ANÁLISE DE GRUPAMENTO

Luiz Campos de Sá Lucas

Tese submetida ao corpo docente da Coordenação dos Programas de Pós-Graduação de Engenharia da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários para obtenção do Grau de Mestre em Ciências (M.Sc.)

Aprovada por:

Cláudio T. Bornstein
Presidente

Annibal P. Sant'Anna

Paulo C. Bravo

Rio de Janeiro, RJ - BRASIL
MARÇO DE 1983

SÁ LUCAS, LUIZ CAMPOS DE

Análise de Grupamento (Rio de Janeiro) 1983.

XI, 161p. 29,7cm (COPPE-UFRJ, M.Sc., Engenharia de Sistemas, 1983)

Tese - Univ. Fed. Rio de Janeiro

1. Análise Multivariada I. COPPE/UFRJ II. Título (Série)

À

Ana,

Maria Lúcia e

João Eduardo.

AGRADECIMENTOS

Cláudio T. Bornstein ,

pela orientação da tese.

AGRADECIMENTOS

Roberto S. Quintanilha,
pelo apoio na elaboração dos programas

Angela Maria B. Alonso e

Elza Veloso de Almeida,
pelo apoio na documentação da tese.

AGRADECIMENTOS

Maria da Glória Alves de Lima

Valéria C. S. de Sant'Ana e

Marlene C. de Moraes,

pela datilografia

Geraldo S. Sonoda,

pelos desenhos

R E S U M O

São apresentados alguns dos métodos mais significativos de análise de grupamento, de forma a permitir ao leitor uma visão clara e prática do assunto. Define-se o problema de análise de grupamento e seus conceitos básicos. Apresentam-se detalhadamente os principais métodos hierárquizados, de realocação iterativa e de programação matemática para a resolução do problema de análise de grupamento. É apresentado, ainda, um exemplo prático de aplicação de análise de grupamento no estudo de etapas de crescimento de larvas. Finalmente, são fornecidas rotinas computacionais em ALGOL para a resolução de problemas pelos principais métodos descritos.

A B S T R A C T

Some of the most significative methods of cluster analysis are presented in order to give to the reader a clear and practical view of the subject. The problem of cluster analysis and its basic concepts are defined. A detailed presentation of the main hierarchical, iterative relocation and mathematical programming procedures for the solution of the cluster problem is given. Yet it is presented a practical example of cluster analysis aplication to the study of larvas' growing stages. Finally, computer programs in ALGOL for the solution of practical problems by the main described methods are given.

S U M Á R I O

I - INTRODUÇÃO

Uma das atividades básicas no processo de conhecimento humano consiste em classificar coisas semelhantes em categorias. Os objetos de conhecimento encontrados usualmente nas atividades diárias são numerosos demais para serem processados mentalmente como entidades isoladas. Os estímulos são, assim, via de regra, descritos primariamente em termos de pertinência a categorias ou grupos.

Evidentemente, tal definição de grupos envolve uma certa dose de arbitrariedade, o que pode implicar em que sejam feitas, às vêzes, certas generalizações indesejáveis.

Existem, pelo menos, três problemas no âmbito da análise estatística multivariada, associados ao estabelecimento de grupos: os problemas de classificação, de análise discriminante e de análise de grupamento (cluster analysis).

Classificação ou identificação é o processo de alocação de um novo item ou observação ao seu próprio lugar num conjunto preestabelecido de categorias. Os atributos essenciais de cada categoria são conhecidos a partir de uma amostra de cada grupo. Há, assim, uma certa incerteza na alocação de uma dada observação. Como exemplos dessa atividade poder-se-ia imaginar um geólogo identificando rochas ou um biólogo catalogando flora ou fauna. O problema de classificação pode, além disso, ser complicado por imperfeições na definição das classes, categorias que se superponham e variações aleatórias nas observações. Na prática, procura-se contornar estatisticamente essa dificuldade calculando a probabilidade de pertinência, a cada categoria, de uma dada observação, alocando-a à categoria mais provável. Apresentações detalhadas do problema de classificação são feitas, por exemplo, em TATSUOKA[39] e em COOLEY e LOHNES[7].

A análise discriminante, por sua vez, também a partir de uma amostra de cada categoria, procura estudar as direções, ou dimensões, ao longo das quais as maiores diferenças entre os grupos ocorrem. Algebricamente, o objetivo é determinar combinações lineares das variáveis originais ao longo das quais se verificam as maiores diferenças entre os grupos. Apresentações detalhadas do método são também feitas em TATSUOKA[39] e COOLEY e LOHNES[7].

No problema de "cluster analysis" (daqui por diante denominada análise de grupamento, a exemplo do que é efetuado em PINHO GAMA[30]), por

outro lado, pouco ou nada se conhece sobre a estrutura dos grupos. Possivelmente, nem o número de grupos é conhecido, sendo disponível apenas uma estimativa mais ou menos aproximada. Em essência, nesse problema, o que é conhecido é apenas o conjunto de observações relativas aos elementos cuja pertinência a categorias é desconhecida. O objetivo é, então, determinar uma estrutura de grupos que se ajuste aos dados disponíveis. Tal ajuste é feito de forma a reunir elementos semelhantes, tendo em vista as várias características observadas, em um mesmo grupo, implicando, assim, em que o grau de associação seja elevado entre os membros de uma mesma categoria e, baixo, entre os elementos de categorias distintas.

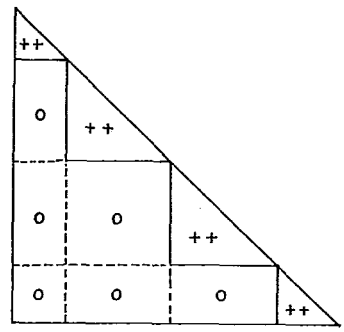
Assim, a análise de grupamento constitui-se em uma técnica a ser utilizada para a descoberta de uma estrutura de grupos e de relações entre esses grupos. Os resultados dessa análise podem contribuir para o desenvolvimento de esquemas censitários de classificação: em botânica e biologia, por exemplo, uma das principais aplicações da análise de grupamento é a construção de taxonomias. Em outras situações pode ser possível, através da análise de grupamento, reduzir um razoável volume de dados a uma descrição compacta através dos grupos formados. Se esse agrupamento "amostral" for adotado para uso operacional, pode-se então tornar a base para a classificação de novas observações.

É interessante observar que essa técnica pode também ser utilizada no sentido inverso, como citou ANDERBERG[1]: se, após a aplicação de um algoritmo de análise de grupamento, os grupos resultantes apresentarem entre si um grau de diferenciação muito pequeno, provavelmente os elementos consistem, apenas, em uma única classe. Aquele autor cita a aplicação dessa idéia à análise de conjuntos de células humanas, à procura de células anormais: uma amostra contendo apenas células normais seria homogênea, ao passo que se houvessem células anormais os resultados da aplicação do método à amostra apresentariam um agrupamento significativo para um certo número de grupos.

A Análise de Grupamento, além disso, pode também ser vista como uma técnica de análise fatorial, como indica HARMAN[16]. A análise fatorial procura reduzir as variáveis originais de um problema a um número menor de fatores, procurando, assim, obter-se uma descrição resumida dos dados do problema.

Nos casos particulares em que a matriz de correlação entre as variáveis originais indique a existência de grupos distintos de variáveis,

ou seja, quando a matriz possa ser escrita de forma:



onde "+ +" e " 0 " representam, respectivamente, uma elevada correlação e correlação nula, a análise de grupamento pode ser utilizada para a detecção desses grupos. Nesse caso, a técnica consistiria em agrupar "variáveis" e não "elementos". Infelizmente, é bastante improvavel que na prática matrizes da forma apresentada na figura anterior venham a ocorrer, o que prejudica a aplicação do método nesse caso.

A análise de grupamento reveste-se, assim, de uma natureza geral, e pode ser aplicada, praticamente, em qualquer área do conhecimento humano, tal como biologia, entomologia, psicologia, educação, economia , pesquisa de mercado, geologia, planejamento urbano e regional, etc.

O primeiro texto sobre o que hoje é conhecido como análise de grupamento, é devido a TRYON[40] e foi publicado em 1939. Desde então, uma vastíssima literatura tem sido apresentada, sendo hoje enorme a variedade de técnicas existentes nessa área. Vários autores (ANDERBERG[1], DURAN e ODELL[10], HARTIGAN[17], PINHO GAMA[30], etc.) têm procurado apresentar um estudo unificado do problema, e se constituem em excelentes referências sobre o assunto. No entanto, dada a amplitude do tema, nenhum desses estudos chega a ser totalmente abrangente, especializando-se, cada texto, como seria de se esperar, numa determinada área.

A presente tese não tem o objetivo de ser um trabalho exaustivo, o que seria impossível. Procurou-se, aqui, efetuar uma apresentação cuidadosa de alguns dos métodos mais significativos e que permitisse ao interessado no assunto uma visão clara e prática de algumas das principais técnicas de análise de grupamento.

Assim, foram aqui incluídos métodos importantes e de publicação recente, que não constam da literatura usualmente disponível e, para melhor entendimento por parte do leitor, foram desenvolvidos para esta tese,

pelo autor, novos teoremas e demonstrações , um método hierarquizado divisivo e novas formalizações para alguns métodos.

De uma maneira geral, o Capítulo II define o problema de análise de grupamento e os conceitos básicos que permitirão a apresentação dos métodos selecionados. No Capítulo III é feita uma descrição sucinta desses métodos (hierarquizados, de realocação iterativa e de programação matemática), cuja apresentação detalhada é feita nos Capítulos IV, V e VI , respectivamente. No Capítulo VII, por sua vez, é apresentado um exemplo prático de aplicação de análise de grupamento no estudo de etapas de crescimento de larvas. Finalmente, no Apêndice I são apresentadas tabelas - resumo das distâncias, medidas de dispersão e funções objetivo utilizados pelos diversos métodos apresentados, sendo, por sua vez, apresentadas no Apêndice II rotinas computacionais em ALGOL para a resolução de problemas práticos pelos principais métodos apresentados neste trabalho.

II - DEFINIÇÃO DO PROBLEMA

II.1 - CONCEITOS PRELIMINARES

Seja $E = \{ e_1, e_2, \ldots, e_n \}$ o conjunto dos n elementos pertencentes à população em estudo. Suponha-se que existam p características observáveis e mensuráveis (quantitativas ou qualitativas), possuídas por cada elemento pertencente a E.

Denotando a medida da k-ésima característica do elemento e_j por x_{kj} e o conjunto das medidas das p características do mesmo elemento e_j pelo vetor coluna

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \cdot \\ \cdot \\ \cdot \\ x_{pj} \end{bmatrix} ,$$

é possível resumir a descrição dos n elementos da população em um conjunto de vetores $X = \{ X_1, X_2, \ldots, X_n \}$. Note-se que o conjunto X pode ser encarado como um conjunto de n pontos no R^p.

II.2 - O PROBLEMA DE ANÁLISE DE GRUPAMENTO

Seja m um número inteiro menor do que n. O problema de análise de grupamento pode ser resumido na seguinte afirmação: "Com base no conjunto X, determinar uma partição P_m dos objetos pertencentes a E em m grupos $g_1, g_2, \ldots, g_m$, alocando cada e_j a apenas um grupo, de forma a que elementos semelhantes sejam reunidos num mesmo agrupamento e objetos não semelhantes sejam alocados a grupos distintos".

Assim, a solução de um problema de análise de grupamento pode ser encarada como uma partição do conjunto E que otimize uma função objetivo $f(P_m)$, função essa que reflita uma medida quantitativa de semelhança "intra" e "entre" grupos.

É possível ilustrar as afirmações acima com um exemplo, citado por DURAN e ODELL[10] . Suponha-se que p=1 característica é medida em cada um dentre n=8 indivíduos, resultando no conjunto X ={ 3,4,7,4,3,3,4,4 }. Deseja-se obter uma partição dos oito elementos em m=3 grupos. Uma medida quantitativa de semelhança poderia ser fornecida pela soma dos desvios quadráticos de cada ponto x_i em relação à media do grupo g_j ao qual esse pon

to fosse alocado. O objetivo então seria minimizar

$$W = \sum_{j=1}^{3} w_j = \sum_{j=1}^{3} \sum_{i \epsilon g_j} (x_i - \bar{x}_j)^2 ,$$

onde $\bar{x}_j$ é a média do grupo g_j.

Evidentemente a solução é dada por

$$g_1 = \{ 3,3,3 \} ,$$

$$g_2 = \{ 4,4,4,4 \} \text{ e}$$

$$g_3 = \{ 7 \},$$

com $W = w_1 + w_2 + w_3 = 0 + 0 + 0 = 0$.

Para que se possa, no entanto, estabelecer funções objetivo convenientes, faz-se necessário discutir algumas medidas de semelhança entre elementos e entre conjuntos de elementos.

II.3 - MEDIDAS DE SEMELHANÇA ENTRE DOIS ELEMENTOS: FUNÇÕES DE DISTÂNCIA, COEFICIENTES DE SIMILARIDADE E COEFICIENTE DE CORRELAÇÃO.

Como foi visto, a solução de um problema de análise de grupamento envolve a quantificação de semelhança e a reunião de elementos semelhantes em um mesmo grupo. Uma maneira de se resolver esse problema seria, por exemplo, atribuir dois elementos e_i e e_j ao mesmo grupo se:

- a distância d_{ij} entre os pontos X_i e X_j fosse suficientemente pequena;

- uma medida de similaridade s_{ij} entre X_i e X_j fosse suficientemente grande; ou

- o coeficiente de correlação r_{ij} entre X_i e X_j fosse suficientemente elevado, atribuindo-se os dois elementos a grupos distintos se d_{ij} fosse elevado ou se s_{ij} ou r_{ij} fossem pequenos.

Tal raciocínio embasa a seguinte definição, dada por DIDAY e SIMON[9] e fundamentada no conceito de distância: um grupo g_s é dito homogêneo se para todos e_i e $e_j \epsilon g_s$ e $e_k \notin g_s$,

$$d_{ij} \leq d_{ik} \quad e \quad d_{ij} \leq d_{jk} ,$$

sendo uma partição $P_m = \{ g_1, g_2, \ldots, g_m \}$ dita homogênea se a propriedade acima for verdadeira para todo $g_s \epsilon P_m$. Definições semelhantes poderiam então ser estabelecidas também para s_{ij} e r_{ij}.

Assim, torna-se interessante analisar mais detalhadamente os conceitos de distância, e de coeficientes de similaridade e de correlação.

II.3.1 - FUNÇÕES DE DISTÂNCIA

Uma função real não-negativa d_{ij} é dita uma função de distância, ou uma métrica, se, para todos X_i, X_j e X_k ϵ R^p, três propriedades são satisfeitas:

i) $d_{ij} = 0$ se e somente se $X_i = X_j$

ii) $d_{ij} = d_{ji}$

iii) $d_{ij} \leq d_{ik} + d_{kj}$

O valor de d_{ij} para X_i e X_j especificados é dito a distância entre X_i e X_j ou, equivalentemente, a distância entre e_i e e_j com relação às p características de interesse.

DURAN e ODELL[10], DIDAY e SIMON[9], PINHO GAMA[30] e outros autores apresentam uma série de funções de distância passíveis de utilização. Dentre estas, foram selecionadas como de interesse para este trabalho as métricas de Minkowsky, ou ℓ_p normas, que tomam a forma:

$$d_{ij} = d_\lambda(X_i, X_j) = \left[\sum_{k=1}^{p} | x_{ki} - x_{kj} |^\lambda \right]^{1/\lambda} \quad , \quad \lambda = 1,2,\ldots$$

Nos casos particulares de $\lambda=1$ e $\lambda=2$, tem-se respectivamente as distâncias expressas nas normas um e euclidiana. A norma um é muito útil em termos de eficiência computacional.

A métrica euclidiana, por sua vez, além de ser bastante conhecida, tem para este trabalho um grande interesse, pois algumas medidas de dispersão dentro de um grupo ou entre grupos a serem apresentadas aqui, têm uma ênfase especial na utilização desse tipo de métrica.

Cabe também observar que o quadrado da distância euclidiana entre e_i e e_j, aqui denotado por $d_2^2(X_i, X_j)$, pode ser escrito da forma:

$$d_2^2(X_i, X_j) = \sum_{k=1}^{p} (x_{ki} - x_{kj})^2 = (X_i - X_j)^t (X_i - X_j).$$

As distâncias d_{ij} podem ser dispostas em matrizes da forma

$D = [\, d_{ij} \,]$ ou $D^2 = [\, d^2_{ij} \,]$, simétricas, de dimensão $n \times n$.

A interpretação do significado da distância como medida de semelhança no caso de medidas quantitativas é evidente. Tal interpretação, no entanto, no caso de medidas qualitativas, deve ser verificada cuidadosamente. Como afirmam DIDAY e SIMON[9], "... como podem dois nomes, ou duas cores, serem adicionados ou multiplicados?". Os mesmos autores, no entanto, sugerem que, ao invés de se criar uma variável qualitativa, que assuma valores inteiros, por exemplo, para cores, se considere cada uma dessas cores como uma variável binária (ou dicotômica, isto é $x_{ki} \in \{0,1\}$) , onde "1" representaria a presença do atributo k no elemento i e "0" sua ausência. Nesse caso, operações aritméticas nos vetores X_i passam a fazer sentido e funções de distâncias podem ser utilizadas sem que se obtenha resultados absurdos.

Por outro lado, o conjunto de elementos $\{ e_1, e_2, \ldots, e_n \}$ é usualmente medido em diferentes unidades. Assim, uma determinada variável pode ser medida, por exemplo, em quilômetros e outra em metros. A utilização , sem maiores cuidados, dos valores "brutos" das medidas implicaria no estabelecimento de uma ponderação implícita nas variáveis: a variável medida em quilômetros teria um peso mil vêzes menor que a característica medida em metros.

É usual, então, procurar-se uma equalização das variáveis, expressando-as de uma forma adimensional. As transformações mais comuns são do tipo:

$$\hat{x}_{ki} = x_{ki} / t_k \ ,$$

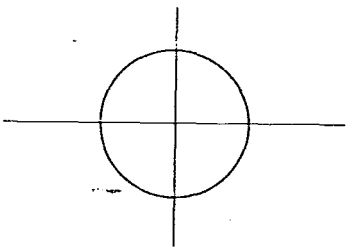
onde t_k é a média, ou a amplitude, ou o desvio padrão da k-ésima variável.

Como aponta ANDERBERG[1], vários autores consideram essa técnica como um método completo de ponderação de variáveis e não dedicam maior atenção a esse tópico. Aquele autor, no entanto, aponta que esse procedimento é uma abdicação da responsabilidade do analista, na medida que isto implica em afirmar que o incremento na dispersão tem a mesma importância para todas as variáveis, qualquer que seja o propósito da análise. Ainda segundo ANDERBERG[1], a escolha dos pesos não depende de técnicas automáticas: na verdade, é o principal meio de que o analista dispõe para adequar a análise aos seus objetivos.

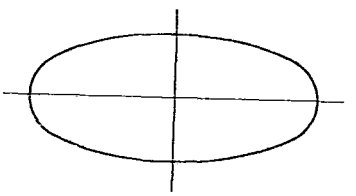
Assim, pode-se considerar a possibilidade de uma transformação de variáveis, anterior ao cálculo das distâncias, da forma:

$$\hat{x}_{ki} = \frac{w_k}{t_k} \; x_{ki} = \alpha_k \, x_{ki}$$

onde w_k representa um peso, atribuído pelo analista, para a k-ésima variável. Geometricamente, e trabalhando com a bola unitária, o efeito seria como descrito na Figura II.1 (v. p. ex. ANDERBERG[1]):



$\hat{x}_{ki} = x_{ki}$



$\hat{x}_{ki} = \alpha_k \, x_{ki}$

Essa transformação provoca um alongamento ou encurtamento nas variáveis, tal que a bola unitária se transforma num elipsóide.

No caso da distância euclidiana, por exemplo, ter-se-ia, após a transformação:

$$d_2^2 \left(\hat{X}_i, \hat{X}_j \right) = \sum_{k=1}^{p} \left(\alpha_k \, x_{ki} - \alpha_k \, x_{kj} \right)^2 = \sum_{k=1}^{p} \alpha_k^2 \left(x_{ki} - x_{kj} \right)^2$$

Evidentemente, nesse caso, como as distâncias calculadas são diferentes para os dados brutos e os dados transformados, o resultado final do processo de agrupamento será, de maneira geral, diferente nos dois casos.

Por outro lado, existem outras transformações, onde $\hat{x}_{kj}$ se torna uma combinação linear das variáveis originais. Uma transformação desse tipo, bastante usual, é aquela obtida pela utilização da análise de componentes principais (v. p.ex. ANDERBERG[1], COOLEY e LOHNES[7], HARMAN[16] ou

TATSUOKA[39]).

Esse método pode, inclusive, permitir uma maior economia na descrição dos dados, uma vez que procura reduzir as p variáveis originais a um número menor de componentes principais. Cabe aqui ressaltar, no entanto, que a solução que vier a ser obtida com os dados assim transformados pode ser criticamente prejudicada no que se refere à interpretação das características dos grupos formados, uma vez que as componentes principais têm apenas uma interpretação geométrica e podem não fazer "sentido" no âmbito do estudo que estiver sendo desenvolvido.

II.3.2 - COEFICIENTES DE SIMILARIDADE

Uma função real não-negativa s_{ij} é dita uma medida de similaridade se, para todos X_i e X_j ε R^p três propriedades são satisfeitas:

i) $0 \leq s_{ij} \leq 1$ se $X_i \neq X_j$

ii) $s_{ii} = 1$

iii) $s_{ij} = s_{ji}$

A quantidade s_{ij} é também denominada coeficiente de similaridade. Os diversos autores de textos sobre Análise de Grupamento apresentam vários coeficientes de similaridade, dos quais foram selecionados para apresentação neste trabalho o Coeficiente de Gower, no caso de variável quantitativa (ver, p.ex. GOWER[14] ou PINHO GAMA[30]), e o Coeficiente de Sokal e Michener (ver, p.ex. SOKAL e MICHENER[37] ou DURAN e ODELL[10]).

II.3.2.1 - COEFICIENTE DE GOWER - VARIÁVEL QUANTITATIVA

Sejam e_i e e_j dois elementos quaisquer de E que se quer comparar em relação a uma característica k.

Define-se então uma quantidade s_{ijk} da forma:

$$s_{ijk} = 1 - \frac{| x_{ki} - x_{kj} |}{R_k} \quad ,$$

onde

x_{ki} = medida da k-ésima característica em e_i,

R_k = amplitude da variável k observada em E

= $\max_{X_r, X_s \varepsilon X} | x_{kr} - x_{ks} |$.

Assim, se e_i e e_j possuem a mesma medida na variável k, $s_{ijk}=1$. Se e_i e e_j são os elementos que mais distam entre s_i na característica k, dentro do conjunto E, $s_{ijk}=0$. No caso geral, $0 \leq s_{ijk} \leq 1$.

O Coeficiente de Gower, neste caso, nada mais é do que a média, no conjunto das p variáveis, das similaridades entre e_i e e_j medidas por s_{ijk}, ou seja,

$$s_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk} .$$

Quando, em todas as características, tem-se $s_{ijk}=1$, o Coeficiente de Gower s_{ij} toma o valor 1. Se todos $s_{ijk}=0$, s_{ij} é igual a 0. No caso geral, $0 \leq s_{ij} \leq 1$.

II.3.2.2 - COEFICIENTE DE SOKAL E MICHENER

Quando todas as p características forem representadas por variáveis binárias, vários coeficientes de similaridade podem ser definidos (ver, p.ex., DURAN e ODELL[10], e DIDAY e SIMON[9]). Um desses coeficientes é o de Sokal e Michener, que pode ser escrito da forma:

$$s_{ij} = \frac{a + b}{p}$$

onde

a = número de características para as quais $x_{ki} = x_{kj} = 1$

b = número de características para as quais $x_{ki} = 0$ e $x_{kj} = 0$

p = número de características.

Assim, se $x_{ki} = x_{kj}$ para todo k, $s_{ij}=1$. Se $x_{ki} \neq x_{kj}$ para todo k, $s_{ij}=0$. No caso geral, $0 \leq s_{ij} \leq 1$.

II.3.2.3 - MÉTRICAS E COEFICIENTES DE SIMILARIDADE

É possível construir-se métricas a partir de coeficientes de similaridade, caso se efetuem as transformações adequadas. GOWER[14] (ver também PINHO GAMA[30]) sugere a seguinte transformação:

$$d_{ij} = (1 - s_{ij})^{1/2}$$

Esta função d_{ij} é uma métrica. Evidentemente, atende às propriedades (i) e (ii) do item II.3.1. No que se refere à propriedade (iii),

tem-se (ver, p.ex. EVERITT[12]):

$$(1 - s_{ij})^{1/2} \leq (1 - s_{ik})^{1/2} + (1 - s_{kj})^{1/2}$$

II.3.3 - COEFICIENTE DE CORRELAÇÃO

O coeficiente de correlação entre X_i e X_j, denotado por r_{ij}, é definido por (ver, p.ex. DIDAY e SIMON[9] ou DURAN e ODELL[10]):

$$r_{ij} = \frac{\sum_{k=1}^{p} (x_{ki} - \bar{x}_{.i})(x_{kj} - \bar{x}_{.j})}{\left[\sum_{k=1}^{p} (x_{ki} - \bar{x}_{.i})^2\right]^{1/2} \left[\sum_{k=1}^{p} (x_{kj} - \bar{x}_{.j})^2\right]^{1/2}},$$

onde $\bar{x}_{.i} = \frac{1}{p} \sum_{k=1}^{p} x_{ki}$ e $\bar{x}_{.j} = \frac{1}{p} \sum_{k=1}^{p} x_{kj}$

Sejam, por outro lado, os vetores Y_i e Y_j, obtidos a partir de X_i e X_j pela transformação:

$$Y_i = X_i - \begin{bmatrix} \bar{x}_{.i} \\ \bar{x}_{.i} \\ \vdots \\ \bar{x}_{.i} \end{bmatrix} = \begin{bmatrix} x_{ki} - \bar{x}_{.i} \end{bmatrix} \text{ e } Y_j = X_j - \begin{bmatrix} \bar{x}_{.j} \\ \bar{x}_{.j} \\ \vdots \\ \bar{x}_{.j} \end{bmatrix} = \begin{bmatrix} x_{kj} - \bar{x}_{.j} \end{bmatrix}.$$

O coeficiente de correlação r_{ij} pode então ser reescrito da forma:

$$r_{ij} = \frac{\sum_{k=1}^{p} y_{ki}\, y_{kj}}{\left[\sum_{k=1}^{p} y_{ki}^2\right]^{1/2} \left[\sum_{k=1}^{p} y_{kj}^2\right]^{1/2}} = \frac{Y_i^t Y_j}{||Y_i||\, ||Y_j||} = \cos \theta_{ij},$$

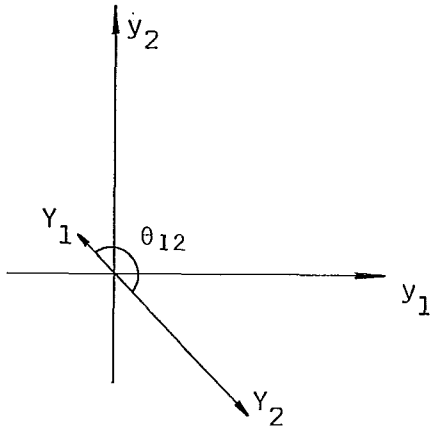
onde θ_{ij} é o ângulo formado entre Y_i e Y_j.

Como exemplo, sejam $X_1^t = \begin{bmatrix} 2 & 4 \end{bmatrix}$ e $X_2^t = \begin{bmatrix} 7 & 3 \end{bmatrix}$. Como $\bar{x}_{.1} = 3$ e $\bar{x}_{.2} = 5$, tem-se $Y_1^t = \begin{bmatrix} -1 & 1 \end{bmatrix}$ e $Y_2^t = \begin{bmatrix} 2 & -2 \end{bmatrix}$. O coeficiente de correla-

ção é então dado por:

$$r_{ij} = \frac{-2 - 2}{\sqrt{2} \cdot \sqrt{8}} \simeq \frac{-4}{4} = -1,0$$

Graficamente, tem-se:



Se, por outro lado, tivermos $X_1^t = [2 \quad 4]$ e $X_2^t = [8 \quad 16]$, é fácil ver que $r_{ij} = 1$. De uma maneira geral (ver, p.ex. DURAN e ODELL[10]), se $X_j = \alpha X_i$, $\alpha > 0$, tem-se $r_{ij} = 1$. Como $-1 \leq r_{ij} = \cos \theta_{ij} \leq +1$, no caso da utilização do coeficiente de correlação como medida de semelhança diz-se que e_i e e_j são semelhantes de forma positiva se r_{ij} for próximo de "+1", de forma negativa se r_{ij} for próximo de "-1" e não semelhantes, se r_{ij} for próximo de "0".

É importante notar aqui que a medida de semelhança fornecida por r_{ij} é bastante diferente daquela fornecida por d_{ij} ou s_{ij}. Nos casos das funções de distância e do coeficiente de similaridade, o maior grau de semelhança entre e_i e e_j é atingido quando $X_i = X_j$. No caso do coeficiente de correlação, a maior semelhança é medida quando $X_j = \alpha X_i$, $\alpha > 0$.

É possível também associar ao coeficiente de correlação uma função de distância. MULVEY e CROWDER[29] definem uma "métrica" de correlação a partir da seguinte transformação:

$$d_{ij} = \left[0.5 \ (1 - r_{ij})\right]^{1/2}$$

Assim, quando dois elementos têm uma correlação positiva perfeita ($r_{ij} = +1$), tem-se $d_{ij} = 0$. Quando $r_{ij} = 0$, tem-se $d_{ij} = 0,71$ e no caso de correlação negativa perfeita ($r_{ij} = -1$), tem-se $d_{ij} = 1$. Note-se que esta função não sa

tisfaz à propriedade (i) do item II.3.1, não sendo assim, a rigor, uma mé trica.

EXEMPLO

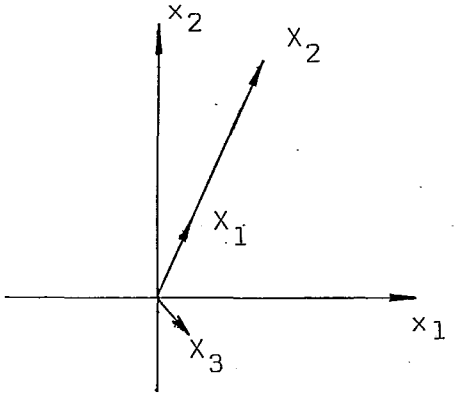
É possível comparar os efeitos da utilização das diversas medidas de semelhança aqui apresentadas através de um exemplo. Suponha-se que a três crianças foram aplicados dois testes, obtendo as crianças os "scores" apresentados no conjunto X

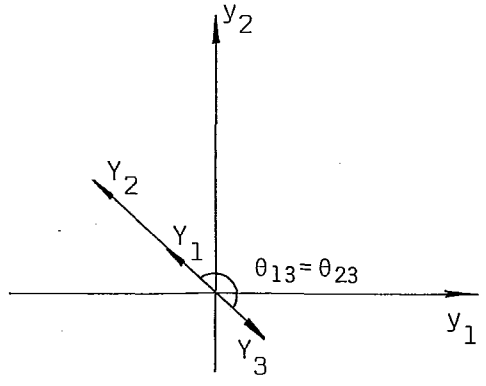
$$X = \{ X_1, X_2, X_3 \} = \{ \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 12 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \}.$$

Como $\bar{x}^t = [\,2 \quad 8 \quad 0\,]$, tem-se:

$$Y = \{ \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -4 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \}.$$

Graficamente:





Calculando-se as matrizes D_2^2 (de quadrados das distâncias euclidianas), G (de coeficientes de Gower), R (de coeficientes de correlação), D_2 (de distâncias euclidianas), D_G (de distâncias associadas ao coeficiente de Gower) e D_R (de métricas de correlação), tem-se:

$$D_2^2 = \begin{bmatrix} 0 & 90 & 16 \\ 90 & 0 & 178 \\ 16 & 178 & 0 \end{bmatrix}, \quad \text{logo} \quad D_2 \simeq \begin{bmatrix} 0 & 9,5 & 4,0 \\ 9,5 & 0 & 13,3 \\ 4,0 & 13,3 & 0 \end{bmatrix}$$

$$G \simeq \begin{bmatrix} 1 & 0.15 & 0.85 \\ 0.15 & 1 & 0.38 \\ 0.85 & 0.38 & 1 \end{bmatrix} \quad \text{logo} \quad D_G \simeq \begin{bmatrix} 0 & 0.92 & 0.39 \\ 0.92 & 0 & 0.79 \\ 0.39 & 0.79 & 0 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \quad \text{logo} \quad D_R = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Assim, caso o objetivo fosse formar dois grupos homogêneos tal como definido por DIDAY e SIMON[9] (ver início deste item II.3) ter-se-ia nos casos da métrica euclidiana e do coeficiente de Gower a partição $P_2 = \{g_1, g_2\} = \{\{1, 3\}, \{2\}\}$ e, no caso do coeficiente de correlação, a partição $P_2 = \{g_1, g_2\} = \{\{1, 2\}, \{3\}\}$

II.4 - MEDIDAS DE DISPERSÃO INTERNA DE UM GRUPO

A solução de um problema de análise de grupamento envolve, como já foi visto, a quantificação de semelhança e a reunião de elementos semelhantes em um mesmo grupo. Uma maneira de se aferir a semelhança entre os elementos reunidos em um determinado grupo é medir a dispersão dos elementos nesse grupo.

Sejam, por exemplo:

$g_J = \{1, 2, 3, 4, 5\}$ e $g_K = \{2, 2, 2, 2, 2\}$.

Se a dispersão fosse aferida por uma medida do tipo $W = \sum_{i \in g} (x_i - \bar{x})^2$, onde $\bar{x}$ fosse a média do grupo g, ter-se-ia

$$W_J = (1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$$

$$W_K = 5 \times (2 - 2)^2 = 0.$$

Nesse sentido, g_K é menos "disperso" que g_J, o que leva a afirmar que os elementos reunidos em g_K são mais semelhantes entre si que os elementos pertencentes a g_J.

De uma maneira geral, seja um grupo $g_I = \{e_1, e_2, \ldots, e_{n_I}\}$ com n_I elementos e $X_I = \{X_1, X_2, \ldots, X_{n_I}\}$ o conjunto de observações efetuadas nas p características, relativo a g_I. É possível definir algumas medidas de dispersão para g_I na forma que se segue.

II.4.1 - SOMA DOS QUADRADOS DENTRO DO GRUPO

Para o conjunto g_I, DURAN e ODELL[10] definem a medida W_I, soma dos quadrados dentro do grupo g_I, da forma :

$$W_I = \sum_{i=1}^{n_I} (X_i - \bar{X}_I)^t (X_i - \bar{X}_I);$$

onde $\bar{X}_I = \frac{1}{n_I} \sum_{i=1}^{n_I} X_i$ é dita a média, ou centróide, do grupo.

Assim, W_I representa a soma dos quadrados das distâncias euclidianas entre cada elemento do grupo e a centróide do mesmo.

Por outro lado, o teorema abaixo estabelece uma importante relação, que permite a determinação de W_I sem que se tenha de proceder ao cálculo da centróide (v. p.ex. DURAN e ODELL[10]):

Teorema II.1:

$$W_I = \sum_{i=1}^{n_I} d_2^2 (X_i , \bar{X}_I) = \frac{1}{n_I} \sum_{i=1}^{n_I} \sum_{j=1}^{i-1} d_2^2 (X_i, X_j).$$

Demonstração:

Primeiramente, pode-se verificar que:

$$\sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I) = \frac{1}{2n_I} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} d_2^2 (X_i, X_j). \quad (1)$$

Para tal, tem-se que, por definição,

$$\sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I) = \sum_{k=1}^{p} \sum_{i=1}^{n_I} (x_{ki} - \bar{x}_{kI})^2$$

$$= \sum_{k=1}^{p} \sum_{i=1}^{n_I} (x_{ki}^2 - 2x_{ki} \bar{x}_{kI} + \bar{x}_{kI}^2).$$

Como $\bar{x}_{kI} = \frac{1}{n_I} \sum_{j=1}^{n_I} x_{kj}$,

$$\sum_{i=1}^{n_I} d_2^2(X_i, \bar{X}_I) = \sum_{k=1}^{p} \sum_{i=1}^{n_I} (x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} + \frac{1}{n_I^2} \sum_{j=1}^{n_I} \sum_{r=1}^{n_I} x_{kj}\, x_{kr})$$

Separando as parcelas entre parênteses e colocando o somatório em p em evidência,

$$= \sum_{k=1}^{p} \left[\sum_{i=1}^{n_I} \{x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj}\} + \right.$$

$$\left. + \sum_{i=1}^{n_I} \{\frac{1}{n_I^2} \sum_{j=1}^{n_I} \sum_{r=1}^{n_I} x_{kj}\, x_{kr}\} \right],$$

ou

$$= \sum_{k=1}^{p} \left[\sum_{i=1}^{n_I} \{ x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} \} + \right.$$

$$\left. + n_I \{ \frac{1}{n_I^2} \sum_{j=1}^{n_I} \sum_{r=1}^{n_I} x_{kj}\, x_{kr} \} \right],$$

ou ainda,

$$= \sum_{k=1}^{p} \left[\sum_{i=1}^{n_I} \{x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj}\} + \right.$$

$$\left. + \frac{1}{n_I} \sum_{j=1}^{n_I} \sum_{r=1}^{n_I} x_{kj}\, x_{kr} \right].$$

Somando e subtraindo $\frac{1}{2} \sum_{i=1}^{n_I} x_{ki}^2 = \frac{1}{2} \sum_{j=1}^{n_I} x_{kj}^2$ e trocando o índice r por i, tem-se:

$$\sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I) = \sum_{k=1}^{p} \left[\sum_{i=1}^{n_I} \{ x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj}\} - \frac{1}{2} \sum_{i=1}^{n_I} x^2_{ki} + \right.$$

$$\left. + \frac{1}{n_I} \sum_{j=1}^{n_I} \sum_{i=1}^{n_I} x_{kj}\, x_{ki} + \frac{1}{2} \sum_{j=1}^{n_I} x_{kj}^2 \right]$$

Colocando em evidência o somatório em i nas três últimas parcelas,

$$\sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I) = \sum_{k=1}^{p} \left[\sum_{i=1}^{n_I} \{ x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj}\} \right.$$

$$\sum_{i=1}^{n_I} \left\{ -\frac{1}{2} x_{ki}^2 + \frac{1}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} + \frac{1}{2n_I} \sum_{j=1}^{n_I} x_{kj}^2 \right\} \Bigg].$$

Resumindo em um único somatório em i,

$$\sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I) = \sum_{k=1}^{p} \sum_{i=1}^{n_I} \left(x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} - \frac{1}{2} x_{ki}^2 + \right.$$

$$\left. + \frac{1}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} + \frac{1}{2n_I} \sum_{j=1}^{n_I} x_{kj}^2 \right) ,$$

ou

$$= \sum_{k=1}^{p} \sum_{i=1}^{n_I} \left(\frac{1}{2} x_{ki}^2 - \frac{2}{2n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} + \frac{1}{2n_I} \sum_{j=1}^{n_I} x_{kj}^2 \right),$$

ou ainda

$$= \frac{1}{2} \sum_{k=1}^{p} \sum_{i=1}^{n_I} \left(x_{ki}^2 - \frac{2}{n_I} x_{ki} \sum_{j=1}^{n_I} x_{kj} + \frac{1}{n_I} \sum_{j=1}^{n_I} x_{kj}^2 \right).$$

como $x_{ki}^2 = \frac{1}{n_I} \sum_{j=1}^{n_I} x_{ki}^2$,

$$d_2^2 (X_i, \bar{X}_I) = \frac{1}{2n_I} \sum_{k=1}^{p} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} (x_{ki}^2 - 2x_{ki} x_{kj} + x_{kj}^2)$$

$$= \frac{1}{2n_I} \sum_{k=1}^{p} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} (x_{ki} - x_{kj})^2$$

$$= \frac{1}{2n_I} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} \sum_{k=1}^{p} (x_{ki} - x_{kj})^2$$

$$= \frac{1}{2n_I} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} d_2^2 (X_i, X_j) ,$$

o que prova a equação (1). No entanto, como

$$d_2^2 (X_i, X_j) = d_2^2 (X_j, X_i) \text{ e } d_2^2 (X_i, X_i) = 0,$$

tem-se

$$\frac{1}{2n_I} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} d_2^2 (X_i, X_j) = \frac{1}{n_I} \sum_{i=1}^{n_I} \sum_{j=1}^{i-1} d_2^2 (X_i, X_j) \text{ , o que}$$

completa a demonstração do teorema.

