ii) $H_{max}(i+j) = -\frac{1}{2}\sum_{k=1}^{r}\left[p_{ik}\log\left(\frac{p_{ik}}{2}\right)+p_{jk}\log\left(\frac{p_{jk}}{2}\right)\right]$

iii) $H_{min}(i+j) = -\frac{1}{2}\sum_{k=1}^{r}\left[p_{ik}\log(p_{ik})+p_{jk}\log(p_{jk})\right]$

iv) $H_{obs}(i+j) = -\frac{1}{2}\sum_{k=1}^{r}(p_{ik}+p_{jk})\log\left(\frac{p_{ik}+p_{jk}}{2}\right)$

Substituindo-se (ii), (iii) e (iv) em [2.16] obtemos [2.18].

$s_2(i,j)$ assume valores no intervalo $[0;1]$. $s_2(i,j)$ assume valor zero quando as amostras são completamente distintas, conforme resultado 2 dado anteriormente. $s_2(i,j)$ assume valor 1 quando as amostras possuem as mesmas proporções, isto é, quando $p_{ik}=p_{jk}$ em ambas as amostras.

EXEMPLO 2.4 - Consideremos aqui novamente para exemplo os dados do Exemplo 2.2, onde serão aplicados os coeficientes de similaridade dados por Morisita e por Horn.

Usando [2.11] e [2.18] com logarítmo na base natural e encontramos S_M e S_H dadas por Morisita e Horn respectivamente, quando aplicadas à tabela de dados 2.3 de distribuição da proporção fenotípica *ABO* por regiões.

$$S_M = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,998 & 1,000 & & & \\ 0,999 & 0,991 & 1,000 & & \\ 0,998 & 0,996 & 0,998 & 1,000 & \\ 0,999 & 0,998 & 0,997 & 0,997 & 1,000 \end{bmatrix} \end{array}$$

$$S_H = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} 1,000 & & & & \\ 0,998 & 1,000 & & & \\ 0,995 & 0,995 & 1,000 & & \\ 0,995 & 0,994 & 0,997 & 1,000 & \\ 0,997 & 0,995 & 0,989 & 0,992 & 1,000 \end{bmatrix} \end{array}$$

De forma análoga ao que fizemos no Exemplo 2.2, consideremos os dendrogramas, resultantes da aplicação do método de agrupamento do vizinho mais próximo sobre S_M e S_H, dados pelas Figuras 5 e 6.

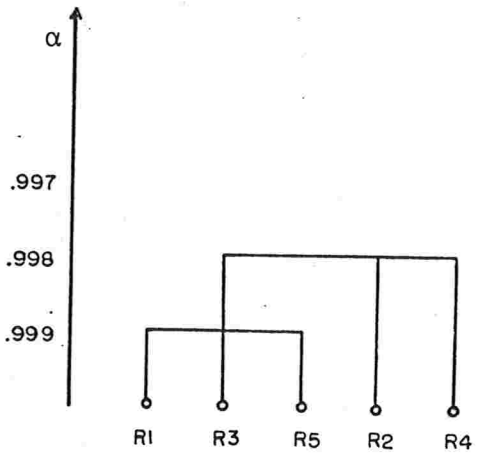


FIGURA 5 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO V.P. SOBRE S_M

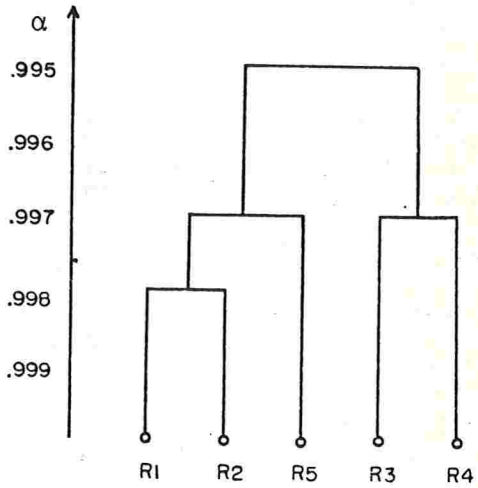


FIGURA 6 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO V.P. SOBRE S_H

2.4 - COMENTÁRIOS

Conforme já foi dito, um dos problemas sério encontrado com análise de agrupamento é o da escolha da medida apropriada a se usar. O exemplo colocado ilustra muito bem esta situação, onde vemos que a maioria das medidas usadas produziram resultados semelhantes, exceto a medida Euclidiana e a similaridade de Morisita. A situação então é a de

que a decisão não ficará somente a cargo do estatístico, mas sim deve existir uma conjunção muito grande entre o pesquisador e o estatístico, para que ao varrer uma lista de possíveis medidas, escolha a de melhor adequação ao problema proposto. Por exemplo, ao observarmos o coeficiente de Morisita verificamos que é uma medida altamente dependente das categorias que apresentam frequência de ocorrência alta, isto é, a probabilidade de ocorrer uma dada categoria, é alta em relação às demais. Isto ocorre porque no desenvolvimento deste coeficiente foi usado o índice de Simpson cujo interesse está voltado para a frequência de ocorrência da categoria de resposta. Já o coeficiente de Horn, esta dependência passa a ser menor, porque para este coeficiente o importante é a ocorrência da categoria de resposta.

Dessa forma então o pesquisador fará a opção por alguma das medidas, dependendo é claro, do interesse que ele tenha para definir o que é proximidade entre objetos.

Algumas das medidas propostas podem ser estendidas como é o caso da distância Sanghvi, a distância Minkowski, com aplicações em Genética, veja C. Radhakrishna Rao (Ryzin, J.V. Classification an Clustering, 1977); a similaridade de Morisita que é apresentada como um caso particular da medida de Grassle e Smith (1976) encontrada no Apêndice II, e também a medida de similaridade de Horn, que tem sua extensão natural para a comparação de três objetos ou mais.

C A P Í T U L O I I I

TÉCNICAS DE AGRUPAMENTO

3.1 - INTRODUÇÃO

A classificação de objetos, elementos ou indivíduos em grupos homogêneos, tem-se destacado consideravelmente com o advento do computador eletrônico, sendo assim possível o desenvolvimento de algorítmos para agrupar, através de critérios de otimização, objetos, que até então era considerado um trabalho quase que impossível, devido as grandes dificuldades de cálculo. Principalmente a partir de 1970, várias foram as técnicas de agrupamento propostas; vejam Everitt (1974), Hartigan (1975), onde se desejava descobrir a existência de um arranjo natural dos elementos componentes da massa de dados em estudo, em grupos homogêneos, passível de

uma interpretação condizente com fenômenos observáveis na natureza.

Neste capítulo abordaremos somente três métodos de agrupamento, dentro de técnicas hierárquicas aglomerativas, que são:

i) Método do vizinho mais próximo;

ii) Método do encadeamento completo;

iii) Método do encadeamento médio.

A escolha dos três métodos, deve-se ao fato de serem estes, algorítmos encontrados em programas prontos, como é o caso do BMDP (Biomedical Computer Programs), e também os dois primeiros métodos por gozarem da propriedade de invariância sob transformações lineares nos dados (Johnson, 1967). Ainda neste capítulo apresentamos dois modos de avaliar o resultado do agrupamento, através do dendrograma resultante da aplicação de cada método de agrupamento sobre a matriz tomada como entrada de dados.

Gostaríamos também de chamar a atenção do leitor, que este capítulo servirá como complementação e ilustração da técnica de agrupar, uma vez que ao pesquisador em técnicas de agrupamento, isto seria considerado como pré-requisito obrigatório; onde maiores detalhes se encontram em Everitt (1974) e Hartigan (1975).

3.2 - TÉCNICAS HIERÁRQUICAS AGLOMERATIVAS

Ao trabalhar-se com técnicas hierárquicas aglomerativas, o procedimento básico é o mesmo. Inicia-se pelo cálculo da matriz de proximidade entre os objetos, que será usada como entrada de dados; e sobre esta aplicam-se os métodos de agrupamento. O modelo hierárquico aglomerativo pode ser descrito como sendo aquele em que começando de uma partição inicial de N grupos (cada objeto é considerado um grupo), a cada estágio reduzimos de um o número deles, até que tenhamos todos os N objetos reunidos num único grupo. Assim o resultado de um processo hierárquico aglomerativo pode ser descrito da seguinte maneira:

Partindo-se inicialmente de N grupos, representados pelos inteiros $1,2,\ldots,N$; no primeiro estágio $N-1$ grupos são formados; onde para isto se considera todas as $d(i, j)$, para os pares (i,j), $i \neq j$ na matriz de proximidade de ordem $N \times N$ e, escolhe-se para a primeira fusão os objetos i,j que otimizam o critério adotado. No segundo estágio $N-2$ grupos são formados de maneira análoga ao procedimento anterior, e assim segue sucessivamente até o último estágio onde os objetos estarão reunidos num único grupo.

Dessa forma, os modelos hierárquicos aglomerativos podem ser caracterizados por um conjunto de partições $P_0, P_1, \ldots, P_{N-1}$ e seus correspondentes valores do critério de

aglomeração, $\alpha_0, \alpha_1, \ldots, \alpha_{N-1}$; onde os estágios subscritados $0, 1, \ldots, N-1$ representam respectivamente os $N, N-1, \ldots, 1$ grupos. Então para a partição P_j existe uma configuração de grupos associada, que é representada por $C_1, C_2, \ldots, C_{N-j}$. A representação gráfica destas configurações e seus respectivos níveis α, dá-se o nome de dendrograma.

A diferença entre os métodos hierárquicos irá depender apenas do critério adotado α, que também é denominado nível de agrupamento:

a) Se numa determinada partição P_j, altos valores do critério implicar em dissimilaridade dos grupos, então o nível de agrupamento α_j é definido pelo menor valor do critério naquela partição, isto é:

$$\alpha_j = \underset{i<m}{mínimo}[d(i,m)]; \quad i,m=1,\ldots,N-j \qquad [3.1]$$

onde $d(i,m)$ é a distância entre os grupos i e m da partição P_j.

b) Se numa determinada partição P_j, altos valores do critério implicar em similaridade dos grupos, então o nível de agrupamento α_j é definido pelo maior valor do critério naquela partição, isto é:

$$\alpha_j = \underset{i<m}{máximo}[d(i,m)]; \quad i,m=1,\ldots,N-j \qquad [3.2]$$

onde $d(i,m)$ é a distância entre os grupos i e m da partição P_j.

3.3 - MÉTODO DE AGRUPAMENTO DO VIZINHO MAIS PRÓXIMO

Este método teve início com Sneath (1957) e Johnson (1967). Aqui partimos de um conjunto com N grupos; cada objeto é considerado como um grupo, e a partir da matriz de proximidade originada pelos N objetos serão fundidos no primeiro estágio os dois grupos mais próximos. A cada fusão decresce de um o número de grupos, e para distância entre eles, definimos como sendo a menor distância entre todas aquelas calculadas entre os componentes dos dois grupos. Para entendimento do processo consideremos o exemplo a seguir.

EXEMPLO 3.1 - No plano cartesiano (X,Y) consideremos seis pontos $A=(-1;2)$, $B=(0;-0,5)$, $C=(1;0)$, $D=(2;0)$, $E=(3;2)$ e $F=(4;0)$ (veja Figura 3.4a). Usando a distância Euclidiana ao quadrado para cálculo da distância entre os pontos, encontramos D_1 dada por:

$$D_1 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & & & \\ 7,25 & 0,00 & & & & \\ 8,00 & 1,25 & 0,00 & & & \\ 13,00 & 4,25 & 1,00 & 0,00 & & \\ 16,00 & 15,25 & 8,00 & 5,00 & 0,00 & \\ 29,00 & 16,25 & 9,00 & 4,00 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

$d(i,j)$ representa a distância entre os pontos da *i-ésima* linha e *j-ésima* coluna. A menor distância encontrada em D_1 ocorre entre os pontos C e D. Desta forma o primeiro grupo se dá com a fusão dos pontos C e D, originando uma nova matriz D_2 onde a distância dos outros pontos a CD é definida por:

$$d(CD,A) = min\{d(C,A);\ d(D,A)\} = d(C,A) = 8,00$$

$$d(CD,B) = min\{d(C,B);\ d(D,B)\} = d(C,B) = 1,25$$

$$d(CD,E) = min\{d(C,E);\ d(D,E)\} = d(D,E) = 5,00$$

$$d(CD,F) = min\{d(C,F);\ d(D,F)\} = d(D,F) = 4,00$$

$$D_2 = \begin{array}{c} \\ A \\ B \\ CD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & CD & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & & \\ 7,25 & 0,00 & & & \\ 8,00 & 1,25 & 0,00 & & \\ 16,00 & 15,25 & 5,00 & 0,00 & \\ 29,00 & 16,25 & 4,00 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

Com procedimento análogo ao anterior obteremos D_3, D_4 e D_5 dadas a seguir:

$$D_3 = \begin{array}{c} \\ A \\ BCD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccc} A & BCD & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & \\ 7,25 & 0,00 & & \\ 16,00 & 5,00 & 0,00 & \\ 29,00 & 4,00 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

$$D_4 = \begin{array}{c} \\ A \\ BCDF \\ E \end{array} \begin{array}{c} \begin{array}{ccc} A & BCDF & E \end{array} \\ \begin{bmatrix} 0,00 & & \\ 7,25 & 0,00 & \\ 16,00 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

$$D_5 = \begin{array}{c} \\ A \\ BCDEF \end{array} \begin{array}{c} \begin{array}{cc} A & BCDEF \end{array} \\ \begin{bmatrix} 0,00 & \\ 7,25 & 0,00 \end{bmatrix} \end{array}$$

A representação gráfica das sucessivas fusões no processo de agrupamento tem como resultado um esquema em árvore denominado dendrograma, que é dado pela Figura 3.1.

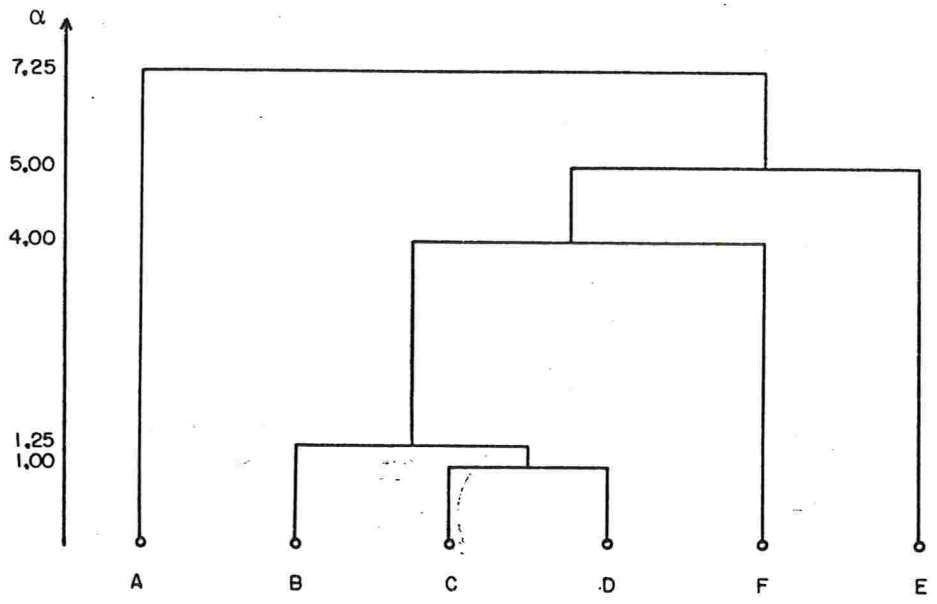


FIGURA 3.1 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO DE AGRUPAMENTO V.P. SOBRE D_1.

3.4 - MÉTODO DE AGRUPAMENTO DO ENCADEAMENTO COMPLETO

Aqui também o precursor deste método foi Johnson (1967). Este método é oposto ao método do vizinho mais próximo, uma vez que a distância entre grupos é aqui definida como a máxima distância entre os elementos componentes dos grupos. Consideremos novamente a matriz D_1 do Exemplo 3.1, para construção do Exemplo 3.2.

EXEMPLO 3.2 - Dada D_1, para obtermos a primeira fusão obedecemos a mesma regra do método anterior. Isto é a menor distância em D_1 ocorre entre C e D. Agora a distância de CD aos demais grupos é definida por:

$$d(CD,A) = máximo\{d(C,A);\ d(D,A)\} = d(D,A) = 13,00$$

$$d(CD,B) = máximo\{d(C,B);\ d(D,B)\} = d(D,B) = 4,25$$

$$d(CD,E) = máximo\{d(C,E);\ d(D,E)\} = d(C,E) = 8,00$$

$$d(CD,F) = máximo\{d(C,F);\ d(D,F)\} = d(C,F) = 9,00$$

Assim então obtemos D_2^* cuja dimensão diminuiu de 1 e é dada por:

$$D_2^* = \begin{array}{c} \\ A \\ B \\ CD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & CD & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & & \\ 7,25 & 0,00 & & & \\ 13,00 & 4,25 & 0,00 & & \\ 16,00 & 15,25 & 8,00 & 0,00 & \\ 29,00 & 16,25 & 9,00 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

Com procedimento análogo obtemos as matrizes D_3^*, D_4^* e D_5^*:

$$D_3^* = \begin{array}{c} \\ A \\ BCD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccc} A & BCD & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & \\ 13,00 & 0,00 & & \\ 16,00 & 15,25 & 0,00 & \\ 29,00 & 16,25 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

$$D_4^* = \begin{array}{c} \\ A \\ BCD \\ EF \end{array} \begin{array}{c} \begin{array}{ccc} A & BCD & EF \end{array} \\ \begin{bmatrix} 0,00 & & \\ 13,00 & 0,00 & \\ 29,00 & 16,25 & 0,00 \end{bmatrix} \end{array}$$

$$D_5^* = \begin{array}{c} \\ ABCD \\ EF \end{array} \begin{array}{c} \begin{array}{cc} ABCD & EF \end{array} \\ \begin{bmatrix} 0,00 & \\ 29,00 & 0,00 \end{bmatrix} \end{array}$$

A representação gráfica das sucessivas fusões é apresentada na Figura 3.2.

Através das Figuras 3.1 e 3.2 observamos que os dois métodos de agrupamento apesar de não provocarem alterações na matriz de dados; como é o caso do método do encadeamento médio que veremos adiante; podem produzir resultados de agrupamentos diferentes.

