REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE
e da Sociedade Brasileira de Estatística

PUBLICAÇÃO TRIMESTRAL

ISSN 0034-7175

Pedidos de assinatura e número avulso ou atrasado para:

Diretoria de Administração — Sergraf: Av. Brasil, 15.671 — Lucas — Rio de Janeiro — Brasil
CEP — 21241
Tel.: 391-7788

Livraria do IBGE: Av. Franklin Roosevelt, 146/loja — Centro — RJ — Brasil
CEP — 20021
Tel.: 220-9147 / 220-8163
DDD: 021

SUMÁRIO

ISSN 0034-7175

R. bras. Estat.	Rio de Janeiro	v. 43	n.º 172	p. 587 a 798	out./dez. 1982

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA — IBGE
Av. Franklin Roosevelt, 166 — Centro
20 021 — Rio de Janeiro, RJ — Brasil

ISSN 0034-7175

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística · — ano 1, n. 1(1940, jan./mar.)- . —

Rio de Janeiro : IBGE, 1940-

Trimestral.

Órgão oficial do IBGE e da Sociedade Brasileira de Estatística.

Continuação de : Revista de economia e estatística.

Índices : autor-assunto, v. 22-24(1961-1963) no v. 25, n. 1 ; v. 25-26 (1964-1965) no v. 27, n. 1 ; v. 27-28(1966-1967) no v. 29, n. 1 ; . Índices anuais de autor-assunto, publicados no n. 1 (jan./mar.) do ano seguinte (período 1968-1977, v. 29-38).

ISSN 0034-7175 = Revista brasileira de estatística.

1. Estatística — Periódicos. I. IBGE.

IBGE. Biblioteca Central
RJ-IBGE/81-41

CDU 31(05)

Impresso no Brasil/Printed in Brazil

ANÁLISE DE GRUPAMENTO*

Luiz Campos de Sá Lucas **

SUMÁRIO

1 — 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.

* Tese submetida ao corpo docente da Coordenação dos Programas de Pós-Graduação de Engenharia da Universidade Federal do Rio de Janeiro — COPPE-UFRJ, como parte dos requisitos necessários para obtenção do Grau de Mestre em Ciências — M. Sc. Análise de Grupamento (Rio de Janeiro) 1983. XI, 161 p., 29,7 cm.

** O autor expressa seus agradecimentos a Cláudio Bornstein, pela orientação da tese, assim como a Roberto Quintanilha, pelo apoio na elaboração dos programas; a Angela Maria Alonso e Elza Velozo de Almeida, pelo apoio na documentação da tese; a Maria da Glória Alves de Lima, Valéria de Sant'Ana e Marlene de Moraes, pela datilografia e a Geraldo Sonoda, pelos desenhos.

Evidentemente, tal definição de grupos envolve uma certa dose de arbitrariedade, o que pode implicar em que sejam feitas, às vezes, 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 (41) e Cooley & 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 (41) e Cooley & Lohnes (7).

No problema de *cluster analysis* (daqui por diante denominado análise de grupamento, a exemplo do que é efetuado em Pinho Gama (32)), 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, que o grau de associação seja elevado entre os membros de uma mesma categoria e, baixo, entre os elementos de categorias distintas.

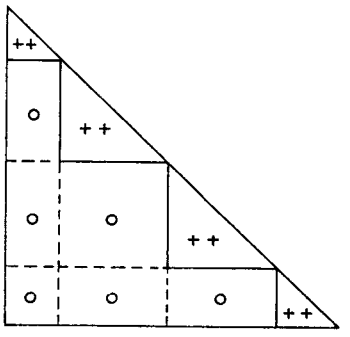
Assim, a análise de grupamento constitui-se em uma técnica a ser utilizada para a descoberta de uma estrutura de grupos e de relações entre esses grupos. Os resultados dessa análise podem contribuir para o desenvolvimento de esquemas censitários de classificação: em botânica

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 (18). 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 "O" 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 improvável 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 (42) 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 & Odell (10), Hartigan (19), Pinho Gama (32), 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 2 define o problema de análise de grupamento e os conceitos básicos que permitirão a apresentação dos métodos selecionados. No Capítulo 3 é feita uma descrição sucinta desses métodos (hierarquizados, de realocação iterativa e de programação matemática), cuja apresentação detalhada é feita nos Capítulos 4, 5 e 6, respectivamente. No Capítulo 7, 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 1 são apresentadas tabelas-resumo das distâncias, medidas de dispersão e funções objetivo utilizados pelos diversos métodos apresentados, sendo, por sua vez, apresentadas no apêndice 2 rotinas computacionais em Algol para a resolução de problemas práticos pelos principais métodos apresentados neste trabalho.

2 — DEFINIÇÃO DO PROBLEMA

2.1 — 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

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

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.

2.2 — 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 & 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 média do grupo g_j ao qual esse ponto fosse alocado. O objetivo então seria minimizar

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

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.

2.3 — 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 & Simon (9) e fundamentada no conceito de distância: um grupo g_s é dito homogêneo se para todos e_i e $e_j \in g_s$ e $e_k \notin g_s$, $d_{ij} \leqslant d_{ik}$ e $d_{ij} \leqslant d_{jk}$, sendo uma partição $P_m = \{g_1, g_2, \ldots, g_m\}$ dita homogênea se a propriedade acima for verdadeira para todo $g_s \in P_m$. 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.

2.3.1 — 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 $X_k \in R^p$, três propriedades são satisfeitas:

1 — $d_{ij} = 0$ se e somente se $X_i = X_j$

2 — $d_{ij} = d_{ji}$

3 — $d_{ij} \leqslant d_{ik} + d_{kj}$.

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 & Odell (10), Diday & Simon (9), Pinho Gama (32) e outros autores, apresentam uma série de funções de distância passíveis de uti-

lização. Dentre estas, foram selecionadas como de interesse para este trabalho as métricas de Minkowsky, ou l_p normas, que tomam a forma

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

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 $D^2 = [d_{ij}^2]$, 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 & 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 vezes 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 me-

dida 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 1 (ver exemplo Anderberg (1)):

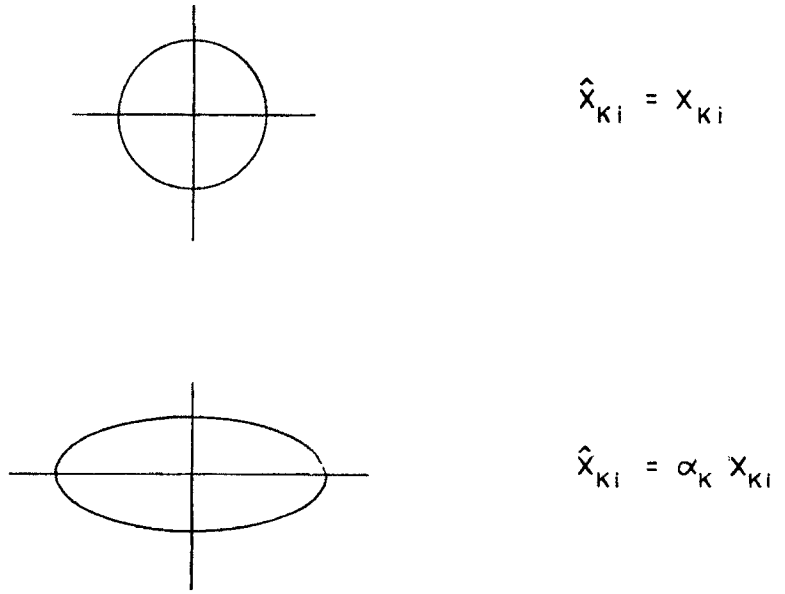


FIGURA 1

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 $d_e^2(\hat{X}_i, \hat{X}_j) = \sum_{k=1}^{p} (\alpha_k x_{ki} - \alpha_k x_{kj})^2 = \sum_{k=1}^{p} \alpha_k^2 (x_{ki} - x_{kj})^2$. 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 (ver por exemplo Anderberg (1), Cooley & Lohnes (7), Harman (18) ou Tatsuoka (41)).

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.

2.3.2 — Coeficientes de similaridade

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

1 — $0 \leqslant s_{ij} \leqslant 1$ se $X_i \neq X_j$

2 — $S_{ji} = 1$

3 — $S_{ij} = s_{ji}$.

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 por exemplo Gower (16) ou Pinho Gama (32)), e o coeficiente de Sokal & Michener (ver por exemplo Sokal & Michener (39) ou Duran & Odell (10)).

2.3.2.1 — 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:

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

onde

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

R_k = amplitude da variável k observada em E

$= máx_{X_r, X_s \in X} |x_{kr} - x_{ks}|$.

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, $0 \leqslant s_{ijk} \leqslant 1$.

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, $0 \leqslant s_{ij} \leqslant 1$.

2.3.2.2 — Coeficiente de Sokal & Michener

Quando todas as p características forem representadas por variáveis binárias, vários coeficientes de similaridade podem ser definidos (ver exemplo Duran & Odell (10) e Diday & Simon (9)). Um desses coeficientes é o de Sokal & 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, $0 \leqslant s_{ij} \leqslant 1$.

2.3.2.3 — Métricas e coeficientes de similaridade

É possível construir-se métricas a partir de coeficientes de similaridade, caso se efetuem as transformações adequadas. Gower (16) (ver também Pinho Gama (32)) sugere a seguinte transformação: $d_{ij} = (1 - s_{ij})^{1/2}$.

Esta função d_{ij} é uma métrica. Evidentemente, atende às propriedades 1 e 2 do item 2.3.1. No que se refere à propriedade 3, tem-se (ver por exemplo Everit (12)): $(1 - s_{ij})^{1/2} \leqslant (1 - s_{ik})^{1/2} + (1 - s_{kj})^{1/2}$.

2.3.3 — Coeficiente de correlação

O coeficiente de correlação entre X_i e X_j, denotado por r_{ij}, é definidc por (ver por exemplo Diday & Simon (9) ou Duran & 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

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

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

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

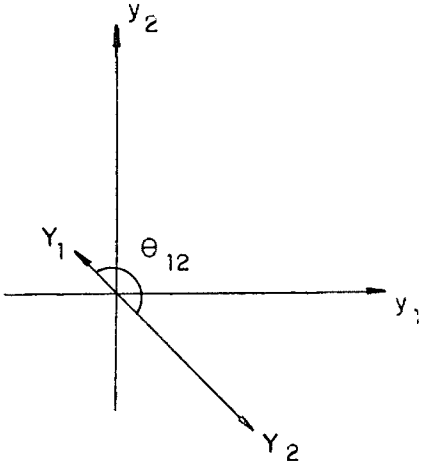
O coeficiente de correlação r_{ij} pode então ser reescrito da forma:

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

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

Como exemplo, sejam $X_1^t = [2\ 4]$ e $X_2^t = [7\ 3]$. Como $\bar{x}_{.1} = 3$ e $\bar{x}_{.2} = 5$, tem-se $Y_1^t = [-1\ 1]$ e $Y_2^t = [2\ -2]$. O coeficiente de correlação é então dado por $r_{ij} = \dfrac{-2\ -2}{\sqrt{2}\ \ \sqrt{8}} = \dfrac{-4}{4} = -\ 1{,}0.$

Graficamente, tem-se:



Se, por outro lado, tivermos $X_1^t = [2\ 4]$ e $X_2^t = [8\ 16]$, é fácil ver que $r_{ij} = 1$. De uma maneira geral (ver exemplo Duran & Odell (11)), se $X_j = \alpha\, X_i$, $\alpha > 0$, tem-se $r_{ij} = 1$. Como $-1 \leqslant r_{ij} = \cos\, \theta_{ij} \leqslant +1$, 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 $X_j = \alpha X_i$, $\alpha > 0$.

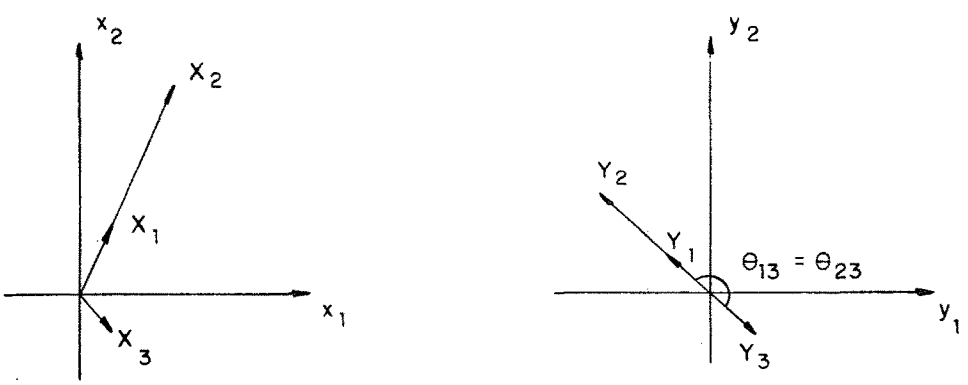
É possível também associar ao coeficiente de correlação uma função de distância. Mulvey & Crowder (31) definem uma "métrica" de correlação a partir da seguinte transformação: $d_{ij} = [0{,}5\ (1 - r_{ij})]^{1/2}$. Assim, quando dois elementos têm uma correlação positiva perfeita ($r_{ij} = +1$), tem-se $d_{ij} = 0$. 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 satisfaz à propriedade 1 do item 2.3.1, não sendo assim, a rigor, uma métrica.

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: $X = \{X_1, X_2, X_3\} = \{[{}^{1}_{3}], [{}^{4}_{12}], [{}^{1}_{-1}]\}$.

Como $\bar{x}^t_. = [2\ 8\ 0]$, tem-se: $Y = \{[{}^{-1}_{1}], [{}^{-4}_{4}], [{}^{1}_{-1}]\}$.

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:

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

$$G = \begin{bmatrix} 1 & 0,15 & 0,85 \\ 0,15 & 1 & 0,38 \\ 0,85 & 0,38 & 1 \end{bmatrix}, \text{ logo } D_G = \begin{bmatrix} 0 & 0,92 & 0,39 \\ 0,92 & 0 & 0,79 \\ 0,39 & 0,79 & 0 \end{bmatrix}$$

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

Assim, caso o objetivo fosse formar dois grupos homogêneos tal como definido por Diday & Simon (9) (ver início do Item 2.3), 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 $P_2 = \{g_1, g_2\} = \{\{1, 2\}, \{3\}\}$.

2.4 — 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 $W = \sum_{i \in g} (x_i - \overline{x})^2$, onde $\overline{x}$ fosse a média do grupo g, ter-se-ia $W_J = (1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$ e $W_K = 5\ x\ (2 - 2)^2 = 0$.

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.

2.4.1 — Soma dos quadrados dentro do grupo

Para o conjunto g_I, Duran & Odell (10) definem a medida W_I, soma dos quadrados dentro do grupo g_I, da forma $W_I = \sum_{i=1}^{n_I} (X_i - \overline{X}_I)^t (X_i - \overline{X}_I)$;

onde $\overline{X}_I = \frac{1}{n_I} \sum_{i=1}^{n_I} X_i$ 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 (ver por exemplo Duran & Odell (10)):

Teorema (2.1):

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

Demonstração:

Primeiramente, pode-se verificar que:

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

para tal, tem-se que, por definição,

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

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

Como

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

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

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

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

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

ou

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

ou ainda,

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

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:

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

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

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

Resumindo em um único somatório em i,

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

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

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

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

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

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

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

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

tem-se

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

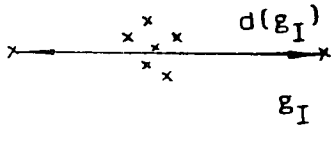
o que completa a demonstração do teorema.

2.4.2 — Variância interna de grupo

Com base na soma dos quadrados dentro do grupo, Duran & Odell (10) definem a variância interna do grupo g_I, denotada por S_I^2, como sendo $S_I^2 = \frac{1}{n_I} W_I$.

2.4.3 — Diâmetro de grupo

Hansen & Delattre (17) 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 2.3. Assim, o grupo apresentado abaixo teria, caso a medida fosse a distância euclidiana, o diâmetro indicado na figura abaixo:

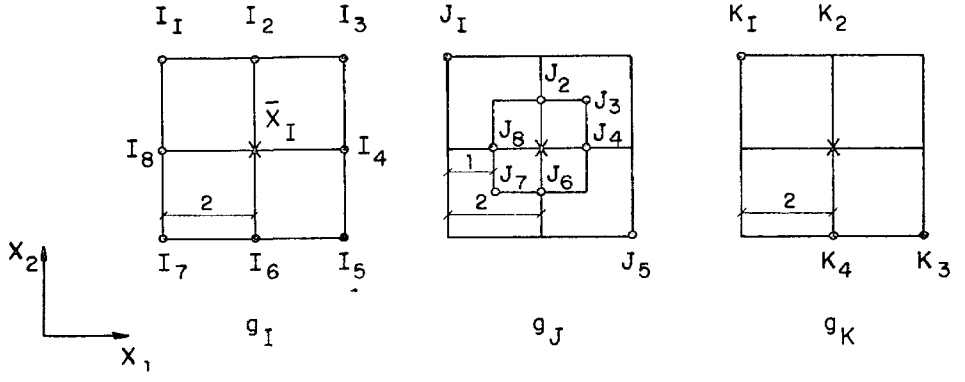


2.4.4 — Dispersão via mediana do grupo

Vinod (43), em um de seus modelos de análise de grupamento via programação inteira (ver Item 6.1) 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 $Z(e_r)$, isto é, $Z(e_r) = \sum_{i=1}^{n_I} d_{ir}$, é mínima. Nesse caso, $Z_{min}(g_I) = Z(e_r) = \min_{e_i \in g_I} Z(e_i)$ 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:

1 — 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;

2 — 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;

3 — 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;

4 — dispersão via mediana de grupo:

Utilizando a distância euclidiana,

— 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 $Z_{min} = 80$.

— 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 $Z_{min} = 32$.

— 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 $Z_{min} = 40$.

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 fez 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 $\overline{X}_J$ 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.

2.5 — Funções objetivo em análise de grupamento

Como foi visto no item 2.2, 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:

Otimizar $f(P_m)$

Sujeito a P_m viável.

A partir das medidas de dispersão apresentadas no item 2.4, é possível definir algumas funções objetivo passíveis de utilização, tais como as funções, a serem minimizadas, descritas a seguir:

1 — soma dos quadrados dentro dos grupos (ver por exemplo Duran & Odell (10)):

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

2 — soma das variâncias internas:

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

3 — diâmetro de grupo (ver por exemplo Hansen & Delattre (17)):

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

4 — dispersão via mediana de grupo (ver por exemplo Vinod (43)):

$$f(P_m) = \sum_{g_I \in P_m} Z_{min}(g_I) = \sum_{g_I \in P_m} \min_{e_j \in g_I} \sum_{i=1}^{n_I} d_{ij}.$$

3 — CONSIDERAÇÕES GERAIS SOBRE OS MÉTODOS DE ANÁLISE DE GRUPAMENTO

3.1 — 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 & Simon (9), Duran & Odell (10), Hartigan (19) e Pinho Gama (32).

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.

3.2 — 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 seqüencialmente 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 (32), 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 pertencam 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 (32) 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 4, 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*.

3.3 — 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 & Odell (10)), K — Médias ou de Macqueen (29), de Jancey (21) e de Forgy (15) (também descritos em Anderberg (1)).

Alguns dos principais métodos de realocação iterativa serão apresentados no Capítulo 5. 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.

3.4 — 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 item 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 & 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 (Bellman (6), Jensen (23) e Rao (35)), teoria dos grafos (Hansen & Delattre (17), Johnson (24), Rohlf (36), Sibson (38) e Zahn (48)) e programação inteira (Rao (35), Roy (37) e Vinod (43)).

Dentre essas técnicas foram selecionados para apresentação nesta tese dois modelos de programação inteira devidos a Vinod (43). Esses modelos apresentam a desvantagem inerente à técnica de programação matemática utilizada: excessivo tempo de computação. No entanto, Rao (35) e Mulvey & Crowder (31), 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 6.

4 — MÉTODOS HIERARQUIZADOS AGLOMERATIVOS

O procedimento padrão nos algoritmos hierarquizados aglomerativos foi descrito sucintamente no item 3.2. Ele consiste basicamente numa técnica iterativa que, em cada passo, reúne 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 2 (exemplo dado por Duran & Odell (10)), apresenta um dendograma para o caso de seis elementos e p características:

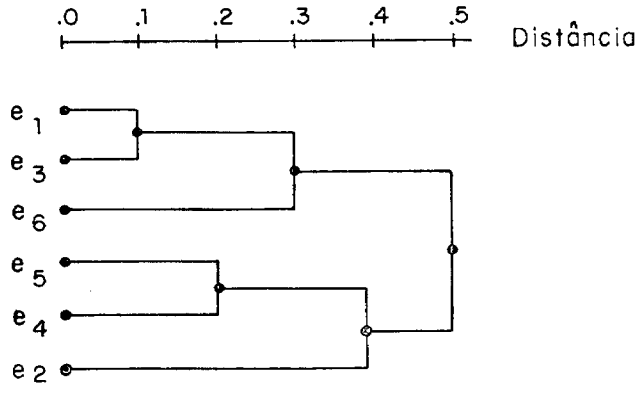


FIGURA 2 — *Exemplo de um dendograma*

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 $\{e_5, e_4\}$ U $\{e_2\}$ 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 2.3.1. Assim, é usual, nesses problemas, definir-se distâncias d^q_{IJ} entre os grupos quaisquer g_I e g_J, distintas das distâncias d_{ij} entre os elementos e_i e e_j. Da mesma forma, d^q_{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^q_{IJ} utilizada. No item 4.3, 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.

4.1 — 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 $g_L = g_I U g_J$, assume-se, para efeito de armazenamento das informações, que, após a união, tem-se $g_I = g_L$ e $g_J = \{\phi\}$.

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 $d^q_{IJ} = \underset{\substack{P \neq Q \\ P,Q \neq \{\phi\}}}{mín} \{d^q_{PQ}\}$.

Passo *2* : (união dos grupos para os quais a distância foi mínima)

Forme o grupo $g_L = g_I U g_J$;
Faça $g_I = g_L$ e $g_J = \{\phi\}$
(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*.

4.2 — Principais métodos hierarquizados aglomerativos

Os métodos a serem apresentados aqui são os seguintes:

— ligação simples;
— ligação completa;
— centróide;
— mediana;
— Ward; e
— média de grupo.

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

4.2.1 — Método da ligação simples

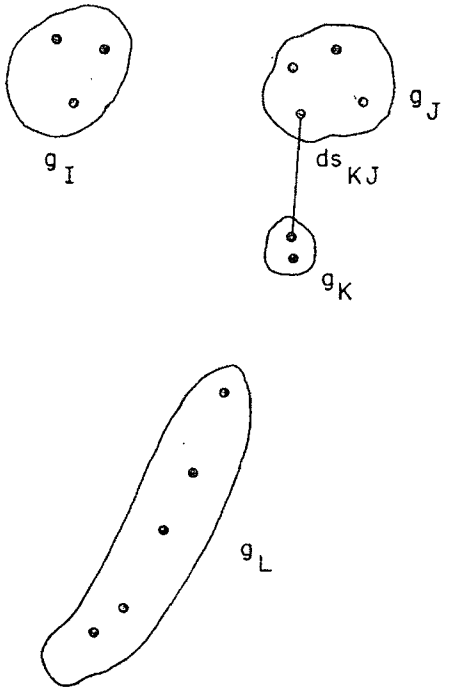
Introduzido por Johnson (24), esse método utiliza a distância "de vizinhança mais próxima" (ver também Duran & Odell (10)), aqui denotada por ds_{IJ}:

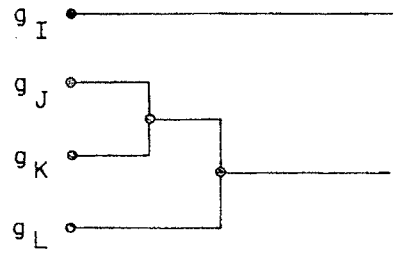
$$d^g_{IJ} = ds_{IJ} = \underset{\substack{e_i \in g_I \\ e_j \in g_J}}{mín}\ d_{ij} \ .$$

Pode-se, assim, nesse caso, fazer uso de quaisquer das distâncias entre elementos d_{ij} mencionadas no item 2.3.

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 2.2.3. 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 exemplo Diday & Simon (9), Duran & Odell (10) e Hansen & Delattre (17)).

4.2.2 — Método da ligação completa

Devido a Macnaughton-Smith (28), o método utiliza a distância "de vizinhança mais afastada" (ver por exemplo Duran & Odell (10)), aqui denotada por dc_{IJ}:

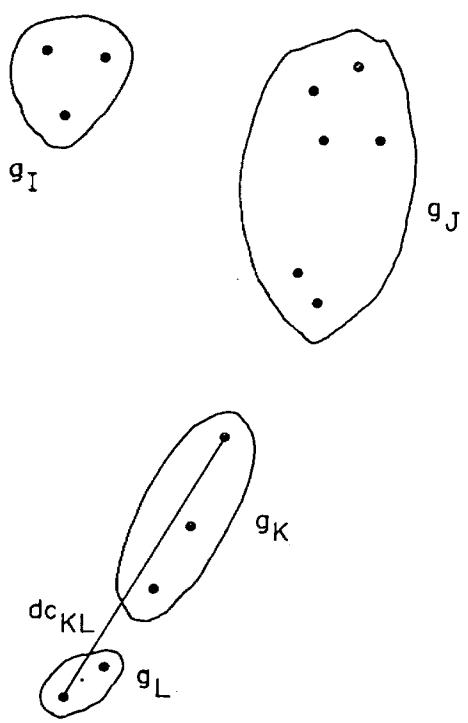
$$d^{g}_{IJ} = dc_{IJ} = \underset{\substack{e_i \in g_I \\ e_j \in g_J}}{máx}\ d_{ij} .$$

Também neste caso, pode-se fazer uso de quaisquer das distâncias mencionadas no item 2.3. Note-se que este método procura minimizar a função objetivo "diâmetro de grupo": $f(P_m) = d(P_m) = \underset{g_I \in P_m}{máx}\ d(g_I)$, descrita no item 2.5.

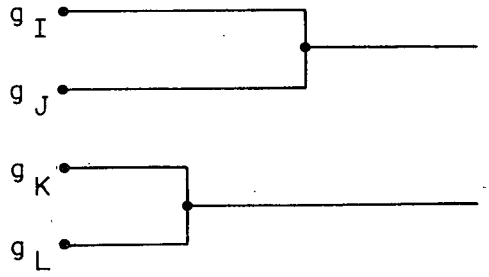
O método da ligação completa forma grupos mais compactos que o método anterior. Aplicado ao exemplo do item 4.2.1, ter-se-ia:

Exemplo:

A partição em quatro grupos (distância euclidiana) seria:



Seqüência de partições, para **m** = 2 :

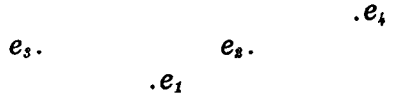


A solução agora é dada por $P_2 = \{g_I, g_J\}, \{g_K, g_L\}$.

Por outro lado, o método não fornece necessariamente soluções ótimas. O exemplo abaixo, citado por Hansen & Delattre (17), ilustra esse fato.

Exemplo:

Sejam quatro 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 $d(P_2') = d_{24} = 3$.

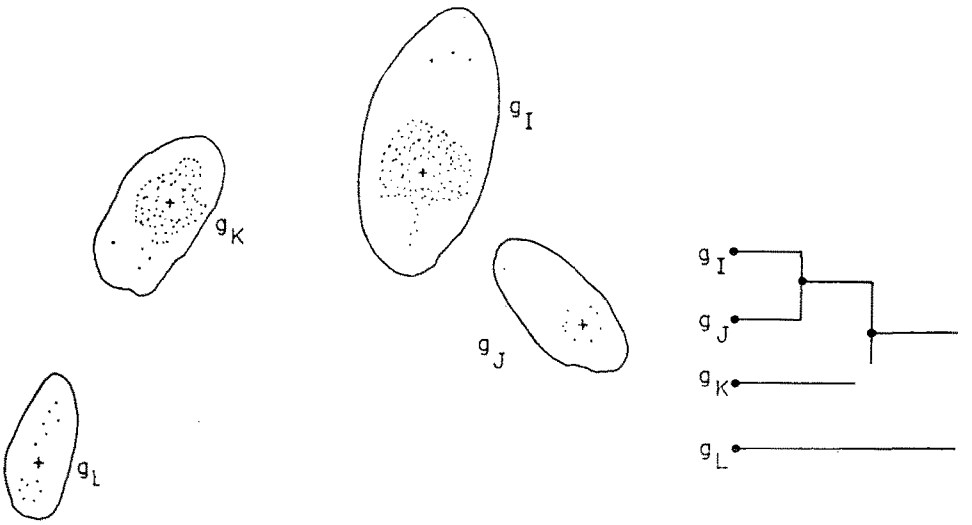
4.2.3 — Método da centróide

No método da centróide, devido a Sokal & Michener (39), a distância entre os grupos g_I e g_J é dada em termos do quadrado da distância euclidiana entre suas centróides (ver por exemplo Duran & Odell (10)). $d_{IJ}^1 = d_{IJ}^2 = d_2^2$ $(\overline{X}_I, \overline{X}_J)$, onde $\overline{X}_I$ e $\overline{X}_J$ são, respectivamente, as médias, ou centróides, dos grupos g_I e g_J. Cabe lembrar aqui que $\overline{X}_I = \frac{1}{n_I} \sum_{i=1}^{n_I} X_i$.

Note-se que se n_I for muito maior que n_J as centróides de $g_L = g_I U g_J$ 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" (ver por exemplo Duran & Odell (10)).

Exemplo:

Sejam seqüência de partições, caso se deseje dois grupos.



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

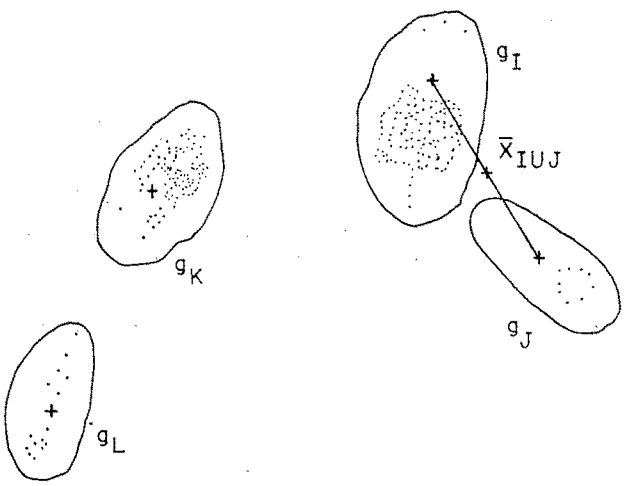
4.2.4 — Método da mediana

Como forma de contornar a dificuldade criada pela formação de grupos ponderados, Lance & Williams (27) 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 $g_M = g_I U g_J$, então $\overline{X}_M = \dfrac{\overline{X}_I + \overline{X}_J}{2}$.

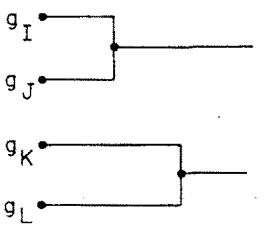
Em outras palavras, nota-se que $\overline{X}_M$ corresponde ao ponto médio do segmento que une $\overline{X}_I$ a $\overline{X}_J$.

Exemplo:

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



seqüência de partições:



4.2.5 — Método de Ward

O método de Ward (44) por sua vez, utiliza como distância entre os grupos a distância estatística (ver exemplo Duran & Odell (10)):

$$d_{IJ}^0 = D_{IJ} = \frac{n_I\, n_J}{n_I + n_J}\, d_2^2\,(\overline{X}_I,\ \overline{X}_J) = \frac{n_I\, n_J}{n_I + n_J}\, d_{IJ}^2 .$$

O objetivo do método de Ward é procurar a minimização da soma dos quadrados dentro dos grupos, definida no item 2.4.1:

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

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

Esse objetivo fica evidente em vista do teorema (4.1), apresentado a seguir, que indica o significado da distância D_{IJ}.

Teorema (4.1):

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

$$W_L = \sum_{i=1}^{n_I} d_2^2\,(X_i,\ M) + \sum_{j=1}^{n_J} d_2^2\,(X_j,\ M), \tag{4.1}$$

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:

1.ª parte:

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

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

Por definição,

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

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

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

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

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

onde

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

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

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

ou seja,

$$\phi = \sum_{k=1}^{p} \phi_k,$$

com

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

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

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

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

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

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

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

$$= W_I + n_I\ d_2^2 (\overline{X}_I, M). \tag{4.2}$$

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

$$\sum_{j=1}^{n_J} d_2^2 (X_j, M) = W_J + n_J\ d_2^2 (\overline{X}_J, M), \tag{4.3}$$

o que completa a primeira parte da demonstração.

2.ª parte:

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

Como foi indicado na equação (4.1),

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

Tendo em vista as equações (4.2) e (4.3), a equação acima pode ser reescrita:

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

Como

$$M = \frac{1}{n_I + n_J} (n_I \, \overline{X}_I + n_J \, \overline{X}_J),$$

tem-se que

$$\begin{aligned} n_I \, d_2^2 (\overline{X}_I, M) &= n_I \sum_{k=1}^{p} (\overline{x}_{kI} - m_k)^2 \\ &= n_I \sum_{k=1}^{p} \left(\overline{x}_{kI} - \frac{n_I \, \overline{x}_{kI} + n_J \, \overline{x}_{kJ}}{n_I + n_J} \right)^2 \\ &= n_I \sum_{k=1}^{p} \left(\frac{n_J \, \overline{x}_{kI} - n_J \, \overline{x}_{kJ}}{n_I + n_J} \right)^2 \\ &= \frac{n_I \, n_J^2}{(n_I + n_J)^2} \sum_{k=1}^{p} (\overline{x}_{kI} - \overline{x}_{kJ})^2 \\ &= \frac{n_I \, n_J^2}{(n_I + n_J)^2} \, d_{IJ}^2 \end{aligned}$$

e, da mesma forma,

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

Assim,

$$\begin{aligned} W_L &= W_I + W_J + \frac{n_I \, n_J^2 + n_I^2 \, n_J}{(n_I + n_J)^2} \, d_{IJ}^2 \\ &= W_I + W_J + \frac{n_I \, n_J \, (n_I + n_J)}{(n_I + n_J)^2} \, d_{IJ}^2 \\ &= W_I + W_J + \frac{n_I \, n_J}{n_I + n_J} \, d_{IJ}^2 . \end{aligned}$$

Finalmente,

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

o que completa a demonstração.

O resultado acima pode ser interpretado da seguinte forma (ver exemplo Duran & 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.