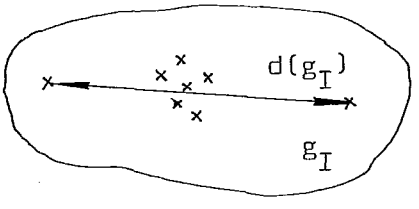
II.4.2 - VARIÂNCIA INTERNA DE GRUPO

Com base na soma dos quadrados dentro do grupo, DURAN e ODELL[10] definem a variância interna do grupo g_I, denotada por S_I^2, como sendo:

$$S_I^2 = \frac{1}{n_I} W_I \text{ .}$$

II.4.3 - DIÂMETRO DE GRUPO

HANSEN e DELATTRE[15] definem como diâmetro do grupo g_I, denotado por $d(g_I)$, a máxima dessemelhança entre os elementos do grupo. Esta dessemelhança pode ser medida, por exemplo, por quaisquer das medidas apresentadas no item II.3. Assim, o grupo apresentado abaixo teria, caso a medida fosse a distância euclidiana, o diâmetro indicado na figura abaixo:



II.4.4 - DISPERSÃO VIA MEDIANA DO GRUPO

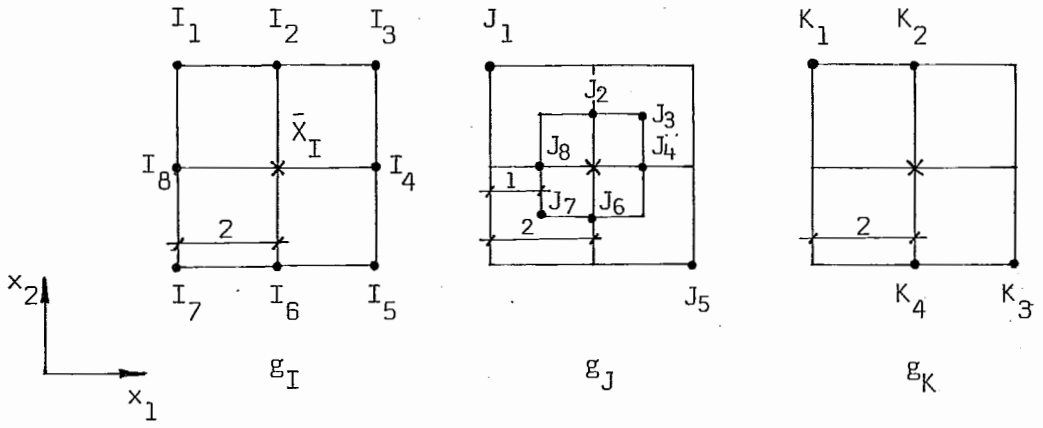
VINOD[41] , em um de seus modelos de Análise de Grupamento via Programação Inteira (ver item VI.1) apresenta um conceito de dispersão interna de grupo semelhante a W_I. Seja o grupo g_I. A mediana do grupo é o elemento e_r para o qual a soma das distâncias (medidas através de uma métrica qualquer) entre ele e os demais elementos de g_I, denotada aqui por $Z(e_r)$,

$$Z(e_r) = \sum_{i=1}^{n_I} d_{ir} \text{ ,}$$

é mínima. Nesse caso, $Z_{mín}(g_I) = Z(e_r) = \min_{e_j \in g_I} Z(e_j)$ mede a dispersão do grupo, dentro do modelo proposto por Vinod.

EXEMPLO:

É possível comparar as diversas medidas de dispersão apresentadas no item anterior por um exemplo. Sejam três grupos g_I, g_J e g_K, onde são medidas duas características, como abaixo:



Tem-se, respectivamente:

i) Soma dos quadrados dentro do grupo:

$$W_I = (2^2+2^2)+2^2+(2^2+2^2)+2^2+(2^2+2^2)+2^2+(2^2+2^2)+2^2 = 48$$

$$W_J = (2^2+2^2)+1^2+(1^2+1^2)+1^2+(2^2+2^2)+1^2+(1^2+1^2)+1^2 = 24$$

$$W_K = (2^2+2^2)+2^2+(2^2+2^2)+2^2 = 24$$

Os grupos menos dispersos internamente seriam g_J e g_K

ii) Variância interna:

$$S_I^2 = 48/8 = 6$$

$$S_J^2 = 24/8 = 3$$

$$S_K^2 = 24/4 = 6$$

O grupo g_J seria o menos disperso internamente.

iii) Diâmetro de grupo:

Tomando como medida de similaridade o quadrado da distância euclidiana:

$$d(g_I) = d_2^2(I_1, I_5) = d_2^2(I_3, I_7) = (2^2+2^2) + (2^2+2^2) = 24$$

$$d(g_J) = d_2^2(J_1, J_5) = 24$$

$$d(g_K) = d_2^2(K_1, K_3) = 24$$

Os três grupos teriam a mesma dispersão interna.

iv) Dispersão via mediana de grupo:

Utilizando a distância euclidiana, ao quadrado,

- Grupo g_I:

$$Z(I_1) = Z(I_3) = Z(I_5) = Z(I_7) =$$

$$= 2^2+4^2+(4^2+2^2)+(4^2+4^2)+(4^2+2^2)+4^2+2^2 = 112$$

$$Z(I_2) = Z(I_4) = Z(I_6) = Z(I_8)$$

$$= 2^2+(2^2+2^2)+(2^2+4^2)+4^2+(4^2+2^2)+(2^2+2^2)+2^2 = 80,$$

logo I_2(ou I_4 ou I_6 ou I_8) é mediana, com $Z_{mín} = 80$.

- Grupo g_J:

$$Z(J_1) = Z(J_5) = (2^2+1^2)+(3^2+1^2)+(3^2+2^2)+(4^2+4^2) +$$

$$+ (2^2+3^2)+(1^2+3^2)+(1^2+2^2) = 88$$

$$Z(J_2) = Z(J_4) = Z(J_6) = Z(J_8) = 1^2+(1^2+1^2)+(3^2+2^2)+2^2 +$$

$$+ (2^2+1^2)+(1^2+1^2)+(1^2+2^2) = 32$$

$$Z(J_3) = Z(J_7) = 1^2+(3^2+1^2)+(2^2+1^2)+(2^2+2^2) +$$

$$+ (1^2+2^2)+(3^2+1^2)+1^2 = 40 \ ,$$

logo J_2(ou J_4 ou J_6 ou J_8) é mediana, com $Z_{mín} = 32$.

- Grupo g_K:

$$Z(K_1) = Z(K_3) = 2^2+(4^2+4^2)+(2^2+4^2) = 56$$

$$Z(K_2) = Z(K_4) = (2^2+4^2)+4^2+2^2 = 40$$

a mediana é K_2 (ou K_4), com $Z_{mín} = 40$.

O grupo menos disperso nesse caso é o grupo g_J.

Em resumo, nota-se que:

- o número de elementos influiu na soma dos quadrados dentro do grupo, o que fêz com que g_J e g_K apresentassem, nessa medida, a mesma dispersão interna. Tal não aconteceu, obviamente, com a medida fornecida pela variância interna;

- no caso do diâmetro de grupo, não há sensibilidade para a distribuição interna dos elementos no grupo: os três agrupamentos apresentaram a mesma dispersão;

- as únicas medidas que apresentaram apenas o grupo g_J como o menos disperso interiormente foram a variância interna e a medida de dispersão via mediana de grupo. Deve-se acrescentar que g_J tem, para um círculo de centro na média $\bar{X}_J$ e raio $\sqrt{2}$, cerca de 75% dos pontos em seu interior ou na fronteira, enquanto os outros dois grupos não têm ponto algum nessa situação.

II.5 - FUNÇÕES OBJETIVO EM ANÁLISE DE GRUPAMENTO

Como foi visto na seção II.2, a solução de um problema de análise de grupamento pode ser vista como uma partição P_m^* do conjunto E que otimize uma função objetivo, ou seja, P_m^* é solução do problema:

$$\text{Otimizar} \quad f(P_m)$$

$$\text{Sujeito a } P_m \text{ viável.}$$

A partir das medidas de dispersão apresentadas no item II.4, é possível definir algumas funções objetivo passíveis de utilização, tais como as funções, a serem minimizadas, descritas a seguir:

(i) soma dos quadrados dentro dos grupos (v. p.ex. DURAN e ODELL[10]):

$$f(P_m) = W = \sum_{g_I \in P_m} W_I = \sum_{g_I \in P_m} \sum_{i=1}^{n_I} d_2^2(X_i, \bar{X}_I);$$

(ii) soma das variâncias internas:

$$f(P_m) = S^2 = \sum_{g_I \varepsilon P_m} S_I^2 = \sum_{g_I \varepsilon P_m} \frac{1}{n_I} W_I \ ;$$

(iii) diâmetro de grupo (v. p.ex. HANSEN e DELATTRE[15]):

$$f(P_m) = d(P_m) = \max_{g_I \varepsilon P_m} d(g_I) = \max_{g_I \varepsilon P_m} \max_{e_i, e_j \varepsilon g_I} d_{ij}; \text{ e}$$

(iv) dispersão via mediana de grupo (v. p.ex. VINOD[41]):

$$f(P_m) = \sum_{g_I \varepsilon P_m} Z_{mín}(g_I) = \sum_{g_I \varepsilon P_m} \text{mín}_{e_j \varepsilon g_I} \sum_{i=1}^{n_I} d_{ij} \ .$$

III - CONSIDERAÇÕES GERAIS SOBRE OS MÉTODOS DE ANÁLISE DE GRUPAMENTO

III.1 - INTRODUÇÃO

A variedade de métodos para a resolução de problemas de análise de grupamento é enorme. Vários textos se dedicaram a uma apresentação do conjunto desses métodos, entre os quais poderiam ser destacados os textos de ANDERBERG[1], DIDAY e SIMON[9], DURAN e ODELL[10], HARTIGAN[17] e PINHO GAMA[30].

O objetivo do presente trabalho é apresentar detalhadamente algumas técnicas selecionadas, classificadas aqui arbitrariamente em três famílias de métodos:

- hierarquizados aglomerativos;
- de realocação iterativa; e
- de programação matemática.

III.2 - MÉTODOS HIERARQUIZADOS AGLOMERATIVOS

Os métodos hierarquizados aglomerativos são inicializados considerando-se os n elementos de E como formando um conjunto de n grupos $\{e_1\}$, $\{e_2\}, \ldots, \{e_n\}$. Seleciona-se então dois elementos e_i e e_j, $i \neq j$, considerados mais semelhantes e une-se esses dois elementos em um único grupamento. Forma-se então uma partição de E em (n-1) grupos $\{e_1\}, \{e_2\}, \ldots, \{e_i, e_j\}, \ldots, \{e_n\}$. O processo é então sequencialmente repetido, isto é, são unidos os grupos mais semelhantes, formando-se (n-2) grupos, (n-3), (n-4), etc., terminando-se o processo quando se determina novamente um único grupamento E.

O termo "hierarquizado" vem do fato de que o processo define uma hierarquia, na medida em que um grupo formado em um determinado passo corresponde a uma união de grupos formados em passos anteriores.

Por outro lado, e como apresentado por PINHO GAMA[30], uma das dificuldades da aplicação desses métodos reside em que eles não apresentam uma "regra de parada" , isto é, são aplicados, em princípio, até que todos os elementos do conjunto E pertençam a um único grupamento.

De uma maneira geral, um importante item na resolução de problemas de análise de grupamento é o da determinação do número de grupos m. PINHO GAMA[30] apresenta alguns métodos para a definição desse número, apontando, no entanto, que este é um problema ainda em aberto na Análise de Grupamento, e que os métodos atualmente existentes são discutíveis quanto à capacidade de determinação de um verdadeiro valor de m.

No capítulo IV, onde serão descritos os principais métodos hierarquizados aglomerativos, serão apresentados alguns procedimentos simples que permitem, na prática, a determinação do número de grupos. Por outro lado, é preciso considerar também que, em muitos casos, m já será determinado a priori.

III.3 - MÉTODOS DE REALOCAÇÃO ITERATIVA

Uma outra maneira de se resolver um problema de análise de grupamento seria, por exemplo, para um valor fixo de m, adotar-se a seguinte técnica: forma-se inicialmente uma partição qualquer $P_m = \{g_1, g_2, \ldots, g_m\}$ do conjunto E; calcula-se então os "centros" de cada grupo, que podem ser, por exemplo, as medianas dos mesmos; gera-se a partir daí um novo grupamento $P'_m = \{g'_1, g'_2, \ldots, g'_m\}$, alocando-se cada elemento e_j ao centro de grupo mais próximo; o processo é repetido (cálculo do novo "centro" e realocação ao "centro" mais próximo) até que algum teste de convergência seja satisfeito.

Esse tipo de procedimento é definido por DIDAY et alii[8] como sendo de "realocação iterativa". Existem, por outro lado, variantes da técnica apresentada acima, tais como os métodos ISODATA (descrito, por exemplo, em DURAN e ODELL[10]), K - Médias ou de MACQUEEN[27], de JANCEY[19] e de FORGY[13] (esses também são descritos em ANDERBERG[1]).

Alguns dos principais métodos de realocação iterativa serão apresentados no Capítulo V. Deve-se, no entanto, esclarecer aqui que para os métodos desse tipo, embora a convergência seja atingida em um número finito de iterações, não existe garantia de que a solução obtida seja ótima.

III.4 - MÉTODOS DE PROGRAMAÇÃO MATEMÁTICA

Os métodos hierarquizados, obviamente, não fornecem necessariamente soluções ótimas, ou seja, não geram obrigatoriamente a melhor dentre todas as partições admissíveis. Como citado no ítem anterior, os métodos de realocação iterativa também não o fazem. Além disso, não é possível, nos dois casos, avaliar-se a qualidade da solução obtida, que pode inclusive possuir um valor de função objetivo muito distante do ótimo.

Uma tentativa de contornar essa dificuldade seria efetuar a análise de grupamento via enumeração completa: para todas as alternativas possíveis de partições P_m do conjunto E, calcula-se o valor da função objetivo $f(P_m)$, escolhendo-se como solução a partição P^*_m tal que $f(P^*_m)$ seja ótima.

No entanto, o número S(n, m) de alternativas de partição dos n elementos de E em m grupos é, segundo DURAN e ODELL[10], dado por:

$$S(n, m) = \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} k^n .$$

Assim, a análise de grupamento por enumeração completa não se mostra prática, a menos que n e m sejam muito pequenos. Por exemplo, se n=16 e m=8, o número de alternativas é de aproximadamente $2{,}1 \times 10^9$, ou seja , 2.100.000.000.

Tal fato levou ao desenvolvimento de uma série de algoritmos baseados em programação dinâmica (p.ex. BELLMAN[6], JENSEN[21] e RAO[33]), teoria dos grafos (p.ex. HANSEN e DELATTRE[15], JOHNSON[22], ROHLF[34], SIBSON[36] e ZAHN[46]) e programação inteira (p.ex. RAO[33], ROY[35] e VINOD[41]).

Dentre essas técnicas, foram selecionados para apresentação nesta tese dois modelos de programação inteira devidos a VINOD[41]. Esses modelos apresentam a desvantagem inerente à tecnica de programação matemática utilizada: excessivo tempo de computação. No entanto, RAO[33] e MULVEY e CROWDER[29], utilizando, respectivamente, programação dinâmica e otimização por subgradientes, demonstraram que é possível, para os dois modelos de Vinod, obter-se uma solução ótima de forma eficiente. Esses métodos serão apresentados no Capítulo VI.

IV - MÉTODOS HIERARQUIZADOS AGLOMERATIVOS

O procedimento padrão nos algoritmos hierarquizados aglomerativos foi descrito sucintamente no item III.2. Ele consiste basicamente numa técnica iterativa que, em cada passo, reune em um único grupo os dois grupamentos mais semelhantes. O algoritmo se inicia considerando cada grupo como formado por um único elemento e, ao seu término, terá reunido todos os elementos em um só grupo. Evidentemente, o algoritmo também poderá ser interrompido quando, em determinado passo, se tiver alcançado um número m de grupos considerado conveniente.

Talvez o método mais utilizado para a representação gráfica dos resultados de um processo hierarquizado aglomerativo seja o que utiliza a idéia do "dendograma". Em sua forma mais usual, o dendograma consiste em um diagrama em forma de árvore, onde os elementos são apresentados verticalmente à esquerda, e os resultados do processo à direita. Os níveis de distâncias em que os grupos são formados, são apresentados horizontalmente acima do diagrama. A figura a seguir (exemplo dado por DURAN e ODELL[10]) apresenta um dendograma para o caso de seis elementos e p características:

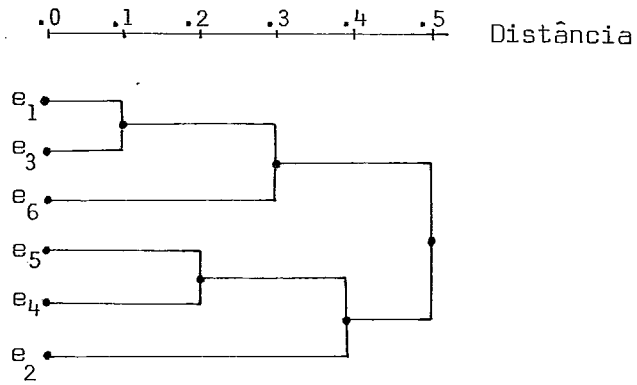


Fig. IV.1 - Exemplo de um dendograma

O dendograma informa que na primeira etapa foram reunidos os elementos e_1 e e_3, com uma distância $d_{13} = 0,1$. Na segunda etapa foram reunidos e_4 e e_5, com $d_{45} = 0,2$. No terceiro passo efetuou-se a reunião de $\{e_1, e_3\}$ e $\{e_6\}$, ao nível de distância 0,3. Posteriormente, tem-se $\{e_5, e_4\} \cup \{e_2\}$ e finalmente todos os elementos estão reunidos em um só grupo.

Como se nota, o algoritmo trabalha com distâncias entre grupos e não distâncias entre elementos, tal como definido no item II.3.1. Assim , é usual, nesses problemas, definir-se distâncias d^g_{IJ} entre os grupos quais

quer g_I e g_J, distintas das distâncias d_{ij} entre os elementos e_i e e_j. Da mesma forma, d^g_{iJ} representa a distância entre o grupo formado por $\{e_i\}$ e o grupo g_J.

Neste capítulo, serão apresentados seis métodos de agrupamento hierarquizado aglomerativo. O algoritmo geral é o mesmo, variando apenas a distância d^g_{IJ} utilizada. No item IV.3, apresenta-se uma tabela-resumo com as distâncias associadas a cada um dos seis métodos. Ao final do capítulo, é feito um estudo comparativo desses métodos, apresentando-se também alguns procedimentos práticos para a determinação do número de grupos a ser adotado na solução do problema.

IV.1 - ALGORITMO GERAL

O algoritmo geral utilizado nos métodos hierarquizados aglomerativos é o que se segue. Deve-se, no entanto, alertar que foi efetuado aqui um artifício, normalmente utilizado quando da implantação do método em computador: ao se unir dois grupos g_I e g_J, formando $g_L = g_I \cup g_J$, assume-se, para efeito de armazenamento das informações, que, após a união, tem-se:

$$g_I = g_L \quad e$$

$$g_J = \{\emptyset\}.$$

Algoritmo:

Passo 0 : (inicialização)
Seja cada grupo formado por um único elemento e_j, $j = 1, \ldots, n$;
Calcule a matriz de distâncias entre grupos;
Faça $k = n$

Passo 1 : (cálculo da distância mínima entre grupos)
Determine g_I e g_J tais que $d^g_{IJ} = \min_{\substack{P \neq Q \\ P,Q \neq \{\emptyset\}}} \{ d^g_{PQ} \}$

Passo 2 : (união dos grupos para os quais a distância foi mínima)
Forme o grupo $g_L = g_I \cup g_J$;
Faça $g_I = g_L$ e $g_J = \{\emptyset\}$
(observe que existem agora k-1 grupos não vazios)

Passo 3 : (regra de parada)

Se k-1 = m, pare;

senão vá para o Passo 4

Passo 4 : (cálculo da nova matriz de distâncias entre os grupos)

Calcule a nova matriz de distâncias entre os grupos não vazios;

faça k = k-1 e volte para o Passo 1.

IV.2 - PRINCIPAIS MÉTODOS HIERARQUIZADOS AGLOMERATIVOS

Os métodos a serem apresentados aqui são os seguintes:

- ligação simples
- ligação completa
- centróide
- mediana
- Ward
- média de grupo.

Todos esses métodos, como já foi dito, seguem a mesma técnica, variando apenas a distância entre grupos d^{g}_{IJ} a ser utilizada.

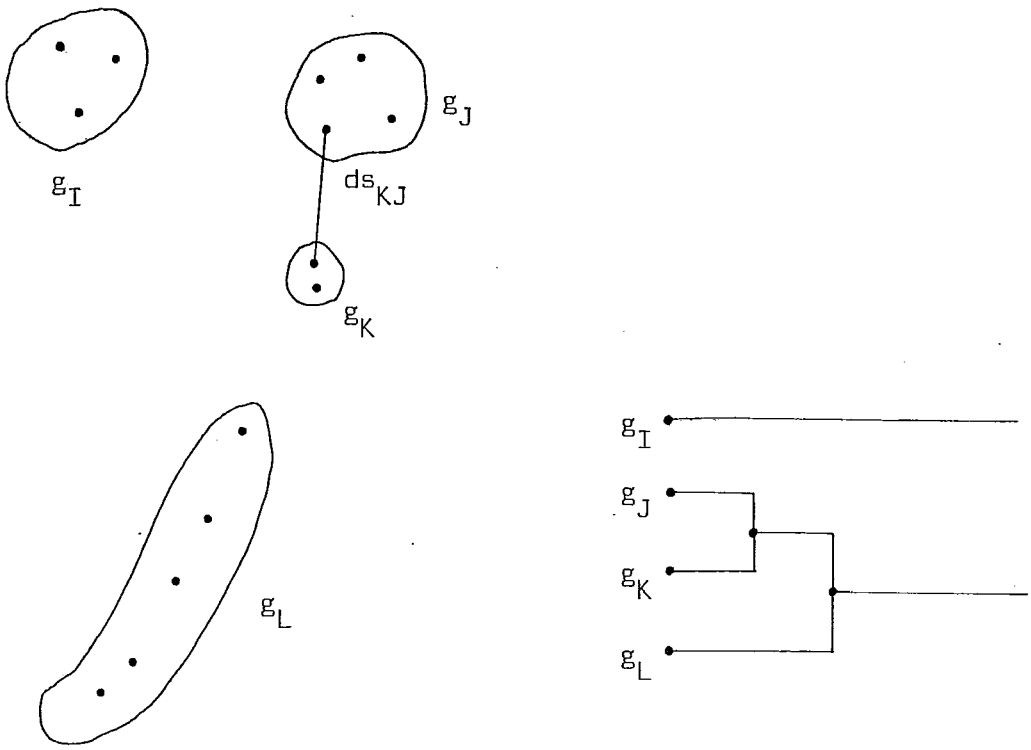
IV.2.1 - MÉTODO DA LIGAÇÃO SIMPLES:

Introduzido por JOHNSON[22], esse método utiliza a distância "de vizinhança mais próxima" (ver também DURAN e ODELL[10]), aqui denotada por ds_{IJ}:

$$d^{g}_{IJ} = ds_{IJ} = \min_{\substack{e_i \in g_I \\ e_j \in g_J}} d_{ij}.$$

Pode-se, assim, nesse caso, fazer uso de quaisquer das distâncias entre elementos d_{ij} mencionadas no item II.3.

Exemplo: Seja a iteração adiante, onde já foram formados quatro grupos (a distância utilizada é a euclidiana), e se deseja determinar um agrupamento final composto por dois grupos:



Assim, a solução é dada por $P_2 = \{g_I\}, \{g_J, g_K, g_L\}$.

Note-se que a partição P_2 não é homogênea, tal como definido no item II.2.3. Além disso, nenhuma partição em dois grupos que seja construída pela união dos grupos g_I, g_J, g_K e g_L o será, dada a conformação alongada do grupo g_L.

De uma maneira geral, o método da ligação simples poderá, ainda nos primeiros estágios, reunir em um grupo elementos bastante dessemelhantes, desde que haja entre eles uma cadeia de outros elementos que sejam, por sua vez, semelhantes entre si. Esta dificuldade é conhecida como "efeito de cadeia" (ver, p.ex. DIDAY e SIMON[9], DURAN e ODELL[10] e HANSEN e DELATTRE[15]).

IV.2.2 - MÉTODO DA LIGAÇÃO COMPLETA

Devido a MACNAUGHTON - SMITH[26], o método utiliza a distância "de vizinhança mais afastada" (v. tb. DURAN e ODELL[10]), aqui denotada por dc_{IJ}:

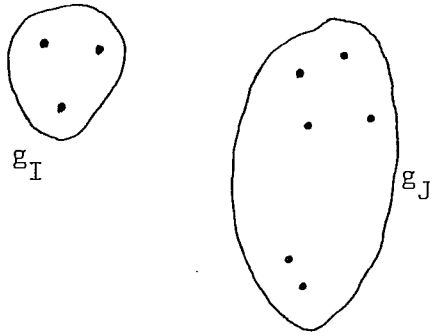
$$d^{g}_{IJ} = dc_{IJ} = \max_{\substack{e_i \in g_I \\ e_j \in g_J}} d_{ij}$$

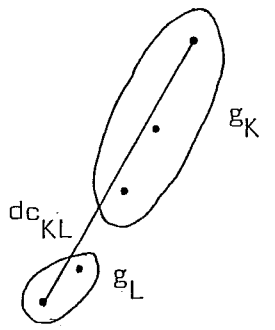
Também neste caso pode-se fazer uso de quaisquer das distâncias mencionadas no item II.3. Note-se que este método procura minimizar a função objetivo "diâmetro de grupo": $f(P_m) = d(P_m) = \max_{g_I \in P_m} d(g_I)$, descrita no item II.5.

O método da ligação completa forma grupos mais compactos que o método anterior. Aplicado ao exemplo do item IV.2.1, ter-se-ia:

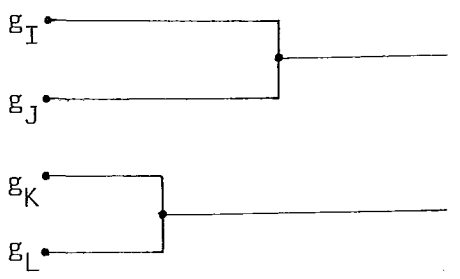
Exemplo

A partição em quatro grupos (distância euclidiana) seria:





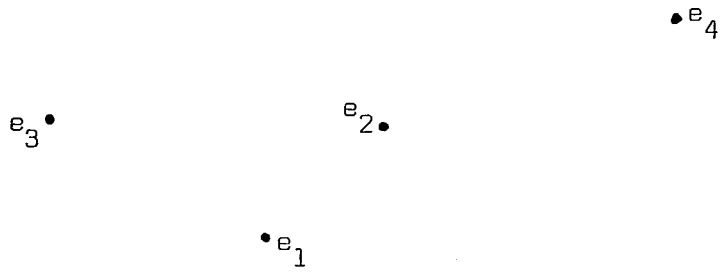
Sequência de partições, para m = 2:



A solução agora é dada por $P_2 = \{g_I, g_J\}, \{g_K, g_L\}$.

Por outro lado, o método não fornece necessariamente soluções ótimas. O exemplo abaixo, citado por HANSEN e DELATTRE[15], ilustra esse fato.

Exemplo: Sejam 4 elementos, com as dessemelhanças $d_{12}=1$, $d_{13}=2$, $d_{24}=3$, $d_{23}=4$, $d_{14}=5$ e $d_{34}=6$:



As partições dadas pelo método da ligação completa seriam:

$P_4 = \{e_1\}, \{e_2\}, \{e_3\}, \{e_4\}$ e $d(P_4) = 0$

$P_3 = \{e_1, e_2\}, \{e_3\}, \{e_4\}$ e $d(P_3) = d_{12} = 1$

$P_2 = \{e_1, e_2, e_3\}, \{e_4\}$ e $d(P_2) = d_{23} = 4$

No entanto, para $P_2' = \{e_1, e_3\}, \{e_2, e_4\}$ tem-se $d(P_2') = d_{24} = 3$.

IV.2.3 - MÉTODO DA CENTRÓIDE

No método da centróide, devido a SOKAL e MICHENER[37], a distância entre os grupos g_I e g_J é dada em termos do quadrado da distância euclidiana entre suas centróides (v. tb. DURAN e ODELL[10]):

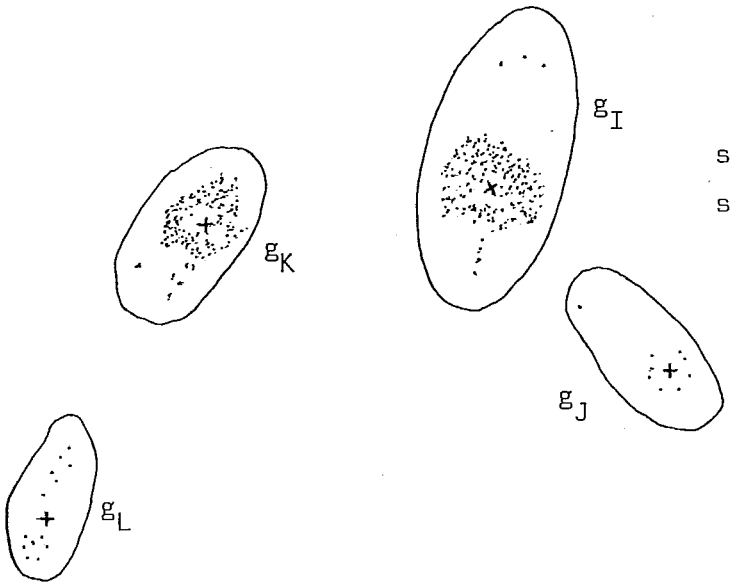
$$d_{IJ}^{g} = d_{IJ}^{2} = d_2^2 (\bar{X}_I, \bar{X}_J),$$

onde $\bar{X}_I$ e $\bar{X}_J$ são, respectivamente, as médias, ou centróides, dos grupos g_I e g_J. Cabe lembrar aqui que:

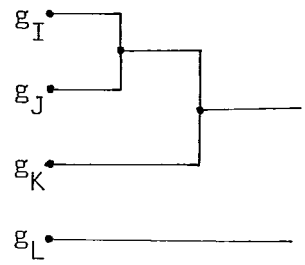
$$\bar{X}_I = \frac{1}{n_I} \sum_{i=1}^{n_I} X_i .$$

Note-se que se n_I for muito maior que n_J as centróides de $g_L = g_I \cup g_J$ e g_I são quase coincidentes, e assim as características do grupo g_J não são levadas em consideração. Isto levou a que se caracterizasse esse método como "técnica de grupos ponderados" (v. DURAN e ODELL[10]).

Exemplo: Sejam



sequência de partições, caso se deseje 2 grupos:



Note-se que tanto pelo método da ligação simples como pelo da ligação completa, a partição P_2 seria $\{ g_I, g_J \}, \{ g_K, g_L \}$.

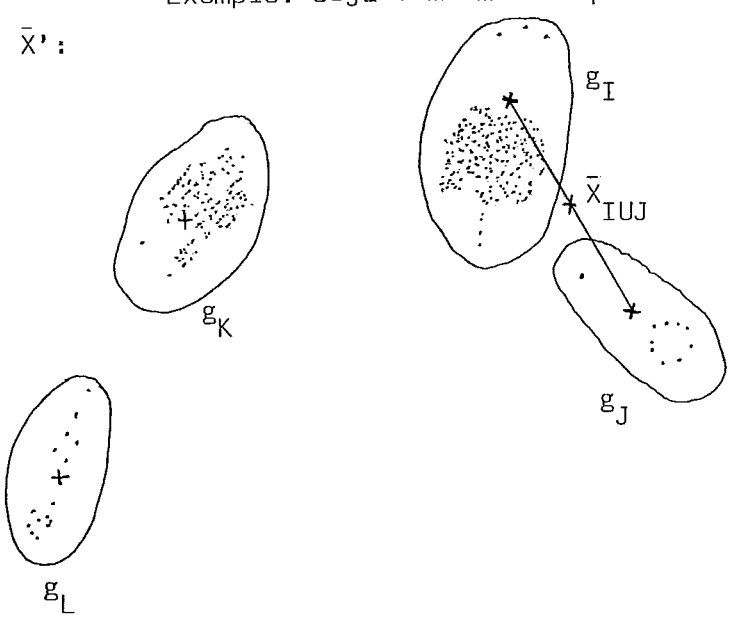
IV.2.4 - MÉTODO DA MEDIANA

Como forma de contornar a dificuldade criada pela formação de grupos ponderados, LANCE e WILLIAMS[25] apresentaram o método da mediana: a filosofia é a mesma, com a exceção de que, na união dos grupos g_I e g_J, assume-se que a centróide do grupo resultante dessa união é dada pela média aritmética das centróides de g_I e g_J. Assim, se $g_M = g_I \cup g_J$, então

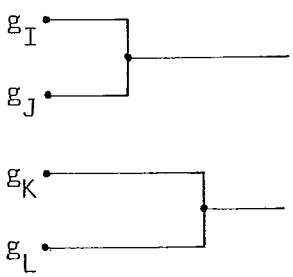
$$\bar{X}_M = \frac{\bar{X}_I + \bar{X}_J}{2} .$$

Em outras palavras, nota-se que $\bar{X}_M$ corresponde ao ponto médio do segmento que une $\bar{X}_I$ a $\bar{X}_J$.

Exemplo: Seja o mesmo exemplo anterior com as novas centróides $\bar{X}'$:



sequência de partições:



IV.2.5 - MÉTODO DE WARD

O método de WARD[42] por sua vez, utiliza como distância entre os grupos a distância estatística (ver tb. DURAN e ODELL[10]):

$$d^{g}_{IJ} = D_{IJ} = \frac{n_I n_J}{n_I + n_J}\, d_2^2\,(\bar{X}_I, \bar{X}_J) = \frac{n_I n_J}{n_I + n_J}\, d^2_{IJ}\ .$$

O objetivo do método de Ward é procurar a minimização da soma dos quadrados dentro dos grupos, definida no item II.4.1:

$$W = \sum_{g_I \varepsilon P_m} W_I = \sum_{g_I \varepsilon P_m} \sum_{i=1}^{n_I} d_2^2\,(X_i, \bar{X}_I),$$

onde $\bar{X}_I$ é a centróide do grupo g_I.

Esse objetivo fica evidente em vista do Teorema IV.1, apresentado a seguir, que indica o significado da distância D_{IJ}.

Teorema IV.1:

Se $g_L = g_I \cup g_J$,

então $W_L = W_I + W_J + D_{IJ}$,

ou seja, D_{IJ} representa o acréscimo na soma dos quadrados dentro dos grupos W quando g_I e g_J são unidos.

Demonstração:

Se $g_L = g_I \cup g_J$, tem-se, por definição, que:

$$W_L = \sum_{i=1}^{n_I} d_2^2\,(X_i, M) + \sum_{j=1}^{n_J} d_2^2\,(X_j, M), \qquad \text{(IV.1)}$$

onde

$$M = \frac{1}{n_I + n_J}\left(\sum_{i=1}^{n_I} X_i + \sum_{j=1}^{n_J} X_j\right)$$

é a centróide do grupo g_L.

O teorema será demonstrado em duas partes:

1a. Parte: $\sum_{i=1}^{n_I} d_2^2\,(X_i, M) = W_I + n_I\, d_2^2\,(\bar{X}_I, M)$ e

$$\sum_{j=1}^{n_J} d_2^2\,(X_j, M) = W_J + n_J\, d_2^2\,(\bar{X}_J, M).$$

Por definição,

$$\sum_{i=1}^{n_I} d_2^2 (X_i, M) = \sum_{i=1}^{n_I} (X_i - M)^t (X_i - M)$$

Somando e subtraindo $\bar{X}_I$,

$$\sum_{i=1}^{n_I} d_2^2 (X_i, M) = \sum_{i=1}^{n_I} (X_i - \bar{X}_I + \bar{X}_I - M)^t (X_i - \bar{X}_I + \bar{X}_I - M)$$

$$= \sum_{i=1}^{n_I} \sum_{k=1}^{p} (x_{ki} - \bar{x}_{kI} + \bar{x}_{kI} - m_k)^2$$

$$= \sum_{i=1}^{n_I} \sum_{k=1}^{p} (x_{ki}^2 - 2x_{ki}\,\bar{x}_{kI} + 2\bar{x}_{kI}^2 - 2\bar{x}_{kI}\,m_k + m_k^2) + 2\emptyset ,$$

onde $\emptyset = \sum_{i=1}^{n_I} \sum_{k=1}^{p} (x_{ki}\,\bar{x}_{kI} - x_{ki}\,m_k - \bar{x}_{kI}^2 + \bar{x}_{kI}\,m_k)$.

No entanto, colocando apenas o somatório em k em evidência, tem-se:

$$\emptyset = \sum_{k=1}^{p} (\bar{x}_{kI} \sum_{i=1}^{n_I} x_{ki} - m_k \sum_{i=1}^{n_I} x_{ki} - n_I\,\bar{x}_{kI}^2 + n_I\,\bar{x}_{kI}\,m_k)$$

ou seja,

$$\emptyset = \sum_{k=1}^{p} \emptyset_k ,$$

com

$$\emptyset_k = \bar{x}_{kI}\,n_I\,\bar{x}_{kI} - m_k\,n_I\,\bar{x}_{kI} - n_I\,\bar{x}_{kI}^2 + n_I\,\bar{x}_{kI}\,m_k = 0.$$

Assim, $\emptyset = 0$ e $\sum_{i=1}^{n_I} d_2^2 (X_i, M)$ se torna:

$$\sum_{i=1}^{n_I} d_2^2 (X_i, M) = \sum_{i=1}^{n_I} \sum_{k=1}^{p} (x_{ki}^2 - 2x_{ki}\,\bar{x}_{kI} + 2\bar{x}_{kI}^2 - 2\bar{x}_{kI}m_k + m_k^2)$$

$$= \sum_{i=1}^{n_I} \sum_{k=1}^{p} \{ (x_{ki}^2 - 2x_{ki}\,\bar{x}_{kI} + \bar{x}_{kI}^2) + (\bar{x}_{kI}^2 - 2\bar{x}_{kI}\,m_k + m_k^2) \}$$

$$= \sum_{i=1}^{n_I} \sum_{k=1}^{p} (x_{ki} - \bar{x}_{kI})^2 + \sum_{i=1}^{n_I} \sum_{k=1}^{p} (\bar{x}_{kI} - m_k)^2$$

$$= \sum_{i=1}^{n_I} \sum_{k=1}^{p} (x_{ki} - \bar{x}_{kI})^2 + n_I \sum_{k=1}^{p} (\bar{x}_{kI} - m_k)^2$$

$$= \sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I) + n_I \, d_2^2 (\bar{X}_I, M)$$

$$= W_I + n_I \, d_2^2 (\bar{X}_I, M). \qquad \text{(IV.2)}$$

Da mesma forma, poder-se-ia mostrar que:

$$\sum_{j=1}^{n_J} d_2^2 (X_j, M) = W_J + n_J \, d_2^2 (\bar{X}_J, M) , \qquad \text{(IV.3)}$$

o que completa a primeira parte da demonstração.

2a. Parte: $W_L = W_I + W_J + D_{IJ}$

Como foi indicado na equação (IV.1) ,

$$W_L = \sum_{i=1}^{n_I} d_2^2 (X_i, M) + \sum_{j=1}^{n_J} d_2^2 (X_j, M).$$

Tendo em vista as equações (IV.2) e (IV.3), a equação acima pode ser reescrita:

$$W_L = W_I + n_I \, d_2^2 (\bar{X}_I, M) + W_J + n_J \, d_2^2 (\bar{X}_J, M).$$

Como $M = \dfrac{1}{n_I + n_J} (n_I \bar{X}_I + n_J \bar{X}_J)$,

tem-se que

$$n_I \, d_2^2 (\bar{X}_I, M) = n_I \sum_{k=1}^{p} (\bar{x}_{kI} - m_k)^2$$

$$= n_I \sum_{k=1}^{p} (\bar{x}_{kI} - \frac{n_I \bar{x}_{kI} + n_J \bar{x}_{kJ}}{n_I + n_J})^2$$

$$= n_I \sum_{k=1}^{p} (\frac{n_J \bar{x}_{kI} - n_J \bar{x}_{kJ}}{n_I + n_J})^2$$

$$= \frac{n_I n_J^2}{(n_I+n_J)^2} \sum_{k=1}^{p} (\bar{x}_{kI} - \bar{x}_{kJ})^2$$

$$= \frac{n_I n_J^2}{(n_I+n_J)^2} d_{IJ}^2$$

e, da mesma forma,

$$n_J \, d_2^2 (\bar{X}_J, M) = \frac{n_I^2 n_J}{(n_I+n_J)^2} d_{IJ}^2 .$$

Assim,

$$W_L = W_I + W_J + \frac{n_I n_J^2 + n_I^2 n_J}{(n_I + n_J)^2} d_{IJ}^2$$

$$= W_I + W_J + \frac{n_I n_J (n_I + n_J)}{(n_I + n_J)^2} d_{IJ}^2$$

$$= W_I + W_J + \frac{n_I n_J}{n_I+n_J} d_{IJ}^2 .$$

Finalmente,

$$W_L = W_I + W_J + D_{IJ} ,$$

o que completa a demonstração.

O resultado acima pode ser interpretado da seguinte forma (v. DURAN e ODELL[10]): a soma total de quadrados das distâncias do novo grupo g_L é igual à soma dos quadrados das distâncias "intra" mais a soma dos quadrados das distâncias "inter" (dada por D_{IJ}) dos grupos g_I e g_J.