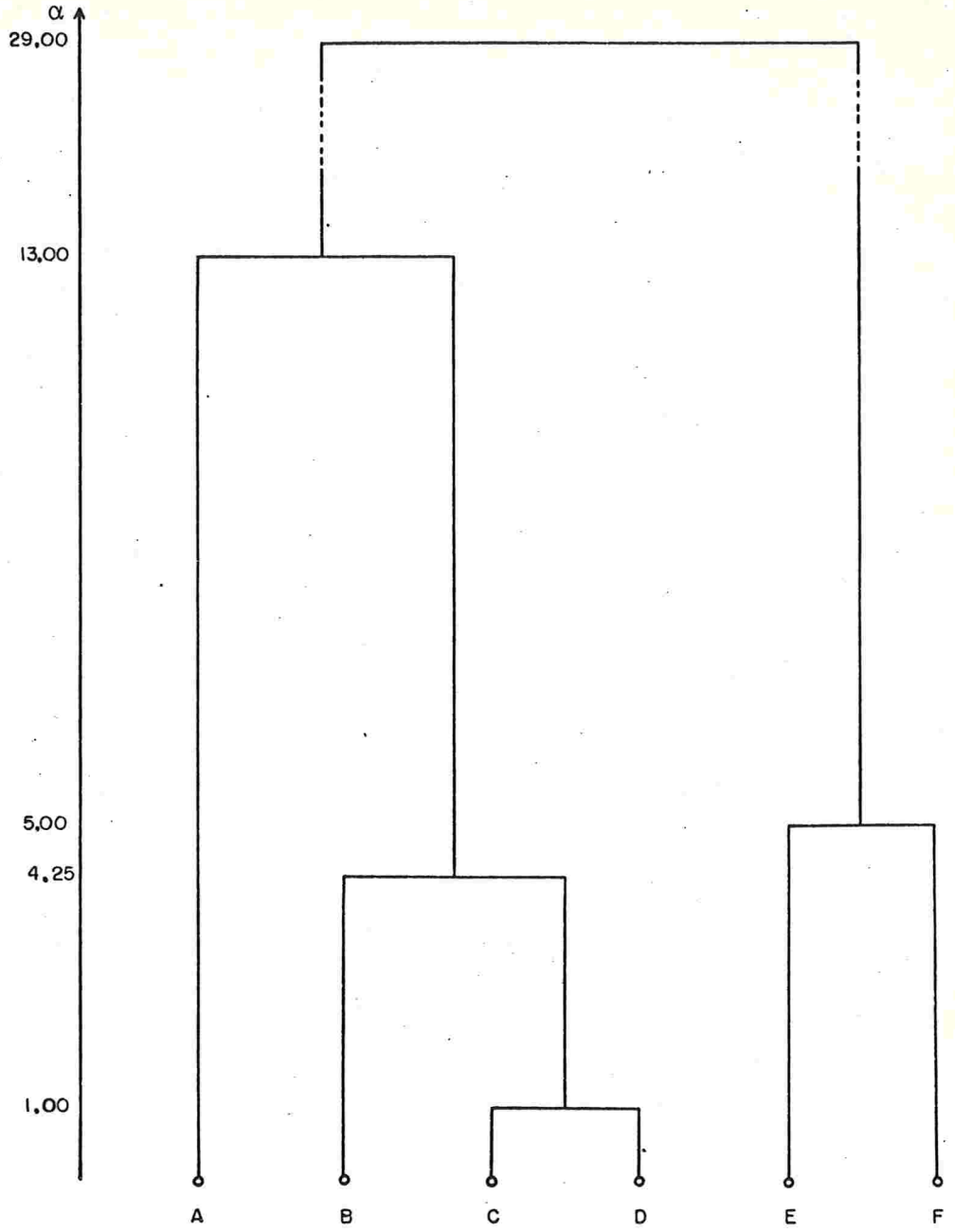


FIGURA 3.2 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO DE AGRUPAMENTO DO ENCADEAMENTO COMPLETO SOBRE D_1.

3.5 - MÉTODO DE AGRUPAMENTO DO ENCADEAMENTO MÉDIO

Com procedimento análogo aos dois métodos anteriormente descritos, aqui também partimos de um conjunto com N grupos, e a partir da matriz de proximidade serão fundidos no primeiro estágio os dois grupos mais próximos. Este grupo então, passará a ser representado pelo objeto cujas coordenadas é a média das coordenadas dos elementos que o compõem. Assim o processo de agrupamento transcorre até que no final tenhamos todos os N elementos iniciais reunidos num único grupo. Para construção do Exemplo 3.3, consideremos novamente os pontos do Exemplo 3.1.

EXEMPLO 3.3 - Consideremos os seis pontos do Exemplo 3.1 com respectiva matriz de proximidade D_1.

GRUPOS	COORDENADAS (X,Y)	
A	-1	2
B	0	-0,5
C	1	0
D	2	0
E	3	2
F	4	0

$$D_1 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & & & \\ 7,25 & 0,00 & & & & \\ 8,00 & 1,25 & 0,00 & & & \\ 13,00 & 4,25 & 1,00 & 0,00 & & \\ 16,00 & 15,25 & 8,00 & 5,00 & 0,00 & \\ 29,00 & 16,25 & 9,00 & 4,00 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

A menor distância encontrada em D_1 é dada entre os pontos C e D, cujo valor é $1,00$. Então no primeiro estágio serão fundidos C com D e os valores para a nova composi

ção de grupos e respectiva matriz de distância D_2^{**} são dados por:

GRUPOS	COORDENADAS (X,Y)	
A	-1	2
B	0	-0,5
CD	1,5	0
E	3	2
F	4	0

$$D_2^{**} = \begin{array}{c} \\ A \\ B \\ CD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & CD & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & & \\ 7,25 & 0,00 & & & \\ 10,25 & 2,5 & 0,00 & & \\ 16,00 & 15,25 & 6,25 & 0,00 & \\ 29,00 & 16,25 & 6,25 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

Com procedimento análogo se tem o segundo, terceiro e quarto estágios cujos resultados são apresentados a seguir.

SEGUNDO ESTÁGIO

GRUPOS	COORDENADAS (X,Y)	
A	-1	2
BCD	1	-0,16
E	3	2
F	4	0

$$D_3^{**} = \begin{array}{c} \\ A \\ BCD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccc} A & BCD & E & F \end{array} \\ \begin{bmatrix} 0,00 & & & \\ 8,66 & 0,00 & & \\ 16,00 & 8,66 & 0,00 & \\ 29,00 & 9,02 & 5,00 & 0,00 \end{bmatrix} \end{array}$$

TERCEIRO ESTÁGIO

GRUPOS	COORDENADAS (X,Y)	
A	-1	2
BCD	1	-0,16
EF	3,5	1

$$D_4^{**} = \begin{array}{c} \\ A \\ BCD \\ EF \end{array} \begin{array}{c} \begin{array}{ccc} A & BCD & EF \end{array} \\ \begin{bmatrix} 0,00 & & \\ 8,66 & 0,00 & \\ 21,25 & 7,59 & 0,00 \end{bmatrix} \end{array}$$

QUARTO ESTÁGIO

GRUPOS	COORDENADAS (X,Y)	
A	*-1*	*2*
BCDEF	*2*	*0,3*

$$D_5^{**} = \begin{array}{c} \\ A \\ BCDEF \end{array} \begin{array}{c} \begin{array}{cc} A & BCDEF \end{array} \\ \begin{bmatrix} 0,00 & \\ 11,89 & 0,00 \end{bmatrix} \end{array}$$

A representação gráfica do processo de agrupamento com respectivos valores para cada estágio é dado pelo dendrograma da Figura 3.3.

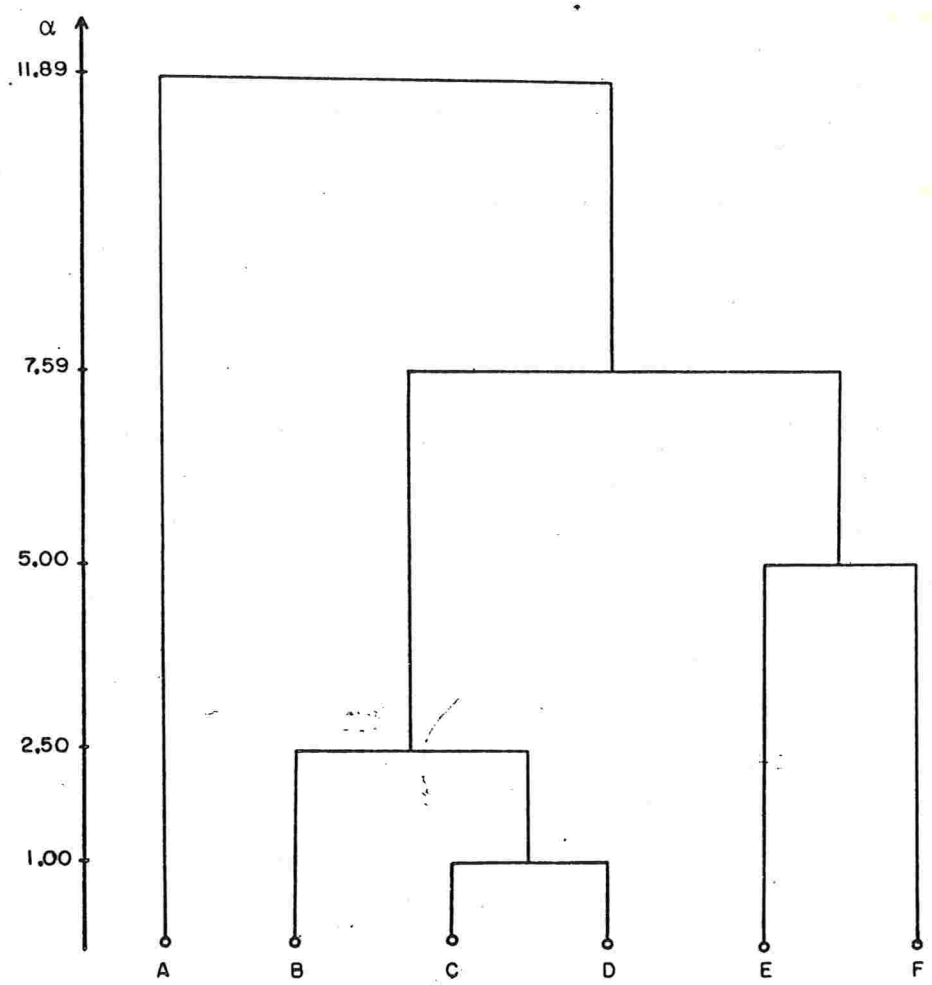


FIGURA 3.3 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO E.M. SOBRE D_1.

As Figuras 3.4a, 3.4b e 3.4c, nos mostra os pontos representados no plano *(X,Y)* e os contornos passo a passo para formação dos grupos, mostrando assim, claramente as diferenças no agrupamento, para os três métodos apresentados, quando aplicados sobre o mesmo conjunto de dados.

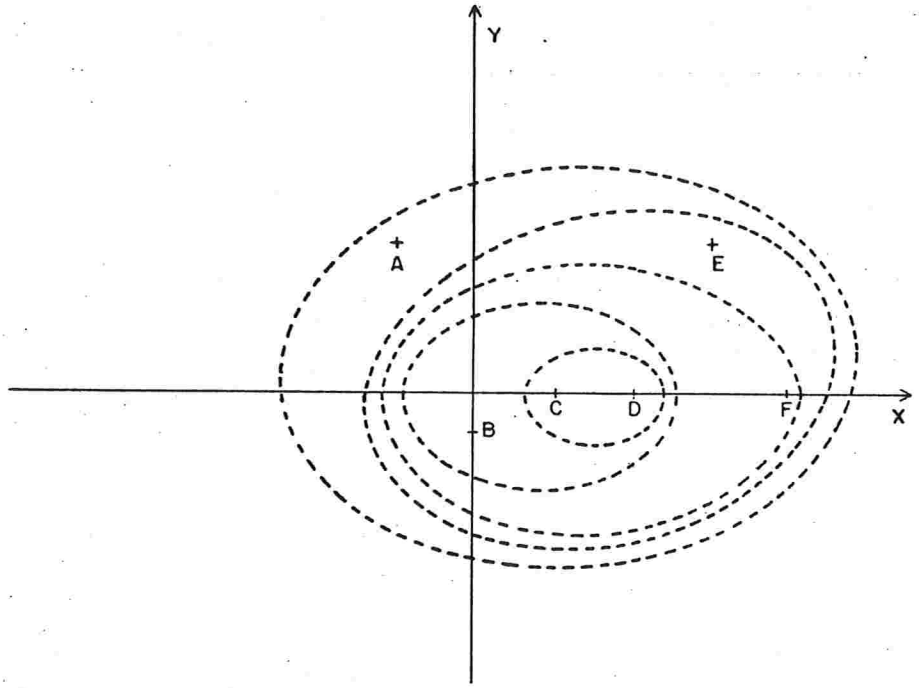


FIGURA 3.4a - REPRESENTAÇÃO DO AGRUPAMENTO OBTIDO PASSO A PASSO PELO MÉTODO VIZINHO MAIS PRÓXIMO

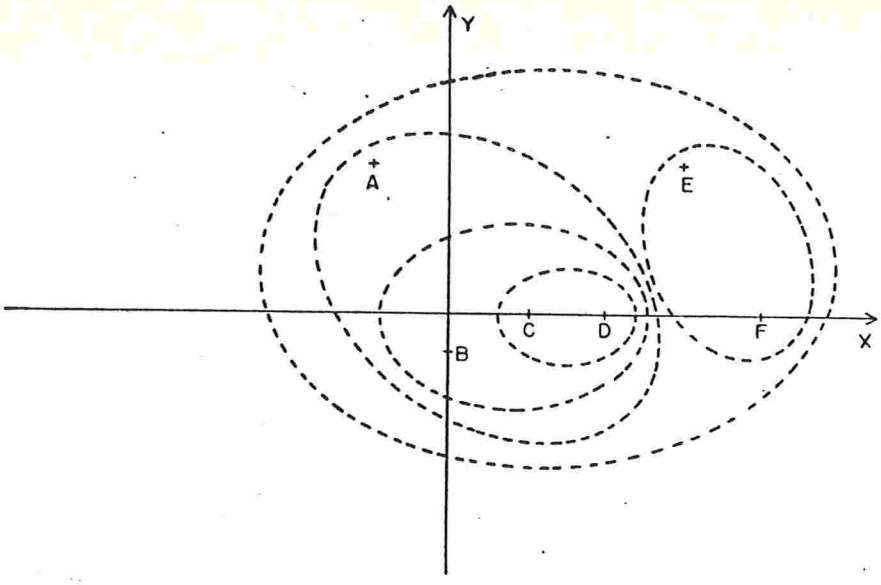


FIGURA 3.4b - REPRESENTAÇÃO DO AGRUPAMENTO OBTIDO PASSO A PASSO PELO MÉTODO DO ENCADEAMENTO COMPLETO.

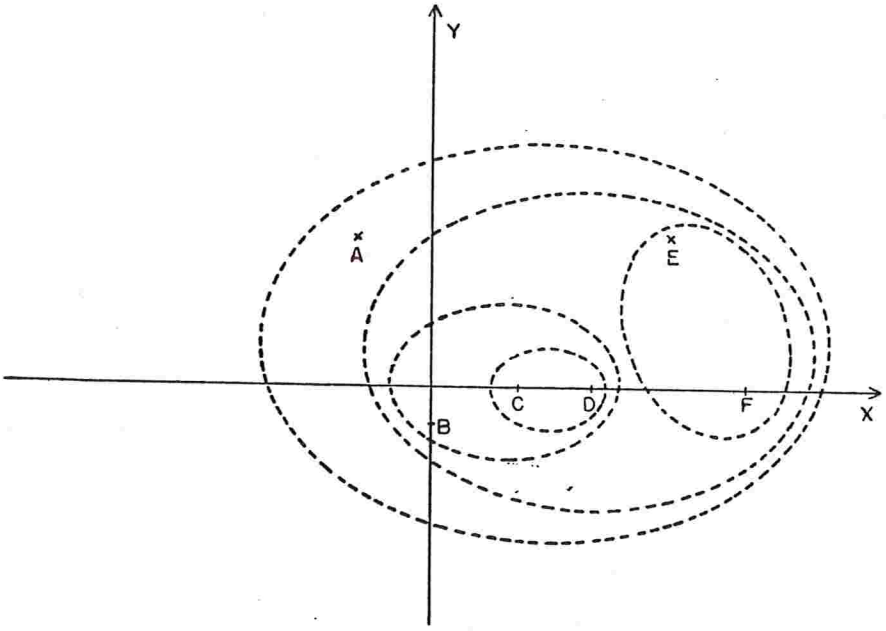


FIGURA 3.4c - REPRESENTAÇÃO DO AGRUPAMENTO OBTIDO PASSO A PASSO PELO MÉTODO DO ENCADEAMENTO MÉDIO.

3.6 - MEDIDA DE AJUSTE DO DENDROGRAMA

3.6.1 - INTRODUÇÃO

O problema de se definir um agrupamento ótimo, tem sido motivo de preocupação a vários autores de análise de conglomerados como Johnson (1967), Hartigan (1967), Jardine e Sibson (1968b), Miller (1969), Farris (1969), Rolf (1970), Levelt (1970) e outros.

Tem ocorrido a vários deles a idéia de definir um agrupamento perfeito como sendo aquele que retém a informação das distâncias originais encontrada na matriz de proximidade, usada como entrada de dados. Medidas de ajuste tem, então, naturalmente sido propostas, baseadas na discrepância entre os valores tomados como entrada de dados, $d(i,j)$, e os valores produzidos pelo algorítmo, $d^*(i,j)$.

Sokal e Rolf (1962), Farris (1970) e Rolf (1970) calcularam o coeficiente de correlação momento produto entre $d(i,j)$ e $d^*(i,j)$ como medida de ajuste, enquanto Johnson (1967) sugeriu a "correlação de postos". Hartigan (1967) propôs uma soma de quadrados ponderada $\Sigma w_{ij}[d(i,j)-d^*(i,j)]^2$; Ogilvie (1972) e Mojena (1977) apresentam duas outras medidas de ajuste, que serão objeto de estudo neste capítulo.

3.6.2 - COEFICIENTE DE CUNNINGHAM E OGILVIE (1972)

Dada uma matriz de proximidade D definida pelo

conjunto de N objetos, com respeito a alguma medida, Ogilvie propôs uma medida de ajuste do dendrograma obtido quando da aplicação do método de agrupamento sobre D, dado por:

$$C_0 = \sum_{i,j} [d(i,j)-d^*(i,j)]^2 / \sum_{i,j} d^2(i,j) \qquad [3.3]$$

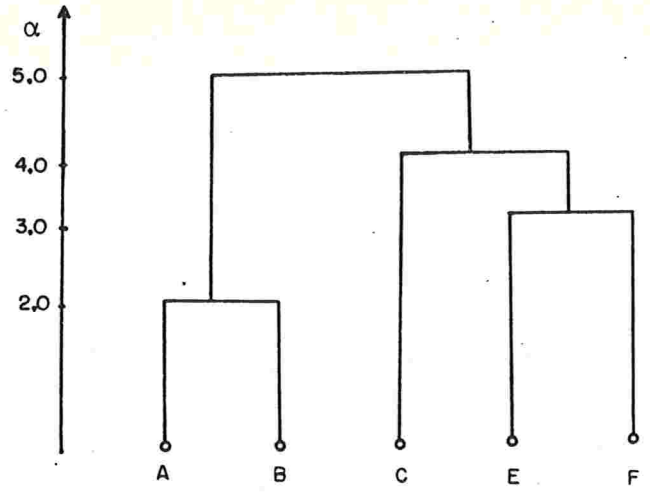
$d(i,j)$ é encontrado na matriz D, e $d^*(i,j)$ é obtido do dendrograma.

Para o cálculo de C_0 faremos uso somente de $N(N-1)/2$ valores, que corresponde exatamente aos elementos da parte triangular superior ou inferior da matriz D excluídos é claro, os elementos da diagonal principal.

Consideremos para ilustração, o exemplo a seguir:

EXEMPLO 3.4 - Considere a matriz D_1 e respectivo dendrograma resultante da aplicação do método de agrupamento vizinho mais próximo sobre D.

$$D = \begin{array}{c} \\ A \\ B \\ C \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & E & F \end{array} \\ \begin{bmatrix} 0,0 & & & & \\ 2,0 & 0,0 & & & \\ 6,0 & 5,0 & 0,0 & & \\ 10,0 & 9,0 & 4,0 & 0,0 & \\ 9,0 & 8,0 & 5,0 & 3,0 & 0,0 \end{bmatrix} \end{array}$$



Através do dendrograma obtemos D^* dada por:

$$D^* = \begin{array}{c} \\ A \\ B \\ C \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & E & F \end{array} \\ \begin{bmatrix} 0,0 & & & & \\ 2,0 & 0,0 & & & \\ 5,0 & 5,0 & 0,0 & & \\ 5,0 & 5,0 & 4,0 & 0,0 & \\ 5,0 & 5,0 & 4,0 & 3,0 & 0,0 \end{bmatrix} \end{array}$$

usando [3.3] obtemos $C_0=0,263$.

Como se pode observar pela [3.3], o coeficiente de Ogilvie apresenta em seu denominador um fator de escalonamento, que torna a medida invariante sob mudança de escala. Este coeficiente assume valores que não estão restritos ao intervalo $[0;1]$, e C_0 será uma boa medida de ajuste quando produzir um valor relativamente pequeno.

3.6.3 - COEFICIENTE DE MOJENA

Se a estrutura base, isto é, a identificação da população dos elementos de um particular conjunto de dados é conhecida, então os resultados do método de agrupamento poderá ser avaliado diretamente. Para um dado conjunto com n elementos, Mojena define uma matriz $n \times n$ simétrica com valores *zero* ou *um* na posição ij, que é chamada matriz de incidência. O valor *zero* na posição ij significa dizer que os elementos i e j não pertencem à mesma população, ou seja não estão no mesmo grupo. Caso contrário, na posição ij o valor será *um*.