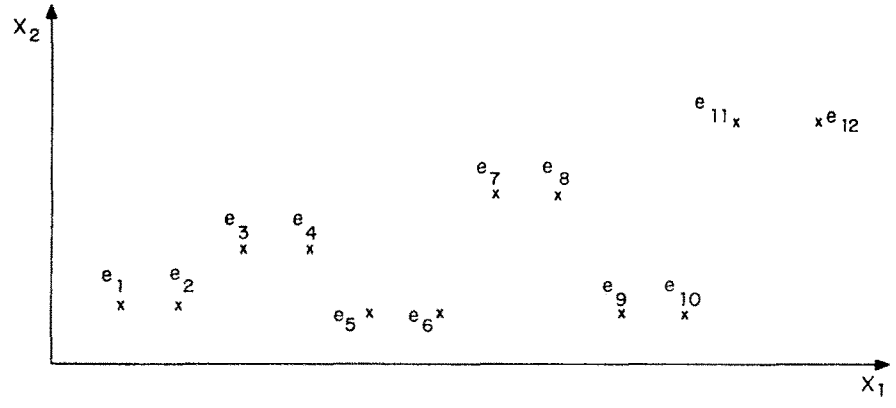
Assim, quando o método de Ward, em cada iteração, escolhe para a formação de um novo grupo os grupos g_I e g_J tais que D_{IJ} é mínimo, ele efetivamente está levando a que, em cada iteração, o acréscimo em W seja mínimo. Logo, o método procura minimizar a soma dos quadrados dentro dos grupos, embora,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

Exemplo:

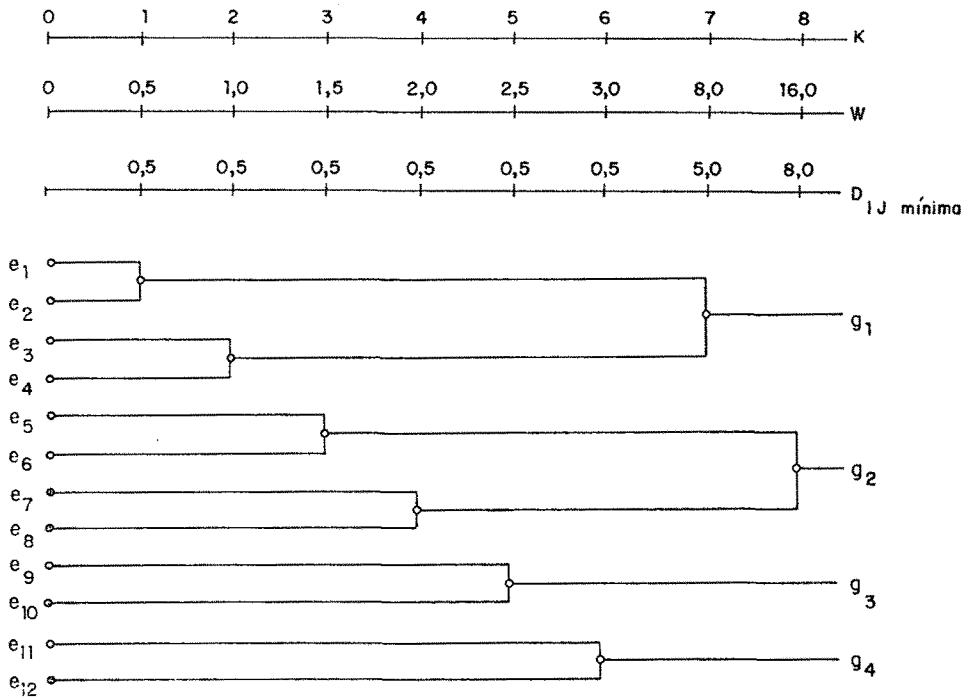
Sejam, por exemplo, $e_1, e_2, \ldots, e_{12}$ tais que:

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

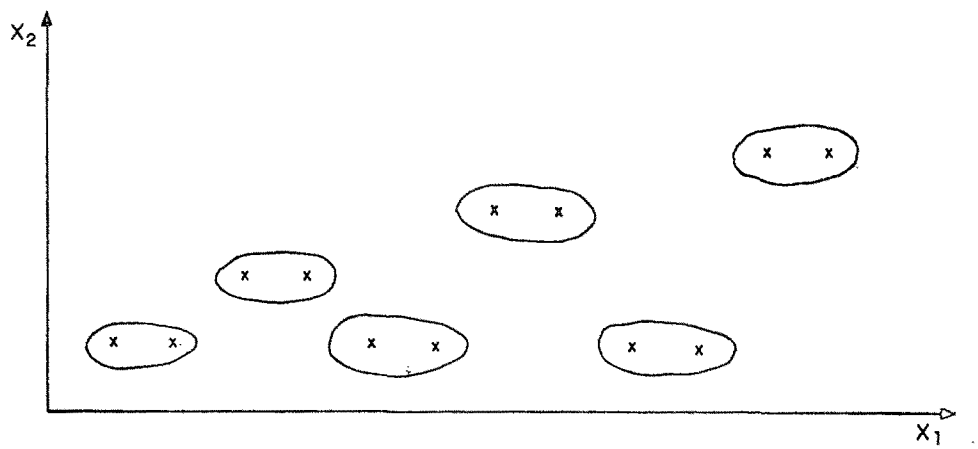
Graficamente, tem-se:



O problema foi resolvido com o auxílio da rotina CLSTROPT, desenvolvida nesta tese e aqui apresentada no apêndice 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:



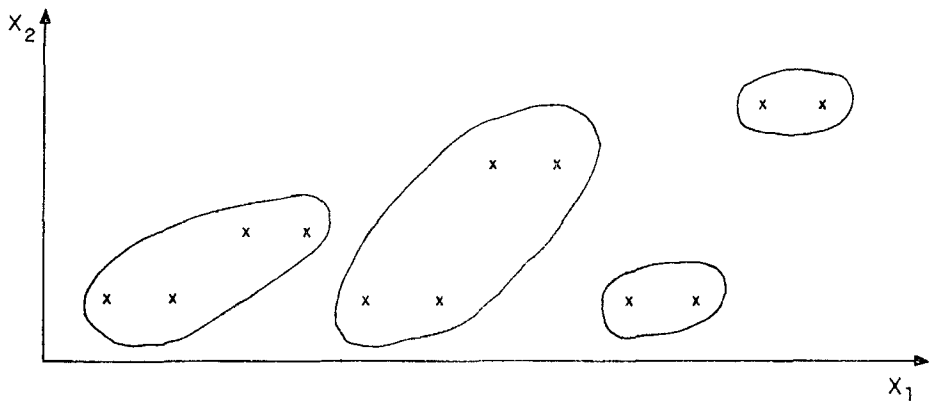
Assim, se fosse desejada uma partição em seis grupos, ter-se-ia:



— Note-se que, ao se formarem esses 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 vezes 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.

4.2.6 — Método da média de grupo

Devido a Sokal & Michener (39), o método da média de grupo utiliza como distância d_{IJ} entre os grupos g_I e g_J, a distância média quadrática (ver exemplo Duran & Odell (10)):

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

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

Teorema (4.2) (ver exemplo Duran & 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 $d_{IJ}^2 = d_2^2\,(\overline{X}_I,\, \overline{X}_J)$ entre as suas centróides.

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

Demonstração:

— o teorema será demonstrado em duas partes:

1.ª 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:

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

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 $\overline{X}_J$,

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

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

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

onde

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

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

$$\phi = \sum_{k=1}^{p} \left\{ \overline{x}_{kJ} \frac{1}{n_J} \sum_{j=1}^{n_J} x_{kj} - x_{ki} \frac{1}{n_J} \sum_{j=1}^{n_J} x_{kj} - \overline{x}^2_{kJ} + \overline{x}_{kJ}\ x_{ki} \right\}$$

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

Assim,

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

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

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

$$= S^2_J + d_2^2 (X_i, \overline{X}_J),$$

completando assim a demonstração da primeira parte do teorema.

2.ª parte:

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

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

$$D^2_{IJ} = \frac{1}{n_I} \sum_{i=1}^{n_I} \left\{ \frac{1}{n_J} \sum_{j=1}^{n_J} d^2_2 (X_i, X_j) \right\}.$$

Assim,

$$\begin{aligned} D^2_{IJ} &= \frac{1}{n_I} \sum_{i=1}^{n_I} D^2_{iJ} \\ &= \frac{1}{n_I} \sum_{i=1}^{n_I} \{S^2_J + d^2_2 (X_i, \overline{X}_J)\} \\ &= \frac{n_I}{n_I} S^2_J + \frac{1}{n_I} \sum_{i=1}^{n_I} d^2_2 (X_i, \overline{X}_J). \end{aligned} \tag{4.4}$$

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

$$D^2_{\overline{X}_J I} = \frac{1}{n_I} \sum_{i=1}^{n_I} d^2_2 (\overline{X}_J, X_i). \tag{4.5}$$

Como resultado da primeira parte da demonstração resulta que

$$D^2_{\overline{X}_J I} = S^2_I + d^2_2 (\overline{X}_J, \overline{X}_I). \tag{4.6}$$

A partir das equações (4.5) e (4.6) e lembrando que

$d^2_{IJ} = d^2_2 (\overline{X}_I, \overline{X}_J)$ tem-se então que

$$\frac{1}{n_I} \sum_{i=1}^{n_I} d^2_2 (X_i, \overline{X}_J) = S^2_I + d^2_{IJ}. \tag{4.7}$$

Finalmente, substituindo (4.7) em (4.4),

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

o que completa a demonstração do teorema.

Teorema (4.3):

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

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

Demonstração:

no item 4.2.5 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:

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

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

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

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

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

Em resumo, o teorema (4.2) 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:

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

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

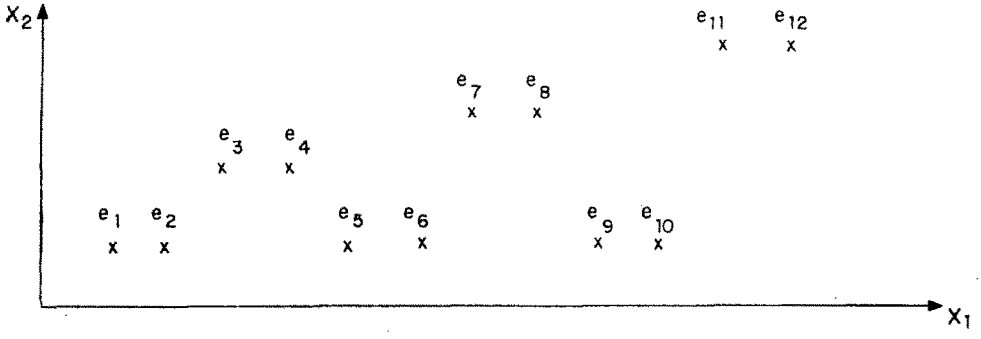
Assim, à primeira vista, poder-se-ia imaginar que o objetivo do método é a minimização da variância do grupo formado.

O teorema (4.3), 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 (4.3) indica é que $S_L^2 \geq \frac{n_I + n_J}{(n_I + n_J)^2} D_{IJ}^2$.

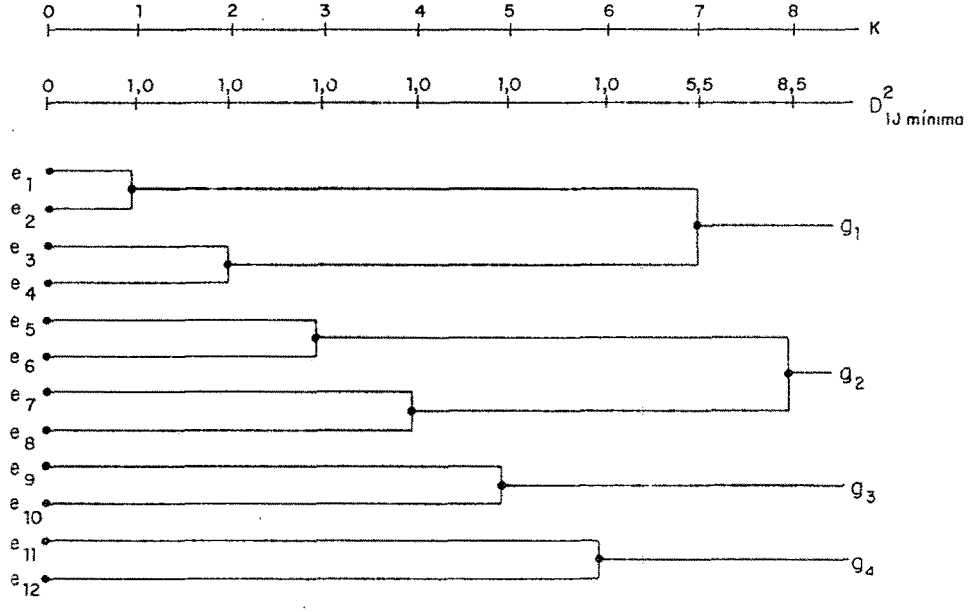
Logo, conhecidos n_I e n_J, D_{IJ}^2 fornece um limite inferior para a variância interna de g_L. Assim, não se pode afirmar que, em cada passo, ao se tomar D_{IJ}^2 mínimo se esteja obtendo S_L^2 mínimo. De uma maneira geral, o que se pode afirmar apenas é que o método tende a formar grupos compactos, 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 apêndice 2, para $m = 4$. Os resultados obtidos são sumarizados abaixo ($D^2_{IJ\,mínima}$ 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 $D^2_{IJ\,mínima}$ 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 (4.3), por sua vez, permite o cálculo da variância interna do grupo formado em cada iteração: se $g_L = g_I U g_J$, então

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

Assim, ao se unir e_1 e e_2 no passo $k = 1$, tem-se $n_1 = n_2 = 1$,

$$S_1^2 = S_2^2 = 0 \quad \text{e}$$

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

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

TABELA 1

VARIÂNCIAS INTERNAS DOS GRUPOS FORMADOS

PASSO	NÚMERO DE 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.

4.3 — Considerações gerais sobre os métodos abordados e apresentação de um método hierarquizado divisivo

A tabela 2, a seguir, resume as distâncias d_{IJ}^q, 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 4.2.4.

TABELA 2

DISTÂNCIAS ENTRE GRUPOS UTILIZADAS EM MÉTODOS HIERARQUIZADOS AGLOMERATIVOS

MÉTODO	DISTÂNCIA d_{IJ}^{2}
Ligação simples	$ds_{IJ} = \underset{\substack{e_i \in g_I \\ e_j \in g_J}}{\text{mín}}\ d_{ij}$
Ligação completa	$dc_{IJ} = \underset{\substack{e_i \in g_I \\ e_j \in g_J}}{\text{máx}}\ d_{ij}$
Centróide	$d_{IJ}^2 = d_2^2\ (\overline{X}_I, \overline{X}_J)$
Ward	$D_{IJ} = \Delta W = \frac{n_I\, n_J}{n_I + n_J}\ d_2^2\ (\overline{X}_I, \overline{X}_J)$
Média de grupo	$D_{IJ}^2 = S_I^2 + S_J^2 + d_2^2\ (\overline{X}_I, \overline{X}_J)$

No método da ligação simples, como visto no item 4.2.1, 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 & Simon (9)):

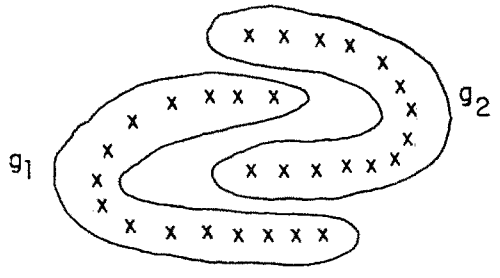


FIGURA 3 — *Grupamentos no método da ligação simples*

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 conseqüência, 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 4.2.2, tem como critério o diâmetro de grupo $d(P_m)$:

$$d(P_m) = \underset{g_I \in P_m}{\text{máx.}}\ d(g_I) = \underset{g_I \in P_m}{\text{máx.}}\ \underset{e_i, e_j \in g_I}{\text{máx.}}\ d_{ij} ;$$

No item 2.4, 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 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 (4.1) do Item 4.2.5), 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 2, é $D^2_{IJ} = S^2_I + S^2_J + d^2_2\ (\overline{X}_I, \overline{X}_J)$.

Como foi comentado no item 4.2.6, 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^2_I e S^2_J de cada um, mas também a distância entre suas centróides.

Por outro lado, nos itens 4.2.5 e 4.2.6 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, examinando-se a evolução dos valores de d^2_{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 4? Dois ou quatro grupos?

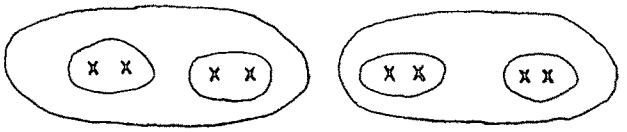


FIGURA 4 — *Grupos e subgrupos*

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:

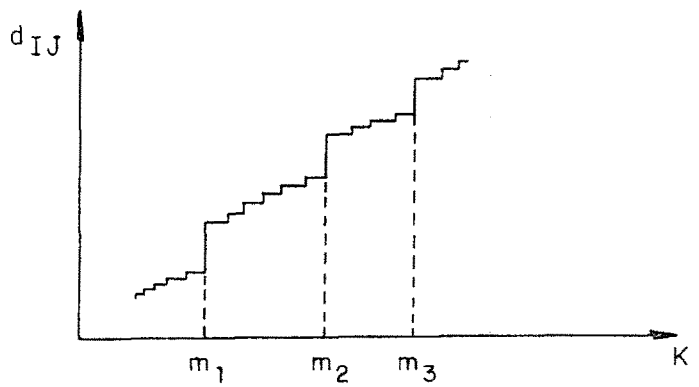


FIGURA 5 — *Evoluções de* d^g_{IJ} *no caso de grupos, subgrupos, etc.*

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 2, for inferior a uma fração α, definida *a priori*, do diâmetro $d(E)$ do conjunto E dos n elementos e_1, e_2, ..., 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 (4.3), 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 $g_L \varepsilon P_k$ tal que

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

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 4.2.5 e 4.2.6), que, nesse caso, corresponde a divisões, e não a aglomerações, como nos casos anteriores.

Exemplo

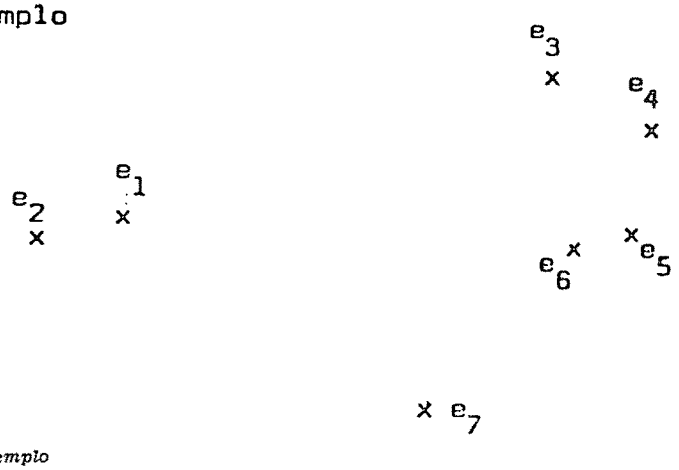


FIGURA 6 — *Exemplo*

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 7:

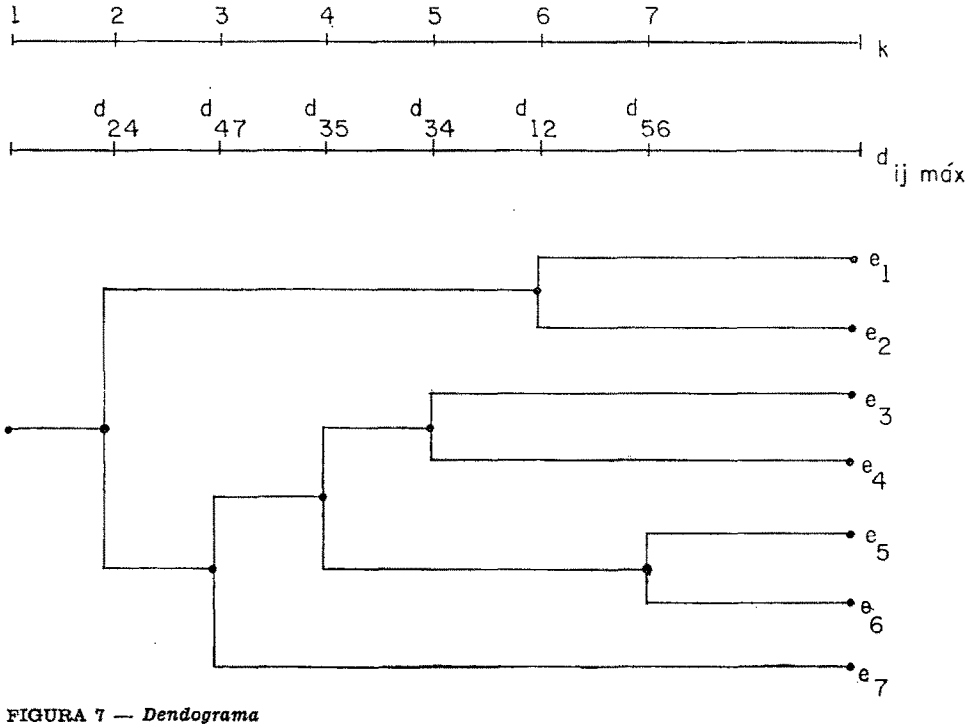


FIGURA 7 — *Dendograma*

4.4 — Algoritmo de Lance & Williams

Lance & Williams (26) e Wishart (46) (ver também Duran & Odell (10) e Diday & 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 $g_L = g_I U g_J$:

$$d^2_{KL} = \alpha_i \, d^2_{KI} + \alpha_j \, d^2_{KJ} + \beta \, d^2_{IJ} + \gamma \, | d^2_{KI} - d^2_{KJ} | ,$$

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

4.4.1 — 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:

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

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

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

4.4.2 — 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 $dc_{KI} < dc_{KJ} \Rightarrow dc_{KL} = dc_{KJ}$.

4.4.3 — Método da centróide

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

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

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

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

$$\gamma = 0.$$

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 (4.4):

Se $g_L = g_I U g_J$, então

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

Demonstração:

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

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

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

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

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

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

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

Como $n_L = n_I + n_J$,

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

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

Novamente, como $n_L = n_I + n_J$,

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

$$\left. + \frac{1}{n_L^2} (n_I^2\, \bar{x}_{rI}^2 + 2 n_I\, n_J\, \bar{x}_{rI}\, \bar{x}_{rJ} + n_J^2\, \bar{x}_{rJ}^2) \right]$$

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

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

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.

4.4.4 — Método da mediana

Como foi visto no item 4.2.4, 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 (4.4), que

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

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

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

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

$$\gamma = 0.$$

4.4.5 — Método de Ward

No caso do método de Ward tem-se:

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

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

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

$$\gamma = 0.$$

A equação básica fornece:

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

como é demonstrado abaixo:

Teorema (4.5):

Se $g_L = g_I U g_J$ 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 4.2.5):

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

tem-se:

$$d^2_{IJ} = \frac{n_I + n_J}{n_I n_J} D_{IJ}. \tag{4.4}$$

Por outro lado, pelo teorema (4.4) do item 4.4.3, se $g_L = g_I U g_J$ e g_K qualquer, tem-se:

$$d^2_{KL} = \frac{n_I}{n_L} d^2_{KI} + \frac{n_J}{n_L} d^2_{KJ} - \frac{n_I n_J}{n^2_L} d^2_{IJ}. \tag{4.5}$$

Pelas equações (4.4) e (4.5), tem-se então:

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

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.

4.4.6 — Método da média de grupo

Toma-se nesse caso:

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

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

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

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 (4.6):

Se $g_L = g_I U g_J$ e g_K é um outro grupo qualquer, então:

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

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.

4.4.7 — Conclusão

Em resumo, o método de Lance & 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 ajustados conforme o método hierarquizado aglomerativo a ser utilizado. Assim, a distância d^q_{KL} entre um grupo g_K e g_L, onde $g_L = g_I U g_J$, é dada por:

$$d^q_{KL} = \alpha_i \; d^q_{KI} + \alpha_j \; d^q_{KJ} + \beta \; d^q_{IJ} + \gamma \, | d^q_{KI} - d^q_{KJ} |,$$

cujos parâmetros são resumidos na tabela 3.

Então, no método de Lance & 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:

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

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,

TABELA 3

PARÂMETROS DO MÉTODO DE LANCE & 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

NOTA — $n_L = n_I + n_J$

nos seis métodos abordados, dada pela própria distância euclidiana entre esses elementos.

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

5 — MÉTODOS DE REALOCAÇÃO ITERATIVA

5.1 — Introdução

No item 3.3 foi descrito o procedimento geral de um método de realocação iterativa: começando-se com uma partição inicial p^o_m do conjunto E dos n elementos em m grupos, a idéia é gerar sucessivamente partições P^k_m tais que:

$$f(P^k_m) \leqslant f(P^{k-1}_m).$$

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.

5.2 — 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 & Williams (27), 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 (47) e Duran & 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 subseqüentes 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 & Hall (4):

- escolha da centróide de E como primeiro ponto semente;
- seleção subseqüente 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 à 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.

5.3 — 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 (15) (e sua variante de Jancey (21)), de Macqueen (29) (também chamado de K-Médias) e o método das K-Medianas (devido a Mulvey & Crowder (31)).

Os algoritmos de Forgy (15), Jancey (21) e Macqueen (29) (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 & Crowder, a ser apresentado no item 6.5.

O método de Forgy consiste na seguinte seqüência de passos (ver também algoritmo α em Diday & 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 seqüência 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 2.4, 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:

$$\phi(P_m, A_1, \ldots, A_m) = \sum_{j=1}^{m} \sum_{e_i \in g_j} d_2^2(X_i, A_j) \qquad (5.1)$$

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 $\phi(., A_1, \ldots, A_m)$. Assim, após o Passo 1, tem-se:

$$\phi(P_m^k, A_1^{k-1}, \ldots, A_m^{k-1}) \leq \phi(P_m^{k-1}, A_1^{k-1}, \ldots, A_m^{k-1}) = f(P_m^{k-1}) \qquad (5.2)$$

Por outro lado, derivando a expressão do lado direito da equação (5.1) 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:

$$\phi(P_m^k, A_1^k, \ldots, A_m^k) = f(P_m^k) \leq \phi(P_m^k, A_1^{k-1}, \ldots, A_m^{k-1}) \qquad (5.3)$$

Assim, por (5.2) e (5.3) tem-se:

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

A variante de Jancey (21) 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 8, baseada em Anderberg (1), ilustra esse procedimento:

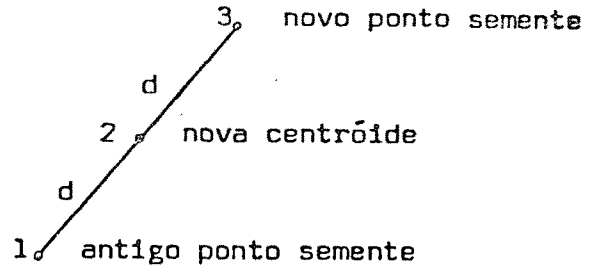


FIGURA 8 — *Determinação do novo ponto semente pelo método de Jancey*

Segundo Anderberg (1), o segmento que une os pontos 1 e 2 da figura 8 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, a medida que a solução viesse a ser obtida, os pontos sementes e as centróides tenderiam a se confundir: as distâncias *d*, da figura 8, 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 à 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 algoritmo 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 (29), 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:

$$f(P_m^1) \leq f(P_m^0). \tag{5.4}$$

Por efetuar apenas duas "iterações", este é o algoritmo mais rápido apresentado no presente estudo. Sua aplicação prática reside em que, como foi dito no início do item 5.1, as maiores mudanças em $f(P_m)$, geralmente ocorrem na primeira iteração. O comportamento do algoritmo depende evidentemente da seqüência 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 & Simon (9). O método K-Médias iterativo é convergente, tendo em vista e generalizando para qualquer iteração a equação (5.4). Evidentemente, os métodos K-Médias procuram minimizar a soma dos quadrados dentro dos grupos.

Como já foi dito anteriormente, dentre os métodos de realocação 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 & Crowder, aqui apresentado no Capítulo 6.

Desenvolvido por Mulvey & Crowder (31), 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 mediana 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 2.4, 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 seqüência:

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

Tomando agora:

$$\phi(v^k) = \phi(P_m^k, L_m^k) = \sum_{j=1}^{m} \sum_{e_i \in Q_j^k} d(X_i, A_j^k) \qquad (5.5)$$

tem-se o seguinte teorema:

Teorema (5.1): a seqüência $\{\phi(v^k)\} = \{f(P_m^k)\}$ é monotonamente decrescente e convergente.

Demonstração:

1 — o conjunto $S(n, m)$ de partições distintas P_m, dos n elementos de E em m grupos, é, como foi visto no item 3.4, finito. Assim, o conjunto $V = \{(P_m, L_m)\}$ também o é;

$$2 - \phi(v^k) = f(P_m^k) \geqslant \phi(v^{k+1}) = f(P_m^{k+1}): \qquad (5.6)$$

Dado L_m^k, a alocação feita no Passo 1, para $k \geqslant 1$, é a que minimiza $\phi(., L_m^k)$. 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:

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

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

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

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

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

O problema pode ser reescrito da forma:

$$\text{Minimizar} \quad \sum_{A_j^k \,\varepsilon\, L_m^k} d_{1j}\, x_j^1 + \ldots + \sum_{A_j^k \,\varepsilon\, L_m^k} d_{nj}\, x_{nj}.$$

$$\text{Sujeito a} \quad \sum_{A_j^k \,\varepsilon\, L_m^k} x_{1j} = 1$$

$$x_{1j} \;\varepsilon\; \{0,1\}$$

$$\vdots$$

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

$$x_{nj} \;\varepsilon\; \{0,1\}.$$

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

$$\text{Minimizar} \quad \sum_{A_j^k \,\varepsilon\, L_m^k} d_{ij}\, x_{ij}.$$

$$\text{Sujeito a} \quad \sum_{A_j^k \,\varepsilon\, L_m^k} x_{ij} = 1$$

$$x_{ij} \;\varepsilon\; \{0,1\}.$$

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:

$$\phi(P_m^k,\ L_m^k) \geq \phi(P_m^{k+1},\ L_m^k). \tag{5.7}$$

Por outro lado e por definição de mediana de grupo,

$$\phi(P_m^{k+1},\ L_m^k) \geq \phi(P_m^{k+1},\ L_m^{k+1}). \tag{5.8}$$

Pelas inequações (5.7) e (5.8), tem-se:

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

o que prova a afirmação (5.6), repetida abaixo:

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

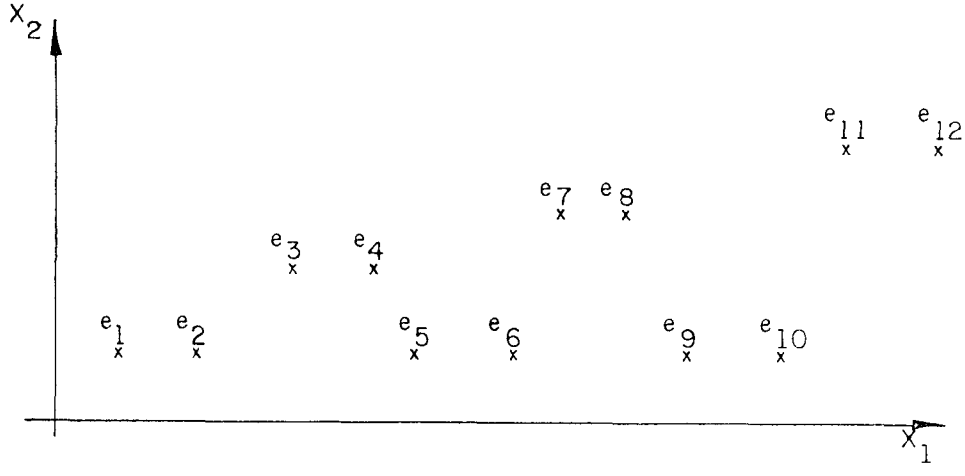
Assim, a seqüência $\{\phi(v^k)\}$ é monotonamente decrescente.

3 — $\{\phi(v^k)\}$ converge: se todos os v^k gerados pelo algoritmo forem distintos, tem-se uma seqüência 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 $f(P_m^k) \leqslant \leqslant f(P_m^{k-1})$, 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 4.2.5:

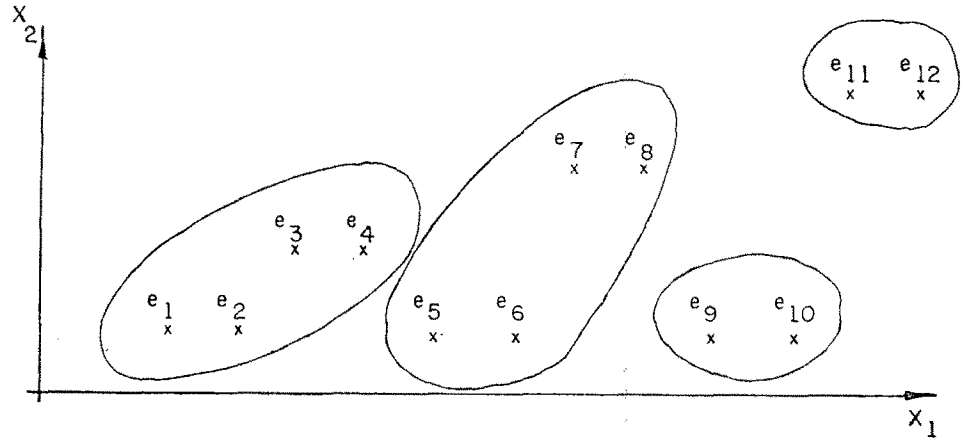
$$X = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 1 & 2 & 2 & 1 & 1 & 3 & 3 & 1 & 1 & 4 & 4 \end{bmatrix}$$



Este problema também foi resolvido com o auxílio da rotina CLSTROPT, apresentada no apêndice 2. Tomou-se $m = 4$ e adotou-se como solução inicial a partição gerada pelo método de Ward (ver Item 4.2.5):

$$g_1 = \{e_1, e_2, e_3, e_4\}$$
$$g_2 = \{e_5, e_6, e_7, e_8\}$$
$$g_3 = \{e_9, e_{10}\}$$
$$g_4 = \{e_{11}, e_{12}\}.$$

Graficamente,



O método convergiu em duas iterações, conforme ilustrado pela tabela 4:

TABELA 4

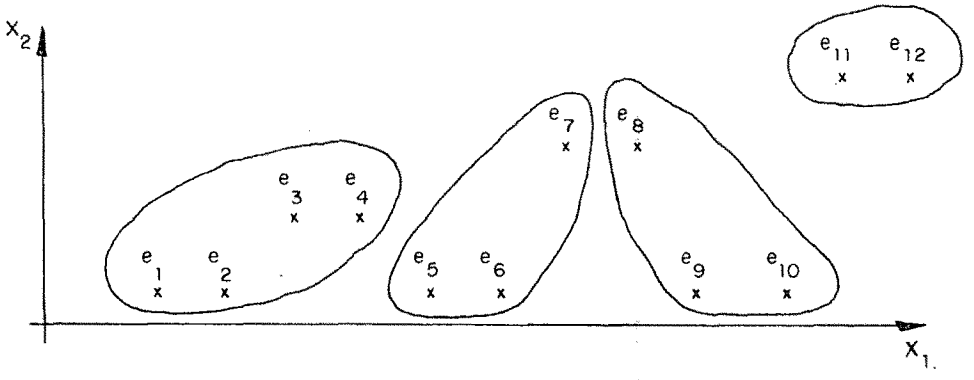
EVOLUÇÃO DA FUNÇÃO OBJETIVO — MÉTODO DAS *K*-MEDIANAS

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 & Crowder, no item 6.4.

5.4 — 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:

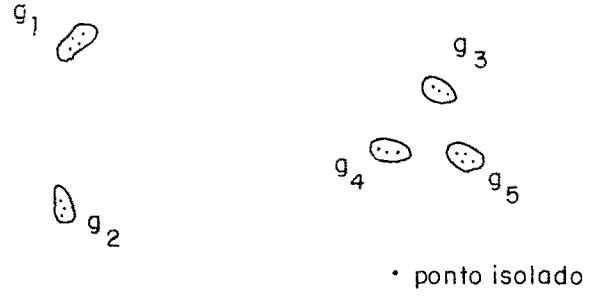


FIGURA 9 — *Configuração de um conjunto E*

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, significativa, 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 9, 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 (29) 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, reúna 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), reúna 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 5.3). 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 $e_1, e_2, \ldots, e_n$, 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 & 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 & 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 aleatoriamente 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 4.2.6) 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 2.4.2) é limitada superiormente:

$$S^2 \leqslant pv.$$

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 conseqüência, a S^2, então é possível que dois grupos fiquem sempre sendo reunidos e separados nos Passos 1 e 2.

O teorema (4.3), 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:

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

ou seja,

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

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

$$pv > \alpha\ R,\ o < \alpha < \infty.$$

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.

5.5 — 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 exemplo Duran & 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 5.1, 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 4.3, para os métodos hierarquizados.

6 — MÉTODOS DE PROGRAMAÇÃO MATEMÁTICA

6.1 — Introdução

Dentre os modelos citados no item 3.4 para a resolução do problema de análise de grupamento, foram selecionados para apresentação neste

trabalho, dois modelos de programação inteira propostos por Vinod (43). 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 6.5). Para esse modelo, Mulvey & Crowder (31) 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 6.3, apenas em sua versão univariada ($p = 1$). Para esse caso univariado, Rao (35) forneceu um algoritmo de programação dinâmica que permite uma solução eficiente para o problema. Embora Vinod (43) 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".