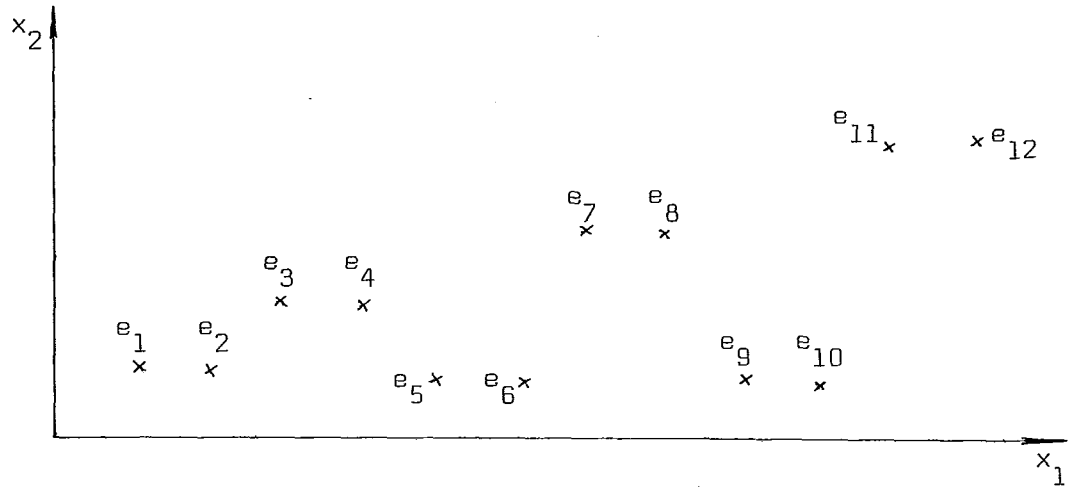
Assim, quando o método de Ward, em cada iteração, escolhe para a formação de um novo grupo os grupos g_I e g_J tais que D_{IJ} é mínimo, ele efetivamente está levando a que, em cada iteração, o acréscimo em W seja mínimo. Logo, o método procura minimizar a soma dos quadrados dentro dos grupos, embora, obviamente, não forneça necessariamente uma solução ótima para o problema.

Exemplo:

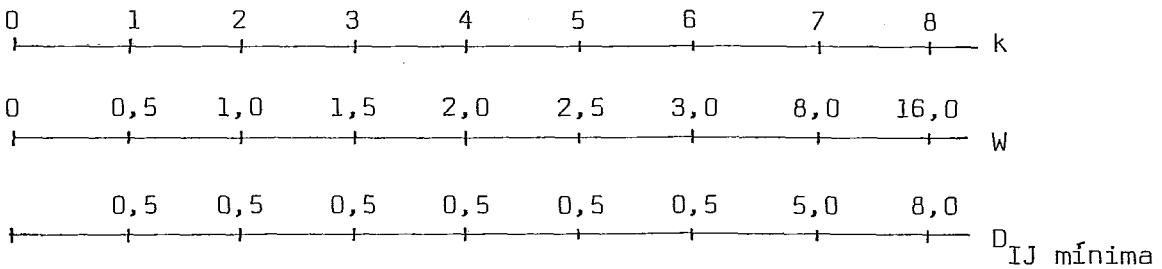
Sejam, por exemplo, $e_1, e_2, \ldots, e_{12}$ tais que:

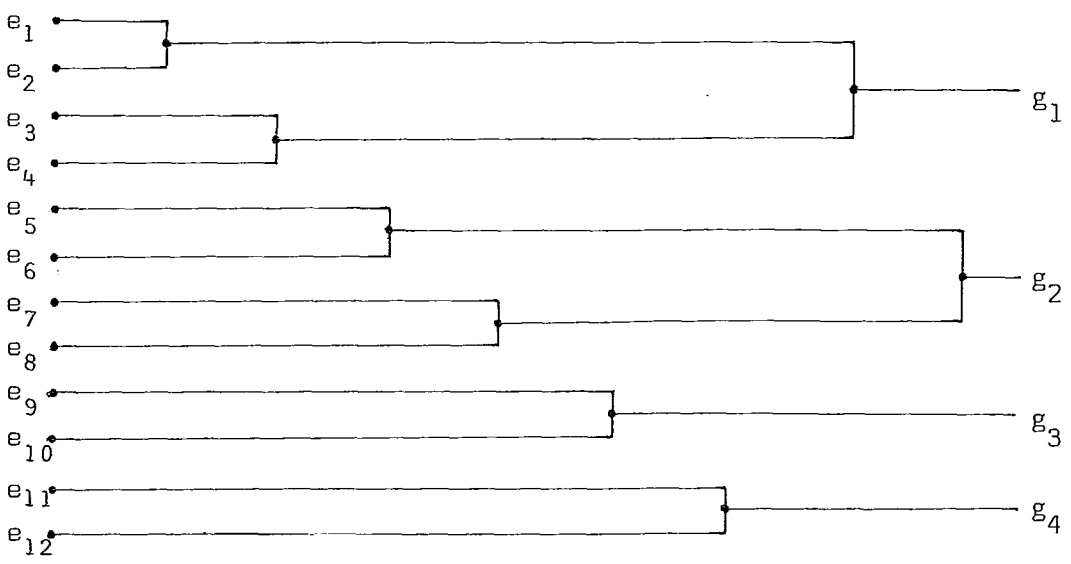
$$X = [X_1 | X_2 | \ldots | X_{12}] = \left[\begin{matrix}1\\1\end{matrix} \middle| \begin{matrix}2\\1\end{matrix} \middle| \begin{matrix}3\\2\end{matrix} \middle| \begin{matrix}4\\2\end{matrix} \middle| \begin{matrix}5\\1\end{matrix} \middle| \begin{matrix}6\\1\end{matrix} \middle| \begin{matrix}7\\3\end{matrix} \middle| \begin{matrix}8\\3\end{matrix} \middle| \begin{matrix}9\\1\end{matrix} \middle| \begin{matrix}10\\1\end{matrix} \middle| \begin{matrix}11\\4\end{matrix} \middle| \begin{matrix}12\\4\end{matrix}\right]$$

Graficamente, tem-se:

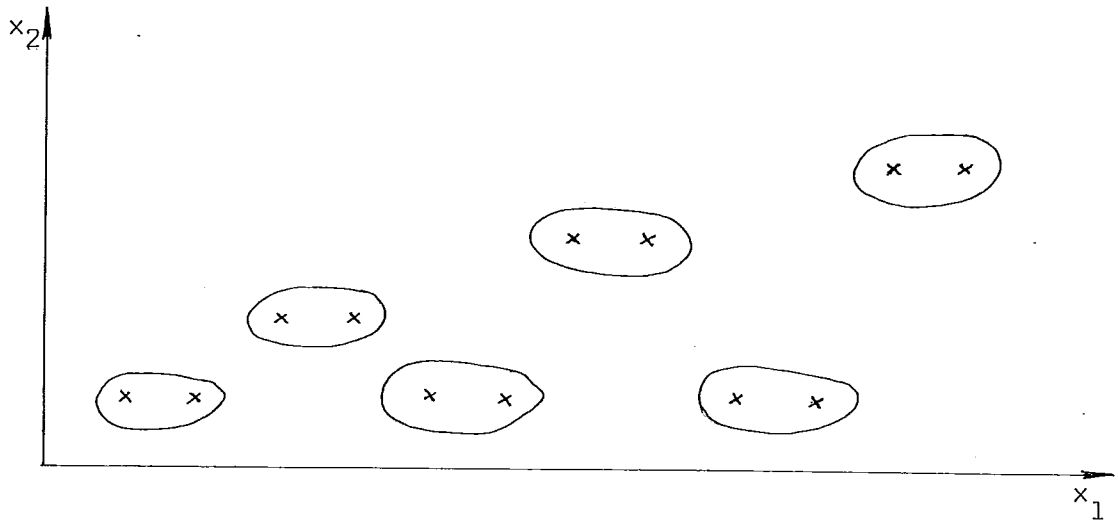


O problema foi resolvido com o auxílio da rotina CLSTROT, desenvolvida nesta tese e aqui apresentada no Apênd. 2. Os resultados obtidos, para m = 4, podem ser sumarizados na forma abaixo, onde estão também indicados o passo k, e a soma dos quadrados dentro dos grupos W e a distância estatística D_{IJ} mínima associadas ao passo k:



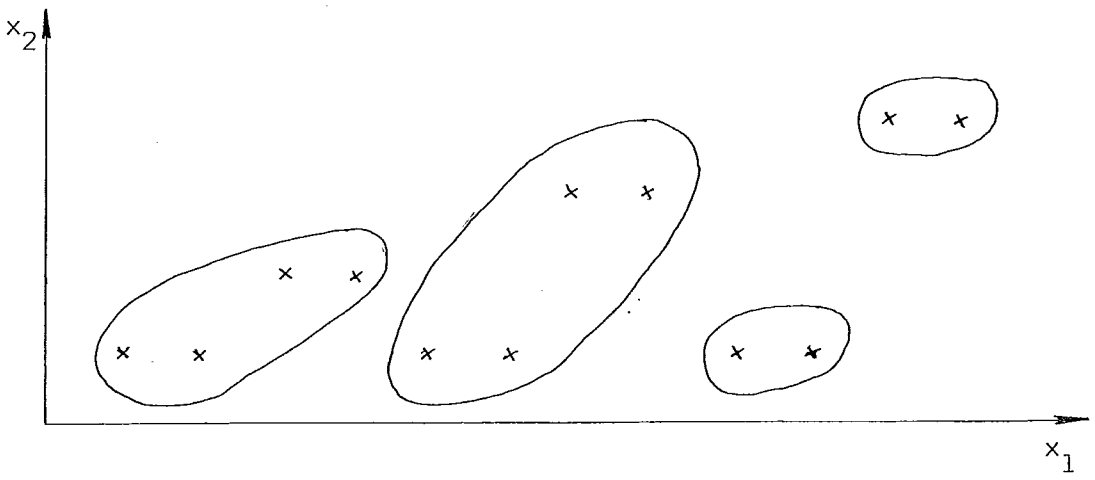


Assim, se fosse desejada uma partição em seis grupos, ter-se-ia:



Note-se que, ao se formarem esse grupos, o nível de distância D_{IJ} entre os elementos que foram unidos foi de D_{IJ}= 0,5, uma vez que n_I = n_J = 1 e d^2_{IJ}= 1 (ver definição de D_{IJ} no início deste item). Ao se formarem cinco grupos, D_{IJ} passa a ser igual a 5,0. Isto é, passou a ser dez vêzes maior que o nível em que eram feitas as uniões de grupos anteriormente. Isso indica que, com uma partição em seis grupos, se obtém grupos "muito" mais similares que com uma partição em cinco ou menos grupos. Uma análise como essa, ou seja, uma verificação do crescimento de D_{IJ} ou de W pode sugerir o número final de grupos a ser definido.

O resultado, para m = 4, seria:



Nota-se assim que a solução mais adequada parece ser a partição em seis grupos.

IV.2.6 - MÉTODO DA MÉDIA DE GRUPO

Devido a SOKAL e MICHENER[37], o método da média de grupo utiliza como distância d_{IJ} entre os grupos g_I e g_J a distância média quadrática (v. tb. DURAN e ODELL[10]):

$$d^{g}_{IJ} = D^2_{IJ} = \frac{1}{n_I n_J} \sum_{i=1}^{n_I} \sum_{j=1}^{n_J} d_2^2 (X_i, X_j).$$

O objetivo deste método pode ser constatado a partir dos seguintes teoremas:

Teorema IV.2 (v. tb. DURAN e ODELL[10]):

A distância média quadrática entre os grupos g_I e g_J é igual à soma das variâncias internas S_I^2 e S_J^2 desses grupos mais o quadrado da distância euclidiana $d_{IJ}^2 = d_2^2 (\bar{X}_I, \bar{X}_J)$ entre as suas centróides.

$$D^2_{IJ} = S^2_I + S^2_J + d^2_{IJ}$$

Demonstração:

O teorema será demonstrado em duas partes:

1a. Parte: Considerando $g_I = \{ e_i \}$, tem-se que a distância média quadrática entre D^2_{iJ} entre o elemento e_i e o grupo g_J é dada por:

$$D^2_{iJ} = S^2_J + d_2^2 (X_i, \bar{X}_J):$$

Por definição, como $n_I = 1$,

$$D^2_{iJ} = \frac{1}{n_J} \sum_{j=1}^{n_J} d_2^2 (X_j, X_i)$$

$$= \frac{1}{n_J} \sum_{j=1}^{n_J} (X_j - X_i)^t (X_j - X_i).$$

Somando e subtraindo $\bar{X}_J$,

$$D^2_{iJ} = \frac{1}{n_J} \sum_{j=1}^{n_J} (X_j - \bar{X}_J + \bar{X}_J - X_i)^t (X_j - \bar{X}_J + \bar{X}_J - X_i)$$

$$= \frac{1}{n_J} \sum_{j=1}^{n_J} \sum_{k=1}^{p} (x_{kj} - \bar{x}_{kJ} + \bar{x}_{kJ} - x_{ki})^2$$

$$= \frac{1}{n_J} \sum_{j=1}^{n_J} \sum_{k=1}^{p} (x_{kj}^2 - 2x_{kj}\,\bar{x}_{kJ} + 2\bar{x}_{kJ}^2 - 2\bar{x}_{kJ}\,x_{ki} + x_{ki}^2) + 2\emptyset,$$

onde

$$\emptyset = \frac{1}{n_J} \sum_{j=1}^{n_J} \sum_{k=1}^{p} (x_{kj}\,\bar{x}_{kJ} - x_{kj}\,x_{ki} - \bar{x}_{kJ}^2 + \bar{x}_{kJ}\,x_{ki}).$$

No entanto, colocando apenas o somatório em p em evidência,

$$\emptyset = \sum_{k=1}^{p} \{\bar{x}_{kJ} \frac{1}{n_J} \sum_{j=1}^{n_J} x_{kj} - x_{ki} \frac{1}{n_J} \sum_{j=1}^{n_J} x_{kj} - \bar{x}_{kJ}^2 + \bar{x}_{kJ}\,x_{ki}\}$$

$$= \sum_{k=1}^{p} (\bar{x}_{kJ}^2 - x_{ki}\,\bar{x}_{kJ} - \bar{x}_{kJ}^2 + \bar{x}_{kJ}\,x_{ki}) = 0.$$

Assim,

$$D_{iJ}^2 = \frac{1}{n_J} \sum_{j=1}^{n_J} \sum_{k=1}^{p} (x_{kj}^2 - 2x_{kj}\,\bar{x}_{kJ} + 2\bar{x}_{kJ}^2 - 2\bar{x}_{kJ}\,x_{ki} + x_{ki}^2)$$

$$= \frac{1}{n_J} \sum_{j=1}^{n_J} \sum_{k=1}^{p} (x_{kj}^2 - 2x_{kj}\,\bar{x}_{kJ} + \bar{x}_{kJ}^2) + \frac{n_J}{n_J} \sum_{k=1}^{p} (\bar{x}_{kJ}^2 - 2\bar{x}_{kJ}\,x_{ki} + x_{ki}^2)$$

$$= \frac{1}{n_J} \sum_{j=1}^{n_J} d_2^2 (X_j, \bar{X}_J) + d_2^2 (\bar{X}_J, X_i)$$

$$= S_J^2 + d_2^2 (X_i, \bar{X}_J),$$

completando assim a demonstração da primeira parte do teorema.

2a. Parte:

$$D_{IJ}^2 = S_I^2 + S_J^2 + d_{IJ}^2.$$

Por definição:

$$D^2_{IJ} = \frac{1}{n_I n_J} \sum_{i=1}^{n_I} \sum_{j=1}^{n_J} d_2^2 (X_i, X_j).$$

Como

$$D^2_{iJ} = \frac{1}{n_J} \sum_{j=1}^{n_J} d_2^2 (X_i, X_j),$$

tem-se

$$D^2_{IJ} = \frac{1}{n_I} \sum_{i=1}^{n_I} \{ \frac{1}{n_J} \sum_{j=1}^{n_J} d_2^2 (X_i, X_j) \} .$$

Assim,

$$D^2_{IJ} = \frac{1}{n_I} \sum_{i=1}^{n_I} D^2_{iJ}$$

$$= \frac{1}{n_I} \sum_{i=1}^{n_I} \{ S_J^2 + d_2^2 (X_i, \bar{X}_J) \}$$

$$= \frac{n_I}{n_I} S_J^2 + \frac{1}{n_I} \sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_J). \qquad \text{(IV.4)}$$

Por outro lado, pela definição de $D^2_{\bar{X}_J I}$ tem-se:

$$D^2_{\bar{X}_J I} = \frac{1}{n_I} \sum_{i=1}^{n_I} d_2^2 (\bar{X}_J, X_i). \qquad \text{(IV.5)}$$

Como resultado da primeira parte da demonstração resulta que

$$D^2_{\bar{X}_J I} = S_I^2 + d_2^2 (\bar{X}_J, \bar{X}_I). \qquad \text{(IV.6)}$$

A partir das equações (IV.5) e (IV.6) e lembrando que

$d^2_{IJ} = d_2^2 (\bar{X}_I, \bar{X}_J)$ tem-se então que

$$\frac{1}{n_I} \sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_J) = S_I^2 + d^2_{IJ}. \qquad \text{(IV.7)}$$

Finalmente, substituindo (IV.7) em (IV.4),

$$D_{IJ}^2 = S_I^2 + S_J^2 + d_{IJ}^2 ,$$

o que completa a demonstração do teorema.

Teorema IV.3:

Quando se une dois grupos g_I e g_J para formar um novo grupo $g_L = g_I \cup g_J$, a variância interna do grupo resultante é dada por:

$$S_L^2 = \frac{1}{(n_I+n_J)^2} (n_I^2 S_I^2 + n_J^2 S_J^2 + n_I n_J D_{IJ}^2).$$

Demonstração:

No item IV.2.5 foi visto que

$$W_L = W_I + W_J + \frac{n_I n_J}{n_I + n_J} d_{IJ}^2 .$$

Assim, a variância interna S_L^2 do grupo g_L é dada por:

$$S_L^2 = \frac{1}{n_L} W_L = \frac{1}{n_I + n_J} (W_I + W_J + \frac{n_I n_J}{n_I + n_J} d_{IJ}^2)$$

$$= \frac{1}{n_I + n_J} (n_I S_I^2 + n_J S_J^2 + \frac{n_I n_J}{n_I + n_J} d_{IJ}^2).$$

Como $D_{IJ}^2 = S_I^2 + S_J^2 + d_{IJ}^2$,

$$d_{IJ}^2 = - S_I^2 - S_J^2 + D_{IJ}^2 ,$$

e assim

$$S_L^2 = \frac{1}{n_I + n_J} \left[(n_I - \frac{n_I n_J}{n_I + n_J}) S_I^2 + (n_J - \frac{n_I n_J}{n_I + n_J}) S_J^2 + \frac{n_I n_J}{n_I + n_J} D_{IJ}^2 \right]$$

$$= \frac{1}{(n_I+n_J)^2} (n_I^2 S_I^2 + n_J^2 S_J^2 + n_I n_J D_{IJ}^2).$$

Em resumo, o teorema (IV.2) indica que a distância média quadrática D_{IJ}^2, utilizada no método da média de grupo, representa a soma das variâncias internas S_I^2 e S_J^2, dos grupos g_I e g_J a serem unidos, mais o quadrado da distância euclidiana entre as centróides desses grupos:

$$D_{IJ}^2 = S_I^2 + S_J^2 + d_2^2 (\bar{X}_I, \bar{X}_J)$$

$$= S_I^2 + S_J^2 + d_{IJ}^2 .$$

Assim, à primeira vista, poder-se-ia imaginar que o objetivo do método é a minimização da variância do grupo formado.

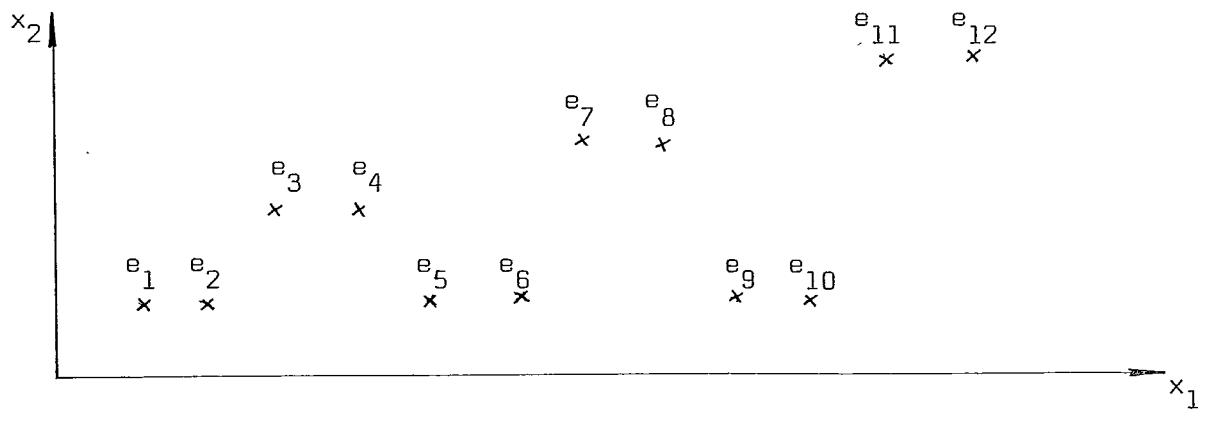
O teorema (IV.3), no entanto, informa que a variância interna do grupo g_L resultante da união de g_I e g_J depende não apenas de D_{IJ}^2, mas também dos tamanhos dos dois grupos a serem reunidos. O que o teorema (IV.3) indica é que

$$S_L^2 \geq \frac{n_I+n_J}{(n_I+n_J)^2} D_{IJ}^2 .$$

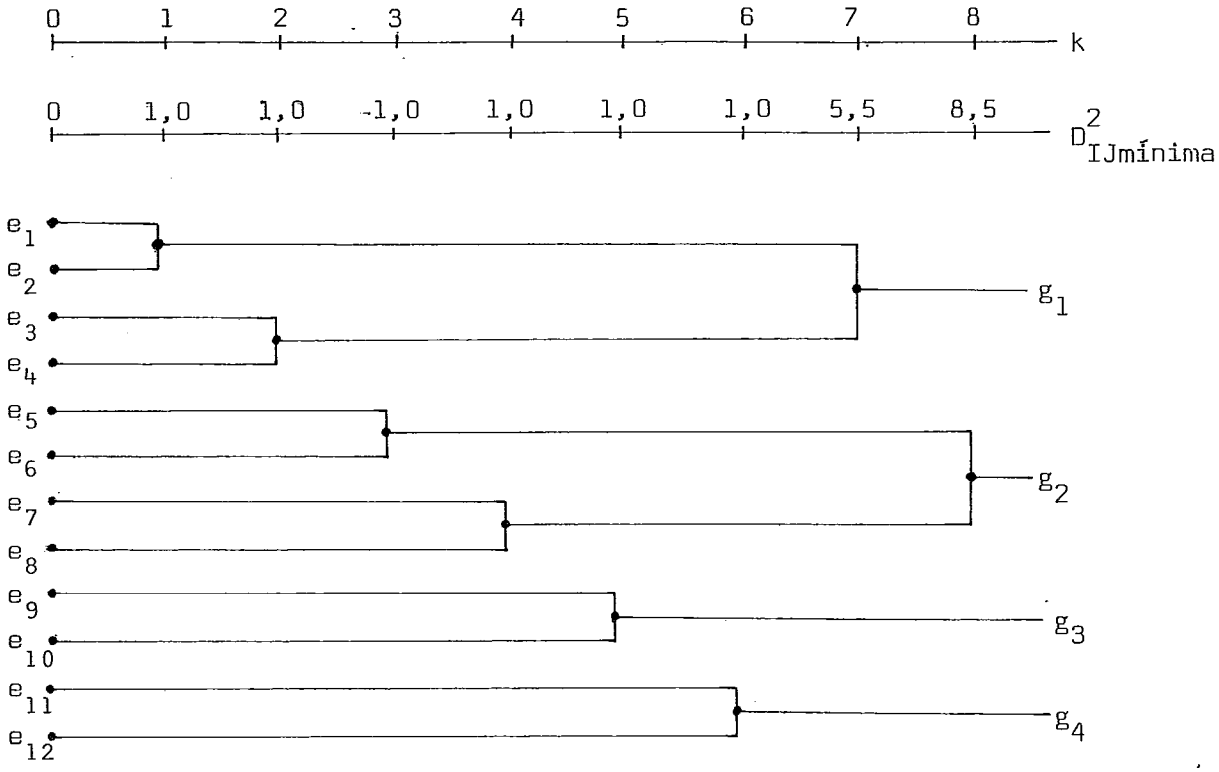
Logo, conhecidos n_I e n_J, D_{IJ}^2 fornece um limite inferior para a variância interna de g_L. Assim, não se pode afirmar que, em cada passo, ao se tomar D_{IJ}^2 mínimo se esteja obtendo S_L^2 mínimo. De uma maneira geral, o que se pode afirmar apenas é que o método tende a formar grupos compactos, ao unir, em cada passo, grupos de pequena variância e que estejam próximos. Não foi possível obter-se uma visão mais clara da função objetivo utilizada neste método.

Exemplo:

Seja o mesmo problema do item anterior:



O problema também foi resolvido pelo método da média de grupo através da rotina CLSTROPT, do Apênd.2, para m = 4. Os resultados obtidos são sumarizados abaixo ($D^2_{IJmínima}$ é a distância média quadrática mínima associada ao passo k):



Coincidentemente, os grupos formados são os mesmos obtidos pelo método de WARD, embora, evidentemente, nada obrigava a que isto acontecesse, uma vez que os critérios de agrupamento são diferentes.

Aqui também, a exemplo do que ocorreu no item anterior, houve um significativo aumento de $D^2_{IJmínimo}$ ao se passar de uma partição em seis grupos para uma partição em cinco, o que sugere a adoção de uma solução com seis "clusters".

O teorema (IV.3), por sua vez, permite o cálculo da variância interna do grupo formado em cada iteração: se $g_L = g_I \cup g_J$, então

$$S^2_L = \frac{1}{(n_I+n_J)^2} (n_I^2 S_I^2 + n_J^2 S_J^2 + n_I n_J D^2_{IJ}) .$$

Assim, ao se unir e_1 e e_2 no passo k=1, tem-se

$$n_1 = n_2 = 1$$

$$S_1^2 = S_2^2 = 0$$

$$D_{12}^2 = d_2^2 (X_1, X_2) = 1,0 .$$

Logo, a variância do grupo formado é igual a 1,0. A tabela (IV.1) abaixo indica as variâncias internas dos grupos formados nos diversos passos:

TABELA IV.1 - Variâncias internas dos grupos formados

Passo	Nº Grupos Existentes	Grupo Formado	Variância Interna
1	11	$\{e_1, e_2\}$	1,0
2	10	$\{e_3, e_4\}$	1,0
3	9	$\{e_5, e_6\}$	1,0
4	8	$\{e_7, e_8\}$	1,0
5	7	$\{e_9, e_{10}\}$	1,0
6	6	$\{e_{11}, e_{12}\}$	1,0
7	5	$\{e_1,e_2,e_3,e_4\}$	1,9
8	4	$\{e_5,e_6,e_7,e_8\}$	2,6

Aqui também se tem um significativo acréscimo ao se passar de uma partição em seis grupos para uma partição em cinco.

IV.3 - CONSIDERAÇÕES GERAIS SOBRE OS MÉTODOS ABORDADOS E APRESENTAÇÃO DE UM MÉTODO HIERARQUIZADO DIVISIVO

A Tabela (IV.2), a seguir, resume as distâncias d_{IJ}^g, entre os grupos g_I e g_J, utilizadas nos diversos métodos aqui expostos. Não é indicada apenas a distância utilizada no método da mediana, uma vez que a sua única diferença em relação ao método da centróide reside na definição das médias dos grupos, como foi visto no item IV.2.4.

TABELA IV.2 - Distâncias entre grupos utilizadas em métodos hierarquizados aglomerativos

Método	Distância d_{IJ}^{g}
Ligação simples	$ds_{IJ} = \min_{\substack{e_i \in g_I \\ e_j \in g_J}} d_{ij}$
Ligação completa	$dc_{IJ} = \max_{\substack{e_i \in g_I \\ e_j \in g_J}} d_{ij}$
Centróide	$d_{IJ}^2 = d_2^2 (\bar{X}_I, \bar{X}_J)$
Ward	$D_{IJ} = \Delta W = \frac{n_I n_J}{n_I + n_J} d_2^2 (\bar{X}_I, \bar{X}_J)$
Média de grupo	$D_{IJ}^2 = S_I^2 + S_J^2 + d_2^2 (\bar{X}_I, \bar{X}_J)$

No método da ligação simples, como visto no item IV.2.1, dois elementos distantes entre si podem ser reunidos em um mesmo grupo, desde que entre eles haja uma cadeia de outros elementos. Este efeito de cadeia faz com que se obtenha soluções do tipo (exemplo dado por DIDAY e SIMON[9]):

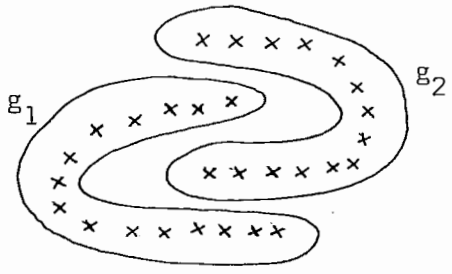


Fig. IV.2 - Grupamentos no método da ligação simples

Pode haver casos em que tais efeitos sejam desejáveis. No caso geral, no entanto, essa característica prejudicará a qualidade da solução obtida.

O método da centróide (e, em consequência, o da mediana) apresenta pouca utilidade prática por não enfocar diretamente a dispersão dos elementos no grupo: aborda apenas as médias e não as variâncias.

Na prática, em geral, os métodos que se mostram mais úteis são os da ligação completa, de Ward e da média de grupo.

O método da ligação completa, como foi visto no item IV.2.2, tem como critério o diâmetro de grupo $d(P_m)$:

$$d(P_m) = \max_{g_I \in P_m} d(g_I) = \max_{g_I \in P_m} \max_{e_i, e_j \in g_I} d_{ij} .$$

No item II.4.6, por outro lado, foi visto que $d(g_I)$ não tem grande sensibilidade para a distribuição interna dos elementos de g_I, enfocando apenas a distância entre seus pontos mais afastados. Apesar dessa dificuldade, o método da ligação completa apresenta sobre os métodos de Ward e da média de grupo duas vantagens. Exige muito menos cálculos que esses dois outros métodos e permite, como se pode notar na Tabela (IV.2), a utilização de qualquer métrica. Assim, se fosse necessária a utilização de uma "métrica de correlação", o método adequado seria o da ligação completa.

O método de Ward tem, por sua vez, a característica de procurar minimizar a soma dos quadrados dentro dos grupos W (ver Teorema (IV.1) do item IV.2.5), tendendo assim a formar grupos compactos (de pequena dispersão).

No caso do método da média de grupo, a distância utilizada, como se observa na Tabela (IV.2), é:

$$D_{IJ}^2 = S_I^2 + S_J^2 + d_2^2 (\bar{X}_I, \bar{X}_J) .$$

Como foi comentado no item IV.2.6, ao se apresentar o método, a função objetivo não fica clara. O método tende a gerar grupos compactos : é levada em conta, na união de dois grupos g_I e g_J, não só a variância interna S_I^2 e S_J^2 de cada um, mas também a distância entre suas centróides.

Por outro lado, nos itens IV.2.5 e IV.2.6 foi indicado como a análise da evolução dos valores das distâncias entre grupos pode, nos métodos de Ward e da média de grupo, sugerir o número de grupos m a ser adotado. No caso mais geral, quando não se dispõe "a priori" de indicação alguma do número m a ser adotado, parece conveniente efetuar-se um ciclo completo (até k=1) de um método hierarquizado aglomerativo adequado, exa

minando-se a evolução dos valores de d^{g}_{IJ}.

É comum, na vida real, a existência de grupos, subgrupos, etc., o que torna a decisão, quanto ao número m a ser adotado, bastante relativa. Por exemplo, quantos grupos se deve tomar no caso apresentado na Figura (IV.3) ? Dois ou quatro grupos?

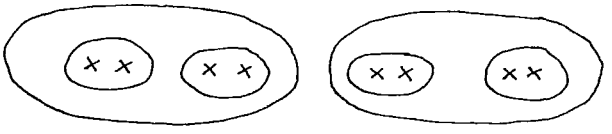


Fig. IV. 3 - Grupos e Subgrupos

Ambas as soluções "fazem sentido" e não parece que a opção possa ser feita independentemente do caso em estudo.

No entanto, se essa divisão em grupos, subgrupos, etc., for bem pronunciada, a evolução dos diversos valores de d^{g}_{IJ}, no desenvolvimento do processo hierarquizado aglomerativo, deverá se apresentar da forma abaixo:

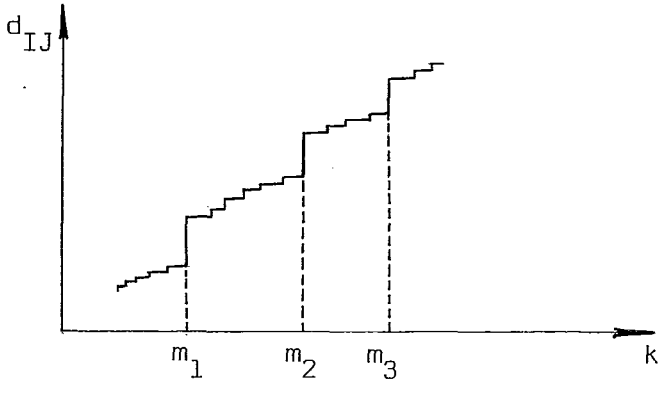


Fig. IV. 4 - Evoluções de d^{g}_{IJ} no caso de grupos, subgrupos, etc.

Isto permitirá a determinação de valores típicos m_1, m_2, m_3,..., que, por sua vez, analisados dentro do contexto do problema prático em questão, poderão indicar um valor adequado para m.

Uma outra solução seria adotar, por exemplo, no caso do método da ligação completa, o seguinte procedimento: dois grupos g_I e g_J só seriam unidos se a distância completa dc_{IJ}, da Tabela (IV.2), for inferior

a uma fração α, definida a priori, do diâmetro d(E) do conjunto E dos n elementos $e_1, e_2, \ldots, e_n$:

$$d(E) = \max_{e_i, e_j \in E} d_{ij}.$$

Isso significa estabelecer um limite superior para o diâmetro de qualquer grupo formado. Quando esse limite fosse alcançado, o algoritmo seria interrompido.

No caso do método de Ward, poder-se-ia estabelecer um limite idêntico para a soma dos quadrados dentro do grupo. No caso da média de grupo, e através do Teorema (IV.3), poder-se-ia estabelecer um limite para a variância interna de cada grupo.

Finalmente, cabe comentar a existência de métodos hierarquizados divisivos, isto é, técnicas em que, em cada passo, um determinado grupo é particionado, ao invés de se efetuar uma aglomeração. Um exemplo disso seria o seguinte algoritmo, onde P_k representa uma partição do conjunto E em k grupos:

Passo 0: (inicialização)

Forme um único grupo, constituído de todos os elementos de E: $P_1 = E$;

Faça k=1

Passo 1: (determinação do grupo g_L a ser particionado)

Determine $g_L \in P_k$ tal que

$$d(g_L) = \max_{e_i, e_j \in g_L} d_{ij} = d(X_p, X_q) = d_{pq}$$

seja máximo

Passo 2: (formação de dois novos grupos por partição de g_L)

Faça $g_I = \{e_p\}$ e $g_J = \{e_q\}$;

Aloque os demais elementos de g_L a g_I ou g_J, conforme for menor a distância a e_p ou e_q;

Faça $P_{k+1} = P_k \cap \overline{\{g_L\}} \cup \{g_I\} \cup \{g_J\}$ e k=k+1

Passo 3: (regra de parada)

Se k = n , pare

Senão , vá para o Passo 1.

Note-se que esse método procura minimizar o diâmetro de grupo e pode ser utilizado com qualquer métrica. Uma outra regra de parada poderia ser dada pelo término do algoritmo quando d_{pq} fosse igual ou inferior a α. d(E), o que implicaria em estabelecer um limite superior para o diâmetro dos grupos formados. O número de grupos poderia ainda ser determinado a partir da análise do dendograma (a exemplo do que foi feito nos itens IV.2.5 e IV.2.6), que, nesse caso, corresponde a divisões, e não a aglomerações, como nos casos anteriores.

Exemplo

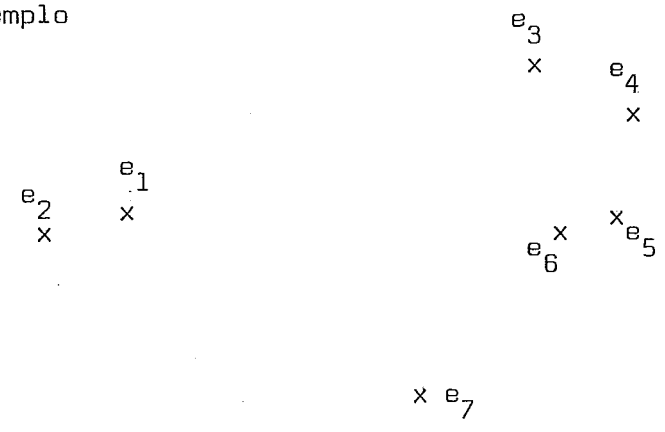


Fig. IV. 5 - Exemplo

No primeiro passo, a distância máxima é dada por d_{24}. Assim, e_2 e e_4 são as "sementes" de dois grupos:$\{e_1,e_2\}$ e $\{e_3,e_4,e_5,e_6,e_7\}$. A resolução do problema é resumida na Figura (IV.6):

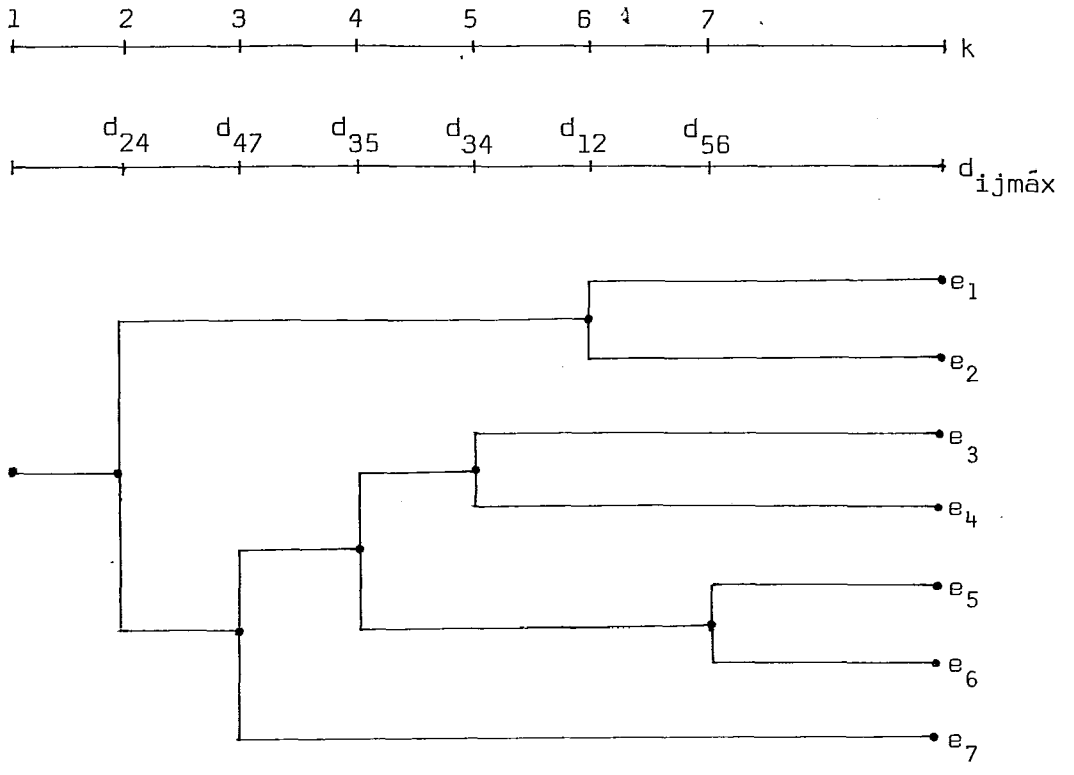


Fig. IV. 6 - Dendograma

IV.4 - ALGORITMO DE LANCE E WILLIAMS

LANCE e WILLIAMS[24] e WISHART[44] (ver também DURAN e ODELL[10] e DIDAY e SIMON[9]), demonstraram que é possível incorporar em um único algoritmo os seis métodos hierarquizados aglomerativos que foram aqui descritos. O método geral parte da seguinte equação básica, que fornece a distância entre um grupo g_K e um grupo g_L, onde $g_L = g_I \cup g_J$:

$$d^g_{KL} = \alpha_i\, d^g_{KI} + \alpha_j\, d^g_{KJ} + \beta\, d^g_{IJ} + \gamma \left| d^g_{KI} - d^g_{KJ} \right| ,$$

onde α_i, α_j, β e γ assumem valores particulares, conforme o método utilizado.