Desta forma, conhecendo-se as matrizes de incidência, uma dada pela estrutura inicial conhecida e, a outra obtida do dendrograma resultante da aplicação do método de agrupamento podemos construir um índice que meça a discrepância entre as duas matrizes. Mojena (1977), então, sugeriu para medir esta discrepância o coeficiente:

$$C_M = \sum_{i,j} \frac{C_{ij}}{n(n-1)/2} \qquad [3.4]$$

onde $C_{ij}=1$ se as correspondentes entradas ij nas duas matrizes têm o mesmo valor, caso contrário $C_{ij}=0$. A divisão por $n(n-1)/2$ deve-se ao fato de estarmos usando somente a parte triangular inferior ou superior das matrizes de incidência, excluindo-se a diagonal principal.

C_M assume valores no intervalo $[0;1]$; indicando nos então que valores próximos de 1 indicam bons resultados ou seja, bons agrupamentos.

Para ilustração desta medida consideremos o exemplo a seguir:

EXEMPLO 3.5 - Sejam $\underset{\sim}{\pi}_1$ e $\underset{\sim}{\pi}_2$ distribuições multinomiais conhecidas

$\underset{\sim}{\pi}_1 = [0,2;\ 0,2;\ 0,2;\ 0,2;\ 0,2]$

$\underset{\sim}{\pi}_2 = [0,4;\ 0,085;\ 0,04;\ 0,085;\ 0,4]$

Para cada distribuição foram selecionadas 3 amostras de tamanho 100 cada uma, e usando como medida de proximidade a medida de Bhattacharyya, fórmula [2.3] obtivemos a matriz de proximidade, cujo dendrograma resultante da aplicação do método de agrupamento do encadeamento completo é dado pela Figura 3.5

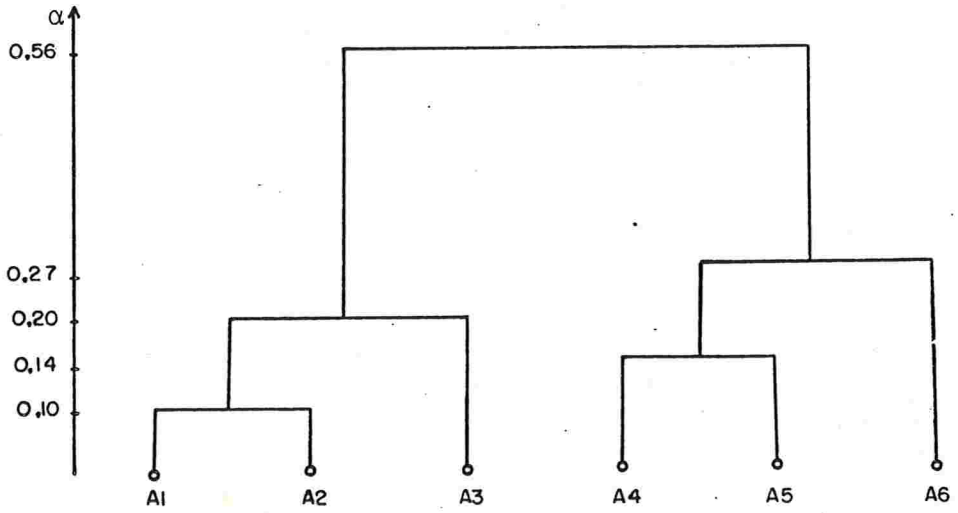


FIGURA 3.5 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO DE AGRUPAMENTO DO ENCADEAMENTO COMPLETO.

OBSERVAÇÃO:

As amostras A_1, A_2 e A_3 são provenientes da população $\underset{\sim}{\pi}_1$, e as amostras A_4, A_5 e A_6 são provenientes da população $\underset{\sim}{\pi}_2$.

A matriz de incidência obtida da estrutura inicial conhecida é dada por:

$$I_{ini} = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{array} \begin{array}{c} \begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \end{array} \\ \begin{bmatrix} & & & & & \\ 1 & & & & & \\ 1 & 1 & & & & \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & 1 & \end{bmatrix} \end{array}.$$

Vamos aqui considerar os agrupamentos passo a passo, e as respectivas matrizes de incidência e o valor de C_M naquele estágio.

PRIMEIRO ESTÁGIO

A_1 se funde com A_2, e as matrizes restantes são consideradas como grupos distintos. Assim a matriz de incidência I_1 é dada por:

$$I_1 = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{array} \begin{array}{c} \begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \end{array} \\ \begin{bmatrix} & & & & & \\ 1 & & & & & \\ 0 & 0 & & & & \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & 0 & \end{bmatrix} \end{array}$$

$$C_M = \frac{10}{15} \cong 0,66.$$

SEGUNDO ESTÁGIO

A_4 se funde com A_5, e agora passamos a ter quatro grupos formados por (A_1,A_2), A_3, (A_4,A_5) e A_6. A matriz de incidência I_2 é dada por:

$$I_2 = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{array} \begin{array}{c} \begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \end{array} \\ \begin{bmatrix} & & & & & \\ 1 & & & & & \\ 0 & 0 & & & & \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 0 & \end{bmatrix} \end{array}$$

$$C_M = \frac{11}{15} \cong 0,73.$$

TERCEIRO ESTÁGIO

A_3 se funde com (A_1,A_2) e agora passamos então a ter três grupos: (A_1,A_2,A_3), (A_4,A_5) e A_6. A matriz de incidência I_3 é dada por:

$$I_3 = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{array} \begin{array}{c} \begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \end{array} \\ \begin{bmatrix} & & & & & \\ 1 & & & & & \\ 1 & 1 & & & & \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 0 & \end{bmatrix} \end{array}$$

$$C_M = \frac{13}{15} \cong 0,86$$

QUARTO ESTÁGIO

Aqui se estabelece a fusão de A_6 ao grupo (A_4,A_5). Dessa forma neste estágio forma-se dois grupos, (A_1,A_2,A_3), (A_4,A_5,A_6), cuja matriz de incidência é dada por I_4:

$$I_4 = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{array} \begin{array}{c} \begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \end{array} \\ \begin{bmatrix} & & & & & \\ 1 & & & & & \\ 1 & 1 & & & & \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & 1 & \end{bmatrix} \end{array}$$

$$C_M = \frac{15}{15} = 1.$$

No quinto estágio as amostras estão todas num único grupo, e neste caso $C_M = \frac{6}{15} = 0,4$.

Através deste exemplo percebemos que o estágio ótimo de parada em termos de C_M, ocorre exatamente no quarto estágio.

O conhecimento da estrutura inicial é que possibilitou a determinação do estágio ótimo de parada, porém na maioria das situações práticas isto não ocorre e, necessitamos então, o estabelecimento de regras de parada, que é apresentado no Capítulo IV.

Com os Exemplos 3.4 e 3.5 vemos claramente a diferença existente entre os coeficientes propostos por Ogilvie e Mojena. O coeficiente de Ogilvie não depende do conhecimento da estrutura inicial, ao passo que o coeficiente de Mojena depende.

C A P Í T U L O I V

REGRA DE PARADA

4.1 - INTRODUÇÃO

Um problema comum a todas as técnicas de agrupamento reside na dificuldade em decidir a respeito do número de grupos presente ao conjunto de dados em estudo. Através dos Exemplos 3.4 e 3.5, vimos que o desconhecimento da estrutura inicial, leva-nos naturalmente a escolher regras de parada que determinem o número de grupos. Para as técnicas onde se procura otimizar algum critério de agrupamento, é geralmente sugerido que se faça o gráfico dos valores do critério contra o número de grupos, que indicará o número correto a ser considerado. Isto tem sido proposto por alguns autores (por exemplo, Friedman e Rubin, 1967); porém, em ge

ral o procedimento tem se mostrado insatisfatório.

Para o modelo hierárquico aglomerativo, conforme caracterização dada no Capítulo III, isto é, um conjunto de N partições denotadas por $P_0, P_1, \ldots, P_{N-1}$ e os correspondentes valores do critério escolhido $\alpha_0, \alpha_1, \ldots, \alpha_{N-1}$; iremos apresentar duas regras de parada propostas por Mojena (1977).

4.2 - PRIMEIRA REGRA

Regras estatísticas para selecionar a partição que melhor se aproxime da estrutura base, pode ser baseada na distribuição dos valores do critério: $\alpha_0, \alpha_1, \ldots, \alpha_{N-1}$ ou em transformações destes. Se um valor relativamente grande de α implicar em objetos dissimilares, então a distribuição dos α's é monotonicamente crescente, caso contrário, é monotonicamente decrescente. Assim, um salto muito grande pode ser definido como um salto "significante", para determinação do número de grupos.

Esta regra utiliza a distribuição dos $N-1$ valores de α, escolhendo um valor significante na cauda superior da distribuição, que pode ser expresso em função da média e o desvio padrão da distribuição.

Mojena (1977) propõe selecionar o valor do critério correspondente ao primeiro estágio j, $j=1,\ldots,N-1$ que sa

tisfaça a desigualdade:

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha} \qquad [4.1]$$

para um dado afastamento padrão k, e onde $\bar{\alpha}$ e s_{α} são respectivamente a média e o desvio padrão dos α.

No caso em que α é decrescente, escolhe-se um valor significante, na cauda inferior da distribuição dos α's, e a desigualdade se inverte.

Se nenhum valor de α satisfizer [4.1], precisamos tomar outras decisões, como por exemplo:

a) escolher o estágio j tal que o estágio $j+1$ produza o maior salto, isto é, o estágio $j+1$ em que $\alpha_{j+1}-\alpha_j$ seja máximo; ou

b) fazer uma análise exploratória dos α's, do tipo daquelas dadas por Tukey (1977), procurando descobrir um valor de α, que possa ser considerado significante.

EXEMPLO 4.1 - Sejam $\underset{\sim}{\pi}_1$ e $\underset{\sim}{\pi}_2$ duas distribuições multinomiais dadas por

$$\underset{\sim}{\pi}_1 = [0,4;\ 0,09;\ 0,02;\ 0,09;\ 0,4]$$
$$\underset{\sim}{\pi}_2 = [0,02;\ 0,09;\ 0,8;\ 0,09;\ 0,02]$$

Para cada distribuição foram retiradas 10 amostras aleatórias, cujos resultados estão na Tabela 4.1.

TABELA 4.1

	AMOSTRA	CATEGORIA					
		PRIMEIRA	SEGUNDA	TERCEIRA	QUARTA	QUINTA	TOTAL
Provenientes de π_1	A_1	39	13	1	12	35	100
	A_2	41	9	3	9	38	100
	A_3	50	5	1	8	36	100
	A_4	38	14	0	12	36	100
	A_5	42	10	3	7	38	100
	A_6	34	12	3	10	41	100
	A_7	42	7	3	6	42	100
	A_8	47	8	1	8	36	100
	A_9	40	9	0	7	44	100
	A_{10}	40	6	3	10	41	100
Provenientes de π_2	A_{11}	2	6	80	12	0	100
	A_{12}	3	12	78	7	0	100
	A_{13}	4	4	80	12	0	100
	A_{14}	1	7	87	5	0	100
	A_{15}	0	10	79	11	0	100
	A_{16}	1	8	78	12	1	100
	A_{17}	3	12	76	8	1	100
	A_{18}	4	10	75	11	0	100
	A_{19}	2	10	78	10	0	100
	A_{20}	0	9	85	6	0	100

Utilizando a medida proposta por Sanghvi, fórmula [2.5] e aplicando sobre a matriz de proximidade o método de agrupamento do encadeamento médio, encontramos a seguinte distribuição dos α:

α	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}
VALOR	0	0,01	0,01	0,02	0,02	0,02	0,02	0,02	0,02	0,03	0,03

α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}	α_{19}
0,04	0,05	0,06	0,06	0,07	0,09	0,09	3,00

OBSERVAÇÃO:

Os valores do critério α, aparecem repetidos devido a aproximação usada.

$\overline{\alpha} = 0,19$ $\qquad s_\alpha = 0,68$

A Tabela 4.2 nos mostra a variação de $\overline{\alpha}+ks_\alpha$ para determinados valores de k:

TABELA 4.2

k	$\overline{\alpha}+ks_\alpha$
1,68	1,33
1,96	1,52
2,00	1,55
2,50	1,89
3,00	2,23

Como se pode observar, o único valor α_{j+1} da sequência acima, que satisfaz a desigualdade [4.1] para os valores de k na Tabela 4.2 é $\alpha_{19}=3,00$. Portanto, o estágio de parada é o décimo oitavo. Para melhor visualização deste resultado observemos o dendrograma apresentado na Figura 4.1.

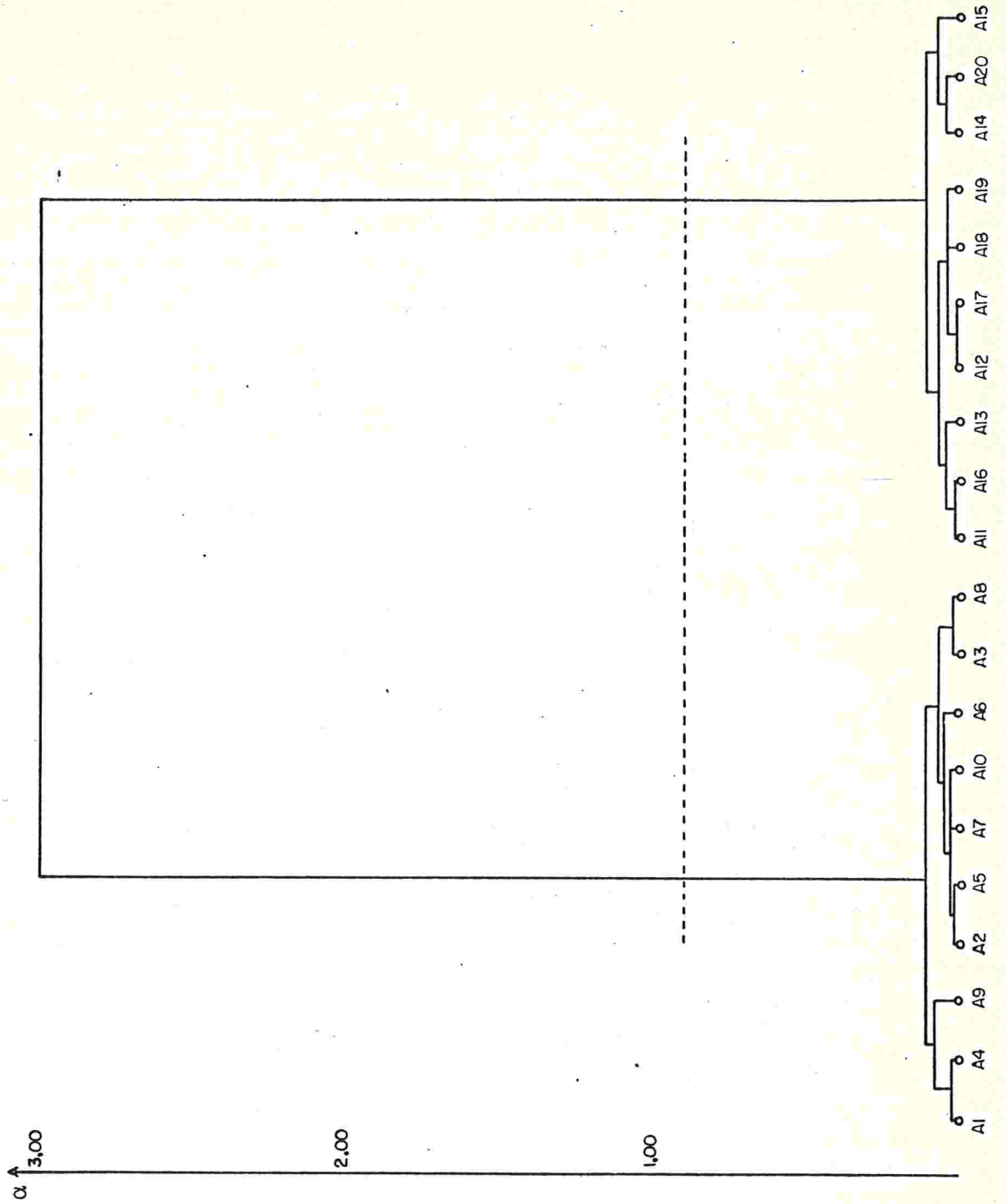


FIGURA 4.1 - DENDROGRAMA RESULTANTE DA APLICAÇÃO DO MÉTODO DO ENCADEAMENTO MÉDIO SOBRE A MATRIZ DE PROXIMIDADE ENCONTRADA A PARTIR DOS DADOS DA TABELA 4.1

EXEMPLO 4.2 - Consideremos o dendrograma apresentado pela Figura 4.2 e respectiva distribuição de α, obtido pela aplicação do método de agrupamento do vizinho mais próximo sobre uma matriz de proximidade, cuja distância utilizada foi a distância Euclidiana, para 18 amostras de areia coletadas na região de Caraguatatuba, para estudos sobre movimento de areia naquela região.

FONTE: Dados apresentados ao SEA, IME-USP, São Paulo, por Olga Cruz, (1979).

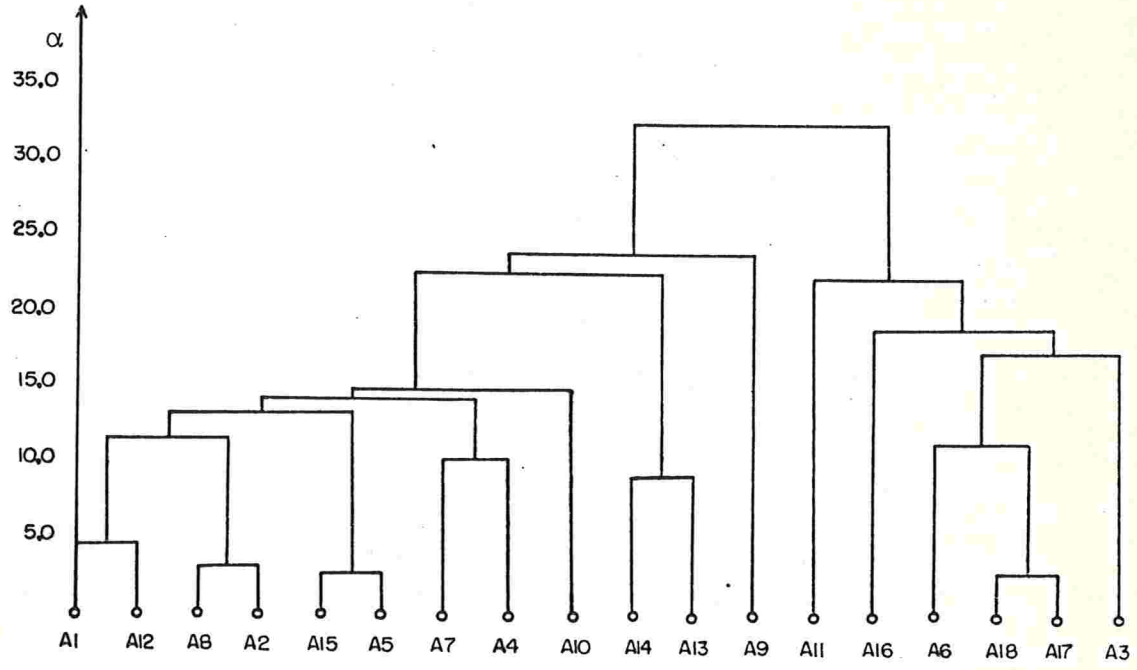


FIGURA 4.2

α	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
VALOR	2,261	2,961	2,979	4,686	9,766	10,145	11,618	11,771

α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}
13,820	14,626	15,014	17,551	18,728	22,547	22,656	24,726	32,252

$\bar{\alpha} = 14,006$ $\qquad s_\alpha = 8,455.$

A Tabela 4.3 nos mostra os valores de $\bar{\alpha}+ks_\alpha$ quando k assume determinados valores.