6.2 — Primeiro modelo de Vinod

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

$$\text{Minimizar} \quad \sum_{i \in I} \sum_{j \in J} d_{ij}\, y_{ij} \tag{6.2.1}$$

$$\text{Sujeito a} \quad \sum_{j \in J} y_{ij} = 1, \quad i \in I \tag{6.2.2}$$

$$\sum_{j \in J} y_{jj} = m, \tag{6.2.3}$$

$$y_{ij} \leqslant y_{jj}, \quad i \in I, \quad j \in J \tag{6.2.4}$$

$$y_{ij} \in \{0,1\}, \quad i \in I, \quad j \in J \tag{6.2.5}$$

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 2.4.4;

— 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 (6.2.2) garante que cada elemento só poderá ser alocado a um grupo;

— a restrição (6.2.3) impõe que o número de grupos seja igual a m; e

— a restrição (6.2.4) 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 realmente é mediana. Isso pode ser visto através da função objetivo (6.2.1). Seja $g_j = \{e_i \mid y_{ij} = 1\}$. Note-se que $e_j \in g_j$. Supondo, por absurdo, que a mediana do grupo fosse $e_k \in g_j$ e não e_j, ter-se-ia:

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

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 $e_k \in g_j$ 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 6.5, 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 & Crowder (31).

6.3 — Segundo modelo de Vinod — caso univariado

6.3.1 — Introdução

Seja $X = \{x, \ldots, x_n\}$, $x_i \in R$, 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 2.4.1):

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

onde

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

Como $x_i \,\varepsilon\, R$, tem-se:

$$W = \sum_{j=1}^{m} W_j = \sum_{j=1}^{m} \sum_{e_i \,\varepsilon\, g_j} (x_i - \bar{x}_j)^2 \qquad (6.3.1)$$

O segundo modelo de Vinod — caso univariado, procura resolver esse problema utilizando a propriedade da cadeia — caso univariado (ver por exemplo Vinod (43), Rao (35) ou Duran & 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:

$$z_1 \leqslant z_2 \leqslant \ldots \leqslant z_n.$$

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:

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

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

Uma discussão detalhada da propriedade da cadeia é apresentada em Vinod (43). 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, $z_i \leqslant z_{i+1}$.

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 (6.3.1)):

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

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 (4.1) da seção 4.2.5, que informa que, ao se unirem dois grupos g_I e g_J para se formar um grupo g_L, tem-se

$$W_L = W_I + W_J + \frac{n_I \, n_J}{n_I + n_J} d^2_{IJ} , \qquad (6.3.2)$$

onde d^2_{IJ} 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 (6.3.2):

$$g_I = \{e'_j\} \quad ; \; n_I = 1 \; ; \; W_I = 0$$
$$g_J = \{e'_{j+1}\} \; ; \; n_J = 1 \; ; \; W_J = 0$$
$$W_L = \frac{1}{2}(z_j - z_{j+1})^2 = \frac{1}{2}(z_{j+1} - z_j)^2.$$

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

$$g_I = \{e'_j \; e'_{j+1}\} \; ; \; n_I = 2 \; ; \; W_I = \frac{1}{2}(z_j - z_{j+1})^2$$
$$g_J = \{e'_{j+2}\} \; ; \; n_J = 1 \; ; \; W_J = 0$$
$$W_L = \frac{1}{2}(z_{j+1} - z_j)^2 + \frac{2}{3}\left(z_{j+2} - \frac{1}{2}\sum_{q=j}^{j+1} z_q\right)^2.$$

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 (6.3.2) ter-se-ia:

$$g_I = \{e'_j, \ldots, e'_{r-1}\} \; ; \; n_I = r - j \; ; \; W_I = \sum_{e'_i \, e \, g_I} (z_i - \bar{z}_I)^2$$
$$g_J = \{e'_r\} \; ; \; n_J = 1 \; ; \; W_J = 0$$

e, finalmente,

$$W_L = W_I + \Delta W_j^r = W_I + \frac{r-j}{(r-j)+1}\left(z_r - \frac{1}{r-j}\sum_{q=j}^{r-1} z_q\right)^2. \quad (6.3.3)$$

Assim, o cálculo de W_k se resume em utilizar a equação (6.3.3) para afirmar que, para o grupo $g_k = \{e'_j, \ldots, e'_i\}$, tem-se:

$$W_k = \sum_{r=j+1}^{i} \Delta W_j^r \quad (6.3.4)$$

6.3.2 — Descrição do modelo

Como, a partir da equação (6.3.3), 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:

$$\Delta W_j^i = \frac{i-j}{(i-j)+1}\left(z_i - \frac{1}{i-j}\sum_{q=j}^{i-1} z_q\right)^2, \quad (6.3.5)$$

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

$$\text{Minimizar} \quad \sum_{j \in J} \sum_{\substack{i \in I \\ i > j}} \Delta W_i^j \, y_{ij} \tag{6.3.6}$$

$$\text{Sujeito a} \quad \sum_{j \in J} y_{ij} = 1 \quad , \quad i \in I \tag{6.3.7}$$

$$\sum_{j \in J} y_{jj} = m \tag{6.3.8}$$

$$y_{jj} \geqslant y_{j+1,j} \geqslant \ldots \geqslant y_{nj}, \quad j \in J \tag{6.3.9}$$

$$y_{ij} = 0 \quad , \quad i < j \tag{6.3.10}$$

$$y_{ij} \in \{0,1\} \quad , \quad i \in I, \; j \in J \tag{6.3.11}$$

onde

I = conjunto dos n elementos;

J = conjunto dos líderes de grupo elegíveis ($J = I$);

ΔW_j^i = dado pela equação (6.3.5);

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 2.4.1, 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 grupos (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 (6.3.7) garante que cada elemento só poderá ser alocado a um grupo;

— a restrição (6.3.8) impõe a formação de m grupos;

— as restrições (6.3.9) e (6.3.10) correspondem à propriedade da cadeia. O que as restrições impõem é que y_{ij} só pode ser igual a 1 se $i \geqslant j$ e $y_{jj} = y_{j+1,j} = \ldots = y_{i-1,j} = 1$.

6.4 — Método de Mulvey & Crowder (31)

6.4.1 — Introdução

O primeiro modelo de Vinod (ver Item 6.2) é descrito como:

$$\text{Problema V1 : Minimizar} \quad \sum_{i \in I} \sum_{j \in J} d_{ij}\, y_{ij} \tag{V1.1}$$

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

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

$$y_{ij} \leqslant y_{jj} \quad , \quad i \in I \quad , \quad j \in J \tag{V1.4}$$

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

Esse problema pode ser reescrito da forma:

$$\text{Problema PV : Minimizar} \quad \sum_{i \in I} \sum_{j \in J} d_{ij}\, y_{ij} \tag{PV.1}$$

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

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

O dual Lagrangiano de PV (ver exemplo Bazaraa & Shetty (5) para uma apresentação de dualidade Lagrangiana) pode, por sua vez, ser descrito como:

$$\text{Problema DV : Maximizar} \quad \theta(u_1, u_2, \ldots, u_n)$$

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

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

A equivalência entre DV e um problema de otimização por subgradientes (para uma discussão detalhada de otimização por subgradientes, ver exemplo Held, Wolfe & Crowder (20)) 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 ($k = 1, \ldots, K$), DV pode ser reescrito como:

$$\text{Problema DV : Maximizar} \quad \theta(u_1, u_2, \ldots, u_n)$$

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

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

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

y_j^i = 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 seqüência de soluções duais $(u_1, u_2, \ldots, u_n)^1, \ldots, (u_1, u_2, \ldots, u_n)^*$, que iria convergir para uma solução ótima $(u_1, \ldots, u_n)^*$ de *DV*.

Pelo Teorema Fraco da Dualidade (ver exemplo Bazaraa & Shetty (5)), um valor de função objetivo associado a qualquer solução do problema dual *DV* é menor que o valor da função objetivo lgiada 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 & Shetty (5). Mulvey & Crowder (31), no entanto, informam que estudos empíricos (inclusive o de Held, Wolfe & Crowder (20)) evidenciaram que a utilidade do método 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 & Crowder (31) 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 seqüências 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 & Crowder (31):

— 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 dez problemas de diferentes portes, soluções ótimas foram obtidas até um máximo de 120 iterações).

6.4.2 — Descrição do método

O método de Mulvey & Crowder (31) se compõe basicamente de três partes:

1 — geração de uma boa solução viável inicial, via técnicas hierarquizadas, sendo a solução posteriormente refinada pelo método das *k*-Medianas;

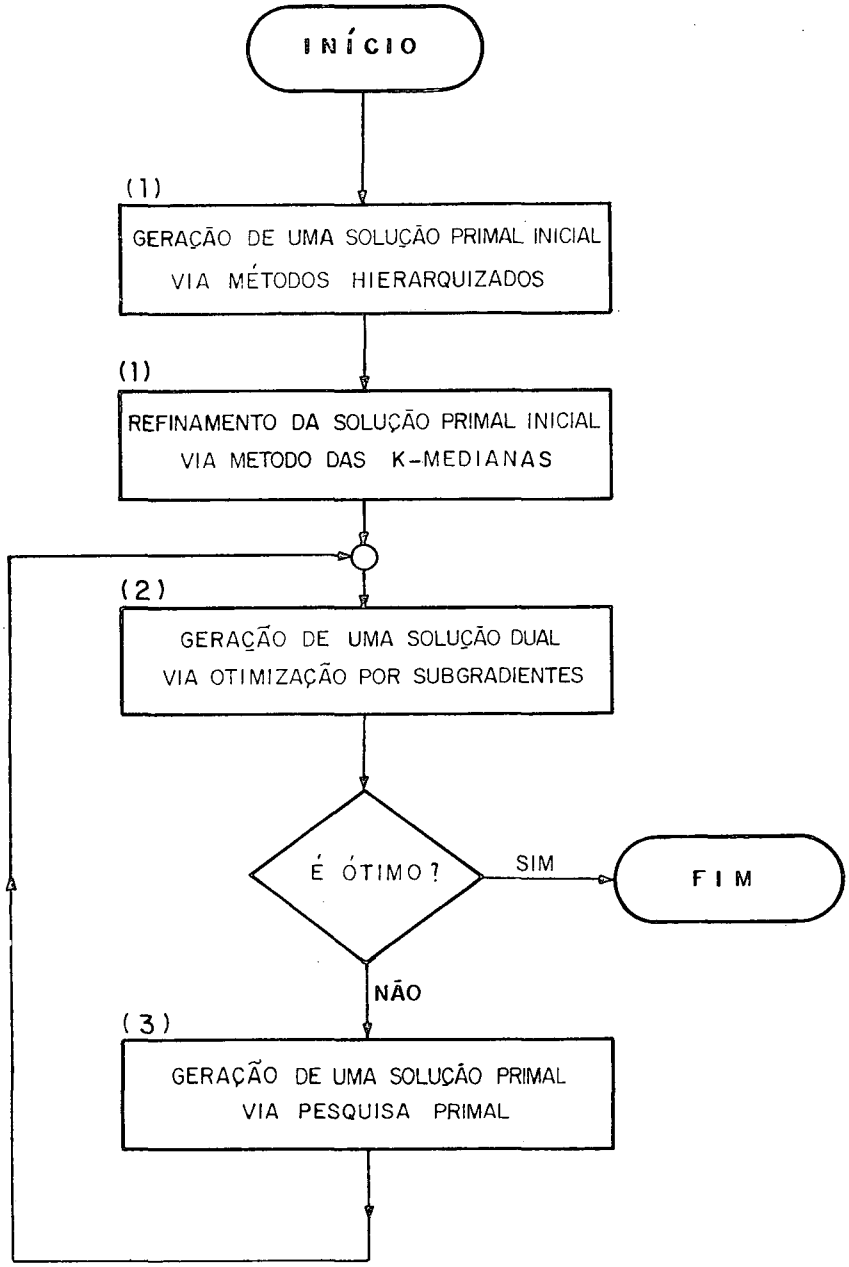
2 — geração de uma solução dual para o problema, via otimização por subgradientes; e

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

A parte 1 se constitui em inicialização do problema. As partes 2 e 3 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 seqüência de limites inferiores, como foi visto no item anterior, converge para a solução do Problema *DV*. A seqüência de soluções primais viáveis não converge necessariamente. Mulvey & Crowder (31) 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 4 e 5 deste trabalho.

— Solução dual:

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

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

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

pode ser reescrito da forma:

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

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

Mulvey & Crowder (31) 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 seqüência $\{(u_1, \ldots, u_n)^j\}$ é gerada pelo método subgradiente.

— Solução primal:

Quanto à solução primal viável do Problema *PV*, Mulvey & Crowder (31) 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:

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

$$y_{ij} = \begin{cases} 1 \text{ se } d_{ij} = \min\limits_{k \in L} \{d_{ik}\}, \text{ com } \sum\limits_{j \in L} y_{ij} = 1 \\ 0 \text{ em caso contrário.} \end{cases}$$

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 & Crowder (31) de "Pesquisa Primal".

Quanto ao ponto de vista de esforço computacional, Mulvey & Crowder (31) 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 & Crowder são equivalentes à resolução do problema por um método hierarquizado aglomerativo. Por outro lado, os resultados obtidos por Mulvey & Crowder (31) 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 5.3 utilizando-se a rotina CLSTROPT 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 5

MÉTODO DE MULVEY & CROWDER APLICADO AO EXEMPLO DO ITEM 5.3

ITER	FOP	FOS	FOD	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 seqüência 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.

6.5 — Relação entre o método de Vinod e o problema de localização não capacitado

O problema de localização não capacitado pode ser descrito (ver exemplo Mateus & Bornstein (30) ou Effroymson & Ray (11)) da forma:

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

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

$$1 \leqslant \sum_{j \in E} y_j \leqslant N \tag{L.3}$$

$$z_{ij} \leqslant y_j \quad , \quad i \in I, \quad j \in E \tag{L.4}$$

$$z_{ij}, y_j \in \{0,1\}, i \in I, j \in E \tag{L.5}$$

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;

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

$y_j = \begin{cases} 1 \text{ se o fornecedor } j \text{ for ativado} \\ 0 \text{ em caso contrário.} \end{cases}$

Então, se d_i é a demanda do centro i, a dimensão do fornecedor j será dada por $\sum_{i \in I} z_{ij}\ d_i$.

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 (ver exemplo Mateus & Bornstein (30) ou Jarvinen, Rajala & Sinervo (22)). 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 & Bornstein (30), 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:

$$\sum_{j \in I} y_j = N.$$

Como se nota, essa formulação do problema das *N*-medianas é equivalente à expressão do primeiro modelo de Vinod.

Mateus & Bornstein (30) 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 & Crowder, determinar intervalos para essa solução [1]. Mulvey & Crowder (31), por sua vez, apontam a necessidade de se desenvolverem estudos comparativos da eficiência dessas duas técnicas na resolução de problemas nas duas áreas.

6.6 — Um modelo eficiente de programação dinâmica para o caso univariado — o modelo de Rao

6.6.1 — Introdução

Na seção 6.3 foi apresentado o segundo modelo de Vinod — caso univariado. O modelo ali descrito, no entanto, é um modelo de progra-

[1] (Para uma descrição de algoritmos gulosos que resolvem essa classe de problemas, ver também Fisher, Nemhauser & Wolsey (13) e (14)).

mação linear inteira e, como foi apontado na introdução deste Capítulo 6, modelos de programação inteira geralmente não são eficientes, do ponto de vista do esforço computacional envolvido para a obtenção de uma solução ótima.

Por outro lado, e baseando-se em Rao (35), 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 6.3):

$$Z = \{2, 2, 4, 5, 6\},$$

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 6.3) 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:

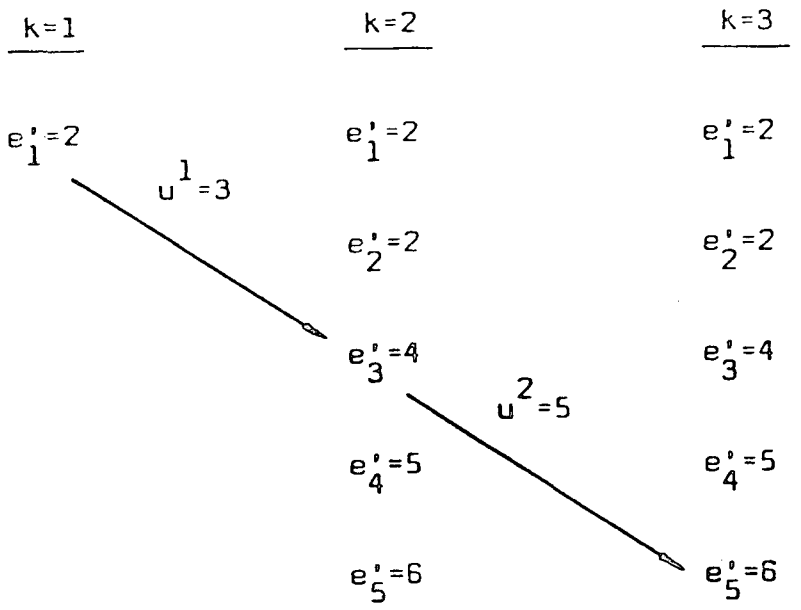
— estágio k : etapa de definição do k-ésimo grupo;

— estado y^k : índice associado ao líder do grupo g_k ($k = 1, \ldots, \ldots, m$) — 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} ($k = 1, \ldots, \ldots, m-1$) — 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:

$$y^{k+1} = u^k; \qquad (6.5.1)$$

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 ($k = 1, \ldots, m - 1$), 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 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. Verifica-se então que, no penúltimo estágio y^{m-1} deve ser inferior a $n : 1 \leqslant y^{m-1} \leqslant n - 1$. No antepenúltimo estágio, $1 \leqslant y^{m-2} \leqslant n - 2$. De uma maneira geral, $1 \leqslant y^{m-i} \leqslant n - i$. Caso $i = m - k$, tem-se finalmente,

$$1 \leqslant y^k \leqslant n - (m - k); \qquad (6.5.2)$$

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:

$$J(y^k, k) = \min_{u^k \text{ viável}} \{W_k + J(u^k, k+1)\}, \qquad (6.5.3)$$

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 (6.3.2) da Seção 6.3.1):

$$J(y^m, m) = J(n-1, m) = W_m = \frac{1}{2}(z_{n-1} - z_n)^2.$$

Em geral, para $y^m = i$, tem-se a formação do grupo

$$g_m = \{e'_i, e'_{i+1}, \ldots, e'_n\}.$$

Na seção 6.3.1, 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 (6.3.3) e (6.3.4)):

$$W_k = \sum_{r=j+1}^{i} \Delta W_j^r = \sum_{r=j+1}^{i} \frac{r-1}{(r-j)+1} \left(z_r - \frac{1}{(r-j)} \sum_{q=j}^{r-1} z_q\right)^2 \qquad (6.5.4)$$

Utilizando, a exemplo do efetuado naquela seção 6.3.1, o teorema (4.1) da seção 4.2.5, que informa que, ao se unirem dois grupos g_I e g_J para se formar o grupo g_L, tem-se:

$$W_L = W_I + W_J + \frac{n_I n_J}{n_I + n_J} d_{IJ}^2, \qquad (6.5.5)$$

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:

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

$$g_J = \{e'_{r+1}, \ldots, e'_n\}; \ n_J = n - r \ ; \ W_J = \sum_{e'_j \in g_J} (z_j - \bar{z}_J)^2$$

e, finalmente, pela equação (6.5.5)

$$W_L = W_J + \delta W_n^r = W_J + \frac{n-r}{1+(n-r)} \left(z_r - \frac{1}{n-r} \sum_{q=r+1}^{n} z_q\right)^2. \qquad (6.5.6)$$

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:

$$W_m = J(y^m, n) = J(i, m) = \sum_{r=i}^{n-1} \delta W_n^r . \qquad (6.5.7)$$

Note-se que a diferença entre as equações (6.5.4) e (6.5.7) reside apenas em que em ΔW_j^r o acréscimo do elemento é feito "à direita", ou seja, considera-se que $g_k = \{e_j', \ldots, e_{i-1}'\}\ U\ \{e_i'\}$ e, em δW_n^r, o acréscimo é feito "à esquerda":

$$g_m = \{e_i'\}\ U\ \{e_{i+1}', \ldots, e_n'\}.$$

Assim, no exemplo em questão, ter-se-ia (ver Equações (6.5.5) e (6.5.6)):

$$g_3 = \{e_5'\} \Rightarrow J(5, 3) = 0$$

$$g_3 = \{e_4', e_5'\} \Rightarrow J(4, 3) = \frac{5-4}{1+(5-4)}(5-6)^2 + J(5,3) = 0{,}50$$

$$g_3 = \{e_3', e_4', e_5'\} \Rightarrow J(3, 3) = \frac{5-3}{1+(5-3)}\left\{4 - \frac{1}{5-3}(5+6)\right\}^2 + J(4, 3)$$

$$= 1{,}5 + 0{,}5 = 2{,}00$$

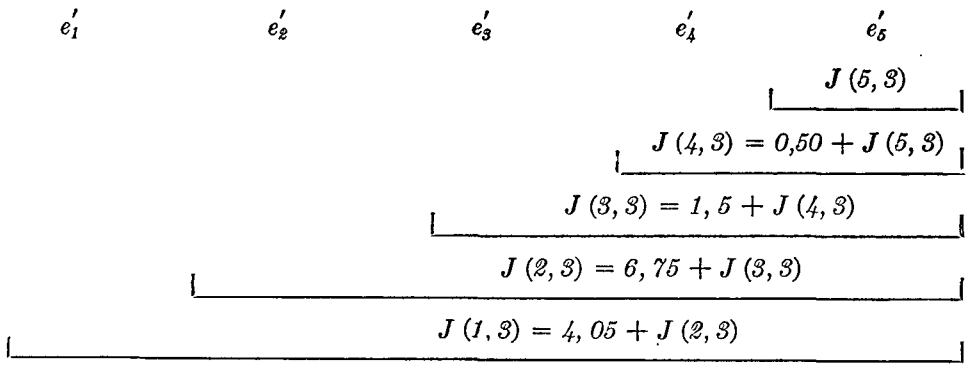
$$g_3 = \{e_2', \ldots, e_5'\} \Rightarrow J(2, 3) = \frac{5-2}{1+(5-2)}\left\{2 - \frac{1}{5-2}(4+5+6)\right\}^2 + J(3, 3)$$

$$= 6{,}75 + 2{,}00 = 8{,}75$$

$$g_3 = \{e_1', \ldots, e_5'\} \Rightarrow J(1, 3) = \frac{5-1}{1+(5-1)}\left\{2 - \frac{1}{5-1}(2+4+5+6)\right\}^2 + J(2, 3)$$

$$= 4{,}05 + 8{,}75 = 12{,}80.$$

Graficamente, tem-se:

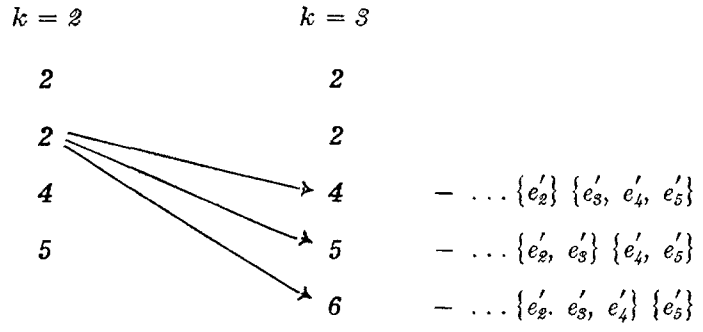


Assim, o cálculo dos $J(y^k, m)$ sempre é feito a partir dos resultados da etapa anterior: $J(3,3)$, por exemplo, é calculado a partir de $J(4,3)$. Além disso, tudo se passa, no último estágio, como se fosse calculado, recursivamente, o custo da formação de um único grupo $\{e'_1, \ldots, e'_n\}$;

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 (6.5.3):

$$J(y^k, k) = \min_{u^k \text{ viável}} \{W_k + J(u^k, k+1)\}.$$

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:

$$u^2 = 3 \Longrightarrow g_2 = \{e'_2\}$$
$$u^2 = 4 \Longrightarrow g_2 = \{e'_2, e'_3\}$$
$$u^2 = 5 \Longrightarrow g_2 = \{e'_2, e'_3, e'_4\}.$$

Note-se que a seqüência de valores admissíveis de u^2 implica no acréscimo de um elemento "à direita" e assim, pela equação (6.5.4) 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:

$$u^2 = 3 \Longrightarrow g_2 = \{e'_2\} \Longrightarrow W_2 = 0{,}00$$

$$u^2 = 4 \Longrightarrow g_2 = \{e'_2,\ e'_3\} \Longrightarrow W_2 = \Delta W_2^3 = 2{,}00$$

$$u^2 = 5 \Longrightarrow g_2 = \{e'_2,\ e'_3,\ e'_4\} \Longrightarrow W_2 = \Delta W_2^4 + \Delta W_2^3 = 2{,}67 + 2{,}00 = 4{,}67.$$

Para o cálculo da decisão ótima tem-se (ver Equação (6.5.3)):

$$u^2 = 3 \Longrightarrow J(y^2,\ 2) = W_2 + J(3{,}3) = 0{,}00 + 2{,}00 = 2{,}00$$

$$u^2 = 4 \Longrightarrow J(y^2,\ 2) = W_2 + J(4{,}3) = 2{,}00 + 0{,}50 = 2{,}50$$

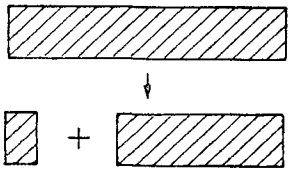
$$u^2 = 5 \Longrightarrow J(y^2,\ 2) = W_2 + J(5{,}3) = 4{,}67 + 0{,}00 = 4{,}67.$$

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 (6.5.7), 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 (6.5.4), o que corresponde a que, em cada estado y^k dos estágios k ($k = 1, \ldots, m-1$) 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:

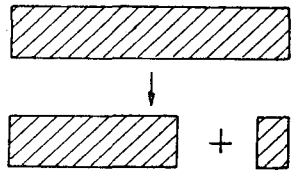
— inclusão "à esquerda" (k = m)



equação (6 .5.7):

$$W_m = \sum_{r=1}^{n-1} \delta W_n^r$$

— inclusão "à direita" ($k = 1, \ldots, m-1$)



equação (6 .5.4):

$$W_k = \sum_{r=j+1}^{i} \Delta W_j^r$$

$k = 1, \ldots, m-1$

6.6.2 — 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, $k = 1, \ldots, m$;

— decisão u^k: índice associado ao líder do grupo g_{k+1}, $k = 1, \ldots, \ldots, (m-1)$;

— estados viáveis $Y(k)$: conjunto de índices associados aos possíveis líderes do grupo g_k. Como visto no item *e* da seção anterior,

$$1 \leqslant y^k \leqslant n-(m-k);$$

— 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: $U(y^k, k) = \{u^k \mid u^k \in \{y^k + 1, \ldots, n-(m-k-1)\}\}$ (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*.

6.6.3 — Eficiência do modelo em relação à enumeração completa

Para $k = m$, como visto nos itens *g* e *i* da seção 6.5.1, 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 $1 \leqslant k \leqslant m-1$, por outro lado, tem-se:

— pela equação (6.5.2), $1 \leqslant y^k \leqslant n-(m-k)$;

— o cálculo da decisão ótima em y^k envolve o cálculo da formação de apenas um grupo (ver Itens *h* e *i* da Seção 6.5.1);

— em $k = 1$ tem-se apenas um estado viável: $y^1 = 1$ (ver Item *a* da Seção 6.5.1); e

— nesse caso ($1 \leqslant k \leqslant m-1$), 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 3.4):

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

Assim, $S(7, 3) = 301$;

— como em cada alternativa de grupamento se efetua o cálculo do custo de formação de três grupos, tem-se:

$$N_{EC} = 3 \times 301 = 903.$$

6.6.4 — 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 exemplo Harman (18)).

Recentemente, Pfeiffer (33) 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 e assim por diante. Maiores detalhes sobre a resolução do problema podem ser obtidos em Pfeiffer (33). 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 2.4.4), 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 (ver exemplo Raj (34) ou Sudman (40)). 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:

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

6.7 — 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 (43) e a Rao (35), 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.

7 — APLICAÇÃO DE ANÁLISE DE GRUPAMENTO À DETERMINAÇÃO DE ETAPAS DE CRESCIMENTO DE LARVAS

7.1 — Descrição do problema

Durante 25 dias foram colocadas, sucessivamente, em cada dia, oito larvas "recém-nascidas" em 25 vidros. No final do 26.º dia recolheu-se os vidros e efetuou-se nas larvas medições das variáveis "cápsula cefálica", "comprimento" e "peso". Obteve-se assim a tabela 6 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 $25 \times 8 = 200$ 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 possuíam 12 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 ínstares 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 ínstares $(i + 1)$ e (i) é constante para todo (i):

$$\frac{CC_i}{CC_{i-1}} = k, \quad \forall\ i \geqslant 2 \tag{7.1}$$

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 6

MEDIDAS EFETUADAS NA AMOSTRA DE LARVAS

(continua)

LARVA	PESO (mg)	COMPRIMENTO (mm)	MEDIDA DA CÁPSULA CEFÁLICA
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
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

TABELA 6

MEDIDAS EFETUADAS NA AMOSTRA DE LARVAS

(continua)

LARVA	PESO (mg)	COMPRIMENTO (mm)	MEDIDA DA CÁPSULA CEFÁLICA
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
10D1	21,00	11.5	1,357
10D2	22,00	11,5	1,214
10D3	17,00	10,0	1,049
10D4	23,00	11,0	1,143
10D5	20,00	11,0	1,236
10D6	32,00	13,5	1,373
10D7	25,00	11,5	1,439
10D8	23,00	11,0	1,291
11D1	30,00	13,5	1,214
11D2	34,00	14,5	1,302
11D3	34,00	13,5	1,346
11D4	32,00	13,0	1,242
11D5	32,00	13,0	1,225
11D6	21,00	10,0	1,488
11D7	21,00	11,5	1,055
11D8	16,00	8,5	1,351
12D1	22,00	12,0	1,203
12D2	28,00	13,5	1,197
12D3	30,00	13,0	1,280
12D4	30,00	13,0	1,340
12D5	25,00	12,0	1,197
12D6	24,00	14,5	1,477
12D7	23,00	12,0	1,142
12D8	30,00	11,5	1,577
13D1	47,00	13,5	1,159
13D2	50,00	15,0	1,395
13D3	44,00	13,0	1,313
13D4	26.00	12,0	1,049
13D5	25,00	10,0	1,494
13D6	25,00	11,0	1,428
13D7	23,00	11,5	1,296
13D8	25,00	9,5	1,428

TABELA 6

MEDIDAS EFETUADAS NA AMOSTRA DE LARVAS

(continua)

LARVA	PESO (mg)	COMPRIMENTO (mm)	MEDIDA DA CÁPSULA CEFÁLICA
14D1	38,00	14,0	1,357
14D2	55,00	16,5	1,373
14D3	23,00	11,0	1,115
14D4	24,00	11,0	1,230
14D5	30,00	12,0	1,511
14D6	33,00	12,5	1,219
14D7	38,00	15,0	1,516
14D8	30,00	13,5	1,230
15D1	38,00	13,5	1,472
15D2	30,00	13,5	1,099
15D3	26,00	13,0	1,099
15D4	33,00	13,5	1,275
15D5	38,00	15,0	1,318
15D6	47,00	15,5	1,296
15D7	45,00	15,0	1,395
15D8	28,00	12,5	1,181
16D1	69,00	19,5	1,813
16D2	59,00	17,0	1,280
16D3	100,00	21,0	1,785
16D4	58,00	17,0	1,522
16D5	85,00	18,5	1,533
16D6	80,00	19,0	1,648
16D7	80,00	17,0	1,550
16D8	76,00	16,0	1,752
17D1	46,00	15,5	1,302
17D2	45,00	15,5	1,648
17D3	50,00	15,0	1,411
17D4	56,00	17,0	1,466
17D5	40,00	13,5	1,373
17D6	52,00	16,0	1,373
17D7	45,00	13,0	1,648
17D8	47,00	13,5	1,648
18D1	77,00	18,0	1,999
18D2	77,00	19,5	1,835
18D3	68,00	15,0	1,714
18D4	65,00	16,0	1,648
18D5	56,00	18,0	1,346
18D6	53,00	17,0	1,488
18D7	60,00	16,0	1,774
18D8	63,00	17,0	1,373
19D1	90,00	20,0	1,604
19D2	68,00	18,0	1,494
19D3	60,00	17,0	1,648
19D4	79,00	19,0	1,846
19D5	70,00	17,0	1,488
19D6	45,00	13,0	1,648
19D7	75,00	18,0	1,873
19D8	73,00	17,5	1,785

TABELA 6

MEDIDAS EFETUADAS NA AMOSTRA DE LARVAS

(conclusão)

LARVA	PESO (mg)	COMPRIMENTO (mm)	MEDIDA DA CÁPSULA CEFÁLICA
20D1	121,00	20,0	1,648
20D2	90,00	17,5	1,901
20D3	133,00	22,0	1,730
20D4	98,00	21,5	1,681
20D5	110,00	20,0	1,785
20D6	92,00	19,0	1,769
20D7	145,00	24,0	1,961
20D8	101,00	19,0	1,648
21D1	122,00	20,5	1,697
21D2	95,00	19,5	1,565
21D3	117,00	20,5	1,769
21D4	82,00	18,0	1,752
21D5	103,00	20,5	1,648
21D6	88,00	19,0	1,593
21D7	125,00	22,0	1,719
21D8	102,00	21,0	1,730
22D1	75,00	18,0	1,648
22D2	63,00	17,0	1,587
22D3	103,00	21,0	1,785
22D4	40,00	12,0	1,439
22D5	58,00	17,0	1,379
22D6	67,00	15,5	1,785
22D7	71,00	20,0	1,472
22D8	85,00	17,5	1,648
23D1	61,00	18,0	1,565
23D2	113,00	21,0	1,648
23D3	78,00	17,0	1,988
23D4	90,00	19,0	1,785
23D5	125,00	20,0	1,944
23D6	98,00	19,5	1,670
23D7	90,00	18,5	1,648
23D8	75,00	17,0	1,544
24D1	113,00	21,0	1,829
24D2	118,00	18,0	1,598
24D3	152,00	24,0	1,917
24D4	83,00	17,0	1,648
24D5	110,00	19,0	1,824
24D6	85,00	16,0	2,236
24D7	138,00	24,0	1,593
24D8	125,00	20,0	1,648
25D1	130,00	20,0	1,824
25D2	113,00	18,0	1,708
25D3	148,00	20,0	1,813
25D4	101,00	17,0	1,862
25D5	100,00	17,0	1,818
25D6	82,00	18,0	1,554
25D7	92,00	17,5	1,648
25D8	128,00	19,5	1,862

7.2 — 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 grupamento, 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 (7.1). 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 (7.1), foi utilizado um modelo de regressão linear simples da forma:

$$\hat{CC}_i = k\ CC_{i-1} \qquad i = 2, \ldots, m,$$

sendo utilizado como medida da qualidade do ajustamento o coeficiente de determinação R^2, que é uma medida do poder "explanatório" da regressão (ver exemplo Wesolowsky (45)):

$$R^2 = 1 - \frac{\sum_{i=2}^{m} (CC_i - \hat{CC}_i)^2}{\sum_{i=2}^{m} (CC_i - \overline{CC})^2}, \tag{7.2}$$

onde CC_i corresponde ao valor médio da cápsula cefálica no grupo $i (i = 2, \ldots, m)$, $\hat{CC}_i$ é o valor estimado pela equação de regressão, $(i = 2, \ldots, m)$ e $\overline{CC}$ corresponde ao valor médio de CC_i $(i = 2, \ldots, m)$.

Pela equação (7.2) se nota que, quanto mais próximo for R^2 de 1, melhor a qualidade do ajustamento.

7.3 — Resolução do problema

7.3.1 — Amostra reduzida

Tendo em vista a utilização do método de Mulvey & 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 três primeiras larvas de cada dia.

Tomando, por outro lado, como variáveis,

$V1$ = peso
$V2$ = comprimento
$V3$ = cápsula cefálica,

TABELA 7

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) M	PESO (mg) DP	COMPRIMENTO (mm) M	COMPRIMENTO (mm) DP	MEDIDA DA CÁPSULA CEFÁLICA M	MEDIDA DA CÁPSULA CEFÁLICA 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	(1)	24,00	(1)	1,92	(1)

NOTA — M = média; DP = desvio padrão.
(1) grupo constituído por apenas um elemento.