IV.4.1 - MÉTODO DA LIGAÇÃO SIMPLES

Seja d uma função de distância qualquer. Tomando $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = -1/2$, tem-se , a partir da equação básica:

- se $ds_{KI} > ds_{KJ} \Rightarrow ds_{KL} = \frac{1}{2} ds_{KI} + \frac{1}{2} ds_{KJ} - \frac{1}{2} ds_{KI} + \frac{1}{2} ds_{KJ} = ds_{KJ}$;

- se $ds_{KI} = ds_{KJ} \Rightarrow ds_{KL} = \frac{1}{2} ds_{KI} + \frac{1}{2} ds_{KJ} = ds_{KI} = ds_{KJ}$; e

- se $ds_{KI} < ds_{KJ} \Rightarrow ds_{KL} = \frac{1}{2} ds_{KI} + \frac{1}{2} ds_{KJ} + \frac{1}{2} ds_{KI} - \frac{1}{2} ds_{KJ} = ds_{KI}$

IV.4.2 - MÉTODO DA LIGAÇÃO COMPLETA

Seja d uma função de distância qualquer. Tomando agora $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = 1/2$, tem-se, usando raciocínio análogo ao do item anterior:

- se $dc_{KI} > dc_{KJ} \Rightarrow dc_{KL} = dc_{KI}$;

- se $dc_{KI} = dc_{KJ} \Rightarrow dc_{KL} = dc_{KI} = dc_{KJ}$; e

- se $dc_{KI} < dc_{KJ} \Rightarrow dc_{KL} = dc_{KJ}$.

IV.4.3 - MÉTODO DA CENTRÓIDE

Quando $g_L = g_I \cup g_J$, $n_L = n_I + n_J$. Sejam então

$$\alpha_i = \frac{n_I}{n_L} \quad ;$$

$$\alpha_j = \frac{n_J}{n_L} \quad ;$$

$$\beta = - \frac{n_I n_J}{n_L^2} \quad ; \quad \text{e}$$

$$\gamma = 0$$

A equação básica fornece:

$$d_{KL}^2 = \frac{n_I}{n_L} d_{KI}^2 + \frac{n_J}{n_L} d_{KJ}^2 - \frac{n_I n_J}{n_L^2} d_{IJ}^2$$

Resta então provar que essa equação é válida:

Teorema IV.4:

Se $g_L = g_I \cup g_J$, então

$$\phi_{KL} = \frac{n_I}{n_L} d^2_{KI} + \frac{n_J}{n_L} d^2_{KJ} - \frac{n_I n_J}{n_L^2} d^2_{IJ} = d^2_{KL} .$$

Demonstração:

$$\phi_{KL} = \frac{n_I}{n_L} d^2_{KI} + \frac{n_J}{n_L} d^2_{KJ} - \frac{n_I n_J}{n_L^2} d^2_{IJ} =$$

$$= \frac{n_I}{n_L} \sum_{r=1}^{p} (\bar{x}_{rK} - \bar{x}_{rI})^2 + \frac{n_J}{n_L} \sum_{r=1}^{p} (\bar{x}_{rK} - \bar{x}_{rJ})^2 - \frac{n_I n_J}{n_L^2} \sum_{r=1}^{p} (\bar{x}_{rI} - \bar{x}_{rJ})^2$$

$$= \frac{n_I}{n_L} \sum_{r=1}^{p} (\bar{x}^2_{rK} - 2\bar{x}_{rK}\bar{x}_{rI} + \bar{x}^2_{rI}) + \frac{n_J}{n_L} \sum_{r=1}^{p} (\bar{x}^2_{rK} - 2\bar{x}_{rK}x_{rJ} + \bar{x}^2_{rJ})$$

$$- \frac{n_I n_J}{n_L^2} \sum_{r=1}^{p} (\bar{x}^2_{rI} - 2\bar{x}_{rI}\bar{x}_{rJ} + \bar{x}^2_{rJ}) .$$

Colocando o somatório em r em evidência e rearranjando-se os termos, tem-se:

$$\phi_{KL} = \sum_{r=1}^{p} \left[\frac{n_I + n_J}{n_L} \bar{x}^2_{rK} - 2 \frac{n_I}{n_L} \bar{x}_{rK}\bar{x}_{rI} - 2 \frac{n_J}{n_L} \bar{x}_{rK}\bar{x}_{rJ} + \left(\frac{n_I}{n_L} - \frac{n_I n_J}{n_L^2} \right) \bar{x}^2_{rI} \right.$$

$$\left. + 2 \frac{n_I n_J}{n_L^2} \bar{x}_{rI}\bar{x}_{rJ} + \left(\frac{n_J}{n_L} - \frac{n_I n_J}{n_L^2} \right) \bar{x}^2_{rJ} \right] .$$

Como $n_L = n_I + n_J$,

$$\phi_{KL} = \sum_{r=1}^{p} \left[\bar{x}^2_{rK} - \frac{2}{n_L} (n_I \bar{x}_{rK}\bar{x}_{rI} + n_J \bar{x}_{rK}\bar{x}_{rJ}) + \left(\frac{n_I n_L - n_I n_J}{n_L^2} \right) \bar{x}^2_{rI} \right.$$

$$\left. + 2 \frac{n_I n_J}{n_L^2} \bar{x}_{rI}\bar{x}_{rJ} + \left(\frac{n_J n_L - n_I n_J}{n_L^2} \right) \bar{x}^2_{rJ} \right] .$$

Novamente, como $n_L = n_I + n_J$,

$$\phi_{KL} = \sum_{r=1}^{p} \left[\bar{x}_{rK}^2 - \frac{2}{n_L} (n_I \bar{x}_{rK}\bar{x}_{rI} + n_J \bar{x}_{rK}\bar{x}_{rJ}) + \frac{1}{n_L^2} (n_I^2 \bar{x}_{rI}^2 + 2n_I n_J \bar{x}_{rI}\bar{x}_{rJ} + n_J^2 \bar{x}_{rJ}^2) \right]$$

$$= \sum_{r=1}^{p} \{ \bar{x}_{rK}^2 - 2\bar{x}_{rK} \frac{1}{n_L} (n_I \bar{x}_{rI} + n_J \bar{x}_{rJ}) + \left[\frac{1}{n_L} (n_I \bar{x}_{rI} + n_J \bar{x}_{rJ}) \right]^2 \}$$

$$= \sum_{r=1}^{p} \left[\bar{x}_{rK} - \frac{1}{n_L} (n_I \bar{x}_{rI} + n_J \bar{x}_{rJ}) \right]^2$$

Como $\bar{x}_{rL} = \frac{1}{n_L} \left(\sum_{i=1}^{n_I} x_{ri} + \sum_{j=1}^{n_J} x_{rj} \right) = \frac{1}{n_L} (n_I \bar{x}_{rI} + n_J \bar{x}_{rJ})$,

$$\phi_{KL} = \sum_{r=1}^{p} (\bar{x}_{rK} - \bar{x}_{rL})^2$$

$$= d_2^2 (\bar{X}_K, \bar{X}_L)$$

$$= d_{KL}^2 ,$$

o que completa a demonstração do teorema.

IV.4.4 - MÉTODO DA MEDIANA

Como foi visto no item IV.2.4, o método da mediana é idêntico ao método da centróide, com a única exceção de que, ao se unir dois grupos , assume-se que são de igual tamanho. Assim, tomando $n_I = n_J = n$ (logo $n_L = 2n$), tem-se, pelo Teorema (IV.4), que

$$d_{KL}^2 = \frac{1}{2} d_{KI}^2 + \frac{1}{2} d_{KJ}^2 - \frac{1}{4} d_{IJ}^2 .$$

Logo, o método da mediana pode ser obtido da equação básica do método de Lance e Williams fazendo-se

$$\alpha_i = \alpha_j = \frac{1}{2} ;$$

$$\beta = -\frac{1}{4} \quad ; \text{ e}$$

$$\gamma = 0$$

IV.4.5 - MÉTODO DE WARD

No caso do método de WARD tem-se:

$$\alpha_1 = \frac{n_K + n_I}{n_K + n_L} ;$$

$$\alpha_j = \frac{n_K + n_J}{n_K + n_L} ;$$

$$\beta = -\frac{n_K}{n_K + n_L} \quad ; \text{ e}$$

$$\gamma = 0 .$$

A equação básica fornece:

$$D_{KL} = \frac{n_K + n_I}{n_K + n_L} D_{KI} + \frac{n_K + n_J}{n_K + n_L} D_{KJ} - \frac{n_K}{n_K + n_L} D_{IJ} ,$$

como é demonstrado abaixo:

Teorema IV.5:

Se $g_L = g_I \cup g_J$ e g_K é um grupo qualquer, então

$$D_{KL} = \frac{n_K + n_I}{n_K + n_L} D_{KI} + \frac{n_K + n_J}{n_K + n_L} D_{KJ} - \frac{n_K}{n_K + n_L} D_{IJ} .$$

Demonstração:

Como, por definição (item IV.2.5):

$$D_{IJ} = \frac{n_I n_J}{n_I + n_J} d_{IJ}^2 ,$$

tem-se :

$$d_{IJ}^2 = \frac{n_I + n_J}{n_I n_J} D_{IJ} . \qquad \text{(IV.4)}$$

Por outro lado, pelo teorema (IV.4) do item IV.4.3, se $g_L = g_I \cup g_J$ e g_K qualquer, tem-se:

$$d_{KL}^2 = \frac{n_I}{n_L} d_{KI}^2 + \frac{n_J}{n_L} d_{KJ}^2 - \frac{n_I n_J}{n_L^2} d_{IJ}^2 . \qquad \text{(IV.5)}$$

Pelas equações (IV.4) e (IV.5) , tem-se então:

$$\frac{n_K + n_L}{n_K n_L} D_{KL} = \frac{n_I}{n_L} \frac{n_K + n_I}{n_K n_I} D_{KI} + \frac{n_J}{n_L} \frac{n_K + n_J}{n_K n_J} D_{KJ} - \frac{n_I n_J}{n_L^2} \cdot \frac{n_I + n_J}{n_I n_J} D_{IJ} ,$$

ou seja,

$$D_{KL} = \frac{n_K + n_I}{n_K + n_L} D_{KI} + \frac{n_K + n_J}{n_K + n_L} D_{KJ} - \frac{n_K}{n_K + n_L} D_{IJ} ,$$

completando a demonstração do teorema.

IV.4.6 - MÉTODO DA MÉDIA DE GRUPO

Toma-se nesse caso:

$$\alpha_i = \frac{n_I}{n_L} ;$$

$$\alpha_j = \frac{n_J}{n_L} ; \quad \text{e}$$

$$\beta = \gamma = 0 \ .$$

A equação básica fornece:

$$D^2_{KL} = \frac{n_I}{n_L} D^2_{KI} + \frac{n_J}{n_L} D^2_{KJ} \ ,$$

que é validada pelo teorema abaixo.

Teorema IV.6:

Se $g_L = g_I \cup g_J$ e g_K é um outro grupo qualquer, então

$$D^2_{KL} = \frac{n_I}{n_L} D^2_{KI} + \frac{n_J}{n_L} D^2_{KJ} \ .$$

Demonstração:

Por definição ,

$$D^2_{KL} = \frac{1}{n_K n_L} \left[\sum_{k=1}^{n_K} \sum_{i=1}^{n_I} d_2^2 (X_k, X_i) + \sum_{k=1}^{n_K} \sum_{j=1}^{n_J} d_2^2 (X_k, X_j) \right]$$

Multiplicando e dividindo as parcelas por n_I e n_J, respectivamente:

$$= \frac{n_I}{n_L} \frac{1}{n_K n_I} \sum_{k=1}^{n_K} \sum_{i=1}^{n_I} d_2^2 (X_k, X_i) + \frac{n_J}{n_L} \frac{1}{n_K n_J} \sum_{k=1}^{n_K} \sum_{j=1}^{n_J} d_2^2 (X_k, X_j)$$

$$= \frac{n_I}{n_L} D^2_{KI} + \frac{n_J}{n_L} D^2_{KJ} \ ,$$

completando-se assim a demonstração.

IV.4.7 - CONCLUSÃO

Em resumo, o método de Lance e Williams permite a resolução de problemas de análise de grupamento via métodos hierarquizados aglomerativos, a partir de uma única equação básica, onde seus parâmetros são ajus

tados conforme o método hierarquizado aglomerativo a ser utilizado. Assim, a distância d^g_{KL} entre um grupo g_K e g_L, onde $g_L = g_I \cup g_J$, é dada por:

$$d^g_{KL} = \alpha_i \, d^g_{KI} + \alpha_j \, d^g_{KJ} + \beta \, d^g_{IJ} + \gamma \left| d^g_{KI} - d^g_{KJ} \right| ,$$

cujos parâmetros são resumidos na Tabela (IV.3).

Então, no método de Lance e Williams, no passo em que se define os grupos g_I e g_J que serão unidos para formar o grupo g_L, é possível calcular as novas distâncias d_{KL}, para os demais grupos g_K, a partir das distâncias definidas no passo anterior:

$$d^g_{KL} = f(d^g_{KI}, d^g_{KJ}, d^g_{IJ}, n_K, n_I, n_J) \quad , \forall K$$

O método especifica assim, como calcular recursivamente a distância entre os grupos sem que, em cada passo, tal cálculo tenha de ser efetuado a partir da estaca zero. No passo inicial, como cada grupo é formado por um único elemento, e caso se utilize como medida de semelhança entre pontos a métrica euclidiana, a distância entre os grupos seria, nos seis métodos abordados, dada pela própria distância euclidiana entre esses elementos.

Finalmente, cabe enfatizar que o método de Lance e Williams se constitui, apenas, numa técnica geral para atualização, em cada iteração, das distâncias d^g_{KL} entre os diversos grupos g_K já formados e o novo grupo $g_L = g_I \cup g_J$. Não é, assim, um método específico hierarquizado aglomerativo de análise de grupamento.

TABELA IV.3 - Parâmetros do método de Lance e Williams

Método	α_i	α_j	β	γ
Ligação Simples	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Ligação Completa	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Centróide	$\frac{n_I}{n_L}$	$\frac{n_J}{n_L}$	$-\frac{n_I n_J}{n_L^2}$	0
Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_K+n_I}{n_K+n_L}$	$\frac{n_K+n_J}{n_K+n_L}$	$-\frac{n_K}{n_K+n_L}$	0
Média de Grupo	$\frac{n_I}{n_L}$	$\frac{n_J}{n_L}$	0	0

Obs.: $n_L = n_I + n_J$

V . MÉTODOS DE REALOCAÇÃO ITERATIVA

V.1 - INTRODUÇÃO

No item III.3 foi descrito o procedimento geral de um método de realocação iterativa: começando-se com uma partição inicial P_m^o do conjunto E dos n elementos em m grupos, a idéia é gerar sucessivamente partições P_m^k tais que

$$f(P_m^k) \leq f(P_m^{k-1}).$$

O processo é interrompido quando algum teste de convergência é satisfeito, indicando que uma melhoria da partição não poderia mais ser obtida pelo método utilizado. Evidentemente, essa solução , para um dado número m de grupos, não corresponderá necessariamente a um ótimo global, constituindo-se, no entanto, pela impossibilidade de melhoria, em um mínimo local.

Segundo ANDERBERG[1], a grande vantagem desses métodos sobre, por exemplo, os métodos hierarquizados, reside em sua extrema eficiência computacional. Geralmente, para um dado número m, a convergência é obtida até um máximo de cinco iterações, havendo realmente grande decréscimo em $f(P_m^k)$ na primeira delas.

Além disso, como se verá adiante, alguns dos métodos de realocação iterativa possuem a flexibilidade de permitir que o número de grupos da solução, ou soluções,também seja fornecido pelo algoritmo. Assim, o texto que se segue aborda duas grandes classes de métodos: com número fixo e variável de grupos. Inicialmente, no entanto, serão abordadas algumas técnicas para a geração de uma solução inicial.

V.2 - GERAÇÃO DE UMA SOLUÇÃO INICIAL

Como se observou no item anterior,os métodos de realocação iterativa partem de uma solução inicial, a ser refinada. Existem, pelo menos, quatro técnicas básicas para a geração dessa solução inicial:

- por escolha intencional;
- por seleção aleatória;
- através de um método hierarquizado; e
- através de pontos "semente".

ANDERBERG[1] comenta que a escolha intencional pode ser feita com base em uma determinada variável que o analista considere como mais importante e que a seleção aleatória não seria uma alternativa atraente, uma vez que os grupos assim formados não possuem necessariamente homogeneidade alguma.

Por outro lado, aquele autor informa que um agrupamento hierarquizado de todo o conjunto dos n elementos não só pode requerer mais esforço computacional que o resto da análise, mas também iria limitar em muito o porte dos problemas a serem considerados. A alternativa, nesse caso, indicada por LANCE e WILLIAMS[25], é aplicar o algoritmo hierarquizado a um subconjunto dos n elementos (de um tamanho de 150 a 250 unidades,por exemplo) utilizando-se os grupos assim formados como núcleos para a alocação dos demais pontos, que seriam atribuídos ao grupo que deles estivesse mais próximo.

Um outro esquema bastante útil é o dos pontos denominados por ANDERBERG[1] de pontos "semente". Tal denominação decorre do fato de que esses pontos, nesse procedimento, se constituem em sementes de novos grupos, que são formados ao se atribuir os diversos elementos de E ao grupo que cuja semente está mais próxima. ANDERBERG[1] cita oito maneiras de se criar pontos sementes:

(a) escolha dos m primeiros elementos de E;

(b) seleção sistemática, com passo [n/m], de m elementos de E;

(c) seleção subjetiva de m elementos de E;

(d) seleção aleatória de m elementos de E;

(e) geração aleatória de m pontos do R^p;

(f) cálculo das centróides de uma partição qualquer de E em m grupos;

(g) análise modal de ASTRAHAN[2] (ver também WISHART[45] e DURAN e ODELL[10]);

- cálculo da "densidade" de cada elemento (número de outros elementos dentro de uma distância predeterminada d_1);
- ordenação dos elementos a partir de sua "densidade" e escolha do elemento de maior "densidade" como primeiro ponto semente;
- escolha de pontos sementes subsequentes em ordem de "densidade" decrescente, sujeita à condição de que cada novo ponto de semente esteja pelo menos a uma distância d_2 dos outros pontos sementes já escolhidos;
- caso haja excesso de pontos sementes, proceder a um agrupamento hierarquizado dessas sementes, até que um número m seja obtido. Caso haja falta de sementes, ajustar d_1 e d_2 e repetir o processo;

(h) técnica de BALL e HALL[4] :

- escolha da centróide de E como primeiro ponto semente;
- seleção subsequente de pontos sementes pelo exame, em ordem

crescente dos índices, dos elementos de E, aceitando como semente qualquer ponto a uma distância maior que um valor predeterminado d, das demais sementes, até que m pontos sejam escolhidos ou o conjunto E seja exaurido (nesse caso, se m pontos não forem escolhidos, diminuir o valor de d e repetir o processo).

A essa lista, que evidentemente não é exaustiva, pode-se acrescer como método de seleção a escolha, como sementes, das medianas de uma partição qualquer de E em m grupos.

Os métodos de (a) a (d), (f) e (g) e o método das medianas têm em comum a propriedade de que os grupos formados a partir dessas sementes têm pelo menos um membro. Os métodos (e) e (h) podem gerar partições iniciais com grupos vazios.

No método (g) a escolha de d_1 e d_2 pode requerer várias tentativas até que se obtenha sucesso. Se d_1 for excessivamente pequeno, muitos pontos serão considerados isolados, ou seja, com "densidade" igual a um. Se d_1 for grande demais, a avaliação do caráter modal dos pontos fica prejudicada, uma vez que todos eles terão "densidades" elevadas. Por outro lado, se d_2 não for igual ou superior ao dobro de d_1, alguns pontos poderão contribuir para as "densidades" de mais de um ponto "semente" escolhido. Essa técnica geralmente exige, como se vê, grande esforço computacional. ASTRAHAN[2] sugere então que seja aplicada a uma amostra de E, no caso de n ser elevado.

Tendo em vista que a determinação do parâmetro d também envolve um processo de tentativa e erro, o método (h) igualmente é desvantajoso no que se refere ao esforço computacional.

A opção por qualquer um desses métodos depende, evidentemente, das condições de cada problema. Tendo, por outro lado, em vista que os métodos de realocação iterativa fornecem mínimos locais, talvez seja de interesse para o analista a resolução do problema para várias soluções iniciais, obtidas por diferentes métodos, analisando-se então a compatibilidade com as soluções finais resultantes.

Os métodos a serem apresentados no item seguinte, com exceção do método K-Médias, admitem qualquer partição inicial. Suas próprias características, no entanto, aliadas, como foi dito, às demais condições do problema, indicarão ao analista as técnicas mais adequadas para a geração dessas partições.

V.3 - MÉTODOS COM NÚMERO FIXO DE GRUPOS

Dentre os métodos com número fixo de grupos ou seja, aqueles em que o número de grupos m, da solução final do problema, é especificado a priori, destacam-se os de FORGY[13] (e sua variante de JANCEY[19]), de MACQUEEN [27] (também chamado de K-Médias) e o método das K-Medianas (devido a MULVEY e CROWDER[29]).

Os algoritmos de FORGY[13], JANCEY[19] e MACQUEEN[27] (também apresentados em ANDERBERG[1]) serão aqui descritos sucintamente, dando-se uma maior ênfase neste trabalho ao método das K-Medianas, por sua importância no desenvolvimento do método de programação matemática de Mulvey e Crowder, a ser apresentado no item VI.5.

O método de Forgy consiste na seguinte sequência de passos (ver também algorítmo α em DIDAY e SIMON[9]):

Passo 0: Faça k = 0

Passo 1: Se k = 0, tome uma partição qualquer P_m^0 de E; senão, forme uma nova partição P_m^k de E, alocando cada elemento ao grupo cujo ponto semente está mais próximo

Passo 2: Calcule as centróides desses grupos (essas centróides serão os pontos sementes dos novos grupos)

Passo 3: Se $f(P_m^k) = f(P_m^{k-1})$, pare;
senão, faça k = k + 1 e vá para o Passo 1.

Como se nota, o algoritmo implica na utilização da alternativa (f) de cálculo de ponto semente para a geração de solução inicial, ou seja, cálculo das centróides de uma partição qualquer. Caso isso não tenha sido feito inicialmente, certamente o será na primeira iteração do Passo 2.

A sequência de Passos 1 e 2 gera soluções P_m^k tais que

$$f(P_m^k) \leq f(P_m^{k-1}),$$

onde a função objetivo é, normalmente, a soma dos quadrados dentro dos grupos, definida no item II.4, uma vez que o objetivo geral é o de minimizar as distâncias dos elementos às centróides desses grupos. Essa melhoria da função objetivo é feita em duas etapas, através dos Passos 1 e 2, na forma descrita a seguir.

Seja $\phi(P_m, A_1, \ldots, A_m)$ a soma dos quadrados das distâncias euclidianas entre os pontos e as sementes de seus grupos:

$$\phi(P_m, A_1, \ldots, A_m) = \sum_{j=1}^{m} \sum_{e_i \in g_j} d_2^2(X_i, A_j) \qquad (V.1)$$

Mais adiante, na demonstração da convergência do método das K-Medianas, será verificado que a alocação efetuada no Passo 1, para $K \geq 1$, é a

que minimiza $\phi(.,A1, \ldots, A_m)$. Assim, após o Passo 1, tem-se:

$$\phi(P_m^k, A_1^{k-1}, \ldots, A_m^{k-1}) \leq \phi(P_m^{k-1}, A_1^{k-1}, \ldots, A_m^{k-1}) = f(P_m^{k-1}) \quad (V.2)$$

Por outro lado, derivando a expressão do lado direito da equação (V.1) com relação a cada componente do vetor A_j, é trivial verificar que as soluções A_j^k que minimizam $(P_m^k, .)$ são as centróides dos grupos de P_m^k. Assim, no Passo 2, tem-se:

$$\phi(P_m^k, A_1^k, \ldots, A_m^k) = f(P_m^k) \leq \phi(P_m^k, A_1^{k-1}, \ldots, A_m^{k-1}) \quad (V.3)$$

Assim, por (V.2) e (V.3) tem-se:

$$f(P_m^k) \leq f(P_m^{k-1}).$$

A variante de JANCEY[19] difere do método proposto por Forgy apenas no que se refere à determinação de pontos sementes. Nessa variante, o primeiro conjunto de pontos sementes ou é definido a priori ou é dado pelas centróides dos grupos da partição inicial. Nas iterações subseqüentes cada novo ponto semente é determinado pela duplicação do segmento de reta que une o antigo ponto semente à nova centróide do grupo. A figura (V.1), baseada em ANDERBERG[1], ilustra esse procedimento:

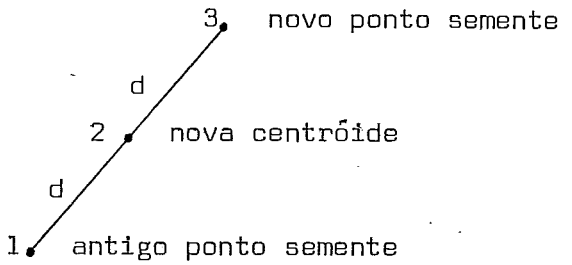


Fig. V.1 - Determinação do novo ponto semente pelo método de Jancey

Segundo ANDERBERG[1], o segmento que une os pontos 1 e 2 da Figura (V.1) pode ser encarado como uma direção em que o ponto semente deveria mover-se para uma maior melhoria da partição. A duplicação do segmento procura acelerar o processo, sugerindo aquele autor que esse procedimento não só deve acelerar a obtenção de uma solução pelo algoritmo mas também possivelmente pode levar a que essa solução que não esteja condicionada apenas a ser um mínimo local. Note-se que nesse método, à medida que a solução viesse a ser obtida, os pontos sementes e as centróides tenderiam a se confundir: as distâncias d, da Figura (V.1), seriam cada vez menores.

Tanto o método de Forgy como a variante de Jancey procuram minimizar, como se vê, a soma dos quadrados dentro dos grupos.

A convergência do método de Forgy pode ser demonstrada de uma maneira bastante semelhante à do método das K-Medianas, a ser apresentada adiante. Deve-se salientar, porém, que alguns autores definem como teste de convergência do algoritmo a condição de que $P_m^k = P_m^{k-1}$. Tal critério, de igualdade de partições, pode prejudicar a convergência do algoritmo: basta que existam pelo menos duas soluções $P_m^i \neq P_m^j$ tais que $f(P_m^i) = f(P_m^j)$. O algaritmo poderia então gerar "eternamente" e de forma alternada as soluções P_m^i e P_m^j, ou seja, o método "ciclaria", não havendo, evidentemente,nesse caso, convergência alguma.

Quanto à variante de Jancey, no entanto, ANDERBERG[1] informa que não se conhece prova de sua convergência nem está disponível um exemplo em que o algoritmo não tenha convergido.

O método das K-Médias, por sua vez, desenvolvido por MACQUEEN[27], consiste na seguinte seqüência de passos:

Passo 0: Tome os primeiros m elementos do conjunto E como pontos sementes

Passo 1: Aloque os (n-m) elementos restantes ao grupo cujo ponto semente está mais próximo. Após cada alocação, calcule a centróide do novo grupo, que passará a ser o novo ponto semente

Passo 2: Ao final dessa alocação, repita o procedimento de atribuição anterior para todo o conjunto E, recalculando, quando for o caso, as centróides do grupo que ganhou um elemento e do grupo que perdeu esse elemento.

Note-se que esse processo não é iterativo, correspondendo, de uma maneira geral, a duas "iterações" apenas. Sua diferença básica em relação ao método de Forgy se deve ao fato de que agora o ponto semente é recalculado após cada atribuição de elemento a grupo, ao passo que no método anterior, como foi visto, os pontos semente permaneceriam constantes durante toda a iteração.

Como a atribuição de um elemento a um novo grupo, no Passo 2, corresponde à formação de uma nova partição e novo cálculo de ponto semente (centróide), tem-se,ao final Passo 2, utilizando os argumentos apresentados para o método de Forgy, que:

$$f(P_m^1) \leq f(P_m^0), \qquad (V.4)$$

Por efetuar apenas duas "iterações", este é o algoritmo mais rápido do aprensentado no presente estudo. Sua aplicação prática reside em que,como

foi dito no início do item V.1, as maiores mudanças em $f(P_m)$, geralmente ocorrem na primeira iteração. O comportamento do algoritmo depende evidentemente da sequência dos elementos $\{e_1, \ldots, e_n\}$. ANDERBERG[1], no entanto, informa que este fato tem pouca importância na solução final quando os grupos são bem separados.

É possível também definir um algoritmo K-Médias convergente, repetindo o processo de realocação característico do método até que $f(P_m^k) = f(P_m^{k-1})$. Este algoritmo corresponde ao algoritmo β, apresentado por DIDAY e SIMON[9]. O método K-Médias iterativo é convergente, tendo em vista e generalizando para qualquer iteração a equação (V.4). Evidentemente, os métodos K-Médias procuram minimizar a soma dos quadrados dentro dos grupos.

Como já foi dito anteriormente, dentre os métodos de realocaçao iterativa com número fixo de grupos assume, para este trabalho, uma importância particular o algoritmo das K-Medianas, pela sua utilização no método de programação matemática de Mulvey e Crowder, aqui apresentado no Capítulo VI.

Desenvolvido por MULVEY e CROWDER[29], o algoritmo das K-Medianas toma a forma:

Passo 0: Faça k = 0

Passo 1: Se k = 0, tome uma partição inicial qualquer de E em m grupos;
senão, forme nova partição alocando cada elemento à media na de grupo mais próxima; Calcule $f(P_m^k)$; Faça K = k + 1

Passo 2: Calcule as medianas dos grupos formados no passo anterior

Passo 3: Repita os Passos 1 e 2 até que $f(P_m^k) = f(P_m^{k-1})$.

Como se nota, esse algoritmo é idêntico ao de Forgy, diferindo apenas na determinação dos pontos sementes (centróides, no caso de Forgy e medianas, no caso de K-Medianas). Evidentemente, o algoritmo K-Medianas procura minimizar a dispersão via mediana de grupo, definida no item II.4, fornecendo uma solução que é um ótimo local para o problema.

A prova da convergência do algoritmo pode ser efetuada como apresentado a seguir. Seja

$$L_m^k = \{A_1^k, \ldots, A_m^k\}$$

o conjunto de medianas definidas na k-ésima iteração. O algoritmo gera então uma sequência

$$\{v^k\} = \{(P_m^k, L_m^k)\}.$$

Tomando agora

$$\phi(v^k) = \phi(P_m^k, L_m^k) = \sum_{j=1}^{m} \sum_{e_i \in g_j^k} d(X_i, A_j^k) \qquad (V.5)$$

tem-se o seguinte teorema:

Teorema V.1: a sequência $\{\phi(v^k)\} = \{f(P_m^k)\}$ é monotonamente decrescente e convergente.

Demonstração:

(i) o conjunto S(n, m), de partições distintas P_m, dos n elementos de E em m grupos, é, como foi visto no item III.4, finito. Assim, o conjunto $V = \{(P_m, L_m)\}$ também o é;

(ii) $\phi(v^k) = f(P_m^k) \geq \phi(v^{k+1}) = f(P_m^{k+1})$: (V.6)

Dado L_m^k, a alocação feita no Passo 1, para $k \geq 1$, é a que minimiza $\phi(\cdot, L_m^k)$. Isso pode ser notado se o problema da alocação dos elementos $\{e_1, \ldots, e_n\}$ a grupos com pontos sementes dados pelas medianas $\{A_1^k, \ldots, A_m^k\}$ for descrito da forma:

Minimizar $\sum_{e_i \in E} \sum_{A_j^k \in L_m^k} d_{ij} x_{ij}$

Sujeito a $\sum_{A_j^k \in L_m^k} x_{ij} = 1, \quad \forall e_i \in E$

$x_{ij} \in \{0, 1\}, \quad \forall e_i \in E, A_j^k \in L_m^k,$

onde d_{ij} = distância do elemento e_i ao ponto semente (mediana) A_j^k

$$x_{ij} = \begin{cases} 1 \text{ se } e_i \text{ for alocado ao grupo que tem } A_j^k \text{ como ponto semente;} \\ 0 \text{ em caso contrário.} \end{cases}$$

O problema pode ser reescrito da forma:

Minimizar $\sum_{A_j^k \in L_m^k} d_{1j} x_{1j} + \ldots + \sum_{A_j^k \in L_m^k} d_{nj} x_{nj}$

Sujeito a

$$\sum_{A_j^k \in L_m^k} x_{1j} = 1$$

$x_{1j} \in \{0,1\}$

$\vdots$

$$\sum_{A_j^k \in L_m^k} x_{nj} = 1$$

$$x_{nj} \in \{0,1\}$$

Nota-se que esse problema corresponde a n problemas separáveis da forma:

Minimizar $$\sum_{A_j^k \in L_m^k} d_{ij} \; x_{ij}$$

Sujeito a

$$\sum_{A_j^k \in L_m^k} x_{ij} = 1$$

$$x_{ij} \in \{0,1\}$$

A solução ótima desse i-ésimo problema separável é trivial: basta tomar como $x_{ij} = 1$ aquele para o qual d_{ij} é mínimo, tornando nulas as demais variáveis.

No entanto, é exatamente isso que o algoritmo das K-Medianas faz em seu primeiro passo, quando aloca o elemento e_i ao ponto semente (mediana) mais próximo.

Assim, como P_m^{k+1} é a partição que minimiza $\phi(.,L_m^k)$, no caso particular tem-se:

$$\phi(P_m^k, L_m^k) \geq \phi(P_m^{k+1}, L_m^k). \qquad (V.7)$$

Por outro lado e por definição de mediana de grupo,

$$\phi(P_m^{k+1}, L_m^k) \geq \phi(P_m^{k+1}, L_m^{k+1}). \qquad (V.8)$$

Pelas inequações (V.7) e (V.8), tem-se:

$$\phi(P_m^k, L_m^k) = f(P_m^k) \geq \phi(P_m^{k+1}, L_m^k) \geq$$

$$\geq \phi(P_m^{k+1}, L_m^{k+1}) = f(P_m^{k+1}),$$

O que prova a afirmação (V.6), repetida abaixo:

$$\phi(v^k) = f(P_m^k) \geq \phi(v^{k+1}) = f(P_m^{k+1}).$$

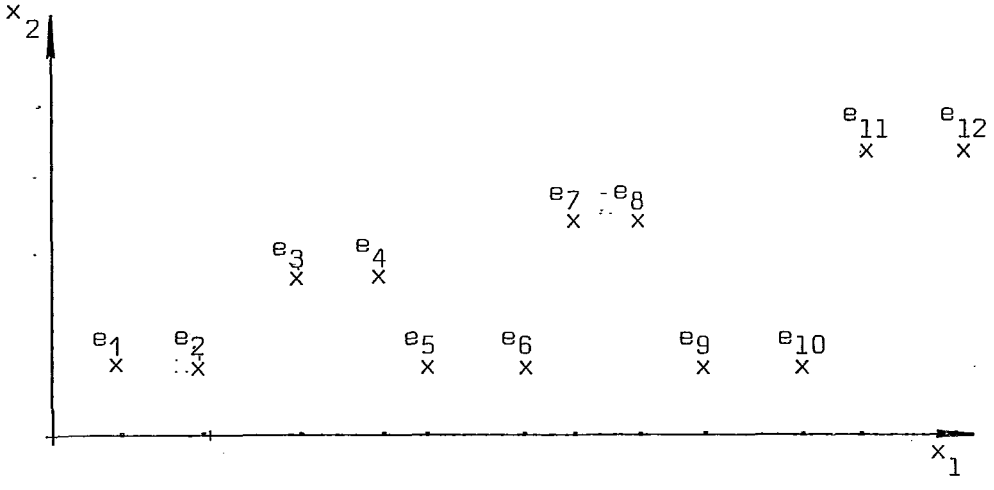
Assim, a seqüência $\{\phi(v^k)\}$ é monotonamente decrescente.

(iii) $\{\phi(v^k)\}$ converge: se todos os v^k gerados pelo algoritmo forem distintos, tem-se uma sequência monótona decrescente em um conjunto finito e assim, neste caso, o método converge. Se, por outro lado, os v^k gerados não forem distintos, isto é, se houver ciclagem e como $f(P_m^k) \leq f(P_m^{k-1})$, a regra de parada $f(P_m^k) = f(P_m^{k-1})$ garante o término do algoritmo. Assim, nos dois casos possíveis o algoritmo termina em um número finito de iterações, isto é, converge, o que completa a demonstração do teorema.

Exemplo:

Seja o mesmo exemplo do item IV.2.5:

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 1 & 2 & 2 & 1 & 1 & 3 & 3 & 1 & 1 & 4 & 4 \end{bmatrix}$$



Este problema também foi resolvido com o auxílio da rotina CLSTROT, apresentada no Anexo 1. Tomou-se m = 4 e adotou-se como solução inicial a partição gerada pelo método de Ward (ver item IV.2.5):

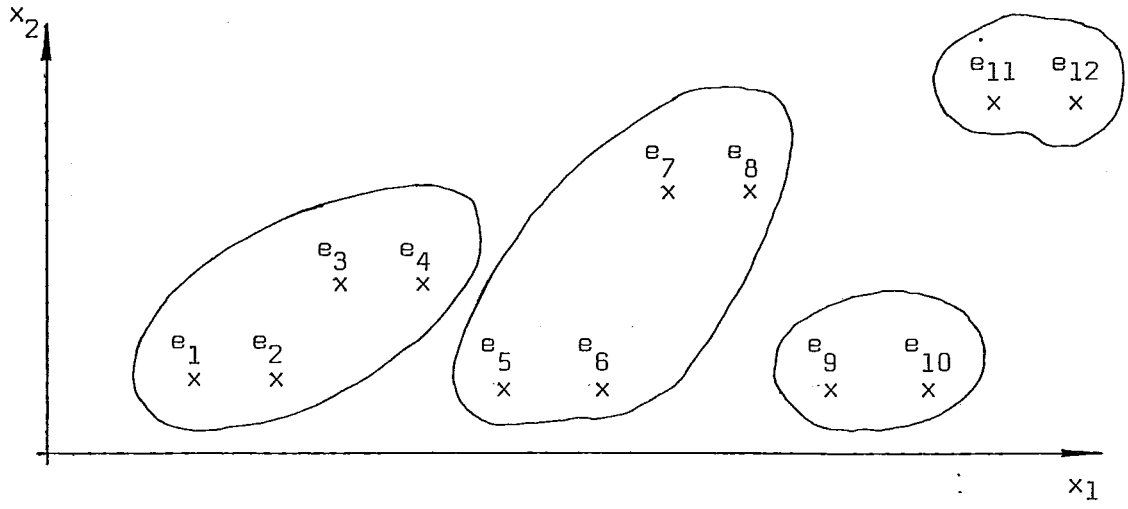
$g_1 = \{e_1, e_2, e_3, e_4\}$

$g_2 = \{e_5, e_6, e_7, e_8\}$

$g_3 = \{e_9, e_{10}\}$

$g_4 = \{e_{11}, e_{12}\}$

Graficamente,



O método convergiu em duas iterações, conforme ilustrado pela Tabela (V.1):

Tabela V.1 - Evolução da função objetivo - Método das K-Medianas

Iteração	Função Objetivo
0	24
1	21
2	21

A solução foi a seguinte:

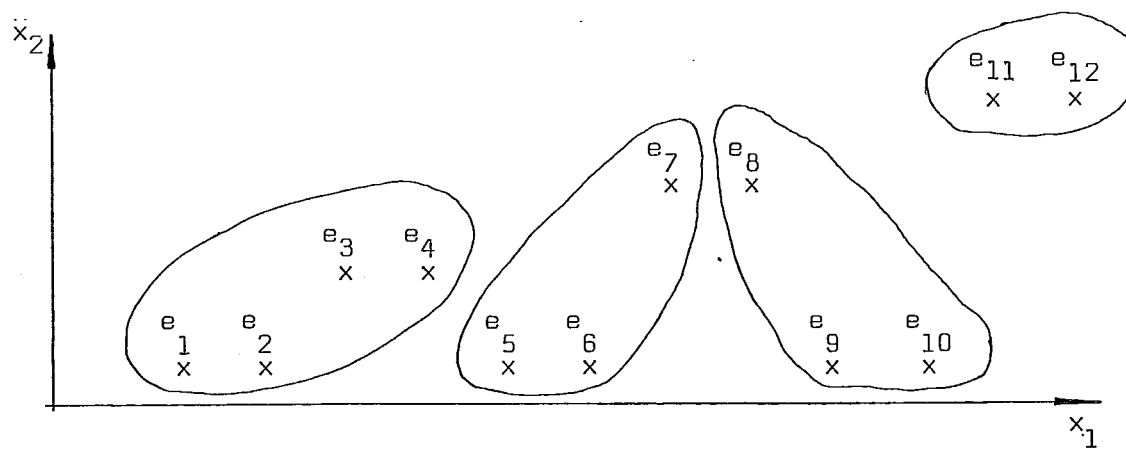
$g_1 = \{e_1, e_2, e_3, e_4\}$

$g_2 = \{e_5, e_6, e_7\}$

$g_3 = \{e_8, e_9, e_{10}\}$

$g_4 = \{e_{11}, e_{12}\}$

Graficamente,



O agrupamento obtido é mais "compacto" que o anterior. Note-se também o decréscimo obtido na primeira iteração, servindo a segunda apenas para satisfazer ao teste de convergência. Não se pode, evidentemente, afirmar, com base nos elementos disponíveis até este Capítulo, se essa solução é um ótimo global. Tal tarefa será executada com base no método de Mulvey e Crowder, no item VI.4.