TABELA 4.3

k	$\bar{\alpha}+ks_\alpha$
1,00	22,461
1,68	28,210
1,96	30,578
2,00	30,916
2,50	35,143
3,00	39,371

Ao compararmos os valores de α, da sequência anterior, com os valores de $\bar{\alpha}+ks_\alpha$ da Tabela 4.3 obteremos os seguintes resultados:

Para $k=1$, $\bar{\alpha}+ks_\alpha=22,461$ e o valor de $\alpha_{j+1}>22,461$ corresponde a $\alpha_{14}=22,547$. Neste caso temos a formação de quatro grupos: $(A_1; A_{12}; A_8; A_2; A_{15}; A_5; A_7; A_4; A_{10})$;$(A_{14}; A_{13})$; (A_{19}); $(A_{11}; A_{16}; A_6; A_{18}; A_{17}; A_3)$.

Para $k=1,96$ e $k=2,00$ o resultado é o mesmo que o obtido para $k=1,68$.

Para $k=2,50$ e $k=3,00$ não encontramos α_{j+1} que satisfaça a desigualdade $\alpha_{j+1} > \overline{\alpha} + ks_{\alpha}$.

4.3 - SEGUNDA REGRA

Aqui o comportamento dos α's é considerado como aqueles de uma série temporal com tendência, que sugere como aproximação o modelo de uma série de período *um*, e definindo um α significante no estágio $j+1$. O modelo proposto é o de média móvel corrigido para a tendência linear. Vejamos então, o caso em que α é crescente:

Procuramos selecionar a partição correspondente ao primeiro estágio j, $j=n,n+1,\ldots,N-2$ que satisfaça a desigualdade

$$\alpha_{j+1} > \overline{\alpha}_j + L_j + b_j + ks_j \qquad [4.2]$$

onde n é o número de observações da média móvel; $\overline{\alpha}_j$ é a média dos elementos até o estágio j; L_j é a correção para a tendência no estágio j; b_j é a estimativa de mínimos quadrados para a inclinação no estágio j; k é o afastamento padrão e s_j é a estimativa do desvio padrão no estágio j.

Portanto, $\hat{\alpha}_{j+1}$, estimativa de α_{j+1} é dada por:

$$\hat{\alpha}_{j+1} = \bar{\alpha}_j + L_j + b_j + ks_j$$

Se nenhum valor de α satisfizer [4.2], então os procedimentos alternativos são os mesmos dados quando apresentamos a primeira regra de parada.

A estimativa de mínimos quadrados para a inclinação no estágio j é dada por:

$$b_j = \frac{6\left[2 \sum_{i=j-n+1}^{j} w_i \alpha_i - (n+1) \sum_{i=j-n+1}^{j} \alpha_i\right]}{n(n^2-1)} \qquad [4.3]$$

onde n e α são definidos como anteriormente; $w_i = w_{i-1}+1$ dado que $w_{j-n+1}=1$; $i=j-n+2,\ldots,j$.

A correção para tendência linear é dada por:

$$L_j = \frac{(n-1)}{2} b_j$$

Na Figura 4.3 apresentamos um esboço gráfico para a estimativa de α_{j+1} segundo o modelo de previsão, $\bar{\alpha}_j + b_j + L_j + ks_j$.

Para obtenção de b_j vejamos o desenvolvimento a seguir:

A inclinação da reta passando pelos pontos $(1,\alpha_1)$; $(2,\alpha_2);\ldots;(n,\alpha_n)$ é dada por:

$$b = \sum_{i=1}^{n} (\alpha_i - \bar{\alpha})(i-\bar{i}) / \sum_{i=1}^{n} (i-\bar{i})^2$$

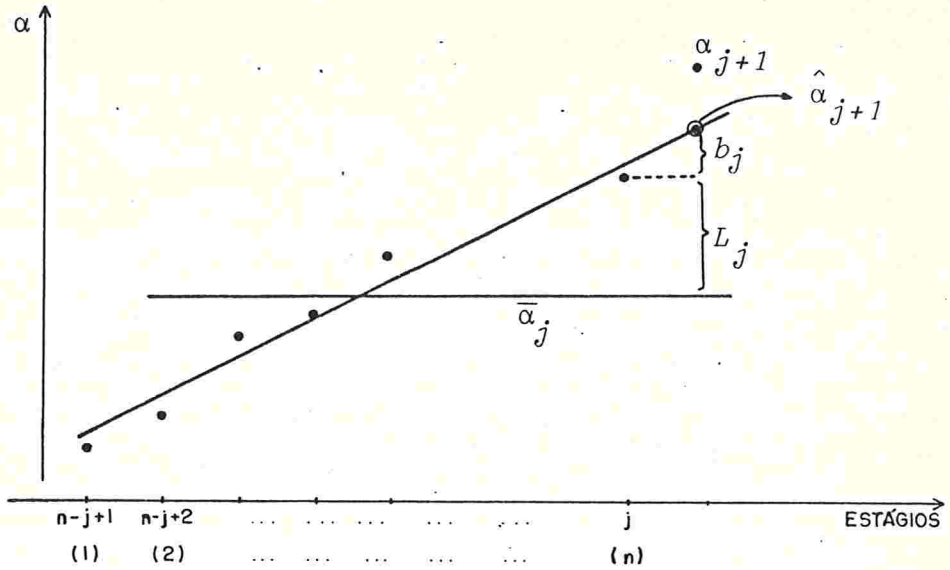


FIGURA 4.3

i) $$\sum_{i=1}^{n}(\alpha_i-\bar{\alpha})(i-\bar{i}) = \sum_{i=1}^{n}\alpha_i i-n\bar{\alpha}\,\bar{i} = \sum_{i=1}^{n}\alpha_i i-\bar{i}\sum_{i=1}^{n}\alpha_i$$

$$\overset{\circledast}{=} \sum_{i=1}^{n}\alpha_i i-\frac{(n+1)}{2}\sum_{i=1}^{n}\alpha_i = [2\sum_{i=1}^{n}\alpha_i i-(n+1)\sum_{i=1}^{n}\alpha_i]/2$$

ii) $$\sum_{i=1}^{n}(i-\bar{i})^2 = \sum_{i=1}^{n}i^2-n\bar{i}^2 = \sum_{i=1}^{n}i^2-\frac{1}{n}(\sum_{i=1}^{n}i)^2 \overset{\circledast}{=} \sum_{i=1}^{n}i^2-\frac{n(n+1)^2}{4} \overset{\circledast\circledast}{=}$$

$$\frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}$$

$\circledast$: $$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$\circledast\circledast$: $$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Por (i) e (ii), então:

$$b = \frac{6\left[2\sum_{i=1}^{n}\alpha_i i-(n+1)\sum_{i=1}^{n}\alpha_i\right]}{n(n^2-1)}$$

Como o modelo proposto é o de médias móveis, onde n é o número de elementos da mesma, fazendo então as substituições obtemos:

$$b_j = \frac{6\left[2\sum_{i=j-n+1}^{j}\alpha_i w_i-(n+1)\sum_{i=j-n+1}^{j}\alpha_i\right]}{n(n^2-1)} \quad \text{com } w_i=i$$

Para ilustrarmos a aplicação desta regra, vejamos um exemplo.

Através da Figura 4.4 notamos que os estágios correspondentes aos valores de α, que poderão ser considerados como salto significante, para determinadas escolhas de valores de n e k são: o 6º, 9º, 12º, 14º e 17º estágios, com a formação de 14, 10, 7, 5 e 2 o número de grupos respectivamente nestes estágios (veja dendrograma da Figura 4.2). A Tabela 4.4 apresenta os valores $\hat{\alpha}_{j+1}$, para diferentes valores de n e k, e também o estágio j e respectivo valor α_{j+1} para comparação com $\hat{\alpha}_{j+1}$ através de [4.2].

EXEMPLO 4.3 - Consideremos a sequência de valores de α do Exemplo 4.2, representada graficamente pela Figura 4.4, com período constante *1*.

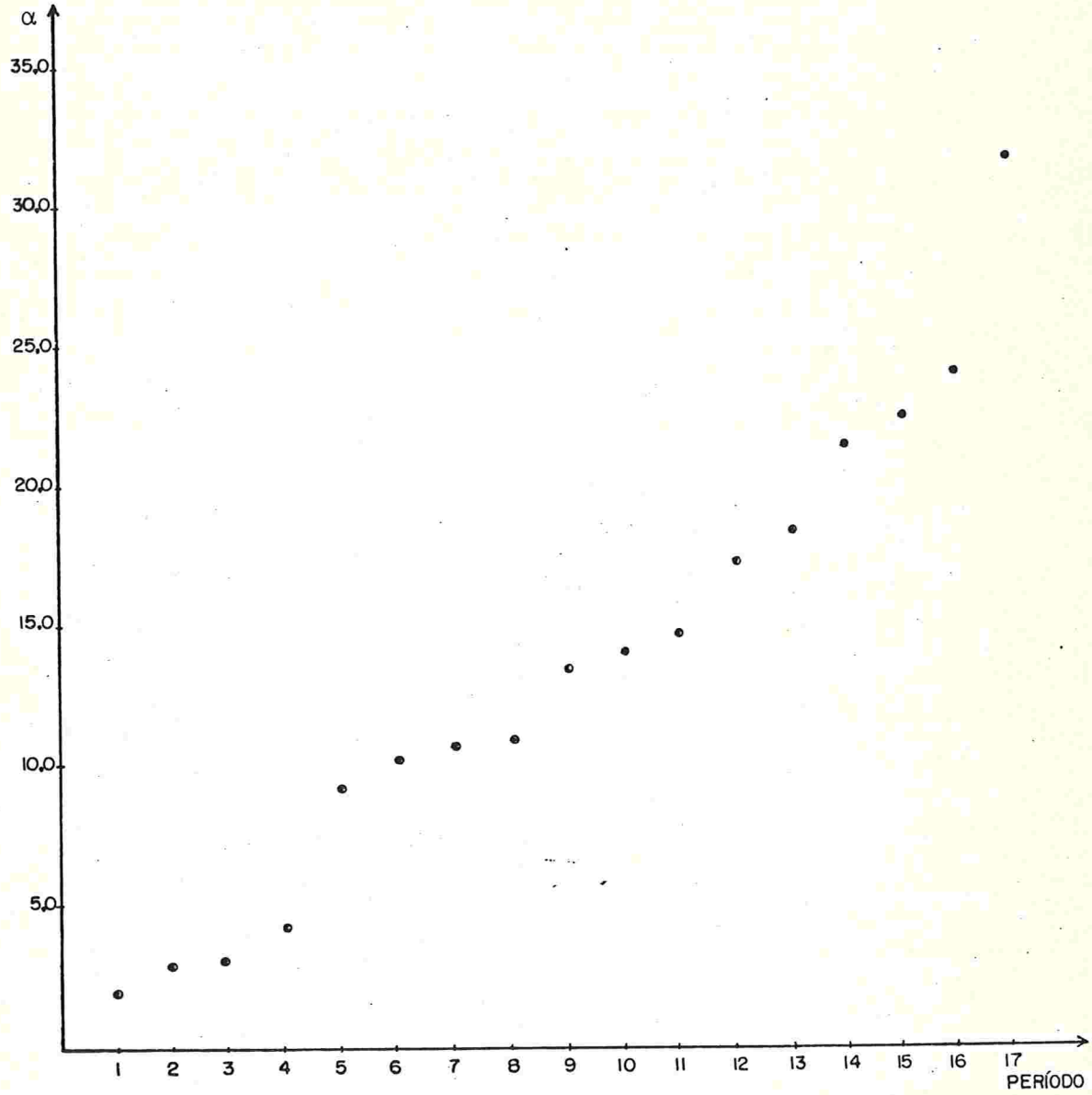


FIGURA 4.4 - GRÁFICO DA SEQUÊNCIA DE α DO EXEMPLO 4.2

TABELA 4.4

n	j	VALORES DE k 1,0	2,0	2,5	3,0	3,5	α_{j+1}
	4	6,07	7,10	7,61	8,13	8,64	9,766
	5	24,02	27,24	28,84	30,45	32,05	10,145
	6	44,69	48,30	50,10	51,90	53,70	11,618
	7	71,67	74,69	76,20	77,42	79,23	11,771
	8	100,30	101,32	101,83	102,34	102,85	13,820
	9	134,52	136,04	136,79	137,55	138,30	14,626
	10	172,71	174,20	174,95	175,70	176,44	15,014
4	11	211,15	212,59	213,32	214,04	214,76	17,551
	12	263,76	265,37	266,17	266,98	267,78	18,728
	13	318,78	320,76	321,75	322,75	323,74	22,547
	14	396,75	399,88	401,45	403,02	404,59	22,656
	15	475,91	478,53	479,84	481,15	482,46	24,726
	16	561,08	563,57	564,82	566,07	567,32	32,252
	8	18,50	22,68	24,76	26,85	28,94	13,820
	9	27,57	31,86	34,01	36,15	38,30	14,626
	10	38,28	42,41	44,47	46,54	48,60	15,014
	11	49,94	53,30	54,98	56,66	58,34	17,551
8	12	65,23	67,90	69,24	70,58	71,92	18,728
	13	83,18	86,15	87,63	89,12	90,60	22,547
	14	106,80	110,52	112,38	114,24	116,09	22,656
	15	130,85	134,86	136,87	138,88	140,89	24,726
	16	158,67	162,84	164,93	167,02	169,11	32,252
	11	22,67	27,61	30,08	32,55	35,02	17,551
	12	30,46	35,45	37,95	40,44	42,94	18,728
11	13	39,59	44,48	46,92	49,37	51,81	22,547
	14	51,70	56,58	59,03	61,47	63,91	22,656
	15	64,65	69,23	71,52	71,81	76,10	24,726
	16	80,43	85,40	87,89	90,37	92,86	32,252
	13	26,09	31,75	34,58	37,41	40,24	22,547
13	14	34,59	40,59	43,59	46,59	49,59	22,656
	15	43,65	49,66	52,66	55,67	58,67	24,726
	16	54,18	60,03	62,96	65,88	68,81	32,252
16	16	32,93	40,19	43,81	47,44	51,07	32,252

Da Tabela 4.4 observamos que os únicos valores de $\alpha_{j+1} > \hat{\alpha}_{j+1}$ são os correspondentes a $n=4$ e $j=4$ para $k=1,0$; $2,0$; $2,5$; $3,0$ e $3,5$, onde $\alpha_{j+1} = \alpha_5 = 9,766$. Neste caso então, paramos no quarto estágio e obtemos a formação de 14 grupos.

4.4 - COMENTÁRIOS

Uma pergunta que surge naturalmente a respeito das regras de parada aqui tratadas, é quanto a adequacidade das mesmas. Conforme podemos observar, ambas dependem do parâmetro k, e a segunda depende também da escolha do parâmetro n.

Não obstante, ainda não se conhecem modelos teóricos que avaliem o desempenho destas regras, assim elas podem ser analisadas através de procedimentos empíricos, ou seja, partindo-se de estruturas conhecidas, simular amostras e analisar o comportamento das mesmas.

Aqui iremos verificar algumas propriedades através desse procedimento.

C A P Í T U L O V

AVALIAÇÕES ATRAVÉS DE SIMULAÇÃO

5.1 - INTRODUÇÃO

Apresentamos neste capítulo os resultados de avaliações feitas sobre simulações realizadas com estruturas multinomiais, que são as que caracterizam a estrutura de dados deste trabalho. Para isto os seguintes passos foram seguidos:

1) Fixamos algumas distribuições multinomiais.

2) Amostras aleatórias das respectivas distribuições foram geradas.

3) Para as amostras obtidas pelo Ítem 2, obtivemos as matrizes de proximidade, utilizando-se das medidas:

D_1: Distância Euclidiana

D_2: Distância Sanghvi

D_3: Distância Bhattacharyya

D_4: Distância Kullback

S_1: Similaridade de Horn

S_2: Similaridade de Morisita.

4) Às matrizes obtidas no Ítem 3, aplicamos os métodos de agrupamento:

E.C.: Encadeamento Completo

V.P.: Vizinho mais Próximo

E.M.: Encadeamento Médio.

5) A partir do dendrograma resultante, obtido pela aplicação de cada método de agrupamento, obtivemos as matrizes para o cálculo do coeficiente de Ogilvie, que avalia o ajuste do dendrograma.

6) À sequência de níveis α_j, $j=1,\ldots,N-1$ resultante da aplicação do método de agrupamento, utilizando-se das regras de parada sugeridas no Capítulo IV, fazendo k (afastamento padrão) e n (número de ítens considerados na média móvel) variar, obtivemos o número de grupos em cada caso.

Observamos aqui, que um número muito grande de combinações de distribuições fixadas a priori, poderão ser escolhidas, mas neste trabalho optamos por tomar distribuições multinomiais com 5 categorias de classificação, sendo que

nos três primeiros casos tomamos a distribuição multinomial uniforme $\underset{\sim}{\pi}_1=[0,2;\ 0,2;\ 0,2;\ 0,2;\ 0,2]$ como padrão, e deixamos as outras distribuições variarem em diferentes direções. Observamos também, que o conhecimento das estruturas iniciais é que nos possibilitam a avaliação dos resultados obtidos pela aplicação das medidas de proximidade, métodos de agrupamento e regras de parada.

5.2 - POPULAÇÃO E AMOSTRAS

Apresentamos aqui as estruturas populacionais escolhidas a priori, para que amostras aleatórias, das mesmas fossem geradas.

Para duas populações, selecionamos dois casos:

Caso *A*:

$$\underset{\sim}{\pi}_1 = [0,2;\ 0,2;\ 0,2;\ 0,2;\ 0,2]$$
$$\underset{\sim}{\pi}_2 = [0,4;\ 0,09;\ 0,02;\ 0,09;\ 0,4]$$

Caso *B*:

$$\underset{\sim}{\pi}_1 = [0,2;\ 0,2;\ 0,2;\ 0,2;\ 0,2]$$
$$\underset{\sim}{\pi}_2 = [0,45;\ 0,25;\ 0,2;\ 0,05;\ 0,05]$$

Para ambos os casos, 10 amostras de tamanho 100 de cada uma das distribuições foram geradas, produzindo assim para cada caso uma matriz de proximidade de dimensão 20×20, conforme a medida de proximidade considerada.

Para três populações consideramos um caso somen

te:

Caso C:

$$\underset{\sim}{\pi}_1 = [0,2;\ 0,2;\ 0,2;\ 0,2;\ 0,2]$$
$$\underset{\sim}{\pi}_2 = [0,02;\ 0,09;\ 0,8;\ 0,09;\ 0,02]$$
$$\underset{\sim}{\pi}_3 = [0,4;\ 0,085;\ 0,03;\ 0,085;\ 0,4]$$

Neste caso 5 amostras aleatórias de tamanho 100 de cada uma das distribuições $\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2$ e $\underset{\sim}{\pi}_3$ foram geradas, produzindo então para cada uma das 6 medidas de proximidade, uma matriz de dimensão 15×15.

Para quatro populações, consideramos também somente um caso:

Caso D:

$$\underset{\sim}{\pi}_1 = [0,05;\ 0,05;\ 0,1;\ 0,65;\ 0,15]$$
$$\underset{\sim}{\pi}_2 = [0,45;\ 0,25;\ 0,2;\ 0,05;\ 0,05]$$
$$\underset{\sim}{\pi}_3 = [0,05;\ 0,1;\ 0,2;\ 0,3;\ 0,35]$$
$$\underset{\sim}{\pi}_4 = [0,02;\ 0,09;\ 0,8;\ 0,09;\ 0,02]$$

Aqui, 5 amostras aleatórias de tamanho 100 de cada uma das distribuições $\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, \underset{\sim}{\pi}_3$ e $\underset{\sim}{\pi}_4$ foram geradas, produzindo então para cada uma das 6 medidas de proximidade considerada uma matriz de proximidade de dimensão 20×20.

5.3 - DESCRIÇÃO DOS DADOS

Para cada um dos casos considerados na seção anterior, foram estudadas as duas regras de parada sugeridas

por Mojena (1977), no Capítulo IV, e a medida de ajuste do dendrograma proposta por Ogilvie (1972) Seção 3.6.2 do Capítulo III, segundo os três métodos de agrupamento e as seis medidas de proximidade. Para a primeira regra de parada foram considerados nove valores distintos de k (afastamento padrão), e para a segunda regra de parada foram considerados somente cinco dos nove valores citados, pois aqui também tivemos que tomar alguns valores de n (número de observações consideradas na média móvel).