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.

TABELA 8

CARACTERÍSTICAS DOS GRUPAMENTOS EFETUADOS NA AMOSTRA DE 75 LARVAS, TOMANDO COMO CARACTERÍSTICA PARA O AGRUPAMENTO, A CÁPSULA CEFÁLICA

NÚMERO DE GRUPOS	GRUPOS	PESO (mg) M	PESO (mg) DP	COMPRIMENTO (mm) M	COMPRIMENTO (mm) DP	MEDIDA DA CÁPSULA CEFÁLICA M	MEDIDA DA CÁPSULA CEFÁLICA 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

NOTA — M = média; DP = desvio padrão.

O problema foi inicialmente resolvido pelo método das *K*-Medianas (apresentado no Item 5.3), para $p = 3$ (peso, comprimento, cápsula cefálica — dados padronizados) e $p = 1$ (cápsula cefálica). Os resultados estão resumidos nas tabelas 7 e 8.

Como se observa na tabela 7, 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 8, de uma maneira geral, uma maior homogeneidade no que diz respeito à cápsula cefálica. Os desvios padrão estão ainda um pouco eleva-

dos (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 & Crowder, uma vez que este método, como foi visto no item 6.4, 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 9.

TABELA 9

APLICAÇÃO DO MÉTODO DE MULVEY & 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 (7.1), através da metodologia apresentada no item 7.2. Os resultados estão resumidos na tabela 10.

TABELA 10

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.

7.3.2 — Amostra integral

Tendo em vista a maior eficiência do algoritmo univariado de programação dinâmica apresentado no item 6.5, 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 11.

TABELA 11

MÉDIA E DESVIO PADRÃO DA CÁPSULA CEFÁLICA, NAS SOLUÇÕES ÓTIMAS OBTIDAS PARA

m = 4, 5, 6 e 7

NÚMERO DE GRUPOS	GRUPO	MÉDIA	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

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 12.

TABELA 12

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".

7.4 — Conclusão

Pela descrição do problema, apresentada no item 7.1, a aplicação da análise de grupamento aos dados da amostra deveria gerar, para um valor de $m = 5$ 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 10:

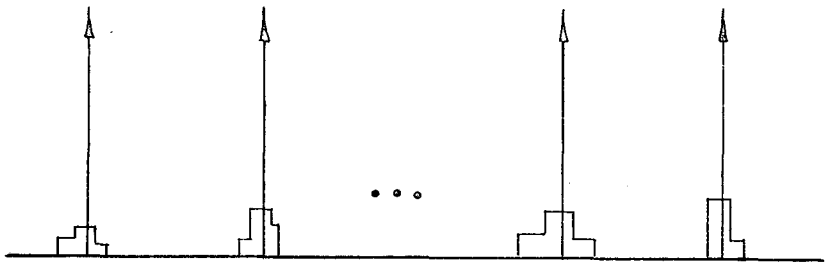


FIGURA 10 — *Distribuição esperada da amostra de larvas*

Um exame do histograma da amostra, indicado na figura 11, indica que tal fenômeno não ocorre:

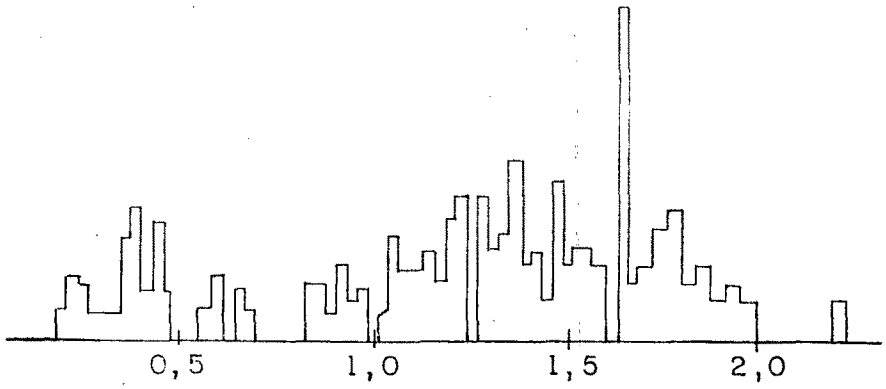


FIGURA 11 — *Distribuição efetiva da amostra de larvas*

Assim, os resultados da análise de grupamento não devem surpreender. A solução ótima do item 7.3.2 apenas informa que não é possível particionar-se a amostra em estudo na forma apresentada na figura 10, fato esse que é sugerido pelo próprio exame da figura 11.

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. Conseqüentemente, 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.

8 — CONCLUSÃO

Em vista do exposto nos Capítulos anteriores, deve ter ficado clara a natureza e as características de cada método apresentado. Restaria no entanto, abordar a questão do porte dos problemas a serem resolvidos 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 agrupados) 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 dificuldade. Como exemplo, seja 250 o número de elementos que os recursos disponíveis permitem agrupar e 5 mil o tamanho do problema original.

Uma das maneiras de se resolver o problema seria tomar uma amostra de 250 elementos, formar um "grupamento semente" com essa amostra e, a partir desse "grupamento semente", classificar os demais 4.750 elementos. Essa classificação poderia ser efetuada alocando-se cada elemento não classificado ao grupo mais próximo ou então poder-se-ia utilizar as técnicas de classificação descritas, por exemplo, em Tatsuoka (41) ou Cooley & 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 análise de grupamento, em um número predeterminado de grupos. Após a resolução do problema em cada bloco, efetua-se uma análise de grupamento no conjunto 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 mil, onde os recursos permitem que se efetue análise de grupamento de até 250 elementos. Pela técnica de Anderberg (1), o problema poderia ser resolvido da seguinte forma:

— o conjunto dos 5 mil 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 $20 \times 12 = 240$ centróides ou medianas seriam, através de análise de grupamento, gerados m grupos; e

— os 5 mil 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.

9 — APÊNDICES

Apêndice 1

TABELA A.1.1

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).....	$d_{ij} = [0{,}5\ (1 - r_{ij})]^{1/2}$

(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.1.2

MEDIDAS DE DISPERSÃO INTERNA DE UM GRUPO

MEDIDAS	EQUAÇÕES
Soma dos quadrados dentro do grupo (1).....	$W_I = \sum_{i=1}^{n_I} d_2^2 (X_i, \overline{X}_I)$
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.....	$Z_{min}(g_I) = \min_{e_j \in g_I} Z(e_j) = \min_{e_j \in g_I} \sum_{i=1}^{n_I} d_{ij}$

(1) $\overline{X}_I$ é a centróide (média) do grupo g_I : $\overline{X}_I = \left(\sum_{i=1}^{n_I} X_i \right) / n_I$;

e $d_2^2 (X_i, \overline{X}_I)$ é o quadrado da distância euclidiana entre X_i e $\overline{X}_I$.

TABELA A.1.3

FUNÇÕES OBJETIVO EM ANÁLISE DE GRUPAMENTO

FUNÇÕES OBJETIVO	EQUAÇÕES
Soma dos quadrados dentro dos grupos ..	$W = \sum_{g_I \in P_m} W_I = \sum_{g_I \in P_m} \sum_{i=1}^{n_I} d_2^2 (X_i, \overline{X}_I)$
Soma das variâncias internas	$S^2 = \sum_{g_I \in P_m} S_I^2 = \sum_{g_I \in P_m} \frac{1}{n_I} W_I$
Diâmetro de grupo	$d(P_m) = \max_{g_I \in P_m} d(g_I) = \max_{g_I \in P_m} \max_{e_i, e_j \in g_I} d_{ij}$
Dispersão via mediana de grupo.....	$Z(P_m) = \sum_{g_I \in P_m} Z_{min}(g_I) = \sum_{g_I \in P_m} \min_{e_j \in g_I} \sum_{i=1}^{n_I} d_{ij}$

TABELA A.1.4

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)	$d_{IJ}^2 = d_2^2 (\overline{X}_I, \overline{X}_J)$
Ward.....	$D_{IJ} = \frac{n_I n_J}{n_I + n_J} d_{IJ}^2$
Média de Grupo(2)	$D_{IJ}^2 = S_I^2 + S_J^2 + d_{IJ}^2$

(1) $\overline{X}_I$ é a centróide (média) do grupo g_I e $d_2^2(\overline{X}_I, \overline{X}_J)$ o quadrado da distância euclidiana entre $\overline{X}_I$ e $\overline{X}_J$. O método da mediana (ver Item 4.2.4) difere do método da centróide apenas no que se refere à determinação das centróides. (2) S_I^2 é a variância interna do grupo g_I (ver Item 2.4.2 ou Tabela A.1.2).

Apêndice 2

A.2.1 — Rotina CLSTROPT

Essa rotina efetua a análise de grupamento através das técnicas hierarquizadas aglomerativas descritas nesta tese e dos métodos *K*-Medianas e Mulvey & Crowder. Uma listagem da rotina é apresentada nas folhas que se seguem.

Os dados de entrada do programa devem ser organizados da seguinte forma:

— 1.º cartão: informar, em 7 campos consecutivos de 5 posições, iniciando-se o conjunto na coluna 1, os seguintes dados, alinhando os valores à direita:

- 1.º campo — número de características;
- 2.º campo — número de elementos;
- 3.º campo — número de grupos;
- 4.º 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 & Crowder (valores das funções objetivo primal e dual, e sua diferença percentual, em cada iteração);
  - — "4" — resultados intermediários dos métodos *K*-Medianas e Mulvey & Crowder;
  - — "5" — resultados intermediários de todos os métodos;

  Obs.: em todos os métodos, qualquer que seja o valor informado nesse 4.º campo, os resultados finais são sempre impressos.
- 5.º campo — número máximo de iterações nos métodos *K*-Medianas e Mulvey & Crowder;
- 6.º campo — distância a ser utilizada. Se informado o valor "1", é calculada a distância euclidiana. Se informado "2", é cal-

culada a métrica de correlação. Se informado um valor diferente de "1" e "2", ocorrerá erro;

- 7.º campo — padronização das variáveis. Se informado "1", os dados são padronizados. Para qualquer outro valor, os dados permanecem sem padronização.

— 2.º 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":

- 1.º campo — Ligação Simples
- 2.º campo — Ligação Completa
- 3.º campo — Mediana
- 4.º campo — Média de Grupo
- 5.º campo — Centróide
- 6.º campo — Ward
- 7.º campo — *K*-Medianas
- 8.º campo — Mulvey & Crowder

Obs.: o método *K*-Medianas exige o processamento anterior de uma técnica hierarquizada. O método Mulvey & Crowder exige o processamento anterior do método *K*-Medianas.

— cartões subseqüentes: informar, para cada elemento:

1.º cartão — código do elemento, com 6 caracteres, iniciando-se na coluna 1.

demais cartões — informar, em seqüência, 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*-Medianas e Mulvey & Crowder, a distância euclidiana em dados padronizados, até um máximo de 40 iterações para os métodos *K*-Medianas

e Mulvey & 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 CLSTROPT da forma (ϕ e ∅ representam, respectivamente, espaço em branco e zero):

```
ϕϕϕϕϕϕ 2 ϕϕϕϕϕϕ 5 ϕϕϕϕϕϕ 3 ϕϕϕϕϕϕ 5 ϕϕϕϕϕ 5∅ ϕϕϕϕϕϕ 1 ϕϕϕϕϕϕ

∅∅∅∅∅ 111

ϕϕϕϕ A1

ϕϕϕ 1∅∅∅∅∅∅ ϕϕϕϕϕϕϕ 35∅

ϕϕϕϕ A2

ϕϕϕϕ 5∅∅∅∅∅ ϕϕϕϕϕϕϕ 42∅

ϕϕϕϕ A3

ϕϕϕϕ 25∅∅∅∅ ϕϕϕϕϕϕϕ 31∅

ϕϕϕϕ A4

ϕϕϕϕ 3∅∅∅∅∅ ϕϕϕϕϕϕϕ 2∅∅

ϕϕϕϕ A5

ϕϕϕϕ 2∅∅∅∅∅ ϕϕϕϕϕϕϕ 18∅
```

BURROUGHS B7700 BDMSALGOL COMPILER, VERSION 32.111.037,
MONDAY, 10/03/83, 03:20 PM.

CLSTROPT

```
%
%  ESTA ROTINA EFETUA UMA ANÁLISE DE AGRUPAMENTO
%  ATRAVÉS DOS MÉTODOS HIERARQUICOS AGLOMERATIVOS,
%  K-MEDIANAS E MULVEY & CROWDER, PADRONIZANDO OU
%  NÃO OS DADOS E UTILIZANDO A MÉTRICA EUCLIDIANA
%  OU A "MÉTRICA" DE CORRELAÇÃO
%
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 CARACTERÍSTICA
%     X [*,K] TENHA MÉDIA O E VARIÂNCIA 1
%
      INTEGER N,          % NÚMERO DE ELEMENTOS
              P;          % NÚMERO DE CARACTERÍSTICAS
      REAL ARRAY X [1,1]; % MATRIZ DE OBSERVAÇÕES

%     ************************************************
          BEGIN
             INTEGER  I,K;            % CONTADORES AUXILIARES
             REAL ARRAY MD [1:P],     % VETOR DE MÉDIAS
                        VAR [1:P];    % VETOR DE VARIÂNCIAS
          %
          %  CÁLCULO  DAS  MÉDIAS
          %
             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;
          %
          %  CÁLCULO  DAS  VARIÂNCIAS
          %
             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;

      %
      %  CÁLCULO  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  DISTÂNCIA  EUCLIDIANA
%  E[I+(J-1)*(J-2)/2], ENTRE X[I,*] E X[J,*], TAIS QUE J>I,I=1,...,N-1
      INTEGER N,              % NÚMERO  DE  ELEMENTOS
              P;              % NÚMERO  DE  CARACTERÍSTICAS
      REAL ARRAY E [1],       % VETOR  DE  QUADRADOS  DAS  DISTÂNCIAS
                 X [1,1];     % MATRIZ  DE  OBSERVAÇÕES

%  ****************************************************
         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  "MÉTRICA"  DE  CORRELAÇÃO  E[I+(J-1)*(J-2)/2]
%  ENTRE  X[I,*] E X[J,*],TAIS QUE J > I,I=1,...,N-1
%
      INTEGER N,              % NÚMERO  DE  ELEMENTOS
              P;              % NÚMERO  DE  CARACTERÍSTICAS
      REAL ARRAY E [1],       % VETOR  DE  DISTÂNCIAS
                 X [1,1];     % MATRIZ  DE  OBSERVAÇÕES
```

```
%  ***************************************************
       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  CLSTHRQ(N,MTD,TOTGRP,W,FO,AGRP,E,NIG,TIMP,CDG);
%                 = = = = = =
%  ESTA ROTINA EFETUA UMA ANÁLISE DE GRUPAMENTO DE N ELEMENTOS EM
%  TOTGRP GRUPOS, ATRAVÉS DO MÉTODO DEFINIDO PELO VALOR DE MTD:
%                  0- LIGAÇÃO  SIMPLES
%                  1- LIGAÇÃO  COMPLETA
%                  2- MEDIANA
%                  3- MÉDIA  DE  GRUPO
%                  4- CENTRÓIDE
%                  5- WARD,
%  SENDO  A  PARTIÇÃO  OBTIDA  ARMAZENDA  NO  VETOR  AGRP [1:N]
%
       INTEGER  N,                    % NÚMERO  DE  ELEMENTOS
                MTD,                  % MÉTODO  DE  GRUPAMENTO
                TOTGRP,               % NÚMERO  TOTAL  DE  GRUPOS
                TIMP;                 % TIPO  DE  IMPRESSÃO
       INTEGER  ARRAY  NIG [1],       % NÚMERO  DE  ELEMENTOS  EM  CADA  GRUPO
                       AGRP [1];      % AGRUPAMENTO  EFETUADO
```

```
    REAL  W,                    % MENOR DISTÂNCIA ENTRE GRUPOS EM CADA
                                  PASSO
          FO;                   % FUNÇÃO OBJETIVO, NO CASO DO MÉTODO DE
                                  WARD
    REAL  ARRAY  E [1],         % VETOR  DE  DISTÂNCIAS
                 CDG [1];       % VETOR  DE  CÓDIGOS  DOS  ELEMENTOS

%  ************************************************************************************************
       BEGIN
%      ********************************************************************************
    PROCEDURE  ATUALIZARD  (IM,JM,N,NIG,MTD,D,P,K);

%                 = = = = = =
%  ESTA  ROTINA  ATUALIZA  O  VETOR  D  DE  DISTÂNCIA  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,                 % DISTÂNCIA  ENTRE  OS  GRUPOS  FOI  MÍNIMA
            N,                  % NÚMERO  DE  ELEMENTOS
            K,                  % NÚMERO  DE  CARACTERÍSTICAS
          MTD;                  % MÉTODO  UTILIZADO
    INTEGER ARRAY P[1],         % VETOR  DE  APONTADORES  DAS  LINHAS  ATIVAS
                  NIG [1];      % NÚMERO  DE  ELEMENTOS  DO  GRUPO
    REAL  ARRAY  D [1];         % VETOR  DE  DISTÂNCIAS  ENTRE  OS  GRUPOS

%  ****************************************************************
       BEGIN
%      ************************************************************************
       PROCEDURE  CLCATL (IM,JM,NIG,MTD,D,LHI,LHJ,LIJ,H,P);
%                 = = = = =
%      ESTA  ROTINA  EFETUA  A  ATUALIZAÇÃO  DO  VETOR  DE  DISTÂNCIAS
%      D,QUANDO  OS  GRUPOS  P [IM]  E  P [JM]  SÃO  UNIDOS,  NO  MÉTODO
%      INDICADO  POR  MTD,  CALCULANDO  A  DISTÂNCIA  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 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 N-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 TÍTULO (MTD);
%               = = =
% ESTA ROTINA IMPRIME O TÍTULO DO MÉTODO, CONFORME O VALOR
% DE MTD
  INTEGER MTD;

% ************************
        BEGIN
        WRITE (OTPT[SKIP 1]);
        CASE    MTD
        OF      BEGIN
                WRITE(OTPT,<T10,"MÉTODO:LIGAÇÃO SIMPLES">);
                WRITE(OTPT,<T10,"MÉTODO:LIGAÇÃO COMPLETA">);
                WRITE(OTPT,<T10,"MÉTODO:MEDIANA">);
                WRITE(OTPT,<T10,"MÉTODO:MÉDIA DE GRUPO">);
```

```
                WRITE(OTPT,<T10,"MÉTODO:CENTRÓIDE">);
                WRITE(OTPT,<T10,"MÉTODO:WARD">);
                END
        END DA PROCEDURE TÍTULO;
        REAL ARRAY D [1:N*(N-1)/2];
        REAL MÍNIMO,
             WAC;
        INTEGER ARRAY P [1:N];
        INTEGER CM,IM,JM;
        INTEGER I,J,K,L;
%
%       INICIALIZAÇÃO
%
        IF TIMP=1 OR TIMP=5
        THEN BEGIN
                TÍTULO (MTD);
                WRITE(OTPT,<///,T5,"ITERAÇÃO",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;
%
% CÁLCULO DOS GRUPOS PARA OS QUAIS A DISTÂNCIA É MÍNIMA
%
FOR     K:=1
STEP    1
UNTIL N-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]<MÍNIMO
                        THEN    BEGIN
                                MÍNIMO:=D[L];
                                IM:=I;
                                JM:=J
                                END
```

```
                    END;
        IF      MTD=5
        THEN    BEGIN
                W:=MÍNIMO/2;
                FO:=*+W;
                END
        ELSE    W:=MÍNIMO;
        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;
%
% ATUALIZAÇÃO  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;

%
% IMPRESSÃO  DOS  RESULTADOS
%
        TÍTULO(MTD);
        IF      MTD=5
        THEN    WRITE (OTPT,<///,T12,"FUNÇÃO  OBJETIVO=",
                E12.5>,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 ANÁLISE DE GRUPAMENTO DE N ELEMENTOS EM
%  M GRUPOS, ATRAVÉS DO MÉTODO DAS K-MEDIANAS, SENDO A PARTIÇÃO
%  OBTIDA ARMAZENADA NO VETOR AGRP[1:N] E AS MEDIANA DOS GRUPOS,
%  POR SUA VEZ, ARMAZENADA NO VETOR MDN[1:M]
  INTEGER M,                % NÚMERO DE GRUPOS
          N,                % NÚMERO DE ELEMENTOS
          NITER,            % NÚMERO MÁXIMO DE ITERAÇÕES
          TIMP;             % TIPO DE IMPRESSÃO
  INTEGER ARRAY MDN[1],     % VETOR DE MEDIANAS
                AGRP[1];    % AGRUPAMENTO EFETUADO
  REAL  FO;                 % VALOR DA FUNÇÃO OBJETIVO
  REAL  ARRAY  E[1],        % VETOR DE DISTÂNCIAS
               CDG[1];      % VETOR DE CÓDIGOS 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
%     FUNÇÃO OBJETIVO ASSOCIADA AO GRUPAMENTO EFETUADO E A
%     DIFERENÇA PREC ENTRE OS VALORES DA FUNÇÃO OBJETIVO
%     ASSOCIADOS A DUAS ITERAÇÕES CONSECUTIVAS
      INTEGER  M,               % NÚMERO DE GRUPOS
               N,               % NÚMERO DE ELEMENTOS
               ITER,            % NÚMERO DE ITERAÇÕES
                 TIMP;          % TIPO DE IMPRESSÃO
      INTEGER ARRAY MDN[1],     % VETOR DE MEDIANAS
                    AGRP [1];   % AGRUPAMENTO EFETUADO
      REAL  FO,                 % VALOR DA FUNÇÃO OBJETIVO
            PREC;               % V. DEFINIÇÃO NO OBJETIVO DA ROTINA
      REAL  ARRAY  E[1];        % VETOR DE DISTÂNCIAS

%     ***************************************************************************
          BEGIN
              INTEGER  I,J,L;
              REAL FA,
                   TOL,
                   CST;
               REAL  ARRAY  CSTMIN[1:M];
          %
          %  INICIALIZAÇÃO
          %
              TOL:=0.0000000001;
              FOR   I:=1  STEP  1       UNTIL  M
              DO    MDN[I]:=0;
          %
          %  CÁLCULO  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;
        %
        %  CÁLCULO  DA  NOVA  FUNÇÃO  OBJETIVO  E  ARMAZENAMENTO
        %  DA  FUNÇÃO  OBJETIVO  DA  ITERAÇÃO  ANTERIOR
        %
           FA:=FO;
           FO:=0;
           FOR     I:=1
           STEP    1
           UNTIL M
           DO      FO:=*+CSTMIN[I];
        %
        %  IMPRESSÃO  DO  VALOR  DA  FUNÇÃO  OBJETIVO
        %
           IF      TIMP=2  OR  TIMP=4  OR  TIMP=5
           THEN    WRITE(OTPT,</,T17,I3,T26,E12.5>,ITER,FO);
        %
        %  CÁLCULO  DA  DIFERENÇA  PERCENTUAL  ENTRE  DOIS  VALORES
        %  SUCESSIVOS  DA  FUNÇÃO  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,                  % NÚMERO  DE  ELEMENTOS
           M;                  % NÚMERO  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 MÍNIMO;
            LABEL SEG;
          %
          % ALOCAÇÃO DE CADA ELEMENTO AO GRUPO CUJA
          % MEDIANA ESTÁ MAIS PRÓXIMA
          %
             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]<MÍNIMO
                         THEN BEGIN
                               AGRP[I]:=J;
                               MÍNIMO:=E[L]
                                END
                         END;
                   SEG:
                   END
      END DA PROCEDURE RLOC;

   INTEGER ITER,
           I,J,K,L;
   REAL PREC,
        MÍNIMO;
   IF    TIMP=2  OR  TIMP=4  OR  TIMP=5
   THEN BEGIN
        WRITE(OTPT[SKIP  1]);
        WRITE(OTPT,<///,T10,"MÉTODO:  K-MEDIANAS",///,T10,
                    "EVOLUÇÃO DA FUNÇÃO OBJETIVO",///,T16,
                    "ITER",T31,"FO",/," ">)
        END;
% INICIALIZAÇÃO
%
   ITER:=0;
   MDNFO(M,MDN,N,AGRP,E,FO,ITER,TIMP,PREC);
% CÁLCULO 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;
%
%   IMPRESSÃO  DOS  RESULTADOS
%
    ITER:=*-1;
    WRITE(OTPT[SKIP  1]);
    WRITE(OTPT,</,T10,"MÉTODO:  K-MEDIANAS",/,T12,
                "NÚMERO  DE  ITERAÇÕES  =  ",I3,/,T12,
                "FUNÇÃO  OBJETIVO  =  ",E12.5,/,T12,
                "PRECISÃO  =  ",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 ANÁLISE DE GRUPAMENTO DE N ELEMENTOS EM
%  M GRUPOS, ATRAVÉS DO MÉTODO DE MULVEY & CROWDER, SENDO A PARTIÇÃO
%  OBTIDA ARMAZENADA NO VETOR AGRPS [1:N] E AS MEDIANAS DOS GRUPOS
%  ARMAZENADAS,  POR  SUA  VEZ,  NO  VETOR  MDNS [1:M]
    INTEGER N,                    % NÚMERO  DE  ELEMENTOS
            M,                    % NÚMERO  DE  GRUPOS
            NITER,                % NÚMERO  MÁXIMO  DE  ITERAÇÕES
            TIMP;                 % TIPO  DE  IMPRESSÃO
    INTEGER ARRAY AGRPS [1],      % AGRUPAMENTO  EFETUADO
                  MDNS [1];       % VETOR  DE  MEDIANAS
    REAL FOPS;                    % VALOR  DA  FUNÇÃO  OBJETIVO  PRIMAL
    REAL ARRAY D [1],             % VETOR  DE  DISTÂNCIAS
               CDG [1];           % VETOR  DE  CÓDIGOS  DOS  ELEMENTOS

%  ************************************************************************************
    BEGIN
%       ***************************************************
        PROCEDURE  PSQPRM (N,M,MDN,AGRP,D,FOP);
%                  = = = = = =
%       ESTA  ROTINA  CALCULA  UMA  SOLUÇÃO  PRIMAL  AGRP [1:N],  COM  SEU
%       CORRESPONDENTE  VALOR  DE  FUNÇÃO  OBJETIVO  FOP,  A  PARTIR  DE
%       UM  CONJUNTO  DE  MEDIANAS  MDN [1:M]
        INTEGER N,                    % NÚMERO  DE  ELEMENTOS
                M;                    % NÚMERO  DE  GRUPOS
        INTEGER ARRAY MDN [1],        % VETOR  DE  MEDIANAS
                      AGRP [1];       % AGRUPAMENTO  EFETUADO
        REAL ARRAY D[1];              % VETOR  DE  DISTÂNCIAS
        REAL FOP;                     % VALOR  DA  FUNÇÃO  OBJETIVO
```

```
%                BEGIN
                      INTEGER  I,J,K,L;

                      REAL  MÍNIMO;
                      LABEL  SEG;
                %
                %  INICIALIZAÇÃO
                %
                      FOP:=0;
                      FOR    J:=1
                      STEP   1
                      UNTIL M
                      DO     AGRP [MDN [J]]:=J;
                %
                %  CÁLCULO  DA  SOLUÇÃO  PRIMAL  E  DA  FUNÇÃO  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]<MÍNIMO
                                    THEN BEGIN
                                           AGRP [I]:=J;
                                           MÍNIMO:=D [L]
                                           END
                                    END;
                             FOP:=*+MÍNIMO;
                             SEG:
                             END
                   END  DA  PROCEDURE  PSQPRM;
%  ***************************************************************************************************
   PROCEDURE  OTSBGR (N,M,AGRPS,MDN,D,V,ITER,LBDA,FOD,FOPS);
%                 = = = = =
%  ESTA  ROTINA  CALCULA  UMA  SOLUÇÃO  DUAL  LBDA [1:N], NO MÉTODO DE
%  MULVEY  &  CROWDER,  BEM  COMO  O  VALOR  DA  FUNÇÃO  OBJETIVO  DUAL
%  FOD  É  O  VETOR  DE  MEDIANAS  MDN [1:M]
   INTEGER N,                    % NÚMERO  DE  ELEMENTOS
           M,                    % NÚMERO  DE  GRUPOS
           ITER;                 % NÚMERO  DE  ITERAÇÕES
   INTEGER ARRAY AGRPS [1],      % AGRUPAMENTO  EFETUADO
                 MDN [1];        % VETOR  DE  MEDIANAS
   REAL FOD,                     % VALOR  DA  FUNÇÃO  OBJETIVO  DUAL
   FOPS;                         % VALOR  DA  FUNÇÃO  OBJETIVO  PRIMAL
   REAL ARRAY D [1],             % VETOR  DE  DISTÂNCIAS
              V [1],             % VETOR  SUBGRADIENTE
              LBDA [1];          % VETOR  DE  VARIÁVEIS  DUAIS
```

```
%  ****************************************************
       BEGIN
%             ****************************************************
              PROCEDURE  QUICKSORT (A,ORD,M,N,);
%                        = = = = = =
%             ESTA  ROTINA  ORDENA  UM  CONJUNTO  A [M:N]
%             PELO  MÉTODO  QUICKSORT  ARMAZENANDO  OS
%             APONTADORES  DOS  ELEMENTOS  ORDENADOS  NO
%             VETOR  ORD [M:N]
              INTEGER M,N;                  % ÍNDICES  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;
                %
                %  PARTIÇÃO  DE  A[M:N]  PARA  A  ORDENAÇÃO
                %
                   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 [1]]<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;       % VARIÁVEIS
      BOOLEAN CHV;                %
      LABEL SEG;                  % AUXILIARES
      BOOLEAN ARRAY MG [1:N]      % VETOR QUE INDICA SE O ELEMEN
                                  % TO [I] É MEDIANA OU NÃO
%
% INICIALIZAÇÃO: CÁLCULO 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+1STEP  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
%
% INICIALIZAÇÃO: ORDENAÇÃO 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);
%
% CÁLCULO DAS MEDIANAS MDN[I],ATRAVÉS 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;
%
% CÁLCULO DA FUNÇÃO 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
    SEP  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;
        %  CÁLCULO DO PASSO T, DO ALGORITMO, E IMPLANTAÇÃO DO
        %  FATOR DE CORREÇÃO DE DIREÇÃO PARA O SUBGRADIENTE
        %  DE 0.6 (VER MULVEY & CROWDER - REFERÊNCIA BIBLIOGRÁ-
        %  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;
        %
        %  CÁLCULO DA VARIÁVEL DUAL LBDA[I] DA ITERAÇÃO SEGUINTE
        %
            FOR     I:=1
            STEP    1
            UNTIL N
            DO      LBDA[I]:=*+T*V[I]
      END DA PROCEDURE OTSBGR;
    INTEGER I,J,K,L,              % VARIÁVEIS AUXILIARES
            ITER;                 % NÚMERO DE ITERAÇÕES
    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,                     % TOLERÂNCIA
         FOP,                     % FUNÇÃO OBJETIVO PRIMAL
         FOD;                     % FUNÇÃO OBJETIVO DUAL
    REAL ARRAY LBDA[1:N],    % VARIÁVEL DUAL
               V[1:N];       % SUBGRADIENTE
    LABEL SEG;                    % VARIÁVEL AUXILIAR
%
%  INICIALIZAÇÃO
%
    TOL:=0.000000001;
    IF      TIMP=3  OR  TIMP=4  OR  TIMP=5
    THEN BEGIN
         WRITE(0TPT[SKIP  1]);
         WRITE(OTPT, <////, T10,"MÉTODO: MULVEY + CROWDER",//,
               T3,"ITER",T10,"FOPRIMAL",T24,"FOSOLUÇÃO",T38,
               "FODUAL",T50,"PRECISÃO",T62,"PRECISÃO %",//,"  ">)
```

```
        END;
    ITER:=1;
    FOR   I:=1
    STEP  1
    UNTIL M
    DO    MDN[I]:=MDNS[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;
%
%  INICIALIZAÇÃO:  CÁLCULO  DA  SOLUÇÃO  DUAL
%
    OTSBGR(N,M,AGRPS,MDN,D,V,ITER,LBDA,FOD,FOPS);
%
%  INICIALIZAÇÃO:  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);
%
%  CÁLCULO  DAS  SUCESSIVAS  SOLUÇÕES  PRIMAIS  E  DUAIS,  COM  ARMAZENA-
%  MENTO  DA  MELHOR  SOLUÇÃO  PRIMAL  OBTIDA  ATÉ  A  ITERAÇÃO  E  CÁLCULO
%  DO  TESTE  DE  OTIMALIDADE
%
    FOR   ITER:=2
    STEP  1
    WHILE ABS(PRECPERC)>0.01  AND  ITER<=NITER
    DO    BEGIN
          PSQPRM(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   0
              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;
%
% IMPRESSÃO  DOS  RESULTADOS
%
   ITER:=*-1;
   WRITE(OTPT[SKIP  1]);
   WRITE(OTPT,<///,T10,"MÉTODO:  MULVEY  +  CROWDER",//,T12,
         "NÚMERO  DE  ITERAÇÕES  =  ",I3,//  T12,"FUNÇÃO  OBJETIVO",
         "PRIMAL  =",E12.5,//,T12,"FUNÇÃO  OBJETIVO  DUAL  ="
         E12.5,//,T12,"PRECISÃO  =  ",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,      % VARIÁVEL  AUXILIAR
           P,        % NÚMERO  DE  CARACTERÍSTICAS
           N,        % NÚMERO  DE  ELEMENTOS
           TIMP,     % TIPO  DE  IMPRESSÃO
           TOTGRP,   % NÚMERO  DE  GRUPOS
           NITER,    % NÚMERO  MÁXIMO  DE  ITERAÇÕES
           MTR,      % CHAVE PARA A ESCOLHA DA MÉTRICA A SER UTILIZADA
           POR,      % CHAVE  PARA  PADRONIZAÇÃO  DAS  VARIÁVEIS
           I,J,K;    % VARIÁVEIS  AUXILIARES
   REAL FO,          % VALOR  DA  FUNÇÃO  OBJETIVO
        W;           % VARIÁVEL  AUXILIAR
```

```
%
% INICIALIZAÇÃO
%
  READ(INPT,<7I5>,P,N,TOTGRP,TIMP,NITER,MTR,PDR);
  BEGIN
       REAL ARRAY E[1:N*(N-1)/2],        % VETOR DE DISTÂNCIAS
                  CDG[1:N],              % VETOR DE CÓDIGOS DOS ELEMENTOS
                  X[1:N,1:P];            % VETOR DE OBSERVAÇÕES
       INTEGER ARRAY MÉTODO [0:7],       % VETOR DE MÉTODOS ESCOLHIDOS
                     NIG [1:N],          % TAMANHO DOS GRUPOS
                     MDN[1:TOTGRP],      % VETOR DE MEDIANAS
                     AGRP[1:N];          % AGRUPAMENTO EFETUADO
       %
       % INICIALIZAÇÃO
       %
         READ(INPT,<8I1>,MÉTODO);
         FOR    I:=1
         STEP   1
         UNTIL  N
         DO     BEGIN
                READ(INPT,<A6>,CDG[I]);
                READ(INPT,<8F10.2>,FOR J:=1 STEP 1 UNTIL P DO X[I,J])
                END;
       %
       % PADRONIZAÇÃO DOS DADOS
       %
         IF     PDR=1
         THEN   PDRZ(N,P,X);
       %
       % CÁLCULO DAS DISTÂNCIAS ENTRE OS ELEMENTOS
       %
         IF     MTR=1
         THEN   ECLD(N,E,X,P);
         IF     MTR=2
         THEN   MTCR(N,E,X,P);
       %
       % CÁLCULO DOS GRUPAMENTOS
       %
         FOR    MTD:=0
         STEP   1
         UNTIL  5
         DO     IF     MÉTODO[MTD]=1
                THEN   BEGIN
                       CLSTHRQ(N,MTD,TOTGRP,W,FO,AGRP,E,NIG,TIMP,CDG);
                       IF     MÉTODO[6]=1
                       THEN   KMDNS(AGRP,N,TOTGRP,E,MDN,FO,NITER,TIMP,CDG);
                       IF     MÉTODO[7]=1
                       THEN   MLVCRD(N,TOTGRP,AGRP,MDN,E,FO,NITER,TIMP,CDG)
                END
     END
END.
==========================================================
NUMBER OF ERRORS DETECTED = 0,
NUMBER OF SEGMENTS = 25, TOTAL SEGMENT SIZE = 1705 WORDS, CORE
ESTIMATE = 2614 WORDS. STACK ESTIMATE = 96
PROGRAM SIZE = 1041 CARDS, 5578 SYNTACTIC ITEMS, 114 DISK SEGMENTS.
PROGRAM FILE NAME: CLSTROPT. B7700 CODE GENERATED.
COMPILATION TIME = 9.509 SECONDS ELAPSED; 5.198 SECONDS PROCESSING; 8.889
SECONDS I/0.
==========================================================
```

A.2.2. — 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:

— 1.º 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 (b e 0 representam, respectivamente, espaço em branco e zero)

```
bbbbbbbbbb3bbbbbbbbb5
bbbAA1bbbb2020
bbbAB1bbbb3000
bbbAA2bbbb4220
bbbAB2bbbb5000
bbbAC1bbbb3560
```

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 $G_J, \ldots, G_M$).