V.4 - MÉTODOS COM NÚMERO VARIÁVEL DE GRUPOS

No Capítulo anterior teceu-se uma série de comentários sobre a existência de grupos, subgrupos etc. e como uma análise da evolução do processo hierarquizado poderia auxiliar o analista na determinação do número m de grupos a ser adotado na solução final.

Essa discussão poderia ser aqui complementada através da apresentação de um exemplo, desenvolvido por ANDERBERG[1] e descrito pela figura a seguir:

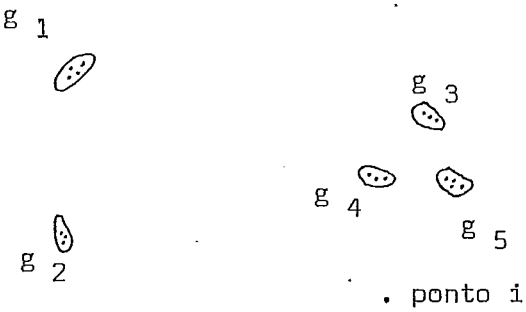


Fig. V.2 - Configuração de um conjunto E.

Esse conjunto E consiste em cinco grupos naturais e um ponto isolado, que não pertence a nenhum desses grupos.

Evidentemente, o analista não saberá a priori da existência de pontos isolados. Caso fosse aplicado a um conjunto como E um algoritmo de realocação iterativa com um número fixo m = 3 de grupos, possivelmente a solução seria:

$$P_3 = \{g_1, g_2\}, \{g_3, g_4, g_5\}, \{\text{ponto isolado}\}.$$

A rigor, esta solução corresponde a uma partição significativa de E em dois grupos e não em três, uma vez que este ponto isolado pode corresponder a uma exceção não significativa ou mesmo a um erro de medição.

Assim, permitindo-se, de uma forma sistemática, incluída no algoritmo, uma variação no número de grupos, poder-se-ia, também, obter uma solução para m = 4 da forma:

$$P_4 = \{g_1\}, \{g_2\}, \{g_3, g_4, g_5\}, \{\text{ponto isolado}\}.$$

Agora foi obtida uma partição de E, signicativa, em três grupos(sem

considerar o ponto isolado).

Por outro lado, pode ser que haja, como já se comentou no Capítulo anterior, mais de um nível de agregação significativo em E. No exemplo da Figura (V.2), caso não se considere o ponto isolado, os grupos g_3, g_4 e g_5 poderiam ser agrupados para se formar uma tipologia significativa com três grupamentos, como se comentou no parágrafo anterior. Por outro lado, um esforço no sentido de se criar quatro ou seis grupos (sem considerar o ponto isolado) iria gerar grupos mal definidos.

Os métodos a serem discutidos a seguir, também apresentados em ANDERBERG[1], foram desenvolvidos para tratar esses problemas. Esses métodos, como se verá adiante, dependem criticamente do valor de certos parâmetros de controle cuja determinação é feita através de tentativa e erro, o que prejudica sensivelmente sua utilização prática. Pela sua própria natureza, são métodos que exigem grande esforço computacional. No entanto, uma apresentação sucinta dessas técnicas ilustra vários aspectos teóricos interessantes, como se verá a seguir.

O primeiro desses métodos deve-se a MACQUEEN[27] e consiste numa variante da técnica K-Médias, vista no item anterior. Os passos do algoritmo são como se segue:

Passo 0: Escolha valores para os três parâmetros M (número inicial de grupos), A (parâmetro de aglomeração) e R (parâmetro de refinamento)

Passo 1: Tome os M primeiros elementos de E como sementes de M grupos iniciais

Passo 2: Calcule as distâncias entre essas M primeiras sementes. Se a menor distância entre duas sementes for igual ou inferior a A, reuna essas duas sementes em um só grupo, calculando as centróides do grupo resultante. Recalcule as distâncias entre essa nova centróide e as demais sementes. Repita esse procedimento até que a menor distância entre sementes (ou centróides) seja superior a A

Passo 3: Existem ainda (n-M) elementos não agrupados. Repita então (n-M) vezes o seguinte procedimento:

- escolha um elemento não agrupado;
- se a distância desse elemento à centróide mais próxima for igual ou superior ao parâmetro de refinamento R, tome esse elemento em questão como semente de um novo grupo;

- se a distância do elemento à centróide mais próxima for inferior a R, aloque o elemento ao grupo associado a essa centróide. Atualize a centróide do novo grupo formado e calcule a distância dessa nova centróide às centróides dos demais grupos. Se agora a menor dessas distâncias for igual ou inferior a A (parâmetro de agrupamento), reuna os dois grupos e repita esse procedimento até que a menor distância entre centróides seja superior a A

Passo 4: Após a alocação de todos os elementos, tome as centróides existentes, sejam quantas forem, como pontos sementes fixos e realoque cada elemento ao ponto semente mais próximo.

Como se nota, os Passos 1, 2 e 3 geram uma partição inicial em que as centróides dos grupos possuem entre si uma distância superior a A (esse fato é garantido pelo último procedimento do Passo 3). Permitindo-se que grupos com centróides próximas (distâncias inferior a A) sejam reunidos, evita-se que se crie distinções finas que dividam artificialmente os grupos. Criando-se novos grupos quando as distâncias são elevadas (superior a R), procura-se a identificação de novos grupos ainda não detectados e de pontos isolados. No entanto, ANDERBERG[1] comenta que, além da observação de que $R > A$, não é disponível nenhuma outra regra prática para a determinação desses parâmetros, a não ser que eles podem ser definidos como frações da dispersão do conjunto E.

O Passo 4 do algoritmo consiste apenas numa iteração de melhoria da função objetivo, em termos de soma dos quadrados dentro dos grupos, tal como efetuado no algoritmo K-Médias (apresentado no item V.3). Esse Passo 4 assume grande importância no algoritmo em virtude do fato de que aqui, nos Passos 1, 2 e 3, a exemplo do que ocorre no K-Médias, o resultado depende da ordem dos elementos e_1, e_2, ..., e_n, no conjunto E.

Finalmente, cabe apresentar também a técnica ISODATA (Iterative Self - Organizing Data Analysis Technique). O método foi desenvolvido no Stanford Research Institute e sua apresentação mais conhecida foi efetuada por BALL e HALL[3]. Existem, no entanto, várias versões do método. ANDERBERG[1], por exemplo, apresenta uma versão com oito parâmetros de controle. Neste trabalho optou-se pela versão indicada por DURAN e ODELL[10], que é bem mais simples.

Nessa versão, os passos do algoritmo são os seguintes:

Passo 0: Gere uma partição inicial em m grupos selecionando aleato

riamente m elementos como pontos semente e alocando os (n - m) elementos restantes ao ponto semente mais próximo

Passo 1: Calcule as distâncias médias quadráticas (ver item IV.2.6) entre os m grupos formados no passo anterior. Combine os grupos que tiverem entre si uma distância média quadrática inferior a R.

Passo 2: Divida, por um método qualquer, em dois grupamentos, os grupos onde a variância s^2 de qualquer variável for superior a um valor limite v. Assim a variância interna S^2 (ver item II.4.2) é limitada superiormente:

$$S^2 \leq pv$$

Passo 3: Repita os Passos 1 e 2 até que a convergência seja obtida, ou seja, uma repetição dos Passos 1 e 2 não muda a configuração dos grupos.

Como se nota, esse algoritmo, a partir de uma configuração inicial qualquer, procura criar grupos cuja dispersão seja controlada pelo usuário (a variância interna é limitada em pv), evitando a formação de grupos mal definidos pela união daqueles considerados próximos (que possuam entre si uma distância inferior a R).

Esse algoritmo apresenta a desvantagem de poder ciclar: se R for bem inferior a pv e, por consequência, a S^2, então é possível que dois grupos fiquem sempre sendo reunidos e separados nos Passos 1 e 2.

O Teorema (IV.3), que permitiu essa observação sobre ciclagem, também indica que não há solução para contornar definitivamente o problema. O Teorema informa que, no caso de distância média quadrática, ao se reunir dois grupos g_I e g_J para formar um grupo g_L, tem-se:

$$S_L^2 \geq \frac{n_I\, n_J}{n_I + n_J}\, D_{IJ}^2 \quad ,$$

ou seja,

$$S_L^2 \geq \alpha\, D_{IJ}^2 \quad , \quad 0 < \alpha < \infty \, .$$

Assim, ter-se-ia que os parâmetros de controle deveriam guardar a relação:

$$pv > \alpha \; R \quad , \quad 0 < \alpha < \infty \, .$$

O parâmetro α depende dos tamanhos n_I e n_J dos grupos, que não é conhecido a priori. Assim, nada mais se pode fazer senão tomar R bastante

inferior a pv e controlar o número de iterações do método, para evitar, mesmo nesse caso, os efeitos da ciclagem e permitir a parada do algoritmo.

V.5 - CONCLUSÃO

Em vista do exposto, fica bastante evidente a extrema eficiência computacional dos métodos de realocação iterativa com número fixo de grupos. Sempre que se puder abrir mão da necessidade de uma solução ótima, deve-se optar por um dos métodos desse tipo, que são, dentre os apresentados neste trabalho, os mais rápidos.

Os métodos com número variável de grupos, como se viu, apesar de serem teoricamente interessantes, apresentam inúmeras dificuldades práticas para a sua implementação, o que não impede que o método ISODATA, por exemplo, seja citado na literatura a respeito como bastante popular (ver, p. ex., DURAN e ODELL[10]).

Finalmente e complementando a discussão que vem sendo efetuada até aqui sobre o número de grupos a serem adotados nas soluções finais, cabe mencionar a seguinte técnica, também citada em ANDERBERG[1], que procura explorar a eficiência computacional dos métodos realocativos com m fixo, sua vantagem sobre os métodos hierarquizados no que se refere à redistribuição de elementos nos grupos e a necessidade de uma avaliação dos níveis significativos de m:

Passo 0: Comece com uma partição qualquer inicial, em um número m de grupos

Passo 1: Utilize um algoritmo qualquer de realocação iterativa com número fixo de grupos, para determinar uma partição de E em m grupos (no caso de utilização de um algoritmo iterativo, defina um número máximo de iterações entre 2 e 5, tendo em vista que, como foi visto no item V.1, é nessas iterações iniciais que as maiores melhorias na função objetivo costumam ocorrer)

Passo 2: Faça m = m - 1. Se m agora for igual a um limite inferior, pare. Senão, calcule as distâncias entre os (m + 1) grupos existentes, una em um só grupo os dois grupamentos mais próximos e vá para o Passo 1.

A análise da função objetivo, nas soluções para os diversos valores de m, pode indicar as partições mais significativas, de uma forma semelhante ao indicado no item IV.3, para os métodos hierarquizados.

VI - MÉTODOS DE PROGRAMAÇÃO MATEMÁTICA

VI.1 - INTRODUÇÃO

Dentre os modelos citados no item III.4 para a resolução do problema de análise de grupamento, foram selecionados para apresentação neste trabalho, dois modelos de programação inteira propostos por VINOD[41]. Tal escolha se deve ao fato de que, apesar da solução via programação inteira desses modelos não ser em geral viável, tendo em vista o excessivo tempo de processamento que os métodos de programação inteira costumam exigir, é possível, para os dois modelos, como se verá adiante, obter uma solução ótima de forma eficiente através de outras técnicas de programação matemática.

O primeiro dos modelos de Vinod, aqui denominado "Primeiro Modelo de Vinod", apresenta, como aponta aquele autor, uma profunda relação com problemas de localização (esse aspecto será abordado no item VI.6). Para esse modelo, MULVEY e CROWDER[29] propuseram um método que, a partir de otimização por subgradientes, pode fornecer de forma eficiente uma solução ótima para o problema.

O aqui denominado "Segundo Modelo de Vinod" será apresentado no item VI.3, apenas em sua versão univariada (p=1). Para esse caso univariado, RAO[33] forneceu um algoritmo de programação dinâmica que permite uma solução eficiente para o problema. Embora VINOD[41] também apresente para esse modelo uma versão multivariada, a opção pelo caso univariado se deve a que, pelo menos ao conhecimento do autor, não é disponível, para p > 1, um método eficiente para a obtenção de solução ótima para o "Segundo Modelo de Vinod".

VI.2 - PRIMEIRO MODELO DE VINOD

De acordo com esse modelo, o problema de análise de grupamento pode ser formulado como se segue:

$$\text{Minimizar} \quad \sum_{i \in I} \sum_{j \in J} d_{ij} \, y_{ij} \qquad \text{(VI.2.1)}$$

$$\text{Sujeito a} \quad \sum_{j \in J} y_{ij} = 1 \quad , \quad i \in I \qquad \text{(VI.2.2)}$$

$$\sum_{j \in J} y_{jj} = m \quad , \qquad \text{(VI.2.3)}$$

$$y_{ij} \leq y_{jj} \quad , \quad i \in I \ , \quad j \in J \qquad \text{(VI.2.4)}$$

$$y_{ij} \in \{0,1\} \quad , \quad i \in I \ , \quad j \in J \qquad \text{(VI.2.5)}$$

onde ,

I = conjunto dos n elementos;

J = conjunto de medianas elegíveis (usualmente $J = I$);

d_{ij} = distância entre os elementos e_i e e_j;

$y_{ij} = \begin{cases} 1 & \text{se } e_i \text{ é alocado ao grupo que tem } e_j \text{ como mediana} \\ 0 & \text{em caso contrário ;} \end{cases}$

$y_{jj} = \begin{cases} 1 & \text{se } e_j \text{ é mediana} \\ 0 & \text{em caso contrário; e} \end{cases}$

m = número de grupos.

Deve-se notar que:

- a função objetivo a ser minimizada é a dispersão via mediana de grupo, apresentada no item II.4.4;
- o objetivo do modelo é selecionar m medianas (centros de grupo) dentre as possíveis, alocando os demais elementos aos grupos representados por essas medianas, de forma a que a soma das distâncias dos elementos aos respectivos centros de grupo seja mínima;
- a restrição (VI.2.2) garante que cada elemento só poderá ser alocado a um grupo;
- a restrição (VI.2.3) impõe que o número de grupos seja igual a m; e
- a restrição (VI.2.4) indica que o elemento e_i só poderá ser alocado ao grupo representado por e_j se e_j for mediana.

Resta mostrar que se $y_{jj} = 1$ na solução ótima do problema, e_j real é mediana. Isso pode ser visto através da função objetivo (VI.2.1) . Seja $g_j = \{ e_i \mid y_{ij} = 1 \}$. Note-se que $e_j \varepsilon g_j$. Supondo, por absurdo , que a mediana do grupo fosse $e_k \varepsilon g_j$ e não e_j, ter-se-ia

$$\sum_{e_i \varepsilon g_j} d_{ij}\, y_{ij} > \sum_{e_i \varepsilon g_j} d_{ik}\, y_{ik} .$$

Assim, fazendo $y_{kk} = 1$ e $y_{jj} = 0$ seria possível obter-se uma melhoria na função objetivo, e $y_{jj}=1$ não corresponderia a uma solução ótima. Como a solução é ótima, não existe $e_k \varepsilon g_j$ que permita tal melhoria e assim e_j é a mediana de g_j.

Esse primeiro modelo de Vinod corresponde a um problema de programação linear inteira, com n^2 variáveis binárias, o que restringe sua aplicação a problemas de porte muito reduzido. No entanto, como se verá no item VI.5, é possível, a partir do problema dual Lagrangiano, obter-se para este modelo de Vinod uma solução eficiente via otimização por subgradientes, na forma proposta por MULVEY e CROWDER[29].

VI.3 - SEGUNDO MODELO DE VINOD - CASO UNIVARIADO

VI.3.1 - INTRODUÇÃO

Seja $X = \{x, \ldots , x_n\}$, $x_i \varepsilon R$, o conjunto de medidas associadas a um conjunto $E = \{e_1, \ldots , e_n\}$ de n elementos, tendo-se como objetivo obter uma partição de E em m grupos que minimize a soma dos quadrados dentro do grupo W (ver item II.4.1):

$$W = \sum_{j=1}^{m} W_j = \sum_{j=1}^{m} \sum_{e_i \varepsilon g_j} d_2^2 (x_i, \bar{x}_j) ,$$

onde

$$\bar{x}_j = \frac{1}{n_j} \sum_{e_i \varepsilon g_j} x_i .$$

Como $x_i \varepsilon R$, tem-se

$$W = \sum_{j=1}^{m} W_j = \sum_{j=1}^{m} \sum_{e_i \varepsilon g_j} (x_i - \bar{x}_j)^2 \qquad \text{(VI.3.1)}$$

O segundo modelo de Vinod - caso univariado, procura resolver esse problema utilizando a propriedade da cadeia - caso univariado (ver, p. ex., VINOD[41], RAO[33] ou DURAN e ODELL[10]). No entanto, para que se possa enunciar a propriedade, faz-se necessária uma redefinição do problema.

Seja $Z = \{z_1, \ldots, z_n\}$ o conjunto resultante da ordenação, em ordem crescente, do conjunto X, isto é, Z é tal que:

$$z_1 \leq z_2 \leq \ldots \leq z_n \ .$$

Seja também $E' = \{e'_1, \ldots, e'_n\}$ o conjunto de elementos associados a Z, resultante da reordenação dos elementos do conjunto inicial E.

A propriedade da cadeia pode então ser enunciada da seguinte forma: " se e'_i pertence, na solução ótima de um problema de análise de grupamento univariado (p=1), ao grupo g_k, do qual e'_j é o primeiro elemento , então todos os elementos e'_r, $r = (j+1), \ldots, (i-1)$, também pertencem a g_k ".

Em outras palavras, a propriedade em questão afirma que, na solução ótima, se e'_j e e'_i pertencem ao mesmo grupo, toda a cadeia de elementos entre e'_j e e'_i, ou seja, tais que

$$z_j \leq z_{j+1} \leq \ldots \leq z_i \quad ,$$

também pertence ao grupo: não existem "buracos" na cadeia.

Uma discussão detalhada da propriedade da cadeia é apresentada em VINOD[41]. De uma maneira geral, a justificativa para a validade da regra se baseia em que, se a cadeia não estivesse preenchida, isto é, se nela houvesse "buracos", sempre se poderia formar grupos mais homogêneos, com menor soma dos quadrados dentro dos grupos, através de uma troca de elementos que restabelecesse a cadeia. Como exemplo, seja o caso abaixo:

$$\ldots \{e'_j, \ldots, e'_{i-1}, e'_{i+1}\} \{e'_i, e'_{i+2}, \ldots\} \ldots$$

É evidente que, ao se colocar e'_i no seu próprio lugar na cadeia, formando-se os grupos

$$\ldots \{e'_j, \ldots, e'_{i-1}, e'_i\} \{e'_{i+1}, e'_{i+2}, \ldots\} \ldots$$

obter-se-ia uma melhoria na soma dos quadrados dentro dos grupos, uma vez que, por construção, $z_i \leq z_{i+1}$.

A partir da regra da cadeia, verifica-se então que encontrar uma partição de E' (ou, evidentemente, de E) que minimize a soma dos quadrados dentro dos grupos, significa determinar uma solução P_m^* ,

$$P_m^* = \{e'_1, e'_2, \ldots\} \ldots \{e'_j, \ldots, e'_i\} \ldots \{\ldots, e'_n\} ,$$

que minimize (ver equação (VI.3.1)) :

$$W = \sum_{j=1}^{m} W_j = \sum_{j=1}^{m} \sum_{e'_i \in g_j} (z_i - \bar{z}_j)^2 .$$

Seja, por outro lado, o grupo g_K tal que

$$g_K = \{e'_j, \ldots, e'_i\} .$$

O elemento e'_j , primeiro elemento de g_K, é dito o líder do grupo g_K. Além disso, W_K, a soma dos quadrados dentro desse grupo, pode ser calculada recursivamente a partir do teorema (IV.1) da seção IV.2.5, que informa que, ao se unirem dois grupos g_I e g_J para se formar um grupo g_L , tem-se

$$W_L = W_I + W_J + \frac{n_I \, n_J}{n_I + n_J} \, d_{IJ}^2 , \qquad \text{(VI.3.2)}$$

onde d_{IJ}^2 corresponde ao quadrado da distância euclidiana entre as centróides (médias) de g_I e g_J .

A partir desse teorema, pode-se calcular recursivamente W_K , a soma dos quadrados dentro de um grupo g_K , da seguinte forma:
seja $g_K = \{e'_j, e'_{j+1}, \ldots, e'_i\}$. Tome-se inicialmente o grupo $\{e'_j\}$. A soma dos quadrados dentro desse grupo composto apenas por um elemento evidentemente é nula.

Ao se formar o grupo $\{e'_j, e'_{j+1}\}$, tem-se, em termos da equação (VI.3.2):

$$g_I = \{e'_j\} \quad ; \quad n_I = 1 \quad ; \quad W_I = 0$$

$$g_J = \{e'_{j+1}\} \quad ; \quad n_J = 1 \quad ; \quad WJ = 0$$

$$W_L = \frac{1}{2}(z_j - z_{j+1})^2 = \frac{1}{2}(z_{j+1} - z_j)^2$$

Formando-se o grupo $\{e'_j, e'_{j+1}, e'_{j+2}\}$ tem-se

$$g_I = \{e'_j, e'_{j+1}\} \quad ; \quad n_I = 2 \quad ; \quad W_I = \frac{1}{2}(z_j - z_{j+1})^2$$

$$g_J = \{e'_{j+2}\} \quad ; \quad n_J = 1 \quad ; \quad W_J = 0$$

$$W_L = \frac{1}{2}(z_{j+1} - z_j)^2 + \frac{2}{3}(z_{j+2} - \frac{1}{2}\sum_{q=j}^{j+1} z_q)^2 .$$

No caso geral, seja ΔW_j^r o acréscimo em W_k devido à inclusão do elemento e'_r no grupo $g_k = \{e'_j, \ldots, e'_{r-1}\}$. Em termos da equação (VI.3.2) ter-se-ia:

$$g_I = \{e'_j, \ldots, e'_{r-1}\} \; ; \; n_I = r-j \; ; \; W_I = \sum_{e'_i \in g_I} (z_i - \bar{z}_I)^2$$

$$g_J = \{e'_r\} \quad ; \quad n_J = 1 \quad ; \quad W_J = 0$$

e, finalmente,

$$W_L = W_I + \Delta W_j^r = W_I + \frac{r-j}{(r-j)+1}(z_r - \frac{1}{r-j}\sum_{q=j}^{r-1} z_q)^2 \quad \text{(VI.3.3)}$$

Assim, o cálculo de W_k se resume em utilizar a equação (VI.3.3) para afirmar que, para o grupo $g_k = \{e'_j, \ldots, e'_i\}$, tem-se:

$$W_k = \sum_{r=j+1}^{i} \Delta W_j^r \qquad \text{(VI.3.4)}$$

VI.3.2 - DESCRIÇÃO DO MODELO

Como, a partir da equação (VI.3.3), o acréscimo ΔW_j^i, na soma dos quadrados dentro dos grupos W, ao se incluir o elemento e'_i ao grupo $\{e'_j, \ldots, e'_{i-1}\}$, é dado por:

$$\Delta W_j^i = \frac{i-j}{(i-j)+1}(z_i - \frac{1}{i-j}\sum_{q=j}^{i-1} z_q)^2 \quad , \quad \text{(VI.3.5)}$$

tem-se então o segundo modelo de Vinod - caso univariado:

$$\text{Minimizar} \quad \sum_{j\in J} \sum_{\substack{i\in I \\ i>j}} \Delta W_j^i \, y_{ij} \qquad \text{(VI.3.6)}$$

$$\text{Sujeito a} \quad \sum_{j\in J} y_{ij} = 1 \quad , \quad i\in I \qquad \text{(VI.3.7)}$$

$$\sum_{j\in J} y_{jj} = m \qquad \text{(VI.3.8)}$$

$$y_{jj} \geq y_{j+1,j} \geq \ldots \geq y_{nj}, \quad j\in J \qquad \text{(VI.3.9)}$$

$$y_{ij} = 0 \quad , \quad i < j \qquad \text{(VI.3.10)}$$

$$y_{ij} \in \{0,1\} \quad , \quad i\in I, \; j\in J \qquad \text{(VI.3.11)}$$

onde

I = Conjunto dos n elementos

J = Conjunto dos líderes de grupo elegíveis ($J=I$)

ΔW_j^i = dado pela equação (VI.3.5)

$$y_{ij} = \begin{cases} 1 & \text{se } e_i' \text{ pertence ao grupo liderado por } e_j' \\ 0 & \text{em caso contrário} \end{cases}$$

$$y_{jj} = \begin{cases} 1 & \text{se } e_j' \text{ é líder} \\ 0 & \text{em caso contrário} \end{cases}$$

m = número de grupos.

Deve-se enfatizar então que:

- a função objetivo a ser minimizada é a soma dos quadrados dentro dos grupos W, definida no item II.4.1, onde ΔW_j^i representa o acréscimo em W_k ao se unir e_i' ao grupo $\{e_j', \ldots, e_{i-1}'\}$, ou

seja, tudo se passa como se as somas dos quadrados dentro dos grupos g_k, isto é, W_k, fossem calculadas recursivamente;

- o objetivo do modelo é selecionar m líderes de grupo (o que , tendo em vista a propriedade da cadeia, determina automaticamente a partição de E' em n grupos) de forma a minimizar a soma dos quadrados dentro dos grupos;

- a restrição (VI.3.7) garante que cada elemento só poderá ser alocado a um grupo;

- a restrição (VI.3.8) impõe a formação de m grupos;

- as restrições (VI.3.9) e (VI.3.10) correspondem à propriedade da cadeia. O que as restrições impõem é que y_{ij} só pode ser igual a 1 se $i \geq j$ e $y_{jj} = y_{j+1,j} = \ldots = y_{i-1,j} = 1$.

VI.4 - MÉTODO DE MULVEY E CROWDER[29]

VI.4.1 - INTRODUÇÃO

O primeiro modelo de Vinod (ver item VI.2) é descrito como:

Problema V1 : Minimizar $\sum_{i \in I} \sum_{j \in J} d_{ij} y_{ij}$ (V1.1)

Sujeito a $\sum_{j \in J} y_{ij} = 1 \quad , \quad i \in I$ (V1.2)

$\sum_{j \in J} y_{jj} = m$ (V1.3)

$y_{ij} \leq y_{jj} \quad , \quad i \in I \; , \; j \in J$ (V1.4)

$y_{ij} \in \{0,1\}$ (V1.5)

Esse problema pode ser reescrito da forma:

Problema PV : Minimizar $\sum_{i\in I} \sum_{j\in J} d_{ij}\ y_{ij}$ (PV.1)

Sujeito a $1 - \sum_{j\in J} y_{ij} = 0 \quad , \quad i\in I$ (PV.2)

$y_{ij} \in C$, sendo o conjunto C definido pelas restrições (V1.3) a (V1.5).

O dual Lagrangiano de PV (ver p.ex. BAZARAA e SHETTY[5] para uma apresentação de dualidade Lagrangiana) pode, por sua vez, ser descrito como:

Problema DV : Maximizar $\theta(u_1, u_2, \ldots, u_n)$

onde $\theta(u_1, u_2, \ldots, u_n) = \min \{ \sum_{i\in I} \sum_{j\in J} d_{ij}\ y_{ij} +$

$+ \sum_{i\in I} u_i(1 - \sum_{j\in J} y_{ij}) : y_{ij} \in C \}$

A equivalência entre DV e um problema de otimização por subgradientes (para uma discussão detalhada de otimização por subgradientes, ver, por exemplo, HELD, WOLFE e CROWDER[18]) pode ser verificada ao se observar que o cálculo de $\theta(u_1, u_2, \ldots, u_n)$, em DV, é um problema de programação linear inteira para o qual a solução se dá em um conjunto finito de soluções viáveis (note-se que o cálculo de $\theta(u_1, \ldots, u_n)$ corresponde a uma minimização feita nas variáveis y_{ij}, considerando-se as variáveis duais u_i como fixas). Tomando k como índice dessas soluções (k=1, ..., K), DV pode

ser reescrito como:

Problema DV : Maximizar $\theta(u_1, u_2, \ldots, u_n)$

onde $\theta(u_1, u_2, \ldots, u_n) = \min\{z^k + \sum_{i=1}^{n} u_i v_i^k : k=1, \ldots, K\}$

$$z^k = \sum_{i\in I} \sum_{j\in J} d_{ij}\, y_{ij}^k , \quad k = 1, \ldots, K$$

$$v_i^k = 1 - \sum_{j\in J} y_{ij}^k , \quad i\in I , \quad k = 1, \ldots, K$$

y_{ij}^k = valor de y_{ij} na k-ésima solução viável.

Assim, DV pode ser resolvido por uma técnica de otimização por subgradientes, técnica essa que produziria uma sequência de soluções duais $(u_1, u_2, \ldots, u_n)^1$, ... , $(u_1, u_2, \ldots, u_n)^*$, que iria convergir para uma solução ótima $(u_1, \ldots, u_n)^*$ de DV.

Pelo Teorema Fraco da Dualidade (ver, por exemplo, BAZARAA e SHETTY[5]), um valor de função objetivo associado a qualquer solução do problema dual DV é menor que o valor da função objetivo ligada a qualquer solução do problema primal PV. Assim, qualquer solução de DV (e particularmente a solução ótima) gera um limite inferior para o valor da função objetivo, na solução ótima, do problema primal PV.

No entanto, não há, nesse caso, qualquer garantia de que o valor das funções objetivo nas soluções ótimas de PV e DV sejam iguais. As condições do Teorema Forte da Dualidade, que garantiriam tal resultado, não são satisfeitas: C não é convexo, uma vez que $y_{ij} \in \{0,1\}$. Uma apresentação detalhada do Teorema Forte da Dualidade é efetuada em BAZARAA e SHETTY[5]. MULVEY e CROWDER[29], no entanto, informam que estudos empíricos (inclusive o de HELD, WOLFE e CROWDER[18]) evidenciaram que a utilidade do método do subgradiente em problemas desse tipo reside na relativa proximidade entre os limites inferiores gerados pelo método e a solução ótima do problema primal.

O método proposto por MULVEY e CROWDER[29] para a resolução do problema PV consiste, essencialmente, em resolver DV, a partir de uma técnica de otimização por subgradientes, utilizando, em cada iteração, as informações da solução dual para se tentar gerar uma solução primal, para PV, melhorada.

Assim, o método pode ser visto como uma técnica primal-dual, que gera duas sequências de soluções: uma converge para a solução do problema dual e a outra, que não converge necessariamente, se constitui em tentativas de geração de soluções primais melhoradas. A partir desse raciocínio ; fica evidente que não há garantia de que o método venha a gerar sempre a solução ótima do problema primal.

Por outro lado, deve-se considerar também que, como indicam MULVEY e CROWDER[29]:

- o método, gerando um limite inferior para a solução ótima de PV, permite que se avalie a qualidade das soluções obtidas por métodos heurísticos como, por exemplo, o método das K-Medianas;

- o problema PV é NP-completo, sendo improvável que um algoritmo exato eficiente venha a ser encontrado; e

- na prática, geralmente, o método fornece a solução ótima de PV em um pequeno número de iterações (quanto a esse aspecto, aqueles autores indicam que, em 10 problemas de diferentes portes, soluções ótimas foram obtidas até um máximo de 120 iterações).

VI.4.2 - DESCRIÇÃO DO MÉTODO

O método de MULVEY e CROWDER[29] se compõe basicamente de três partes:

(i) geração de uma boa solução viável inicial, via técnicas hierarquizadas, sendo a solução posteriormente refinada pelo método das k-Medianas;

(ii) geração de uma solução dual para o problema, via otimização por subgradientes; e

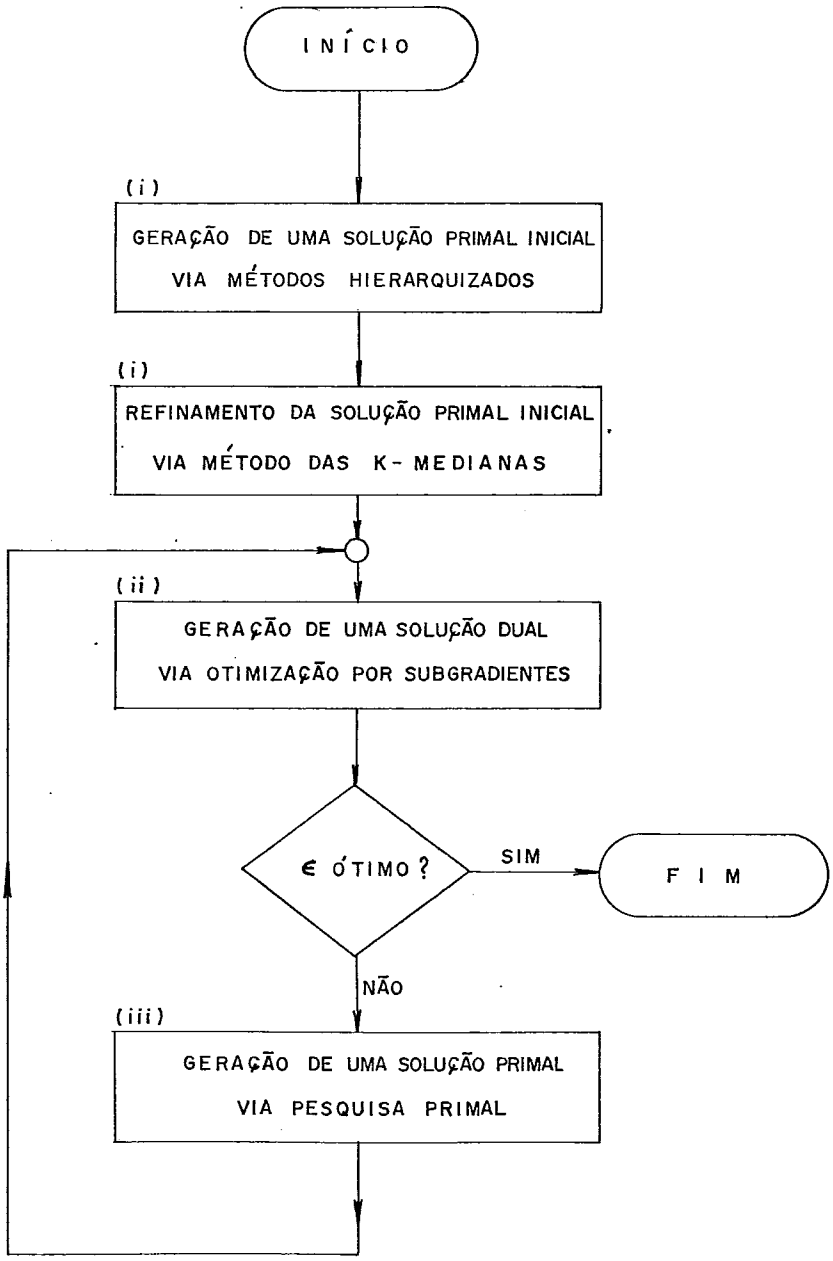
(iii) geração de uma solução primal viável melhorada, a partir de informações fornecidas pela solução dual (Pesquisa Primal).

A parte (i) se constitui em inicialização do problema. As partes (ii) e (iii) constituem o corpo do método, que gera, de forma iterativa, limites inferiores (via solução dual) e superiores (via solução primal), terminando-se o algoritmo quando a diferença ε entre os limites é atingida

ou quando o número de iterações atinge um valor predeterminado.

A sequência de limites inferiores, como foi visto no item anterior, converge para a solução do problema DV. A sequência de soluções primais viáveis não converge necessariamente. MULVEY e CROWDER[29] indicam que, caso se desejasse, seria possível implantar na Pesquisa Primal uma rotina de programação dinâmica que garantiria a obtenção dessa solução primal ótima, mas que isso é desnecessário, uma vez que sua experiência indica que soluções aceitáveis ($\varepsilon \simeq 0{,}01$) são geralmente obtidas pela Pesquisa Primal que será definida adiante.

O algoritmo pode ser melhor entendido através do fluxograma que se segue:



O detalhamento dessas etapas é feito a seguir:

- Inicialização:

Os métodos hierarquizados e o método das k-Medianas foram apresentados detalhadamente nos Capítulos IV e V deste trabalho.

- Solução Dual:

Quanto à otimização por subgradientes, note-se que o problema DV, do item anterior:

Problema DV : Maximizar $\theta(u_1, \ldots, u_n)$

onde $\theta(u_1, \ldots, u_n) = \min\{ \sum_{i\in I} \sum_{j\in J} d_{ij}y_{ij} + \sum_{i\in I} u_i(-\sum_{j\in J} y_{ij}+1) :$

$y_{ij} \in C \}$

pode ser reescrito da forma:

Problema DV : Maximizar $\theta(u_1, \ldots, u_n)$

onde $\theta(u_1, \ldots, u_n) = \min\{ \sum_{i\in I} u_i + \sum_{i\in I} \sum_{j\in J} (d_{ij}-u_i)y_{ij} : y_{ij}\in C\}.$

MULVEY e CROWDER[29] sugerem que, para um dado conjunto $(\hat{u}_1,\ldots,\hat{u}_n)$ o cálculo do mínimo acima pode facilmente ser efetuado tomando $y_{jj} = 1$ para as m menores somas de colunas definidas no conjunto

$$\{ S(j) : S(j) = \sum_{i\in I} \min (d_{ij} - \hat{u}_i, 0) \}$$

e $y_{jj} = 0$ nos demais casos. Empates nas somas de colunas poderiam ser resolvidos arbitrariamente ou através da técnica a ser descrita adiante para a Pesquisa Primal. Designando então o conjunto dos j para os quais $y_{jj} = 1$ como L, determina-se as demais variáveis tomando $y_{ij} = 1$ se $j \in L$ e $(d_{ij} - \hat{u}_i) < 0$, e $y_{ij} = 0$ nos demais casos. Isso significa escolher m medianas de acordo com seus "custos reduzidos" S(.) e alocar um elemento a uma mediana se o custo associado a essa alocação for negativo. Note-se que assim um objeto pode ser alocado a apenas uma mediana, a várias ou a nenhuma.

Por outro lado, a sequência $\{(u_1, \ldots, u_n)^j\}$ é gerada pelo método do subgradiente.

- Solução Primal:

Quanto à solução primal viável do problema PV, MULVEY e CROWDER[29] sugerem que se tome como medianas aquelas determinadas pela solução dual , isto é, o conjunto L ao qual estão associados os m menores valores de S(.), atribuindo-se os demais elementos à mediana mais próxima. Assim, a solução primal viável seria dada por:

$$y_{jj} = \begin{cases} 1 & \text{se } j \in L, \\ 0 & \text{em caso contrário.} \end{cases}$$

$$y_{ij} = \begin{cases} 1 & \text{se } d_{ij} = \min_{k \in L} \{d_{ik}\} , \text{ com } \sum_{j \in L} y_{ij} = 1 \\ 0 & \text{em caso contrário} \end{cases}$$

Uma complicação nesse processo pode ocorrer se dois elementos, por exemplo, e_p e e_q , estiverem muito próximos. Nesse caso, $S(p) \simeq S(q)$, e os dois elementos tenderão a ser escolhidos como medianas, pertencendo , assim ao conjunto L de medianas. Para resolver essa dificuldade, um dos elementos é então eliminado do conjunto L, por exemplo, substituindo-o pelo (m+1)-ésimo menor elemento de {S(.)}. Esse processo de identificação de vizinhança de elementos e de eliminação das variáveis é denominado por MULVEY e CROWDER[29] de "Pesquisa Primal".

Quanto ao ponto de vista de esforço computacional, MULVEY e CROWDER[29] indicam que cada iteração do método requer $n^2 + mn$ operações, cerca de metade, no caso de problemas maiores, das operações requeridas para se obter a solução por um método hierarquizado aglomerativo (cerca de $2n^2$ operações). Assim, de uma maneira geral, duas iterações do método de Mulvey e Crowder são equivalentes à resolução do problema por um método hierarquizado aglomerativo. Por outro lado, os resultados obtidos por MULVEY e CROWDER[29] indicam que o número total de iterações de seu método não parece depender criticamente dos valores de m e n.

EXEMPLO

Aplicando o método ao exemplo resolvido no item V.3 , utilizando-se a rotina CLSTROT apresentada no Apêndice 2, a convergência foi obtida em seis iterações. Os resultados estão apresentados na tabela a seguir, onde a solução primal de inicialização é a obtida pelo algoritmo K-Medianas:

TABELA VI.1 - Método de Mulvey e Crowder aplicado ao exemplo do item V.3

I T E R	F O P	F O S	F O D	ε (%)
1	-	21 , 000	3 , 000	6 , 000
2	49 , 000	21 , 000	18 , 500	0 , 135
3	49 , 000	21 , 000	20 , 008	0 , 049
4	49 , 000	21 , 000	20 , 544	0 , 022
5	49 , 000	21 , 000	20 , 763	0 , 011
6	49 , 000	21 , 000	20 , 869	0 , 006

onde

ITER - iteração;

FOP - função objetivo associada à solução primal gerada pela Pesquisa Primal;

FOS - função objetivo associada à melhor solução primal obtida até a iteração (na primeira iteração, FOS é o valor da função objetivo associada à solução obtida pelo método das K-Medianas) ;

FOD - função objetivo associada à solução dual;

ε (%) - (FOS - FOD) / FOD.

Note-se que a sequência de soluções geradas pela Pesquisa Primal não convergiu nem gerou uma solução melhor que a obtida pelo método das K-Medianas que, nesse caso, é ótima.