Para o Caso A apresentamos na Tabela 5.1, os resultados obtidos pela aplicação da primeira regra de parada (Seção 4.2 do Capítulo IV), onde encontramos para cada valor de k, medida de proximidade e método de agrupamento; o número de grupos resultantes, bem como as porcentagens marginais de acertos.

Da Tabela 5.1 podemos observar que, para qualquer valor de k, $1,0 \leq k \leq 3,5$, encontramos o número ótimo de grupos, uma vez que a estrutura inicial é composta de duas populações. Nas tabelas marginais, notamos que entre métodos de agrupamento não existe diferença, e entre medidas de proximidade, S_1 (Similaridade de Horn) foi a que produziu melhor resultado, enquanto que D_3 (Distância de Bhattacharyya) foi a pior. Da Tabela 5.2, onde encontramos os valores do coeficiente de Ogilvie (Seção 3.5 do Capítulo III), vemos que o método de agrupamento que produziu melhor resultado, para as

seis medidas de proximidade consideradas, foi o do encadeamento médio.

A Tabela 5.3 nos mostra o comportamento da segunda regra de parada, segundo as mesmas características apresentadas na Seção 4.3 do Capítulo IV. Nesta tabela, encontramos os valores de:

n: número de observações consideradas para fazer a regressão

k: afastamento padrão

j: estágio de parada

α_{j+1}: valor do critério (nível) no estágio $j+1$

Ng: número de grupos que se obteve no estágio j

Na: casos em que a desigualdade [4.2] não é satisfeita.

Lembremos que esta regra de parada estabelece que devemos parar o processo de agrupamento no primeiro estágio j tal que a desigualdade $\alpha_{j+1} > \overline{\alpha}_j + L_j + b_j + ks_j$ esteja satisfeita (caso em que α é crescente).

TABELA 5.1 - NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA 1ª REGRA DE PARADA AO CASO A DA SEÇÃO 5.2 DO CAPÍTULO V

MÉTODO	MEDIDAS	VALORES DE k								
		0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
ENCADEAMENTO COMPLETO (E.C.)	D_1	7	4	2	2	2	2	2	2	1
	D_2	5	2	2	2	2	2	2	2	2
	D_3	6	4	2	2	2	2	2	2	2
	D_4	4	2	2	2	2	2	2	2	2
	S_1	4	2	2	2	2	2	2	2	2
	S_2	3	2	2	2	2	2	2	2	2
VIZINHO MAIS PRÓXIMO (V.P.)	D_1	3	2	2	2	2	2	2	2	2
	D_2	2	2	2	2	2	2	2	2	1
	D_3	4	3	2	2	2	2	2	2	1
	D_4	4	3	2	2	2	2	2	2	1
	S_1	2	2	2	2	2	2	2	2	2
	S_2	3	2	2	2	2	2	2	2	2
ENCADEAMENTO MÉDIO (E.M.)	D_1	5	3	2	2	2	2	2	2	1
	D_2	5	2	2	2	2	2	2	2	2
	D_3	6	3	2	2	2	2	2	2	1
	D_4	5	2	2	2	2	2	2	2	1
	S_1	3	2	2	2	2	2	2	2	2
	S_2	4	2	2	2	2	2	2	2	2

MEDIDA	% ACERTO
D_1	74,07
D_2	88,88
D_3	70,37
D_4	77,77
S_1	92,59
S_2	88,88

MÉTODO	% ACERTO
E.C.	83,34
V.P.	83,34
E.M.	83,34

TABELA 5.2 - VALORES DO COEFICIENTE DE OGILVIE PARA O CASO A

MEDIDA	MÉTODOS		
	E.C.	V.P.	E.M.
D_1	0,1450	0,1229	0,0227
D_2	0,6112	0,3016	0,0631
D_3	0,1149	0,1405	0,0202
D_4	2,6153	0,6868	0,2239
S_1	0,0145	0,0082	0,0016
S_2	0,0972	0,0116	0,0102

TABELA 5.3 - VALORES $\hat{\alpha}_{j+1}$ E O NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA SEGUNDA REGRA DE PARADA PARA O CASO A DA SEÇÃO 5.2

$n=10$		VALORES DE k							Ng	$n=15$		VALORES DE k							Ng
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}				1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	
E.C.	D_1								Na	E.C.	D_1	0,38	0,43	0,46	0,48	0,51	18	0,55	2
	D_2	0,89	0,95	0,98	1,01	1,05	18	1,60	2		D_2	0,37	0,43	0,47	0,50	0,53	18	1,60	2
	D_3								Na		D_3	0,47	0,53	0,56	0,59	0,62	18	0,71	2
	D_4	0,93	1,01	1,05	1,09	1,13	18	2,30	2		D_4	0,39	0,47	0,51	0,55	0,59	18	2,30	2
	S_1	1,00	0,99	0,99	0,99	0,99	10	0,99	10		S_1	1,00	0,99	0,98	0,98	0,97	15	0,99	5
	S_2								Na		S_2	1,02	1,01	1,00	0,99	0,99	15	0,99	5
V.P.	D_1								Na	V.P.	D_1	0,20	0,21	0,22	0,22	0,23	18	0,28	2
	D_2	0,28	0,28	0,29	0,29	0,30	18	0,45	2		D_2	0,11	0,12	0,13	0,13	0,14	18	0,45	2
	D_3								Na		D_3	0,28	0,30	0,31	0,32	0,33	18	0,35	2
	D_4								Na		D_4	0,10	0,12	0,12	0,13	0,14	18	0,19	2
	S_1	1,00	0,99	0,99	0,99	0,99	10	0,99	10		S_1	1,00	0,99	0,99	0,99	0,98	10	0,99	10
	S_2								Na		S_2								Na
E.M.	D_1								Na	E.M.	D_1	0,30	0,33	0,34	0,36	0,38	18	0,41	2
	D_2	0,56	0,59	0,61	0,63	0,64	18	0,91	2		D_2	0,23	0,26	0,28	0,30	0,32	18	0,91	2
	D_3								Na		D_3	0,38	0,42	0,44	0,46	0,48	18	0,54	2
	D_4	0,55	0,59	0,60	0,62	0,64	18	0,85	2		D_4	0,23	0,27	0,28	0,30	0,32	18	0,85	2
	S_1								Na		S_1	1,00	0,99	0,99	0,98	0,98	15	0,99	5
	S_2	1,02	1,01	1,01	1,00	0,99	10	0,99	10		S_2	1,02	1,01	1,00	1,00	1,00	15	1,00	5

CONTINUAÇÃO DA TABELA 5.3

n=17		VALORES DE k								n=18		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng			1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1	0,28	0,33	0,36	0,38	0,41	18	0,55	2	E.C.	D_1	0,24	0,30	0,33	0,35	0,38	18	0,55	2
	D_2	0,28	0,35	0,38	0,41	0,45	18	1,60	2		D_2	0,25	0,31	0,35	0,38	0,42	18	1,60	2
	D_3	0,34	0,40	0,43	0,46	0,49	18	0,71	2		D_3	0,30	0,37	0,40	0,43	0,46	18	0,71	2
	D_4	0,30	0,38	0,41	0,45	0,49	18	2,30	2		D_4	0,27	0,34	0,38	0,42	0,46	18	2,30	2
	S_1	1,00	0,98	0,98	0,97	0,97	17	0,96	3		S_1	1,00	0,98	0,98	0,97	0,96	18	0,67	2
	S_2	1,02	1,01	1,01	0,99	0,97	17	0,96	3		S_2	1,02	1,01	1,00	0,98	0,97	18	0,85	2
V.P.	D_1	0,14	0,16	0,17	0,17	0,18	18	0,28	2	V.P.	D_1	0,12	0,14	0,15	0,16	0,18	18	0,28	2
	D_2	0,08	0,09	0,10	0,11	0,11	18	0,45	2		D_2	0,07	0,09	0,09	0,10	0,11	18	0,45	2
	D_3	0,20	0,22	0,24	0,25	0,26	18	0,35	2		D_3	0,17	0,20	0,22	0,23	0,25	18	0,35	2
	D_4	0,08	0,09	0,10	0,10	0,11	18	0,19	2		D_4	0,07	0,08	0,09	0,10	0,10	18	0,19	2
	S_1	1,00	0,99	0,99	0,99	0,98	17	0,99	3		S_1	1,00	0,99	0,99	0,99	0,99	18	0,91	2
	S_2	1,02	1,01	1,00	1,00	0,99	17	1,01	3		S_2	1,02	1,01	1,00	0,99	0,98	18	0,96	2
E.M.	D_1	0,21	0,25	0,27	0,28	0,30	18	0,41	2	E.M.	D_1	0,19	0,22	0,24	0,26	0,28	18	0,41	2
	D_2	0,17	0,21	0,23	0,25	0,26	18	0,91	2		D_2	0,15	0,19	0,21	0,23	0,25	18	0,91	2
	D_3	0,27	0,32	0,34	0,36	0,38	18	0,54	2		D_3	0,24	0,29	0,31	0,34	0,36	18	0,54	2
	D_4	0,17	0,21	0,23	0,25	0,27	18	0,82	2		D_4	0,15	0,19	0,21	0,23	0,25	18	0,85	2
	S_1	1,00	0,99	0,99	0,98	0,98	17	0,97	3		S_1	1,00	0,99	0,98	0,98	0,98	18	0,89	2
	S_2	1,02	1,01	1,01	1,00	0,98	17	0,98	3		S_2	1,02	1,01	1,00	0,99	0,99	18	0,84	2

A Tabela 5.4 nos fornece a porcentagem de acertos para os diferentes valores de n.

TABELA 5.4

VALORES DE n	PORCENTAGEM DE ACERTOS
10	27,77
15	66,66
17	66,66
18	100,00

Observamos então, que a previsão do número de grupos melhora com o valor de n crescendo; entretanto, esta é uma dificuldade a mais para o uso desta regra, pois o pesquisador não tem informações sobre quantos pontos usar na média móvel.

Para os casos B, C e D, os resultados de aplicação das regras de parada e os valores do coeficiente de Ogilvie se encontram nas Tabelas Ap.I.1, Ap.I.2, Ap.I.3, Ap.I.4, Ap.I.5 e Ap.I.6 do Apêndice I.

5.4 - ANÁLISE DOS RESULTADOS

Das Tabelas 5.1, 5.2 e 5.3 da seção anterior, e das Tabelas Ap.I.1 a Ap.I.6 do Apêndice I, construímos as Ta

belas 5.5, 5.6, 5.7 e 5.8.

TABELA 5.5 - TABELA DA PORCENTAGEM DE ACERTOS SEGUNDO OS VALORES DE k, OBTIDOS PARA A PRIMEIRA REGRA DE PARADA (FONTE: TABELAS 5.1, AP.I.1, AP.I.3 E AP.I.5)

CASOS	VALORES DE k								
	0,0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
A (2 POP.)	11,11	66,66	100,00	100,00	100,00	100,00	100,00	100,00	61,11
B (2 POP.)	16,66	94,44	100,00	100,00	100,00	100,00	100,00	94,44	61,11
C (3 POP.)	77,77	72,22	5,00	0	0	0	0	0	0
D (4 POP.)	61,11	94,44	61,11	16,66	0	0	0	0	0

Desta tabela observamos que a escolha dos k, que produzem valores corretos para o número de grupos, irá depender do número de populações constituintes da estrutura de dados, e que a medida que este número cresce, os valores de k decrescem.

TABELA 5.6 - COMPORTAMENTO DAS SEIS MEDIDAS DE PROXIMIDADE POR ORDEM DE MAIORES PORCENTAGENS DE ACERTO (FONTE: TABELAS MARGINAIS 5.1, AP.I.1, AP.I.3 e AP.I.5)

ORDEM	CASO A	CASO B	CASO C	CASO D
1º	S_1	S_1	S_1 e S_2	S_2
2º	D_2 e S_2	S_2	D_1 e D_3	D_1
3º	D_4	D_2 e D_4	D_2	D_2, D_3 e S_1
4º	D_1	D_3	D_4	D_4
5º	D_3	D_1		

Com os postos atribuídos a cada medida e cada um dos casos, da Tabela 5.6 temos que a soma dos postos para cada medida é dada por:

D_1: $4+5+2+2 = 13$

D_2: $2+3+3+3 = 11$

D_3: $5+4+2+3 = 14$

D_4: $3+3+4+4 = 14$

S_1: $1+1+1+3 = 6$

S_2: $2+2+1+1 = 6$

Assim, elegemos S_1 e S_2 como as medidas que produziram melhor resultado e D_2 a que produziu pior resultado.

TABELA 5.7 - COMPORTAMENTO DOS MÉTODOS DE AGRUPAMENTO POR ORDEM DE MAIORES PORCENTAGENS DE ACERTO (FONTE: TABELAS MARGINAIS 5.1, AP.I.1, AP.I.3 E AP.I.5)

ORDEM	DADOS			
	CASO A	CASO B	CASO C	CASO D
1º	*E.C., V.P., E.M.*	*V.P.*	*E.M.*	*E.M.*
2º		*E.M., E.C.*	*E.C.*	*E.C.*
3º			*V.P.*	*V.P.*

Da mesma forma como procedemos anteriormente, a soma dos postos para cada um dos três métodos de agrupamento é dada por:

E.C.: 1+2+2+2 = 7

V.P.: 1+1+3+3 = 8

E.M.: 1+2+1+1 = 5

Portanto, o melhor método de agrupamento para os dados em questão, é o do encadeamento médio, e o pior, o método do vizinho mais próximo.

A Tabela 5.8 nos mostra as porcentagens de acertos, para o caso da segunda regra de parada, com os diferentes valores de n para cada um dos casos considerados na Seção 5.2.

TABELA 5.8 - PORCENTAGENS DE ACERTOS PARA CADA CASO, PARA OS DIFERENTES VALORES DE n CONSIDERADOS NAS TABELAS 5.3 DA SEÇÃO 5.3 E AP.I.2, AP.I.4 E AP.I.6 DO APÊNDICE I RESPECTIVAMENTE

DADOS	VALOR DE n	PORCENTAGEM DE ACERTOS
CASO *A* (2 POPULAÇÕES)	10	27,77
	15	66,66
	17	66,66
	18	100,00
CASO *B* (2 POPULAÇÕES)	10	33,33
	15	61,11
	17	66,66
	18	94,44
CASO *C* (3 POPULAÇÕES)	10	55,55
	11	55,55
	12	83,33
CASO *D* (4 POPULAÇÕES)	10	11,11
	12	27,77
	15	44,44

Da Tabela 5.8 observamos que a porcentagem de acertos para a previsão do número de grupos, cresce a medida que n cresce, mas esta porcentagem também decresce a medida que o número de populações fixadas a priori aumenta. Assim sendo, teremos dificuldades em escolher os parâmetros necessários à utilização prática desta regra.

5.5 - COMENTÁRIOS

Embora o número de casos aqui estudados seja pequeno em relação ao número de combinações de distribuições fixadas a priori, que poderiam ser consideradas, já serviu para nos dar uma idéia do comportamento das medidas de proximidade, métodos de agrupamento e regras de parada aqui considerados. Através dos resultados apresentados na Seção 5.4, um caminho que poderá ser seguido como um guia ao usuário interessado em técnicas de agrupamento, com dados caracterizados pela distribuição multinomial com r categorias de resposta, é apresentado:

1) Para medida de proximidade a ordem de preferência é dada por:

1º lugar, usaríamos S_1 e S_2

2º lugar, usaríamos D_1 ou D_3 ou D_4

3º e último lugar, usaríamos D_2.

2) Para método de agrupamento usaríamos:

1º lugar, método do encadeamento médio

2º lugar, método do encadeamento completo

3º e último lugar, método do vizinho mais próximo.

3) Para regra de parada a preferência é dada para a primeira regra, uma vez que a segunda regra apresentará dificuldades na escolha dos parâmetros necessários à sua utilização.

C A P Í T U L O V I

CONCLUSÕES

Neste trabalho procuramos fornecer ao usuário interessado, um guia prático de utilização das técnicas de agrupamento aqui estudadas, cujos dados podem ser considerados como amostras de populações multinomiais com r categorias de resposta. Na Seção 1.2 do Capítulo I, foram levantados três problemas comuns a todas as técnicas de agrupamento, ou sejam:

i) a escolha da medida de proximidade adequada;

ii) a escolha do método de agrupamento adequado;

iii) a determinação do número de grupos.

No Capítulo V, através de simulações, obtivemos alguns resultados, que poderão ser seguidos como sugestão, na tentativa de resolver os problemas levantados. No caso do ítem (i); a escolha da medida de proximidade adequada, a não

ser que o pesquisador tenha razões teóricas fortes pela escolha de uma medida pré determinada, poderá então a sugestão apresentada na Seção 5.5 do Capítulo V, ser seguida.

Neste trabalho apresentamos algumas medidas de proximidade, porém outras poderão ser estudadas, como por exemplo, a medida de Grassle e Smith apresentada no Apêndice II, que trata da ponderação para ocorrência de espécies raras, poderá ser estudada com maior profundidade, uma vez que em trabalhos aplicados na área de Ecologia, estes tipos de dados aparecem com grande frequência. Outras medidas aparecem, veja por exemplo, Hartigan (1967).

Com respeito ao ítem (ii), a escolha do método de agrupamento adequado; neste trabalho nossa atenção esteve voltada para somente três deles, ou seja:

. Método do Encadeamento Completo - E.C.

. Método do Vizinho mais Próximo - V.P.

. Método do Encadeamento Médio - E.M.

Isto porque, como já foi mencionado anteriormente, devido a facilidade em se dispor destes três métodos em programas prontos como: BMDP (Biomedical Computer Programs, SAS (Statistical Analysis System), IMSL (International Mathematical Statistics Library), etc. Pelos resultados apresentados no Capítulo V, vimos que a sugestão de utilização para os três métodos de agrupamento, por ordem de preferência é: E.M., E.C. e V.P. Observamos porém, que do ponto de vista prático, a uti

lização do método do encadeamento médio poderá ser problemático, uma vez que este método de agrupamento exige uma quantidade de memória do computador, bem maior que a necessária para qualquer um dos outros dois métodos. Caso o problema em estudo provoque estouro de memória, então a sugestão será a de se usar o método do encadeamento completo como primeira opção.

Além dos três métodos de agrupamento por nós abordados, outros poderão ser utilizados. Everitt (1974) apresenta o modelo generalizado de Lance e William's (1967), que foi estendido por Wishart em (1969) para incluir o método de Ward's:

$$d_{rs} = a_p d_{ps} + a_q d_{qs} + b d_{pq} + g \left| d_{ps} - d_{qs} \right| \qquad [6.1]$$

onde r representa o índice para o novo grupo obtido pela fusão dos grupos p e q; d representa a medida de proximidade; s representa um outro grupo; a_p, a_q, b e g representam parâmetros cujos valores irão depender do método adotado para se definir novas associações a cada estágio. Para maiores detalhes, vejamos a Tabela 6.1.

A respeito do ítem (iii), determinação do número de grupos, além das duas regras de parada aqui apresentadas, outros caminhos poderão ser seguidos:

. Por exemplo, ao invés de ajustarmos uma reta, como o fizemos no caso da segunda regra de parada, procura

mos aqueles pontos onde haja mudança de estrutura. Consideremos por exemplo, o caso em que o modelo regressivo é composto de dois submodelos; caso mais simples aquele que duas retas se cruzam num determinado ponto, indicando existir aí uma tal mudança. Para desenvolvimento de estudos nesse sentido, vejam Pait (1979).

. Um outro caminho é sugerido por Friedman e Rubin (1967), onde solicita-se a construção do gráfico de número de grupos versus valores do critério de agrupamento, de maneira que a forma da curva indique o número correto de grupos.

TABELA 6.1 - MÉTODOS DE AGRUPAMENTO E RESPECTIVOS VALORES DOS PARÂMETROS DA EQUAÇÃO 6.1

MÉTODO	a_p	a_q	b	g
Vizinho mais Próximo	1/2	1/2	0	-1/2
Encadeamento Completo	1/2	1/2	0	1/2
Encadeamento Médio	1/2	1/2	0	0
Encadeamento Médio Ponderado	n_p/n_r	n_q/n_r	0	0
Mediana	1/2	1/2	-1/4	0
Centróide	n_p/n_r	n_q/n_r	$-n_p.n_q/n_r^2$	0
Ward's	n_s+n_p/n_s+n_r	n_s+n_q/n_s+n_r	$-n_s/n_s+n_r$	0

Dos resultados apresentados no Capítulo V, a sugestão é de que devemos usar a primeira regra de parada, uma

vez que a segunda regra apresenta dificuldades na escolha dos parâmetros necessários a sua utilização.