BURROUGHS B7700 BDMSALGOL COMPILER, VERSION 32.111.037,
MONDAY, 10/03/83, 03:19 PM.

CLSTPD

```
% ESTA ROTINA EFETUA UMA ANÁLISE DE AGRUPAMENTO
% ATRAVÉS DO MÉTODO DE RAO — PROGRAMAÇÃO DINÂMICA
%
BEGIN
      FILE, INPT (KIND=READER),
            OTPT (KIND=PRINTER);
%     ************************************************************
      PROCEDURE QUICKSORT(A,ORD,M,N);
%                   = = = = = =
%       ESTA ROTINA ORDENA UM CONJUNTO A[M:N],
%       PELO MÉTODO QUICKSORT, ARMAZENANDO OS
%       APONTADORES DOS ELEMENTOS ORDENADOS NO
%       VETOR ORD[M:N]
%
      INTEGER M,N;                % ÍNDICES AUXILIARES
      INTEGER ARRAY ORD[1];       % VETOR DE APONTADORES
      REAL ARRAY A[1];            % VETOR A SER ORDENADO
%     ************************************************************************
            BEGIN
                  INTEGER AUX,I,K; % VARIÁVEIS
                  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;
               %
               % PARTIÇÃO DE A[M:N] PARA A ORDENAÇÃO
               %
                  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;
%
% INICIALIZAÇÃO
%
    INTEGER M, % NÚMERO MÁXIMO DE GRUPOS
              N; % NÚMERO DE ELEMENTOS
    READ(INPT,<2I10>,M,N);
    BEGIN
          REAL ARRAY X[1:N],                % VETOR DE OBSERVAÇÕES
                     W[1:N,2:M],            % MATRIZ DE CUSTOS ÓTIMOS
                     CDG[1:N];              % VETOR DE CÓDIGOS DOS ELEMENTOS
          INTEGER ARRAY ORD[1:N],           % VETOR DE ÍNDICES PARA ORDENAÇÃO
                        U[1:N,2:M];         % MATRIZ DE ESTRATÉGIA ÓTIMA
          REAL MD,      % VAR. AUX. P/ CÁLCULO DE MÉDIAS
               CSTDCS,  % VAR. AUX. P/ CÁLCULO DE CUSTO DE DECISÕES
               W1,      % CUSTO ASSOCIADO A PARTIÇÃO EM M GRUPOS
               WAC;     % VAR. AUX. P/ CÁLCULO DE CUSTO DE DECISÕES
          INTEGER U1,   % LÍDER DO GRUPO 2 NA PARTIÇÃO EM M GRUPOS
                  NJ,   % NÚMERO DE ELEMENTOS DO GRUPO J
                  I,J,IND,K; % CONTADORES AUXILIARES
%
% LEITURA DAS OBSERVAÇÕES
%
    FOR   I:=1
    STEP  1
    UNTIL N
    DO    READ(INPT,<A6,F8.3>,CDG[I],X[I]);
%
% ORDENAÇÃO DOS ELEMENTOS
%
    FOR   I:=1
    STEP  1
    UNTIL N
    DO    ORD[I]:=I;
    I:=1;
    QUICKSORT(X,ORD,I,N);
%
```

```
%  IMPRESSÃO  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]]);
%
%  INICIALIZAÇÃO  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;
%
%  CÁLCULO  DOS  CUSTOS  NO  ÚLTIMO  ESTÁGIO
%
    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;
%
%  CÁLCULO  DAS  DECISÕES  ÓTIMAS  NOS  ESTÁGIOS  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;
                  WAC:=0;

                  FOR    K:=I+2
                  STEP   1
                  UNTIL IND+1
                  DO     BEGIN
                         WAC:=*+((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:=*+1
                        END
                END
    END;
%
% CÁLCULO DA DECISÃO ÓTIMA NO ESTÁGIO 1
%
    U1:=0;
    W1:=W[2,2];
    MD:=X[ORD[1]];
    NJ:=1;
    WAC:=0;
    FOR   K:=3
    STEP  1
    UNTIL IND
    DO    BEGIN
          WAC:=*+((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:=*+1
          END;
%
% IMPRESSÃO 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 M-1
          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.
```

```
NUMBER OF ERROS DETECTED = 0.
NUMBER OF SEGMENTS = 9. TOTAL SEGMENT SIZE = 386 WORDS. CORE
ESTIMATE = 1322 WORDS. STACK ESTIMATE = 31
PROGRAM SIZE = 214 CARDS, 1099 SYNTACTIC ITEMS, 30 DISK SEGMENTS.
PROGRAM FILE NAME: CLSTPD. B7700 CODE GENERATED.
COMPILATION TIME = 4.292 SECONDS ELAPSED; 1.189 SECONDS
PROCESSING; 2.243 SECONDS I/O.
```

10 — 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. Stanford, USA, Stanford University, 1970.

3 — BALL, G. H. & HALL, D. J. — *ISODATA, a novel method of data analysis and pattern classification;* technical report. Menlo Park, USA, Stanford Research Institute, 1965. 72 p.

4 — ————. PROMENADE — *an on-line pattern recognition system;* Rep. No. RADC — TR — 67-310, AD 822174. Menlo Park, USA, Stanford Research Institute, 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. 1., *18*:311-2, 1973.

7 — COOLEY, W. W. & 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 — ————. & SIMON, J. C. — Clustering analysis. *Communication and Cybernetics,* Heidelberg, *10*:46-94, 1976.

10 — DURAN, B. S. & 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*:361-8, 1966.

12 — EVERITT, B. — *Cluster Analysis.* London, Heinemann Educational Books, 1974.

13 — FISHER, M. L.; NEMHAUSER, G. L. & WOLSEY, L. A. — An analysis of approximations for Maximizing Submodular Set Functions — I. *Mathematical Programming,* Amsterdam, *14*:265-94, 1978.

14 — ————. An analysis of approximations for Maximizing Submodular Set Functions — II. *Mathematical Programming Study,* Amsterdam, *8*:73-87, 1978.

15 — FORGY, E. W. — Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics,* Washington D. C., *21*(3):768, 1965.

16 — GOWER, J. C. — A general coefficient of similarity and some of its properties. *Biometrics,* Washington D. C., *27*:857-74, 1971.

17 — HANSEN, P. & DELATTRE, M. — Complete-link cluster by graph coloring. *Journal of the American Statistical Association,* Washington D. C., *73*(362):397-403, 1978.

18 — HARMAN, H. H. — *Modern Factor Analysis.* Chicago, The University of Chicago Press, 1976. 487 p.

19 — HARTIGAN, J. A. — *Clustering Algorithms.* New York, John Wiley and Sons, 1971.

20 — HELD, M.; WOLFE, P. & CROWDER, H. P. — Validation of Subgradient Optimization. *Mathematical Programming,* Amsterdam, *6*:62-88, 1974.

21 — JANCEY, R. C. — Mutidimensional group analysis. *Austral. J. Botany,* s. 1., *14*(1):127-30, 1966.

22 — JARVINEN, P.; RAJALA, J. & SINERVO, M. — A branch-and-bound algorithm for seeking the p-median. *Operations Research,* Baltimore, *20*:173-8, 1972.

23 — JENSEN, R. E. — A dynamic programming algorithm for cluster analysis. *Operations Research,* Baltimore, *17*:1034-57, 1969.

24 — JOHNSON, S. C. — Hierarchical clustering schemes. *Psychometrica,* Williansburg, *32*:241-54, 1967.

25 — LANCE, G. N. & WILLIAMS, W. T. — A generalized sorting strategy for computer classifications. *Nature,* New York, : 212-18, 1966.

26 — ————. A general theory of classificatory sorting strategies. I — Hierarchical Systems. *Computer Journal,* London, *9*(4):373-80, 1966.

27 — ————. A general theory of classificatory sorting strategies. II — Clustering Systems. *Computer Journal,* London, *10*(3):276, 1967.

28 — MACNAUGHTON-Smith, P. — Some statistical and other numerical techniques for classifying individuals. *Home Office Research Unit Report No. 6,* London, HMSO, 1965.

29 — MACQUEEN, J. B. — Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probab., 5th, Berkeley,* Berkeley, University of California, *1*:281-97, 1967.

30 — MATEUS, G. R. & 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.

31 — MULVEY, J. M. & CROWDER, H. P. — Cluster analysis: an application of lagrangian relaxation. *Management Science,* Providence, *25*(4):329-40, 1979.

32 — PINHO GAMA, M. — *Bases da Análise de Grupamento.* Brasília, Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Estatística, 1980. 229 p.

33 — PFEIFFER, D. — Disparidades de Desenvolvimento no Brasil — um exemplo da análise de cluster. *Revista Brasileira de Estatística,* Rio de Janeiro, IBGE, *41*(164):559-76, out./dez. 1980.

34 — RAJ, D. — *Sampling Theory.* New Delhi, Tata Mcgraw — Hill Publishing, 1978. 302 p.

35 — RAO, M. R. — Cluster analysis and mathematical programming. *Journal of the American Statistical Association,* Washington D. C., *66*:622-6, 1971.

36 — ROHLF, F. J. — Hierarchical clustering using the minimum spanning tree. *The Computer Journal,* London, *16*:93-5, 1973.

37 — ROY, B. — An algorithm for a general constrained set covering problem; *in*: *Praph Theory and Computing.* New York, Academic Press, : 267-83, 1972.

38 — SIBSON, R. — SLINK: an optimally efficient algorithm for the single — link cluster method. *The Computer Journal,* London, *16*:30-4, 1973.

39 — SOKAL, R. R. & MICHENER, C. D. — A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletim,* Kansas, : 1409-38, 1958.

40 — SUDMAN, S. — *Applied Sampling.* New York, Academic Press, 1976. 251 p.

41 — TATSUOKA, M. M. — *Multivariate analysis: techniques for educational and psychological research.* New York, John Wiley and Sons, 1971. 310 p.

42 — TRYON, R. C. — *Cluster Analysis.* Ann Arbor, Michigan, Edwards Bros., 1939.

43 — VINOD, H. D. — Integer programming and the theory of grouping. *Journal of the American Statistical Association,* Washington D. C., *64*:506-19, 1969.

44 — WARD, J. H. — Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* Washington D. C., *58*(301):236-44, 1963.

45 — WESOLOWSKY, G. O. — *Multiple regression and analysis of variance.* New York, John Wiley and Sons, 1976. 292 p.

46 — WISHART, D. — An algorithm for hierarchical classifications. *Biometrics,* Washington D. C., *22*(1):165-70, 1969.

47 — WISHART, D. — Mode analysis: a generalization of nearest neighbor which reduces chaining effects, em A. J. Cole (ed.) *in: Numerical Taxanomy.* New York, Academic Press, : 282-319, 1969.

48 — ZAHN, C. T. — Graph — theoretical methods for detecting gestalt clusters. *IEEE Transactions on Computers, 20*:68-86, 1971.

RESUMO

São apresentados alguns dos métodos mais significativos de análise de grupamento, define-se o problema de análise de grupamento e seus conceitos básicos e apresentam-se os principais métodos hierarquizados 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 e são fornecidas rotinas computacionais em ALGOL para a resolução de problemas pelos principais métodos descritos.

ABSTRACT

Some of the most significative methods of cluster analysis are presented and the problem of cluster analysis and its basic concepts are defined. A 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 larvas growing stages and computer programs in ALGOL for the solution of practical problems by the main described methods are given.

CUSTOS COMPARATIVOS NA AGRICULTURA BRASILEIRA - ANÁLISE DE ALGUNS PRODUTOS A NÍVEL DE MESORREGIÃO HOMOGÊNEA

Jairo Augusto Silva

SUMÁRIO

R. bras. Estat., Rio de Janeiro, *43*(172) : 725-794, out./dez. 1982

1 — INTRODUÇÃO

Os estudos sobre custos da atividade agrícola tratam normalmente da compilação e análise, em termos monetários, das despesas direta ou indiretamente relacionadas aos fenômenos envolvidos na produção de produtos agrícolas. Usualmente, o propósito mais imediato de tais estudos é a determinação do custo unitário de um produto ou de uma cesta de produtos, a fim de possibilitar, a nível de empresa ou de estabelecimento, o conhecimento de sua viabilidade econômica e, do ponto de vista das autoridades governamentais, possibilitar a introdução de políticas de controle e incentivo à produção, quer seja através da fixação de preços mínimos, tabelamento de preços, regularização do comércio, quer seja através da criação de subsídios, taxas sobre o comércio exterior, etc.

Neste trabalho, pelas próprias características dos dados utilizados, dados censitários, procurou-se determinar a estrutura dos custos de produção de alguns produtos agrícolas, com a preocupação fundamental de análise comparativa entre diversas regiões produtoras sem mais rigidez na estimativa pura e simples do custo unitário de produção. Na verdade, o simples confronto entre a estrutura de custos e os aspectos da produção agrícola de diversas regiões produtoras nacionais pode levantar subsídios para um melhor direcionamento de políticas de desenvolvimento regional e para o próprio zoneamento da atividade agrícola no País.

2 — METODOLOGIA

2.1 — Tipologia agrícola e seleção de produtos

A metodologia utilizada na estimação dos custos de produção se baseou no relativamente alto grau de especialização, no cultivo de alguns produtos, atingido nas últimas décadas em certas regiões produtoras do País. Por grau de especialização pode-se entender duas relações, segundo se queira considerar a produção independentemente do número de estabelecimentos produtores, ou vice-versa. Assim, pode-se conceituar um índice de especialização da produção em determinada área como sendo a relação entre as quantidades colhidas de um produto qualquer nos estabelecimentos especializados, em que o valor da produção deste produto representa mais de 80% do valor da produção total do estabelecimento, e as quantidades colhidas na área como um todo. Similarmente, pode-se conceituar um índice de especialização dos estabelecimentos, na mesma área, representando a relação entre o número de estabelecimentos especializados na produção de um produto qualquer e o número total de estabelecimentos produtores.

O primeiro índice, da produção, indica a participação dos produtores especializados na produção total da região em análise, não podendo ser, portanto, superior a um. Quanto menor o índice de especialização da produção, evidentemente menor será esta participação. O índice dos estabelecimentos, por sua vez, indica a importância do número de produtores especializados no número de produtores da região. Na verdade, este índice, analisado isoladamente, pode sugerir a extensão das explorações especializadas, sem, no entanto, qualquer indicação quanto ao seu peso na produção, dado pelo primeiro índice.

O produto dos dois índices, da produção e dos estabelecimentos, é que realmente pode fornecer uma informação mais geral acerca da especialização dos cultivos numa região, considerando tanto a produção obtida quanto o número de estabelecimentos envolvidos. Este novo índice, que pode ser chamado de índice "verdadeiro" de especialização dos produtos, varia de zero a um, segundo a inexistência de estabelecimentos especializados, ou conforme todos os estabelecimentos produtores de uma determinada região sejam especializados. Observe-se que esse índice nada mais é do que a relação entre o número de produtores especializados e o número total de produtores, ponderada pela participação dos primeiros na produção total da região. A importância do índice de especialização dos cultivos, neste trabalho, se deve não apenas à sua utilidade na distinção de desigualdades na exploração agrícola entre regiões, como, principalmente, à sua utilização como critério básico para a seleção dos estabelecimentos a serem analisados, como será descrito no comentário sobre os dados utilizados.

É claro que se o critério de especialização é satisfatório para a seleção de estabelecimentos, não o é para a escolha das áreas e dos produtos a serem estudados. No caso das áreas a serem consideradas, optou-se pela seleção das mesorregiões que apresentaram significativos excedentes de produção, conforme resultados apresentados no trabalho "Balanço uso-disponibilidade de 15 produtos agrícolas alimentares — uma análise a nível mesorregional" [1]. Tal procedimento visou não apenas a uma maior limitação deste estudo, em termos de volume de dados a serem analisados, como também a fornecer maiores informações acerca das principais zonas produtoras excedentárias, responsáveis por significativa parcela da produção agrícola nacional.

Na escolha dos produtos agrícolas a serem analisados, por outro lado, quatro critérios básicos foram estudados. O primeiro foi considerar a importância relativa do produto, em função do seu valor estratégico no abastecimento interno, quer seja como matéria-prima para agroindústria, quer seja como alimento *in natura*. Segundo este critério, obviamente pouco restritivo, um grande número de produtos deveria

[1] Silva, Jairo Augusto & Rocha, Sonia — *Balanço uso-disponibilidade de 15 produtos agrícolas alimentares — uma análise a nível mesorregional.*

ser considerado, sendo que, para parte significativa deles, os resultados conseguidos provavelmente seriam desnecessários ou mesmo sem sentido, frente às limitações dos dados censitários a serem utilizados. É o caso dos produtos de lavouras permanentes em que os gastos efetuados na produção se distribuem pelo número de anos correspondentes ao ciclo produtivo da espécie explorada. Portanto, o segundo critério para a seleção dos produtos deveria se relacionar às restrições impostas à análise pela utilização dos dados censitários. Como terceiro critério, em vista do objetivo de análise comparativa, optou-se pelos produtos com maior distribuição espacial em seus cultivos, justamente visando a uma melhor determinação das características e tipicidades das explorações regionais. Finalmente, como quarto critério, foram considerados os produtos responsáveis por mais de 80% do valor da produção total de um número de estabelecimentos suficientemente grande e expressivo dentro de cada mesorregião homogênea[2].

Assim, foram selecionados apenas cinco produtos das lavouras temporárias: o arroz em casca, a batata-inglesa, a cana-de-açúcar, a cebola e o milho em grão.

Tomando-se, portanto, os dados agregados dos estabelecimentos especializados na produção de cada um desses produtos e em cada uma das mesorregiões selecionadas, pode-se chegar a uma estimativa das despesas totais despendidas nessa produção, assim como a um perfil médio da estrutura de custos nas mesorregiões consideradas.

Os principais indicadores utilizados nessa análise comparativa foram construídos segundo os critérios usualmente utilizados nas análises de custo a nível de empresas e em análises globais, em sua maioria experimentais, já que raramente são encontrados na bibliografia especializada.

Nos tópicos seguintes, procurar-se-á melhor esclarecer alguns dos aspectos metodológicos até agora sucintamente levantados.

2.2 — Os dados

Os dados utilizados foram os de uma listagem especial do Censo Agropecuário de 1975, em que foram selecionados, por mesorregião homogênea, todos os estabelecimentos agropecuários em que o valor da produção dos produtos considerados se equivalesse ou fosse superior a 80% do valor da produção total dos estabelecimentos, definido como o somatório dos produtos das quantidades de bens produzidos pelos preços médios de venda.

Obviamente, a utilização dos dados censitários para uma análise particular de cada estabelecimento, ou de pequenos grupos de estabelecimentos, pode levar a erros significativos e mesmo não avaliáveis em

[2] A soja e o trigo não foram considerados, devido a não satisfazerem, em 1975, ao terceiro critério; o feijão, por não satisfazer ao quarto critério, etc.

termos estatísticos, por várias razões, dentre as quais a imensa diversidade de combinações de fatores de produção na atividade agropecuária e o próprio sistema de levantamentos de dados podem ser consideradas as principais.

Assim, se diferenças substanciais nas condições climáticas e meteorológicas, no sentido amplo, podem condicionar o emprego de diferentes formas tecnológicas de produção, exigindo maior extensão de áreas pesquisadas, também as técnicas de levantamento de dados condicionam a validade das análises realizadas, quer sejam individuais por estabelecimento, quer sejam agregadas por grupos de estabelecimentos.

Na realidade, a explicação desse último condicionamento, no caso das estatísticas censitárias nacionais, está na forma de levantamento, tradicionalmente subjetivo do Censo Agropecuário e no seu período de referência, que é o ano civil imediatamente anterior ao período de coleta de dados. É claro que diversos outros fatores também importantes, como a própria defasagem de tempo existente entre o período de referência e o período de coleta de dados, o pouco preparo de uma grande parte dos recenseadores e a própria infra-estrutura da rede de coleta, influenciam na qualidade final dos dados. No entanto, parece que os dois primeiros fatores levantados, ou sejam, a informação subjetiva e o ano civil como período de referência, é que podem originar maiores tendências e erros nos dados censitários.

Há de se considerar porém, que o exposto deve ser válido e sopesado nos casos de análises individuais de estabelecimentos e em conjunto de estabelecimentos não estatisticamente significantes, não sendo necessariamente válido para as análises em que a ciência estatística é utilizada adequadamente.

Outro problema apresentado pela utilização dos dados censitários refere-se à especificação das despesas arroladas como tal no questionário geral do Censo.

Estas despesas, em 1975, consistiam no pagamento de salários ao pessoal temporário e permanente empregado nos estabelecimentos, na quota-parte entregue a parceiros, no arrendamento de terras, na aquisição de adubos e corretivos, de sementes e mudas, de defensivos agrícolas, de medicamentos para animais, de rações, de sal, no pagamento dos aluguéis de máquinas e equipamentos, dos serviços de empreitadas com equipamentos e mão-de-obra ou apenas mão-de-obra, do transporte da produção, dos juros e despesas bancárias, dos impostos e taxas, e finalmente, de outras despesas efetuadas durante o ano civil no estabelecimento, e diretamente relacionados à atividade.

Cada uma dessas despesas apresenta certas impropriedades para uma análise de custo de produção, impropriedades estas que podem ser gerais ou comuns a todos os itens arrolados e derivadas de critérios adotados nos levantamentos censitários ou específicas de cada uma das

despesas. Assim, como impropriedades comuns que mais diretamente afetam as estimativas de custo da produção, pode-se citar o critério de não se incluírem nas despesas levantadas, os valores dos insumos de produção própria do estabelecimento e a instrução para a avaliação dos pagamentos feitos em espécie. A primeira destas impropriedades deve afetar significativamente a atividade pecuária, pela não consideração da alimentação animal produzida no estabelecimento e, a atividade agrícola, pela não consideração dos adubos orgânicos e das sementes e mudas de produção própria. O mesmo deve ocorrer com o transporte da produção e outros itens das despesas, em maior ou menor grau, segundo a importância das mesmas na exploração agropecuária. A segunda impropriedade deve afetar a análise de custos através, principalmente, dos salários pagos e da quota-parte entregue a parceiros, apesar de, em princípio, todos os pagamentos realizados pelo produtor rural poderem ser feitos em espécie.

Com respeito às impropriedades específicas de cada um dos itens de despesas, se originam elas, na maioria das vezes, da dificuldade de avaliação de alguns de seus componentes. Assim, no caso dos salários pagos, estão incluídos apenas os pagamentos aos trabalhadores contratados pelo produtor rural, inclusive ao administrador ou responsável pela direção do estabelecimento, justamente pela dificuldade de avaliação do trabalho executado pelo proprietário e membros de sua família. A mesma dificuldade de avaliação é que justifica a não inclusão, como despesa, do valor alternativo da terra utilizada na produção. Aliás a consideração ou não desse valor como custo de produção, não é ainda totalmente aceita na administração rural, devido à inexistência de critérios claramente definidos e gerais de avaliação.

O mesmo acontece com a imputação das despesas com máquinas e equipamentos utilizados. Na verdade, o arrendamento de terras e o aluguel de máquinas desvirtuam as análises de custos comparativos, desde que a distribuição destas despesas não é uniforme pelos estabelecimentos agropecuários. A solução desse problema será discutida nos próximos tópicos.

2.3 — Composição dos custos

Numa função de produção agrícola, simplificadamente, a produção obtida é função da área cultivada, da fertilidade da área, das condições climáticas, dos insumos utilizados, do capital empregado, do trabalho exigido e do nível de tecnologia existente.

Nas condições de análise agregada de dados censitários, evidentemente é difícil, senão impossível, a consideração de aspectos ligados à fertilidade da terra utilizada e mesmo das condições meteorológicas ocorridas no ano base do estudo. Assim, em princípio, numa derivação da função de produção agrícola de cada uma das mesorregiões em

estudo, pode-se aceitar a existência de uma função de custos de produção em que apenas os fatores terra, capital e trabalho são considerados. A especificação dos insumos agrícolas, como sendo fatores de capital ou como recursos naturais, é de menor importância para os objetivos pretendidos e, portanto, não deverá ser considerada.

2.3.1 — A terra

O custo da terra na produção agropecuária é da maior importância, particularmente nas regiões mais densamente povoadas do País.

É sabido que, na última década, o preço das terras apresentou um grande crescimento, causado quer pela conjuntura inflacionária, que transformou as áreas rurais em disputados bens para a reserva de valor, quer pelas facilidades de financiamento e incentivos dados pelo Governo às atividades rurais. A consideração de parcela do valor da terra utilizada como componente do custo de produção agropecuária, é baseada usualmente nos critérios da contabilidade nacional para efeitos de comparações internacionais e nos manuais de administração rural, ou na argumentação de que os investimentos realizados na aquisição de terras teriam alternativas de rendimentos positivos fora do setor agropecuário, donde a imputação desses rendimentos não auferidos como custos adicionais à produção agropecuária. Ora, tal raciocínio parece perfeitamente ajustado para uma análise de investimentos e numa situação em que a terra apareça como um bem seguro para a reserva de valor, mas não necessariamente numa conjuntura em que a terra se apresente como um atraente investimento especulativo. Assim, pode-se decompor o valor da terra em dois componentes: o primeiro, referente às suas potencialidades de produção e de geração de renda agrícola e o segundo, referente a possíveis ganhos em outras atividades econômicas. Deve ficar claro que apenas o primeiro componente do valor das terras é que deve ser considerado na avaliação dos custos de produção agropecuária uma vez que se procure analisar somente esta atividade.

Observe-se que estas colocações contrariam frontalmente o critério, usualmente utilizado, para a imputação do custo da terra no custo total da produção através da aplicação de uma taxa média de retorno do capital investido, taxa esta consistente com a normalmente obtida no mercado. A avaliação da parcela do custo de produção agropecuária derivada do valor da terra deve ser, portanto, relacionada com a potencialidade de produção desse fator, sendo que um dos melhores indicadores de tal potencialidade é o valor dos arrendamentos das terras para a produção agropecuária.

Neste trabalho procurar-se-á imputar no custo da produção obtida o custo da terra utilizada, estimado segundo os valores médios dos arrendamentos em cada uma das mesorregiões estudadas. O fato de ter sido levantado pelo Censo Agropecuário de 1975 o valor dos arrenda-

mentos dos estabelecimentos como um todo e não conforme a utilização das áreas, possivelmente deve originar subestimações no custo da terra agricultável, o que no entanto, não deve invalidar as estimações feitas neste trabalho.

2.3.2 — O capital

Uma das parcelas mais difíceis para se calcular no custo de produção agropecuária é a referente à utilização do capital agrário. Para efeito de melhor ordenamento do trabalho, pode-se entender o capital como sendo de dois tipos, quanto ao seu comportamento na produção: fixo e circulante.

Por capital fixo será entendido o capital auxiliar da produção, de uma maneira contínua, servindo a mais de um ciclo produtivo. São os edifícios, as plantações permanentes, as máquinas e implementos, as instalações, o gado leiteiro, de trabalho, os reprodutores, etc. O capital circulante, por sua vez, será o que auxilia a produção de um modo instantâneo, servindo a um só ciclo técnico produtivo, como sementes, adubos, forragens, etc[3].

2.3.2.1 — Capital fixo

"A disponibilidade de capital implica em quatro tipos de custos: juros, conservação, riscos e depreciação ou amortização". "Usualmente, a todo capital empregado na produção, quer de propriedade do empresário, quer obtido por via do crédito, deve-se atribuir um juro, calculado a uma taxa normal, em aplicações de risco equivalente"[4]. A conservação ou manutenção dos bens de capital, por sua vez, representa o "custo anual necessário para manter os bens em condições de uso". Os custos de riscos representam os gastos efetuados para a cobertura de danos imprevistos na atividade, enquanto que a depreciação consiste no "custo necessário para a substituição dos bens de capital inutilizados pelo tempo, ou pelo uso, ou simplesmente, tornados obsoletos devido a inovações tecnológicas".

Neste trabalho, apenas serão imputados como custos de produção, o valor da depreciação do capital existente nos estabelecimentos, desprezando-se a imputação dos juros sobre o capital próprio do produtor, por considerá-los embutidos no lucro líquido obtido na atividade. Os juros sobre o capital obtido por empréstimo, serão considerados como custos de produção e foram levantados explicitamente pelo Censo Agropecuário de 1975.

---

[3] Oliveira, Contalício Preto de. *Economia e Administração Rural.*

[4] Hoffmann, Rodolfo *et alii. Administração da Empresa Agrícola.*

Da mesma maneira, foram levantados os custos com a conservação ou manutenção do capital, nos pagamentos dos seguros contra danos e perdas imprevistos, podendo-se dizer não ser tal prática muito difundida no País, a não ser as parcelas correspondentes ao seguro dos "financiamentos de custeio ou investimento concedidos por instituições financeiras e de parcela de recursos próprios do produtor, prevista no instrumento de crédito, segundo critérios aprovados pelo Conselho Monetário Nacional" [5]. Tais pagamentos de seguros, conforme as instruções do Censo Agropecuário de 1975, igualmente devem ter sido arrolados pelos produtores como despesas de produção.

Na estimação dos custos de produção relacionados ao capital agrário, particularmente no caso da depreciação dos bens, há de se considerar que a intensidade de utilização dos mesmos é que deve condicionar o tempo de duração desses bens; antes até que a própria ação da natureza ou obsolescência técnica do capital. Assim, o emprego de períodos de tempo médios para o cálculo da depreciação, como será utilizado neste trabalho, sem a consideração da intensidade de uso nas diferentes mesorregiões em análise, deve originar erros na avaliação dos custos de produção, aparecendo como uma das limitações às estimativas de custo total. O próprio período de duração dos bens, sob condições normais de uso, varia de região para região e mesmo na opinião de técnicos e estudiosos do assunto, como mostrado no quadro 1.

QUADRO 1

TAXAS DE DEPRECIAÇÃO DE BENS POR FONTES DIVERSAS

BENS	FONTES		
	Contabilidade agrícola (Armando Aloe e Francisco Valle) (anos)	*Administração de* empresa agrícola (anos)	Normas contábeis e traços financeiros (anos)
Construções	20 (5%)	10 a 50	25 (4%)
Benfeitorias	20 (5%)	15 a 50	—
Instalações agrícolas	13,2 (7,5%)	15 a 50	10 (10%)
Cafezal	25 (4%)	—	—
Tratores	6,66 (15%)	10	4 a 5 (20 a 25%)
Máquinas agrícolas	10 (10%)	10 a 30	10 (10%)
Implementos	10 (10%)	5 a 20	—
Veículos a motor	6,666 (15%)	10	6,66 a 3 (15 a 33%)
Carros e arreios	10 (10%)	10	5 (20%)
Lonas e encerados	5,88 (17%)	6	—
Móveis e utensílios	10 (10%)	—	10 (10%)
Forraria	10 (10%)	—	6,66 (15%)
Carpintaria	10 (10%)	—	—
Animais de trabalho	10 (10%)	5 a 12	5 (20%)
Bois de carro	6,66 (15%)	5	—
Cavalos de sela	10 (10%)	8	—

[5] Almeida, Manoel Lizardo de. Seguro agrícola e proagro. *Lavoura Arrozeira.*

Neste trabalho, considerando-se os bens levantados pelo Censo Agropecuário de 1975, serão utilizados as seguintes taxas de depreciação anual (Quadro 2):

QUADRO 2

TAXAS DE DEPRECIAÇÃO UTILIZADAS

BENS	TAXA DE DEPRECIAÇÃO ANUAL (%)
Prédios residenciais e para fins sociais.....	2
Culturas permanentes.....	4
Veículos e outros meios de transporte.....	10
Máquinas e instrumentos agrários.....	10
Instalações e outras benfeitorias.....	5

O valor das terras, das matas plantadas e dos animais de criação e trabalho, não serão considerados. O primeiro, por já ter sido avaliado como fator de produção independente e, o segundo, justamente pela dificuldade de avaliação, além do que, em princípio, poder-se ajuizar que as matas plantadas tendem a ter seus valores crescentes com o tempo, dependendo da espécie explorada, o que certamente deveria ser calculado e imputado ao se estimar a riqueza do setor, o que foge aos objetivos do trabalho.

Finalmente, quanto ao valor dos animais de criação e de trabalho, o alto valor residual desses animais ao final de suas vidas úteis, assim como a existência de elementos subjetivos de avaliação do mesmo, justifica em parte, a não consideração desse valor no cálculo da depreciação. Ademais, a não especificação do número de animais por espécie e idade, impede qualquer exercício mais sério de avaliação da depreciação desses bens.

O método a ser utilizado para o cálculo da depreciação dos bens agropecuários será o linear, ou das cotas fixas, uma vez que se considerou a depreciação simplesmente, como a desvalorização dos bens ao longo do tempo.

O cálculo da depreciação será feito segundo a fórmula:

$$Di = \frac{\sqrt{i} - \sqrt{F}}{n}, \text{ onde:}$$

$Di =$ valor da depreciação do bem, no ano i,

$\sqrt{i} =$ valor do bem ao final do ano i,

$\sqrt{F}$ = valor residual ou de sucata do bem ao final do seu período de utilização,

n = número de anos de duração ou de vida útil do bem.

Apenas para os veículos e outros meios de transporte e para as máquinas e instrumentos agrários é que será estipulado um valor residual, ou de sucata, correspondente a 10% do valor desses bens.

2.3.2.2 — Capital circulante

O capital circulante, entendido como as despesas ligadas à produção obtida no ano, foi levantado pelo Censo Agropecuário de 1975, compreendendo as despesas realizadas com adubos e corretivos, sementes e mudas, defensivos agrícolas, medicamentos e alimentação dos animais, transporte da produção, impostos e taxas, juros e despesas bancárias, aluguel de máquinas e equipamentos e outras despesas. O capital circulante é de fundamental importância na atividade agropecuária, uma vez que é "absorvido" no processo produtivo não devendo, portanto, ser superior ao valor da produção obtido.

2.3.3 — O trabalho

Como fator de produção agropecuária, o trabalho, a ser considerado nesta análise, será apenas o trabalho humano, distinguindo-se quatro avaliações, segundo as relações do trabalhador com a terra: salários pagos aos trabalhadores permanentes e temporários, quota-parte entregue a parceiros, serviços de empreitada, "pró-labore" e/ou o valor do trabalho executado pelo produtor e familiares. Os salários pagos aos trabalhadores permanentes e temporários foram devidamente levantados pelo Censo Agropecuário de 1975, não apresentando, em princípio, maiores dificuldades na utilização desses valores. Quanto à quota-parte entregue a parceiros, também levantada pelo Censo, há algumas restrições em se considerá-la como pagamento ao fator trabalho, desde que nas diversas formas de parceria existentes (meeiros, terceiros, etc.), geralmente estão consideradas remunerações a outros fatores de produção que não o trabalho, tais como sementes, adubos, defensivos, etc. Numa consideração mais geral, pode-se, no entanto, admitir que parcela substancial da quota-parte entregue a parceiros seja destinada a remunerar o trabalho. Os serviços de empreitada, da mesma maneira que a quota-parte entregue a parceiros, incluem também pagamentos a outros fatores de produção que não o trabalho humano.

O Censo Agropecuário de 1975 pesquisou dois tipos de pagamentos a serviços de empreitada: os serviços que utilizaram equipamentos e mão-de-obra e os serviços que usaram apenas mão-de-obra. Para o Brasil, como um todo, o primeiro tipo de despesa representou cerca de 36% do total dos pagamentos feitos a empreiteiros, podendo-se aceitar ser este

percentual, o limite superior dos pagamentos a fatores de produção que não o trabalho.