VI.5 - RELAÇÃO ENTRE O MÉTODO DE VINOD E O PROBLEMA DE LOCALIZAÇÃO NÃO CAPACITADO:

O problema de localização não capacitado pode ser descrito (ver p. ex. MATEUS e BORNSTEIN[28] ou EFFROYMSON e RAY[11]) da forma:

Problema L : Minimizar $$\sum_{i\epsilon I} \sum_{j\epsilon E} c_{ij}\, z_{ij} + \sum_{j\epsilon E} b_j\, y_j \qquad (L.1)$$

Sujeito a $$\sum_{j\epsilon E} z_{ij} = 1 \quad , \quad i\epsilon I \qquad (L.2)$$

$$1 \leq \sum_{j\epsilon E} y_j \leq N \qquad (L.3)$$

$$z_{ij} \leq y_j \quad , \quad i\epsilon I \ , \ j\epsilon E \qquad (L.4)$$

$$z_{ij} \ , \ y_j \ \epsilon \{0,1\} \ , \ i\epsilon I \ , \ j\epsilon E \qquad (L.5)$$

onde

I = Conjunto de m centros de demanda, a serem abastecidos por uma série de fornecedores, cuja localização e dimensionamento se deseja determinar;

E = Conjunto das n possíveis localizações desses fornecedores;

c_{ij} = custo associado ao fato do fornecedor j abastecer o centro de demanda i;

b_j = custo fixo de instalação do fornecedor j;

$$z_{ij} = \begin{cases} 1 & \text{se o centro i é abastecido pelo fornecedor j} \\ 0 & \text{em caso contrário} \end{cases}$$

$$y_j = \begin{cases} 1 & \text{se o fornecedor j for ativado} \\ 0 & \text{em caso contrário} \end{cases}$$

Então, se d_i é a demanda do centro i, a dimensão do fornecedor j será dada por $\sum_{i\epsilon I} z_{ij}\, d_i$.

A semelhança entre os Problemas V1 (modelo de Vinod) e L (problema de localização não capacitado) é evidente. O Problema V1 poderia então ser visto como um caso particular do Problema L, onde b_j=0 (ou seja , os custos de implantação não são levados em consideração), E é subconjunto de I e o número de fornecedores é fixo e conhecido a priori.

Por outro lado, existe também um outro caso particular do problema L que seria interessante analisar: o problema das N-medianas (v.p. ex. MATEUS e BORNSTEIN[28] ou JARVINEN, RAJALA e SINERVO[20]). Esse problema consiste em localizar N facilidades (armazéns, fornecedores, etc.) num grafo ou rede, de forma a minimizar a soma das distâncias entre cada nó da rede e a facilidade mais próxima. Como indicam MATEUS e BORNSTEIN[28], pode-se demonstrar que a solução ótima desse problema implica em localizar as facilidades sempre em nós do grafo, de maneira que o conjunto desses nós representa o conjunto de locais candidatos para a localização das facilidades.

O problema das N-medianas pode então ser descrito a partir do problema L, tomando I=E, eliminando os custos fixos em (L.1), tomando como c_{ij} a distância d_{ij} e transformando (L.3) em igualdade:

$$\sum_{j\varepsilon I} y_j = N \ .$$

Como se nota, essa formulação do problema das N-medianas é equivalente à expressão do primeiro modelo de Vinod.

MATEUS e BORNSTEIN[28] apresentam um algoritmo guloso para a resolução do Problema L que, embora nem sempre determine uma solução ótima, permite, no entanto, tal como o método de Mulvey e Crowder, determinar intervalos para essa solução(*). Mulvey e Crowder[29], por sua vez, apontam a necessidade de se desenvolverem estudos comparativos da eficiência dessas duas técnicas na resolução de problemas nas duas áreas.

VI.6 - UM MODELO EFICIENTE DE PROGRAMAÇÃO DINÂMICA PARA O CASO UNIVARIADO - O MODELO DE RAO

VI.6.1 - INTRODUÇÃO

Na seção VI.3 foi apresentado o Segundo Modelo de Vinod - caso univariado. O modelo ali descrito, no entanto, é um modelo de programação linear inteira e, como foi apontado na introdução deste Capítulo VI, modelos de programação inteira geralmente não são eficientes, do ponto de vista do esforço computacional envolvido para a obtenção de uma solução ótima.

(*) (para uma descrição de algoritmos gulosos que resolvem essa classe de problemas, ver também FISHER, NEMHAUSER e WOLSEY[47 e 48])

Por outro lado, e baseando-se em RAO[33], é possível formular-se um modelo de programação dinâmica extremamente eficiente, para a obtenção de uma solução ótima para o citado Segundo Modelo de Vinod - caso univariado.

Seja, por exemplo, o conjunto X = {2,6,4,5,2} , onde se deseja particionar o conjunto E, associado a X, em três grupos, minimizando-se a soma dos quadrados dentro dos grupos. Para que se utilize a propriedade da cadeia definida para o Segundo Modelo de Vinod - caso univariado, forma-se o conjunto Z, que se constitui numa ordenação de X em ordem crescente (ver item VI.3):

$$Z = \{2,2,4,5,6\}$$

definindo-se assim o conjunto $E' = \{e'_1, \ldots, e'_n\}$, associado a Z, correspondente à ordenação efetuada.

Lembrando que um líder de grupo é o primeiro elemento de um grupo (ver item VI.3) e que a determinação desses líderes de grupo implica na definição dos próprios grupos, uma vez que a propriedade da cadeia impõe que a solução ótima do problema de formação de m grupos corresponde a m partições da forma:

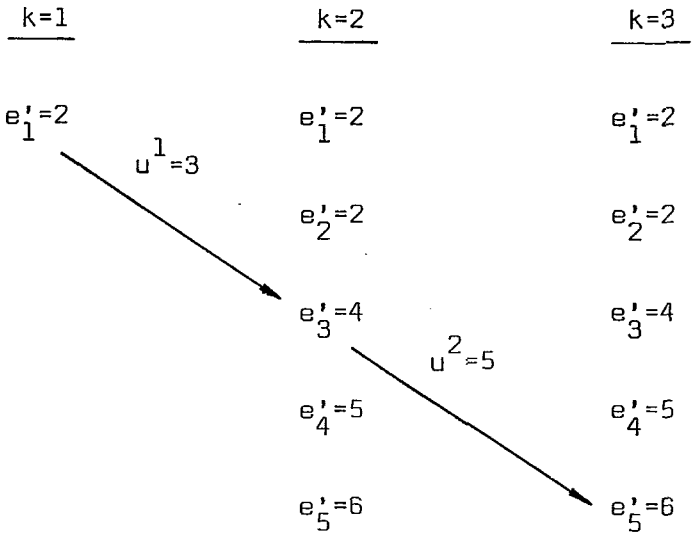
$$\{e'_1, e'_2, \ldots\} \ldots \{e'_j, e'_{j+1}, \ldots, e'_i\}\{e'_{i+1}, \ldots\} \ldots \{\ldots, e'_n\}$$

pode-se então definir, para o exemplo em questão:

- estágio k : etapa de definição do k-ésimo grupo.
- estado y^k : índice associado ao líder do grupo g_k (k=1,...,m)- se e'_j é líder do grupo g_k, então $y^k = j$; e
- decisão u^k: índice associado ao líder do grupo g_{k+1} (k=1, ..., m-1) - note-se que as decisões só são tomadas até k = m-1 , uma vez que em k = m não faz sentido calcular-se o líder do grupo seguinte, que seria g_{m+1} (o objetivo é formar m grupos).

Assim, como no estágio k define-se o líder do grupo g_k (dado pelo estado y^k) e o líder do grupo g_{k+1} (dado pela decisão u^k), ficam, nesse estágio, perfeitamente definidos todos os elementos que compõem g_k.

O problema pode ser ilustrado graficamente da forma:



onde as setas indicam as decisões tomadas, ou seja, os índices dos líderes dos grupos formados. Nota-se então que:

(a) no estágio k=1, formou-se o grupo $g_1 = \{ e'_1, e'_2 \}$, ao se definir que g_2 teria como líder e'_3 (pois $u^1 = 3$). Como o líder de g_1 só pode ser e'_1, só existe um estado viável no estágio k=1 : $y^1 = 1$;

(b) no segundo estágio $y^2 = u^1$. De uma maneira geral, tem-se:

$$y^{k+1} = u^k \quad ; \qquad \text{(VI.5.1)}$$

(c) pelas decisões indicadas na figura, os grupos formados seriam $\{e'_1, e'_2\}$, $\{e'_3, e'_4\}$, $\{e'_5\}$;

(d) ao se permitir que y^2 e y^3 fossem iguais a 1, permitiu-se que a solução do problema através da Programação Dinâmica fornecesse também a partição de E' em 2 e 1 grupos, respectivamente. Por exemplo, se $y^2 = 1$ e $u^2 = 3$, tem-se os grupos $\{e'_1, e'_2\}$, $\{e'_3, e'_4, e'_5\}$. Assim, a definição, através do cálculo , no sentido inverso, das decisões ótimas, em cada estágio k (k=1, ..., m-1), para todos os estados $y^k = 1$, permite a determinação de todas as partições ótimas de 1 a m grupos;

(e) a rigor, não faz sentido permitir-se que $y^2 = 5$, uma vez que, neste caso, u^2 não pode ser definido: como $y^{k+1} = u^k$, não existe, para $y^2 = 5$, valor de u^k que seja admissível. Verifi

ca-se então que, no penúltimo estágio y^{m-1} deve ser inferior a n : $1 \leq y^{m-1} \leq n-1$. No antepenúltimo estágio, $1 \leq y^{m-2} \leq n-2$. De uma maneira geral, $1 \leq y^{m-i} \leq n-i$. Caso $i = m-k$, tem-se finalmente,

$$1 \leq y^{k} \leq n-(m-k) \; ; \qquad \text{(VI.5.2)}$$

(f) denotando por $J(y^k, k)$ o custo acumulado da formação dos grupos $\{e'_{y^k}, \ldots\}\{e'_{u^k}, \ldots\}, \ldots$, ou seja,

$$J(y^k, k) = \sum_{j=k}^{m} W_j \; ,$$

onde W_j é a soma dos quadrados dentro do grupo g_j, o problema de programação dinâmica seria resolvido no sentido inverso através da seguinte equação recursiva de otimalidade:

$$J(y^k, k) = \min_{u^k \text{ viável}} \{W_k + J(u^k, k+1)\}, \qquad \text{(VI.5.3)}$$

para $k = 1, \ldots, (m-1)$;

(g) para k=m, ou seja, no último estágio, o valor de $J(y^m, m)$, para cada y^m, corresponde à soma dos quadrados dentro do grupo g_m que tem y^m como líder. Assim, para $y^m = n$, tem-se

$$J(y^n, m) = J(n, m) = W_m = 0 \; ,$$

pois W_m, no caso, corresponde à soma dos quadrados dentro de um grupo constituído por um único elemento = e'_n. Para $y^m = n-1$, tem-se o grupo $g_m = \{e'_{n-1}, e'_n\}$. A soma dos quadrados dentro desse grupo é (ver equação (VI.3.2) da seção VI.3.1) :

$$J(y^m, m) = J(n-1, m) = W_m = \frac{1}{2}(z_{n-1} - z_n)^2$$

Em geral, para $y^m = i$, tem-se a formação do grupo

$$g_m = \{e'_i, e'_{i+1}, \ldots, e'_n\}.$$

Na seção VI.3.1, foi visto que a soma dos quadrados dentro desse grupo g_m poderia ser efetuada recursivamente. Naquela seção, verificou-se que a soma dos quadrados dentro do grupo $g_k = \{e'_j, \ldots, e'_i\}$ é dada por [ver equações (VI.3.3) e (VI.3.4)] :

$$W_k = \sum_{r=j+1}^{i} \Delta W_j^r = \sum_{r=j+1}^{i} \frac{r-j}{(r-j)+1} \left(z_r - \frac{1}{(r-j)} \sum_{q=j}^{r-1} z_q\right)^2 \qquad \text{(VI.5.4)}$$

Utilizando, a exemplo do efetuado naquela seção VI.3.1, o teorema (IV.1) da seção IV.2.5, que informa que, ao se unirem dois grupos g_I e g_J para se formar o grupo g_L, tem-se

$$W_L = W_I + W_J + \frac{n_I n_J}{n_I + n_J} d_{IJ}^2 \quad , \qquad \text{(VI.5.5)}$$

onde d_{IJ}^2 corresponde ao quadrado da distância euclidiana entre as centróides (médias) de g_I e g_J, tem-se, ao incluir o elemento e'_r no grupo $\{e'_{r+1}, \ldots, e'_n\}$, que:

$$g_I = \{e'_r\} \quad ; \quad n_I = 1 \quad ; \quad W_I = 0$$

$$g_J = \{e'_{r+1}, \ldots, e'_n\}; \; n_J = n-r \; ; \; W_J = \sum_{e'_j \in g_J} (z_j - \bar{z}_J)^2$$

e, finalmente, pela equação (VI.5.5)

$$W_L = W_J + \delta W_n^r = W_J + \frac{n-r}{1+(n-r)} \left(z_r - \frac{1}{n-r} \sum_{q=r+1}^{n} z_q\right)^2 \qquad \text{(VI.5.6)}$$

Assim, δW_n^r representa o acréscimo na soma dos quadrados dentro dos grupos devido à inclusão de e'_r no grupo $\{e'_{r+1}, \ldots, e'_n\}$. De uma maneira geral, pode-se calcular, recursivamente, a soma dos quadrados dentro do grupo $g_m = \{e'_i, \ldots, e'_n\}$ fazendo r variar, retroativamente, de (n-1) até (i), tendo-se, finalmente, então:

$$W_m = J(y^m, n) = J(i, m) = \sum_{r=i}^{n-1} \delta W_n^r \qquad \text{(VI.5.7)}$$

Note-se que a diferença entre as equações (VI.5.4) e (VI.5.7) reside apenas em que em ΔW_j^r o acréscimo do elemento é feito "à direita" , ou seja, considera-se que $g_k = \{e'_j,\ldots,e'_{i-1}\} \cup \{e'_i\}$ e, em δW_n^r , o acréscimo é feito "à esquerda" :

$$g_m = \{e'_i\} \cup \{e'_{i+1}, \ldots, e'_n\} .$$

Assim, no exemplo em questão, ter-se-ia (ver equações (VI. 5 .5) e (VI.5.6) :

$$g_3 = \{e'_5\} \Rightarrow J(5, 3) = 0$$

$$g_3 = \{e'_4, e'_5\} \Rightarrow J(4, 3) = \frac{5-4}{1+(5-4)}(5-6)^2 + J(5, 3) = 0{,}50$$

$$g_3 = \{e'_3, e'_4, e'_5\} \Rightarrow J(3, 3) = \frac{5-3}{1+(5-3)}\{4-\frac{1}{5-3}(5+6)\}^2 + J(4, 3)$$

$$= 1{,}5 + 0{,}5 = 2{,}00$$

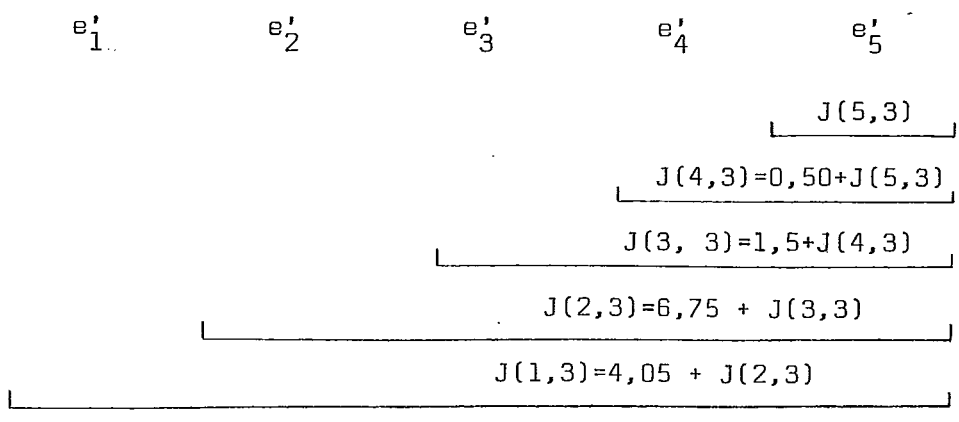
$$g_3 = \{e'_2, \ldots, e'_5\} \Rightarrow J(2,3) = \frac{5-2}{1+(5-2)}\{2-\frac{1}{5-2}(4+5+6)\}^2 + J(3, 3)$$

$$= 6{,}75 + 2{,}00 = 8{,}75$$

$$g_3 = \{e'_1, \ldots, e'_5\} \Rightarrow J(1,3) = \frac{5-1}{1+(5-1)}\{2-\frac{1}{5-1}(2+4+5+6)\}^2 + J(2,3)$$

$$= 4{,}05 + 8{,}75 = 12{,}80.$$

Graficamente, tem-se

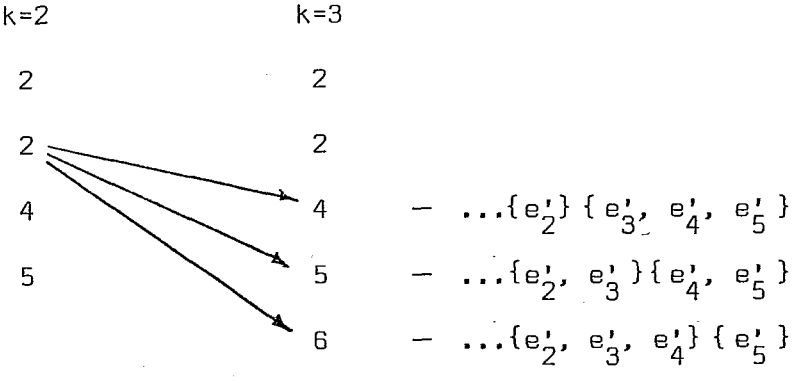


Assim, o cálculo dos $J(y^k, m)$ sempre é feito a partir dos resultados da etapa anterior: J(3,3), por exemplo, é calculado a partir de J(4,3). Além disso, tudo se passa, no último estágio, como se fosse calculado, recursivamente, o custo da formação de um único grupo $\{e'_1, \ldots, e'_n\}$.

(h) calculados os $J(y^k, m)$, a determinação das decisões ótimas em cada estado dos demais estágios pode ser efetuada a partir da equação (VI.5.3):

$$J(y^k, k) = \min_{u^k \text{ viável}} \{W_k + J(u^k, k+1)\} .$$

Seja, por exemplo, k=2 e $y^2 = 2$. São possíveis, nesse caso três alternativas de grupamento



Os cálculos dos custos associados aos possíveis grupos g_3:

$$g_3 = \{e'_3, e'_4, e'_5\}$$

$$g_3 = \{e'_4, e'_5\}$$

$$g_3 = \{e'_5\}$$

foram efetuados no item (g) anterior e correspondem aos valores de J(3,3)=2,00 , J(4,3)=0,50 e J(5,3)=0 , respectivamente. Por outro lado, para cada valor admissível de u^2, ter-se-ia:

$$u^2 = 3 \Rightarrow g_2 = \{e'_2\}$$

$$u^2 = 4 \Rightarrow g_2 = \{e'_2, e'_3\}$$

$$u^2 = 5 \Rightarrow g_2 = \{ e'_2,\ e'_3,\ e'_4 \}$$

Note-se que a sequência de valores admissíveis de u^2 implica no acréscimo de um elemento "à direita" e assim, pela equação (VI.5.4) do item (g) anterior, tem-se:

$$W_2 = \sum_{r=3}^{u^2-1} \Delta W_2^r = \sum_{r=3}^{u^2-1} \frac{r-2}{(r-2)+1} \left(z_r - \frac{1}{r-2} \sum_{q=2}^{r-1} z_q \right)^2$$

Voltando ao exemplo, tem-se:

$$u^2 = 3 \Rightarrow g_2 = \{ e'_2 \} \Rightarrow W_2 = 0{,}00$$

$$u^2 = 4 \Rightarrow g_2 = \{ e'_2,\ e'_3 \} \Rightarrow W_2 = \Delta W_2^3 = 2{,}00$$

$$u^2 = 5 \Rightarrow g_2 = \{ e'_2,\ e'_3,\ e'_4 \} \Rightarrow W_2 = \Delta W_2^4 + \Delta W_2^3 = 2{,}67 + 2{,}00 = 4{,}67$$

Para o cálculo da decisão ótima tem-se (ver equação (VI.5.3):

$$u^2 = 3 \Rightarrow J(y^2,2) = W_2 + J(3,3) = 0{,}00 + 2{,}00 = 2{,}00$$

$$u^2 = 4 \Rightarrow J(y^2,2) = W_2 + J(4,3) = 2{,}00 + 0{,}50 = 2{,}50$$

$$u^2 = 5 \Rightarrow J(y^2,2) = W_2 + J(5,3) = 4{,}67 + 0{,}00 = 4{,}67$$

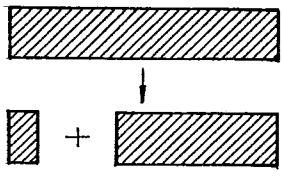
Nota-se que o mínimo de $J(y^2,2)$ ocorre para $u^2 = 3$, que é assim a decisão ótima para $k=2$ e $y^2 = 2$. Como se verifica, o cálculo do custo da formação do grupo g_2 para $y^2 = 2$ envolve apenas o cálculo recursivo do custo da formação de um único grupo $\{ e'_2,\ e'_3,\ e'_4 \}$. Evidentemente esse procedimento de cálculo da decisão ótima pode ser generalizado para qualquer estágio k ($k=1, \ldots, m-1$) e qualquer estado y^k viável.

(i) Assim, em resumo, nota-se pelo exposto no item (g), que é mais eficiente calcular-se a soma dos quadrados dentro dos grupos, no último estágio ($k=m$), através de inclusões "à esquerda", isto é, através da equação (VI.5.7), o que corresponderia ao cálculo do custo de formação de apenas um grupo. Nos outros estágios ($k=1, \ldots, m-1$), como se viu no item(h), é mais conveniente considerar-se a inclusão de elementos "à direita", calculando-se a soma dos quadrados dentro dos grupos através da equação (VI.5.4), o que corresponde a que ,

em cada estado y^K dos estágios k(k=1, ..., m-1) se calcule o custo de formação de apenas um grupo.

Graficamente, os procedimentos de cálculo da soma dos quadrados dentro dos grupos poderiam ser ilustrados da forma:

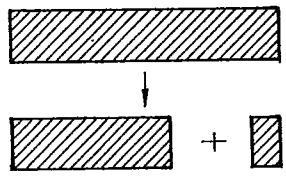
- inclusão "à esquerda" (k=m)



equação (VI.5.7):

$$W_m = \sum_{r=1}^{n-1} \delta W_n^r$$

inclusão "à direita" (k=1, ..., m-1)



equação (VI.5.4)

$$W_k = \sum_{r=j+1}^{i} \Delta W_j^r$$

k=1, ..., m-1

VI.6.2 - DESCRIÇÃO DO MODELO

As observações contidas na seção anterior podem ser resumidas no seguinte modelo de programação dinâmica:

Estágio k - etapa de definição do k-ésimo grupo;

Estágio y^k - índice associado ao líder do grupo g_k, k=1,...,m ;

Decisão u^k - índice associado ao líder do grupo g_{k+1}, k=1, ... , (m-1);

Estados viáveis Y(k) - conjunto de índices associados aos possíveis líderes do grupo g_k. Como visto no item (e) da seção anterior,

$$1 \leq y^k \leq n-(m-k) ;$$

Estado inicial y^1 - de acordo com o item (a) anterior,

$$y^1 = 1 ;$$

Equação de transição de estado - pelo item (b) anterior,

$$y^{k+1} = u^k \ ;$$

Decisões admissíveis - $U(y^k, k) = \{u^k | u^k \in \{y^k+1, \ldots, n-(m-k-1)\}\}$

(ver definição de estados viáveis: como $y^{k+1} = u^k$, é necessário que y^{k+1} seja viável).

Equação recursiva de otimalidade : (ver itens (f) a (h) anteriores):

$$J(y^k, k) = \min_{u^k \text{ viável}} \{ W_k + J(u^k, k+1) \} ,$$

com $J(y^k, m)$ calculado na forma descrita no item (g).

VI.6.3 - EFICIÊNCIA DO MODELO EM RELAÇÃO À ENUMERAÇÃO COMPLETA

Para k=m, como visto nos itens (g) e (i) da seção VI.5.1, o cálculo de $J(y^m, m)$, para todo y^m, equivale ao cálculo do custo da formação de apenas um grupo: $\{ e'_1, \ldots, e'_n \}$.

Para $1 \leq k \leq m-1$, por outro lado, tem-se:

- pela equação (VI.5.2), $1 \leq y^k \leq n-(m-k)$;
- o cálculo da decisão ótima em y^k envolve o cálculo da formação de apenas um grupo (ver itens (h) e (i) da seção (VI.5.1);
- em k=1 tem-se apenas um estado viável: $y^1 = 1$ (ver item (a) da seção (VI.5.1); e
- nesse caso ($1 \leq k \leq m-1$), o número de cálculos de formação de grupos é dado pelo número total de estados admissíveis: $1 + \sum_{k=2}^{m-1} \{n-(m-k)\}$

Assim, o número total de cálculos de custo de formação de grupos, N_R, é dado por

$$N_R = 2 + \sum_{k=2}^{m-1} \{ n-(m-k)\}.$$

Por exemplo, se n=7 e m=3 , o número de cálculos de formação de grupos seria:

$$N_R = 2 + \{7-(3-1)\} = 8 .$$

No caso da enumeração completa de todas as alternativas de grupamento, ter-se-ia:

- o número de alternativas de grupamento é dado pela seguinte equação (ver item III.4):

$$S(n, m) = \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} k^n .$$

  Assim, S(7, 3) = 301.

- como em cada alternativa de grupamento se efetua o cálculo do custo de formação de três grupos, tem-se

$$N_{EC} = 3 \times 301 = 903.$$

VI.6.4 - CONCLUSÃO

Apesar de sua extrema eficiência, o método é de reduzida aplicação, pois só pode ser utilizado no caso univariado ou em casos em que as p variáveis originais possam ser reduzidas a um único fator, através do uso de alguma técnica de análise fatorial (para uma apresentação detalhada das principais técnicas ver, p.ex., HARMAN[16]).

Recentemente, PFEIFFER[31] apresentou um exemplo prático em que tal fenômeno ocorreu. Analisando as disparidades de desenvolvimento no Brasil, aquele autor, através de análise de grupamento, criou uma tipologia das unidades da Federação, em quatro grupos, a partir das seguintes características:

- capacidade de produção econômica;
- assistência médica;
- educação;
- habitação; e
- alimentação.

Cada característica foi avaliada por uma variável. Assim, a capacidade de produção econômica foi descrita pela renda per capita em 1970 ,

a assistência médica pelo número de médicos por 100.000 habitantes em 1975 é assim por diante. Maiores detalhes sobre a resolução do problema podem ser obtidos em PFEIFFER[31]. O que cabe ressaltar aqui é que a matriz dos coeficientes de correlação entre as cinco variáveis escolhidas indicou uma elevadíssima correlação entre todas elas, o que permitiria selecionar - se uma "variável mediana" do grupo de variáveis (o equivalente, para variável, ao "elemento mediano", utilizado na dispersão via mediana de grupo , definida no item II.4.4), para a qual poderia ser aplicado o método de RAO. Apenas como informação vale registrar que, nesse caso, a variável mediana seria aquela associada à característica habitação.

Um outro problema prático em que esse método se mostra bastante útil é o caso da amostragem estratificada com uma única variável de estratificação (v.p.ex. RAJ[32] ou SUDMAN[38]). Nesse problema, procura-se particionar uma população a ser amostrada em grupos, dentro dos quais a variância da variável de estratificação é mínima. Para que a estratificação seja efetuada, com a minimização da variância ao invés da soma dos quadrados dentro dos grupos, basta que sejam implantadas no algoritmo as modificações convenientes lembrando que:

$$S_I^2 = \frac{1}{n_I} W_I \ .$$

VI. 7 - CONCLUSÃO

Como foi visto, os modelos de Vinod, embora não permitam uma solução do problema de análise de grupamento de forma eficiente, estabelecem uma base na qual podem se apoiar métodos mais eficazes. Permanece, no entanto, pelo menos na literatura consultada pelo autor, um tópico ainda em aberto nessa área: a possibilidade de implementação, por exemplo, de um método subgradiente para outros modelos devidos a VINOD[41] e a RAO[33], que permitissem uma solução eficiente para o problema de análise de grupamento minimizando a soma dos quadrados dentro dos grupos ou a variância interna.

Por outro lado, fica evidente que a obtenção de uma solução ótima para o problema é feita a um preço bem mais elevado do que o das técnicas heurísticas. Assim, a utilização desses métodos exige do usuário uma análise cuidadosa, não só do porte do problema a ser resolvido mas também da necessidade efetiva de obtenção de uma solução ótima.

VII - APLICAÇÃO DE ANÁLISE DE GRUPAMENTO À DETERMINAÇÃO DE ETAPAS DE CRESCIMENTO DE LARVAS

VII.1 - DESCRIÇÃO DO PROBLEMA

Durante vinte e cinco dias foram colocadas, sucessivamente, em cada dia, oito larvas "recém-nascidas" em vinte e cinco vidros. No final do vigésimo sexto dia recolheu-se os vidros e efetuou-se nas larvas medições das variáveis "cápsula cefálica", "comprimento" e "peso". Obteve-se assim a Tabela VII.1 que, como se pode observar, não apresenta resultados para cinco e quatro das larvas com um e dois dias de vida, o que reduz a amostra de 25 x 8 = 200 para 200 - 9 = 191 larvas. Cada larva na tabela é caracterizada pelo código XXDY, onde XX representa o dia de vida da larva e Y o número de ordem da larva no dia. Assim, 12D4 representa a quarta larva dentre as que possuiam doze dias de vida.

Sabe-se que as lagartas crescem segundo diversas etapas de crescimento, denominadas "ínstares". Depois de cada etapa elas mudam de pele, havendo de uma fase para outra um salto no que diz respeito principalmente à cápsula cefálica, que permanece constante em cada instar.

O objetivo do estudo é determinar a quantos instares correspondem as medidas tomadas, qual a duração média de cada instar e quais são as características (média, variância, etc.) de cada instar no que diz respeito ao comprimento, ao peso e ao diâmetro da cápsula cefálica. A utilização, nesse caso, da análise de grupamento se deve ao fato de que a observação, em intervalos de tempo predeterminados, de uma mesma lagarta, é infrutífera (ela come a própria pele) e sua medição constante (que exige a administração à larva de certas substâncias) altera sensivelmente o seu crescimento.

Acredita-se que existam cinco ou seis ínstares e que o aumento da cápsula cefálica deve seguir a "Regra de Dyat", pela qual a razão entre os diâmetros da cápsula cefálica nos instares (i+1) e (i) é constante para todo (i):

$$\frac{CC_i}{CC_{i-1}} = k, \quad \forall\, i \geq 2 \qquad \text{(VII.1)}$$

Um outro objetivo correlato ao estudo é o de, ao se tomar uma lagarta ao acaso no campo, determinar em que instar ela se encontra.

TABELA VII .1 - Medidas efetuadas na amostra de larvas

Larva	Peso (mg)	Comprimento (mm)	Medida da Cáp. Cef.
01D1	0,055	1,5	0,236
01D2	0,055	1,5	0,187
01D3	0,055	1,5	0,258
02D1	0,075	1,5	0,308
02D2	0,075	1,5	0,221
02D3	0,075	1,5	0,260
02D4	0,075	1,5	0,221
03D1	0,437	3,0	0,332
03D2	0,437	3,5	0,405
03D3	0,437	2,5	0,353
03D4	0,437	3,0	0,399
03D5	0,437	2,0	0,388
03D6	0,437	2,5	0,285
03D7	0,437	3,0	0,353
03D8	0,437	2,5	0,357
04D1	0,625	2,5	0,395
04D2	0,625	3,0	0,370
04D3	0,625	4,0	0,397
04D4	0,625	3,5	0,382
04D5	0,625	3,0	0,397
04D6	0,625	3,0	0,585
04D7	0,625	4,0	0,568
04D8	0,625	3,5	0,355
05D1	1,250	5,0	0,444
05D2	1,250	5,0	0,421
05D3	1,250	4,5	0,438
05D4	1,250	5,5	0,425
05D5	1,250	5,0	0,448
05D6	1,250	5,5	0,713
05D7	1,250	5,0	0,661
05D8	1,250	5,5	0,609

TABELA VII.1 - Medidas efetuadas na amostra de larvas (cont.)

Larva	Peso (mg)	Comprimento (mm)	Medida da Cáp. cef.
06D1	1,25	4,0	0,615
06D2	1,25	5,0	0,450
06D3	1,25	5,0	0,467
06D4	1,25	4,0	0,434
06D5	1,25	4,0	0,676
06D6	1,25	4,0	0,472
06D7	1,25	4,5	0,626
06D8	1,25	4,5	0,456
07D1	7,00	7,0	0,862
07D2	10,00	9,0	0,917
07D3	5,00	6,5	0,967
07D4	10,00	8,0	0,923
07D5	5,00	6,0	0,873
07D6	4,00	7,0	0,917
07D7	8,00	8,0	0,846
07D8	6,00	6,5	0,840
08D1	18,00	11,0	1,203
08D2	15,00	10,0	1,126
08D3	14,00	9,5	1,055
08D4	18,00	10,0	1,401
08D5	17,00	11,0	1,038
08D6	14,00	9,0	1,121
08D7	20,00	11,5	1,368
08D8	17,00	10,0	1,154
09D1	21,00	12,0	1,176
09D2	20,00	11,0	1,236
09D3	27,00	13,0	1,351
09D4	23,00	11,0	1,055
09D5	16,00	10,5	0,983
09D6	14,00	8,5	1,104
09D7	20,00	12,0	1,275
09D8	16,00	10,0	0,956

TABELA VII.1 - Medidas efetuadas na amostra de larvas (cont.)

Larva	Peso (mg)	Comprimento (mm)	Medida da Cáp. Cef.
10D1	21	11,5	1,357
10D2	22	11,5	1,214
10D3	17	10,0	1,049
10D4	23	11,0	1,143
10D5	20	11,0	1,236
10D6	32	13,5	1,373
10D7	25	11,5	1,439
10D8	23	11,0	1,291
11D1	30	13,5	1,214
11D2	34	14,5	1,302
11D3	34	13,5	1,346
11D4	32	13,0	1,242
11D5	32	13,0	1,225
11D6	21	10,0	1,488
11D7	21	11,5	1,055
11D8	16	8,5	1,351
12D1	22	12,0	1,203
12D2	28	13,5	1,197
12D3	30	13,0	1,280
12D4	30	13,0	1,340
12D5	25	12,0	1,197
12D6	24	14,5	1,477
12D7	23	12,0	1,142
12D8	30	11,5	1,577
13D1	47	13,5	1,159
13D2	50	15,0	1,395
13D3	44	13,0	1,313
13D4	26	12,0	1,049
13D5	25	10,0	1,494
13D6	26	11,0	1,428
13D7	23	11,5	1,296
13D8	25	9,5	1,428

TABELA VII.1 - Medidas efetuadas na amostra de larvas (cont.)

Larva	Peso (mg)	Comprimento (mm)	Medida da Cáp. Cef.
14D1	38	14,0	1,357
14D2	55	16,5	1,373
14D3	23	11,0	1,115
14D4	24	11,0	1,230
14D5	30	12,0	1,511
14D6	33	12,5	1,219
14D7	38	15,0	1,516
14D8	30	13,5	1,230
15D1	38	13,5	1,472
15D2	30	13,5	1,099
15D3	26	13,0	1,099
15D4	33	13,5	1,275
15D5	38	15,0	1,318
15D6	47	15,5	1,296
15D7	45	15,0	1,395
15D8	28	12,5	1,181
16D1	69	19,5	1,813
16D2	59	17,0	1,280
16D3	100	21,0	1,785
16D4	58	17,0	1,522
16D5	85	18,5	1,533
16D6	80	19,0	1,648
16D7	80	17,0	1,550
16D8	76	16,0	1,752
17D1	46	15,5	1,302
17D2	45	15,5	1,648
17D3	50	15,0	1,411
17D4	56	17,0	1,466
17D5	40	13,5	1,373
17D6	52	16,0	1,373
17D7	45	13,0	1,648
17D8	47	13,5	1,648

TABELA VII.1 - Medidas efetuadas na amostra de larvas (cont.)

Larva	Peso (mg)	Comprimento (mm)	Medida da Cáp. Cef.
18D1	77	18,0	1,999
18D2	77	19,5	1,835
18D3	68	15,0	1,714
18D4	65	16,0	1,648
18D5	56	18,0	1,346
18D6	53	17,0	1,488
18D7	60	16,0	1,774
18D8	63	17,0	1,373
19D1	90	20,0	1,604
19D2	68	18,0	1,494
19D3	60	17,0	1,648
19D4	79	19,0	1,846
19D5	70	17,0	1,488
19D6	45	13,0	1,648
19D7	75	18,0	1,873
19D8	73	17,5	1,785
20D1	121	20,0	1,648
20D2	90	17,5	1,901
20D3	133	22,0	1,730
20D4	98	21,5	1,681
20D5	110	20,0	1,785
20D6	92	19,0	1,769
20D7	145	24,0	1,961
20D8	101	19,0	1,648
21D1	122	20,5	1,697
21D2	95	19,5	1,565
21D3	117	20,5	1,769
21D4	82	18,0	1,752
21D5	103	20,5	1,648
21D6	88	19,0	1,593
21D7	125	22,0	1,719
21D8	102	21,0	1,730

TABELA VII.1 - Medidas efetuadas na amostra de larvas (cont.)

Larva	Peso (mg)	Comprimento (mm)	Medida da Cáp. Cef.
22D1	75	18,0	1,648
22D2	63	17,0	1,587
22D3	103	21,0	1,785
22D4	40	12,0	1,439
22D5	58	17,0	1,379
22D6	67	15,5	1,785
22D7	71	20,0	1,472
22D8	85	17,5	1,648
23D1	61	18,0	1,565
23D2	113	21,0	1,648
23D3	78	17,0	1,988
23D4	90	19,0	1,785
23D5	125	20,0	1,944
23D6	98	19,5	1,670
23D7	90	18,5	1,648
23D8	75	17,0	1,544
24D1	113	21,0	1,829
24D2	118	18,0	1,598
24D3	152	24,0	1,917
24D4	83	17,0	1,648
24D5	110	19,0	1,824
24D6	85	16,0	2,236
24D7	138	24,0	1,593
24D8	125	20,0	1,648
25D1	130	20,0	1,824
25D2	113	18,0	1,708
25D3	148	20,0	1,813
25D4	101	17,0	1,862
25D5	100	17,0	1,818
25D6	82	18,0	1,554
25D7	92	17,5	1,648
25D8	128	19,5	1,862

VII.2 - METODOLOGIA UTILIZADA

Tendo em vista os objetivos definidos no item anterior, a metodologia adotada para a resolução do problema foi a de, via análise de agrupamento, tentar criar uma tipologia de larvas, em cinco ou seis grupos, que satisfaçam à Regra de Dyat.

Assim, o problema foi resolvido para m = 4,5,6 e 7, com o objetivo de se adotar, dentre as soluções para m = 5 ou 6, aquela que atende às condições da citada regra: pequena variância, em cada grupo, na cápsula cefálica e satisfação da equação (VII.1). O problema também foi resolvido para m = 4 e 7 apenas visando uma confirmação de que nesse caso os resultados seriam incompatíveis com as hipóteses adotadas.

Para a verificação da condição imposta pela equação (VII.1), foi utilizado um modelo de regressão linear simples da forma:

$$\hat{\hat{CC}}_i = k\ CC_{i-1} \qquad i=2, \ldots, m ,$$

sendo utilizado como medida da qualidade do ajustamento o coeficiente de determinação R^2, que é uma medida do poder "explanatório" da regressão (ver, p.ex., WESOLOWSKY[43]):

$$R^2 = 1 - \frac{(CC_i - \hat{\hat{CC}}_i)^2}{(CC_i - \overline{CC})^2} , \qquad \text{(VII.2)}$$

onde CC_i corresponde ao valor médio da cápsula cefálica no grupo i(i = 2, ... , m) , $\hat{\hat{CC}}_i$ é o valor estimado pela equação de regressão, (i=2,...,m) e $\overline{CC}$ corresponde ao valor médio de CC_i (i=2,..., m).

Pela equação (VII.2) se nota que, quanto mais próximo for R^2 de 1, melhor a qualidade do ajustamento.

VII.3 - RESOLUÇÃO DO PROBLEMA

VII.3.1 - AMOSTRA REDUZIDA

Tendo em vista a utilização do método de Mulvey e Crowder e a disponibilidade de tempo de processamento, procurou-se diminuir o porte do problema através de uma redução inicial da amostra de larvas para 75 elementos, que correspondiam às 3 primeiras larvas de cada dia.