Em face aos problemas aqui colocados, e das alternativas para tentar solucioná-los, esperamos que o pesquisador interessado possa se utilizar das técnicas apresentadas para facilitar a compreensão e interpretação das observações. Embora Everitt (1979) tenha ponderado, "..., interpretabilidade e simplicidade são importantes na análise de dados, e qualquer inferência rígida do número ótimo de grupos, à luz de valores observados em um índice numérico de bondade de ajustamento, pode ser improdutivo, ..."; devemos ter em mente que o bom emprego do método de agrupamento requer a aliação da técnica numérica com o conhecimento teórico e experimental do pesquisador.

Para finalizar, apresentamos um exemplo, onde os passos sugeridos na Seção 5.5 do Capítulo V são seguidos, isto é, utilizamos respectivamente:

1) Medida de Proximidade: similaridade de Horn, fórmula [2.18];

2) Método de Agrupamento: encadeamento médio;

3) Regra de Parada: primeira regra de parada, Seção 4.2 do Capítulo IV;

4) Medida de Ajuste do Dendrograma: coeficiente de Ogilvie, fórmula [3.3] do Capítulo III.

EXEMPLO 6.1 - Consideremos a matriz de proximidade, dada abaixo, calculada para dezoito amostras de areia, coletadas na região da praia de Caraguatatuba, para estudos sobre o movimento de areia naquela região.

FONTE: Dados apresentados ao SEA, IME-USP, São Paulo, por Olga Cruz, (1979).

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}
A_1	1,000																	
A_2	0,982	1,000																
A_3	0,794	0,765	1,000															
A_4	0,917	0,985	0,795	1,000														
A_5	0,958	0,962	0,638	0,964	1,000													
A_6	0,761	0,714	0,950	0,561	0,619	1,000												
A_7	0,948	0,964	0,753	0,963	0,934	0,711	1,000											
A_8	0,976	0,995	0,847	0,952	0,968	0,676	0,992	1,000										
A_9	0,527	0,563	0,433	0,950	0,905	0,375	0,579	0,579	1,000									
A_{10}	0,961	0,931	0,976	0,873	0,927	0,826	0,965	0,910	0,778	1,000								
A_{11}	0,870	0,865	0,876	0,805	0,900	0,884	0,917	0,837	0,673	0,933	1,000							
A_{12}	0,996	1,000	0,770	1,000	0,966	0,735	0,933	0,971	0,830	0,961	0,846	1,000						
A_{13}	1,000	1,000	0,707	0,875	0,888	0,681	0,876	0,970	0,812	0,872	0,764	0,955	1,000					
A_{14}	0,940	0,941	0,807	0,823	0,865	0,787	0,852	0,934	0,686	0,877	0,800	0,932	0,973	1,000				
A_{15}	0,962	0,975	0,689	0,960	1,000	0,642	1,000	0,974	0,891	0,911	0,884	0,969	0,927	0,871	1,000			
A_{16}	1,000	0,749	0,649	0,558	0,619	0,935	0,714	0,711	0,373	0,783	0,849	0,735	0,712	0,829	0,506	1,000		
A_{17}	0,842	0,847	0,945	0,652	0,733	1,000	0,842	0,838	0,534	0,942	0,937	0,894	0,853	0,943	0,819	0,996	1,000	
A_{18}	0,850	0,806	0,956	0,637	0,717	0,971	0,770	0,772	0,461	0,869	0,880	0,829	0,789	0,930	0,728	0,978	0,945	1,000

Na Figura 6.1, encontramos o dendrograma resultante da aplicação do método do encadeamento médio sobre D.

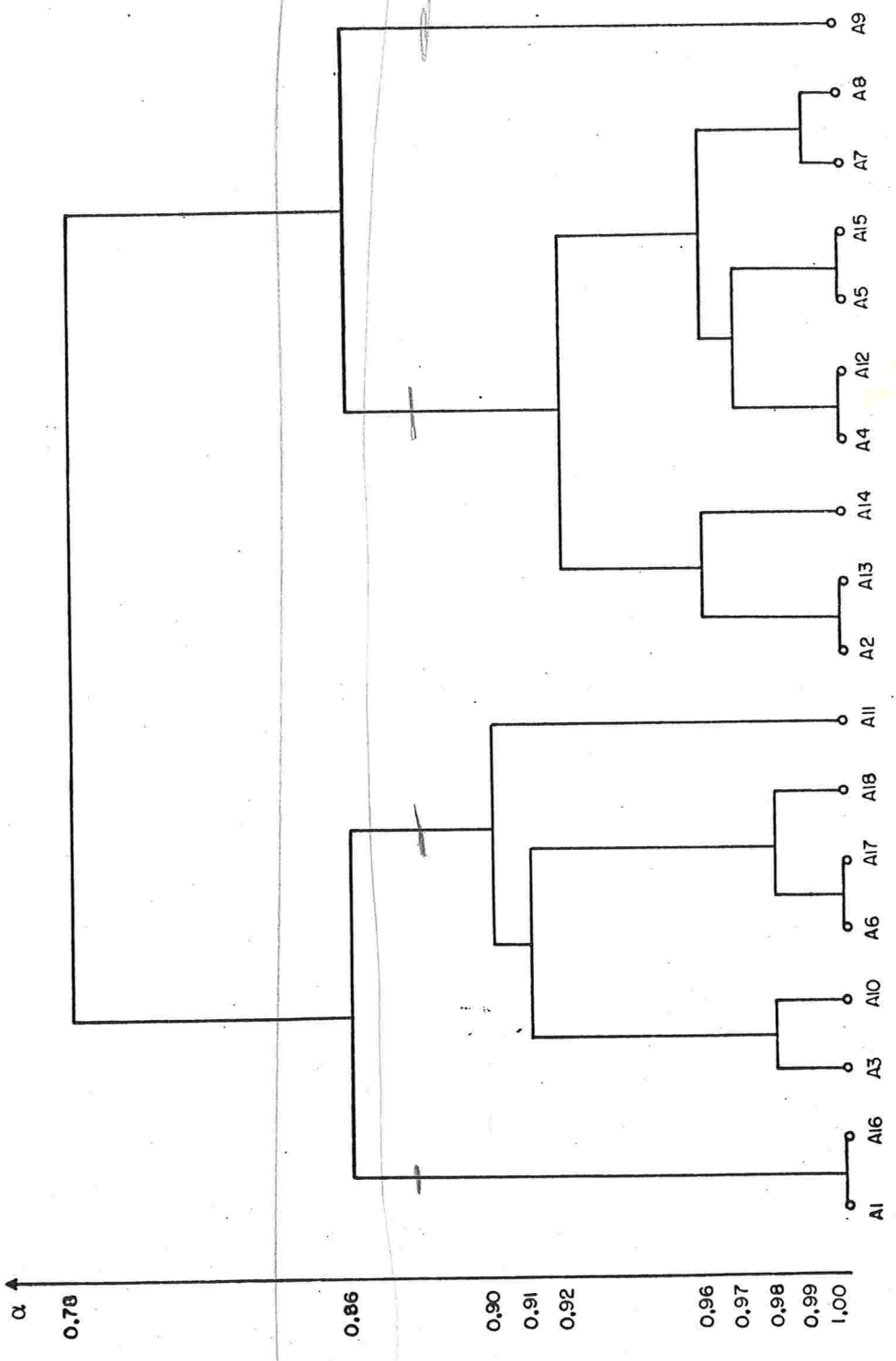


FIGURA 6.1

Para a sequência de níveis α_j, $j=1,\ldots,17$, dada por:

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}
1,00	1,00	1,00	1,00	1,00	0,99	0,98	0,98	0,97	0,96	0,96	0,92

α_{13}	α_{14}	α_{15}	α_{16}	α_{17}
0,91	0,90	0,86	0,86	0,78

Encontramos: $\bar{\alpha} = 0,94$ e $s_\alpha = 0,06$.

A primeira regra de parada estabelece que devemos parar o processo de agrupamento no primeiro estágio j, tal que, a desigualdade $\alpha_{j+1} < \bar{\alpha} - ks_\alpha$ esteja satisfeita; onde

α_{j+1}: valor do nível de agrupamento no estágio $j+1$

$\bar{\alpha}$: média da sequência de níveis α_j, $j=1,\ldots,17$

s_α: desvio padrão da sequência α_j, $j=1,\ldots,17$

k: afastamento padrão.

Na Tabela 6.2, apresentamos os valores de $\hat{\alpha}_{j+1} = \bar{\alpha} - ks_\alpha$, para vários valores de k e também os correspondentes números de grupos para cada caso.

TABELA 6.2 - VALORES DE k, $\hat{\alpha}_{j+1}$ E O NÚMERO DE GRUPOS QUE SE OBTÊM EM CADA CASO

k	0	0,5	1,0	1,5	2,0	2,5	3,0	3,5
$\hat{\alpha}_{j+1} = \bar{\alpha} - ks_\alpha$	0,94	0,91	0,88	0,85	0,82	0,79	0,76	0,73
NÚMERO DE GRUPOS	7	5	4	2	2	2	1	1

Para determinação do número de grupos, basta olharmos o dendrograma da Figura 6.1. Vejamos o caso em que $k=1$. Neste caso $\hat{\alpha}_{j+1}=0,88$, e o primeiro valor de α_{j+1} menor que $0,88$ é $\alpha_{15}=0,86$. Então de acordo com a primeira regra, paramos o processo de agrupamento no décimo quarto estágio, obtendo assim a formação de 4 grupos.

Para o coeficiente de Ogilvie, fórmula [3.3], que mede o ajuste do dendrograma encontramos:

$C_0 = 0,016$.

De acordo com os resultados produzidos pela Tabela 6.2, estabelecemos que o número de grupos é 2, que corresponde aos valores de k; $1,5 \leqq k \leqq 2,5$; e para $C_0=0,016$ observamos ser um valor ótimo de ajuste do dendrograma, uma vez que o campo de variação de C_0 é $C_0 \geqq 0$.

A P Ê N D I C E I

No trabalho de simulação, emprego dos métodos de agrupamento, cálculo do coeficiente de Ogilvie e utilização das regras de parada, contamos com os recursos:

- PDP/11 Mod.45 do ICMSC (Instituto de Ciências Matemáticas de São Carlos, USP).
- IBM/370 Mod.145 da UFSCar (Universidade Federal de São Carlos).

As Tabelas Ap.I.1, Ap.I.3 e Ap.I.5 nos dão os resultados obtidos pelo emprego da primeira regra de parada e cálculo do coeficiente de Ogilvie, e as Tabelas Ap.I.2, Ap.I.4 e Ap.I.6 nos dão os resultados obtidos pelo emprego da segunda regra de parada.

TABELA AP.I.1 - NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA 1ª REGRA DE PARADA AO CASO B DA SEÇÃO 5.2 DO CAPÍTULO IV

MÉTODO	MEDIDAS	VALORES DE k								
		0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
ENCADEAMENTO COMPLETO (E.C.)	D_1	7	4	2	2	2	2	2	1	1
	D_2	6	2	2	2	2	2	2	2	1
	D_3	7	5	2	2	2	2	2	2	2
	D_4	5	2	2	2	2	2	2	2	2
	S_1	4	2	2	2	2	2	2	2	2
	S_2	4	2	2	2	2	2	2	2	1
VIZINHO MAIS PRÓXIMO (V.P.)	D_1	5	3	2	2	2	2	2	2	2
	D_2	4	2	2	2	2	2	2	2	1
	D_3	4	3	2	2	2	2	2	2	2
	D_4	4	4	2	2	2	2	2	2	2
	S_1	2	2	2	2	2	2	2	2	2
	S_2	2	2	2	2	2	2	2	2	2
ENCADEAMENTO MÉDIO (E.M.)	D_1	7	7	3	2	2	2	2	2	1
	D_2	6	2	2	2	2	2	2	2	2
	D_3	6	4	2	2	2	2	2	2	1
	D_4	5	2	2	2	2	2	2	2	1
	S_1	2	2	2	2	2	2	2	2	2
	S_2	3	2	2	2	2	2	2	2	2

MEDIDA	% ACERTO
D_1	62,96
D_2	81,48
D_3	74,07
D_4	81,48
S_1	96,29
S_2	88,88

MÉTODO	% ACERTO
E.C.	79,63
V.P.	85,18
E.M.	79,63

VALORES DO COEFICIENTE DE OGILVIE

MEDIDA	MÉTODOS		
	E.C.	V.P.	E.M.
D_1	0,2181	0,1248	0,0378
D_2	0,5624	0,2712	0,0679
D_3	0,1533	0,1117	0,0244
D_4	0,8568	0,3333	0,0934
S_1	0,0052	0,0035	0,0007
S_2	0,0222	0,0116	0,0025

TABELA AP.I.2 - VALORES $\hat{\alpha}_{j+1}$ E NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA SEGUNDA REGRA DE PARADA AO CASO B (2 POPULAÇÕES) DA SEÇÃO 5.2 DO CAPÍTULO V

$n=10$		VALORES DE k								$n=15$		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng			1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1								Na	E.C.	D_1	0,37	0,42	0,44	0,46	0,49	18	0,47	2
	D_2	0,84	0,90	0,93	0,96	0,99	18	1,00	2		D_2	0,35	0,41	0,44	0,47	0,50	18	1,00	2
	D_3								Na		D_3	0,43	0,51	0,53	0,56	0,59	18	0,54	2
	D_4	0,90	0,96	0,99	1,02	1,05	18	1,32	2		D_4	0,37	0,44	0,47	0,50	0,54	18	0,54	2
	S_1	1,00	0,99	0,99	0,99	0,99	10	0,99	10		S_1	1,00	0,99	0,98	0,98	0,97	15	0,97	5
	S_2	1,03	1,02	1,02	1,01	1,01	10	1,01	10		S_2	1,02	1,01	1,00	1,00	0,99	15	0,98	5
V.P.	D_1								Na	V.P.	D_1	0,21	0,22	0,23	0,23	0,24	19	0,23	1
	D_2	0,26	0,27	0,28	0,28	0,29	18	0,31	2		D_2	0,10	0,11	0,12	0,13	0,13	18	0,31	2
	D_3								Na		D_3	0,24	0,26	0,26	0,27	0,28	18	0,28	2
	D_4	0,27	0,28	0,28	0,29	0,29	18	0,33	2		D_4	0,10	0,12	0,12	0,13	0,14	18	0,33	2
	S_1	1,00	1,00	0,99	0,99	0,99	10	0,99	10		S_1	1,00	0,99	0,99	0,99	0,99	15	0,99	5
	S_2								Na		S_2								Na
E.M.	D_1								Na	E.M.	D_1	0,29	0,32	0,34	0,35	0,36	18	0,33	2
	D_2	0,53	0,56	0,57	0,59	0,60	18	0,58	2		D_2	0,22	0,25	0,26	0,28	0,30	18	0,58	2
	D_3								Na		D_3	0,35	0,38	0,40	0,41	0,43	18	0,40	2
	D_4	0,55	0,57	0,59	0,60	0,62	18	0,68	2		D_4	0,22	0,26	0,27	0,29	0,31	18	0,68	2
	S_1								Na		S_1								Na
	S_2	1,03	1,02	1,02	1,02	1,01	10	0,99	10		S_2	1,02	1,01	1,01	1,00	1,00	15	0,98	5

CONTINUAÇÃO DA TABELA AP.I.2

n=17		VALORES DE k								n=18		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng			1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1	0,27	0,32	0,34	0,36	0,39	18	0,47	2	E.C.	D_1	0,23	0,28	0,30	0,33	0,35	18	0,47	2
	D_2	0,26	0,32	0,35	0,39	0,42	18	1,00	2		D_2	0,23	0,29	0,32	0,35	0,38	18	1,00	2
	D_3	0,32	0,38	0,41	0,43	0,46	18	0,54	2		D_3	0,28	0,33	0,36	0,39	0,42	18	0,54	2
	D_4	0,28	0,35	0,38	0,41	0,45	18	1,32	2		D_4	0,25	0,31	0,35	0,38	0,41	18	1,32	2
	S_1	1,00	0,99	0,98	0,97	0,97	17	0,96	3		S_1	1,00	0,98	0,98	0,97	0,96	18	0,80	2
	S_2	1,02	1,00	0,99	0,98	0,97	17	0,96	3		S_2	1,02	1,00	0,98	0,97	0,96	18	0,63	2
V.P.	D_1	0,14	0,16	0,16	0,17	0,18	18	0,23	2	V.P.	D_1	0,12	0,13	0,13	0,15	0,16	18	0,23	2
	D_2	0,07	0,09	0,09	0,10	0,11	18	0,31	2		D_2	0,06	0,08	0,08	0,09	0,10	18	0,31	2
	D_3	0,16	0,18	0,19	0,20	0,20	18	0,28	2		D_3	0,13	0,15	0,16	0,17	0,18	18	0,28	2
	D_4	0,07	0,09	0,09	0,10	0,11	18	0,33	2		D_4	0,06	0,08	0,08	0,09	0,10	18	0,33	2
	S_1	1,00	0,99	0,99	0,99	0,99	17	0,99	3		S_1	1,00	0,99	0,99	0,99	0,98	18	0,99	2
	S_2	1,02	1,01	1,01	1,00	0,99	17	1,01	3		S_2	1,02	1,01	1,00	0,99	0,99	18	0,92	2
E.M.	D_1	0,21	0,24	0,25	0,27	0,28	18	0,33	2	E.M.	D_1	0,17	0,21	0,22	0,24	0,25	18	0,33	2
	D_2	0,16	0,19	0,21	0,23	0,24	18	0,58	2		D_2	0,14	0,17	0,19	0,21	0,22	18	0,58	2
	D_3	0,24	0,28	0,29	0,31	0,33	18	0,40	2		D_3	0,20	0,24	0,26	0,28	0,29	18	0,40	2
	D_4	0,17	0,20	0,22	0,23	0,25	18	0,68	2		D_4	0,14	0,18	0,20	0,21	0,23	18	0,68	2
	S_1	1,00	0,99	0,99	0,98	0,98	17	0,98	3		S_1	1,00	0,99	0,99	0,98	0,98	17	0,98	3
	S_2	1,02	1,01	1,01	1,00	0,00	17	0,98	3		S_2	1,02	1,01	1,00	0,99	0,99	18	0,80	2

TABELA AP.I.3 - NÚMERO DE GRUPOS OBTIDOS DA APLICAÇÃO DA 1ª REGRA DE PARADA AO CASO C DA SEÇÃO 5.2 DO CAPÍTULO V

MÉTODO	MEDIDAS	VALORES DE k 0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
ENCADEAMENTO COMPLETO (E.C.)	D_1	4	3	2	2	2	2	2	1	1
	D_2	3	3	2	2	2	2	2	1	1
	D_3	4	3	3	2	2	2	2	1	1
	D_4	3	3	2	2	2	2	2	1	1
	S_1	3	3	2	2	2	2	2	1	1
	S_2	3	3	2	2	2	2	2	1	1
VIZINHO MAIS PRÓXIMO (V.P.)	D_1	3	3	2	2	2	2	2	1	1
	D_2	5	4	3	2	2	2	2	1	1
	D_3	6	3	2	2	2	2	2	1	1
	D_4	3	2	2	2	2	2	2	1	1
	S_1	3	3	3	3	2	2	2	1	1
	S_2	3	2	2	2	2	2	2	1	1
ENCADEAMENTO MÉDIO (E.M.)	D_1	3	3	3	2	2	2	2	1	1
	D_2	3	3	2	2	2	2	2	1	1
	D_3	3	3	3	2	2	2	2	1	1
	D_4	3	2	2	2	1	1	1	1	1
	S_1	3	2	2	2	2	2	2	1	1
	S_2	3	3	3	3	2	2	2	1	1