Na impossibilidade de se avaliar a verdadeira participação dos pagamentos ao trabalho nesse tipo de empreitadas, aceitaremos ser este fator exclusivamente responsável por estas despesas. A fim de se atenuar um pouco tal hipótese, deve-se observar que a participação das despesas com empreitadas, com o uso de equipamentos e mão-de-obra representava em 1975 para o Brasil, como um todo, apenas 3,3% das despesas totais realizadas pelos estabelecimentos e cerca de 14% dos pagamentos realizados ao fator trabalho tal como definido neste trabalho (salários + quota-parte entregue a parceiros + pagamentos a serviços de empreitadas).

Mesmo considerando estes percentuais como limites superiores dos pagamentos a outros fatores, eles são significativos e devem ser levados em conta na análise comparativa entre regiões.

Como último dos componentes dos custos com o fator trabalho no setor agropecuário, há de se considerar o valor do trabalho despendido pelo produtor e membros de sua família.

Este valor, pelo alto grau de subjetividade embutido em sua estimação ou avaliação, não é levantado pelos Censos Agropecuários, devendo portanto ser imputado em uma análise de custos, mesmo no nível de agregação e num estudo experimental como é este trabalho.

Considerado o produtor proprietário, pode-se supor que ele desempenha cinco funções do ponto de vista de levantamento de custo: a primeira, é a capitalista, "cuja remuneração caberia sobre os juros sobre o capital", já considerada anteriormente. A segunda, é a "iniciativa da produção, conjugando e harmonizando dentro da empresa todos os fatores que concorrem para a atividade produtiva". A terceira, seria o "exercício da direção da empresa despendendo energia física e esforço intelectual, sujeitando-se a uma série de tensões psíquicas e preocupações morais, assim como renunciando a empregar o seu tempo e a sua capacidade em outras atividades menos penosas". A quarta, seria a sujeição consciente "aos riscos inerentes às operações técnico-agrícolas e comerciais e, portanto, a sofrer prejuízos" [6]. A quinta, finalmente, seria a de auxiliar e executar serviços normalmente desenvolvidos por seus empregados.

As quatro primeiras funções, que poderiam representar parcelas do custo de produção, estão normalmente embutidas no lucro líquido de atividade, significando a remuneração paga ao empresário e capitalista, não sendo, portanto, consideradas nesta análise.

A quinta função, no entanto, representa a execução de um trabalho perfeitamente passível de ser substituído através da contratação dos

[6] Aloe, Armando & Valle, Francisco. *Contabilidade Agrícola.*

serviços de outras pessoas e, como tal, deve ser imputado como pagamentos de salários nos custos de produção. A metodologia para a avaliação desse valor, apesar de apresentar sérias inconveniências, será a utilização do salário médio pago aos trabalhadores permanentes e temporários como a remuneração paga pelo trabalho do produtor e membros de sua família.

Observe-se que as despesas derivadas das quatro primeiras funções do produtor são condicionadas à existência de lucro líquido na atividade, não sendo consideradas em situações com resultados negativos. A avaliação do trabalho do produtor e de seus familiares, no entanto, deve ser considerada independentemente do desempenho econômico-financeiro do estabelecimento agropecuário.

3 — OS PRODUTOS ANALISADOS

3.1 — Arroz

As mesorregiões selecionadas para este trabalho, segundo o critério de importância da produção excedentária exposto na metodologia geral, foram as Mesorregiões: Triângulo Mineiro, no Estado de Minas Gerais, Goiânia, em Goiás, Leste Rio-Grandense e Oeste Gaúcho, no Estado do Rio Grande do Sul [7]. Estas quatro Mesorregiões representavam, em 1975, cerca de 31,58% do valor da produção total de arroz no País, 17,24% da área e 28,86% da quantidade colhida e, apenas, 3,88% do número de estabelecimentos produtores de arroz. A nível de Unidade da Federação (UF) e, apenas considerando o valor da produção do produto, na Mesorregião Triângulo Mineiro representava 29,58% do total do Estado, na de Goiânia 30,17%, na Leste Rio-Grandense 51,32% e na Oeste Gaúcho 39,85%, sendo que estas duas últimas, juntas, foram responsáveis por mais de 91% da produção de arroz no Estado do Rio Grande do Sul.

Considerando apenas os estabelecimentos especializados na produção de arroz, conforme conceituados na metodologia geral, as relações entre o valor da produção desses estabelecimentos e o valor da produção nas Mesorregiões em estudo e nas UFs decrescem significantemente de 13,47 para 3,98% na Triângulo Mineiro, de 24,45 para 7,38% na de Goiânia, de 44,02 para 22,59% na Leste Rio-Grandense e de 38,83 para 15,48% na Oeste Gaúcho, respectivamente. Observe-se que, a propósito da primeira relação, ela indica o grau de especialização da produção nas regiões consideradas, podendo-se afirmar ter sido a Mesorregião Leste Rio-Grandense a mais especializada na produção de arroz, dentre as quatro consideradas, vindo a seguir a Oeste Gaúcho, Goiânia e Triângulo Mineiro.

[7] Nas duas últimas Mesorregiões predomina a exploração do arroz irrigado, enquanto que o arroz de sequeiro é o mais cultivado nas mesorregiões do centro-sul do País.

É claro que, se a relação entre o valor da produção do produto nos estabelecimentos especializados e o valor da produção do produto nas Mesorregiões em questão, indica o grau de especialização dessa produção, quer seja em termos de valor, quer seja em termos de quantidades, nada pode informar sobre o grau de especialização da produção entre os estabelecimentos produtores. Para este grau de especialização pode-se, no entanto, usar a razão entre o número de estabelecimentos especializados e o número total dos estabelecimentos produtores, em cada uma das Mesorregiões analisadas. Assim, ambos os índices de especialização, da produção e do número de estabelecimentos, podem ser usados como indicadores do nível de desenvolvimento tecnológico alcançado nas regiões em estudo, como da própria margem de aceitação das análises agregadas que são propostas neste trabalho. Na utilização desses índices como indicadores de desenvolvimento tecnológico está, evidentemente, a inferência de que os produtores que se dedicam à exploração de um produto, de maneira preponderante, tendem a um melhor conhecimento

das técnicas produtivas e das forças de mercado, do que os produtores com maior diversificação em sua produção.

Para as Mesorregiões consideradas, o quadro a seguir, mostra os índices de especialização encontrados:

QUADRO 3.1.1

ÍNDICES DE ESPECIALIZAÇÃO NA CULTURA DO ARROZ

ESPECIFICAÇÃO	ÍNDICES (%)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Índice de especialização da produção...	13,47	24,45	44,02	38,83
Índice de especialização dos estabelecimentos.....	8,51	14,69	16,74	18,92
Índice verdadeiro de especialização....	1,15	3,59	7,38	7,35

Cada um desses itens do custo total de produção deve ser analisado sucintamente, a fim de se caracterizar as diferenças básicas na estrutura de custos das Mesorregiões em estudo.

QUADRO 3.1.2

ESTRUTURA DE CUSTOS DA PRODUÇÃO NOS ESTABELECIMENTOS ESPECIALIZADOS NA CULTURA DO ARROZ

ESPECIFICAÇÃO	CUSTOS DA PRODUÇÃO (%)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
CUSTO TOTAL.....	**100,00**	**100,00**	**100,00**	**100,00**
Custo da terra.....	29,96	19,59	20,31	16,68
Custo do capital fixo.....	6,60	5,24	9,78	12,45
Custo do capital circulante.....	25,61	37,64	48,72	51,89
Custo do trabalho.....	37,83	37,53	21,19	18,98

De início, observe-se o alto custo relativo da terra na Mesorregião Triângulo Mineiro (cerca de 30% do custo total) em contraposição aos aproximadamente 20% na de Goiânia e na Leste Rio-Grandense e 17% na Oeste Gaúcho. Na verdade, mesmo considerando uma não avaliável margem de erro, derivada, tanto do método de estimação do custo da terra baseada no valor dos arrendamentos médios, quanto de possíveis desvios provenientes das formas de levantamentos estatísticos, o diferencial desse custo na Mesorregião Triângulo Mineiro em relação às demais Mesorregiões é por demais acentuado, de onde não se pode descartar a hipótese de supervalorização das terras devido a fenômenos não ligados à produção agrícola. A relação entre o valor dos arrendamentos por área arrendada e o valor da produção obtida de arroz por hectare colhido também reforça esta hipótese, uma vez que de 0,45 na Mesorregião Triângulo Mineiro, passa a 0,25 na de Goiânia, 0,24 na Leste Rio-Grandense e 0,17 na Oeste Gaúcho.

O custo do capital fixo, por sua vez, representado pelo custo de depreciação dos bens, é bastante significativo nas duas Mesorregiões gaúchas, 9,78% na Leste Rio-Grandense e 12,45% na Oeste Gaúcho. Nesta última Mesorregião, o peso da depreciação dos bens no custo total da produção foi superior em mais de 88% ao peso correspondente na Triângulo Mineiro e mais de 137% na de Goiânia. Considerando que o item de maior peso no custo do capital fixo é "máquinas e instrumentos de trabalho" (47 na Mesorregião Triângulo Mineiro, 55 na de Goiânia, 69 na Leste Rio-Grandense e 76% na Oeste Gaúcho), pode-se bem evidenciar o maior emprego de máquinas e instrumentos poupadores de trabalho nas Mesorregiões gaúchas.

O custo do capital circulante juntamente com o custo do trabalho é que condicionam na verdade, o resultado financeiro da exploração agrícola. Isto porque ambos os custos, devem ser cobertos dentro do período produtivo pelo valor da produção obtida no período. Ademais, é a utilização de insumos modernos e de trabalho que determina, em condições normais, os níveis de produtividade de uma atividade. Observe-se que a participação do custo do capital circulante no custo total da produção na Mesorregião Triângulo Mineiro é bastante baixa (25,61%), aumentando porém na de Goiânia (37,64%), na Leste Rio-Grandense (48,73%) e chegando aos 51,89% na Oeste Gaúcho. Numa desagregação maior do capital circulante (Tabela 5.1.6), os insumos modernos, entendidos como sendo adubos e corretivos, as sementes e mudas e os defensivos agrícolas, apresentaram participações relativas no total do custo do capital circulante bastante próximas nas Mesorregiões Triângulo Mineiro, Goiânia e Leste Rio-Grandense (39,31, 41,95 e 45,77%, respectivamente), enquanto que na Oeste Gaúcho esta participação era de apenas 35,03%. O baixo percentual nesta última Mesorregião apenas aparentemente é que pode levar à inferência de pouca utilização de insumos modernos em seus estabelecimentos, pois, antes de tudo, ele indica uma maior

participação dos outros componentes do capital circulante, especialmente das outras despesas, o que será abordado na seqüência do trabalho. Aliás, a utilização de insumos modernos pode melhor ser avaliada considerando-se a relação entre o valor das despesas com estes insumos e a área total colhida, como segue:

QUADRO 3.1.3

DESPESAS COM INSUMOS MODERNOS POR ÁREA COLHIDA

DESPESAS	VALOR (Cr$/ha)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Despesas com insumos modernos por área colhida.....	194,70	282,00	1 059,80	768,40

Como se observa, na Mesorregião Goiânia utilizaram-se cerca de 45% a mais de insumos modernos por área colhida do que na Mesorregião Triângulo Mineiro, enquanto que nas Leste Rio-Grandense e Oeste Gaúcho empregaram-se mais de cinco e cerca de quatro vezes mais, respectivamente, do que na região mineira. É claro que a hipótese de igualdade dos preços médios dos insumos nas quatro Mesorregiões, está subentendida nessa análise comparativa.

Quanto aos demais componentes do capital circulante, deve-se notar a alta participação das despesas com impostos e taxas no custo do capital circulante na Mesorregião Goiânia (12,15%) contra apenas 5,24% na Triângulo Mineiro, 3,38% na Leste Rio-Grandense e 3,26% na Oeste Gaúcho. Os gastos com juros e despesas bancárias, no entanto, têm uma participação semelhante nas quatro Mesorregiões, apesar de, se considerados em relação à área colhida, apresentarem diferenças substanciais entre as mesmas — Cr$ 59,80 por hectare na Mesorregião Triângulo Mineiro, Cr$ 8,42 por hectare na de Goiânia, Cr$ 27,16 por hectare na Leste Rio-Grandense e Cr$ 25,47 por hectare na Oeste Gaúcho —, podendo isto refletir um maior envolvimento dos produtores das duas regiões gaúchas no mercado financeiro da região. Finalmente, quanto às outras despesas, que compreendem também os gastos com combustíveis e lubrificantes, perfizeram cerca de 32,6% do total do capital circulante na Mesorregião Triângulo Mineiro, 25,6% na de Goiânia, 30,81% na Leste Rio-Grandense e 43,04% na Oeste Gaúcho. Nesta última Mesorregião, então, as despesas com combustíveis e lubrificantes

representaram cerca de 67% das outras despesas, inferindo a importância da mecanização agrícola nesta área.

Quanto ao custo do fator trabalho, deve-se considerar a sua participação no custo total da produção, por um lado, e, a sua própria composição, por outro lado. Assim, pelo quadro 3.1.2, observa-se que o custo do trabalho contribuiu com 37,83% do custo total na Mesorregião Triângulo Mineiro, 37,53% na de Goiânia, 21,19% na Leste Rio-Grandense e apenas 18,98% na Oeste Gaúcho. É importante notar que o custo do trabalho constitui-se no principal componente do custo de produção na Mesorregião mineira, assim como na goiana, sendo que, nesta última, é apenas ligeiramente inferior ao custo do capital circulante (78.144 milhares de cruzeiros como custo do trabalho e 78.375 milhares de cruzeiros como custo do capital circulante). Este fato vem novamente mostrar a predominância nas Mesorregiões gaúchas, do fator capital sobre o fator trabalho nas explorações orizícolas, com uma situação inversa nas mesorregiões centrais do País.

No referente à composição dos custos do fator trabalho, o quadro a seguir fornece informações também relevantes.

QUADRO 3.1.4

COMPOSIÇÃO PERCENTUAL DOS CUSTOS DO FATOR TRABALHO NA CULTURA DO ARROZ

DISPÊNDIOS	CUSTOS (%)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
TOTAL	**100,00**	**100,00**	**100,00**	**100,00**
Salários pagos	25,05	17,94	40,61	47,37
Pagamentos imputados relativos ao trabalho do produtor e membros da família	40,10	42,33	22,49	27,41
Quota-parte entregue a parceiros	12,11	14,64	12,62	8,80
Empreitadas apenas com mão-de-obra	5,88	8,39	14,12	7,06
Empreitadas com mão-de-obra, máquinas e equipamentos	16,86	16,70	10,16	9,36

Os dados deste quadro evidenciam as substanciais diferenças existentes entre as regiões centrais e sulinas do País. Repare-se que o valor imputado ao trabalho do produtor e dos membros de sua família atin-

giu os 40,10% do custo total do fator trabalho na Mesorregião Triângulo Mineiro e 42,33% na de Goiânia, contra apenas 22,49% na Leste Rio-Grandense e 27,41% na Oeste Gaúcho. Inversamente a esse valor, a participação dos salários pagos foi baixa nas duas primeiras Mesorregiões (25,05 e 17,94%) e relativamente elevada nas duas últimas (40,61 e 47,37%). Estes percentuais mostram a predominância do trabalho familiar na produção orizícola na Mesorregião Triângulo Mineiro e na Goiânia, enquanto que nas regiões gaúchas, a figura do assalariado predomina nas relações de trabalho.

A estrutura de produção, se por um lado fornece importantes informações relativas à combinação dos fatores e de certas características da produção, não permite que se formem inferências acerca da eficiência e dos custos reais da produção. Para tanto, procurou-se analisar a relação entre cada um dos custos estimados e do valor da produção e a área total colhida nas quatro regiões consideradas.

QUADRO 3.1.5

CUSTOS E VALOR DA PRODUÇÃO POR ÁREA COLHIDA NA CULTURA DO ARROZ

ESPECIFICAÇÃO	VALOR (Cr$/ha)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
CUSTO TOTAL	**1 933,20**	**1 786,10**	**4 751,60**	**4 227,30**
Custo da terra	579,00	349,80	964,90	705,20
Custo do capital fixo	127,60	93,60	464,60	526,10
Custo do capital circulante	495,20	672,30	2 315,40	2 193,50
Custo do trabalho	731,40	670,40	1 006,70	802,50
Valor da Produção	**1 416,50**	**1 700,00**	**6 805,70**	**6 483,40**

A primeira observação a ser feita sobre os dados apresentados refere-se ao diferencial entre o valor da produção obtido e o custo total por área colhida, que, de negativo nas Mesorregiões Triângulo Mineiro (Cr$ — 516,70) e Goiânia (Cr$ — 86,10), passa a positivo nas Mesorregiões Leste Rio-Grandense (Cr$ 2.054,10) e Oeste Gaúcho (Cr$ 2.256,10).

Estes diferenciais, que podem ser chamados de prejuízos líquidos, quando negativo, ou lucros líquidos, quando positivos, são função dos preços do produto obtido e dos fatores de produção, das quantidades empregadas desses fatores e das combinações efetuadas com eles, que, em última análise, representam o nível de tecnologia empregado na produção e, por último e não menos importante, das condições meteorológicas, no sentido lato, vigente no período produtivo da atividade.

Sob as hipóteses de igualdade nos preços dos fatores componentes do capital, fixo e circulante, de condições climáticas normais e homogeneidade dos fatores, verifica-se que, justamente a intensidade de aplicação do capital é que pode explicar os relativamente altos lucros obtidos nas duas Mesorregiões gaúchas frente aos prejuízos líquidos nas duas outras Mesorregiões. Isto fica claro ao se considerar que o fator trabalho é muito mais utilizado na Mesorregião Triângulo Mineiro, 4,88 hectares de área colhida por pessoa empregada, e na de Goiânia, 4,32 hectares por pessoa empregada, contra 5,85 hectares de área colhida por pessoa empregada na Leste Rio-Grandense e 7,27 hectares na Oeste Gaúcho. Maiores dispêndios com os fatores de capital devem ter permitido, portanto, a obtenção de maiores produtividades físicas por área colhida, maior valor de produção e melhores remunerações aos fatores envolvidos na atividade, como se pode observar pelas tabelas em anexo.

3.2 — Batata-inglesa

Para a análise comparativa dos custos de produção da batata-inglesa foram selecionadas as Mesorregiões Sudoeste Mineiro, em Minas Gerais, Curitiba e Leste Paranaense, no Estado do Paraná. Estas três Mesorregiões juntas, representavam em 1975, cerca de 42,67% da produção de batata, 35,58% da área colhida e 8,21% do número de estabelecimentos produtores no País como um todo.

Em relação às UFs, a Mesorregião Sudoeste Mineiro representava a quase totalidade da produção mineira (94%), cerca de 90% da área colhida e 69% do número de estabelecimentos produtores, enquanto que as duas Mesorregiões paranaenses respondiam por 94% da produção, 95% da área colhida e 66% do número de estabelecimentos produtores no Estado do Paraná.

Os estabelecimentos especializados na produção de batata-inglesa, por sua vez, representavam parcelas significativas da produção e do número de estabelecimentos, determinando um alto índice de especialização em termos de valor da produção, nas três Mesorregiões consideradas.

QUADRO 3.2.1

ÍNDICES DE ESPECIALIZAÇÃO NA CULTURA DA BATATA-INGLESA

ESPECIFICAÇÃO	ÍNDICES (%)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Índice de especialização da produção.....	66,62	50,79	20,99
Índice de especialização dos estabelecimentos...	18,49	12,65	2,66
Índice "verdadeiro" de especialização.....	13,32	6,42	0,56

Mais acentuadamente que no caso analisado do arroz, as relações entre as produtividades físicas da produção de batata-inglesa nos estabelecimentos especializados e nos estabelecimentos produtores de cada uma das Mesorregiões, inferem uma tecnologia de produção mais moderna nos primeiros estabelecimentos, cujos índices de produtividade foram 3,6 vezes maiores que a média na Mesorregião Sudoeste Mineiro, 4 vezes na de Curitiba e cerca de 7,9 vezes na Leste Paranaense.

Aceitando-se como significativos os números encontrados de estabelecimentos especializados, 838 na Mesorregião mineira, 532 na de Curitiba e 181 na Leste Paranaense, seria a seguinte a estrutura de custos da produção de batata-inglesa:

QUADRO 3.2.2

ESTRUTURA DE CUSTOS DE PRODUÇÃO NOS ESTABELECIMENTOS ESPECIALIZADOS NA CULTURA DA BATATA-INGLESA

ESPECIFICAÇÃO	CUSTOS DA PRODUÇÃO (%)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
CUSTO TOTAL.....	**100,00**	**100,00**	**100,00**
Custo da terra.....	7,74	10,77	8,93
Custo do capital fixo.....	4,63	12,16	8,22
Custo do capital circulante.....	65,20	55,38	68,25
Custo do trabalho.....	22,43	21,69	14,60

À primeira vista, a estrutura de custos de produção das três Mesorregiões indica uma maior concentração dos custos nos dispêndios com o capital e com o trabalho. Assim, o custo da terra, que no caso da produção de arroz, representava de 16 a 30% do custo de produção, ficou entre 7,74% na Mesorregião Sudoeste Mineiro, 8,93% na Leste Paranaense e 10,77% na de Curitiba, se bem que o valor médio dos arrendamentos, base para a estimativa do custo da terra, tenha sido 20% superior nos estabelecimentos produtores de arroz do que os pagos pelos produtores de batata-inglesa (Cr$ 790/ha no primeiro caso, contra Cr$ 660/ha no segundo). Nas três Mesorregiões produtoras deste último produto, no entanto, os valores médios dos arrendamentos não apresentaram grandes variações entre si, não servindo, portanto, como justificativa para as diferenças das estruturas de custos apresentadas entre as Mesorregiões em estudo.

O custo do capital fixo, por outro lado, merece algumas considerações, já que com um peso de apenas 4,63% no custo de produção na

Mesorregião Sudoeste Mineiro, passa a 12,16% na de Curitiba e 8,22% na Leste Paranaense. Os dados da tabela 5.2.5 — custo do capital fixo, evidencia a importância dos custos de depreciação das "máquinas e instrumentos de trabalho" e "veículos e outros meios de transporte" no custo total de depreciação. A relação entre estes dois custos e a área total colhida infere uma mais baixa utilização desses capitais na Mesorregião Sudoeste Mineiro (Cr$ 333/ha) em comparação à de Curitiba (Cr$ 504/ha) e à Leste Paranaense (Cr$ 488/ha). O diferencial desse custo por área colhida entre as Mesorregiões, no entanto, não é suficiente para explicar a baixa participação do custo do capital fixo no custo total de produção na Sudoeste Mineiro, participação esta mais devida ao alto volume das despesas com o capital circulante nessa Mesorregião. Estas despesas perfizeram 65,7% do custo de produção nessa Mesorregião, 55,39% na de Curitiba e 68,25% na Leste Paranaense. Numa análise mais desagregada do capital circulante nestas três regiões, mostra mais semelhanças do que diferenças entre elas na composição do capital circulante. Assim, as despesas com insumos modernos (adubos e corretivos, sementes e mudas e defensivos agrícolas) respondiam com cerca de 70% do custo do capital circulante na Mesorregião Sudoeste Mineiro, 67% na de Curitiba e 66% na Leste Paranaense. Se também considerados a componente "outras despesas", estas participações passam a 83,76, 93,94 e 87,80%, respectivamente.

Apesar da similaridade da composição dos custos do capital circulante entre as regiões analisadas, observa-se que a intensidade de sua aplicação por área colhida variou bastante, sendo que na Mesorregião Sudoeste Mineiro o custo dos insumos modernos por área colhida foi cerca de 90% superior ao da de Curitiba e 20% superior ao da Leste Paranaense. Em compensação, estas duas últimas Mesorregiões con-

QUADRO 3.2.3

DESPESAS COM INSUMOS MODERNOS, COMBUSTÍVEIS E CAPITAL CIRCULANTE, POR ÁREA COLHIDA DA CULTURA DA BATATA-INGLESA

DISPÊNDIOS	VALOR (Cr$/ha)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Despesas com insumos modernos.....	4 003,10	2 109,10	3 328,50
Despesas com combustíveis.....	410,70	448,20	481,70
Custo do capital circulante.....	5 702,60	3 125,10	5 020,50

sumiram mais combustíveis por área colhida, podendo indicar uma maior utilização de maquinarias agrícolas. A respeito, considerando-se a relação entre as despesas com combustíveis e o número de tratores existentes, que, em milhares de cruzeiros, foi de 15/trator na Mesorregião Sudoeste Mineiro, 12/trator na de Curitiba e 15/trator na Leste Paranaense, pode-se afirmar ter havido uma maior intensidade de utilização dessas máquinas na primeira e na última Mesorregião, aparecendo na de Curitiba com um menor índice de uso.

É de se esperar que, a disponibilidade e a utilização de máquinas nas atividades agrícolas, devam se refletir nas quantidades insumidas dos demais fatores de produção, especialmente no trabalho. O peso deste fator no custo estimado da produção foi de 22,43% na Mesorregião Sudoeste Mineiro, 21,69% na de Curitiba e 14,6% na Leste Paranaense.

A análise de seus componentes, no entanto, é que mostra as substanciais diferenças na exploração do produto nas três regiões.

QUADRO 3.2.4

COMPOSIÇÃO PERCENTUAL DOS CUSTOS DO FATOR TRABALHO NA CULTURA DA BATATA-INGLESA

DISPÊNDIOS	CUSTOS (%)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
TOTAL	**100,00**	**100,00**	**100,00**
Salários pagos	54,77	37,99	46,44
Pagamentos imputados relativos ao trabalho do produtor e membros da família	28,29	47,45	37,54
Quota-parte entregue a parceiros	10,15	0,85	0,04
Empreitadas apenas com mão-de-obra	3,41	8,26	9,44
Empreitadas com mão-de-obra, máquinas e equipamentos	3,39	5,45	6,54

Uma explicação para a baixa participação do valor imputado ao trabalho familiar nos custos do trabalho na Mesorregião Sudoeste Mineiro (28,29%) em contraposição às relativamente altas participações na de Curitiba (47,45%) e na Leste Paranaense (37,54%) é complexa, frente aos dados disponíveis, mas, provavelmente, a condição do produtor em relação à posse da terra pode, em parte, justificar tais diferenças. Assim, enquanto que na Mesorregião mineira o número de produtores proprietários correspondia a apenas 397 estabelecimentos

num total de 838 produtores (47,4%), na de Curitiba eles representavam cerca de 86,6% dos produtores (532 estabelecimentos) e 78,4% na Leste Paranaense (num total de 181 produtores). Evidentemente, o custo relativo à quota-parte entregue a parceiros, de 10,15% na Sudoeste Mineiro e desprezível nas demais Mesorregiões, é essencialmente função das relações de posse da terra vigentes.

As relações entre os custos de produção e a área colhida do produto em questão, permitem melhores inferências acerca dos aspectos relevantes já indicados pela estrutura de custos.

QUADRO 3.2.5

CUSTOS E VALOR DA PRODUÇÃO POR ÁREA COLHIDA NA CULTURA DA BATATA-INGLESA

ESPECIFICAÇÃO	VALOR (Cr$/ha)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
CUSTO TOTAL	**8 746,50**	**5 642,40**	**7 356,40**
Custo da terra	676,80	607,80	656,70
Custo do capital fixo	405,10	685,90	604,90
Custo do capital circulante	5 702,60	3 125,10	5 020,50
Custo do trabalho	1 962,00	1 223,60	1 074,30
Valor da produção	**12 727,90**	**7 657,00**	**9 690,80**

Partindo-se da diferença entre o valor da produção e o custo total, por hectare de área colhida, de Cr$ 3.981,40 na Mesorregião Sudoeste Mineiro, Cr$ — 2.014,60 na de Curitiba e Cr$ — 2.334,40 na Leste Paranaense, pode-se aparentemente inferir que a estrutura de custos de produção praticada na Mesorregião mineira foi mais eficiente do que a praticada nas duas outras Mesorregiões, tudo o mais permanecendo constante. Há, no entanto, de se considerar que os preços recebidos pelos produtores foram significantemente superiores na Mesorregião mineira, Cr$ 1.222,10 por tonelada, contra Cr$ 934,70 por tonelada na Leste Paranaense e Cr$787 por tonelada na de Curitiba. As razões para estes diferenciais podem ser várias, desde diferenças nas variedades do produto até especificidades do calendário agrícola nas regiões produtoras. No caso da Mesorregião Sudoeste Mineiro, há de se supor que a obtenção de maiores preços para o produto também está relacionada à excelente posição geográfica da região, entre os dois maiores centros consumidores do

País, — as Cidades de São Paulo e Rio de Janeiro —, o que deve reduzir custos de transporte e de comercialização do produto.

Numa simulação, supondo-se o custo da terra, o custo do trabalho e o preço do produto constantes e iguais aos ocorridos na Mesorregião Sudoeste Mineiro, o lucro líquido obtido na Mesorregião de Curitiba passaria a Cr$ 5.244,60 por hectare colhido e a Cr$ 4.398,60 na Mesorregião Leste Paranaense, bastante superiores ao obtido na região mineira. É claro que nada se pode afirmar, segundo esta simulação, acerca da excelência ou não da alocação de fatores nas três Mesorregiões, desde que as estruturas de custos da produção tenderiam a se ajustar às variações nos preços relativos dos fatores.

3.3 — Cana-de-açúcar

O critério para a seleção das mesorregiões produtoras de cana-de-açúcar a serem analisadas neste trabalho, foi diferente do adotado para a escolha das mesorregiões produtoras dos demais produtos, baseado na produção excedentária de cada uma delas. Para a cana-de-açúcar foram

identificadas as quatro UFs maiores produtoras, selecionando-se, em cada uma delas, a mesorregião com maior produção. O objetivo desse critério foi o de, ao mesmo tempo, abranger o máximo possível da produção nacional e distinguir as três grandes áreas produtoras mais diferenciadas em termos de disponibilidade e uso dos fatores de produção — a zona da mata nordestina, o norte fluminense e o nordeste paulista.

As UFs selecionadas foram: Pernambuco, Alagoas, Rio de Janeiro e São Paulo, que juntas, representavam cerca de 16% do número total de produtores de cana-de-açúcar, 84% da quantidade obtida e 79% da área colhida no Brasil, em 1975. As Mesorregiões de maior produção nestes Estados foram a Mata Pernambucana, com 84% da produção estadual, Mata Alagoana, com 85% da produção de Alagoas, Norte Fluminense, com 97% da produção do Rio de Janeiro, Campinas e Ribeirão Preto, com 77% da produção paulista. A representatividade dessas Mesorregiões na produção brasileira foi bastante alta, mais de 68% da quantidade obtida no ano em questão.

Igualmente alta foi a representatividade dos estabelecimentos especializados na produção de cana-de-açúcar, que, somados nas quatro Mesorregiões, perfizeram mais de 62% da produção nacional e cerca de 58% da área colhida no País. Evidentemente o índice de especialização no cultivo do produto foi bastante elevado, como se pode observar no quadro 3.3.1.

Considerando-se os dois índices de especialização apresentados, observa-se que na Mata Pernambucana 95,81% da produção estavam concentrados nos estabelecimentos predominantemente dedicados ao cultivo da cana-de-açúcar que representavam, no entanto, apenas

QUADRO 3.3.1

ÍNDICES DE ESPECIALIZAÇÃO NA CULTURA DA CANA-DE-AÇÚCAR

ESPECIFICAÇÃO	ÍNDICES (%)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Índice de especialização da produção...	95,81	98,32	90,32	87,03
Índice de especialização dos estabelecimentos.....	56,40	83,41	68,44	56,26
Índice "verdadeiro" de especialização	54,04	82,01	61,82	48,96

56,4% do número de estabelecimentos produtores. Foi a Mata Alagoana, porém, que apresentou os mais altos índices, significando que a maioria dos produtores (83,41%) se dedicava quase que exclusivamente à produção da cana-de-açúcar.

Tomando-se a relação entre os dois índices de especialização como um indicador da coexistência de diferentes tecnologias e escalas de produção, já que representa a diferencial (em percentagem) entre o valor da produção, por estabelecimento, obtido nos estabelecimentos especializados em questão, pode-se levantar algumas questões importantes relativas à produção de cana-de-açúcar. Assim, na Mata Pernambucana, o valor da produção obtido nos estabelecimentos especializados foi, aproximadamente, 70% superior à média da Mesorregião, considerando que o número de estabelecimentos especializados perfazia cerca de 56,4% do total de produtores, indica um valor da produção por número de estabelecimentos não especializados ao redor de apenas 10% da média da Mesorregião [8].

Na Mata Alagoana, a razão entre os dois índices de especialização foi relativamente baixa, com o valor médio da produção nos estabelecimentos especializados superior em cerca de 18% ao valor médio da região, passando a 32% na Mesorregião Norte Fluminense e 55% na de Campinas e Ribeirão Preto.

Dados os índices de especialização dos estabelecimentos em cada uma dessas Mesorregiões, pode-se avaliar a relação entre o valor médio da produção nos estabelecimentos não especializados e o valor médio conseguido nessas Mesorregiões, que foi de 10% na Mata Alagoana, 31% na Norte Fluminense e 30% na de Campinas e Ribeirão Preto.

As razões para tão grandes diferenças entre estabelecimentos especializados e não especializados talvez pudessem ser inferidas em uma análise mais desagregada, por classes de área, o que, no entanto, foge às pretensões deste trabalho.

A composição dos custos de produção da cana-de-açúcar nas Mesorregiões em estudo evidencia as diferenças e semelhanças em suas estruturas produtivas.

De início, deve-se reparar na semelhança da composição dos custos nas duas regiões da Mata Nordestina e a grande diferença entre estas e a Norte Fluminense e a de Campinas e Ribeirão Preto. Assim, o custo da terra que representava 4,65 e 5,47% do custo total de produção nas regiões nordestinas, participava com 17,27% deste na Norte Flu-

[8] A razão entre o valor médio da produção nos estabelecimentos não especializados e o valor médio da produção nos estabelecimentos produtores de cada mesorregião pode ser encontrada segundo a fórmula $x = \frac{1 - a}{1 - b}$, onde a é o índice de especialização da produção e b o índice de especialização dos estabelecimentos.

QUADRO 3.3.2

ESTRUTURA DE CUSTOS DA PRODUÇÃO NOS ESTABELECIMENTOS ESPECIALIZADOS NA CULTURA DA CANA-DE-AÇÚCAR

ESPECIFICAÇÃO	CUSTOS DA PRODUÇÃO (%)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
CUSTO TOTAL	**100,00**	**100,00**	**100,00**	**100,00**
Custo da terra	4,65	5,47	17,27	14,31
Custo do capital fixo	4,23	6,02	8,33	6,51
Custo do capital circulante	41,07	42,60	35,85	45,51
Custo do trabalho	50,05	45,91	38,55	33,67

minense e 14,31% na Mesorregião paulista. O alto preço da terra nestas duas últimas regiões, certamente foi o responsável por tais diferenças.

Quanto ao custo do capital fixo, observa-se que o seu peso na composição do custo total na Mesorregião Norte Fluminense foi relativamente alto, quando comparado aos das demais Mesorregiões. Este fato, na verdade, era esperado nas regiões nordestinas, onde, reconhecidamente, é problemática a mecanização agrícola em grande parte de suas áreas produtoras, bastante acidentadas, mas não em relação à Mesorregião paulista. Considerando apenas o custo da depreciação dos veículos e outros meios de transporte e das máquinas e instrumentos de trabalho por área colhida, como indicador da utilização de capital fixo, a Mesorregião Norte Fluminense ainda assim surpreende, com uma relação de Cr$ 181 por hectare, contra Cr$ 161 por hectare na de Campinas e Ribeirão Preto, Cr$ 155 por hectare na Mata Alagoana e Cr$ 87 por hectare na Mata Pernambucana. Nestas duas últimas regiões, ao contrário das do sudoeste brasileiro, o peso do custo de depreciação dos veículos e outros meios de transporte nessas relações é bem maior do que o peso correspondente das máquinas e instrumentos de trabalho, como era de se esperar, ou seja, 60% em Pernambuco, 53% em Alagoas, 39% no Rio de Janeiro e 37% em São Paulo.