Tomando, por outro lado, como variáveis,

TABELA VII.2 - Características dos grupamentos efetuados na amostra de 75 larvas; tomando como características para o agrupamento o peso, o comprimento e a cápsula cefálica (dados padronizados)

Número de Grupos	Grupos	Peso (mg)		Comp. (mm)		Medida Cáp. Cef.	
		M	DP	M	DP	M	DP
4	1	1,15	1,75	3,47	1,75	0,42	0,20
	2	18,45	4,08	10,77	1,01	1,15	0,12
	3	36,63	13,70	13,97	0,91	1,33	0,17
	4	96,30	28,59	19,24	1,84	1,71	0,17
5	1	1,15	1,75	3,47	1,75	0,42	0,20
	2	18,60	4,27	10,85	1,03	1,16	0,12
	3	35,63	14,50	13,76	1,32	1,31	0,18
	4	73,43	22,27	17,36	1,34	1,64	0,23
	5	114,71	22,55	20,71	1,17	1,74	0,10
6	1	0,27	0,24	2,14	0,78	0,30	0,07
	2	2,24	2,20	5,11	1,02	0,56	0,21
	3	18,45	4,08	10,77	1,01	1,15	0,12
	4	36,63	13,70	13,97	0,91	1,33	0,17
	5	76,42	20,84	17,50	0,56	1,65	0,22
	6	112,20	23,82	20,63	1,17	1,75	0,10
7	1	0,27	0,24	2,14	0,78	0,30	0,07
	2	2,24	2,20	5,11	1,02	0,56	0,21
	3	18,60	4,27	10,85	1,03	1,16	0,18
	4	35,54	14,07	13,75	1,29	1,32	0,18
	5	76,42	20,84	17,50	0,56	1,64	0,22
	6	109,36	21,91	20,39	0,74	1,74	0,09
	7	152,00	(*)	24,00	(*)	1,92	(*)

Obs. - M - média; DP - desvio padrão

(*) grupo constituído por apenas um elemento

TABELA VII.3 - Características dos grupamentos efetuados na amostra de 75 larvas, tomando como característica para o grupamento, a cápsula cefálica.

Número de Grupos	Grupos	Peso (mg)		Comp. (mm)		Medida Cáp. Cef.	
		M	DP	M	DP	M	DP
4	1	1,57	2,58	3,74	2,09	0,44	0,22
	2	50,53	32,87	14,94	3,19	1,42	0,26
	3	113,50	30,14	20,33	2,38	1,69	0,04
	4	112,40	20,02	20,44	1,33	1,86	0,05
5	1	0,61	0,50	3,11	1,42	0,36	0,11
	2	7,33	2,51	7,50	1,32	0,91	0,05
	3	50,53	32,87	14,94	3,19	1,41	0,26
	4	113,50	30,14	20,33	2,38	1,69	0,04
	5	112,40	20,02	20,40	1,33	1,86	0,05
6	1	0,12	0,14	1,71	0,57	0,26	0,05
	2	0,93	0,37	4,00	1,00	0,43	0,07
	3	7,33	2,52	7,50	1,32	0,91	0,05
	4	50,53	32,87	14,94	3,19	1,41	0,26
	5	113,50	30,14	20,33	2,38	1,69	0,04
	6	112,40	20,02	20,40	1,33	1,86	0,05
7	1	0,12	0,14	1,71	0,57	0,26	0,05
	2	0,93	0,37	4,00	1,00	0,43	0,07
	3	7,33	2,52	7,50	1,32	0,91	0,05
	4	45,73	20,99	14,65	3,16	1,39	0,26
	5	97,25	23,00	17,75	2,06	1,66	0,06
	6	113,50	20,02	20,33	1,33	1,69	0,51
	7	112,40	30,14	20,40	3,38	1,86	0,44

Obs. - M - média; DP - desvio padrão

V1 = peso

V2 = comprimento

V3 = cápsula cefálica,

obteve-se, para essa amostra reduzida, a seguinte matriz de correlação, onde r_{ij} representa o coeficiente de correlação entre V_i e V_j:

$$R = \begin{bmatrix} 1,00 & 0,89 & 0,84 \\ 0,89 & 1,00 & 0,96 \\ 0,84 & 0,96 & 1,00 \end{bmatrix}$$

Nota-se então que a variável V3, "cápsula cefálica", é altamente correlacionada com as demais, isto é, o comportamento das outras duas variáveis é muito semelhante ao do diâmetro da cápsula cefálica.

Isto quer dizer que grupamentos homogêneos formados apenas com base na cápsula cefálica também tenderão a ser homogêneos no que diz respeito ao comprimento e ao peso.

O problema foi inicialmente resolvido pelo método das K-Medianas (apresentado no item V.3), para p=3 (peso, comprimento, cápsula cefálica - dados padronizados) e p=1 (cápsula cefálica). Os resultados estão resumidos nas Tabelas (VII.2) e (VII.3).

Como se observa na Tabela (VII.2), para p=3 os desvios padrão da cápsula cefálica nos diversos grupos são elevados. Os grupos são bastante homogêneos no que diz respeito ao comprimento, que foi assim a variável decisiva nessa classificação. De qualquer forma, os grupamentos obtidos para p=3 não satisfazem ao requisito de pequena variância na cápsula cefálica (o desvio padrão, em alguns casos, chega a ser próximo de 50% da média).

Os resultados para p=1 já apresentam, como se nota na Tabela (VII.3), de uma maneira geral, uma maior homogeneidade no que diz respeito à cápsula cefálica. Os desvios padrão estão ainda um pouco elevados (ainda aqui existem casos em que o desvio padrão é de cerca de 50% da média), mas são inferiores aos do caso anterior.

Ilustrando os aspectos computacionais envolvidos, cabe comentar que as soluções iniciais para a utilização do algoritmo das K-Medianas foram fornecidas pelo método de Ward, tendo o algoritmo das K-Medianas convergido, em todos os casos, em uma única iteração. No caso específico de p=1 e m=4 e 5, tentou-se avaliar a qualidade da solução obtida através do

método de Mulvey e Crowder, uma vez que este método, como foi visto no item VI.4, fornece um limite superior (função objetivo primal) e um limite inferior (função objetivo dual) para o valor ótimo da função objetivo. A tentativa foi infrutífera, pois o método, nos dois exemplos, apesar de atingir 40 iterações, forneceu valores de funções objetivo primal e dual ainda bastante defasadas, como se pode notar pela Tabela (VII.4):

TABELA VII.4 - Aplicação do Método de Mulvey e Crowder para $p = 1$ e $m = 4$ e 5

Número de Grupos	Número de Iterações	Função Objetivo Primal	Função Objetivo Dual
4	40	$3,80 \times 10^6$	$-3,05 \times 10^6$
5	40	$3,02 \times 10^6$	$-1,49 \times 10^5$

Ainda que o requisito de homogeneidade não tenha sido satisfeito, procedeu-se, no caso de $p = 1$, a uma verificação das condições impostas pela Regra de Dyat, descrita pela equação (VII.1), através da metodologia apresentada no item VII.2. Os resultados estão resumidos na Tabela (VII.5) que se segue:

TABELA VII.5 - Verificação da Regra de Dyat para $p = 1$ e $m = 4, 5, 6$ e 7

Número de Grupos	K	R^2
4	1,22	0,97
5	1,23	0,99
6	1,23	0,93
7	1,16	0,92

Como se nota, em todos os casos o ajustamento foi de boa qualidade, sem que se possa, a rigor, optar por qualquer um deles como sendo indiscutivelmente o melhor.

Para solucionar o impasse, tentou-se, através de um algoritmo que pudesse oferecer uma solução ótima, gerar uma partição da amostra de larvas que não apresentasse os problemas até aqui descritos.

VII.3.2 - AMOSTRA INTEGRAL

Tendo em vista a maior eficiência do algoritmo univariado de programação dinâmica apresentado no item VI.5, procedeu-se a uma partição , através do citado algoritmo, de toda a amostra de 191 larvas em 4, 5, 6 e 7 grupos, utilizando como variável de classificação o diâmetro da cápsula cefálica e como função objetivo a soma dos quadrados dentro dos grupos. As características da solução, em termos de medida da cápsula cefálica, estão resumidas na Tabela (VII.6).

Como se nota, não houve uma melhoria acentuada na homogeneidade dos grupos. Os desvios padrão ainda que ligeiramente menores, são ainda elevados: ainda existem grupos cujo desvio padrão é de cerca de 30% da média. Ainda assim, novamente se verificou o comportamento dos grupamentos efetuados, em termos de satisfação da Regra de Dyat. Os resultados são apresentados na Tabela (VII.7), a seguir:

TABELA VII.7 - Verificação da Regra de Dyat no caso de Soluções Ótimas para p=1 e m = 4,5,6 e 7

Número de Grupos	K	R^2
4	1,33	0,95
5	1;24	0,97
6	1,22	0,99
7	1,19	1,00

Aqui novamente não é possível adotar-se um valor indiscutível de m: todos são "adequados".

VII.4 - CONCLUSÃO

Pela descrição do problema, apresentada no item VII.1, a aplica

TABELA VII.6 - Média e desvio padrão da cápsula cefálica, nas soluções ótimas obtidas para m = 4, 5, 6 e 7.

Número de Grupos	Grupo	M é d i a	Desvio Padrão
4	1	0,42	0,13
	2	1,09	0,12
	3	1,41	0,09
	4	1,75	0,12
5	1	0,42	0,13
	2	1,02	0,10
	3	1,31	0,08
	4	1,59	0,07
	5	1,83	0,10
6	1	0,36	0,08
	2	0,75	0,14
	3	1,14	0,07
	4	1,38	0,06
	5	1,62	0,06
	6	1,85	0,10
7	1	0,36	0,08
	2	0,63	0,05
	3	0,96	0,08
	4	1,21	0,06
	5	1,41	0,06
	6	1,63	0,05
	7	1,85	0,10

ção da análise de grupamento aos dados da amostra deveria gerar, para um valor de m igual 5 ou 6, uma solução com ínfima variância na cápsula cefálica em cada grupo, apresentando na verificação da Regra de Dyat um valor para o coeficiente de determinação R^2 próximo de 1. Para qualquer outro valor de m, esse coeficiente de determinação deveria ser bem inferior e as variâncias bem maiores.

Graficamente, isso significa dizer que os dados relativos ao diâmetro da cápsula cefálica deveriam distribuir-se de uma forma semelhante ao indicado na Figura (VII.1):

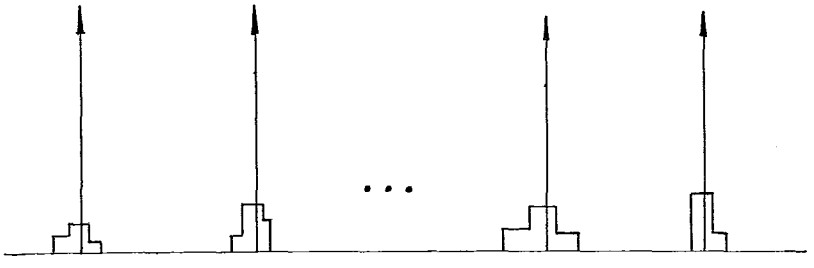


Figura VII.1 - Distribuição esperada da amostra de larvas

Um exame do histograma da amostra, indicado na Figura (VII.2) indica que tal fenômeno não ocorre:

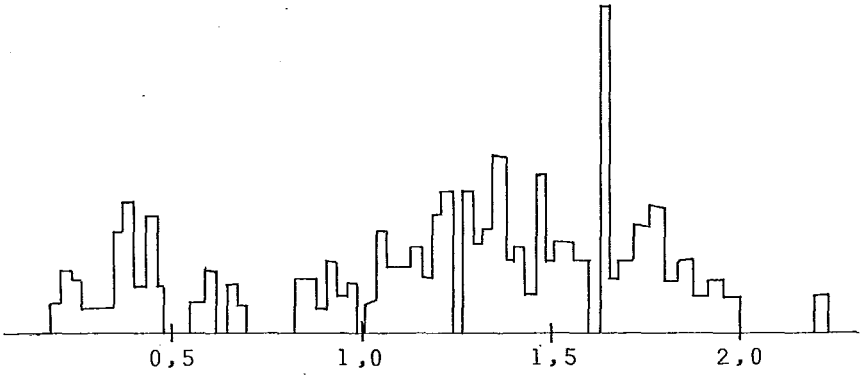


Figura VII.2 - Distribuição efetiva da amostra de larvas

Assim, os resultados da análise de grupamento não devem surpreender. A solução ótima do item VII.3.2 apenas informa que não é possível particionar-se a amostra em estudo na forma apresentada na Figura (VII.1), fato esse que é sugerido pelo próprio exame da Figura (VII.2).

Isso leva a supor, finalmente, que não existem valores típicos para o diâmetro da cápsula cefálica em cada instar. Ainda que a Regra de

Dyat seja satisfeita, tal fato ocorre para cada larva, individualmente , não sendo assim possível a caracterização de um diâmetro de cápsula cefálica típico para cada instar. Consequentemente, também não é possível, à luz dos resultados obtidos, tomar uma larva ao acaso no campo e determinar em que instar ela se encontra.

Cabe ressaltar que os resultados aqui expostos baseiam-se meramente na aplicação de métodos de análise de grupamento. No sentido de validar ou não esses resultados, seria interessante aplicar outras técnicas de análise ao problema em questão.

VIII - CONCLUSÃO

Em vista do exposto nos capítulos anteriores, deve ter ficado clara a natureza e as características de cada método apresentado. Resta - ria no entanto, abordar a questão do porte dos problemas a serem resolvi dos pelos diversos algoritmos aqui descritos.

Se, por um lado, deve ter ficado evidente a melhor aptidão dos métodos de realocação iterativa para tratar problemas de maior porte,dada a sua grande rapidez, por outro é preciso levar em conta que não se pode, a priori, indicar o tamanho (em termos de número de elementos a serem agru pados) dos problemas que cada método pode tratar: isto é função inclusive da disponibilidade de memória e tempo de processamento em computador para cada usuário.

Assim, é bastante provável que na prática ocorram problemas cujo porte é superior aos recursos disponíveis.

Existem pelo menos duas maneiras de se contornar essa dificulda de. Como exemplo, seja 250 o número de elementos que os recursos disponí- veis permitem agrupar e 5.000 o tamanho do problema original.

Uma das maneiras de se resolver o problema seria tomar uma amos tra de 250 elementos, formar um "grupamento semente" com essa amostra e, a partir desse "grupamento semente", classificar os demais 4.750 elementos.Es sa classificação poderia ser efetuada alocando-se cada elemento não classi ficado ao grupo mais próximo ou então poder-se-ia utilizar as técnicas de classificação descritas, por exemplo, em TATSUOKA[39] ou CODLEY e LOHNES[7].

Outra técnica para a resolução do problema de grande porte é devida a ANDERBERG[1] e procura particionar o conjunto original em blocos cuja análise de grupamento seja viável. Em cada bloco é efetuada uma análi se de grupamento, em um número predeterminado de grupos. Após a resolução do problema em cada bloco, efetua-se uma análise de grupamento no conjun to de todas as centróides ou medianas dos grupos definidos anteriormente . Terminada essa etapa, os elementos do conjunto original são alocados ao grupo a que pertence sua centróide ou mediana (definida quando do problema restrito a cada bloco).

Como exemplo, seja novamente o problema de tamanho 5.000, onde os recursos permitem que se efetue análise de grupamento de até 250 elemen tos. Pela técnica de ANDERBERG[1], o problema poderia ser resolvido da se guinte forma :

- o conjunto dos 5.000 elementos seria particionado em 20 blocos de 250 elementos;

- em cada bloco seria efetuada uma análise de grupamento em 12 grupos, gerando-se, em cada problema, 12 centróides ou medianas;

- a partir das 20 x 12 = 240 centróides ou medianas seriam, através de análise de grupamento, gerados m grupos; e

- os 5.000 elementos originais seriam classificados nos m grupos conforme a pertinência da centróide ou mediana definida na segunda etapa.

Como se nota, foram resolvidos 21 problemas de análise de grupamento.

REFERÊNCIAS BIBLIOGRÁFICAS

1 - ANDERBERG, M.R. - Cluster analysis for applications. New York Academic Press, 1973. 361 p.

2 - ASTRAHAN, M.M. - Speech analysis by clustering, or the hyperphoneme method; Stanford Artificial Intelligence Proj. Mem., AIM - 124, AD 709067. Stanf., Calif., Stanford University, 1970.

3 - BALL, G.H. e HALL, D.J. - ISODATA, a novel method of data analysis and pattern classification; technical report. Menlo Park, California, Stanford-Research Institute, 1965. 72 p.

4 - BALL, G.M. & HALL, D.J. - PROMENADE - an on-line pattern recognition system; Rep. No. RADC - TR - 67-310, AD 822174. Menlo Park, California, Stanford Res. Inst., 1967. 124 p.

5 - BAZARAA, M.S. & SHETTY, C.M. - Nonlinear Programming. New York, John Wiley and Sons, 1979.

6 - BELLMAN, R. - A note on cluster analysis and dynamic programming. Mathematical Biosciences, s.l., 18:311-312, 1973.

7 - COOLEY, W.W. e LOHNES, P.R. - Multivariate Data Analysis. New York, John Wiley and Sons, 1971. 364 p.

8 - DIDAY, E. et alii - A new kind of representation in clustering. S.n.t. 7 p.

9 - DIDAY, E. & SIMON J.C. - Clustering Analysis. Communication and Cybernetics, Heidelberg, 10:46-94, 1976.

10 - DURAN, B.S. e ODELL, P.L. - Cluster Analysis - a Survey. Heidelberg , Springer-Verlag Berlin, 1974. 137 p.

11 - EFFROYMSON, M.A. & RAY, T.L. - A branch-and-bound algorithm for plant

location. Operations Research, Baltimore, 14:361-368, 1966.

12 - EVERITT, B. - Cluster Analysis. London, Heinemann Educational Books Ltd., 1974.

13 - FORGY, E.W. - Cluster analysis of multivariate data : efficiency versus interpretability of classifications. B i o m e t r i c s, Washington D.C., 21(3):768, 1965.

14 - GOWER, J.C. - A general coefficient of similarity and some of its properties. Biometrics, Washington D.C., 27:857-874, 1971.

15 - HANSEN, P. & DELATTRE, M. - Complete-link cluster by graph coloring. Journal of the American Statistical Association, Washington D. C., 73(362):397-403, 1978.

16 - HARMAN, H.H. - Modern Factor Analysis. Chicago, The University of Chicago Press, 1976. 487 p.

17 - HARTIGAN, J.A. - Clustering Algorithms. New York, John Wiley and Sons, 1971.

18 - HELD, M., WOLFE, P. & CROWDER, H.P. - Validation of Subgradient Optimization. Mathematical Programming, Amsterdam, 6:62-88, 1974.

19 - JANCEY, R.C. - Multidimensional group analysis. Austral. J. Botany, s.l., 14(1):127-130, 1966.

20 - JARVINEN, P., RAJALA, J. e SINERVO, M.- A branch-and-bound algorithm for seeking the p-median. Operations Research, Baltimore, 20:173 -178, 1972.

21 - JENSEN, R.E. - A dynamic programming algorithm for cluster analysis. Operations Research, Baltimore, 17:1034-1057, 1969.

22 - JOHNSON, S.C. - Hierarchical clustering schemes. Psychometrica , Williansburg, 32:241-254, 1967.

23 - LANCE, G.N. & WILLIAMS, W.T. - A generalized sorting strategy for computer classifications. Nature, New York, 212:218, 1966.

24 - LANCE, G.N. & WILLIAMS, W.T. - A general theory of classificatory sorting strategies. I- Hierarchical Systems. Computer Journal , London, 9(4):373-380, 1966.

25 - LANCE, G.N. & WILLIAMS, W.T. - A general theory of classificatory sorting strategies. II- Clustering Systems. Computer Journal , London, 10(3):276, 1967.

26 - MACNAUGHTON-SMITH, P. - Some statistical and other numerical techniques for classifying individuals. Home Office Research Unit Report No. 6, London, HMSO, 1965.

27 - MACQUEEN, J.B. - Some methods for classification and analysis of multivariate observations. Proc. Symp. Math. Statist. and Probab., 5th, Berkeley, Berkeley, Univ. of Calif., 1:281-297, 1967.

28 - MATEUS, G.R. e BORNSTEIN, C.T. - Um algoritmo guloso para o problema de localização não capacitado; Relatório Técnico do Programa de Engenharia de Sistemas e Computação ES 06-87. Rio de Janeiro, COPPE/UFRJ, 1981. 24 p.

29 - MULVEY, J.M. & CROWDER, H.P. - Cluster analysis: an application of lagrangian relaxation. Management Science, Providence, 25(4):329 -340, 1979.

30 - PINHO GAMA, M. - Bases da Análise de Grupamento. Brasília, Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Estatística, 1980. 229 p.

31 - PFEIFFER, D. - Disparidades de Desenvolvimento no Brasil - um exemplo da análise de cluster. Revista Brasileira de Estatística, Rio de Janeiro, 41(164):559-576, out./dez. 1980.

32 - RAJ, D. - Sampling theory. New Delhi, Tata Mcgraw - Hill Publishing

Company Ltd., 1978. 302 p.

33 - RAO, M.R. - Cluster analysis and mathematical programming. Journal of the American Statistical Association, Washington D.C., 66:622 -626, 1971.

34 - ROHLF, F.J. - Hierarchical clustering using the minimum spanning tree. The Computer Journal, London, 16:93-95, 1973.

35 - ROY, B. - An algorithm for a general constrained set covering problem; IN: Graph Theory and Computing. New York, Academic Press, 1972. p. 267-283

36 - SIBSON, R. - SLINK: an optimally efficient algorithm for the single -link cluster method. The Computer Journal, London, 16:30-34 , 1973.

37 - SOKAL, R.R. e MICHENER, C.D. - A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletim, Kansas, 1409-1438, 1958.

38 - SUDMAN, S. - Applied Sampling. New York, Academic Press, 1976. 251 p.

39 - TATSUOKA, M.M. - Multivariate analysis: techniques for educational and psychological research. New York, John Wiley and Sons, 1971. 310 p.

40 - TRYON, R.C. - Cluster Analysis. Ann Arbor , Michigan, Edwards Bros., 1939.

41 - VINOD, H.D. - Integer programming and the theory of grouping. Journal of the American Statistical Association, Washington D.C., 64:506-519, 1969.

42 - WARD, J.H. - Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association ,

Washington D.C., 58(301):236-244, 1963.

43 - WESOLOWSKY, G.O. - Multiple regression and analysis of variance. New York, John Wiley and Sons, 1976. 292 p.

44 - WISHART, D. - An algorithm for hierarchical classifications. Biometrics, Washington D.C., 22(1):165-170, 1969.

45 - WISHART, D. - Mode analysis: a generalization of nearest neighbor which reduces chaining effects, em A.J. Cole (ed.) IN: Numerical Taxonomy. New York, Academic Press, 1969. p. 282-319

46 - ZAHN, C.T. - Graph - theoretical methods for detecting gestalt clusters. IEEE Transactions on Computers, 20:68-86, 1971.

47 - FISHER, M.L., NEMHAUSER, G.L. e WOLSEY, L.A. - An analysis of approximations for Maximizing Submodular Set Functions - I". Mathematical Programming, Amsterdam, 14:265-294, 1978.

48 - FISHER, NEMHAUSER e WOLSEY - An analysis of approximations for Maximizing Submodular Set Functions - II". Mathematical Programming Study, Amsterdam, 8:73-87, 1978.

A P Ê N D I C E I

TABELA A.1 - Distâncias entre pontos

Distâncias	Equações
Métricas de Minkowsky (1)	$d_{ij} = d_{\lambda}(X_i, X_j) = \left[\sum_{k=1}^{p} \lvert x_{ki} - x_{kj} \rvert^{\lambda} \right]^{1/\lambda}$
Métricas de Similaridade (2)	$d_{ij} = (1 - s_{ij})^{1/2}$
"Métrica" de Correlação (3)	$d_{ij} = \left[0.5 (1 - r_{ij}) \right]^{1/2}$

Obs. - (1) Quando $\lambda = 2$, a métrica de Minkowsky se transforma na distância euclidiana.

(2) s_{ij} - Coeficiente de similaridade entre e_i e e_j .

(3) r_{ij} - Coeficiente de correlação entre e_i e e_j .

TABELA A.2 - Medidas de dispersão interna de um grupo

Medidas	Equações
Soma dos quadrados dentro do grupo (1)	$W_I = \sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I)$
Variância interna de grupo	$S_I = \frac{1}{n_I} W_I$
Diâmetro de grupo	$d(g_I) = \max_{e_i, e_j \in g_I} d_{ij}$
Dispersão via mediana de grupo	$Z_{min}(g_I) = \min_{e_j \in g_I} Z(e_j) = \min_{e_j \in g_I} \sum_{i=1}^{n_I} d_{ij}$

Obs. - (1) - $\bar{X}_I$ é a centróide (média) do grupo g_I : $\bar{X}_I = (\sum_{i=1}^{n_I} X_i) / n_I$; e

- $d_2^2 (X_i, \bar{X}_I)$ é o quadrado da distância euclidiana entre X_i e $\bar{X}_I$.

TABELA A.3 - Funções objetivo em análise de grupamento

Funções Objetivo	Equações
Soma dos quadrados dentro dos grupos	$W = \sum_{g_I \varepsilon P_m} W_I = \sum_{g_I \varepsilon P_m} \sum_{i=1}^{n_I} d_2^2 (X_i, \bar{X}_I)$
Soma das variâncias internas	$S^2 = \sum_{g_I \varepsilon P_m} S_I^2 = \sum_{g_I \varepsilon P_m} \frac{1}{n_I} W_I$
Diâmetro de grupo	$d(P_m) = \max_{g_I \varepsilon P_m} d(g_I) = \max_{g_I \varepsilon P_m} \max_{e_i, e_j \varepsilon g_I} d_{ij}$
Dispersão via mediana de grupo	$Z(P_m) = \sum_{g_I \varepsilon P_m} Z_{min}(g_I) = \sum_{g_I \varepsilon P_m} \min_{e_j \varepsilon g_I} \sum_{i=1}^{n_I} d_{ij}$

TABELA A.4 - Distâncias entre grupos utilizadas em métodos hierarquizados aglomerativos

Métodos	Distâncias
Ligação Simples	$ds_{IJ} = \min_{\substack{e_i \in g_I \\ e_j \in g_J}} d_{ij}$
Ligação Completa	$dc_{IJ} = \max_{\substack{e_i \in g_I \\ e_j \in g_J}} d_{ij}$
Centróide (1)	$d^2_{IJ} = d^2_2 (\bar{X}_I, \bar{X}_J)$
Ward	$D_{IJ} = \frac{n_I n_J}{n_I + n_J} d^2_{IJ}$
Média de Grupo (2)	$D^2_{IJ} = S^2_I + S^2_J + d^2_{IJ}$

Obs. - (1) $\bar{X}_I$ é a centróide (média) do grupo g_I e $d^2_2 (\bar{X}_I, \bar{X}_J)$ o quadrado da distância euclidiana entre $\bar{X}_I$ e $\bar{X}_J$. O método da mediana (v. item IV.2.4) difere do método da centróide apenas no que se refere à determinação das centróides.

(2) S^2_I é a variância interna do grupo g_I (v. item II.4.2 ou Tabela A.2).

A P Ê N D I C E II

A2.1 - Rotina CLSTOT

Essa rotina efetua a análise de grupamento através das técnicas hierarquizadas aglomerativas descritas nesta tese e dos métodos K-Medianas e Mulvey e Crowder. Uma listagem da rotina é apresentada nas folhas que se seguem.

Os dados de entrada do programa devem ser organizados da seguinte forma:

- 1º cartão: informar, em 7 campos consecutivos de 5 posições, iniciando-se o conjunto na coluna 1, os seguintes dados, alinhando os valores à direita:

  . 1º campo - número de características

  . 2º campo - número de elementos

  . 3º campo - número de grupos

  . 4º campo - tipo de impressão, na seguinte forma:

  - "1" - resultados intermediários dos métodos hierarquizados (dendogramas e distância entre os grupos unidos em cada iteração).

  - "2" - resultados intermediários do método K - Medianas (valor da função objetivo em cada iteração).

  - "3" - resultados intermediários do método Mulvey e Crowder (valores das funções objetivo primal e dual, e sua diferença percentual, em cada iteração).

  - "4" - resultados intermediários dos métodos K-Medianas e Mulvey e Crowder.

  - "5" - resultados intermediários de todos os métodos.

  Obs.: - em todos os métodos, qualquer que seja o valor informado nesse 4º campo, os resultados finais são sempre impressos.

  5º campo - número máximo de iterações nos métodos K - Me

dianas e Mulvey e Crowder.

6º campo - distância a ser utilizada. Se informado o valor "1", é calculada a distância euclidiana. Se informado "2", é calculada a métrica de correlação. Se informado um valor diferente de "1" e "2", ocorrerá erro.

7º campo - padronização das variáveis. Se informado "1", os dados são padronizados. Para qualquer outro valor, os dados permanecem sem padronização.

2º cartão: informar, em 8 campos consecutivos de 1 posição, iniciando-se o conjunto na coluna 1, os métodos a serem utilizados. O método será utilizado se na posição correspondente for informado "1":

. 1º campo - Ligação Simples

. 2º campo - Ligação Completa

. 3º campo - Mediana

. 4º campo - Média de Grupo

. 5º campo - Centróide

. 6º campo - Ward

. 7º campo - K-Medianas

. 8º campo - Mulvey e Crowder

Obs.: O método K-Medianas exige o processamento anterior de uma técnica hierarquizada. O método Mulvey e Crowder exige o processamento anterior do método K-Medianas.

- cartões subseqüentes: informar, para cada elemento:

1º cartão - código do elemento, com 6 caracteres, iniciando

do-se na coluna 1.

demais cartões - informar, em seqüencia, os valores das p observações para o elemento, em campos de 10 posições, com 2 casas decimais, iniciando-se o conjunto na coluna 1.

Exemplo: Seja m = 3, n = 5, p = 2, utilizando-se os métodos de Ward, K-Mediana e Mulvey e Crowder, a distância euclidiana em dados padronizados , até um máximo de 40 iterações para os métodos K-Medianas e Mulvey e Crowder, imprimindo-se todos os passos intermediários, com os seguintes dados:

Elemento	Observações	
A1	10.000	3,5
A2	5.000	4,2
A3	2.500	3,1
A4	3.000	2,0
A5	2.000	1,8

Esses dados seriam lidos na rotina CLSTOT da forma (Ø e Ø representam, respectivamente, espaço em branco e zero):

ØØØØØØ 2 ØØØØØØ 5 ØØØØØØ 3 ØØØØØØ 5 ØØØØØ 5Ø ØØØØØØ 1 ØØØØØØ 1
ØØØØØ 111
ØØØØ A1
ØØØ 1ØØØØØØ ØØØØØØØ 35Ø
ØØØØ A2
ØØØØ 5ØØØØØ ØØØØØØØ 42Ø
ØØØØ A3
ØØØØ 25ØØØØ ØØØØØØØ 31Ø
ØØØØ A4
ØØØØ 3ØØØØØ ØØØØØØØ 2ØØ
ØØØØ A5
ØØØØ 2ØØØØØ ØØØØØØØ 18Ø