MEDIDA	% ACERTO
D_1	22,22
D_2	18,51
D_3	22,22
D_4	14,81
S_1	25,92
S_2	25,92

MÉTODO	% ACERTO
E.C.	20,37
V.P.	16,66
E.M.	25,92

VALORES DO COEFICIENTE DE OGILVIE

MEDIDA	MÉTODOS E.C.	V.P.	E.M.
D_1	0,0865	0,1154	0,0348
D_2	0,2597	0,2438	0,0825
D_3	0,1409	0,1180	0,0422
D_4	0,9843	0,4506	0,1911
S_1	0,0871	0,0707	0,0258
S_2	0,0982	0,1443	0,1071

TABELA AP.I.4 - VALORES $\hat{\alpha}_{j+1}$ E O NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA SEGUNDA REGRA DE PARADA AO CASO C (3 POPULAÇÕES) DA SEÇÃO 5.2 DO CAPÍTULO V

$n=10$		VALORES DE k								$n=11$		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng			1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1	0,41	0,47	0,49	0,52	0,54	12	0,43	3	E.C.	D_1	0,32	0,38	0,41	0,43	0,46	12	0,43	3
	D_2	0,33	0,38	0,41	0,44	0,46	12	0,91	3		D_2	0,26	0,32	0,34	0,37	0,40	12	0,91	3
	D_3	0,49	0,55	0,58	0,60	0,63	12	0,59	3		D_3	0,38	0,44	0,47	0,49	0,52	12	0,59	3
	D_4	0,34	0,39	0,42	0,45	0,48	12	1,26	3		D_4	0,27	0,33	0,35	0,38	0,41	12	1,26	3
	S_1	1,00	0,99	0,98	0,98	0,98	10	0,98	5		S_1	0,99	0,99	0,98	0,98	0,97	11	0,96	4
	S_2	1,00	0,99	0,99	0,99	0,98	10	0,98	5		S_2	1,00	0,99	0,99	0,98	0,98	11	0,94	4
V.P.	D_1	0,41	0,45	0,47	0,49	0,51	13	0,57	2	V.P.	D_1	0,32	0,36	0,38	0,41	0,43	13	0,57	2
	D_2	0,19	0,21	0,23	0,24	0,25	12	0,22	3		D_2	0,15	0,17	0,19	0,20	0,21	12	0,22	3
	D_3	0,51	0,56	0,59	0,61	0,64	13	0,66	2		D_3	0,40	0,45	0,47	0,50	0,52	13	0,66	2
	D_4	0,14	0,15	0,16	0,17	0,18	12	0,23	2		D_4	0,10	0,12	0,13	0,14	0,15	12	0,23	3
	S_1								Na		S_1								Na
	S_2	1,02	1,01	1,00	0,99	0,99	10	1,00	5		S_2	1,01	1,00	1,00	0,99	0,99	11	1,00	4
E.M.	D_1	0,32	0,35	0,36	0,38	0,39	12	0,38	3	E.M.	D_1	0,25	0,27	0,29	0,30	0,32	12	0,78	3
	D_2	0,24	0,27	0,28	0,29	0,31	12	0,78	3		D_2	0,19	0,22	0,23	0,24	0,26	12	0,78	3
	D_3	0,38	0,40	0,42	0,43	0,44	12	0,46	3		D_3	0,28	0,31	0,32	0,34	0,35	12	0,46	3
	D_4	0,26	0,29	0,30	0,32	0,33	12	0,92	3		D_4	0,20	0,23	0,25	0,26	0,28	12	0,92	3
	S_1	0,99	0,99	0,98	0,98	0,98	10	0,98	5		S_1	0,99	0,99	0,98	0,98	0,98	11	0,98	4
	S_2	1,01	1,00	1,00	0,99	0,99	10	0,99	5		S_2	1,01	1,00	1,00	0,99	0,98	11	0,98	4

CONTINUAÇÃO DA TABELA AP.I.4

$n=12$		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1	0,26	0,32	0,35	0,38	0,41	12	0,43	3
	D_2	0,75	1,00	1,12	1,25	1,37	13	3,18	2
	D_3	0,30	0,37	0,40	0,43	0,46	12	0,59	3
	D_4	0,22	0,28	0,31	0,34	0,37	12	1,26	3
	S_1	0,99	0,98	0,98	0,97	0,97	12	0,80	3
	S_2	0,99	0,98	0,97	0,96	0,95	12	0,69	3
V.P.	D_1	0,17	0,20	0,22	0,23	0,25	12	0,20	3
	D_2	0,12	0,15	0,16	0,18	0,19	12	0,22	3
	D_3	0,31	0,36	0,39	0,41	0,43	13	0,66	2
	D_4	0,09	0,11	0,12	0,13	0,14	12	0,23	3
	S_1								Na
	S_2	1,01	1,00	1,00	0,99	0,99	12	0,95	3
E.M.	D_1	0,19	0,22	0,23	0,25	0,27	12	0,38	3
	D_2	0,15	0,18	0,19	0,21	0,22	12	0,78	3
	D_3	0,21	0,24	0,26	0,27	0,29	12	0,46	3
	D_4	0,16	0,19	0,21	0,23	0,24	12	0,92	3
	S_1	0,99	0,98	0,98	0,98	0,98	12	0,85	3
	S_2	1,01	1,00	0,99	0,99	0,98	12	0,74	3

TABELA AP.I.5 - NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA 1ª REGRA DE PARADA AO CASO D DA SEÇÃO 5.2 DO CAPÍTULO V

MÉTODO	MEDIDAS	VALORES DE k 0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
ENCADEAMENTO COMPLETO (E.C.)	D_1	5	4	4	3	2	1	1	1	1
	D_2	4	4	3	3	3	2	2	1	1
	D_3	5	4	4	3	3	2	1	1	1
	D_4	4	4	3	3	3	2	2	1	1
	S_1	4	4	3	3	3	2	1	1	1
	S_2	4	4	4	3	3	2	1	1	1
VIZINHO MAIS PRÓXIMO (V.P.)	D_1	5	4	4	4	3	2	1	1	1
	D_2	4	4	4	3	3	2	1	1	1
	D_3	5	4	4	4	3	2	1	1	1
	D_4	6	3	3	3	3	2	1	1	1
	S_1	4	4	4	3	3	2	2	1	1
	S_2	4	4	4	3	3	2	1	1	1
ENCADEAMENTO MÉDIO (E.M.)	D_1	5	4	4	4	3	3	2	1	1
	D_2	4	4	3	3	3	1	1	1	1
	D_3	5	4	4	3	3	2	1	1	1
	D_4	4	4	3	3	3	2	2	1	1
	S_1	4	4	3	3	2	2	2	1	1
	S_2	4	4	4	3	3	2	2	1	1

MEDIDA	% ACERTO
D_1	29,63
D_2	25,92
D_3	25,92
D_4	14,81
S_1	25,92
S_2	33,33

MÉTODO	% ACERTO
E.C.	24,07
V.P.	22,22
E.M.	25,92

MEDIDA	MÉTODOS E.C.	V.P.	E.M.
D_1	0,1199	0,1142	0,0247
D_2	0,2240	0,2521	0,0751
D_3	0,1933	0,0826	0,0179
D_4	0,7615	0,6894	0,1380
S_1	0,1408	0,8143	0,0198
S_2	0,1477	0,7948	0,0340

TABELA AP.I.6 - VALORES $\hat{\alpha}_{j+1}$ E O NÚMERO DE GRUPOS OBTIDOS PELA APLICAÇÃO DA SEGUNDA REGRA DE PARADA AO CASO D (4 POPULAÇÕES) DA SEÇÃO 5.2 DO CAPÍTULO V

$n=10$		VALORES DE k								$n=12$		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng			1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1								Na	E.C.	D_1	0,59	0,65	0,68	0,70	0,73	16	0,64	4
	D_2	2,20	2,37	2,45	2,54	2,62	18	3,26	2		D_2	1,50	1,66	1,75	1,83	1,92	18	3,26	2
	D_3								Na		D_3								Na
	D_4	1,14	1,26	1,33	1,39	1,45	16	1,20	4		D_4	0,77	0,89	0,96	1,02	1,08	16	1,20	4
	S_1	0,99	0,98	0,98	0,97	0,97	10	0,97	10		S_1	0,99	0,98	0,97	0,97	0,96	12	0,95	8
	S_2	1,02	1,01	1,01	1,00	1,00	10	1,01	10		S_2	1,02	1,01	1,01	1,00	1,00	12	1,00	8
V.P.	D_1								Na	V.P.	D_1	0,45	0,49	0,51	0,52	0,54	16	0,46	4
	D_2								Na		D_2	0,52	0,56	0,58	0,60	0,62	17	0,69	3
	D_3								Na		D_3								Na
	D_4								Na		D_4	0,43	0,49	0,51	0,54	0,57	17	0,50	3
	S_1	0,99	0,98	0,98	0,98	0,98	10	0,98	10		S_1	0,99	0,98	0,98	0,98	0,97	12	0,98	8
	S_2	1,03	1,02	1,01	1,00	1,00	10	1,01	10		S_2	1,02	1,01	1,01	1,00	1,00	12	1,00	8
E.M.	D_1								Na	E.M.	D_1								Na
	D_2								Na		D_2	0,66	0,73	0,77	0,81	0,85	16	0,99	4
	D_3								Na		D_3								Na
	D_4	0,80	0,88	0,92	0,95	0,99	16	0,85	4		D_4	0,54	0,62	0,66	0,70	0,74	16	0,85	4
	S_1	0,99	0,98	0,98	0,97	0,97	10	0,97	10		S_1	0,99	0,98	0,98	0,97	0,97	10	0,96	10
	S_2	1,02	1,01	1,01	1,00	1,00	10	1,01	10		S_2	1,02	1,01	1,01	1,00	1,00	10	1,00	10

CONTINUAÇÃO DA TABELA AP.I.6

$n=15$		VALORES DE k							
		1,0	2,0	2,5	3,0	3,5	j	α_{j+1}	Ng
E.C.	D_1	0,34	0,40	0,43	0,46	0,49	16	0,64	4
	D_2	0,92	1,08	1,17	1,25	1,34	18	3,26	2
	D_3	0,83	0,98	1,05	1,12	1,19	17	1,03	3
	D_4	0,48	0,60	0,66	0,72	0,78	16	1,20	4
	S_1	0,99	0,97	0,95	0,94	0,93	15	0,92	5
	S_2	1,02	1,00	0,99	0,98	0,97	15	0,94	5
V.P.	D_1	0,25	0,29	0,32	0,34	0,36	16	0,46	4
	D_2	0,30	0,35	0,37	0,40	0,42	17	0,69	3
	D_3	0,37	0,42	0,44	0,47	0,49	16	0,45	4
	D_4	0,26	0,32	0,34	0,37	0,39	17	0,50	3
	S_1	0,99	0,98	0,98	0,97	0,97	15	0,96	5
	S_2	1,02	1,01	1,00	1,00	0,99	15	0,99	5
E.M.	D_1	0,29	0,34	0,37	0,39	0,42	16	0,55	4
	D_2	0,39	0,47	0,51	0,56	0,60	16	0,99	4
	D_3	0,44	0,52	0,55	0,59	0,63	16	0,56	4
	D_4	0,33	0,41	0,45	0,49	0,53	16	0,85	4
	S_1	1,00	1,00	1,00	0,99	0,99	15	0,98	5
	S_2	1,02	1,00	1,00	0,99	0,98	15	0,97	5

A P Ê N D I C E I I

UMA MEDIDA DE SIMILARIDADE SENSÍVEL À CONTRIBUIÇÃO DAS ESPÉCIES RARAS

1. INTRODUÇÃO

Em Ecologia, medidas quantitativas de similaridade entre duas populações tem sido de grande interesse e, a preocupação em se levar em conta a contribuição das espécies raras e espécies dominantes têm levado os pesquisadores a sugerirem várias medidas. Por exemplo têm sido usadas as medidas propostas por Morisita (1959), Horn (1966), coeficiente de correlação de Pearson, etc. Acontece porém, que do ponto de vista de aplicação prática, se tem notado que estas medidas são altamente dependentes das categorias dominantes, levando muitas vezes, a resultados indesejáveis. Com o objetivo então, em se dar maior importância às espécies raras é que Grassle e Smith (1976) propuseram uma família de medidas de similaridade baseadas no número de espécies que participam de amostras de tamanho m fixado, selecionadas aleatoriamente de cada uma das populações.

Assim, se as amostras são selecionadas aleatoriamente de cada população, o número de espécies que participam das mesmas variam com cada par selecionado, e se o experimento for repetido muitas vezes, o número médio de espécies que participam das amostras, pela Lei dos Grandes Números, aproximar-se-á de um número fixado, que formará a base da medida proposta. A contribuição de cada espécie para esta medida é determinada pela probabilidade de que ela apareça na amostra de tamanho m. Para valores pequenos de m, a medida é fortemente influenciada pelas espécies dominantes, mas ao passo que m cresce, as espécies raras passam a dar maior contribuição.

Para desenvolvimento da família de medidas propostas por Grassle e Smith, necessitamos de conhecer o conceito de diversidade das espécies, que é dado a seguir.

2. MEDIDA DE DIVERSIDADE DAS ESPÉCIES

Na literatura, uma medida que envolve o conceito de espécies raras e espécies dominantes é conhecida por medida de diversidade das espécies. Enorme confusão tem ocorrido entre pesquisadores a respeito de tal medida, em face da multiplicidade de definições criadas em torno do conceito; veja Peet (1974). Estaremos aqui, usando o conceito dado

por Smith e Grassle (1977).

Uma das principais medidas de diversidade é o conhecido índice de Simpson, dado por [2.6]. Na área Biológica, críticas tem surgido a respeito desta medida, pois a contribuição de cada espécie é dada pela probabilidade de que a mesma apareça numa amostra de tamanho dois. Então, é claro que esta medida é fortemente influenciada pelas espécies dominantes. A medida de Morisita dada por [2.7] também leva sua crítica em razão do mesmo fato.

Uma outra medida cujas propriedades amostrais tem sido revista por Bowman, Hutcheson, Odum e Shenton (1969), é a medida de informação estatística de Shanon-Wiener dada por [2.12], onde conseguiram mostrar que em média a informação estatística estimada através das amostras subestima a verdadeira informação populacional. Isto quer dizer, a estimativa da medida é tendenciosa. A tendenciosidade depende do tamanho da amostra, mas a medida que esta cresce, a tendenciosidade diminui.

A medida de Shanon-Wiener também não deixou de ser fortemente influenciada pelas espécies dominantes, e por consequência também o será a de Horn dada por [2.13].

Com o intuito de minimizar a influência das espécies dominantes e estudar as propriedades amostrais é que Hurlbert (1971), propôs uma generalização do índice de Simpson. A idéia de Hurlbert foi a de se usar o número espe

rado de espécies em uma amostra de m indivíduos como medida de diversidade das espécies.

Suponha que tenhamos uma população finita consistindo de r espécies com N_i indivíduos da espécie i. O vetor $\underset{\sim}{N}=(N_1,N_2,\ldots,N_r)$ com $\sum_{i=1}^{r} N_i=N$, representa a população toda. A variável aleatória S denotará o número de espécies na amostra de m indivíduos.

O número esperado de espécies em uma amostra aleatória sem reposição, dada a população finita $\underset{\sim}{N}$ é dado por:

$$E[S/N] = \sum_{i=1}^{r} [1 - \frac{\binom{N-N_i}{m}}{\binom{N}{m}}] \qquad \text{[Ap.II.1]}$$

Fazendo uma extensão à idéia de Hurlbert, suponhamos que temos uma população multinomial onde π_i é a proporção de indivíduos da espécie i na população e $\underset{\sim}{\pi}=(\pi_1,\pi_2,\ldots,\pi_r)$ o vetor que descreve a população. Para o modelo de amostragem, vamos supor que a população seja infinita, onde a identificação de cada indivíduo amostrado passa a independer de qualquer outro que seja amostrado.

Sob este regime de amostragem, o número esperado de espécies em uma amostra de m indivíduos, é dado por:

$$s(m) = E[S|\underset{\sim}{\pi}] = \sum_{i=1}^{r} [1-(1-\pi_i)^m] \qquad \text{[Ap.II.2]}$$

Para $m=2$, $s(2) = 2-\sum_{i=1}^{r} \pi_i^2$ e a diversidade de Simpson é igual a $\sum_{i=1}^{r} \pi_i^2$.

Para pequenos valores de m, $s(m)$ enfatiza as espécies mais abundantes, ao passo que com o crescimento de m as espécies raras também passam a contribuir mais para o índice. Para m grande, obviamente que $s(m)$ se aproxima do número de espécies presente na população.

A grande vantagem prática que a medida de Hurlbert e o índice de Simpson têm sobre outras medidas de diversidade, é que ambas têm estimadores não tendenciosos de variância mínima.

Vejamos o caso da medida de Hurlbert:

Suponha que tenhamos uma amostra de n indivíduos provenientes da população multinomial $\underset{\sim}{\pi}$, com n_i indivíduos da espécie i. Seja $\underset{\sim}{n}=(n_1,n_2,\ldots,n_r)$ o vetor que descreve a amostra aleatória.

Através da amostra desejamos estimar $s(m)$, o número esperado de espécies na amostra de m indivíduos provenientes da população.

O estimador não tendencioso de variância mínima é dado por:

$$\hat{s}(m) = E[S|\underset{\sim}{n}] = \sum_{i=1}^{r} [1 - \frac{\binom{n-n_i}{m}}{\binom{n}{m}}] \qquad \text{[Ap.II.3]}$$

onde $\sum_{i=1}^{r} n_i = n$.

De fato:

Pelo teorema de Rao Blackwell (Fraser 1958, pag. 220), se $\hat{\theta}$ é um estimador não tendencioso do parâmetro θ, para encontrarmos o estimador não tendencioso de variância mínima de θ basta encontrarmos o valor esperado de $\hat{\theta}$ condicionado a uma estatística suficiente e completa. No caso em questão, S é um estimador não tendencioso de $s(m)$ e $\underset{\sim}{n}=(n_1,\ldots,n_r)$ é uma estatística suficiente e completa. Portanto, o estimador não tendencioso de variância mínima para o número esperado de espécies em uma amostra aleatória proveniente da população é:

$$E[S|\underset{\sim}{n}] = \sum_{i=1}^{r} [1 - \frac{\binom{n-n_i}{m}}{\binom{n}{m}}]$$

3. MEDIDA DE SIMILARIDADE SENSÍVEL A CONTRIBUIÇÃO DAS ESPÉCIES RARAS (GRASSLE E SMITH)

Consideremos uma população multinomial $\underset{\sim}{\pi}$, com r espécies, ou seja, com r categorias de resposta onde a proporção dos elementos da *i-ésima* categoria é denotada por π_i com

$$\sum_{i=1}^{r} \pi_i = 1, \quad \underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_r).$$

Vamos agora supor que tenhamos duas populações multinomiais $\underset{\sim}{\pi}_1$ e $\underset{\sim}{\pi}_2$; com as mesmas características da população $\underset{\sim}{\pi}$ acima descrita. De cada uma das populações, vamos selecionar uma amostra aleatória de tamanho m.

Seja X a variável aleatória que denota o número de espécies que as duas amostras têm em comum. Então a expectância de X é dada por:

$$E[X|\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m] = \sum_{i=1}^{r} [1-(1-\pi_{1i})^m][1-(1-\pi_{2i})^m] \qquad \text{[Ap.II.4]}$$

De fato:

Seja

$$\delta_{1i} = \begin{cases} 1 \text{ se ocorre a } i\text{-ésima espécie na amostra da população} \\ \underset{\sim}{\pi}_1 \text{ com probabilidade } [1-(1-\pi_{1i})^m] \\ 0 \text{ caso contrário} \end{cases}$$

$$\delta_{2i} = \begin{cases} 1 \text{ se ocorre a } i\text{-ésima espécie na amostra da população} \\ \underset{\sim}{\pi}_2 \text{ com probabilidade } [1-(1-\pi_{2i})^m] \\ 0 \text{ caso contrário} \end{cases}$$

O número de espécies que as duas amostras tem em comum X, pode ser dado por:

$$X = \sum_{i=1}^{r} \delta_{1i}\delta_{2i}$$

então

$$E[X] = E[X|\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m] = \sum_{i=1}^{r} [1-(1-\pi_{1i})^m][1-(1-\pi_{2i})^m].$$

Para construção de uma medida cuja variação esteja restrita ao intervalo $[0;1]$, necessitamos então de um fator de normalização. Para isto, considere então, o problema de selecionarmos duas amostras de tamanho m de uma mesma população $\underset{\sim}{\pi}$ e, determinemos o número esperado de espécies em comum.