A análise do custo do capital fixo evidentemente, deve ser complementada pelos estudos da composição e do volume do capital circulante empregado. Os dados do quadro 3.3.2, mostram que a participação relativa dos custos do capital circulante no custo total de produção não apresentou grandes diferenças nas Mesorregiões nordestinas (41,07 na Mata Pernambucana e 42,6% na Mata Alagoana) e paulista

(45,5% na de Campinas e Ribeirão Preto), diferindo destas três, significantemente, a participação relativa estimada para a Mesorregião Norte Fluminense (35,85%).

As semelhanças na composição dos custos das três primeiras Mesorregiões ainda continuam a um nível mais desagregado do capital circulante, especialmente na utilização de insumos modernos e nos dispêndios com juros e despesas bancárias. Aliás, são estes dois componentes do capital circulante, mais os impostos e taxas e as outras despesas, que determinam as principais diferenças na composição do capital circulante entre as Mesorregiões.

QUADRO 3.3.3

COMPOSIÇÃO DO CAPITAL CIRCULANTE NA CULTURA DA CANA-DE-AÇÚCAR

ITENS DA DESPESA	PARTICIPAÇÃO DAS DESPESAS (%)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
TOTAL	**100,00**	**100,00**	**100,00**	**100,00**
Insumos modernos	43,25	45,58	30,19	45,44
Medicamentos e alimentos para animais	2,43	1,40	1,04	1,00
Transporte da produção	10,51	9,34	10,14	14,06
Impostos e taxas	17,09	7,37	2,55	3,96
Juros e despesas bancárias	7,28	8,15	14,08	6,66
Aluguel de máquinas e equipamentos	0,57	0,15	4,60	2,38
Outras despesas	18,87	28,01	37,40	26,50

A elevada participação das despesas com insumos modernos nos custos do capital circulante, nas Mesorregiões nordestinas e na paulista, frente à baixa participação encontrada na Norte Fluminense, não necessariamente deve indicar uma maior utilização desses insumos nas três primeiras regiões, mas, no caso do cultivo da cana-de-açúcar isto deve ter ocorrido, já que as despesas com insumos modernos por área colhida foi bem inferior na região fluminense: Cr$ 297,50 por hectare, contra Cr$ 533,40 por hectare na Mata Pernambucana, Cr$ 605,30 por hectare na Mata Alagoana e Cr$ 745,80 por hectare na de Campinas e Ribeirão Preto.

Quanto aos juros e despesas bancárias, o seu peso foi muito alto, comparativamente às outras regiões, podendo indicar uma maior dependência desses estabelecimentos a recursos não próprios. Na realida-

de, se considerarmos a relação entre os gastos com juros e despesas bancárias e as despesas com insumos modernos por área colhida, pode-se verificar que enquanto essa relação não passava de 17% na Mata Pernambucana, 18% na Mata Alagoana e 15% na de Campinas e Ribeirão Preto, ela se aproximava dos 47% na Mesorregião Norte Fluminense. Isto significa que, para cada cruzeiro despendido em insumos modernos por área colhida nesta região, gastava-se Cr$ 0,47 em pagamentos a bancos e instituições financeiras e mesmo a particulares, por empréstimos solicitados.

Em princípio, nada se pode afirmar acerca de perniciosidade ou não dessas altas despesas, o que não ocorre com os dispêndios com impostos e taxas, que representavam 17,09% do custo do capital circulante na Mata Pernambucana e 7,37% na Mata Alagoana. Se por um lado estes impostos e taxas significavam uma maior contribuição da atividade para as receitas do governo, representando, em tese, uma redistribuição intersetorial da renda nos dois estados nordestinos, por outro lado significavam também um ônus adicional aos estabelecimentos produtores. Numa comparação entre as Mesorregiões em estudo, essas despesas por área colhida foram de Cr$ 210,60 por hectare na Mata Pernambucana, Cr$ 9,79 por hectare na Mata Alagoana, Cr$ 6,58 por hectare na de Campinas e Ribeirão Preto e apenas Cr$ 2,51 por hectare na Norte Fluminense.

No componente "outras despesas" do custo do capital circulante, o importante é verificar-se o comportamento dos custos com combustíveis e lubrificantes como um indicador de mecanização da atividade e do grau de utilização de maquinarias agrícolas.

QUADRO 3.3.4

DESPESAS COM COMBUSTÍVEIS POR ÁREA E VALOR DAS MÁQUINAS, EQUIPAMENTOS E VEÍCULOS NA CULTURA DA CANA-DE-AÇÚCAR

DESPESAS	VALOR			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Despesas com combustíveis por área colhida (Cr$ 1 000/ha).....	0,1297	0,1971	0,2126	0,1908
Despesas com combustíveis por valor das máquinas, equipamentos e veículos (em Cr$ 10 000).....	0,1500	0,1271	0,1177	0,1187

Como indicador de mecanização, observe-se que na Mesorregião Norte Fluminense foram gastos mais combustíveis por área colhida do que nas demais Mesorregiões, apresentando a Mata Pernambucana a menor despesa por área. Por outro lado, foi esta última região que apresentou o maior grau de utilização de maquinarias, conceituado como sendo a relação entre as despesas com combustíveis e o valor das máquinas, equipamentos agrícolas e veículos. As imperfeições desse indicador são óbvias e não parece necessário discutí-las, apesar de sua, teoricamente, inversa relação com o volume de trabalho utilizado na atividade. Este ponto será novamente abordado após a discussão da composição do trabalho nas mesorregiões.

QUADRO 3.3.5

COMPOSIÇÃO PERCENTUAL DOS CUSTOS DO FATOR TRABALHO NA CULTURA DA CANA-DE-AÇÚCAR

DISPÊNDIOS	CUSTOS (%)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
TOTAL	**100,00**	**100,00**	**100,00**	**100,00**
Salários pagos	73,12	71,81	45,06	54,08
Pagamentos imputados relativos ao trabalho do produtor e membros da família	12,03	10,13	31,81	12,80
Quota-parte entregue a parceiros	0,02	0,08	0,46	0,69
Empreitadas apenas com mão-de-obra	10,77	15,25	16,72	21,15
Empreitadas com mão-de-obra, máquinas e equipamentos	4,06	2,73	5,95	11,28

Além da inexpressividade do peso da quota-parte entregue a parceiros, deve-se notar a marcante diferença entre a participação da mão-de-obra familiar no custo estimado do trabalho nas Mesorregiões nordestinas e na paulista (12,03, 10,13 e 12,8%, respectivamente), da encontrada para a Norte Fluminense (31,81%). Isto pode indicar desde uma estrutura mais empresarial da produção nas três primeiras, com maior recorrência ao mercado de trabalho regional, até uma estrutura de produção mais eficiente com melhores combinações de fatores na Norte Fluminense.

As relações entre cada um dos componentes do custo de produção pode esclarecer melhor este ponto.

QUADRO 3.3.6

CUSTOS E VALOR DA PRODUÇÃO POR ÁREA COLHIDA NA CULTURA DA CANA-DE-AÇÚCAR

ESPECIFICAÇÃO	VALOR (Cr$/ha)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
CUSTO TOTAL	**3 002,10**	**3 117,00**	**2 748,60**	**3 654,70**
Custo da terra	139,60	170,60	474,80	523,10
Custo do capital fixo	127,00	187,70	228,90	238,00
Custo do capital circulante	1 232,90	1 328,00	985,50	1 662,90
Custo do trabalho	1 502,60	1 430,70	1 059,40	1 230,70
Valor da produção	**3 169,20**	**3 404,90**	**2 771,40**	**3 455,70**

O quadro 3.3.6 é muito elucidativo, permitindo tirar algumas conclusões, acerca da estrutura de custos comparativos de produção nas áreas em estudo. A primeira constatação a ser feita é que, se a produção é diretamente relacionada com os volumes de fatores empregados, o lucro líquido da exploração deveria crescer da Mesorregião Norte Fluminense para a Mata Pernambucana, para a Mata Alagoana e para a de Campinas e Ribeirão Preto, sob a hipótese de preços uniformes. Tal não aconteceu, tendo o lucro líquido por área colhida decrescido no sentido da Mata Alagoana (Cr$ 287,90/ha), para a Mata Pernambucana (Cr$ 167/ha), a Norte Fluminense (Cr$ 22,80/ha) e a de Campinas e Ribeirão Preto (Cr$ — 199/ha).

É claro que a hipótese de igualdade nos preços da cana-de-açúcar entre as regiões não é válida, sabendo-se que são fixados pelo Governo segundo critérios de política global de desenvolvimento tendentes a beneficiar mais as áreas produtoras nordestinas. Sem qualquer juízo crítico dessa política, pode-se verificar que as diferenças de preços entre as regiões foram bastante significativas, sendo, realmente, as responsáveis pelas diferenças relativas nos lucros líquidos das Mesorregiões em estudo.

QUADRO 3.3.7

PRODUTIVIDADE FÍSICA E PREÇOS MÉDIOS OBTIDOS NA CULTURA DA CANA-DE-AÇÚCAR

VARIÁVEIS	MESORREGIÕES			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Produtividade física (t/ha)[1]	43,8999	41,8516	41,9426	52,0714
Preços médios do produto (Cr$/t).....	97	97	79	79

1 Produtividade física por área efetivamente colhida do produto.

Numa simulação simples, supondo-se que os preços recebidos pelos produtores nordestinos vigorassem também no centro-sul do País, o lucro líquido por área colhida obtido na Mesorregião Norte Fluminense seria de Cr$ 616 por hectare e de Cr$ 562 por hectare na de Campinas e Ribeirão Preto.

3.4 — Cebola

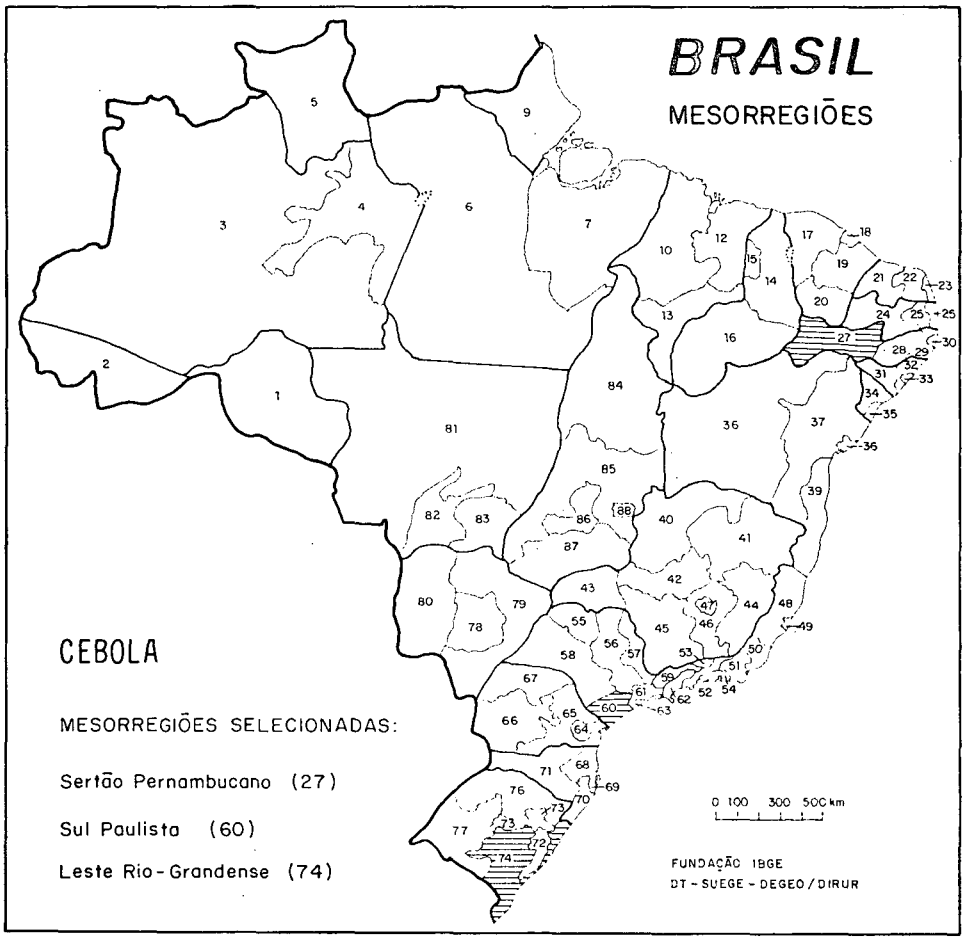
Foram selecionadas três Mesorregiões para a análise comparativa dos custos da produção de cebola: a Leste Rio-Grandense, a Sertão Pernambucano e a Sul Paulista, respectivamente situadas em três grandes Estados produtores: Rio Grande do Sul, Pernambuco e São Paulo. Estas três Mesorregiões juntas, apesar de representarem, em 1975, apenas 19,83% do número de produtores, produziram cerca de 56,73% da quantidade colhida e 56,62% do valor da produção obtida com a cebola, no País como um todo. Em relação às UFs, a Leste Rio-Grandense representava 25,47% do número de produtores e 92,47% da produção gaúcha, enquanto que a Sertão Pernabucano concentrava o número de produtores (86,16%) e o valor da produção (99,29%) de Pernambuco. A Sul Paulista, por sua vez e contrariamente às demais Mesorregiões, representava uma maior participação no número de produtores (51,54%) do que no valor da produção paulista (54,02%).

Observe-se que as três Mesorregiões selecionadas têm o calendário agrícola a diferenciá-las significantemente, o que por certo deve afetar os resultados econômicos do cultivo da cebola. Assim, e segundo os dados do Censo Agropecuário de 1975, os principais meses de colheita do produto na Sertão Pernambucano (junho, julho, agosto e setembro), antecediam os de colheita paulista (novembro e dezembro) e gaúcha (dezembro e janeiro). É, pois, claro que o volume da colheita em cada

uma dessas áreas deve condicionar o comportamento dos preços do produto ao longo do ano muitas vezes com reflexos negativos na atividade em uma, duas ou até nas três Mesorregiões. Este fato deve, portanto, ser sempre considerado na análise de custos comparativos, principalmente devido à relativamente alta perecibilidade do produto, que impede a sua estocagem por períodos de tempo suficientes para o melhor aproveitamento de condições favoráveis do mercado.

O nível de especialização encontrado nas três mesorregiões em análise foi bastante alto, com índices superiores nas duas áreas do centro-sul do País.

Pelos índices apresentados, pode-se inferir uma distribuição mais homogênea da produção na Mesorregião Sul Paulista e uma maior concentração na Leste Rio-Grandense e na Sertão Pernambucano. As estruturas de custos nestas Mesorregiões, no entanto, é que evidenciam as profundas diferenças na produção da cebola no País, mostrando, certamente, a validade das políticas regionalizadas de incentivo à produção.



QUADRO 3.4.1

ÍNDICES DE ESPECIFICAÇÃO NA CULTURA DA CEBOLA

ESPECIFICAÇÃO	ÍNDICES (%)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Índice de especialização da produção.....	62,20	47,33	57,19
Índice de especialização dos estabelecimentos...	25,77	21,07	30,77
Índice "verdadeiro" de especialização.....	16,03	9,97	17,60

QUADRO 3.4.2

ESTRUTURA DE CUSTOS DA PRODUÇÃO NOS ESTABELECIMENTOS ESPECIALIZADOS NA CULTURA DA CEBOLA

ESPECIFICAÇÃO	CUSTOS DA PRODUÇÃO (%)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
CUSTO TOTAL.....	**100,00**	**100,00**	**100,00**
Custo da terra.....	8,12	16,54	43,13
Custo do capital fixo.....	4,76	2,50	5,93
Custo do capital circulante.....	24,36	17,42	29,22
Custo do trabalho.....	62,76	63,54	21,72

De início, o peso do custo da terra no custo total da produção na Mesorregião paulista, surpreende pela sua disparidade em relação às demais Mesorregiões, tendo sido mais de cinco vezes superior ao da Leste Rio-Grandense e quase três vezes ao da Sertão Pernambucano. Na verdade, o valor médio dos arrendamentos de terras, variável básica para a estimativa do custo da terra, foi de aproximadamente Cr$ 9.412 por hectare na Sul Paulista, contra apenas Cr$ 1.168 e Cr$ 2.275 por hectare nas Mesorregiões gaúcha e pernambucana, respectivamente.

A participação do custo do capital fixo no custo total da produção se confrontada com a participação do custo do trabalho leva à inferência dos diferentes níveis de utilização de maquinarias nas regiões.

Na Mesorregião Sul Paulista, como era de se esperar, o custo do capital fixo por hectare colhido (Cr$ 1.295,10/ha) foi bem superior ao despendido na Leste Rio-Grandense (Cr$ 684,80/ha) e na Sertão Pernambucano (Cr$ 344/ha), com participação do custo com máquinas e instrumentos de trabalho de 65,6, 18,7 e 33,4%, respectivamente. O maior nível tecnológico nas explorações paulistas fica evidenciado ao se considerar, desagregadamente, o custo do capital circulante por área colhida.

QUADRO 3.4.3

CUSTO DO CAPITAL CIRCULANTE POR ÁREA COLHIDA NA CULTURA DA CEBOLA

ITENS DAS DESPESAS	VALOR (Cr$/ha)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Insumos modernos	1 807,40	1 263,70	3 611,80
Medicamentos e alimentos para animais	143,10	61,90	21,30
Transporte da produção	192,20	71,40	251,50
Impostos e taxas	323,40	17,30	50,20
Juros e despesas bancárias	81,20	104,80	238,70
Aluguel de máquinas e equipamentos	3,90	9 963,10	163,50
Outras despesas	950,70	815,50	2 040,50

Além da substancial diferença nas despesas com insumos modernos entre as Mesorregiões, os dados mostram as significativas despesas com impostos e taxas, efetuadas na Leste Rio-Grandense, como uma característica toda especial desta Mesorregião, desde que são despesas institucionalizadas, passíveis de serem alteradas, segundo os objetivos das políticas econômicas, em nível regional ou federal.

Como último, e não menos importante componente dos custos de produção de cebola, o custo do trabalho empregado na exploração foi bastante expressivo na região gaúcha, justamente pela intensidade de uso do trabalho nessa região. A alta participação desse fator no custo da produção na Mesorregião Sertão Pernambucano, por outro lado, se deveu mais à metodologia utilizada na estimação do custo da mão-de-obra familiar, já que o salário médio utilizado para tal, foi significantemente superior ao das demais regiões em análise. Tomando-se como indicador de intensidade de uso da mão-de-obra a relação entre a área colhida e o número total de trabalhadores em cada Mesorregião,

a Sertão Pernambucano aparece mesmo como a menos intensiva, com 0,69 hectares por pessoa empregada, contra 0,50 na Sul Paulista e apenas 0,26 hectares na Leste Rio-Grandense.

Se na composição dos custos de produção, a utilização de máquinas e equipamentos e as despesas com insumos modernos indicam uma exploração mais tecnificada na Mesorregião Sul Paulista, a comparação entre os custos e o valor da produção de cebola por hectare colhido mostra a obtenção de ganhos no cultivo do produto apenas na Leste Rio-Grandense.

QUADRO 3.4.4

CUSTOS E VALOR DA PRODUÇÃO POR ÁREA COLHIDA NA CULTURA DA CEBOLA

ESPECIFICAÇÃO	VALOR (Cr$/ha)		
	Messoregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
CUSTO TOTAL	**14 374,60**	**13 761,20**	**21 823,80**
Custo da terra	1 167,70	2 275,60	9 411,90
Custo do capital fixo	684,80	344,00	1 295,10
Custo do capital circulante	3 502,00	2 397,60	6 377,50
Custo do trabalho	9 020,10	8 744,00	4 739,30
Valor da produção	**16 305,90**	**10 035,70**	**18 565,10**

Como se observa, nesta Mesorregião, o valor da produção de cebola foi 13,44% superior ao custo total por hectare colhido, enquanto que na Sertão Pernambucana e na Sul Paulista foram inferiores em 27,07 e 14,93%, respectivamente. Estes resultados, aparentemente estranhos, têm explicações simples, no caso da Sul Paulista, em que os preços médios obtidos pelo produtor foram muito inferiores aos obtidos nas demais regiões. Assim, se na Leste Rio-Grandense o preço recebido pelos produtores foi de Cr$ 1.690 por tonelada e na Sertão Pernambucano foi de Cr$ 1.908,40 por tonelada, na Sul Paulista o preço recebido era de apenas Cr$ 1.403,10 por tonelada, o que, certamente, originou a situação desfavorável da produção paulista. Se nesta região tivessem vigorado preços próximos aos das demais regiões, os resultados seriam francamente favoráveis aos produtores da Sul Paulista cuja produtividade física superava em 49,84 e 47,37% as produtividades obtidas na Sertão Pernambucano e na Leste Rio-Grandense, respectivamente.

Na região nordestina, por outro lado, mesmo supondo-se uma superestimação do valor do trabalho familiar, os resultados dificilmente seriam positivos como na Mesorregião Leste Rio-Grandense, ainda que se considere a similaridade das produtividades físicas obtidas nas duas regiões.

3.5 — Milho em grão

Para o estudo comparativo da produção de milho em grão foram selecionadas apenas duas Mesorregiões: a Oeste Paranaense e a Sul Goiano. Elas representavam, em 1975, parcela significativa do número de produtores, da quantidade produzida, da área colhida e do valor da produção de seus respectivos estados. Tal representatividade, na ordem citada, era de 52, 62, 58 e 61% na Oeste Paranaense e 23, 65, 54 e 65% na Sul Goiano. Quanto aos estabelecimentos especializados, as participações nos totais das suas Mesorregiões foram bem menores na Oeste Paranaense do que na Sul Goiano.

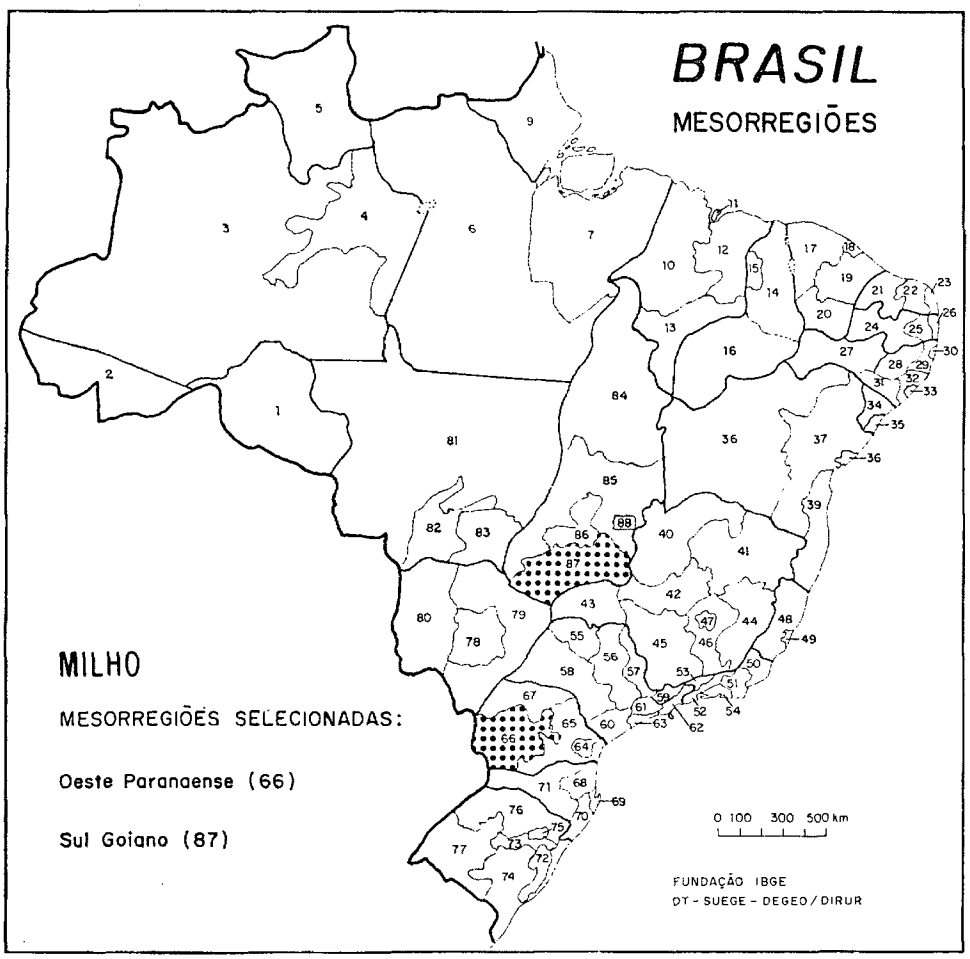
QUADRO 3.5.1

ÍNDICES DE ESPECIFICAÇÃO NA CULTURA DO MILHO

ESPECIFICAÇÃO	ÍNDICES (%)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
Índice de especialização da produção.....	7,74	29,80
Índice de especialização dos estabelecimentos...	3,42	5,52
Índice "verdadeiro" de especialização.....	0,26	1,64

O alto índice de especialização da produção nessa última região é, até certo ponto, surpreendente, considerada a forma tradicional de exploração do produto, predominantemente voltada para o consumo no estabelecimento ou para o abastecimento de pequenas indústrias locais.

Certamente, a grande expansão da indústria de rações para animais no País, principal consumidora de milho em grão, ocorrida a partir do primeiro qüinqüênio da década de 70, assim como a larga disponibilidade de terras na Mesorregião, é que podem em parte explicar tal concentração da produção. No caso da Oeste Paranaense, talvez a limitação das terras e a concorrência da soja é que possam justificar o relativamente baixo índice de especialização da produção encontrado.



Observe-se que as relações entre os índices de especialização da produção e os índices de especialização dos estabelecimentos, nas duas Mesorregiões, são bastante altos, — 0,95 na Oeste Paranaense e 0,74 na Sul Goiano —, significando uma relativamente pequena diferença entre as produtividades conseguidas nos estabelecimentos especializados e as médias das Mesorregiões (de 5% na primeria e 26% na segunda).

É na composição dos custos de produção e nos dispêndios por área colhida que se observam, no entanto, as especificidades no cultivo do milho nas duas áreas em estudo.

As diferenças fundamentais na composição dos custos entre as duas Mesorregiões estão, como se vê, nas participações relativas do custo do capital circulante e no custo do trabalho. Este custo é de maior peso na Oeste Paranaense (57,89%), enquanto que a importância do capital circulante é maior na Sul Goiano (46,64%). A maior utilização do trabalho, na primeira Mesorregião, reflete menor disponibilidade

QUADRO 3.5.2

ESTRUTURA DE CUSTOS DA PRODUÇÃO NOS ESTABELECIMENTOS ESPECIALIZADOS NA CULTURA DO MILHO

ESPECIFICAÇÃO	CUSTOS DA PRODUÇÃO (%)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
CUSTO TOTAL	**100,00**	**100,00**
Custo da terra	21,26	22,07
Custo do capital fixo	5,39	6,77
Custo do capital circulante	15,46	46,64
Custo do trabalho	57,89	24,52

de máquinas e de terras na região. Assim, na Oeste Paranaense, o valor das máquinas e instrumentos de trabalho por área colhida foi de apenas Cr$ 380,02 por hectare contra Cr$ 614,10 por hectare na Sul Goiano, enquanto que a média da área colhida por estabelecimento foi de 11,2 hectares na primeira região e de 82,4 hectares na segunda. Considerando ainda, que na Oeste Paranaense a relação entre a área colhida e o número de trabalhadores empregados era de 1,9 contra 10,4 na Sul Goiano, pode-se bem avaliar a intensidade de uso do trabalho na produção paranaense.

Uma rápida análise da composição do capital circulante nas duas Mesorregiões reforça as distinções entre elas.

QUADRO 3.5.3

COMPOSIÇÃO DO CAPITAL CIRCULANTE NA CULTURA DO MILHO

ITENS DAS DESPESAS	PARTICIPAÇÃO DAS DESPESAS (%)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
TOTAL	**100,00**	**100,00**
Insumos modernos	18,00	47,64
Medicamentos e alimentos para animais	7,94	2,53
Transporte da produção	4,92	6,56
Impostos e taxas	10,28	3,29
Juros e despesas bancárias	9,03	10,34
Aluguel de máquinas e equipamentos	5,21	0,90
Outras despesas	44,62	28,74

A alta participação das despesas com insumos modernos na Sul Goiano, cerca de 2,6 vezes a da Oeste Paranaense, não é suficiente para mostrar a sua importância na produção da primeira Mesorregião que, por área colhida, despendeu, em média, cerca de Cr$ 304,60 contra apenas Cr$ 45,30 na segunda, ou seja, aproximadamente sete vezes mais.

Dos demais componentes do custo do capital circulante mereceram ainda destaque as despesas com medicamentos e alimentos para animais, bastante expressivo na Oeste Paranaense (7,94%) e as "outras despesas". Quanto à primeira, é sempre válido lembrar uma das mais fortes restrições à análise de variáveis econômicas e financeiras baseadas em dados censitários agregados como feita neste trabalho, que é a não consideração do valor da variação nos estoques de animais como componente do valor da produção da pecuária. É, portanto, presumível uma margem de erro nas análises dos resultados apresentados em ambas Mesorregiões, mormente se considerarmos a importância do cultivo do milho como atividade complementar de explorações dedicadas preponderantemente à produção pecuária.

Nas "outras despesas", por outro lado, vale ressaltar os gastos com combustíveis e lubrificantes que novamente mostram a maior utilização de máquinas e equipamentos na Sul Goiano, que despendia Cr$ 101,90 por hectare colhido contra Cr$ 57,40 por hectare na Oeste Paranaense.

É interessante observar que as estruturas de custo da produção do milho em grão por área colhida, permitem evidenciar a importância do equilíbrio na alocação dos fatores produtivos.

Esta evidência é simplesmente demonstrada se aceita a hipótese de homogeneidade dos fatores e dos seus preços. Assim, *coeteris paribus*, se na Oeste Paranaense o dispêndio com fatores de produção por área colhida foi superior ao efetivado na Sul Goiano, seria de se esperar

QUADRO 3.5.4

CUSTOS E VALOR DA PRODUÇÃO POR ÁREA COLHIDA NA CULTURA DO MILHO

ESPECIFICAÇÃO	VALOR (Cr$/ha)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
CUSTO TOTAL	**1 625,90**	**1 371,40**
Custo da terra	345,70	302,70
Custo do capital fixo	87,60	92,90
Custo do capital circulante	251,30	639,60
Custo do trabalho	941,30	336,20
Valor da produção	**1 420,10**	**1 519,90**

maiores produtividades físicas naquela região do que nessa última. Tal não ocorreu, tendo sido as quantidades por área colhida superior na Mesorregião goiana (2,336 t/ha) do que na paranaense (2,262 t/ha).

Considerando que os preços efetivamente pagos, aos produtores pelo milho em grão na Oeste Paranaense foram inferiores aos pagos aos produtores na Sul Goiano, fica claro que, ainda sob a homogeneidade dos fatores e sob condições meteorológicas normais, a combinação dos fatores nesta Mesorregião se mostrou bem mais eficiente do que na Oeste Paranaense. O prejuízo líquido por hectare colhido nesta região, de — Cr$ 205,70, frente ao lucro líquido, Cr$ 148,50, na região goiana, parece ser, portanto, o resultado da má combinação ou limitação de fatores.

4 — CONCLUSÃO

Dos resultados apresentados, para as quinze Mesorregiões, pode-se tirar algumas relevantes conclusões acerca da atividade agrícola no País. A primeira delas e, a mais óbvia, refere-se às acentuadas diferenças de tecnologia de produção nas diversas Mesorregiões, principalmente em termos de combinação e intensidade de usos dos diversos fatores, monetariamente avaliados. Como é lógico, diferenciadas, também, serão as funções de custo da produção, em que se podem realçar três componentes básicos para a elaboração de políticas relativas ao setor, a saber: o custo da terra, o custo do trabalho e os custos do capital circulante, mais especialmente, os de caráter institucional.

O custo da terra, considerado isoladamente, foi bastante significativo na produção do arroz nas quatro Mesorregiões consideradas, na produção do milho, na produção da cana-de-açúcar no Estado do Rio de Janeiro e na produção de cebola nas Mesorregiões Sertão Pernambucano e Sul Paulista. Nesta última, então, o custo da terra superou os 43% do custo total da produção, significando parcelas substanciais de recursos dos produtores sem a propriedade da terra, evidenciando os altos custos sociais derivados das relações fundiárias existentes.

Quanto ao custo do fator trabalho, observou-se a sua estreita relação com o nível de utilização de maquinarias, com o tipo de produto explorado e com os níveis salariais vigentes nas regiões. Assim, a distinção na participação dos custos do trabalho no custo total da produção do arroz entre as regiões centrais do País (Triângulo Mineiro e Goiânia) e as regiões sulinas (Leste Rio-Grandense e Oeste Gaúcho), bem evidenciam essas relações. O mesmo ocorreu na produção da cana-de-açúcar, com nítidas distinções entre a produção nordestina e a do sudoeste do País, na produção da cebola e na do milho.

Os custos do capital circulante, por sua vez, além de mostrarem as diferentes intensidades de utilização de insumos agrícolas, especial-

mente os considerados modernos, evidenciaram a importância dos custos aqui chamados de institucionais. Na verdade estas variáveis se constituem em campos de manobra para a execução de políticas de incentivo à produção agrícola e merecem uma maior observação por parte dos governos em todos os níveis: federal, estadual ou municipal.

Neste trabalho ficou evidenciada, também, a importância dos níveis de preços obtidos pelos produtores rurais nos resultados econômico-financeiros das explorações rurais. Assim, das seis Mesorregiões que apresentaram resultados negativos, Goiânia, Campinas e Ribeirão Preto, Sertão Pernambucano, Sul Paulista, Oeste Paranaense e Triângulo Mineiro, apenas estas duas últimas é que continuariam a apresentar prejuízos após uma simulação em que a produção agrícola de todas as Mesorregiões fosse avaliada segundo o preço mais alto nelas encontrado. O caso destas duas Mesorregiões, no entanto, parece estar mais ligado à possível existência de atividades complementares, como a pecuária, ou a exploração com maiores ciclos produtivos como as culturas permanentes, o que certamente tenderia a subestimar o valor da produção obtido.

Como última observação, vale notar o papel dos financiamentos obtidos pelos produtores nas Mesorregiões com resultados negativos (Quadro 4.1), que seria, justamente, o de cobrir tais resultados, mostrando a alta dependência desses produtores ao sistema de crédito rural.

Deve-se ainda notar, aqui como um reparo a estas conclusões e a este trabalho como um todo, que trata-se de um estudo exploratório, quanto à utilização de dados censitários, com sérias restrições para uma análise econômico-financeira de agregados de estabelecimentos agrícolas, apesar de permitirem a criação e o aperfeiçoamento de estudos relativos à tipologia agrícola e à análise comparativa da produção agropecuária entre regiões, como aqui tentado.

Quadro 4.1

PREJUÍZO LÍQUIDO E FINANCIAMENTOS OBTIDOS POR ÁREA COLHIDA

MESORREGIÕES	PRODUTOS E VARIÁVEIS (Cr$/ha)							
	Arroz		Cana-de-açúcar		Cebola		Milho	
	Prejuízo líquido por ha	Financiamento por ha	Prejuízo líquido por ha	Financiamento por ha	Prejuízo líquido por ha	Financiamento por ha	Prejuízo líquido por ha	Financiamento por ha
Triângulo Mineiro.....	516,70	784,00	—	—	—	—	—	—
Goiânia.....	86,10	891,00	—	—	—	—	—	—
Sertão Pernambucano...	—	—	—	—	3 725,50	1 754,00	—	—
Sul Paulista.....	—	—	—	—	3 258,70	3 096,00	—	—
Oeste Paranaense.....	—	—	—	—	—	—	205,70	244,60
Campinas e Ribeirão Preto.....	—	—	199,00	1 601,00	—	—	—	—

5 — ANEXO

Todos os dados das tabelas deste anexo foram conseguidos ou derivados diretamente de listagens especiais do Censo Agropecuário de 1975.