C L S T O T

```
%
%   ESTA ROTINA EFETUA UMA ANALISE DE AGRUPAMENTO
%   ATRAVES DOS METODOS HIERARQUICOS AGLOMERATIVOS,
%   K-MEDIANAS E WOLVEY E CROWDER, PADRONIZANDO OU
%   NAO OS DADOS E UTILIZANDO A METRICA EUCLIDIANA
%   OU A "METRICA" DE CORRELACAO
%
BEGIN

      FILE INPT  (KIND=READER),

           OTPT  (KIND=PRINTER);

%     **********************
      PROCEDURE PDRZ(N,P,X);
%               ====
%     ESTA ROTINA PADRONIZA OS DADOS DE UMA MATRIZ
%     X[1:N,1:P],FAZENDO COM QUE CADA CARACTERISTICA
%     X[*,K] TENHA MEDIA 0 E VARIANCIA 1
%
      INTEGER N,          %NUMERO DE ELEMENTOS
              P;          %NUMERO DE CARACTERISTICAS
      REAL ARRAY X[1,1]; %MATRIZ DE OBSERVACOES
%     **********************
           BEGIN
              INTEGER I,K;          %CONTADORES AUXILIARES

              REAL ARRAY MD[1:P],   %VETOR DE MEDIAS
                         VAR[1:P]; % VETOR DE VARIANCIAS
%
% CALCULO DAS MEDIAS
%
              FOR   K:=1
              STEP  1
              UNTIL P
              DO    BEGIN
                    FOR   I:=1
                    STEP  1
                    UNTIL N
                    DO    MD[K]:=*+X[I,K];
                    MD[K]:=*/N
                    END;
%
% CALCULO DAS VARIANCIAS
%
              FOR   K:=1
              STEP  1
              UNTIL P
              DO    BEGIN
```

```
                FOR   I:=1
                STEP  1
                UNTIL N
                DO    VAR[K]:=*+(X[I,K]-MD[K])**2;
                VAR[K]:=*/N
                END;
        %
        % CALCULO DOS DADOS PADRONIZADOS
        %
          FOR   I:=1
          STEP  1
          UNTIL N
          DO    FOR   K:=1
                STEP  1
                UNTIL P
                DO    X[I,K]:=(X[I,K]-MD[K])/VAR[K]
      END DA PROCEDURE PDRZ;

%     ************************
      PROCEDURE ECLD(N,E,X,P);
%               ====
%     ESTA ROTINA CALCULA O QUADRADO DA DISTANCIA EUCLIDIANA
%     E[I+(J-1)*(J-2)/2],ENTRE X[I,*] E X[J,*],TAIS QUE J>1,I=1,...,N-1
      INTEGER N,            %NUMERO DE ELEMENTOS
              P;            %NUMERO DE CARACTERISTICAS
      REAL ARRAY E[1],      %VETOR DE QUADRADOS DAS DISTANCIAS
                 X[1,1];    %MATRIZ DE OBSERVACOES
%     ************************
      BEGIN
            INTEGER I,J,K,L; %CONTADORES AUXILIARES

      FOR   I:=1
      STEP  1
      UNTIL N-1
      DO    FOR   J:=I+1
            STEP  1
            UNTIL N
            DO    BEGIN
                  L:=I+(J-1)*(J-2)/2;
                  E[L]:=0;
                  FOR   K:=1
                  STEP  1
                  UNTIL P
                  DO    E[L]:=*+(X[I,K]-X[J,K])**2;
                  END;
      END DA PROCEDURE ECLD;

%     ************************
      PROCEDURE MTCR(N,E,X,P);
%               ====
%     ESTA ROTINA CALCULA A "METRICA" DE CORRELACAO E[I+(J-1)*(J-2)/2]
%     ENTRE X[I,*] E X[J,*], TAIS QUE J > I,I=1,...,N-1
%
      INTEGER N,            %NUMERO DE ELEMENTOS
              P;            %NUMERO DE CARACTERISTICAS
      REAL ARRAY E[1],      %VETOR DE DISTANCIAS
                 X[1,1];    %MATRIZ DE OBSERVACOES
%     ************************
      BEGIN
            INTEGER I,J,K,L;
```

```
            REAL ARRAY MD[1:N];
            REAL AUX1,AUX2,AUX3,AUX4,AUX5;
            FOR   I:=1
            STEP  1
            UNTIL N
            DO    BEGIN
                  FOR   K:=1
                  STEP  1
                  UNTIL P
                  DO    MD[K]:=*+X[I,K];
                  MD[I]:=*/P
                  END;
            FOR   I:=1
            STEP  1
            UNTIL N-1
            DO    FOR   J:=I+1
                  STEP  1
                  UNTIL N
                  DO    BEGIN
                        AUX1:=AUX2:=AUX3:=AUX4:=AUX5:=0;
                        L:=I+(J-1)*(J-2)/2;
                        FOR   K:=1
                        STEP  1
                        UNTIL P
                        DO    BEGIN
                              AUX4:=X[I,K]-MD[I];
                              AUX5:=X[J,K]-MD[J];
                              AUX1:=*+AUX4*AUX5;
                              AUX2:=*+AUX4**2;
                              AUX3:=*+AUX5**2
                              END;
                        E[L]:=AUX1/(SQRT(AUX2)*SQRT(AUX3));
                        E[L]:=SQRT((1-E[L])/2)
                        END
      END DA PROCEDURE MTCR;

%     ************************************************************
      PROCEDURE CLSTHRU(N,MTD,TOTGRP,W,FO,AGRP,E,NIG,TIMP,COG);
%           ========
%     ESTA ROTINA EFETUA UMA ANALISE DE GRUPAMENTO DE N ELEMENTOS EM
%     TOTGRP GRUPOS,ATRAVES DO METODO DEFINIDO PELO VALOR DE MTD:
%                     0= LIGACAO SIMPLES
%                     1= LIGACAO COMPLETA
%                     2= MEDIANA
%                     3= MEDIA DE GRUPO
%                     4= CENTROIDE
%                     5= WARD,
%     SENDO A PARTICAO OBTIDA ARMAZENDA NO VETOR AGRP[1:N]
%
      INTEGER N,                  %NUMERO DE ELEMENTOS
              MTD,                %METODO DE GRUPAMENTO
              TOTGRP,             %NUMERO TOTAL DE GRUPOS
              TIMP;               %TIPO DE IMPRESSAO
      INTEGER ARRAY NIG[1],       %NUMERO DE ELEMENTOS EM CADA GRUPO
                    AGRP[1];      %AGRUPAMENTO EFETUADO
      REAL W,                     %MENOR DISTANCIA ENTRE GRUPOS EM CADA PASSO
           FO;                    %FUNCAO OBJETIVO,NO CASO DO METODO DE WARD
      REAL ARRAY E[1],            %VETOR DE DISTANCIAS
                 COG[1];          %VETOR DE CODIGOS DOS ELEMENTOS
```

```
%     ***********************************************************
      BEGIN
%     ***********************************************
      PROCEDURE ATUALIZARD(IM,JM,N,NIG,MTD,D,P,K);

%               ==========
%     ESTA ROTINA ATUALIZA O VETOR D DE DISTANCIA ENTRE OS GRUPOS,
%     QUANDO OS GRUPOS INDICADOS NAS COLUNAS(OU LINHAS) P[IM]  E
%     P[JM]
      INTEGER IM,              %LINHA IM E COLUNA JM PARA AS QUAIS A
              JM,              %DISTANCIA ENTRE OS GRUPOS FOI MINIMA
              N ,              %NUMERO DE ELEMENTOS
              K,               %NUMERO DE CARACTERISTICAS
             MTD;              %METODO UTILIZADO
      INTEGER ARRAY P[1],      %VETOR DE APONTADORES DAS LINHAS ATIVAS
                    NIG[1];    %NUMERO DE ELEMENTOS DO GRUPO
      REAL ARRAY D[1];         %VETOR DE DISTANCIAS ENTRE OS GRUPOS
%     ***************************************
            BEGIN
%           ***************************************************
            PROCEDURE CLCATL(IM,JM,NIG,MTD,D,LHI,LHJ,LIJ,H,P);

%                     ======
%           ESTA ROTINA EFETUA A ATUALIZACAO DO VETOR DE DISTANCIAS
%           D,QUANDO OS GRUPOS P[IM] E P[JM] SAO UNIDOS,NO METODO
%           INDICADO POR MTD,CALCULANDO A DISTANCIA ENTRE P[H] E
%           P[IM] U P[JM]
            INTEGER IM,JM,MTD,LHI,LHJ,LIJ,H;
            INTEGER ARRAY P[1],
                          NIG[1];
            REAL ARRAY D[1];
%           ***********************************************
                  BEGIN
                  REAL AUX1,AUX2;

                        CASE MTD
                        OF    BEGIN
                              D[LHI]:=(D[LHI]+D[LHJ]-ABS(D[LHI]-
                               D[LHJ]))/2;
                              D[LHI]:=(D[LHI]+D[LHJ]+ABS(D[LHI]-
                               D[LHJ]))/2;
                              D[LHI]:=(D[LHI]+D[LHJ])/2-
                               D[LHJ]/4;
                              BEGIN
                              D[LHI]:=NIG[P[IM]]*D[LHI]+NIG[P[JM]]*
                               D[LHJ];
                              D[LHI]:=*/(NIG[P[IM]]+NIG[P[JM]])
                              END;
                              BEGIN
                              AUX1:=NIG[P[IM]]*D[LHI]+NIG[P[JM]]*
                               D[LHJ];
                              AUX1:=*/(NIG[P[IM]]+NIG[P[JM]]);
                              AUX2:=NIG[P[IM]]*NIG[P[JM]]*D[LIJ];
                              AUX2:=*/(NIG[P[IM]]+NIG[P[JM]])**2;
                              D[LHI]:=AUX1-AUX2
                              END;
                              BEGIN
                              AUX1:=(NIG[P[H]]+NIG[P[IM]])*D[LHI]+
                               (NIG[P[H]]+NIG[P[JM]])*D[LHJ];
                              AUX1:=*-NIG[P[H]]*D[LIJ];
```

```
                        AUX2:=NIG[P[H]]+NIG[P[IM]]+NIG[P[JM]];
                        D[LHI]:=AUX1/AUX2
                        END
                        END
            END DA PROCEDURE CLCATL;

            INTEGER H,
                    LHI,LHJ,LIJ;
            LIJ:=P[IM]+(P[JM]-1)*(P[JM]-2)/2;
            FOR   H:=1
            STEP  1
            UNTIL H=IM-1
            DO    BEGIN
                  LHI:=P[H]+(P[IM]-1)*(P[IM]-2)/2;
                  LHJ:=P[H]+(P[JM]-1)*(P[JM]-2)/2;
                  CLCATL(IM,JM,NIG,MTD,D,LHI,LHJ,LIJ,H,P);
                  END;
            FOR   H:=IM+1
            STEP  1
            UNTIL H=K+1
            DO    IF    H=JM
                  THEN BEGIN
                       LHI:=P[IM]+(P[H]-1)*(P[H]-2)/2;
                       IF    H<JM
                       THEN LHJ:=H+(JM-1)*(JM-2)/2
                       ELSE LHJ:=JM+(H-1)*(H-2)/2;
                       CLCATL(IM,JM,NIG,MTD,D,LHI,LHJ,LIJ,H,P);
                       END
      END DA PROCEDURE ATUALIZARD;

%     ************************
      PROCEDURE TITULO(MTD);
%               =======
%     ESTA ROTINA IMPRIME O TITULO DO METODO,CONFORME O VALOR
%     DE MTD
      INTEGER MTD;
%     ************************
            BEGIN
            WRITE(OTPT[SKIP 1]);
            CASE MTD
            OF    BEGIN
                  WRITE(OTPT,<T10,"METODO:LIGACAO SIMPLES">);
                  WRITE(OTPT,<T10,"METODO:LIGACAO COMPLETA">);
                  WRITE(OTPT,<T10,"METODO:MEDIANA">);
                  WRITE(OTPT,<T10,"METODO:MEDIA DE GRUPO">);
                  WRITE(OTPT,<T10,"METODO:CENTROIDE">);
                  WRITE(OTPT,<T10,"METODO:WARD">);
                  END
            END DA PROCEDURE TITULO;
            REAL ARRAY D[1:N*(N-1)/2];
            REAL MINIMO,
                 AC;
            INTEGER ARRAY P[1:N];
            INTEGER CM,IM,JM;
            INTEGER I,J,K,L;
          %
          % INICIALIZACAO
          %
            IF    TIMP=1 OR TIMP=5
            THEN BEGIN
```

```
                TITULO(MTD);
                WRITE(OTPT,<////,T5,"ITERACAO",T16,"GRUPO ",
                 "FORMADO",T37,"H",T49,"HACUM">)
          END;
    FOR   K:=1
    STEP  1
    UNTIL N*(N-1)/2
    DO    D[K]:=E[K];
    FOR   J:=1
    STEP  1
    UNTIL N
     DO    BEGIN
           NIG[J]:=1;
           P[J]:=J
           END;
    FO:=0;
    FOR   K:=1
    STEP  1
    UNTIL N
    DO    AGRP[K]:=K;
%
% CALCULO DOS GRUPOS PARA OS QUAIS A DISTANCIA E MINIMA
%
    FOR   K:=1
    STEP  1
    UNTIL K=TOTGRP
    DO    BEGIN
          CM:=0;
          FOR  J:=2
          STEP  1
          UNTIL N=K+1
          DO    FOR   I:=1
                STEP  1
                UNTIL J=1
                DO    BEGIN
                      CM:=*+1;
                      L:=P[I]+(P[J]-1)*(P[J]-2)/2;
                      IF    CM=1 OR D[L]<MINIMO
                      THEN BEGIN
                           MINIMO:=D[L];
                           IM:=I;
                           JM:=J
                           END
                      END;
          IF    MTD=5
          THEN BEGIN
               W:=MINIMO/2;
               FO:=*+W;
               END
          ELSE W:=MINIMO;
          IF    TIMP=1 OR TIMP=5
          THEN BEGIN
               WAC:=*+W;
               WRITE(OTPT,<T7,I3,X5,A6,X1,"U",X1,A6,X2,
                     2(E12.5,X2)>,K,CDG[P[IM]],CDG[P[JM]],
                     W,WAC)
               END;
%
% ATUALIZACAO DOS DADOS
%
```

```
                ATUALIZARD(IM,JM,N,NIG,MTD,D,P,K);
                NIG[P[IM]]:=**NIG[P[JM]];
                FOR   I:=1
                STEP  1
                UNTIL N
                DO    IF    AGRP[I]=P[JM]
                      THEN AGRP[I]:=P[IM];
                FOR   I:=JM
                STEP  1
                UNTIL N=K
                DO    P[I]:=P[I+1]
                END;
        %
        % IMPRESSAO DOS RESULTADOS
        %
                TITULO(MTD);
                IF   MTD=5
                THEN WRITE(OTPT,<///,T12,"FUNCAO OBJETIVO = ",
                           F12.5>,FO);
                WRITE(OTPT,<//,T12,"GRUPOS FORMADOS :",/," ">);
                FOR   I:=1
                STEP  1
                UNTIL TOTGRP
                DO    BEGIN
                      WRITE(OTPT,<//,T15,"GRUPO ",I3>,I);
                      FOR   K:=1
                      STEP  1
                      UNTIL N
                      DO    IF    AGRP[K]=P[I]
                            THEN BEGIN
                                 WRITE(OTPT,<T16,A6>,CDG[K]);
                                 AGRP[K]:=I
                                 END
                      END
    END DA PROCEDURE CLSTHRQ;

%   ***********************************************
    PROCEDURE KMDNS(AGRP,N,M,E,MDN,FO,NITER,TIMP,CDG);
%           =====
%   ESTA ROTINA EFETUA UMA ANALISE DE GRUPAMENTO DE N ELEMENTOS EM
%   M GRUPOS,ATRAVES DO METODO DAS K-MEDIANAS,SENDO A PARTICAO OBTIDA
%   ARMAZENADA NO VETOR AGRP[1:N] E AS MEDIANAS DOS GRUPOS,POR SUA VEZ,
%   ARMAZENADA  NO VETOR MDN[1:M]
    INTEGER M,              %NUMERO DE GRUPOS
            N,              %NUMERO DE ELEMENTOS
            NITER,          %NUMERO MAXIMO DE ITERACOES
            TIMP;           %TIPO DE IMPRESSAO
    INTEGER ARRAY MDN[1],   %VETOR DE MEDIANAS
                  AGRP[1];  %AGRUPAMENTO EFETUADO
    REAL FO;                %VALOR DA FUNCAO OBJETIVO
    REAL ARRAY E[1],        %VETOR DE DISTANCIAS
               CDG[1];      %VETOR DE CODIGOS DOS ELEMENTOS
%   ***********************************************
        BEGIN
%           ***********************************************
            PROCEDURE MDNFO(M,MDN,N,AGRP,E,FO,ITER,TIMP,PREC);

%                   =====
%           ESTA ROTINA CALCULA AS MEDIANAS MDN[I],O VALOR FO DA
%           FUNCAO OBJETIVO ASSOCIADA AO GRUPAMENTO EFETUADO E A
```

```
%               DIFERENCA  PREC ENTRE OS VALORES DA FUNCAO OBJETIVO
%               ASSOCIADOS A DUAS ITERACOES CONSECUTIVAS
                INTEGER M,              %NUMERO DE GRUPOS
                        N,              %NUMERO DE ELEMENTOS
                        ITER,           %NUMERO DE ITERACOES
                        TIMP;           %TIPO DE IMPRESSAO
                INTEGER ARRAY MDN[1], % VETOR DE MEDIANAS
                              AGRP[1];%AGRUPAMENTO EFETUADO
                REAL FO,                %VALOR DA FUNCAO OBJETIVO
                     PREC;              %V. DEFINICAO NO OBJETIVO DA ROTINA
                REAL ARRAY E[1];        %VETOR DE DISTANCIAS
%               ************************************************
                BEGIN
                      INTEGER I,J,L;

                      REAL FA,
                           TOL,
                           CST;
                      REAL ARRAY CSTMIN[1:M];
                %
                % INICIALIZACAO
                %
                  TOL:=0.0000000001;
                  FOR   I:=1 STEP 1     UNTIL M
                  DO    MDN[I]:=0;
                %
                % CALCULO DAS MEDIANAS DOS GRUPOS
                %
                  FOR   I:=1
                  STEP  1
                  UNTIL N
                  DO    BEGIN
                        CST:=0;
                        FOR   J:=1
                        STEP  1
                        UNTIL N
                        DO    BEGIN
                              IF    AGRP[J]=AGRP[I] AND J^=I
                              THEN BEGIN
                                   IF    J>I
                                   THEN BEGIN
                                        L:=I+(J-1)*(J-2)/2;
                                        CST:=*+E[L]
                                        END
                                   ELSE BEGIN
                                        L:=J+(I-1)*(I-2)/2;
                                        CST:=*+E[L]
                                        END
                                   END
                              END;
                        IF    MDN[AGRP[I]]<1 OR CST<CSTMIN[AGRP[I]]
                        THEN BEGIN
                             MDN[AGRP[I]]:=I;
                             CSTMIN[AGRP[I]]:=CST
                             END
                        END;
                %
                % CALCULO DA NOVA FUNCAO OBJETIVO E ARMAZENAMENTO
                % DA FUNCAO OBJETIVO DA ITERACAO ANTERIOR
                %
```

```
        FA:=FO;
        FO:=0;
        FOR   I:=1
        STEP  1
        UNTIL M
        DO    FO:=*+CSTMIN[I];
      %
      % IMPRESSAO DO VALOR DA FUNCAO OBJETIVO
      %
        IF    TIMP=2 OR TIMP=4 OR TIMP=5
        THEN WRITE(OTPT,</,T17,I3,T26,E12.5>,ITER,FO);
      %
      % CALCULO DA DIFERENCA PERCENTUAL ENTRE DOIS VALORES
      % SUCESSIVOS DA FUNCAO OBJETIVO
      %
        IF    ITER<1
        THEN FA:=FO+1;
        IF    ABS(FA)<TOL
        THEN PREC:=0
        ELSE PREC:=ABS(FA-FO)/FA;
      END DA PROCEDURE MDNFO;

%     ********************************
      PROCEDURE RLOC(N,M,MDN,E,AGRP);
%               ====
%     ESTA ROTINA CALCULA UM NOVO GRUPAMENTO AGRP[I],A PARTIR
%     DE UM CONJUNTO DE MEDIANAS MDN[I]
      INTEGER N,                %NUMERO DE ELEMENTOS
              M;                %NUMERO DE GRUPOS
      INTEGER ARRAY MDN[1], %VETOR DE MEDIANAS
                    AGRP[1];%AGRUPAMENTO EFETUADO
      REAL ARRAY E[1];        %VETOR DE MEDIANAS
%     ********************************
        BEGIN
          INTEGER I,J,K,L;

          REAL MINIMO;
          LABEL SEG;
        %
        % ALOCACAO DE CADA ELEMENTO AO GRUPO CUJA
        % MEDIANA ESTA MAIS PROXIMA
        %
          FOR   I:=1
          STEP  1
          UNTIL N
          DO    BEGIN
                FOR   J:=1
                STEP  1
                UNTIL M
                DO    IF    I=MDN[J]
                      THEN GO TO SEG;
                FOR  J:=1
                STEP  1
                UNTIL M
                DO    BEGIN
                      K:=MDN[J];
                      IF    K>I
                      THEN L:=I+(K-1)*(K-2)/2
                      ELSE L:=K+(I-1)*(I-2)/2;
                      IF    J=1 OR E[L]<MINIMO
```

```
                                THEN BEGIN
                                     AGRP[I]:=J;
                                     MINIMO:=E[L]
                                     END
                                END;
                           SEG;
                           END
          END DA PROCEDURE RLOC;

     INTEGER ITER,
             I,J,K,L;
     REAL PREC,
          MINIMO;
     IF   TIMP=2 OR TIMP=4 OR TIMP=5
     THEN BEGIN
          WRITE(OTPT[SKIP 1]);
          WRITE(OTPT,<///,T10,"METODO: K-MEDIANAS",///,T10,
                     "EVOLUCAO DA FUNCAO OBJETIVO",///,T16,
                     "ITER",T31,"FO",/," ">)
          END;
% INICIALIZACAO
%
     ITER:=0;
     MDNFO(M,MDN,N,AGRP,E,FO,ITER,TIMP,PREC);
% CALCULO DOS GRUPAMENTOS E MEDIANAS
%
     FOR  ITER:=1
     STEP 1
     WHILE ITER<=NITER AND PREC>=0.01
     DO   BEGIN
          RLOC(N,M,MDN,E,AGRP);
          MDNFO(M,MDN,N,AGRP,E,FO,ITER,TIMP,PREC);
          END;
%
% IMPRESSAO DOS RESULTADOS
%
          ITER:=ITER-1;
          WRITE(OTPT[SKIP 1]);
          WRITE(OTPT,<///,T10,"METODO: K-MEDIANAS",//,T12,
                     "NUMERO DE ITERACOES = ",I3,//,T12,
                     "FUNCAO OBJETIVO = ",E12.5,//,T12,
                     "PRECISAO = ",E12.5,
                     //,T12,"GRUPOS FORMADOS :",/," ">,ITER,

                     FO,PREC);
          FOR  J:=1
          STEP 1
          UNTIL M
          DO   BEGIN
               WRITE(OTPT,<//,T15,"GRUPO ",I3,X2,"MEDIANA :"
                          ,A6>,J,CDG[MDN[J]]);
               FOR  I:=1
               STEP 1
               UNTIL N
               DO   IF   AGRP[I]=J
                    THEN WRITE(OTPT,<T16,A6>,CDG[I]);
               END
END DA PROCEDURE KMDNS;

%     ************************************************
```

```
      PROCEDURE MLVCRD(N,M,AGRPS,MDNS,D,FOPS,NITER,TIMP,CDG);
%               =======
%     ESTA ROTINA EFETUA UMA ANALISE DE GRUPAMENTO DE N ELEMENTOS EM
%     M GRUPOS,ATRAVES DO METODO DE MULVEY E CROWDER,SENDO A PARTICAO
%     OBTIDA ARMAZENADA NO VETOR AGRPS[1:N] E AS MEDIANAS DOS GRUPOS
%     ARMAZENADAS,POR SUA VEZ,NO VETOR MDNS[1:M]
      INTEGER N,                %NUMERO DE ELEMENTOS
              M,                %NUMERO DE GRUPOS
              NITER,            %NUMERO MAXIMO DE ITERACOES
              TIMP;             %TIPO DE IMPRESSAO
      INTEGER ARRAY AGRPS[1],   %AGRUPAMENTO EFETUADO
                    MDNS[1];    %VETOR DE MEDIANAS
      REAL FOPS;                %VALOR DA FUNCAO OBJETIVO PRIMAL
      REAL ARRAY D[1],          %VETOR DE DISTANCIAS
                 CDG[1];        %VETOR DE CODIGOS DOS ELEMENTOS
%     ************************************************
      BEGIN
%           ***************************************
            PROCEDURE PSOPRM(N,M,MDN,AGRP,D,FOP);

%                     =======
%           ESTA ROTINA CALCULA UMA SOLUCAO PRIMAL AGRP[1:N],COM SEU
%           CORRESPONDENTE VALOR DE FUNCAO OBJETIVO FOP,A PARTIR  DE
%           UM CONJUNTO DE MEDIANAS MDN[1:M]
            INTEGER N,                %NUMERO DE ELEMENTOS
                    M;                %NUMERO DE GRUPOS
            INTEGER ARRAY MDN[1],     %VETOR DE MEDIANAS
                          AGRP[1];    %AGRUPAMENTO EFETUADO
            REAL ARRAY D[1];          %VETOR DE DISTANCIAS
            REAL FOP;                 %VALOR DA FUNCAO OBJETIVO
%           ***************************************
                  BEGIN
                        INTEGER I,J,K,L;

                        REAL MINIMO;
                        LABEL SEG;
                  %
                  % INICIALIZACAO
                  %
                        FOP:=0;
                        FOR   J:=1
                        STEP  1
                        UNTIL M
                        DO    AGRP[MDN[J]]:=J;
                  %
                  % CALCULO DA SOLUCAO PRIMAL E DA FUNCAO OBJETIVO
                  %
                        FOR   I:=1
                        STEP  1
                        UNTIL N
                        DO    BEGIN
                              FOR   J:=1
                              STEP  1
                              UNTIL M
                              DO    IF   I=MDN[J]
                                    THEN GO TO SEG;
                              FOR   J:=1
                              STEP  1
                              UNTIL M
                              DO    BEGIN
```

```
                    K:=MDN[J];
                    IF   K>I
                    THEN L:=I+(K-1)*(K-2)/2
                    ELSE L:=K+(I-1)*(I-2)/2;
                    IF   J=1 OR D[L]<MINIMO
                    THEN BEGIN
                         AGRP[I]:=J;
                         MINIMO:=D[L]
                         END
                    END;
               FOP:=X+MINIMO;
               SEG:
               END
     END DA PROCEDURE PSOPRM;

*****************************************************
PROCEDURE OTSBGR(N,M,AGRPS,MDN,D,V,ITER,LBDA,FOD,FOPS);
          =======
ESTA ROTINA CALCULA UMA SOLUCAO DUAL LBDA[1:N],NO METODO DE
MULVEY E CROWDER,BEM COMO O VALOR DA FUNCAO OBJETIVO DUAL
FOD E O VETOR DE MEDIANAS MDN[1:M]
INTEGER N,                  %NUMERO DE ELEMENTOS
        M,                  %NUMERO DE GRUPOS
        ITER;               %NUMERO DE ITERACOES
INTEGER ARRAY AGRPS[1],     % AGRUPAMENTOEFETUADO
              MDN[1];       %VETOR DE MEDIANAS
                            %VALOR DA FUNCAO OBJETIVO DUAL
REAL FOD,
                            %VALOR DA FUNCAO OBJETIVO PRIMAL
FOPS;
REAL ARRAY D[1],            %VETOR DE DISTANCIAS
          V[1],             %VETOR SUBGRADTENTE
          LBDA[1];          %VETOR DE VARIAVEIS DUAIS
*****************************************************
     BEGIN
          *********************************
          PROCEDURE QUICKSORT(A,ORD,M,N);

                    ==========
          ESTA ROTINA ORDENA UM CONJUNTO A[M:N]
          PELO METODO QUICKSORT ARMAZENANDO OS
          APONTADORES DOS ELEMENTOS ORDENADOS NO
          VETOR ORD[M:N]
          INTEGER M,N;              %INDICES AUXILIARES
          INTEGER ARRAY ORD[1];%VETOR DE APONTADORES
          REAL ARRAY A[1];          %VETOR A SER ORDENADO
          *********************************
               BEGIN
                    INTEGER AUX,I,K;

                    BOOLEAN CHV;
                    LABEL SEG;
                  %
                  % RETORNO DOS ELEMENTOS ORDENADOS
                  %
                    IF   N<=M
                    THEN GO TO SEG;
                    IF   N=M+1
                    THEN BEGIN
                         IF   A[ORD[M]]>A[ORD[N]]
                         THEN BEGIN
                              AUX:=ORD[N];
```

```
                    ORD[N]:=ORD[M];
                    ORD[M]:=AUX
                    END;
                GO TO SEG
                END;
        %
        % PARTICAO DE A[M:N] PARA A ORDENACAO
        %
          K:=M+ENTIER((N-M)/2);
          CHV:=TRUE;
          WHILE CHV
          DO    BEGIN
                CHV:=FALSE;
                FOR   I:=M
                STEP  1
                UNTIL K-1
                DO    IF    A[ORD[I]]>A[ORD[K]]
                      THEN BEGIN
                           AUX:=ORD[K];
                           ORD[K]:=ORD[I];
                           ORD[I]:=AUX;
                           CHV:=TRUE
                           END;
                FOR   I:=K+1
                STEP  1
                UNTIL N
                DO    IF    A[ORD[I]]<A[ORD[K]]
                      THEN BEGIN
                           AUX:=ORD[K];
                           ORD[K]:=ORD[I];
                           ORD[I]:=AUX;
                           CHV:=TRUE
                           END
                END;
          QUICKSORT(A,ORD,M,K-1);
          QUICKSORT(A,ORD,K+1,N);
          SEG:
      END DA PROCEDURE QUICKSORT;

  REAL ARRAY S[1:N],     %VETOR DE CUSTOS REDUZIDOS
             VA[1:N];    %VET. SUBGRAD. ITER. ANTERIOR
  REAL AUX,              %VAR. AUXILIAR
       T;                %PASSO DO ALGOR. SUBGRADIENTE
  INTEGER I,J,K,L,ITER; %VARIAVEIS
  BOOLEAN CHV;           %
  LABEL SEG;             %AUXILIARES
  BOOLEAN ARRAY MG[1:N];%VETOR QUE INDICA SE O ELEMEN
                         %TO [I] E MEDIANA OU NAO
%
%INICIALIZACAO: CALCULO DOS CUSTOS REDUZIDOS S[J]
%
  FOR   J:=1
  STEP  1
  UNTIL N
  DO    BEGIN
        AUX:=-LBDA[J];
        S[J]:=MIN(AUX,0);
        FOR I:=1   STEP 1 UNTIL J-1,
               J+1 STEP 1 UNTIL N
        DO  BEGIN
```

```
                IF    I<J
                THEN L:=I+(J-1)*(J-2)/2
                ELSE L:=J+(I-1)*(I-2)/2;
                AUX:=D[L]-LBDA[I];
                S[J]:=**MIN(AUX,0)
                END
          END;
  BEGIN
%
%INICIALIZACAO: ORDENACAO DOS S[J]
%
        INTEGER ARRAY ORD[1:N];

        FOR   J:=1
        STEP  1
        UNTIL N
        DO    ORD[J]:=J;
        J:=1;
        QUICKSORT(S,ORD,J,N);
%
%CALCULO DAS MEDIANAS MDN[I],ATRAVES DA PESQUISA
%PRIMAL
%
        FOR   I:=1
        STEP  1
        UNTIL M
        DO    MDN[I]:=ORD[I];
        FOR   I:=1
        STEP  1
        UNTIL M-1
        DO    FOR   J:=I+1
              STEP  1
              UNTIL M
              DO    IF    ABS(S[ORD[I]]-S[ORD[J]])<0.001
                        AND AGRPS[ORD[I]]=AGRPS[ORD[J]]
                    THEN BEGIN
                         MDN[J]:=ORD[M+1];
                         K:=ORD[J];
                         ORD[J]:=ORD[M+1];
                         FOR   L:=M+1
                         STEP  1
                         UNTIL N-1
                         DO    ORD[L]:=ORD[L+1];
                         ORD[N]:=K
                         END

  END;

%
%CALCULO DA FUNCAO OBJETIVO DUAL FOD E DO SUBGRADIENTE
%V[I]
%
  FOD:=0;
  FOR   I:=1
  STEP  1
  UNTIL N
  DO    BEGIN
        VA[I]:=V[I];
        MG[I]:=FALSE;
        FOD:=**LBDA[I];
        V[I]:=1;
```

```
                FOR   J:=1
                STEP  1
                UNTIL M
                DO    IF   MDN[J]=I
                      THEN BEGIN
                           V[I]:=*=1;
                           FOD:=*=LBDA[I];
                           MG[I]:=TRUE;
                           GO TO SEG
                           END;
                SEG:
                END;
          FOR   I:=1
          STEP  1
          UNTIL N
          DO    IF   ¬MG[I]
                THEN BEGIN
                     FOR   J:=1
                     STEP  1
                     UNTIL M
                     DO    BEGIN
                           K:=MDN[J];
                           IF   I<K
                           THEN L:=I+(K=1)*(K=2)/2
                           ELSE L:=K+(I-1)*(I-2)/2;
                           IF   D[L]=LBDA[I]<=0,001
                           THEN BEGIN
                                FOD:=*+D[L]=LBDA[I];
                                V[I]:=*=1
                                END
                           END
                     END;
        %CALCULO DO PASSO T,DO ALGORITMO,E IMPLANTACAO DO
        %FATOR DE CORRECAO DE DIRECAO PARA O SUBGRADIENTE
        %DE 0,6 (VER MULVEY E CROWDER = REFERENCIA BIBLIOGRA=
        %FICA 29,PAG.334)
        %
          AUX:=0;
          FOR   I:=1
          STEP  1
          UNTIL N
          DO    BEGIN
                V[I]:=*+0,6*VA[I];
                AUX:=*+V[I]**2
                END;
          T:=(FOPS-FOD)/AUX;
        %
        %CALCULO DA VARIAVEL DUAL LBDA[I] DA ITERACAO SEGUINTE
        %
          FOR   I:=1
          STEP  1
          UNTIL N
          DO    LBDA[I]:=*+T*V[I]
     END DA PROCEDURE OTSBGR;

INTEGER I,J,K,L,          %VARIAVEIS AUXILIARES
        ITER;             %NUMERO DE ITERACOES
INTEGER ARRAY MDN[1:M], %VETOR DE MEDIANAS
              AGRP[1:N];%AGRUPAMENTO EFETUADO
REAL PREC,                %DIF. ENTRE SOL. PRIMAL E DUAL
```

```
      PRECPERC,            %DIF. PREC EM TERMOS PERCENTUAIS
      TOL,                 %TOLERANCIA
      FOP,                 %FUNCAO OBJETIVO PRIMAL
      FOD;                 %FUNCAO OBJETIVO DUAL
  REAL ARRAY LBDA[1:N],    %VARIAVEL DUAL
       V[1:N];             %SUBGRADIENTE
  LABEL SEG;               %VARIAVEL AUXILIAR
%
% INICIALIZACAO
%
  TOL:=0.000000001;
  IF   TIMP=3 OR TIMP=4 OR TIMP=5
  THEN BEGIN
       WRITE(OTPT[SKIP 1]);
       WRITE(OTPT,<////, T10,"METODO: MULVEY + CROWDER",//,
             T3,"ITER",T10,"FOPRIMAL",T24,"POSOLUCAO",T38,
             "FODUAL",T50,"PRECISAO",T62,"PRECISAO %",//," ">)
       END;
  ITER:=1;
  FOR  I:=1
  STEP 1
  UNTIL N
  DO   MDN[I]:=MDNS[I];
  FOR  I:=1
  STEP 1
  UNTIL N
  DO   AGRP[I]:=AGRPS[I];
  FOR  I:=1
  STEP 1
  UNTIL N
  DO   BEGIN
       FOR  J:=1
       STEP 1
       UNTIL M
       DO   IF   AGRP[I]=AGRP[MDN[J]] AND I^=MDN[J]
            THEN BEGIN
                 K:=MDN[J];
                 IF   I<K
                 THEN L:=I+(K-1)*(K-2)/2
                 ELSE L:=K+(I-1)*(I-2)/2;
                 LBDA[I]:=*+D[L]+1;
                 GO TO SEG
                 END;
       SEG:
       END;
%
% INICIALIZACAO: CALCULO DA SOLUCAO DUAL
%
  OTSGGR(N,M,AGRPS,MDN,D,V,ITER,LBDA,FOD,FOPS);
%
% INICIALIZACAO: TESTE DE OTIMALIDADE
%
  PREC:=FOPS-FOD;
  IF   ABS(FOPS)<TOL
  THEN PRECPERC:=0
  ELSE BEGIN;
       IF   ABS(FOD)<TOL
       THEN PRECPERC:=9.9999
       ELSE PRECPERC:=PREC/FOD;
       END;
```

```
  IF    TIMP=3 OR TIMP=4 OR TIMP=5
  THEN WRITE(OTPT,<T3,I3,T7,5(X1,E12.5)>,ITER,FOP,FOPS,FOD,
            PREC,PRECPERC);
%
%CALCULO DAS SUCESSIVAS SOLUCOES PRIMAIS E DUAIS,COM ARMAZENA-
%MENTO DA MELHOR SOLUCAO PRIMAL OBTIDA ATE A ITERACAO E CALCULO
%DO TESTE DE OTIMALIDADE
%
  FOR   ITER:=2
  STEP  1
  WHILE ABS(PRECPERC)>0.01 AND ITER<=NITER
  DO    BEGIN
        PSOPRM(N,M,MDN,AGRP,D,FOP);
        OTSBGR(N,M,AGRPS,MDN,D,V,ITER,LBDA,FOD,FOPS);
        IF    FOP<FOPS
        THEN BEGIN
             FOR   I:=1
             STEP  1
             UNTIL N
             DO    AGRPS[I]:=AGRP[I];
             FOR   I:=1
             STEP  1
             UNTIL M
             DO    MDNS[I]:=MDN[I];
             FOPS:=FOP
             END;
        PREC:=FOPS-FOD;
        IF    ABS(FOPS)<TOL
        THEN PRECPERC:=0
        ELSE BEGIN;
             IF    ABS(FOD)<TOL
             THEN PRECPERC:=9,9999
             ELSE PRECPERC:=PREC/FOD
             END;
        IF    TIMP=3 OR TIMP=4 OR TIMP=5
        THEN WRITE(OTPT,<T3,I3,T7,5(X1,E12.5)>,ITER,FOP,FOPS,
                   FOD,PREC,PRECPERC)
        END;
%
%IMPRESSAO DOS RESULTADOS
%
  ITER:=ITER-1;
  WRITE(OTPT[SKIP 1]);
  WRITE(OTPT,<///,T10,"METODO: MULVEY + CROWDER",//,T12,
        "NUMERO DE ITERACOES = ",I3,// T12,"FUNCAO OBJETIVO ",
        "PRIMAL =",E12.5,//,T12,"FUNCAO OBJETIVO DUAL =",
        E12.5,//,T12,"PRECISAO = ",E12.5,//,T12,"GRUPOS ",
        "FORMADOS:",/," ">,ITER,FOPS,FOD,PRECPERC);

  FOR   J:=1
  STEP  1
  UNTIL M
  DO    BEGIN
        WRITE(OTPT,<//,T15,"GRUPO ",I3,X2,"MEDIANA : ",A6>,J,
              CDG[MDNS[J]]);
        FOR   I:=1
        STEP  1
        UNTIL N
        DO    IF    AGRPS[I]=J
              THEN WRITE(OTPT,<T16,A6>,CDG[I])
```

```
            END
  END DA PROCEDURE MLVCRD;

  INTEGER MTD,    %VARIAVEL AUXILIAR
          P,      %NUMERO DE CARACTERISTICAS
          N,      %NUMERO DE ELEMENTOS
          TIMP,   %TIPO DE IMPRESSAO
          TOTGRP, %NUMERO DE GRUPOS
          NITER,  %NUMERO MAXIMO DE ITERACOES
          MTR,    %CHAVE PARA A ESCOLHA DA METRICA A SER UTILIZADA
          PDR,    %CHAVE PARA PADRONIZACAO DAS VARIAVEIS
          I,J,K;  %VARIAVEIS AUXILIARES
  REAL FO,        %VALOR DA FUNCAO OBJETIVO
       W;         %VARIAVEL AUXILIAR
%
%INTCIALIZACAO
%
  READ(INPT,<7I5>,P,N,TOTGRP,TIMP,NITER,MTR,PDR);
  BEGIN
       REAL ARRAY E[1:N*(N-1)/2],    %VETOR DE DISTANCIAS

                  CDG[1:N],          %VETOR DE CODIGOS DOS ELEMENTOS
                  X[1:N,1:P];        %VETOR DE OBSERVACOES
       INTEGER ARRAY METODO [0:7],   %VETOR DE METODOS ESCOLHIDOS
                     NIG[1:N],       %TAMANHO DOS GRUPOS
                     MDN[1:TOTGRP],  %VETOR DE MEDIANAS
                     AGRP[1:N];      %AGRUPAMENTO EFETUADO
     %
     % INICIALIZACAO
     %
       READ(INPT,<8I1>,METODO);
       FOR   I:=1
       STEP  1
       UNTIL N
       DO    BEGIN
             READ(INPT,<A6>,CDG[I]);
             READ(INPT,<5F10.2>,FOR J:=1 STEP 1 UNTIL P DO X[I,J])
             END;
     %
     % PADRONIZACAO DOS DADOS
     %
       IF    PDR=1
       THEN PDRZ(N,P,X);
     %
     % CALCULO DAS DISTANCIAS ENTRE OS ELEMENTOS
     %
       IF    MTR=1
       THEN ECLD(N,E,X,P);
       IF    MTR=2
       THEN MTCR(N,E,X,P);
     %
     % CALCULO DOS GRUPAMENTOS
     %
       FOR   MTD:=0
       STEP  1
       UNTIL 5
       DO    IF    METODO[MTD]=1
             THEN BEGIN
                  CLSTHRQ(N,MTD,TOTGRP,W,FO,AGRP,E,NIG,TIMP,CDG);
                  IF    METODO[6]=1
```

```
                THEN KMDNS(AGRP,N,TOTGRP,E,MDN,FO,NITER,TIMP,CDG);
                IF    METODO[7]=1
                THEN MLVCRO(N,TOTGRP,AGRP,MDN,E,FO,NITER,TIMP,CDG)
                END
        END
END.
```

A2.2 - Rotina CLSTPD

Essa rotina efetua a análise de grupamento através do método de Rao - Programação Dinâmica. Uma listagem da rotina é apresentada nas folhas que se seguem.

Os dados de entrada do programa devem ser organizados da seguinte forma:

— 1º cartão: informar o número máximo de grupos e o número de elementos, em dois campos consecutivos de 10 posições, iniciando-se na coluna 1. Os dados devem ser alinhados à direita no campo.

— Cartões seguintes: informar, usando um cartão para cada elemento, o código do elemento e o valor da observação para esse elemento. O campo de código tem 6 posições, iniciando-se na coluna 1, e o da informação tem 8 posições, das quais três correspondem a casas decimais.

Exemplo: Suponha m = 3, n = 5 e que as observações sejam

CÓDIGO	MEDIÇÃO
AA1	2,02
AB1	3,00
AA2	4,22
AB2	5,00
AC1	3,56

Esses dados seriam lidos na rotina CLSTPD da seguinte forma (Ø e ø representam, respectivamente, espaço em branco e zero)

ØØØØØØØØØ3ØØØØØØØØ5
ØØØAA1ØØØØ2ø2ø
ØØØAB1ØØØØ3øøø
ØØØAA2ØØØØ422ø
ØØØAB2ØØØØ5øøø
ØØØAC1ØØØØ356ø

Os resultados dessa rotina são apresentados da forma:

U(I,J) — decisão ótima do estado I, no estágio J; e

W(I,J) — custo associado a U(I,J) (custo acumulado da formação dos grupos G_J, ..., G_M).

C L S T P D
= = = = = =

```
%  ESTA ROTINA EFETUA UMA ANALISE DE AGRUPAMENTO
%  ATRAVES DO METODO DE RAO - PROGRAMACAO DINAMICA
%
BEGIN

     FILE INPT (KIND=READER),

          OTPT (KIND=PRINTER);

%    ***********************************************
     PROCEDURE QUICKSORT(A,ORD,M,N);
%              =========
%     ESTA ROTINA ORDENA UM CONJUNTO A[M:N],
%     PELO METODO QUICKSORT,ARMAZENANDO  OS
%     APONTADORES DOS ELEMENTOS ORDENADOS NO
%     VETOR ORD[M:N]
%
     INTEGER M,N;              %INDICES AUXILIARES
     INTEGER ARRAY ORD[1];     %VETOR DE APONTADORES
     REAL ARRAY A[1];          %VETOR A SER ORDENADO
%    ***********************************************
          BEGIN
               INTEGER AUX,I,K; % VARIAVEIS

               BOOLEAN CHV;      %
               LABEL SEG;        % AUXILIARES
          %
          % RETORNO DOS ELEMENTOS ORDENADOS
          %
               IF   N<=M
               THEN GO TO SEG;
               IF   N=M+1
               THEN BEGIN
                    IF   A[ORD[M]]>A[ORD[N]]
                    THEN BEGIN
                          AUX:=ORD[N];
                         ORD[N]:=ORD[M];
                         ORD[M]:=AUX
                          END;
                    GO TO SEG
                    END;
          %
          % PARTICAO DE A[M:N] PARA A ORDENACAO
          %
               K:=M+ENTIER((N-M)/2);
               CHV:=TRUE;
               WHILE CHV
               DO    BEGIN
                     CHV:=FALSE;
```

```
                    FOR    I:=M
                    STEP   1
                    UNTIL  K=1
                    DO     IF    A[ORD[I]]>A[ORD[K]]
                     THEN BEGIN
                           AUX:=ORD[K];
                           ORD[K]:=ORD[I];
                           ORD[I]:=AUX;
                           CHV:=TRUE
                           END;
                    FOR    I:=K+1
                    STEP   1
                    UNTIL  N
                    DO     IF    A[ORD[I]]<A[ORD[K]]
                           THEN BEGIN
                           AUX:=ORD[K];
                           ORD[K]:=ORD[I];
                           ORD[I]:=AUX;
                           CHV:=TRUE
                           END
                    END;
              QUICKSORT(A,ORD,M,K-1);
              QUICKSORT(A,ORD,K+1,N);
              SEG;
        END DA PROCEDURE QUICKSORT;

%
% INICIALIZACAO
%
  INTEGER M, % NUMERO MAXIMO DE GRUPOS
          N; % NUMERO DE ELEMENTOS
  READ(INPT,<2I10>,M,N);
  BEGIN
       REAL ARRAY X[1:N],              % VETOR DE OBSERVACOES

                  W[1:N,2:M],          % MATRIZ DE CUSTOS OTIMOS
                  COD[1:N];            % VETOR DE CODIGOS DOS ELEMENTOS
       INTEGER ARRAY ORD[1:N],         % VETOR DE INDICES PARA ORDENACAO
                     U[1:N,2:M];       % MATRIZ DE ESTRATEGIA OTIMA
       REAL MD,      % VAR. AUX. P/ CALCULO DE MEDIAS
            CSTDCS,  % VAR. AUX. P/ CALCULO DE CUSTO DE DECISOES
            W1,      % CUSTO ASSOCIADO A PARTICAO EM M GRUPOS
            WAC;     % VAR. AUX. P/ CALCULO DE CUSTO DE DECISOES
       INTEGER U1,   % LIDER DO GRUPO 2 NA PARTICAO EM M GRUPOS
               NJ,   % NUMERO DE ELEMENTOS DO GRUPO J
               I,J,IND,K; % CONTADORES AUXILIARES
%
% LEITURA DAS OBSERVACOES
%
  FOR    I:=1
  STEP   1
  UNTIL  N
  DO     READ(INPT,<A6,F8.3>,COD[I],X[I]);
%
% ORDENACAO DOS ELEMENTOS
%
  FOR    I:=1
  STEP   1
  UNTIL  N
  DO     ORD[I]:=I;
```

```
   I:=1;
   QUICKSORT(X,ORD,I,N);
%
% IMPRESSAO DOS DADOS ORDENADOS
%
   WRITE(OTPT[SKIP 1]);
   FOR    I:=1
   STEP   1
   UNTIL N
   DO     WRITE(OTPT,<///,T10,I3,T15,A6,T25,F8.3>,
                I,CDG[ORD[I]],X[ORD[I]]);
%
% INICIALIZACAO DO PROBLEMA
%
   FOR    I:=1
   STEP   1
   UNTIL N
   DO     FOR    J:=2
          STEP   1
          UNTIL M
          DO     BEGIN
                 U[I,J]:=0;
                 W[I,J]:=0
                 END;
   MD:=X[ORD[N]];
   NJ:=1;
%
% CALCULO DOS CUSTOS NO ULTIMO ESTAGIO
%
   FOR    I:=N-1
   STEP   -1
   UNTIL 1
   DO     BEGIN
          U[I,M]:=N+1;
          W[I,M]  :=   (NJ/(NJ+1))*(X[ORD[I]]-MD)**2+W[I+1,M];
          MD:=((NJ*MD)+X[ORD[I]])/(NJ+1);
          NJ:=*+1
          END;
%
% CALCULO DAS DECISOES OTIMAS NOS ESTAGIOS DE (M-1) A 2
%
   FOR    J:=M-1
   STEP   -1
   UNTIL 2
   DO     BEGIN
          IND:=N-(M-J);
          U[IND,J]:=IND+1;
          W[IND,J]:=W[IND+1,J+1];
          FOR     I:=IND-1
          STEP    -1
          UNTIL   1
          DO      BEGIN
                  U[I,J]:=I+1;
                  W[I,J]:=W[I+1,J+1];
                  MD:=X[ORD[I]];
                  NJ:=1;
                  SAC:=0;

                  FOR    K:=I+2
                  STEP   1
```

```
                UNTIL IND+1
                DO    BEGIN
                      WAC:=W+((NJ/(NJ+1))*(X[ORD[K-1]]-MD)**2);
                      CSTDCS:=WAC+W[K,J+1];
                      IF    CSTDCS<W[I,J]
                      THEN BEGIN
                           U[I,J]:=K;
                           W[I,J]:=CSTDCS
                           END;
                      MD:=((NJ*MD)+X[ORD[K-1]])/(NJ+1);
                      NJ:=W+1
                      END
                END;
        END;
%
% CALCULO DA DECISAO OTIMA NO ESTAGIO 1
%
   U1:=2;
   W1:=W[2,2];
   MD:=X[ORD[1]];
   NJ:=1;
   WAC:=0;
   FOR   K:=3
   STEP  1
   UNTIL IND
   DO    BEGIN
         WAC:=W+((NJ/(NJ+1))*(X[ORD[K-1]]-MD)**2);
         CSTDCS:=WAC+W[K,2];
         IF    CSTDCS<W1
         THEN BEGIN
              U1:=K;
              W1:=CSTDCS;
              END;
         MD:=((NJ*MD)+X[ORD[K-1]])/(NJ+1);
         NJ:=W+1
         END;
%
% IMPRESSAO DOS RESULTADOS
%
   WRITE(OTPT [SKIP 1]);
   WRITE(OTPT,<///,T10,"U1 = ",I4,T30,"W1 = ",E12.5>,U1,W1);
   FOR   I:=1
   STEP  1
   UNTIL N
   DO    FOR   J:=2
         STEP  1
         UNTIL N
         DO    WRITE(OTPT,<//,T10,"U[",I4,",",I2,"] = ",I4,T30,"W[",
                     I4,",",I2,"] = ",E12.5>,I,J,U[I,J],I,J,W[I,J])
   END
END.
```