Este número é dado por:

$$E[\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_1, m] = \sum_{i=1}^{r} [1-(1-\pi_i)^m]^2 \qquad \text{[Ap.II.5]}$$

usando a média aritmética encontramos:

$$\{E[\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_1, m] + E[\underset{\sim}{\pi}_2, \underset{\sim}{\pi}_2, m]\}/2 \qquad \text{[Ap.II.6]}$$

Por [Ap.II.4], [Ap.II.5] e [Ap.II.6], define-se a medida de similaridade $N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$ dada por:

$$N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m) = \frac{2E[X|\pi_1, \pi_2, m]}{E[\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_1, m] + E[\underset{\sim}{\pi}_2, \underset{\sim}{\pi}_2, m]} \qquad \text{[Ap.II.7]}$$

$$N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m) = \frac{2\sum_{i=1}^{r} [1-(1-\pi_{1i})^m][1-(1-\pi_{2i})^m]}{\sum_{i=1}^{r} [1-(1-\pi_{1i})^m]^2 + \sum_{i=1}^{r} [1-(1-\pi_{2i})^m]^2}$$

Desta forma [Ap.II.7] fica restrito ao intervalo $[0,1]$. De fato:

$$\sum_{i=1}^{r} [1-(1-\pi_{1i})^m][1-(1-\pi_{2i})^m] \leq$$

$$\left\{ \sum_{i=1}^{r} [1-(1-\pi_{1i})^m]^2 \sum_{i=1}^{r} [1-(1-\pi_{2i})^m]^2 \right\}^{1/2} \leq$$

$$\frac{1}{2} \left\{ \sum_{i=1}^{r} [1-(1-\pi_{1i})^m]^2 + \sum_{i=1}^{r} [1-(1-\pi_{2i})^m]^2 \right\}$$

As desigualdades são derivadas através de Holders e a desigualdade entre a média geométrica e a média aritmética. (Abramowitz e Stegum, 1972, pag. 10-11). $N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)=1$ somente quando $\underset{\sim}{\pi}_1 = \underset{\sim}{\pi}_2$.

Do ponto de vista prático, as coisas ficam um pouco mais complicada, pois não conhecemos os parâmetros populacionais $\underset{\sim}{\pi}$ e o que se faz é estimar $N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$.

Seja então $\underset{\sim}{n}_1$ uma amostra aleatória de $\underset{\sim}{\pi}_1$, e n_{1i} a variável aleatória que denota o número de elementos pertencente à *i-ésima* categoria. Da mesma forma, seja $\underset{\sim}{n}_2$, uma amostra aleatória da população $\underset{\sim}{\pi}_2$ e n_{2i} a variável aleatória que denota o número de elementos pertencente à *i-ésima* categoria, e

$$\sum_{i=1}^{r} n_{1i} = n_1 \; ; \qquad \sum_{i=1}^{r} n_{2i} = n_2 .$$

Ao selecionarmos subamostras de tamanho m, sem reposição, das amostras maiores $\underset{\sim}{n}_1$ e $\underset{\sim}{n}_2$, iremos obter como estimador não tendencioso de variância mínima do número esperado de espécies que as subamostras têm em comum

$$\hat{E}[X|\underset{\sim}{n}_1,\underset{\sim}{n}_2] = \sum_{i=1}^{r} [1 - \frac{\binom{n_1-n_{1i}}{m}}{\binom{n_1}{m}}][1 - \frac{\binom{n_2-n_{2i}}{m}}{\binom{n_2}{m}}] \qquad \text{[Ap.II.8]}$$

O estimador não tendencioso de variância mínima para o número esperado de espécies em comum em duas subamostras disjuntas da mesma amostra finita $\underset{\sim}{n}$ é dado por:

$$\hat{E}[\underset{\sim}{n},\underset{\sim}{n},m] = \sum_{i=1}^{r} [1-2\frac{\binom{n-n_i}{m}}{\binom{n}{m}} + \frac{\binom{n-n_i}{m}}{\binom{n}{2m}}] \qquad \text{[Ap.II.9]}$$

Com [Ap.II.8] e [Ap.II.9], podemos construir um estimador para $N(\underset{\sim}{\pi}_1,\underset{\sim}{\pi}_2,m)$, denotado por $\hat{N}(\underset{\sim}{\pi}_1,\underset{\sim}{\pi}_2,m)$

$$\hat{N}(\underset{\sim}{\pi}_1,\underset{\sim}{\pi}_2,m) = \frac{2\hat{E}[X|\underset{\sim}{n}_1,\underset{\sim}{n}_2,m]}{\hat{E}[\underset{\sim}{n}_1,\underset{\sim}{n}_1,m]+\hat{E}[\underset{\sim}{n}_2,\underset{\sim}{n}_2,m]} \qquad \text{[Ap.II.10]}$$

$$\hat{N}(\underset{\sim}{\pi}_1,\underset{\sim}{\pi}_2,m) = \frac{2\sum_{i=1}^{r}[1-\binom{n_1-n_{1i}}{m}/\binom{n_1}{m}][1-\binom{n_2-n_{2i}}{m}/\binom{n_2}{m}]}{\sum_{i=1}^{r}[1-2\frac{\binom{n_1-n_{1i}}{m}}{\binom{n_1}{m}}+\frac{\binom{n_1-n_{1i}}{m}}{\binom{n_1}{2m}}]+\sum_{i=1}^{r}[1-2\frac{\binom{n_2-n_{2i}}{m}}{\binom{n_2}{m}}+\frac{\binom{n_2-n_{2i}}{2m}}{\binom{n_2}{2m}}}$$

[Ap.II.11]

OBSERVAÇÕES:

1) Embora para o numerador e denominador de [Ap.II.7] encontramos estimadores não tendenciosos de variância mínima, $\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$ dado por [Ap.II.10] é tendencioso; porém cabe notarmos que quando comparado com outros índices como "Percent Similarity" (Whittaker, 1952), medida de Horn (1966), Metrica de Camberra (Sepkoski, 1974, pág. 147; Williams, 1973) $\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$ tem se mostrado mais eficiente mesmo para pequenos valores de m.

2) Para $m=1$ temos:

i) $$\hat{E}[\underset{\sim}{n}, \underset{\sim}{n}, 1] = \sum_{i=1}^{r} \frac{n_i(n_{i-1})}{n(n-1)} = \lambda \quad \text{(Índice de Simpson)}$$

ii) $$2\hat{E}[X|\underset{\sim}{n}_1, \underset{\sim}{n}_2, 1] = 2\sum_{i=1}^{r} \frac{n_{1i}}{n_1} \cdot \frac{n_{2i}}{n_2}$$

Por (i) e (ii) encontramos:

$$\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, 1) = \frac{2\sum_{i=1}^{r} n_{1i}n_{2i}}{(\lambda_1+\lambda_2)n_1n_2}$$

similaridade de Morisita dada em [2.8].

Para elucidação da medida de Grassle e Smith consideremos o exemplo a seguir:

EXEMPLO AP.II.1 - Consideremos os dados da Tabela Ap.II.1 referentes a amostras coletadas nas regiões

da Base Aérea de Santos e no cais da Alamoa. Estas amostras consistiam de lamínulas, onde para cada lamínula foi contado o número de indivíduos das diferentes espécies de ciliados sésseis.

TABELA AP.II.1

ESPÉCIES	AMOSTRAS											
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
	Mai. 1979	Jun. 1979	Ago. 1979	Out. 1979	Dez. 1979	Jan. 1980	Mai. 1979	Jun. 1979	Ago. 1979	Out. 1979	Dez. 1970	Jan. 1980
Ephlota Gemmipara	1290	404	48	4040	2180	4830	2880	916	302	13320	3820	5
Podophrya Sp A	0	0	0	1	4	0	0	0	0	0	0	0
Acineta Tuberosa	3	17	13	244	2412	7	3	0	0	13	289	4185
Zoothamnium Commune	152	187	474	921	522	2560	11	21	15	70	162	1110
Zoothamnium Spec	88	0	25	1303	0	0	0	0	0	3	4	0
Vorticella Nebulifera	12	25	11	16	4	366	5	0	0	4	48	21
Vorticella Sp C	58	2	0	3	55	6	0	0	0	0	0	0
Cothurna Maritima	91	104	18	42	15	85	0	0	2	1	5	13
Cothurna Cf. Cordylo Phorae	0	0	0	1	1	0	1	0	0	0	0	0
Pyxicola Socialis	0	0	0	1	0	0	0	0	0	0	1	0
Vaginicola Crystallina	0	0	0	5	71	64	0	0	0	0	0	0
TOTAL EM TODA LAMÍNULA	1694	711	589	6577	5564	7918	2900	937	319	13411	4340	5334
	Região da Alamoa						Região da Base Aérea de Santos					

FONTE: Dados apresentados ao SEA, IME-USP, São Paulo (1979).

Usando [Ap.II.11] para $m=1$ encontramos a matriz de similaridade $GS(1)$ dada por:

$$GS(1) = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \\ A_8 \\ A_9 \\ A_{10} \\ A_{11} \\ A_{12} \end{array} \begin{array}{c} \begin{array}{cccccccccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} & A_{11} & A_{12} \end{array} \\ \begin{bmatrix} 1{,}000 & & & & & & & & & & & \\ 0{,}925 & 1{,}000 & & & & & & & & & & \\ 0{,}220 & 0{,}469 & 1{,}000 & & & & & & & & & \\ 0{,}951 & 0{,}910 & 0{,}314 & 1{,}000 & & & & & & & & \\ 0{,}625 & 0{,}700 & 0{,}314 & 0{,}692 & 1{,}000 & & & & & & & \\ 0{,}921 & 0{,}973 & 0{,}547 & 0{,}917 & 0{,}682 & 1{,}000 & & & & & & \\ 0{,}957 & 0{,}807 & 0{,}102 & 0{,}857 & 0{,}578 & 0{,}913 & 1{,}000 & & & & & \\ 0{,}962 & 0{,}822 & 0{,}120 & 0{,}866 & 0{,}585 & 0{,}841 & 0{,}999 & 1{,}000 & & & & \\ 0{,}975 & 0{,}839 & 0{,}148 & 0{,}880 & 0{,}599 & 0{,}875 & 0{,}998 & 0{,}999 & 1{,}000 & & & \\ 0{,}957 & 0{,}080 & 0{,}103 & 0{,}858 & 0{,}578 & 0{,}829 & 1{,}000 & 0{,}999 & 0{,}998 & 1{,}000 & & \\ 0{,}980 & 0{,}857 & 0{,}143 & 0{,}900 & 0{,}663 & 0{,}872 & 0{,}989 & 0{,}991 & 0{,}994 & 0{,}989 & 1{,}000 & \\ 0{,}033 & 0{,}138 & 0{,}281 & 0{,}107 & 0{,}726 & 0{,}130 & 0{,}003 & 0{,}006 & 0{,}013 & 0{,}003 & 0{,}084 & 1{,}000 \end{bmatrix} \end{array}$$

Para $m=10$ obtemos $GS(10)$ dada por:

$$GS(10) = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \\ A_8 \\ A_9 \\ A_{10} \\ A_{11} \\ A_{12} \end{array} \begin{array}{c} \begin{array}{cccccccccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} & A_{11} & A_{12} \end{array} \\ \begin{bmatrix} 1,000 & & & & & & & & & & & \\ 0,798 & 1,000 & & & & & & & & & & \\ 0,601 & 0,661 & 1,000 & & & & & & & & & \\ 0,808 & 0,623 & 0,585 & 1,000 & & & & & & & & \\ 0,690 & 0,697 & 0,539 & 0,712 & 1,000 & & & & & & & \\ 0,846 & 0,820 & 0,733 & 0,716 & 0,756 & 1,000 & & & & & & \\ 0,728 & 0,510 & 0,289 & 0,545 & 0,567 & 0,668 & 1,000 & & & & & \\ 0,673 & 0,518 & 0,327 & 0,538 & 0,554 & 0,657 & 0,796 & 1,000 & & & & \\ 0,809 & 0,646 & 0,431 & 0,637 & 0,660 & 0,809 & 0,879 & 0,761 & 1,000 & & & \\ 0,734 & 0,515 & 0,298 & 0,552 & 0,573 & 0,337 & 0,499 & 0,798 & 0,883 & 1,000 & & \\ 0,719 & 0,623 & 0,459 & 0,649 & 0,827 & 0,734 & 0,791 & 0,686 & 0,780 & 0,792 & 1,000 & \\ 0,326 & 0,456 & 0,459 & 0,447 & 0,775 & 0,467 & 0,038 & 0,115 & 0,230 & 0,046 & 0,470 & 1,000 \end{bmatrix} \end{array}$$

4. COMENTÁRIOS

Conforme mencionamos anteriormente $\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$ é um estimador tendencioso, mas mesmo assim quando comparado com outros coeficientes, apresenta-se como sendo mais eficiente. O cálculo do vício deste estimador é bastante trabalhoso e não é aqui apresentado. Grassle e Smith compararam seu estimador $\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$, com os coeficientes "similaridade percent", métrica de Camberra e o coeficiente de Horn (1966) através de amostras simuladas a partir de uma distribuição multinomial conhecida, baseando-se em que: se $\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$ apresenta um vício pequeno, então a similaridade média entre a amostra e a população, deve estar próximo de um. A Tabela Ap.II.2 apresenta o resultado de 100 amostras simuladas através de $\underset{\sim}{\pi}=(0,5;\ 0,2;\ 0,2;\ 0,05;\ 0,05)$ com cada amostra contendo 30 indivíduos, mostrando assim a eficiência do estimador $\hat{N}(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, m)$ quando m cresce.

TABELA AP.II.2

	MÉDIA AMOSTRAL	DESVIO PADRÃO AMOSTRAL	INTERV.DE CONF., DE 95%,PARA A MÉDIA AMOSTRAL
Similaridade Percent	0,820	0,085	0,803 —— 0,837
Métrica Camberra	0,683	0,133	0,657 —— 0,709
Similaridade Horn	0,942	0,042	0,934 —— 0,950
$N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, 1)$	0,997	0,067	0,984 —— 1,010
$N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, 2)$	0,997	0,052	0,987 —— 1,007
$N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, 4)$	0,997	0,046	0,988 —— 1,006
$N(\underset{\sim}{\pi}_1, \underset{\sim}{\pi}_2, 10)$	0,998	0.059	0,986 —— 1,010

A P Ê N D I C E I I I

Apresentamos aqui algumas referências de programas prontos em análise de agrupamento.

1) BMDP (Biomedical Computer Programs, série P, 1979):

Nesta versão encontramos dois dos programas em análise de agrupamento que podem facilmente ser utilizados.

. P1M - Análise de agrupamento (John Hartigan, Yale University).

. P2M - Análise de agrupamento (Laszlo Engelman, HSCF).

Em nosso trabalho utilizamos o programa P1M, cuja entrada de dados foram as matrizes de proximidades e os métodos de agrupamento:

. Método do encadeamento completo (Complete Linkage)

. Método do vizinho mais próximo (Single Linkage)

. Método do encadeamento médio (Average Linkage).

O programa P2M, forma grupos, utilizando-se somente do método do encadeamento médio.

2) SAS (Statistical Analysis System, 1979):

O procedimento de análise de agrupamento deste programa e o do encadeamento completo, descrito por: Johnson, Stephen C. "Hierarchical Clustering Schemes", Psychometrika XXXII 1967, pp. 241-254.

A entrada de dados para este programa são os valores observados, e a medida de proximidade é a distância Euclidiana. O programa gera então a matriz de proximidade, que é a matriz de proximidade Euclidiana.

3) IMSL (International Mathematical Statistics Library, 1979):

Este pacote é constituído de um conjunto de subrotinas, que possibilitam o estabelecimento dos programas desejados.

No caso de análise de agrupamento, utilizamo-nos das subrotinas "oclink" e "ustree". A subrotina "oclink" permite a utilização dos métodos de agrupamento do encadeamento completo e do vizinho mais próximo, uma vez que seja fornecida a matriz de proximidade, independentemente da medida escolhida. A subrotina "ustree" trata da saída (output) a ser impressa em forma de árvore, que denominamos dendrograma.

R E F E R Ê N C I A S

ACERO, P.H.C. (1976) - *Caracterização Demográfica e Estrutura Genética de uma População Migrante Brasileira*, Tese de Doutoramento, Instituto de Biociências da USP, São Paulo.

BALAKRISHNAN, V., and SANGHVI, L.D. (1968)-"Distance Between Populations on the Basis of Atribute Data", *Biometrics*, 24, 859-865.

BHATTACHARYYA, A. (1946) - "On a Measure of Divergence Between Two Multinomial Populations", *Sankhyã*, 7, 401-406.

CUNNINGHAM, K.M., and OGILVIE, J.C. (1972) - "Evaluation of Hierarchical Grouping Techniques", *A Preliminary Study-The Computer Journal*, 15, 209-213.

EDWARDS, A.W.F., and CAVALLI-SFORZA, L.L. (1964) - "Reconstruction of Evolutionary Trees. Phenetic and Phylogenetic Classification", *The Systematics Assoc. Publ.* 6, 67-76.

EDWARDS, A.W.F. (1971) - "Distance Between Populations on the Basis of Gene Frequencies", *Biometrics*, 27, 373-381.

EVERITT, B.S. (1974) - *Cluster Analysis*, Halstead Press, London.

EVERITT, B.S. (1979) - "Unresolved Problems in Cluster Analysis", *Biometrics*, 35, 169-181.

GOODMAN, M.M. (1972) - "Distance Analysis in Biology", *Syst. Zool.*, 21 (2), 174-186.

GRASSLE, J.F., and SMITH, W. (1976) - "A Similarity Measure Sensitive to the Contribution of Rare Species end its Use in Investigation of Variation in Marine Benthic Communities", *Oecologia (Berl.)*, 25, 13-22.

HARTIGAN, J.A. (1975) - *Clustering Algorithms*, John Wiley & Sons.

HILLS, M. (1967) - "Discrimination and Allocation with Discrete Data", *Applied Statistics*, 16, 237-250.

HORN, H.S. (1966) - "Measurement of 'Overlap' in Comparative Ecological Studies", *The American Naturalist*, V.100, 914, 419-424.

JOHNSON, S.C. (1967) - "Hierarchical Clustering Schemes", *Psychometrika*, V.32, 3, 241-254.

KENDALL, M.G., and STUART, A. (1961) - *The Advanced Theory of Statistics*, V.III, Charles Griffin & Co., Ltd., London (1961).

KULLBACK, S. (1952) - "An Aplication of Information Theory

to Multivariate Analysis", *Ann. Math. Statist.*, V.23, 88 102.

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979) - *Multivariate Analysis*, Academic Press, Cap.13, 360-394.

MIAZAKI, E.S. (1979) - *Mistura de Multinormais como Técnica de Análise de Conglomerados*, Tese de Mestrado, IME, USP, São Paulo.

MOJENA, R. (1977) - "Hierarchical Grouping Methods and Stopping Rules an Evaluation", *The Computer Journal*, V.20, 4, 359-363.

MORISITA, M. (1959b) - "Measuring of Interspecific Association and Similarity Between Communities", Memoirs of the Faculty of Science, Kyushu *Univ*. Series E (Biology), 3, 65-80.

PAIT, R.G. (1979) - *Alguns Testes para Detectar Mudanças em Modelos de Regressão*, Tese de Mestrado, IME-USP, São Paulo.

PEET, R.K. (1974) - "The Measurement of Species Diversity", Annual Review of Ecology and Systematics, 5, 285-307.

PIELOU, E.C. (1969) - *An Introduction to Mathematical Ecology*, John Wiley & Sons.

RYZIN, J.V. (1977) - *Classification and Clustering*, Academic

Press, pp. 175-197.

SOKAL, R.R., and SNEATH, P.H. (1963) - *Principles of Numerical Taxionomy*, W. H. Freeman, San Francisco and London.

SOKAL, R.R., and ROHLF, F.J. (1962) - "The Comparison of Dendrograms by Objetive Methods", *Taxionomy*, V.2, 33-40.