TABELA 5.1.1

VALOR DOS BENS NA CULTURA DO ARROZ — 1975

BENS	VALOR DOS BENS (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Valor total dos bens	**611 394**	**1 115 759**	**5 268 734**	**2 710 742**
Prédios residenciais e para fins sociais	17 450	46 148	140 420	97 073
Terras (próprias, arrendadas ou ocupadas)	500 674	886 658	3 798 128	1 645 351
Culturas permanentes	19 054	831	1 993	1 709
Matas plantadas	2 353	52	60 113	12 469
Veículos e outros meios de transporte	4 946	15 405	82 471	74 911
Animais de criação e de trabalho	25 951	58 175	261 568	148 127
Máquinas e instrumentos de trabalho	22 485	59 802	598 633	561 741
Instalações e outras benfeitorias	18 481	48 688	325 408	169 361

TABELA 5.1.2

MÃO-DE-OBRA EMPREGADA NA CULTURA DO ARROZ — 1975

CATEGORIA DA MÃO-DE-OBRA	PESSOAL OCUPADO			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Responsável e membros não remunerados da família	3 000	10 074	7 592	4 572
Empregados permanentes	499	1 111	6 370	3 858
Empregados temporários	1 375	3 158	7 338	4 044
Parceiros	438	2 042	153	121
Outra condição	49	198	58	86

NOTA — Pessoal ocupado com 14 anos e mais.

TABELA 5.1.3

VALOR DO TRABALHO FAMILIAR NA CULTURA DO ARROZ — 1975

ESPECIFICAÇÃO	VALOR DO TRABALHO FAMILIAR (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Número de empregados permanentes e temporários (1)	1 874	4 269	13 708	7 902
Total dos salários pagos	6 859	14 017	76 781	—53 581
Salário médio	3 660,10	3 283,40	5 601,20	6 780,70
Valor do trabalho familiar	10 980	33 077	42 524	31 001

1) Pessoal ocupado.

TABELA 5.1.4

CUSTO DA TERRA NA CULTURA DO ARROZ — 1975

ESPECIFICAÇÃO	CUSTO DA TERRA (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Área plantada com cultivos permanentes e temporárias (ha)	37 433	116 571	187 827	140 931
Área arrendada (ha)	10 295	22 085	137 507	138 961
Valor dos arrendamentos pagos	5 961	7 726	132 683	97 995
Valor médio de arrendamento (Cr$/ha)	579,00	349,80	964,90	705,20
Custo da terra	21 674	40 777	181 234	99 385

TABELA 5.1.5

CUSTO DO CAPITAL FIXO NA CULTURA DO ARROZ — 1975

BENS DE CAPITAL	DEPRECIAÇÃO DO CAPITAL FIXO (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
CUSTO TOTAL DO CAPITAL FIXO	**4 778**	**10 910**	**87 268**	**74 142**
Prédios residenciais e para fins sociais...	349	923	2 808	1 941
Culturas permanentes.....	762	33	80	68
Veículos e outros meios de transportes	495	1 540	8 247	7 491
Máquinas e instrumentos de trabalho..	2 248	5 980	59 863	56 174
Instalações e outras benfeitorias.....	924	2 434	16 270	8 468

TABELA 5.1.6

CUSTO DO CAPITAL CIRCULANTE NA CULTURA DO ARROZ — 1975

ITENS DAS DESPESAS	CUSTO DO CAPITAL CIRCULANTE (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
TOTAL.....	**18 537**	**78 375**	**434 897**	**309 127**
Adubos e corretivos.....	4 672	28 742	106 965	48 208
Sementes e mudas.....	1 812	2 714	59 456	42 666
Defensivos agrícolas.....	803	1 420	32 689	17 411
Medicamentos e alimentos para animais	722	1 937	6 500	4 543
Transporte da produção.....	525	1 472	25 839	14 571
Impostos e taxas.....	972	9 521	14 704	10 082
Juros e despesas bancárias.....	2 240	9 815	51 023	35 890
Aluguel de máquinas e equipamentos..	748	2 689	3 797	2 697
Outras despesas.....	6 043	20 065	133 984	133 059

TABELA 5.1.7

CUSTO DO TRABALHO NA CULTURA DO ARROZ — 1975

CATEGORIAS DE REMUNERAÇÃO	CUSTO DO TRABALHO (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
TOTAL	**27 380**	**78 144**	**189 088**	**113 100**
Salários pagos	6 859	14 017	76 781	53 581
Pagamentos imputados	10 980	33 077	42 524	31 001
Quota-parte entregue a parceiros	3 316	11 441	23 859	9 975
Empreitadas apenas com mão-de-obra	1 609	6 560	26 705	7 985
Empreitadas com mão-de-obra e equipamentos	4 616	13 049	19 219	10 558

TABELA 5.1.8

CUSTO TOTAL NA CULTURA DO ARROZ — 1975

ESPECIFICAÇÃO	CUSTOS (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
CUSTO TOTAL	**72 369**	**208 206**	**892 487**	**595 754**
Custo da terra	21 674	40 777	181 234	99 385
Custo do capital fixo	4 778	10 910	87 268	74 142
Custo do capital circulante	18 537	78 375	434 897	309 127
Custo do trabalho	27 380	78 144	189 088	113 100

TABELA 5.1.9

LUCRO LÍQUIDO NA ATIVIDADE NA CULTURA DO ARROZ — 1975

ESPECIFICAÇÃO	VALORES (Cr$ 1 000)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Valor total das receitas	53 025	195 158	1 251 258	887 211
Valor total da produção	53 007	198 173	1 278 291	913 713
Custo total	72 369	208 206	892 487	595 754
Lucro líquido (receitas — custo total)	— 19 344	— 13 048	358 771	291 457
Lucro líquido (produção — custo total)	— 19 362	— 10 033	385 803	317 959

TABELA 5.1.10

INDICADORES DE DESEMPENHO POR ÁREA PLANTADA NA CULTURA DO ARROZ — 1975

ESPECIFICAÇÃO	INDICADORES DE DESEMPENHO (Cr$/ha)			
	Mesorregiões			
	Triângulo Mineiro	Goiânia	Leste Rio-Grandense	Oeste Gaúcho
Lucro líquido por valor de produção	—365,30	—50,60	301,80	348,00
Lucro líquido por valor de bens	—31,70	—9,00	73,20	117,40
Valor da produção por valor de bens	86,70	177,60	242,70	337,40
Custo do capital fixo	127,60	93,60	464,60	526,10
Custo do capital circulante	495,20	672,30	2 315,40	2 193,50
Custo do trabalho	731,40	670,40	1 006,70	802,50
Custo da terra	579,00	349,80	964,90	705,20
Valor da produção	**1 416,50**	**1 700,00**	**6 805,70**	**6 483,40**

TABELA 5.2.1

VALOR DOS BENS NA CULTURA DA BATATA-INGLESA — 1975

BENS	VALOR DOS BENS (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
VALOR TOTAL DOS BENS	**200 191**	**187 597**	**46 964**
Prédios residenciais e para fins sociais	8 786	24 936	6 196
Terras (próprias, arrendadas ou ocupadas)	139 717	112 208	24 176
Culturas permanentes	5 236	102	51
Matas plantadas	2 809	720	10
Veículos e outros meios de transporte	9 147	12 124	5 007
Animais de criação e de trabalho	9 987	4 682	1 481
Máquinas e instrumentos de trabalho	19 776	19 864	6 877
Instalações e outras benfeitorias	4 733	12 961	3 166

TABELA 5.2.2

MÃO-DE-OBRA EMPREGADA NA CULTURA DA BATATA-INGLESA — 1975

CATEGORIA DA MÃO-DE-OBRA	PESSOAL OCUPADO		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Responsável e membros não remunerados da família	1 629	1 867	573
Empregados permanentes	654	60	73
Empregados temporários	2 500	1 435	636
Parceiros	36	1	1
Outra condição	8	1	0

NOTA — Pessoal ocupado com 14 anos e mais.

TABELA 5.2.3

VALOR DO TRABALHO FAMILIAR NA CULTURA DA BATATA-INGLESA — 1975

ESPECIFICAÇÃO	VALOR DO TRABALHO FAMILIAR (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Número de empregados permanentes e temporários (1)	3 154	1 495	709
Total dos salários pagos	9 323	2 948	1 215
Salário médio	2 955,90	1 971,90	1 713,70
Valor do trabalho familiar	4 815	3 682	982

(1) Pessoal ocupado.

TABELA 5.2.4

CUSTO DA TERRA NA CULTURA DA BATATA-INGLESA — 1975

ESPECIFICAÇÃO	CUSTO DA TERRA (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Área plantada com cultivos permanentes e temporárias (ha)	8 676	6 341	2 435
Área arrendada (ha)	2 927	872	568
Valor dos arrendamentos pagos	1 981	530	375
Valor médio de arrendamento (Cr$/ha)	676,80	607,80	656,70
Custo da terra	5 872	3 854	1 599

TABELA 5.2.5

CUSTO DO CAPITAL FIXO NA CULTURA DA BATATA-INGLESA — 1975

BENS DO CAPITAL	DEPRECIAÇÃO DO CAPITAL FIXO (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
CUSTO TOTAL DO CAPITAL FIXO....	**3 515**	**4 349**	**1 473**
Prédios residenciais e para fins sociais.....	176	499	124
Culturas permanentes.....	209	4	2
Veículos e outros meios de transporte.....	915	1 212	501
Máquinas e instrumentos de trabalho.....	1 978	1 986	688
Instalações e outras benfeitorias.....	237	648	158

TABELA 5.2.6

CUSTO DO CAPITAL CIRCULANTE NA CULTURA DA BATATA-INGLESA — 1975

ITENS DAS DESPESAS	CUSTO DO CAPITAL CIRCULANTE (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
TOTAL.....	**49 476**	**19 816**	**12 225**
Adubos e corretivos.....	20 197	8 995	4 933
Sementes e mudas.....	9 434	2 715	1 806
Defensivos agrícolas.....	5 100	1 664	1 366
Medicamentos e alimentos para animais.....	739	143	62
Transporte da produção.....	3 769	38	646
Impostos e taxas.....	841	206	215
Juros e despesas bancárias.....	1 752	669	455
Aluguel de máquinas e equipamentos.....	935	145	114
Outras despesas.....	6 709	5 241	2 628

TABELA 5.2.7

CUSTO DO TRABALHO NA CULTURA DA BATATA-INGLESA — 1975

CATEGORIAS DE REMUNERAÇÃO	CUSTO DO TRABALHO (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
TOTAL	**17 022**	**7 759**	**2 616**
Salários pagos	9 323	2 948	1 215
Pagamentos imputados	4 815	3 682	982
Quota-parte entregue a parceiros	1 728	65	1
Empreitadas apenas com mão-de-obra	581	641	247
Empreitadas com mão-de-obra e equipamentos	575	423	171

TABELA 5.2.8

CUSTO TOTAL NA CULTURA DA BATATA-INGLESA — 1975

ESPECIFICAÇÃO	CUSTOS (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
CUSTO TOTAL	**75 885**	**35 778**	**17 913**
Custo da terra	5 872	3 854	1 599
Custo do capital fixo	3 515	4 349	1 473
Custo do capital circulante	49 476	19 816	12 225
Custo do trabalho	17 022	7 759	2 616

TABELA 5.2.9

LUCRO LÍQUIDO NA ATIVIDADE NA CULTURA DA BATATA-INGLESA — 1975

ESPECIFICAÇÃO	VALORES (Cr$ 1 000)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Valor total das receitas	109 624	46 544	23 065
Valor total da produção	110 427	48 553	23 597
Custo total	75 885	35 778	17 913
Lucro líquido (receitas — custo total)	33 739	10 766	5 152
Lucro líquido (produção — custo total)	34 542	12 775	5 684

TABELA 5.2.10

INDICADORES DE DESEMPENHO POR ÁREA PLANTADA NA CULTURA DA BATATA-INGLESA — 1975

ESPECIFICAÇÃO	INDICADORES DE DESEMPENHO (Cr$/ha)		
	Mesorregiões		
	Sudoeste Mineiro	Curitiba	Leste Paranaense
Lucro líquido por valor de produção	312,80	263,10	240,90
Lucro líquido por valor de bens	172,60	68,10	121,00
Valor da produção por valor de bens	551,70	259,00	502,40
Custo do capital fixo	405,10	685,90	604,90
Custo do capital circulante	5 702,60	3 125,10	5 020,50
Custo do trabalho	1 962,00	1 223,60	1 074,30
Custo da terra	676,80	607,80	656,70
Valor da produção	**12 727,90**	**7 657,00**	**9 690,80**

TABELA 5.3.1

VALOR DOS BENS NA CULTURA DA CANA-DE-AÇÚCAR — 1975

BENS	VALOR DOS BENS (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
VALOR TOTAL DOS BENS	**3 796 335**	**3 644 967**	**4 529 070**	**7 113 644**
Prédios residenciais e para fins sociais	451 022	276 699	144 676	1 093 146
Terras (próprias, arrendadas ou ocupadas)	2 579 255	2 721 667	3 661 500	4 457 747
Culturas permanentes	38 820	13 640	1 032	128 743
Matas plantadas	2 336	50	247 540	113 636
Veículos e outros meios de transporte	185 290	227 269	118 410	328 492
Animais de criação e de trabalho	336 171	138 987	67 288	125 979
Máquinas e instrumentos de trabalho	124 459	205 831	184 987	555 860
Instalações e outras benfeitorias	78 982	60 824	103 637	310 041

TABELA 5.3.2

MÃO-DE-OBRA EMPREGADA NA CULTURA DA CANA-DE-AÇÚCAR — 1975

CATEGORIA DA MÃO-DE-OBRA	PESSOAL OCUPADO			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Responsável e membros não remunerados da família	11 302	8 478	11 100	9 783
Empregados permanentes	50 127	42 127	8 891	28 402
Empregados temporários	18 585	17 961	6 833	12 907
Parceiros	88	134	90	293
Outra condição	2 179	324	15	95

NOTA — Pessoal ocupado com 14 anos e mais.

TABELA 5.3.3

VALOR DO TRABALHO FAMILIAR NA CULTURA DA CANA-DE-AÇÚCAR — 1975

ESPECIFICAÇÃO	VALOR DO TRABALHO FAMILIAR (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Número de empregados permanentes e temporários (1)	68 712	60 117	15 724	41 309
Total dos salários pagos	393 584	286 916	80 207	366 087
Salário médio	5 728,00	4 772,60	5 100,90	8 862,20
Valor do trabalho familiar	64 738	40 462	56 620	86 699

(1) Pessoal ocupado.

TABELA 5.3.4

CUSTO DA TERRA NA CULTURA DA CANA-DE-AÇÚCAR — 1975

ESPECIFICAÇÃO	CUSTO DA TERRA (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Área plantada com cultivos permanentes e temporários (ha)	358 186	279 273	168 020	550 163
Área arrendada (ha)	208 911	81 584	9 644	96 594
Valor dos arrendamentos pagos	29 174	13 920	4 579	50 524
Valor médio de arrendamento (Cr$/ha)	139,60	170,60	474,80	523,10
Custo da terra	50 003	47 644	79 776	287 790

TABELA 5.3.5

CUSTO DO CAPITAL FIXO NA CULTURA DA CANA-DE-AÇÚCAR — 1975

BENS DE CAPITAL	DEPRECIAÇÃO DO CAPITAL FIXO (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
CUSTO TOTAL DO CAPITAL FIXO	**45 497**	**52 431**	**38 457**	**130 950**
Prédios residenciais e para fins sociais	9 020	5 534	2 894	21 863
Culturas permanentes	1 553	546	41	5 150
Veículos e outros meios de transporte	18 529	22 727	11 841	32 849
Máquinas e instrumentos de trabalho	12 446	20 583	18 499	55 586
Instalações e outras benfeitorias	3 949	3 041	5 182	15 502

TABELA 5.3.6

CUSTO DO CAPITAL CIRCULANTE DA CULTURA DA CANA-DE-AÇÚCAR — 1975

ITENS DAS DESPESAS	CUSTO DO CAPITAL CIRCULANTE (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
TOTAL	**441 620**	**370 864**	**165 583**	**914 849**
Adubos e corretivos	179 525	163 751	35 838	330 311
Sementes e mudas	3 154	2 101	11 587	29 809
Defensivos agrícolas	8 373	3 179	2 564	50 201
Medicamentos e alimentos para animais	10 721	5 179	1 726	9 151
Transporte da produção	46 414	34 642	16 797	134 076
Impostos e taxas	75 451	27 338	4 221	36 227
Juros e despesas bancárias	32 145	30 224	23 310	60 898
Aluguel de máquinas e equipamentos	2 497	553	7 620	21 746
Outras despesas	83 340	103 897	61 920	242 430

TABELA 5.3.7

CUSTO DO TRABALHO NA CULTURA DA CANA-DE-AÇÚCAR — 1975

CATEGORIAS DE REMUNERAÇÃO	CUSTO DO TRABALHO (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
TOTAL	**538 216**	**399 553**	**178 001**	**677 085**
Salários pagos	393 584	286 916	80 207	366 087
Pagamentos imputados	64 738	40 462	56 620	86 699
Quota-parte entregue a parceiros	98	321	813	4 657
Empreitadas apenas com mão-de-obra	57 970	60 936	29 769	143 233
Empreitadas com mão-de-obra e equipamentos	21 826	10 918	10 592	76 409

TABELA 5.3.8

CUSTO TOTAL NA CULTURA DA CANA-DE-AÇÚCAR — 1975

ESPECIFICAÇÃO	CUSTOS (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
CUSTO TOTAL	**1 075 336**	**870 492**	**461 817**	**2 010 674**
Custo da terra	50 003	47 644	79 776	287 790
Custo do capital fixo	45 497	52 431	38 457	130 950
Custo do capital circulante	441 620	370 864	165 583	914 849
Custo do trabalho	538 216	399 553	178 001	677 085

TABELA 5.3.9

LUCRO LÍQUIDO NA ATIVIDADE NA CULTURA DA CANA-DE-AÇÚCAR — 1975

ESPECIFICAÇÃO	VALORES (Cr$ 1 000)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Valor total das receitas	1 124 248	949 849	462 694	1 903 847
Valor total da produção	1 135 160	950 909	465 657	1 901 207
Custo total	1 075 336	870 492	461 817	2 010 674
Lucro líquido (receitas — custo total)	48 912	79 357	877	—106 827
Lucro líquido (produção — custo total)	59 824	80 417	3 840	—109 467

TABELA 5.3.10

INDICADORES DE DESEMPENHO POR ÁREA PLANTADA NA CULTURA DA CANA-DE-AÇÚCAR — 1975

ESPECIFICAÇÃO	INDICADORES DE DESEMPENHO (Cr$/ha)			
	Mesorregiões			
	Mata Pernambucana	Mata Alagoana	Norte Fluminense	Campinas e Ribeirão Preto
Lucro líquido por valor de produção	52,70	84,60	8,20	—57,60
Lucro líquido por valor de bens	15,80	22,10	0,90	—15,40
Valor da produção por valor de bens	299,00	260,90	103,70	267,40
Custo do capital fixo	127,00	187,70	228,90	238,00
Custo do capital circulante	1 232,90	1 328,00	985,50	1 662,90
Custo do trabalho	1 502,60	1 430,70	1 059,40	1 230,70
Custo da terra	139,60	170,60	474,80	523,10
Valor da produção	3 169,20	3 404,90	2 771,40	3 455,70

TABELA 5.4.1

VALOR DOS BENS NA CULTURA DA CEBOLA — 1975

BENS	VALOR DOS BENS (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
VALOR TOTAL DOS BENS	**274 368**	**21 057**	**183 203**
Prédios residenciais e para fins sociais	38 161	1 881	12 651
Terras (próprias, arrendadas ou ocupadas)	151 659	8 767	129 217
Culturas permanentes	425	151	405
Matas plantadas	4 498	0	497
Veículos e outros meios de transporte	13 366	567	7 371
Animais de criação e de trabalho	32 803	2 074	2 497
Máquinas e instrumentos de trabalho	7 838	1 932	24 722
Instalações e outras benfeitorias	25 618	5 685	5 843

TABELA 5.4.2

MÃO-DE-OBRA EMPREGADA NA CULTURA DA CEBOLA — 1975

CATEGORIA DA MÃO-DE-OBRA	PESSOAL OCUPADO		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Responsável e membros não remunerados da família	11 260	1 020	2 867
Empregados permanentes	369	13	109
Empregados temporários	413	127	928
Parceiros	179	328	108
Outra condição	77	2	2

NOTA — Pessoal ocupado com 14 anos e mais.

TABELA 5.4.3

VALOR DO TRABALHO FAMILIAR NA CULTURA DA CEBOLA — 1975

ESPECIFICAÇÃO	VALOR DO TRABALHO FAMILIAR (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Número de empregados permanentes e temporários (1)	782	140	1 037
Total dos salários pagos	3 476	1 499	3 426
Salário médio	4 445,00	10 707,10	3 303,80
Valor do trabalho familiar	50 051	10 921	9 472

(1) Pessoal ocupado.

TABELA 5.4.4

CUSTO DA TERRA NA CULTURA DA CEBOLA — 1975

ESPECIFICAÇÃO	CUSTO DA TERRA (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Área plantada com cultivos permanentes e temporárias (ha)	6 107	1 680	2 911
Área arrendada (ha)	3 364	374	437
Valor dos arrendamentos pagos	3 928	851	4 113
Valor médio de arrendamento (Cr$/ha)	116,77	227,54	941,19
Custo da Terra	7 131	3 823	27 398

TABELA 5.4.5

CUSTO DO CAPITAL FIXO NA CULTURA DA CEBOLA — 1975

BENS DO CAPITAL	DEPRECIAÇÃO DO CAPITAL FIXO (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
CUSTO TOTAL DO CAPITAL FIXO	**4 182**	**578**	**3 770**
Prédios residenciais e para fins sociais..	763	38	253
Culturas permanentes.....	17	6	16
Veículos e outros meios de transporte..	1 337	57	737
Máquinas e instrumentos de trabalho..	784	193	2 472
Instalações e outras benfeitorias.....	1 281	284	292

TABELA 5.4.6

CUSTO DO CAPITAL CIRCULANTE NA CULTURA DA CEBOLA — 1975

ITENS DAS DESPESAS	CUSTO DO CAPITAL CIRCULANTE (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
TOTAL.....	**21 387**	**4 028**	**18 565**
Adubos e corretivos.....	8 217	1 085	7 101
Sementes e mudas.....	1 846	441	1 439
Defensivos agrícolas.....	975	597	1 974
Medicamentos e alimentos para animais	874	104	62
Transporte da produção.....	1 174	120	732
Impostos e taxas.....	1 975	29	146
Juros e despesas bancárias.....	496	176	695
Aluguel de máquinas e equipamentos..	24	106	476
Outras despesas.....	5 806	1 370	5 940

TABELA 5.4.7

CUSTO DO TRABALHO NA CULTURA DA CEBOLA — 1975

CATEGORIAS DE REMUNERAÇÃO	CUSTO DO TRABALHO (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
TOTAL	**55 086**	**14 690**	**13 796**
Salários pagos	3 476	1 499	3 426
Pagamentos imputados	50 051	10 921	9 472
Quota-parte entregue a parceiros	1 347	2 171	762
Empreitadas apenas com mão-de-obra	29	88	86
Empreitadas com mão-de-obra e equipamentos	183	11	50

TABELA 5.4.8

CUSTO TOTAL NA CULTURA DA CEBOLA — 1975

ESPECIFICAÇÃO	CUSTOS (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
CUSTO TOTAL	**87 786**	**23 119**	**63 529**
Custo da terra	7 131	3 823	27 398
Custo do capital fixo	4 182	578	3 770
Custo do capital circulante	21 387	4 028	18 565
Custo do trabalho	55 086	14 690	13 796

TABELA 5.4.9

LUCRO LÍQUIDO NA ATIVIDADE NA CULTURA DA CEBOLA — 1975

ESPECIFICAÇÃO	VALORES (Cr$ 1 000)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Valor total das receitas.....	95 580	16 908	54 231
Valor total da produção.....	99 580	16 860	54 043
Custo total.....	87 786	23 119	63 529
Lucro líquido (receitas — custo total)..	7 794	—6 211	—9 298
Lucro líquido (produção — custo total)	11 794	—6 259	—9 486

TABELA 5.4.10

INDICADORES DE DESEMPENHO POR ÁREA PLANTADA NA CULTURA DA CEBOLA

ESPECIFICAÇÃO	INDICADORES DE DESEMPENHO (Cr$/ha)		
	Mesorregiões		
	Leste Rio-Grandense	Sertão Pernambucano	Sul Paulista
Lucro líquido por valor de produção..	118,40	—371,20	—172,00
Lucro líquido por valor de bens.....	43,00	—297,20	—51,80
Valor da produção por valor de bens..	363,10	800,70	295,00
Custo do capital fixo.....	684,80	344,00	1 295,10
Custo do capital circulante.....	3 502,00	2 397,60	6 377,50
Custo do trabalho.....	9 020,10	8 744,00	4 739,30
Custo da terra.....	1 167,70	2 275,60	9 411,90
Valor da produção.....	16 305,90	10 035,70	18 565,10

TABELA 5.5.1

VALOR DOS BENS NA CULTURA DO MILHO — 1975

BENS	VALOR DOS BENS (Cr$ 1000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
VALOR TOTAL DOS BENS	**997 745**	**1 026 452**
Prédios residenciais e para fins sociais	46 970	21 541
Terras (próprias, arrendadas ou ocupadas)	772 532	851 825
Culturas permanentes	6 722	340
Matas plantadas	72 984	700
Veículos e outros meios de transporte	13 911	12 721
Animais de criação e de trabalho	32 666	52 368
Máquinas e instrumentos de trabalho	28 808	59 910
Instalações e outras benfeitorias	23 152	27 047

TABELA 5.5.2

MÃO-DE-OBRA EMPREGADA NA CULTURA DO MILHO — 1975

CATEGORIA DA MÃO-DE-OBRA	PESSOAL OCUPADO	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
Responsável e membros não remunerados da família	18 564	2 832
Empregados permanentes	997	1 261
Empregados temporários	2 058	1 789
Parceiros	229	124
Outra condição	53	122

NOTA — *Pessoal ocupado com 14 anos e mais.*

TABELA 5.5.3

VALOR DO TRABALHO FAMILIAR NA CULTURA DO MILHO — 1975

ESPECIFICAÇÃO	VALOR DO TRABALHO FAMILIAR (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
Número de empregados permanentes e temporários (1)	3 055	3 050
Total dos salários pagos	8 979	10 546
Salário médio	2 939,10	2 457,70
Valor do trabalho familiar	54 561	9 792

(1) Pessoal ocupado.

TABELA 5.5.4

CUSTO DA TERRA NA CULTURA DO MILHO — 1975

ESPECIFICAÇÃO	CUSTO DA TERRA (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
Área plantada com cultivos permanentes e temporárias (ha)	75 767	97 555
Área arrendada (ha)	16 289	28 495
Valor dos arrendamentos pagos	5 631	8 626
Valor médio de arrendamento (Cr$/ha)	345,70	302,70
Custo da terra	26 193	29 530

TABELA 5.5.5

CUSTO DO CAPITAL FIXO NA CULTURA DO MILHO — 1975

BENS DO CAPITAL	DEPRECIAÇÃO DO CAPITAL FIXO (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
CUSTO TOTAL DO CAPITAL FIXO	**6 635**	**9 060**
Prédios residenciais e para fins sociais.....	936	431
Culturas permanentes	269	14
Veículos e outros meios de transporte.....	1 391	1 272
Máquinas e instrumentos de trabalho.....	2 881	5 991
Instalações e outras benfeitorias.....	1 158	1 352

TABELA 5.5.6

CUSTO DO CAPITAL CIRCULANTE NA CULTURA DO MILHO — 1975

ITENS DAS DESPESAS	CUSTO DO CAPITAL CIRCULANTE (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
TOTAL.....	**19 039**	**62 398**
Adubos e corretivos.....	1 291	23 453
Sementes e mudas.....	1 839	4 673
Defensivos agrícolas.....	299	1 588
Medicamentos e alimentos para animais.....	1 511	1 581
Transporte da produção.....	936	4 095
Impostos e taxas.....	1 957	2 050
Juros e despesas bancárias.....	1 720	6 452
Aluguel de máquinas e equipamentos.....	991	574
Outras despesas.....	8 495	17 932

TABELA 5.5.7

CUSTO DO TRABALHO NA CULTURA DO MILHO — 1975

CATEGORIAS DE REMUNERAÇÃO	CUSTO DO TRABALHO (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
TOTAL	**71 316**	**32 801**
Salários pagos	8 979	10 546
Pagamentos imputados	54 561	9 792
Quota-parte entregue a parceiros	905	807
Empreitadas apenas com mão-de-obra	5 142	3 664
Empreitadas com mão-de-obra e equipamentos	1 729	7 992

TABELA 5.5.8

CUSTO TOTAL NA CULTURA DO MILHO — 1975

ESPECIFICAÇÃO	CUSTOS (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
CUSTO TOTAL	**123 183**	**133 789**
Custo da terra	26 193	29 530
Custo do capital fixo	6 635	9 060
Custo do capital circulante	19 039	62 398
Custo do trabalho	71 316	32 801

TABELA 5.5.9

LUCRO LÍQUIDO NA ATIVIDADE NA CULTURA DO MILHO — 1975

ESPECIFICAÇÃO	VALORES (Cr$ 1 000)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
Valor total das receitas	103 225	150 567
Valor total da produção	107 593	148 272
Custo total	123 183	133 789
Lucro líquido (receitas — custo total)	—19 958	16 778
Lucro líquido (produção — custo total)	—15 590	14 483

TABELA 5.5.10

INDICADORES DE DESEMPENHO POR ÁREA PLANTADA NA CULTURA DO MILHO — 1975

ESPECIFICAÇÃO	INDICADORES DE DESEMPENHO (Cr$ / ha)	
	Mesorregiões	
	Oeste Paranaense	Sul Goiano
Lucro líquido por valor de produção	—144,90	97,70
Lucro líquido por valor de bens	—15,60	14,10
Valor da produção por valor de bens	107,90	144,50
Custo do capital fixo	87,60	92,90
Custo do capital circulante	251,30	639,60
Custo do trabalho	941,30	336,20
Custo da terra	345,70	302,70
Valor da produção	1 420,10	1 519,90

6 — BIBLIOGRAFIA

ALMEIDA, Manoel Lizardo. — Seguro agrícola e proagro. *Lavoura Arroense EXPOINTER/80, 33* (323).

ALOE, Armando & VALLE, Francisco. — *Contabilidade Agrícola.* 6.ª ed. São Paulo, Atlas, 1978. 237 p.

HOFFMANN, Rodolfo *et alii. Administração da Empresa Agrícola.* São Paulo, Pioneira, 1976. 323 p.

OLIVEIRA, Contalício Preto de. *Economia e Administração Rural.* Porto Alegre, Sulina, 1969. 162 p.

SILVA, Jairo Augusto & ROCHA, Sonia Maria Rodrigues. — Balanço uso-disponibilidade de 15 produtos agrícolas alimentares — uma análise a nível mesorregional. Rio de Janeiro, IBGE, (no prelo).

BIBLIOGRAFIA

PUBLICAÇÕES DE INTERESSE PARA A ESTATÍSTICA EDITADAS PELO IBGE NO PERÍODO DE ABRIL A SETEMBRO DE 1982 *

ANUÁRIO ESTATÍSTICO DO BRASIL. Rio de Janeiro, 1981.

ARMAZENAGEM E ESTOCAGEM A SECO E A FRIO — 1978. Rio de Janeiro, 1982. v. 3, t. 1 — Região Norte.

———. t. 2 — Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba.

———. t. 3 — Pernambuco, Alagoas, Sergipe, Bahia.

———. t. 4 — Minas Gerais.

———. t. 5 — Espírito Santo, Rio de Janeiro.

———. t. 6 — São Paulo.

———. t. 7 — Paraná.

———. t. 8 — Santa Catarina, Rio Grande do Sul.

———. t. 9 — Região Centro-Oeste.

———. t. 10 — Brasil.

Brasil; síntese de dados 1981. Rio de Janeiro, 1982. n. p.

CENSO DEMOGRÁFICO — DADOS DISTRITAIS 1980. Rio de Janeiro, 1982. n. 1. Rondônia, Roraima, Amapá.

———. n. 2 — Acre.

———. n. 6 — Piauí.

———. n. 8 — Rio Grande do Norte.

———. n. 9 — Paraíba.

———. n. 11 — Alagoas.

———. n. 15 — Espírito Santo.

———. n. 17 — São Paulo.

* Preparado no Departamento de Documentação e Referência da Biblioteca Central do IBGE por Hesperia Zuma de Rosso.

———. n. 20 — Rio Grande do Sul.

———. n. 22 — Mato Grosso.

CENSO INDUSTRIAL — 1975. Rio de Janeiro, 1982. v. 2, pt. 2 — Brasil; produção física.

Dracena, SP. Rio de Janeiro, 1982. 24 p. (Coleção de monografias, 616).

Emprego, subemprego e desemprego — Rio de Janeiro 1981. Rio de Janeiro, 1982. 64 p.

ESTATÍSTICAS DA SAÚDE; ASSISTÊNCIA MÉDICO-SANITÁRIA — 1979. Rio de Janeiro, 1982. v. 4.

ESTATÍSTICAS ECONÔMICAS DO GOVERNO ESTADUAL E MUNICIPAL; atividade empresarial, estadual e municipal — 1975. Rio de Janeiro, 1982. v. 1, t. 4.

———. 1976. Rio de Janeiro, 1982. v. 2, t. 4.

———. 1977. Rio de Janeiro, 1982. v. 3, t. 4.

———; balanços municipais, versão analítica da despesa realizada — 1976. Rio de Janeiro, 1982. v. 2, t. 3.

———; balanços estaduais — 1976. Rio de Janeiro, 1982. v. 2, t. 2.

Flórida Paulista, SP. Rio de Janeiro, 1982. 20 p. (Coleção monografias, 617).

Força do trabalho no Brasil: uma análise de mobilidade ocupacional. Rio de Janeiro, 1982. 79 p. (Série estudos e pesquisas, 8).

Itapiranga, SC. Rio de Janeiro, 1982. 20 p. (Coleção de monografias, 614).

Nomenclatura dos alimentos consumidos no Brasil. Rio de Janeiro, 1981. (Estudo nacional da despesa familiar, v. 3, t. 3, pt. 2 — Animais).

Passo Fundo, RS. Rio de Janeiro, 1982. 20 p. (Coleção de monografias, 618).

Perfil estatístico de crianças e mães no Brasil — aspectos nutricionais 1974/1975. Rio de Janeiro, 1982. 267 p.

——— — *características sócio-demográficas 1970/1977.* Rio de Janeiro, 1982. 424 p.

PESQUISA INDUSTRIAL — 1978; produção física. Rio de Janeiro, 1982. t. 6 — Brasil.

PRODUÇÃO DA PECUÁRIA MUNICIPAL — 1980. Rio de Janeiro, 1982. v. 8, t. 1 — Região Norte.

———. t. 2 — Região Nordeste.

———. t. 3 — Região Sudeste.

———. t. 4 — Regiões Sul e Centro-Oeste.

———. t. 5 — Brasil.

PRODUÇÃO EXTRATIVA VEGETAL — 1980; Brasil, Grandes Regiões, Unidades da Federação, Mesorregiões, Microrregiões Homogêneas, Municípios. Rio de Janeiro, 1982. v. 8.

São Leopoldo ,RS. Rio de Janeiro, 1982. 24 p. (Coleção de monografias, 619).

Serro, MG. Rio de Janeiro, 1982. 11 p. (Coleção de monografias, 613).

Sinopse estatística da Região Norte. Rio de Janeiro, 1982. 241 p.

Sinopse estatística da Região Sul. Rio de Janeiro, 1982. 248 p.

SINOPSE PRELIMINAR DO CENSO AGROPECUÁRIO — 1980. Rio de Janeiro, 1982. n. 1 — Brasil.

———. n. 2 — Roraima, Rondônia, Amapá.

———. n. 3 — Acre, Amazonas, Pará.

———. n. 4 — Maranhão, Piauí.

———. n. 5 — Ceará, Rio Grande do Norte.

———. n. 9 — Minas Gerais.

———. n. 10 — Espírito Santo, Rio de Janeiro.

———. n. 11 — São Paulo.

———. n. 12 — Paraná, Santa Catarina.

———. n. 13 — Rio Grande do Sul.

———. n. 14 — Mato Grosso do Sul, Mato Grosso, Goiás, Distrito Federal.

Tabelas de composição de alimentos. 2.ª ed. Rio de Janeiro, 1981. 216 p. (Estudo nacional da despesa familiar, v. 3, t. 1).